Cloud Analytics mit Microsoft Azure

Entwickeln moderner Data Warehouses mit der kombinierten Leistung von Analytics und Azure

Has Altaiar, Jack Lee und Michael Peña

Packt
www.packt.com

Cloud Analytics mit Microsoft Azure

Entwickeln moderner Data Warehouses mit der kombinierten Leistung von Analytics und Azure

Has Altaiar
Jack Lee
Michael Peña

Packt›

Cloud Analytics mit Microsoft Azure

Autoren: Has Altair, Jack Lee und Michael Peña

Technische Lektoren: Aram Golbaghikoukia und Aaditya Pokkunuri

Leitender Editor: Aditya Datar

Akquisitions-Editor: Alicia Wooding

Produktions-Editor: Deepak Chavan

Redaktionsteam: Sonali Anubhavne, Vishal Bodwani, Ewan Buckingham, Megan Carlisle, Alex Mazonowicz und Jonathan Wray

Erste Veröffentlichung: Oktober 2019

Produktionsreferenz: 2071119

ISBN: 978-1-83921-640-4

Veröffentlicht von Packt Publishing Ltd.

Livery Place, 35 Livery Street

Birmingham B3 2PB, UK

Table of Contents

Vorwort

Über

In diesem Abschnitt werden der Autor, die vom Kurs abgedeckten Themen, die für den Einstieg benötigten technischen Fähigkeiten sowie die für den Abschluss aller enthaltenen Aktivitäten und Übungen benötigten Hardware- und Softwarekomponenten kurz vorgestellt.

Über Cloud Analytics mit Microsoft Azure

Da Daten mit exponentieller Geschwindigkeit erzeugt werden, migrieren Unternehmen weltweit ihre Infrastruktur zur Cloud. Die Anwendungsverwaltung wird sehr viel einfacher, wenn Sie eine Cloud Platform verwenden, um Ihre Dienste und Anwendungen zu entwickeln, zu verwalten und bereitzustellen.

Cloud Analytics mit Microsoft Azure bietet alles, was Sie benötigen, um anhand Ihrer Daten nützliche Insights zu gewinnen. Sie werden mit Big-Data-Analytics, dem Internet der Dinge (Internet of Things, IoT), Machine Learning, künstlicher Intelligenz und DataOps entdecken, was Daten leisten können. Darüber hinaus werden Sie sich eingehend mit Daten-Analytics befassen und Anwendungsfälle untersuchen, die sich auf den Gewinn verwertbarer Insights aus Fast-Echtzeitdaten konzentrieren. Im weiteren Verlauf werden Sie erfahren, wie Sie mithilfe von Machine Learning und Deep Learning eine durchgängige Analytics-Pipeline in der Cloud entwickeln.

Wenn Sie am Ende dieses Buchs angelangt sind, werden Sie über fundierte Kenntnisse der Daten-Analytics mit Azure und ihrer praktischen Umsetzung verfügen.

Über die Autoren

Has Altaiar ist im Herzen Softwareentwickler und von Beruf Berater. Has lebt im australischen Melbourne und ist Executive Director bei vNEXT Solutions. Bei seiner Arbeit konzentriert er sich auf Daten, IoT und KI auf Microsoft Azure. Für zwei seiner neuesten IoT-Projekte erhielt er mehrere Auszeichnungen. Has ist darüber hinaus Microsoft Azure MVP und regelmäßiger Organisator und Referent auf lokalen und internationalen Konferenzen, darunter Microsoft Ignite, NDC und ServerlessDays. Er ist auch Vorstandsmitglied von Global AI Community (http://globalai.community). Sie können den Blog von Has lesen oder ihm auf Twitter unter **@hasaltaiar** folgen.

Jack Lee ist ein erfahrener Azure-zertifizierter Berater und Azure Practice Lead mit einer Leidenschaft für Softwareentwicklung, Cloud und DevOps-Innovationen. Er trägt aktiv zur technischen Microsoft-Community bei und hat für verschiedene Benutzergruppen und Konferenzen Vorträge gehalten, einschließlich des Global Azure Bootcamp von Microsoft Canada. Jack ist ein erfahrener Mentor und Schiedsrichter bei Hackathons und auch der Präsident einer Benutzergruppe, die sich auf Azure, DevOps und Softwareentwicklung konzentriert. Jack wurde für seine Beiträge zur technischen Community als Microsoft MVP ausgezeichnet. Sie können Jack auf Twitter unter **@jlee_Consulting** folgen.

Michael Peña ist ein erfahrener technischer Berater und lebt im australischen Sydney. Er ist Microsoft MVP und zertifizierter Experte mit mehr als 10 Jahren Erfahrung in den Bereichen Daten, mobile Anwendungen, Cloud, Web und DevOps. Während dieser Jahre übte er verschiedene Funktionen aus, betrachtete sich jedoch im Herzen als Entwickler. Er ist auch ein internationaler Referent, der bei zahlreichen Veranstaltungen wie Microsoft Ignite, NDC, DDD, Cross-Platform Summit und verschiedenen persönlichen und virtuellen Treffen Vorträge gehalten hat. Michael war Praktikant bei Microsoft und ist auch ein Microsoft Student Partner Alumnus. Sie können ihm auf Twitter unter **@mjtpena** folgen.

Lernziele

Am Ende dieses Kurses werden Sie zu Folgendem in der Lage sein:

- Entdecken der Konzepte moderner Data Warehouses und Datenpipelines
- Ermitteln verschiedener Entwurfsüberlegungen während der Anwendung einer Cloud Analytics-Lösung
- Entwickeln einer umfassenden Analytics-Pipeline in der Cloud
- Unterscheiden von strukturierten, semistrukturierten und unstrukturierten Daten
- Auswählen eines cloudbasierten Diensts für Daten-Analytics-Lösungen
- Verwenden der Azure-Dienste zum Erfassen, Speichern und Analysieren von Daten jeglichen Umfangs

Publikum

Wenn Sie die Einführung des Cloud Analytics-Modells für Ihr Unternehmen planen, wird Ihnen dieses Buch helfen, die Designs und geschäftlichen Überlegungen zu verstehen, die Sie berücksichtigen müssen. Auch wenn dies nicht notwendig ist, wird Ihnen ein grundlegendes Verständnis von Daten-Analytics-Konzepten wie Datenstreaming, Datentypen, Machine Learning-Lebenszyklus und Docker-Containern helfen, optimalen Nutzen aus diesem Buch zu ziehen.

Ansatz

„Cloud Analytics mit Microsoft Azure“ erklärt komplexe Konzepte auf leicht verständliche Weise. Das Buch enthält auch mehrere Quick-Start-Anleitungen, die wichtige Konzepte behandeln und praktische Erfahrungen ermöglichen.

Hardwareanforderungen

Um Teilnehmern eine optimale Erfahrung zu bieten, sollten Sie folgende Hardwarekonfiguration verwenden:

- Prozessor: Intel Core i5 oder entsprechend
- Arbeitsspeicher (RAM): mindestens 1 GB verfügbar, 1,5 GB oder mehr empfohlen
- Anzeige: mindestens 1440 x 900 oder 1600 x 900 (16:9) empfohlen
- CPU: x86- oder x64-Bit-Prozessor mit 1 Gigahertz (GHz) oder schneller empfohlen

Softwareanforderungen

Sie sollten darüber hinaus die folgenden Softwareanwendungen im Voraus installieren:

- Windows 7/Windows Server 2008 R2 oder höher
- .NET 4.5
- Internet Explorer 10 oder höher

Konventionen

Codewörter im Text, Namen von Datenbanktabellen, Ordnernamen, Dateinamen, Dateierweiterungen, Pfadnamen, Dummy-URLs, Benutzereingaben und Twitter-Handles werden wie folgt formatiert:

„Mit dem folgenden Codeausschnitt wird eine Tabelle namens **`TweetsStream`** in Azure Synapse Analytics erstellt, die den Stream empfangen soll. Diese Tabelle umfasst zwei einfache Spalten: eine für den **`Zeitstempel`** und eine für den Wert, der aus dem Datenstrom empfangen wird. Im folgenden Beispiel wird **`ROUND_ROBIN`** dieser Tabelle als Verteilungsrichtlinie zugewiesen.“

Ein Codeblock wird wie folgt formatiert:

```
CREATE TABLE [dbo].[TweetsStream]
(
  [timestamp] DATETIME NULL,
  [Value] BIGINT NULL
)
WITH
(
  DISTRIBUTION = ROUND_ROBIN,
  CLUSTERED INDEX ([timestamp])
)
```

Installation und Einrichtung

Sie können Power BI Desktop installieren und interaktive Berichte erstellen.

Steigen Sie jetzt bei Cloud-Analytics ein.

1

Einführung in Analytics auf Azure

Bei einer Umfrage von Dresner Advisory Service 2019 gaben 48 % der befragten Organisationen – so viele wie nie zuvor – an, dass Business Intelligence in der Cloud für ihre Geschäftstätigkeit entscheidend oder sehr wichtig ist. Darüber hinaus ergab die Umfrage mit dem Titel *Cloud Computing and Business Intelligence Market Study*, dass Vertriebs- und Marketingteams am meisten von Analytics profitieren.

Im Zuge des Unternehmenswachstums generieren sie tagtäglich riesige Datenmengen. Diese Daten stammen aus verschiedenen Quellen wie beispielsweise Mobiltelefonen, Sensoren im **Internet der Dinge** (**Internet of Things, IoT**) und verschiedenen **Software-as-a-Service**(**SAAS**)-Produkten wie beispielsweise **CRM** (**Customer Relationship Manager, Kundenbeziehungsmanager**). Unternehmen müssen ihre Datenarchitektur und -infrastruktur skalieren und modernisieren, um der Nachfrage gerecht zu werden und in ihrer Branche wettbewerbsfähig zu bleiben.

Die entscheidende Strategie zur Verwirklichung dieses Wachstums besteht in Analytics-Funktionen im Cloudmaßstab. Wenn Sie die Vorteile der Cloud nutzen, können Sie den Anwendern leichteren Zugang zu Ihren Unternehmen ermöglichen und müssen kein eigenes Rechenzentrum verwalten. Mithilfe eines Clouddienstanbieters wie Microsoft Azure können Sie Ihre Verfahren für Daten-Analytics ohne die Einschränkungen Ihrer IT-Infrastruktur beschleunigen. Die Zeiten haben sich geändert, was die Verwaltung von IT-Infrastrukturen anbelangt, da **Data Lakes** und das Data Warehouse in der Cloud in der Lage sind, riesige Datenmengen zu speichern und zu verwalten.

Daten einfach nur zu erfassen, bringt keinen Mehrwert für Ihr Unternehmen mit sich. Sie müssen Insights daraus gewinnen und Ihrem Unternehmen mit Daten-Analytics zu Wachstum verhelfen. Azure ist nicht nur ein Hub zum Erfassen von Daten, sondern eine unschätzbare Ressource für Daten-Analytics. Daten-Analytics bieten Ihnen die Möglichkeit, bessere Kenntnisse Ihres Unternehmens und Ihrer Kunden zu erlangen. Durch Anwendung verschiedener Data-Science-Konzepte wie Machine Learning, Regressionsanalysen, Klassifikationsalgorithmen und Zeitreihenprognosen können Sie Ihre Hypothesen testen und datengesteuerte Entscheidungen für die Zukunft treffen. Dabei sind die Unternehmen jedoch ständig mit der schwierigen Frage konfrontiert, wie sie diese Möglichkeiten der analytischen Modellierung bei der Verarbeitung von Milliarden von Datenzeilen schnell nutzen können. Hier können ein modernes Data Warehouse und eine Datenpipeline eine große Hilfe sein (mehr dazu in den nächsten Abschnitten).

Daten-Analytics können in vielfacher Hinsicht zum Erfolg Ihres Unternehmens beitragen. Wenn Sie im Einzelhandel Ihre Kunden besser verstehen, haben Sie eine bessere Vorstellung davon, welche Produkte Sie wo, wann und wie verkaufen sollten. Im Finanzsektor unterstützen Daten-Analytics die Bekämpfung von Straftaten durch die Erkennung betrügerischer Transaktionen und die Bereitstellung fundierterer Risikobewertungen auf der Grundlage strafrechtlicher Informationen aus der Vergangenheit.

In diesem Kapitel geht es um grundsätzliche Aspekte des Potenzials von Daten. Dabei werden **Big Data Analytics**, **IoT**, **Machine Learning** (**ML**) und **künstliche Intelligenz** (**KI**) sowie **DataOps** behandelt. Sie erfahren auch, was Microsoft Azure zu einer perfekten Plattform für Analytics in der Cloud macht. Darüber hinaus werden Sie die grundlegenden Konzepte eines modernen Data Warehouse sowie von Datenpipelines erkunden.

Das Potenzial der Daten

Als Verbraucher haben Sie bereits erlebt, wie der Einzug der Daten unseren Alltag beeinflusst hat und es noch immer tut. Beliebte Unterhaltungsanwendungen wie YouTube bieten jetzt eine individuelle User Experience mit Features wie Videoempfehlungen auf der Grundlage unserer Interessen sowie Suchprotokollinformationen. Sie erkennen jetzt im Handumdrehen neue Inhalte, die Ihren bevorzugten Inhalten ähnlich sind. Auch neue, beliebte Trendinhalte sind ganz leicht auffindbar.

Durch die enorme Entwicklung im Bereich der tragbaren Technologie ist es auch möglich geworden, gesundheitliche Daten zu verfolgen, also beispielsweise Herzfrequenz, Blutdruck usw. zu überwachen. Die Geräte geben dann auf der Grundlage der erfassten statistischen Daten eine individuelle Empfehlung aus. Diese individuellen Gesundheitsstatistiken sind jedoch nur ein Beispiel für die weltweit tagtäglich erfolgende massive Datenerfassung – an der wir aktiv mitwirken.

Jeden Tag verwenden Millionen Menschen auf der ganzen Welt Social-Networking-Plattformen und Suchmaschinen. Internetgiganten wie Facebook, Instagram und Google nutzen Clickstream-Daten, um neue Innovationen zu entwickeln und ihre Dienste zu verbessern.

Eine umfangreiche Datensammlung erfolgt auch im Rahmen von Projekten wie dem *Great Elephant Census* oder *eBird*, die den Artenschutz fördern sollen. Selbst für Projekte zum Schutz der Tiger in Indien wurden datengesteuerte Techniken eingeführt.

Diese Techniken spielen sogar eine unschätzbare Rolle bei den weltweiten Bemühungen, Fakten zum Klimawandel, den Ursachen hierfür und den möglichen Reaktionen darauf zu sammeln, um die Temperatur der Meeresoberfläche nachvollziehen zu können, Naturkatastrophen wie Überflutungen von Küstenregionen sowie die Muster der Erderwärmung zu analysieren und so gemeinsam unser Ökosystem zu retten.

Organisationen wie Global Open Data for Agriculture and Nutrition (GODAN), die von Landwirten, Viehzüchtern und Verbrauchern gleichermaßen genutzt werden können, tragen ebenfalls zu dieser unaufhörlichen Datenerfassung bei.

Darüber hinaus unterstützt die Datenanalyse (wie dies auch beim Einzug der tragbaren Technologie der Fall war) bahnbrechende Fortschritte im Gesundheitswesen. Patientendatensätze werden analysiert, um Krankheitsmuster und frühe Symptome von Krankheiten zu erkennen und so bessere Lösungen für bekannte Gesundheitsprobleme zu finden.

Wir sprechen hier von riesigen Datenmengen – daher ist häufig von „Big Data“ die Rede, um die nutzbare Leistung dieses enormen Datenvolumens zu beschreiben.

> **Hinweis**
>
> Weitere Informationen hierzu finden Sie hier.

Big Data Analytics

Häufig wird der Begriff „Big Data“ verwendet, um riesige Datenmengen zu beschrieben, mit denen herkömmliche Tools nicht fertig werden. Big Data lassen sich mit fünf Vs charakterisieren:

- **Volume (Menge)**: Hier geht es um die Menge an Daten, die für Big Data Analytics analysiert werden müssen. Wir haben es heute mit größeren Datasets als je zuvor zu tun. Möglich wurde dies durch die Verfügbarkeit elektronischer Produkte wie mobiler Geräte und IoT-Sensoren, die weltweit vielfach für kommerzielle Zwecke eingesetzt werden.
- **Velocity (Geschwindigkeit)**: Dies bezieht sich auf das Tempo der Datengenerierung. Geräte und Plattformen wie die soeben erwähnten produzieren ständig in großem Maßstab und mit hoher Geschwindigkeit Daten. Dies macht eine schnelle Erfassung, Verarbeitung, Analyse und Bereitstellung der Daten notwendig.
- **Variety (Vielfalt)**: Hier geht es um die Struktur der generierten Daten. Die Datenquellen sind uneinheitlich. Manche sind strukturiert, andere nicht. (Auf den folgenden Seiten werden Sie mehr hierzu erfahren.)
- **Value (Wert)**: Dieser Aspekt betrifft den Wert der extrahierten Daten. Nicht immer sind verfügbare Daten auch wertvoll. Mit den richtigen Tools können Sie auf kostengünstige und skalierbare Weise einen Mehrwert aus den Daten erzielen.
- **Veracity (Richtigkeit)**: Hier geht es um die Qualität oder Vertrauenswürdigkeit von Daten. Ein unformatiertes Dataset wird in der Regel viele **Stördaten** und **Verzerrungen** aufweisen und muss bereinigt werden. Ein großes Dataset ist nicht hilfreich, wenn die meisten Daten nicht korrekt sind.

Big Data Analytics bezeichnet die Suche nach Mustern, Trends und Korrelationen in unstrukturierten Daten, um aussagekräftige Insights zu gewinnen, die geschäftliche Entscheidungen bestimmen werden. Diese unstrukturierten Daten weisen in der Regel eine hohe Dateigröße auf (z. B. Bilder, Videos und soziale Graphen).

Das heißt nicht, dass relationale Datenbanken für Big Data nicht relevant sind. Moderne Data-Warehouse-Plattformen wie Azure Synapse Analytics (früher Azure SQL Data Warehouse) unterstützen strukturierte und semistrukturierte Daten (z. B. JSON) und können beliebig skaliert werden, um Datenmengen in einer Größenordnung von mehreren Terabyte bis Petabyte zu unterstützen. Mit Microsoft Azure können Sie flexibel eine beliebige Plattform auswählen. Die Technologien können sich ergänzen, sodass eine stabile Daten-Analytics-Pipeline entsteht.

Einige erstklassige Anwendungsfälle von Big Data Analytics sind:

- **Social-Media-Analysen**: Auf Social-Media-Websites wie Twitter, Facebook und Instagram können Unternehmen erfahren, was Kunden über ihre Produkte und Dienstleistungen sagen. Social-Media-Analysen helfen Unternehmen, ihre Zielgruppen den Anwenderpräferenzen und Markttrends gemäß gezielt anzusprechen. Schwierigkeiten bereiten in diesem Zusammenhang die enorme Datenmenge und die Unstrukturiertheit von Tweets und Posts.
- **Betrugsprävention**: Hierbei handelt es sich um einen der bekanntesten Anwendungsfälle von Big Data. Eines der herausragenden Merkmale von Big Data Analytics in Zusammenhang mit Betrugsprävention ist die Möglichkeit, Anomalien in einem Dataset zu erkennen. Ein Beispiel hierfür ist die Validierung von Kreditkartentransaktionen durch Nachvollziehen von Transaktionsmustern wie Standortdaten und Kategorien gekaufter Artikel. Die größte Herausforderung besteht hierbei darin, sicherzustellen, dass die KI/ML-Modelle einwandfrei und unvoreingenommen sind. Es könnte sein, dass das Modell nur für einen bestimmten Parameter wie beispielsweise das Herkunftsland des Anwenders trainiert wurde. Das Modell konzentriert sich dann darauf, nur anhand des Standorts des Anwenders Muster zu erkennen, und lässt andere Parameter möglicherweise außer Acht.
- **Preisoptimierung**: Mithilfe von Big Data Analytics können Sie anhand historischer Marktdaten prognostizieren, bei welchen Preispunkten die besten Ergebnisse erzielt werden. So können Unternehmen dafür sorgen, dass sie den Preis für ihre Artikel nicht zu hoch oder zu niedrig ansetzen. Das Schwierige dabei ist, dass sich viele Faktoren auf die Preise auswirken können. Durch die Fokussierung auf einen bestimmten Faktor wie beispielsweise den Preis von Wettbewerbern könnte Ihr Modell letztendlich trainiert werden, nur diesen Bereich zu berücksichtigen, und andere Faktoren wie Wetter- und Verkehrsdaten vernachlässigen.

Mit Big Data für Unternehmen ist in der Regel das Konzept einer IoT-Infrastruktur verbunden, in der Hunderte, Tausende oder sogar Millionen von Geräten mit einem Netzwerk verbunden sind, das ständig Daten an einen Server sendet.

Internet der Dinge (Internet of Things, IoT)

Das IoT spielt eine entscheidende Rolle für eine über Ihre aktuellen Datenquellen hinausgehende Skalierung Ihrer Anwendung. Es handelt sich hier einfach um eine Vernetzung von Geräten, die zu einem einzigen Zweck in Objekte um uns herum eingebettet sind, um Daten zu senden und zu empfangen. Mithilfe des IoT können wir ständig mehr Daten über „Dinge“ erfassen, ohne sie manuell in einer Datenbank zu codieren.

Eine Smartwatch ist ein gutes Beispiel für ein IoT-Gerät, das die Vitalparameter Ihres Körpers ständig misst. Anstatt ein Messgerät zu erhalten und dieses in einem System zu codieren, können Sie die Daten mit einer Smartwatch automatisch aufzeichnen. Ein weiteres gutes Beispiel ist ein Gerätetracker für eine Anlage, der Informationen zu Standort, Temperatur und Feuchtigkeit erfasst. So können Logistikunternehmen ihre durchlaufenden Posten überwachen und auf diese Weise die Qualität und Effizienz ihrer Dienste gewährleisten.

In großem Maßstab generieren diese IoT-Geräte überall Daten in einer Größenordnung von mehreren Gigabyte bis Terabyte. Diese Daten werden in der Regel in unformatiertem, unstrukturiertem Format in einem Data Lake gespeichert und später analysiert, um daraus Business Insights zu gewinnen. Ein Data Lake ist ein zentrales Repository aller strukturierten, semistrukturierten und unstrukturierten Daten. In dem oben genannten Beispiel eines Logistikunternehmens könnten Muster (z. B. die besten Lieferwege) generiert werden. Mithilfe der Daten könnten auch Anomalien wie Datenlecks oder mutmaßliche betrügerische Aktivitäten nachvollzogen werden.

Machine Learning und künstliche Intelligenz

Mit zunehmender Datenmenge eröffnen sich den Unternehmen viele über das Verständnis von Geschäftstrends und Mustern hinausgehende Möglichkeiten. Machine Learning und künstliche Intelligenz sind Beispiele für Innovationen, die Sie mit Ihren Daten nutzen können. Funktionen der künstlichen Intelligenz und Machine Learning zu entwickeln, ist jetzt aufgrund der Verfügbarkeit von Technologien und der Möglichkeit, Speicher und Computing in der Cloud zu skalieren, relativ einfach.

Die Begriffe ML und KI werden oft miteinander verwechselt. Kurz gesagt: Machine Learning ist eine Unterform (oder Anwendung) künstlicher Intelligenz. Bei Machine Learning geht es darum, Systemen die Möglichkeit zu bieten, aus früheren Datasets zu lernen und ohne menschliche Unterstützung automatisch Anpassungen vorzunehmen. Zu diesem Zweck stehen verschiedene Algorithmen zur Verfügung, die auf das Dataset angewendet werden. Die Algorithmen analysieren die Daten nahezu in Echtzeit und schlagen dann mögliche Aktionen vor, basierend auf der aus früheren Erfahrungen abgeleiteten Genauigkeit oder Zuverlässigkeit.

Das Wort „Learning“ besagt, dass das Programm ständig aus den ihm zugeführten Daten lernt. Ziel von Machine Learning ist Genauigkeit, nicht so sehr Erfolg. Es gibt zwei Hauptkategorien von Machine-Learning-Algorithmen: **überwachte** und **unüberwachte** Algorithmen.

Überwachte Machine-Learning-Algorithmen erstellen eine Zuordnungsfunktion für den Abgleich der Eingabevariablen mit der Ausgabevariablen. Der Algorithmus trainiert sich selbst mithilfe der vorhandenen Datasets, um die Ausgabe zu prognostizieren. Klassifizierung ist eine Form des überwachten ML, die in Anwendungen wie der Bildkategorisierung oder Kundensegmentierung verwendet werden kann und für gezielte Marketingkampagnen zum Einsatz kommt.

Unüberwachtes Machine Learning findet andererseits statt, wenn Sie das Programm ohne irgendwelche Kennzeichnungen eigenständig ein Muster finden lassen. Ein gutes Beispiel hierfür ist die Analyse der Kaufmuster von Kunden beim Produktkauf. Sie erhalten inhärente Gruppierungen (**Clustering**) gemäß dem Kaufverhalten, und das Programm kann Kunden und Produkte nach Kaufmustern zuordnen. So könnten Sie beispielsweise feststellen, dass Kunden, die Produkt A kaufen, auch gerne Produkt B kaufen. Dies ist ein Beispiel für einen Algorithmus für anwenderbasierte Empfehlungen sowie eine marktbasierte Analyse. Für die Anwender würde dies letztendlich heißen, dass sie beim Kauf eines bestimmten Artikels wie z. B.eines Buchs auch angeregt werden, andere Bücher derselben Serie, desselben Genres oder derselben Kategorie zu kaufen.

Künstliche Intelligenz geht über die Möglichkeiten von Machine Learning hinaus. Hier geht es darum, Entscheidungen zu treffen. Es kommt auf den Erfolg, weniger auf Genauigkeit an. Sie können sich das so vorstellen: Ziel von Machine Learning ist es, Wissen zu erwerben, während es bei künstlicher Intelligenz um Weisheit oder Intelligenz geht. Ein Beispiel für KI in der Praxis wäre etwa der Atlas-Roboter von Boston Dynamic, der sich frei in der offenen Welt bewegen und Hindernisse ohne menschliche Steuerung vermeiden kann. Der Roboter ist für seine Bewegungen nicht vollständig auf die historischen Kartendaten angewiesen. Bei Machine Learning dagegen geht es darum, aus der Analyse historischer Daten ein Muster zu erstellen oder zu prognostizieren. Ähnlich wie bei der Navigation des Roboters handelt es sich darum, den optimalen Weg zu finden, indem anhand historischer und Crowdsourcing-Verkehrsdaten Muster erstellt werden.

Die Einrichtung eines modernen Data Warehouse mit Cloud Analytics ist entscheidend bei der Vorbereitung der ML-/KI-Ausführung. Ohne Migration der Workloads zur Cloud werden bei der Entwicklung von ML-/KI-Modellen verschiedene Hindernisse auftreten, wenn es darum geht, den geschäftlichen Nutzen dieser neuen Technologien zu maximieren. Ein modernes Data Warehouse und eine Analytics-Pipeline bilden das Rückgrat, das dies möglich macht.

Microsoft ist ein führender Anbieter von Machine Learning und künstlicher Intelligenz und hat in seinen Produkten und Tools viele Innovationen gefördert, wie z. B. die digitale Assistentin von Windows, Cortana, sowie die Live-Untertitel und Bildunterschriften von Office 365. Das Angebot umfasst eine Reihe von Produkten, Tools und Diensten wie Microsoft Cognitive Services, ML Studio, Azure Machine Learning Service und ML.NET.

Mit seiner Initiative *AI for Good* setzt Microsoft ein Zeichen. Ziel der Initiative ist es, die Welt mit KI nachhaltiger und zugänglicher zu machen. Besonders interessant ist das Projekt *AI for Snow Leopards*, bei dem Microsoft KI-Technologie zum Aufspüren von Schneeleoparden (die im Schnee fast unsichtbar sind) einsetzt, um diese gefährdete Tierart zu schützen.

Die Erforschung von künstlicher Intelligenz und Deep Learning, insbesondere die Data-Science- und Formelaspekte, sind nicht Schwerpunktthema dieses Buchs. Sie werden sich jedoch in späteren Kapiteln gelegentlich mit einigen Konzepten befassen (weitere Informationen hierzu in *Kapitel 3, Verarbeitung und Visualisierung von Daten*).

DataOps

Für eine effiziente und agile Implementierung von Daten-Analytics in Ihrem Unternehmen benötigen Sie die richtige Kultur und die entsprechenden Prozesse. Hier kommt das **DataOps**-Konzept ins Spiel. Mithilfe von DataOps lässt sich die Koordinationsbarriere zwischen den Datenteams (Analysten, Dateningenieure und Datenwissenschaftler) und den Betriebsteams (Administratoren und Betriebsleiter) aus dem Weg räumen, um schnelle und präzise Daten-Analytics zu ermöglichen.

Bei DataOps geht es um eine Kultur der Zusammenarbeit zwischen verschiedenen Rollen und Funktionen. Datenwissenschaftler haben Zugang zu Echtzeitdaten, die sie erkunden, vorbereiten und bereitstellen. Automatisierte Prozesse und Abläufe sind für diese Zusammenarbeit zwischen Analysten und Entwicklern von unschätzbarem Wert, da sie mithilfe von Visualisierungstools einfachen Zugriff auf diese Daten bieten. Die relevanten Daten sollten den Anwendern über Webanwendungen oder mobile Anwendungen bereitgestellt werden. Dies ist gewöhnlich über eine **Anwendungsprogrammierschnittstelle** (**API**) möglich. Für CEOs bedeutet DataOps eine schnellere Entscheidungsfindung, da sie ihr Unternehmen auf hoher Ebene überwachen können, ohne auf die Berichte der Teamleiter warten zu müssen. Die folgende Abbildung soll die Idee einer DataOps-Kultur der Zusammenarbeit veranschaulichen:

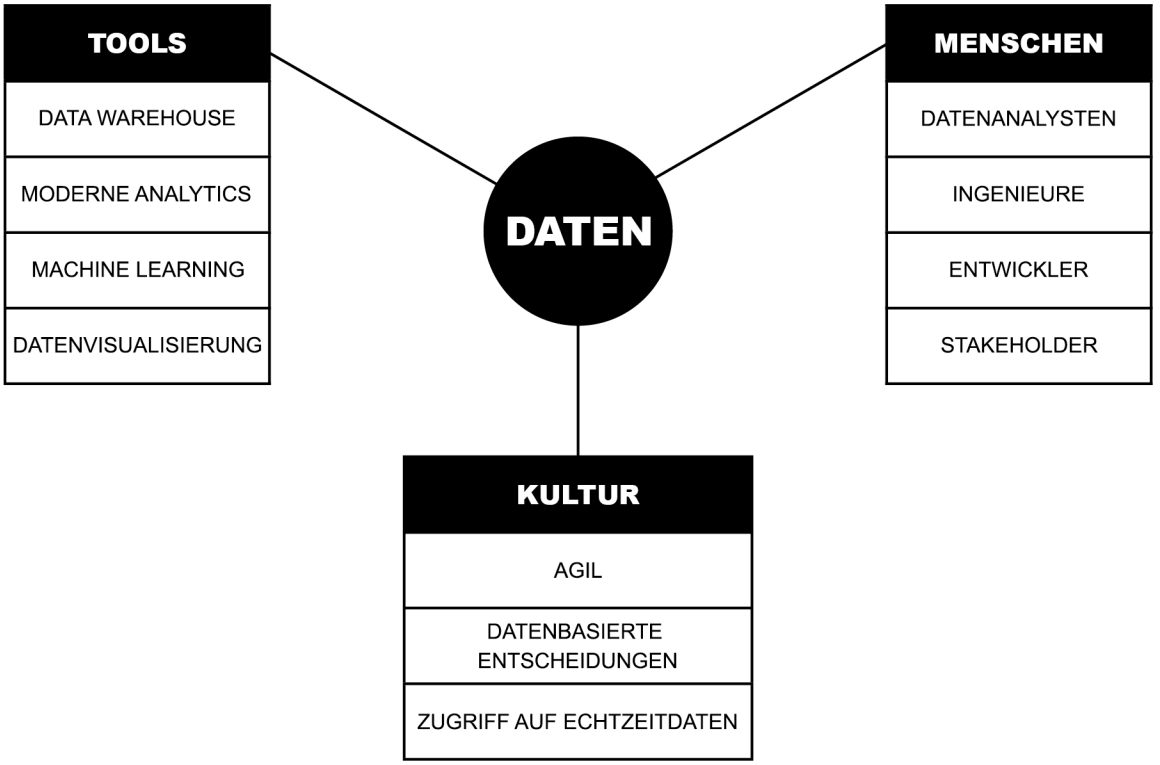

Abbildung 1.1: DataOps-Prozess

Wenn ein Team die gewünschte Geschwindigkeit und Präzision beim Testen seiner Hypothesen (z. B. der Wahrscheinlichkeit, dass jemand aufgrund seiner Eigenschaften und seines Verhaltens ein Produkt kauft) erreicht hat, kann es bessere Insights gewinnen. Bessere Insights ermöglichen besser umsetzbare und angemessenere Entscheidungspunkte für die Stakeholder des Unternehmens und damit eine Minimierung der Risiken und eine Maximierung der Gewinne.

Warum Microsoft Azure?

Microsoft Azure ist eine Gruppe von Cloud Computing-Diensten auf Unternehmensniveau, die von Microsoft über eigene verwaltete Rechenzentren angeboten werden. Azure ist die einzige Cloud mit einer wirklichen End-to-End-Analytics-Lösung. Mit Azure können Analysten in Sekundenschnelle Insights aus allen Unternehmensdaten gewinnen. Azure bietet einen ausgereiften, soliden Datenfluss ohne Einschränkungen in puncto Nebenläufigkeit.

Azure unterstützt **Infrastructure-as-a-Service** (**IAAS**), **Platform-as-a-Service** (**PAAS**) und **SAAS**. Viele staatliche Einrichtungen weltweit sowie 95 % der Fortune 500-Unternehmen nutzen Azure, von Branchen wie dem Gesundheitswesen und Finanzdienstleistungen bis hin zu Einzelhandel und Fertigung.

Microsoft ist ein Technologiekonzern, der es seit Jahrzehnten vielen Menschen ermöglicht, mit ihrer Software, ihren Tools und Plattformen mit weniger Aufwand mehr zu erreichen. Azure bietet Flexibilität. Bekannte Tools und Infrastrukturen von Microsoft (z. B. SQL Server, Windows Server, **Internetinformationsdienste** (**Internet Information Services, IIS**) und .NET) oder Tools wie MySQL, Linux, PHP, Python, Java oder andere Open Source Technologien können alle in der Azure-Cloud ausgeführt werden. Die Zeiten, in denen Sie nur mit einer geschlossenen Gruppe von Tools und Technologien arbeiten konnten, sind vorbei.

Azure bietet Ihnen je nach Bedarf verschiedene Produkte und Dienste. Sie können alles nach Maß ausführen, von der Verwaltung Ihrer IaaS unter Einsatz von Windows Server Virtual Machines mit installiertem Enterprise SQL Server bis hin zur Verwendung eines verwalteten PaaS-Angebots wie Azure Synapse Analytics (mehr hierzu in *Kapitel 2, Entwicklung Ihres modernen Data Warehouse*).

Die folgende Abbildung zeigt das breite Spektrum an datenspezifischen Azure-Tools und -Diensten, die zum Erstellen von End-to-End-Datenpipelines verwendet werden können:

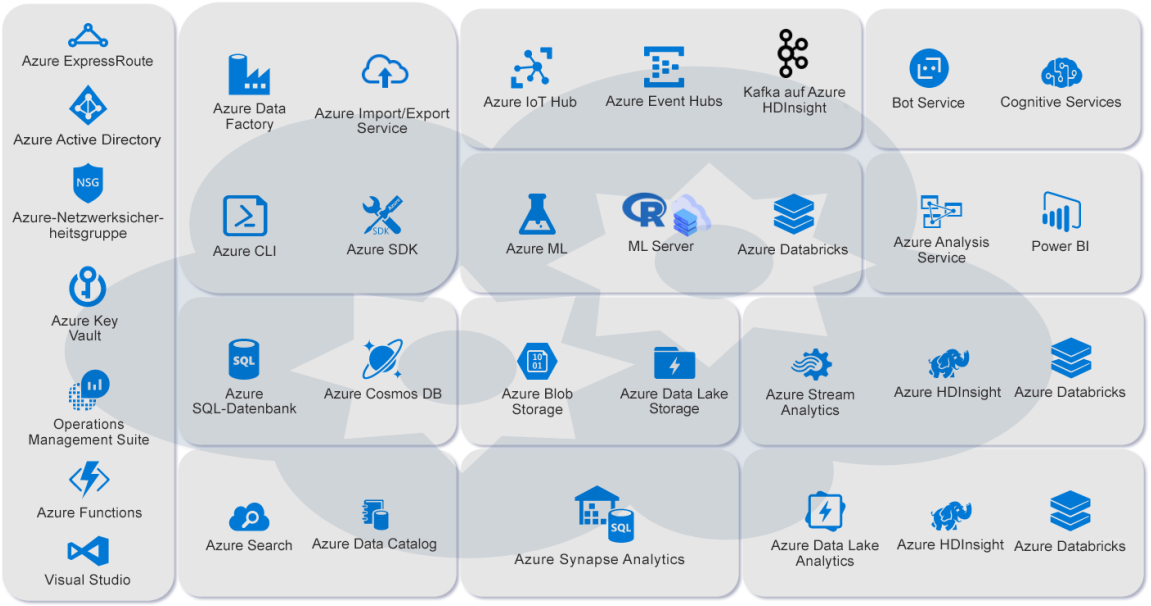

Abbildung 1.2: Datenbezogene Microsoft Azure-Dienste

Azure bietet Ihnen die Flexibilität, den besten Ansatz zur selbstständigen Lösung Ihres Problems zu wählen. Sie sind nicht mehr gezwungen, ein weniger anpassungsfähiges Produkt zu „verbiegen“, damit es eine unnatürliche Funktion ausführen kann. Sie sind auch nicht nur auf SQL Server beschränkt. Sie können zudem flexibel andere Arten von Datenbanken oder Speicher wählen, sei es über einen Dienst, der auf einem Linux-Server oder einer Containerlösung installiert ist, oder über eine verwaltete Plattform (wie z. B. Azure Cosmos DB für Ihre Cassandra- und MongoDB-Instanzen). Dies ist sehr wichtig, da in der Praxis verschiedene Szenarien unterschiedliche Lösungen, Tools und Produkte erfordern.

Microsoft Azure bietet Ihnen eine End-to-End-Plattform, von Azure Active Directory für die Verwaltung der Anwenderidentität und des Zugriffs bis hin zu Azure IoT-Angeboten (wie z. B. IoT Hub) zur Erfassung von Daten von Hunderten, ja Tausenden von IoT-Geräten. Auch Dienste wie Entwicklungstools und Cloud Hosting-Optionen, mit denen Ihre Entwickler nach modernsten Methoden arbeiten, sowie verschiedene Analytics- und Machine-Learning-Tools, mit denen Datenwissenschaftler, Dateningenieure und Datenanalysten produktiver arbeiten können, sind verfügbar (mehr hierzu in *Kapitel 3, Verarbeitung und Visualisierung von Daten*).

Das gesamte Spektrum der Azure-Dienste ist zu umfangreich, um hier behandelt zu werden. Daher konzentriert sich dieses Buch auf die zentrale Suite von Data Warehousing- und Business Intelligence-Produkten: Azure Data Factory, Azure Data Lake, Azure Synapse Analytics, Azure Databricks, Azure Analysis Services, Power BI und Azure Machine Learning (siehe *Kapitel 2*, *Entwicklung Ihres modernen Data Warehouse*, und *Kapitel 3*, *Verarbeitung und Visualisierung von Daten*).

Sicherheit

Für Microsoft hat Sicherheit höchste Priorität. Im Hinblick auf Daten sind Privatsphäre und Sicherheit nicht verhandelbar. Es wird immer Bedrohungen geben. Azure verfügt über fortschrittlichste Sicherheits- und Datenschutzfunktionen im Analytics-Bereich. Azure-Dienste unterstützen den Datenschutz mithilfe von **Virtual Networks** (**VNets**). Obwohl sich die Datenpunkte in der Cloud befinden, sind sie somit nicht über das öffentliche Internet zugänglich. Nur die Anwender in demselben VNet können miteinander kommunizieren. Für Webanwendungen erhalten Sie eine **Web Application Firewall** (**WAF**), die von Azure Application Gateway bereitgestellt wird. Diese sorgt dafür, dass nur gültige Anfragen in Ihr Netzwerk gelangen können.

Mit rollenbasierter Zugriffssteuerung (**Autorisierung**) können Sie sicherstellen, dass nur die Anwender mit den entsprechenden Rollen, also beispielsweise Administratoren, Zugriff auf bestimmte Komponenten und die Funktionen verschiedener Ressourcen haben. Mit der **Authentifizierung** wird dagegen sichergestellt, dass Sie nicht auf eine Ressource zugreifen können, wenn Sie nicht über die entsprechenden Anmeldeinformationen (z. B. Kennwörter) verfügen. Autorisierung und Authentifizierung sind mithilfe von Azure Active Directory in verschiedene Dienste und Komponenten von Microsoft Azure integriert.

Azure bietet auch einen Dienst namens **Azure Key Vault**. Mit Key Vault können Sie geheime Schlüssel und Kennwörter sicher speichern und verwalten, Verschlüsselungsschlüssel erstellen und Zertifikate verwalten, sodass die Anwendungen keinen direkten Zugriff auf private Schlüssel haben. Wenn Sie dieses Muster mit Key Vault befolgen, müssen Sie Ihre geheimen Schlüssel und Kennwörter nicht in Ihrem Quellcode und Skriptrepository hartcodieren.

Azure Synapse Analytics verwendet ML und KI zum Schutz Ihrer Daten. In Azure SQL bietet Microsoft erweiterte Datensicherheit, um den Schutz Ihrer Daten zu gewährleisten. Hierzu gehört auch, zu erkennen, ob Ihre Datenbank Sicherheitslücken aufweist, z. B. öffentlich verfügbare Portnummern. Diese Funktionen ermöglichen ferner eine bessere Konformität mit verschiedenen Standards, wie z. B. der **DSGVO** (**Datenschutz-Grundverordnung**), indem sie eine Klassifizierung der Kundendaten sicherstellen, die als vertraulich eingestuft werden. Für Azure SQL wurden kürzlich zudem neue Features – **Sicherheit auf Zeilenebene** (**Row-Level Security, RLS**) und **Sicherheit auf Spaltenebene** (**Column-Level Security, CLS**) – angekündigt, die den Zugriff auf die Zeilen und Spalten einer Datenbanktabelle basierend auf den Anwendereigenschaften steuern sollen.

Microsoft investiert jedes Jahr mindestens 1 Milliarde US-Dollar in den Bereich für Cybersicherheit, einschließlich der Azure-Plattform. Azure verfügt über verschiedene Referenzen und Auszeichnungen von unabhängigen Bewertungsstellen – ein Beweis, dass Sie in Bezug auf alle Sicherheitsaspekte auf Azure bauen können – von der physischen Sicherheit (kein physischer Zugang zu Rechenzentren für nicht autorisierte Anwender) bis hin zur Sicherheit auf Anwendungsebene.

Dies sind einige Sicherheitsfeatures, die Sie berücksichtigen müssen, wenn Sie Ihr eigenes Rechenzentrum unterhalten.

Cloud Scale

Azure hat mit kostengünstigen Daten-Analytics eine Veränderung der Branche bewirkt. Vor der großflächigen Einführung von Cloud Computing mussten Sie angemessen planen und dafür sorgen, dass Sie über das erforderliche Kapital verfügten, wenn Sie Daten-Analytics für mehrere Terabyte oder sogar Petabyte Daten durchführen wollten. Dies bedeutete erhebliche Vorabkosten für die Infrastruktur und professionelle Dienstleistungen, um überhaupt loslegen zu können. Mit Azure dagegen können Sie klein anfangen (für viele der Dienste gibt es kostenfreie Tarife). Sie können Ihre Cloud-Ressourcen problemlos innerhalb von Minuten vertikal und horizontal hoch- und herunterskalieren. Azure hat Skalierbarkeit demokratisiert und für jeden wirtschaftlich tragbar und zugänglich gemacht, wie in *Abbildung 1.3* dargestellt:

Abbildung 1.3: Microsoft Azure-Regionen

Zurzeit gibt es 54 Microsoft Azure-Regionen, die 140 Länder unterstützen. Bei manchen Unternehmen und in manchen Branchen müssen die Daten in dem Land gehostet werden, in dem die Geschäftstätigkeit erfolgt. Angesichts der Verfügbarkeit verschiedener Rechenzentren weltweit ist es einfach für Sie, in andere Regionen zu expandieren. Dieses Konzept mehrerer Regionen ist auch von Vorteil, wenn es um die hohe Verfügbarkeit Ihrer Anwendungen geht.

Die wahre Stärke der Cloud ist ihre Elastizität. So können Sie Ressourcen nicht nur hochskalieren, sondern bei Bedarf auch wieder herunterskalieren. Im Bereich Data Science ist dies sehr hilfreich, da Data Science mit variablen Workloads verbunden ist. Wenn Datenwissenschaftler und -ingenieure Datasets analysieren, müssen beispielsweise mehr Berechnungen durchgeführt werden. Azure macht mithilfe von Databricks (mehr hierzu in *Kapitel 2*, *Entwicklung Ihres modernen Data Warehouse*) eine Skalierung nach Bedarf möglich. In Zeiten mit geringerer Auslastung (z. B. von 19 Uhr bis 7 Uhr an Wochentagen und an Wochenenden), wenn die Datenwissenschaftler und -ingenieure keine Verarbeitungsleistung für Datenanalysen benötigen, können Sie Ihre Ressourcen herunterskalieren. Sie müssen dann nicht rund um die Uhr für aktive Ressourcen zahlen. Databricks bietet im Grunde einen Dienst mit nutzungsbasierter Bezahlung („Pay-as-you-go“ oder „Pay-what-you-use“).

Azure bietet zudem ein **Service Level Agreement** (**SLA**) für seine Dienste als Verpflichtung, die Betriebszeit und Konnektivität für seine Produktionskunden zu gewährleisten. Wenn es zu Ausfallzeiten oder einem Zwischenfall kommt, werden Servicegutschriften (Rabatte) auf die betroffenen Ressourcen angewendet. So haben Sie die Gewissheit, dass Ihre Anwendung immer verfügbar sein wird, bei minimalen Ausfallzeiten.

Microsoft Azure bietet verschiedene Skalierungsansätze und -muster:

- **Vertikale Skalierung**: Hier werden derselben Instanz (Server oder Dienst) mehr Ressourcen hinzugefügt. Ein Beispiel hierfür ist die Hochskalierung einer virtuellen Maschine von 4 GB RAM auf 16 GB RAM. Es handelt sich hier um einen einfachen und unkomplizierten Ansatz für die Skalierung Ihrer Anwendung. Es gibt jedoch eine technische Obergrenze für die Skalierung einer Instanz, außerdem ist dies der teuerste Skalierungsansatz.

- **Horizontale Skalierung**: Hier stellen Sie Ihre Anwendung in mehreren Instanzen bereit. Dies würde folgerichtig eine unendliche Skalierbarkeit Ihrer Anwendung bedeuten, da Sie für Ihre Operationen nicht einen einzelnen Computer verwenden. Diese Flexibilität führt auch zu einer gewissen Komplexität. Um dieser Komplexität entgegenzuwirken, werden in der Regel mehrere Muster ausgeführt und verschiedene Orchestrierungstechnologien eingesetzt, wie z. B. Docker und Kubernetes.

- **Geografische Skalierung**: Hier skalieren Sie Ihre Anwendungen auf verschiedene geografische Standorte, wofür es zwei wichtige Gründe gibt: Resilienz und geringere Latenz. Resilienz ermöglicht die freie Ausführung Ihrer Anwendung in der betreffenden Region, ohne dass alle Ressourcen mit einer Masterregion verbunden sind. Geringere Latenz heißt, dass Anwender dieser Region ihre Webanforderungen aufgrund ihrer Nähe zum Rechenzentrum schneller erhalten.
- **Sharding**: Hierbei handelt es sich um eines der Verfahren zum Verteilen riesiger Mengen zusammenhängender, strukturierter Daten auf mehrere unabhängige Datenbanken.
- **Development, Testing, Acceptance, and Production** (**Entwicklung, Test, Akzeptanz und Produktion, DTAP**): Dies ist ein Ansatz, bei dem sich mehrere Instanzen in unterschiedlichen logischen Umgebungen befinden. Dies geschieht in der Regel, um Entwicklungs- und Testserver von den Staging- und Produktionsservern zu trennen. Azure DevTest Labs bietet eine Entwicklungs- und Testumgebung, die mit Gruppenrichtlinien konfiguriert werden kann.

Ein weiterer Vorteil der Cloudfähigkeit Ihres Unternehmens ist die Verfügbarkeit der Dienste. Mit Azure ist es einfacher, Ihre Infrastruktur und Ressourcen georedundant, also für mehrere Regionen und Rechenzentren weltweit, verfügbar zu machen. Angenommen, Sie möchten Ihr Unternehmen von Australien nach Kanada expandieren. Zu diesem Zweck können Sie Ihren SQL Server **georedundant** machen, damit kanadische Anwender nicht die Anwendungs- und Datenbankinstanz in Australien abfragen müssen.

Zwar ist Azure eine gemeinsame Suite von Produkten und Dienstangeboten, Sie müssen jedoch nicht komplett umsteigen. Sie können also zunächst eine hybride Architektur mit einer Kombination von On-Premises-Rechenzentren und der Cloud (Azure) implementieren. Eine Hybridlösung umfasst verschiedene Ansätze und Technologien, beispielsweise die Verwendung von **Virtual Private Networks** (**VPNs**) und Azure ExpressRoute, wenn Sie dedizierten Zugriff benötigen.

Mit Azure Data Factory (weitere Informationen hierzu in *Kapitel 2, Entwicklung Ihres modernen Data Warehouse*) können Sie eine Momentaufnahme der Datenquellen von Ihrem On-Premises-SQL Server aus erhalten. Dasselbe Prinzip gilt, wenn Sie über andere Datenquellen von anderen Cloudanbietern oder SaaS-Produkten verfügen. Sie haben die Möglichkeit, eine Kopie dieser Daten in Ihren Azure Data Lake zu bringen. Diese Flexibilität ist sehr vorteilhaft, da Sie auf diese Weise nicht in eine **Anbieterabhängigkeit** geraten, die eine vollständige Migration erforderlich macht.

Die wichtigsten Geschäftsfaktoren für die Einführung von Daten-Analytics in der Cloud

Die einzelnen Unternehmen haben jeweils unterschiedliche Gründe für die Einführung von Daten-Analytics mittels einer Public Cloud wie Microsoft Azure. In den meisten Fällen lassen sich jedoch drei Hauptgründe ermitteln: schnelles Wachstum und Skalierbarkeit, Kostensenkung und Innovationsförderung.

Schnelles Wachstum und Skalierbarkeit

Unternehmen müssen ihren digitalen Fußabdruck schnell vergrößern. Angesichts des rasanten Wachstums mobiler Anwendungen – insbesondere von Medientypen (z. B. Bildern und Videos), IoT-Sensoren und Social-Media-Daten – gibt es unendlich viele Daten zu erfassen. Die Unternehmen müssen daher ihre Infrastruktur skalieren, um diesen massiven Anforderungen gerecht zu werden. Die Größe der Unternehmensdatenbanken nimmt kontinuierlich zu, von mehreren Gigabyte an Daten zu Terabyte oder sogar Petabyte.

Die Anwender sind heute anspruchsvoller als je zuvor. Wenn Ihre Anwendung nicht innerhalb von Sekunden reagiert, werden sich die Anwender eher gegen Ihren Dienst oder Ihr Produkt entscheiden.

Skalierung betrifft nicht nur die Verbraucher der Anwendungen. Sie ist auch für Datenwissenschaftler, -ingenieure und -analysten wichtig, damit diese die Daten eines Unternehmens analysieren können. Die Skalierung der Infrastruktur ist entscheidend, da Sie nicht erwarten können, dass Ihre Dateningenieure auf einem einzelnen Computer riesige Datenmengen (in einer Größenordnung von mehreren Gigabyte bis Terabyte) bearbeiten und Skripte zum Testen Ihrer Datenmodelle ausführen. Selbst wenn Sie diese in einer einzelnen Hochleistungs-Serverinstanz bereitstellen, dauert es Wochen oder Tage, bis der Test abgeschlossen ist. Ganz zu schweigen von den Leistungsengpässen, die die Anwender erleben werden, wenn sie die betreffende Datenbank nutzen.

Mit einem modernen Data Warehouse wie Azure Synapse Analytics stehen Ihnen einige verwaltete Funktionen zum Skalieren zur Verfügung, beispielsweise eine dedizierte Cachingebene. Caching bietet Analysten, Dateningenieuren und -wissenschaftlern die Möglichkeit schnellerer Abfragen.

Geringere Kosten

Aufgrund der Skalierungsanforderungen müssen Unternehmen über einen Mechanismus verfügen, um ihre Dateninfrastruktur kostengünstig und finanziell tragfähig zu erweitern. Es ist zu teuer, ein Data Warehouse On-Premises einzurichten. Nachfolgend sind nur einige Kostenüberlegungen aufgeführt:

- Wartezeit für die Serverbereitstellung und damit verbundene interne Beschaffungsprozesse
- Kosten für das Netzwerk und sonstige physische Infrastruktur, wie z. B. Hardwarekühlung und Rechenzentrumsgebäude
- Kosten für professionelle Dienstleistungen im Zusammenhang mit der Einrichtung und Wartung dieser Server
- Lizenzierungskosten (falls zutreffend)
- Produktivitätsverlust der Mitarbeiter und Teams, die ihre Produkte nicht schneller versenden können

Bei einem modernen Data Warehouse können Sie bei Bedarf neue Hochleistungsserver mit leistungsstarken Grafikkarten hochfahren. Und wenn Sie einen Cloudanbieter wie Microsoft Azure nutzen, müssen Sie nur für die Zeit bezahlen, in der Sie diese Server verwenden. Sie können sie herunterfahren, wenn Sie sie nicht mehr benötigen. Sie können sie nicht nur bei Bedarf deaktivieren, sondern haben auch die Möglichkeit, die betreffenden Ressourcen zu löschen und einfach einen anderen Dienst bereitzustellen, wenn sich herausstellt, dass ein bestimmter Dienst für Ihre Anforderungen nicht geeignet ist.

Azure bietet auch einen Rabatt für „reservierte“ Instanzen, die Sie für einen bestimmten Zeitraum verwenden möchten. Diese sind sehr hilfreich für Datenbanken, Speicherlösungen und Anwendungen, die rund um die Uhr bei minimalen Ausfallzeiten betriebsbereit sein müssen.

Innovationsförderung

Unternehmen müssen in diesem sehr wettbewerbsintensiven Markt ständig innovativ sein, da ansonsten andere ihren Marktanteil übernehmen werden. Natürlich kann aber niemand die Zukunft mit hundertprozentiger Genauigkeit vorhersagen. Daher müssen Unternehmen über einen Mechanismus verfügen, um neue Dinge basierend auf vorhandenen Kenntnissen zu erkunden.

Gute Beispiele hierfür sind das Outsourcing von Geschäftsprozessen (Business Process Outsourcing, BPO) und die Telekommunikationsbranche. Hier gibt es mehrere Petabyte Daten, die möglicherweise noch nicht erforscht wurden. Mit dem modernen Data Warehouse von Microsoft Azure verfügen die Akteure in diesen Branchen über die Infrastruktur für das Durchsuchen von Daten. Mit Azure Data Lake, Azure Data Factory, Azure Synapse Analytics, Azure Databricks, Power BI und Azure Machine Learning können sie ihre Daten erkunden, um geschäftliche Entscheidungen zu fördern. Möglicherweise können sie ein Datenmodell entwickeln, das betrügerische Aktionen erkennen oder Kundenpräferenzen und Erwartungen besser verstehen kann, um so bessere Zufriedenheitsbewertungen zu erzielen. Mit Advanced Analytics können diese Unternehmen Entscheidungen treffen, die heute (und möglicherweise in der Zukunft) relevant sind und sich nicht auf die Analyse historischer Daten beschränken.

Angenommen, Sie möchten ein autonomes Fahrzeug entwickeln. Sie werden ein robustes Data Warehouse benötigen, um Ihre Datasets und eine enorme Menge an Datenverarbeitung zu speichern. Sie müssen riesige Datenmengen erfassen – über Bilder oder Videos, die das Auto kontinuierlich aufzeichnet – und auf Grundlage Ihrer Datensätze und Algorithmen fast sofort eine Antwort finden.

Mit einem Cloudanbieter wie Microsoft Azure können Sie Ihre Ideen frühzeitig testen und validieren, ohne massive Investitionen tätigen zu müssen. Mit Azure können Sie Ihre Ideen schnell in Prototypen umsetzen und verschiedene Möglichkeiten erkunden. Doch was passiert, wenn sich herausstellt, dass das Produkt oder der Dienst, an dem Sie oder Ihr Team gerade arbeiten, nicht wirklich gut ankommt? Wenn Sie On-Premises arbeiten, müssen Sie neben den zugehörigen Lizenzierungs- und Servicekosten weiterhin hohe Haftpflicht- und Betriebskosten tragen, da Sie die Infrastruktur physisch besitzen.

Warum benötigen Sie ein modernes Data Warehouse?

Ein Data Warehouse ist ein zentrales Repository, das verschiedene (oft unterschiedliche) Datenquellen vereinigt. Der Hauptunterschied zwischen einem Data Warehouse und einer Datenbank besteht darin, dass Data Warehouses für OLAP (Online Analytical Processing – analytische Onlineverarbeitung), Datenbanken dagegen für OLTP (Online Transaction Processing – Online-Transaktionsverarbeitung) bestimmt sind. OLAP bedeutet, dass Data Warehouses in erster Linie verwendet werden, um Analytics, Business Intelligence und sogar Machine Learning-Modelle zu generieren. OLTP besagt, dass Datenbanken hauptsächlich für Transaktionen verwendet werden. Bei diesen Transaktionen handelt es sich um die täglichen Arbeitsgänge von Anwendungen, bei denen diese gleichzeitig Daten in Datenbanken lesen und schreiben.

Ein Data Warehouse ist unerlässlich, wenn Sie Ihre Big Data analysieren möchten, da es auch historische Daten enthält (oft als kalte Daten bezeichnet). Zu den meisten gespeicherten Daten gibt es Legacy-Informationen, beispielsweise Daten, die vor 5, 10 oder sogar 15 Jahren gespeichert wurden. Wahrscheinlich möchten Sie nicht, dass die Datenbankinstanz, die Ihre Anwender abfragen, auch diese historischen Daten enthält, da dies die Leistung bei großen Datenmengen beeinträchtigen könnte.

Einige Vorteile eines modernen Data Warehouse:

- Unterstützt jede Datenquelle
- Hoch skalierbar und verfügbar
- Bietet Insights aus Analyse-Dashboards in Echtzeit
- Unterstützt eine Machine-Learning-Umgebung

Die verschiedenen Tools und Dienste, die das moderne Data Warehouse bilden, sind wie folgt miteinander verbunden:

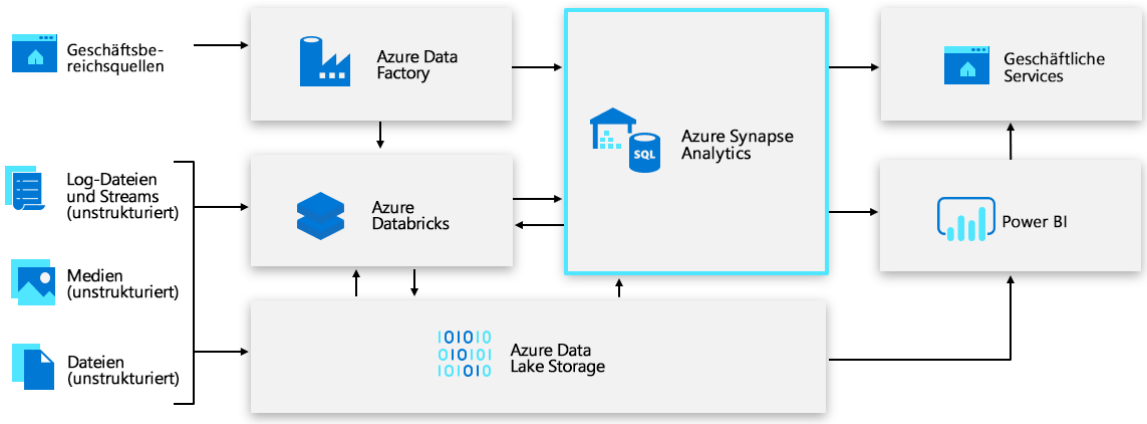

Abbildung 1.4: Architektur eines modernen Data Warehouse

Es gibt viele neue Muster und Architekturen für Data Warehousing, am beliebtesten sind jedoch diejenigen, die eine Trennung der Aufgaben und Verantwortlichkeiten in verschiedenen Phasen der Datenpipeline unterstützen (mehr dazu im Abschnitt *Erstellen einer Datenpipeline*).

Um zu verstehen, was ein modernes Data Warehouse ausmacht, müssen Sie sich zunächst bewusst machen, wie Sie ein herkömmliches Data Warehouse erstellen und verwalten. Hierbei sind zwei wichtige Konzepte zu nennen:

- **Computing**: Hier geht es um die Fähigkeit, die Daten zu verarbeiten und ihre Bedeutung zu verstehen. Dies kann in Form einer Datenbankabfrage erfolgen, um die Ergebnisse für eine andere Schnittstelle, wie z. B. Webanwendungen, zugänglich zu machen.
- **Speicher**: Hier geht es um die Fähigkeit, Daten zu speichern, damit sie auch in Zukunft jederzeit verfügbar sind.

In einem modernen Data Warehouse sind Computing und Speicher kostengünstig getrennt. Anders als bei SQL Server und **SQL Server Integration Services** (**SSIS**) üblich, beinhaltet das Preismodell sowohl die Speicherkapazität als auch die Rechenleistung zur Analyse von Daten. Azure bietet als erster Cloudanbieter ein Data Warehouse, das Computing und Speicher trennt.

Eine weitere Veränderung besteht darin, dass das herkömmliche **ETL**-Modell (**Extrahieren-Transformieren-Laden**) des Data Warehousing jetzt in **ELT** (**Extrahieren-Laden-Transformieren**) geändert wurde. Beim herkömmlichen ETL-Modell sind es Analysten gewohnt, zunächst die Datentransformation abwarten zu müssen, da sie keinen direkten Zugriff auf alle Datenquellen haben. In einem modernen Data Warehouse können riesige Datenmengen in einem Data Lake oder Data Warehouse gespeichert und von den Analysten jederzeit transformiert werden, ohne dass diese auf die Bereitstellung der Daten durch Dateningenieure oder Datenbankadministratoren warten müssen.

Natürlich sind für eine Modernisierung Ihres Data Warehouse weitere Faktoren zu berücksichtigen, beispielsweise die Erweiterbarkeit, Notfallwiederherstellung und Verfügbarkeit. In diesem Abschnitt geht es jedoch zunächst um Computing.

Zusammenbringen Ihrer Daten

In der Vergangenheit waren Datenbanken oft die einzige Datenquelle für Ihre Anwendungen. Heute gibt es jedoch Hunderte, ja Tausende verschiedene Datenquellen. Die Daten aus diesen verschiedenen Quellen haben unterschiedliche Datentypen – einige sind strukturiert, andere unstrukturiert.

Strukturierte Daten: Das Wort „strukturiert“ deutet darauf hin, dass es ein leicht zu interpretierendes Muster gibt. Zu diesen Daten gibt es in der Regel einen vordefinierten Satz von Modellen und ein Schema. Ein **Managementsystem für relationale Datenbanken** (**Relational Database Management System – RDBMS**) wie Microsoft SQL Server ist ein gängiges Beispiel für eine strukturierte Datenspeicherlösung. Das liegt daran, dass es ein Datenbankschema und Tabellenspalten umfasst, die die von Ihnen gespeicherten Daten definieren.

Einige Beispiele für strukturierte Datentypen:

- Kundenname(n)
- Adresse(n)
- Geolocation
- Datum und Uhrzeit
- Mobil- und Festnetznummern
- Kreditkartennummern
- Produktnamen und Stock Keeping Units (SKUs)
- Allgemeine Transaktionsinformationen wie „Von“ und „Bis“ mit Zeitstempeln und Betragswerten

Ein gutes Beispiel für strukturierte Daten sind die Informationen, die Anwender bei der ersten Anmeldung bei einer Anwendung bereitstellen. Den Anwendern wird ein Formular zum Ausfüllen angezeigt. Wenn sie auf die Schaltfläche „Senden“ klicken, werden die Daten an eine Datenbank gesendet und in eine Anwendertabelle mit vordefinierten Spalten für Namen, Adressen und andere Angaben eingefügt. Die Anwender können sich dann bei der Anwendung anmelden, da das System nun die vorhandenen Datensätze für die registrierten Anwender in der Datenbank suchen kann.

Sie können auf die Anwendung zugreifen und Transaktionen ausführen, z. B. Geld überweisen und Vermögenswerte übertragen. Im Laufe der Zeit werden die Anwender eine Reihe von Transaktionen generieren, die Ihre Datenbank letztendlich vergrößern werden. Auch Ihr Datenbankschema wird erweitert, um verschiedene Geschäftsanforderungen zu unterstützen.

Sobald genügend Daten vorhanden sind, können Sie eine Datendurchsuchung durchführen. Zu diesem Zeitpunkt beginnen Sie, nach Mustern in den Daten zu suchen. Sie können betrügerische Transaktionen identifizieren und Hypothesen testen, indem Sie große und wiederholte Transaktionsbeträge von demselben Anwender analysieren.

Die Datendurchsuchung ist begrenzt, da sie nur auf einem strukturierten Dataset in semantischer Form basieren kann. Was aber, wenn Sie auch andere, unstrukturierte Datenquellen, wie z. B. Freitext, berücksichtigen möchten? Ein Beispiel ist eine Transaktionsbeschreibung, in der die Art oder der Empfänger der Transaktion angegeben sein kann. Sie möchten nicht jede Transaktionsbeschreibung manuell lesen und in die rechte Spalte einer Datenbanktabelle einfügen. Sie möchten wahrscheinlich nur die relevanten Informationen extrahieren und in ein strukturiertes Format transformieren. Hier nun kommen unstrukturierte Daten ins Spiel.

Unstrukturierte Daten: Dieser Datentyp ist mehr oder weniger „der Rest“ – d. h. alles, wobei es sich nicht um strukturierte Daten handelt. Dies liegt vor allem daran, dass Sie auf nichts beschränkt sind. Für unstrukturierte Datentypen gibt es in der Regel kein vordefiniertes Datenmodell, das direkt in eine Datenbank passt. Unstrukturierte Daten können „textlastig“ sein und werden in der Regel pro Zeile gelesen oder durch Leerzeichen getrennt. Einige Beispiele für unstrukturierte Datenquellen:

- Bilddateien
- Videos
- E-Mail-Nachrichten und Dokumente
- Log-Dateien
- IoT-Geräte und -Sensoren
- NoSQL-Datenbanken wie MongoDB
- Social Media und Microsoft Graph

Bilddateien und Videos werden aufgrund ihrer dynamischen Beschaffenheit als unstrukturierte Daten klassifiziert. Zwar können ihre Metadaten als strukturiert angesehen werden (beispielsweise Titel, Künstler, Dateiname usw.), der Inhalt selbst ist jedoch unstrukturiert. Mit modernen Tools und Technologien für Daten-Analytics können Sie diese Daten jetzt untersuchen und verstehen. Ein gängiges Beispiel hierfür ist die Gesichtserkennung in Bildern oder Videos.

E-Mails, Dokumente und Log-Dateien verfügen alle über Metadaten. Interessanter für Sie ist jedoch der Inhalt dieser Dateien. In der Regel werden die Daten in E-Mails, Dokumenten und Log-Dateien pro Zeile getrennt und die Nachrichten sind unstrukturiert. Hier würden Sie gerne den Inhalt beschreiben, ohne alles (eventuell Hunderte oder sogar Millionen von Dateien) manuell zu lesen. Ein Beispiel hierfür ist die Stimmungsanalyse von Inhalten, um festzustellen, ob die vorherrschende Emotion Freude, Traurigkeit oder Wut ist. Bei Log-Dateien möchten Sie wahrscheinlich die Fehlermeldungen, Zeitstempel (Datumsangaben) und Messungen (Traces) zwischen den Nachrichten trennen.

IoT-Geräte und -Sensoren werden ähnlich wie Log-Dateien verwendet, um Messungen und Fehler zu einem bestimmten Element zu erfassen. Der Hauptunterschied besteht darin, dass diese Geräte in der Regel in großen Clustern (Hunderte bis Tausende von Geräten) verwendet werden und kontinuierlich Daten streamen. Die von diesen Geräten generierten Daten sind semistrukturiert oder unstrukturiert, da sie im JSON- oder XML-Format vorliegen. Moderne Technologien, wie z. B. Azure IoT-Dienste, beheben diese Komplexität bereits mit Diensten wie Azure IoT Hub, der all diese Daten von verschiedenen Sensoren aggregiert und kontinuierlich in eine Datenquelle exportiert. Manchmal können Sie diese Daten als semistrukturiert klassifizieren, da die Traces und Log-Dateien für ein System leicht verständlich sind.

Sowohl Social-Media-Plattformen als auch Microsoft Graph bieten semistrukturierte Daten. Sie werden in dieser Art klassifiziert, da es nicht ausreicht, einfach nur alle Tweets auf Twitter zu einem Thema abzufragen. Die Ergebnisse sind nicht wirklich sinnvoll, bis Sie einige Analysen durchführen. Der Hauptfokus liegt auf der Unterscheidung von Mustern und Anomalien. Vielleicht möchten Sie Trends zu Nachrichten und Themen ermitteln, aber auch irrelevante Daten, wie z. B. von gefälschten Konten stammende Tweets, entfernen.

Interessanterweise bieten einige **Branchenanwendungen** (**LOB-Anwendungen**) sowohl strukturierte als auch unstrukturierte Daten. Microsoft Dynamics CRM und Salesforce bieten beispielsweise strukturierte Daten, die leicht interpretiert und in Ihre SQL Database-Tabellen exportiert werden können, wie z. B. Daten zu Produkten, ihren Mengen und ihrem Wert. Sie unterstützen jedoch auch unstrukturierte Daten wie Bilder, Videos und „Hinweistext“. Auch wenn es sich bei „Hinweistext“ um eine Zeichenfolge handelt, kann dies dennoch als unstrukturierte Daten angesehen werden, da der Text als „freier Text“ konzipiert ist. Der Text muss keinem angemessenen Format entsprechen, ist jedoch trotzdem eine Untersuchung wert. Diese Art von Text wird häufig verwendet, wenn es darum geht, herauszufinden, warum Verkäufe nicht erfolgreich waren.

Erstellen einer Datenpipeline

Wenn Sie Ihre Datenquellen ermittelt haben, erstellen Sie im nächsten Schritt eine Datenpipeline (manchmal auch als Datenfluss bezeichnet). Abstrakt formuliert beinhaltet dies folgende Schritte: Datenerfassung, Datenspeicherung, Datenvorbereitung und -training, Datenmodellierung und -bereitstellung sowie Datenvisualisierung.

Mit diesem Ansatz entwickeln Sie eine hochgradig skalierbare Architektur, von der alle Anwender des Systems profitieren: von den Endanwendern, Dateningenieuren und Datenwissenschaftlern, die die Daten durchsuchen, über die Analysten, die die Daten für das Unternehmen interpretieren, bis hin zum CEO, der in Echtzeit sehen möchte, wie die Geschäfte laufen:

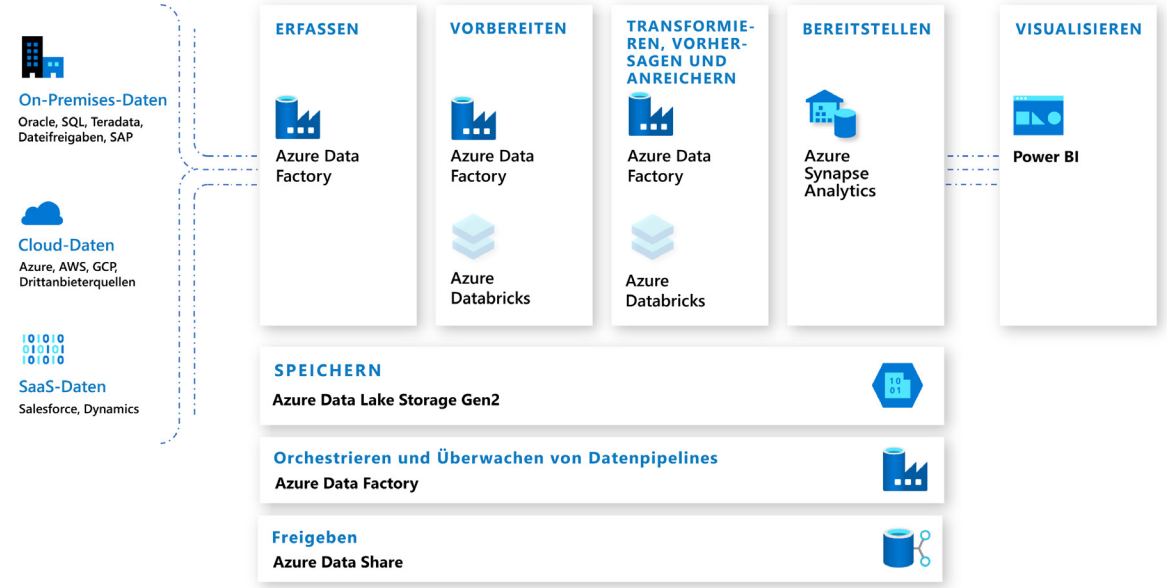

Abbildung 1.5: Beispieldatenpipeline

Datenerfassung

Datenerfassung bezeichnet die Übertragung von (strukturierten und unstrukturierten) Daten von der Quelle in Ihren Speicher, Ihren Data Lake oder Ihr Data Warehouse.

Hierfür wird eine Komponente wie beispielsweise Azure Data Factory (mehr hierzu in *Kapitel 2, Entwicklung Ihres modernen Data Warehouse*) benötigt, die die Daten aus verschiedenen Quellen, wie z. B. On-Premises-Datenbanken und SaaS-Produkten, in einen Data Lake überträgt. Mit diesem Schritt können Sie Ihre **ETL**-Workflows (**Extrahieren-Transformieren-Laden**) und **ELT**-Workflows (**Extrahieren-Laden-Transformieren**) verwalten, ohne dass hierzu eine manuelle Abstimmung erforderlich ist.

Dies ist kein einmaliger Prozess. Idealerweise planen Sie dies oder legen die Auslösung dieser Maßnahme fest, damit Ihr Data Lake von Zeit zu Zeit eine historische Momentaufnahme erhält. Ein Beispiel hierfür ist eine Verbindung von Ihren CRM-Tools, wie z. B. Microsoft Dynamics CRM, mit Azure Data Lake mithilfe von Azure Data Factory. Datenwissenschaftler und Dateningenieure haben damit die Möglichkeit, die Daten in verschiedenen Zeitintervallen zu erkunden, ohne die eigentliche CRM-Anwendung zu unterbrechen.

Datenspeicher

Nach ihrer Erfassung aus verschiedenen Datenquellen werden alle Daten in einem Data Lake gespeichert. Die Daten in dem Data Lake liegen immer noch im Rohdatenformat vor und umfassen sowohl strukturierte als auch unstrukturierte Datenformate. Zu diesem Zeitpunkt sind die Daten nicht sehr hilfreich, um Business Insights zu gewinnen.

Orchestrieren und Überwachen von Datenpipelines

In einem Szenario mit einem modernen Data Warehouse ist es sehr wichtig, dass die Datenquellen und Dienste die Daten effizient von der Quelle zum Ziel übertragen. Azure Data Factory ist ein Orchestrator, mit dessen Hilfe die Dienste die Datenmigration oder -übertragung durchführen können. Das Tool führt nicht die eigentliche Übertragung durch, sondern weist einen Dienst an, sie durchzuführen. So wird beispielsweise ein Hadoop-Cluster angewiesen, eine Hive-Abfrage vorzunehmen.

Mithilfe von Azure Data Factory können Sie zudem Benachrichtigungen und Metriken erstellen, um informiert zu werden, wenn die Dienstorchestrierung funktioniert. Sie können eine Benachrichtigung per E-Mail erstellen, falls eine Datenübertragung von der Quelle zum Ziel nicht erfolgreich war.

Datenfreigabe

In einer modernen Data-Warehouse-Struktur sollte eine nahtlose und sichere Freigabe von Daten stattfinden. Dies kann oft über FTP (File Transport Protocol), E-Mails oder APIs erfolgen, um nur einige Optionen zu nennen. Die Freigabe von Daten in großem Rahmen ist mit einem hohen Verwaltungsaufwand verbunden. Mit Azure Data Share können Sie Ihre Big Data sicher verwalten und an andere Personen und Organisationen weitergeben. Der Datenanbieter hat die volle Kontrolle darüber, wer auf die Datensätze zugreifen kann und wer zu welchen Aktionen berechtigt ist. Dies macht es für die abhängigen Unternehmen leichter, Insights zu gewinnen und KI-Szenarien zu erkunden.

Datenvorbereitung

Nach der Datenerfassung folgt als nächster Schritt die Datenvorbereitung. In dieser Phase werden die Daten aus verschiedenen Datenquellen zu Zwecken der Daten-Analytics vorverarbeitet. Ein Beispiel hierfür ist das Abfragen von Daten von einer API und das Einfügen der Daten in eine Datenbanktabelle. Mit Azure Data Factory können Sie diese Datenvorbereitung orchestrieren. Auch Azure Databricks kann Ihnen bei der Datenvorbereitung helfen, da es Cluster gleichzeitig ausführen kann, um riesige Datenmengen innerhalb von Sekunden oder Minuten zu verarbeiten.

Transformation, Prognose und Anreicherung von Daten

Manchmal erfordert die Datenvorbereitung mehr als nur „Kopieren und Einfügen". Hier kommt die Datentransformation ins Spiel. Es gibt Fälle, in denen Sie zunächst eine benutzerdefinierte Logik auf die Rohdaten anwenden möchten – beispielsweise durch Anwendung von Filtern –, bevor Sie sich entschließen, die Daten in ein Data Warehouse zu übertragen. Azure Data Factory und Azure Databricks können auch in einer solchen Situation hilfreich sein.

Darüber hinaus können Sie die Batchdaten nach Bedarf anreichern, indem Sie einen Azure Machine Learning Service aufrufen, der Echtzeitprognosen zu den Daten vornimmt. Dies kann in Form eines zusätzlichen Features in Ihrer Datenpipeline in Azure Data Factory erfolgen. Mehr über Azure Machine Learning erfahren Sie in *Kapitel 3, Verarbeitung und Visualisierung von Daten*.

Datenbereitstellung

Nach der Vorbereitung und dem Training Ihrer Daten können Sie die Daten modellieren und den Verbrauchern bereitstellen. In dieser Phase modellieren Sie die Daten im Grunde so, dass sie für die Systeme leicht verständlich sind. Dies beinhaltet in der Regel die Durchführung komplexer Abfragen, die Sie aus der Datenvorbereitungs- und -trainingsphase generiert haben, und das Einfügen dieser Datensätze in eine Datenbank, sodass die Daten in einer definierten Tabelle und einem Schema strukturiert sind.

Alle analytischen Daten Ihres Unternehmens werden in einem Data Warehouse gespeichert. Möglicherweise nutzen Hunderte oder gar Tausende Anwender, Berichte und Dashboards gleichzeitig ein einziges Data Warehouse.

In der Regel führen Sie Datenmodellierungen und Dienstintegrationen mit einer Data-Warehouse-Plattform wie Microsoft Azure Synapse Analytics durch. Komplexe und umfassende Abfragen können Stunden oder Tage dauern. Mit dem Potenzial der Cloud können Sie Azure Synapse Analytics jedoch skalieren, um diese Abfragen schneller durchzuführen. So werden aus Tagen Stunden und aus Stunden Minuten (weitere Informationen hierzu in *Kapitel 2, Entwicklung Ihres modernen Data Warehouse*).

Datenvisualisierung

Die Datenvisualisierung ist eine effiziente Möglichkeit der Leistungsanalyse mithilfe von Grafiken und Diagrammen. Dies wird als Business Intelligence bezeichnet. Tools wie Power BI helfen Analysten, die Daten optimal auszuschöpfen. Die Datenvisualisierung ermöglicht eine ergiebige und aussagekräftige Darstellung Ihrer Daten, die Ihnen und Ihren Kunden einen geschäftlichen Mehrwert bringt. Das Team kann Trends, Ausreißer und Muster erkennen. Dies erleichtert datengesteuerte Entscheidungen.

Nach der Analyse der einzelnen Leistungsparameter können verschiedene Stakeholder innerhalb des Unternehmens zusammenarbeiten. Verkauft Ihr Unternehmen seine Produkte gut? In welchen Regionen erzielen Sie den größten Teil Ihres Umsatzes? Wenn Ihre Annahmen auf umfangreiche Daten gestützt sind, können Stakeholder wie etwa CEOs angemessene datengesteuerte Entscheidungen treffen und so die Risiken minimieren. Welche Produktlinien sollten Sie erweitern? Wo sollten Sie weiter expandieren? Dies sind einige häufige Fragen, die Sie beantworten können, wenn Sie über umfassendere Daten-Analytics verfügen.

Analysten können mithilfe von Desktop- oder Webanwendungstools aussagekräftige Darstellungen ihrer Daten erstellen. Im Folgenden sehen Sie ein Beispiel einer Desktopansicht von Power BI. Hier können die Anwender die Daten ihres Unternehmens analysieren und in Grafiken visualisieren:

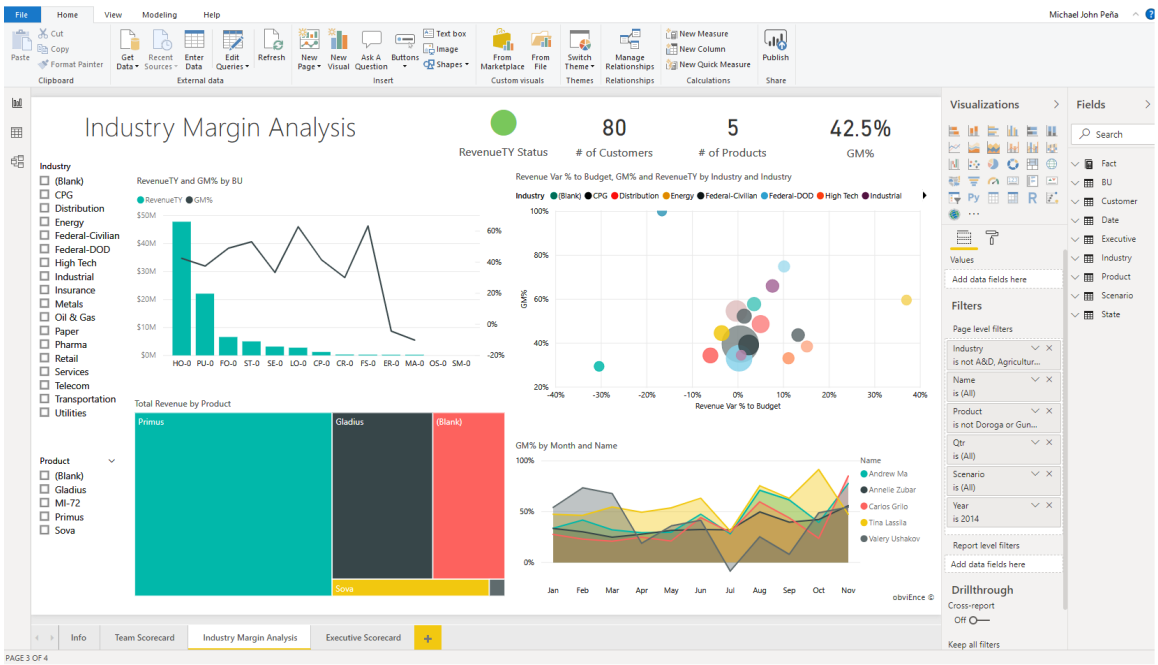

Abbildung 1.6: Power BI-Dashboard auf dem Desktop

Nach ihrer Generierung können die Berichte in einen Arbeitsbereich exportiert werden, in dem die Mitarbeiter gemeinsam an einer Verbesserung der Berichte arbeiten können. Nachfolgend finden Sie eine Beispielansicht desselben Berichts in einer mobilen Anwendung. Anwender können Kommentare und Anmerkungen zu dem Bericht hinzufügen, wodurch eine schnellere Feedbackschleife für die Analysten ermöglicht wird:

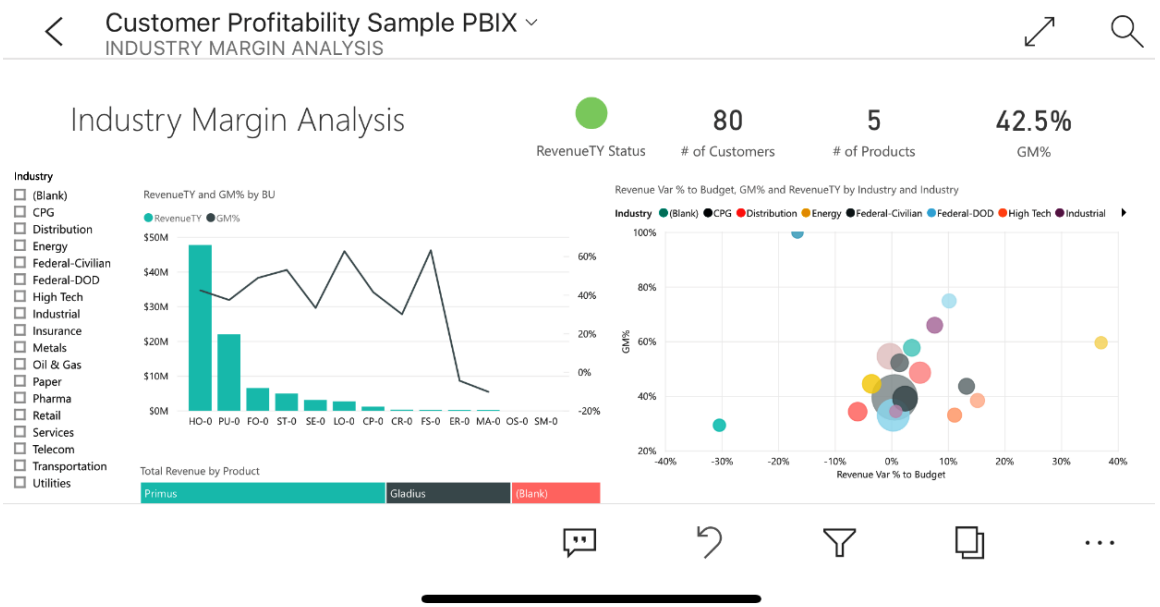

Abbildung 1.7: Power BI-Dashboard für Mobilgeräte

Mit Power BI können Sie umfangreiche personalisierte Dashboards erstellen, die Ihren Anforderungen und Ihrer Marke am besten entsprechen. Im Gegensatz zu Präsentationen, die eine Momentaufnahme einer Grafik aus der letzten Woche oder dem letzten Monat enthalten, können Sie mit diesem Mechanismus denselben Bericht immer wieder aktualisieren.

Intelligentere Anwendungen

Machine Learning hat den Unternehmen geholfen, Anwendungen und Produkte wie beispielsweise Chatbots zu entwickeln, die bestimmte Aufgaben für Anwender ohne menschliches Eingreifen ausführen. Ein typisches Beispiel sind Sprach-Assistenten, wie z. B. Cortana, die aktiv lernen, uns dabei zu unterstützen, bei unseren täglichen Aufgaben produktiver zu werden.

Ein anderes Beispiel sind Onlinespiele, bei denen Sie Ihre Leistung im Vergleich zu anderen Spielern ganz leicht weltweit verfolgen können. Sie können leichter erkennen, wo Sie verglichen mit anderen Spielern stehen, in welchen Bereichen Sie besonders gut abschneiden, wo Sie sich verbessern müssen und wie Sie sich verbessern können.

Die Möglichkeiten der Nutzung umfangreicher Daten sind praktisch grenzenlos. Sie brauchen jedoch den richtigen Ansatz und die richtige Infrastruktur, um eine Skalierung in großem Maßstab zu bewältigen.

Zusammenfassung

In diesem Kapitel wurde die Bedeutung von Daten-Analytics erläutert. Zudem wurden mehrere Aspekte genannt, die Microsoft Azure zu einer idealen Plattform machen, um Business Intelligence-Möglichkeiten in der Cloud zu nutzen. Auch einige grundlegende Konzepte rund um Big Data, Machine Learning und DataOps wurden angesprochen. Darüber hinaus haben Sie einige Geschäftsfaktoren für die Einführung von Daten-Analytics in der Cloud kennengelernt. Schließlich haben Sie eine Übersicht der Voraussetzungen für ein modernes Data Warehouse gewonnen.

Im nächsten Kapitel erfahren Sie, wie Sie mithilfe von Azure Data Factory, Azure Databricks, Azure Data Lake, Azure Synapse Analytics und verwandten Technologien mit der Entwicklung eines modernen Data Warehouse beginnen können.

2

Entwicklung Ihres modernen Data Warehouse

In den letzten Jahren haben Big Data weltweit deutlich an Bedeutung gewonnen. Angesichts einer derart großen Datenmenge werden für die Verarbeitung und Analyse der Daten spezielle Plattformen, Tools und Speicher benötigt.

In *Kapitel 1*, *Einführung in Analytics auf Azure*, haben wir Ihnen Azure vorgestellt. Sie haben die verschiedenen Arten von Plattformen, Tools und Ressourcen kennengelernt, die Azure zum leichteren Erstellen von Data-Warehouse-Lösungen bietet.

In diesem Kapitel werden wir uns eingehender mit den folgenden vier Schlüsseltechnologien befassen:

- Azure Synapse Analytics (früher Azure SQL Data Warehouse)
- Azure Data Factory
- Azure Data Lake Storage Gen2
- Azure Databricks

Am Ende dieses Kapitels erfahren Sie, wie Sie mit diesen Technologien Ihre eigene moderne Data-Warehouse-Lösung entwickeln können. Schnallen Sie sich an, es geht los.

Was ist ein modernes Data Warehouse?

In einem modernen Data Warehouse können Sie Daten aus verschiedenen Datenquellen in jeder Größenordnung – ob On-Premises oder in der Cloud – erfassen, um wertvolle Insights für Ihr Unternehmen zu gewinnen. Der Vorteil eines modernen Data Warehouse besteht darin, dass es sich bei der Datenquelle um strukturierte, semistrukturierte oder unstrukturierte Daten handeln kann. Beispiele für die einzelnen Datenquellentypen finden Sie in der folgenden Abbildung:

Typ der Datenquelle	Beispiele
Strukturierte Daten	Kauftransaktionen in einer Datenbank
Semistrukturierte Daten	Log-Dateien mit Ereignissen und Nachverfolgungsnachrichten aus dem Anwendungsserver
Unstrukturierte Daten	Beiträge in natürlicher Sprache aus Feeds in sozialen Medien (z. B. Twitter, Facebook usw.)

Abbildung 2.1: Beispiele für verschiedene Datenquellentypen

Hier sehen Sie ein typisches Architektur- und Datenflussdiagramm eines modernen Data Warehouse:

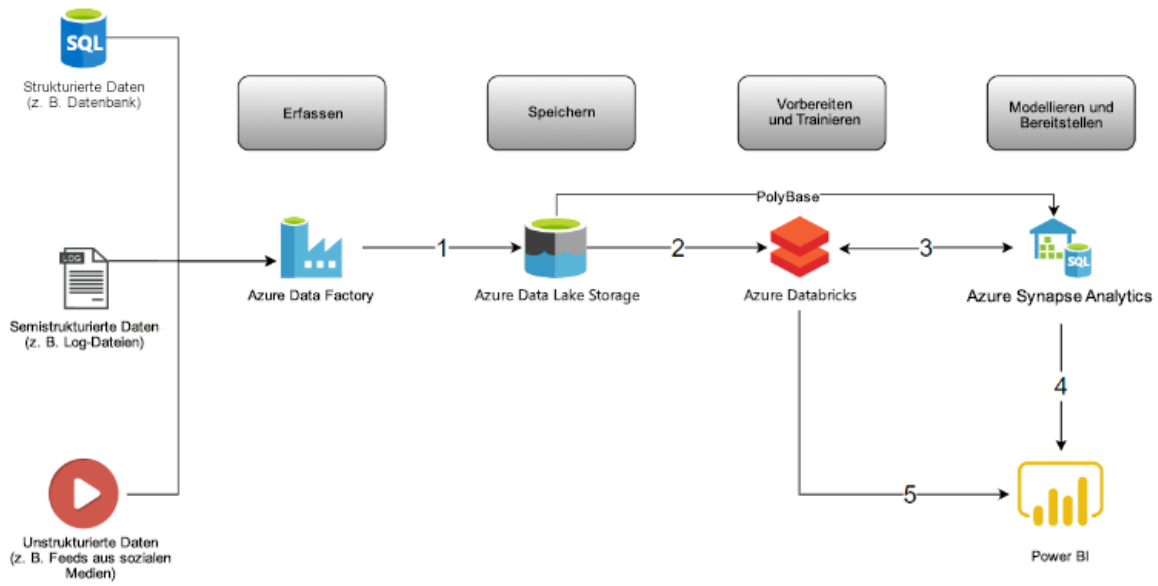

Abbildung 2.2: Architektur eines modernen Data Warehouse

Für die Architektur und den Datenfluss gilt Folgendes:

1. Der Lebenszyklus eines modernen Data Warehouse beginnt mit der Phase der Datenerfassung. Mithilfe von **Azure Data Factory** können Sie Daten aus verschiedenen Quellen zusammenführen – unabhängig davon, ob sie strukturiert, semistrukturiert oder unstrukturiert sind – und die Daten in **Azure Data Lake Storage Gen2** speichern.
2. Zur Datenvorbereitung können Sie die in **Azure Data Lake Storage Gen2** gespeicherten Daten mithilfe von skalierbaren Analytics unter Verwendung von **Azure Databricks** bereinigen und transformieren.
3. Nach der vollständigen Bereinigung und Transformation können die Daten an die vorhandenen Daten in der Azure Warehouse-Datenbank angehängt werden. Sie können diese Daten mithilfe von Connectors zwischen **Azure Synapse Analytics** und **Azure Databricks** abfragen oder verschieben.
4. Die wie in *Punkt* 3 beschrieben vorbereiteten Daten können in Form von Berichten oder einer Art Dashboard für Daten-Analytics genutzt werden.
5. Um den Workflow des modernen Data Warehouse abzuschließen, können Sie direkt in **Azure Databricks** Ad-hoc-Abfragen für die Daten ausführen und die Ergebnisse in **Power BI** visualisieren. (Dies wird in Kapitel 3, Verarbeitung und Visualisierung von Daten, erörtert.)

In diesem Kapitel konzentrieren Sie sich speziell auf die *Schritte* 1, 2 und 3. Die *Schritte* 4 und 5 werden im nächsten Kapitel behandelt.

In den folgenden Abschnitten werden die oben genannten Azure-Dienste, ihre Features und Vorteile vorgestellt.

Azure Synapse Analytics

Azure Synapse Analytics (früher als Azure SQL Data Warehouse bezeichnet) ist ein unbegrenzter Analytics-Dienst, der Warehouse und Big Data auf Unternehmensebene vereinigt. Der Dienst ist in der Lage, riesige Datenmengen gleichzeitig zu verarbeiten. In diesem Abschnitt erhalten Sie eine Übersicht der Features und Vorteile von Azure Synapse Analytics. An späterer Stelle in diesem Kapitel im Abschnitt „Quick-Start-Leitfaden“ werden Sie lernen, Azure Synapse Analytics bereitzustellen.

Features

Azure Synapse Analytics bietet folgende Features:

- Möglichkeit zur gleichzeitigen Ausführung von bis zu 128 Abfragen mit Massively Parallel Processing (MPP)
- Trennung von Computing und Speicher
- Äußerst kostengünstiges Cloud Data Warehouse
- Möglichkeit, Datenbanken temporär zu stoppen und innerhalb weniger Sekunden fortzusetzen
- Möglichkeit zum reibungslosen Entwickeln eigener Aufträge und eines Hubs mit Connectors für Datenintegrations- und -visualisierungsdienste
- Kompatibilität mit Datenschutzgesetzen in mehr als 30 Ländern weltweit
- Integrierte Zwischenspeicherung von Daten, dadurch schnellere Datenabfragen und bessere Leistung

Vorteile

Azure Synapse Analytics bietet Ihnen folgende Vorteile:

- Einfache Skalierbarkeit gemäß Ihren Workload-Anforderungen
- Kosteneffizienz aufgrund des Azure-Konzepts der nutzungsbasierten Bezahlung („Pay-as-you-go“)
- 99,9%ige Verfügbarkeit garantiert

Azure Data Factory

Azure Data Factory (ADF) ist ein vollständig verwaltetes, hochgradig skalierbares, hochverfügbares und anwenderfreundliches Tool zum Erstellen von Integrationslösungen sowie zur Implementierung von ETL-Phasen (Extrahieren-Transformieren-Laden).

Nach der Datenerfassung können Sie die Daten mit Azure Data Factory unter Verwendung des nativen Datenkonnektors in Azure Data Lake Storage Gen2 übertragen. Mithilfe von Drag & Drop können Sie problemlos neue Pipelines in Azure Data Factory erstellen, ohne Code zu schreiben. Für eine erweiterte Implementierung können Sie eigenen Code in Ihren bevorzugten Sprachen schreiben, um Azure Data Factory Ihren spezifischen Anforderungen entsprechend weiter anzupassen.

Azure Data Factory vereinfacht die Datenintegration für Anwender jedes Kenntnisstands.

Features

Azure Data Factory bietet die folgenden Features:

- Möglichkeit einer Verbindung mit verschiedenen Datenquellen, unabhängig davon, ob sich diese On-Premises oder in der Cloud befinden
- Möglichkeit zum Verschieben von Daten aus On-Premises- und Cloud-Datenspeichern in einen zentralisierten Datenspeicher auf Azure mithilfe der Kopieraktivität in der Datenpipeline
- Codefreie Erfassung
- Codefreie Datentransformation
- Möglichkeit zur Verarbeitung und Transformation der Daten aus einem zentralisierten Datenspeicher auf Azure
- Kontrollierter Zeitplan zum Erstellen einer vertrauenswürdigen Datenquelle, die von Produktionsumgebungen genutzt werden soll
- Möglichkeit, Daten zu transformieren, zu bereinigen und in Azure Synapse Analytics zu laden, damit sie von Business-Intelligence-Tools und Analytics-Modulen genutzt werden können
- Qualitätssicherung durch kontinuierliche Pipeline-Überwachung
- Integrierte Überwachungsfeatures wie Azure Monitor oder Azure PowerShell für die Verwaltung von Ressourcen

Vorteile

Azure Data Factory bietet Ihnen folgende Vorteile:

- Orchestrierung anderer Azure-Dienste. Azure Data Factory kann beispielsweise gespeicherte Prozeduren in Azure Synapse Analytics aufrufen oder Azure Databricks-Notebooks ausführen.
- Vollständig verwaltet und serverlos
- Anwenderfreundliches Tool zum Erstellen von Integrationslösungen
- Hochgradig skalierbar In der folgenden Tabelle ist zu sehen, welche Kopierdauer Azure Data Factory je nach Datengröße und Bandbreite erzielt:

Datengröße/ Bandbreite	50 MBit/s	100 MBit/s	500 MBit/s	1 GBit/s	5 GBit/s	10 GBit/s	50 GBit/s
1 GB	2,7 Min.	1,4 Min.	0,3 Min.	0,1 Min.	0,03 Min.	0,01 Min.	0,0 Min.
10 GB	27,3 Min.	13,7 Min.	2,7 Min.	1,3 Min.	0,3 Min.	0,1 Min.	0,03 Min.
100 GB	4,6 Std.	2,3 Std.	0,5 Std.	0,2 Std.	0,05 Std.	0,02 Std.	0,0 Std.
1 TB	46,6 Std.	23,3 Std.	4,7 Std.	2,3 Std.	0,5 Std.	0,2 Std.	0,05 Std.
10 TB	19,4 Tage	9,7 Tage	1,9 Tage	0,9 Tage	0,2 Tage	0,1 Tage	0,02 Tage
100 TB	194,2 Tage	97,1 Tage	19,4 Tage	9,7 Tage	1,9 Tage	1 Tag	0,2 Tage
1 PB	64,7 Mon.	32,4 Mon.	6,5 Mon.	3,2 Mon.	0,6 Mon.	0,3 Mon.	0,06 Mon.
10 PB	647,3 Mon.	323,6 Mon.	64,7 Mon.	31,6 Mon.	6,5 Mon.	3,2 Mon.	0,6 Mon.

Abbildung 2.3: ADF-Kopierdauer basierend auf Datengröße und Bandbreite

Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2 bietet kostengünstige skalierbare Datenspeicherlösungen, die auf der Grundlage von Azure Blob Storage-Technologie entwickelt wurden. Azure Data Lake Storage Gen2 wurde speziell für Big Data Analytics entwickelt und ermöglicht es Anwendern, strukturierte, semistrukturierte und unstrukturierte Daten aus verschiedenen Quellen zu speichern. Zu diesen Quellen gehören relationale Datenbanken, CRM-Systeme (Customer Relationship Management – Kundenbeziehungsmanagement), mobile Anwendungen, Desktopanwendungen, IoT-Geräte und mehr.

In Azure Data Lake Storage Gen2 gespeicherte strukturierte Daten können mit Azure Data Factory, Azure Databricks, PolyBase oder dem Befehl COPY in Azure Synapse Analytics geladen werden.

Features

Azure Data Lake Storage Gen2 bietet folgende Features:

- Datenzugriff und -verwaltung über Hadoop Distributed File System (HDFS)
- ABFS-Treiber (Azure BLOB File System), der den Datenzugriff auf Azure Data Lake Storage Gen2 von allen Apache-Hadoop-Umgebungen, wie z. B. Azure Synapse Analytics, Azure Databricks und Azure HDInsight, aus ermöglicht
- Unterstützung für ACL- und POSIX-Berechtigungen neben zusätzlichen Azure Data Lake Storage Gen2-Berechtigungen
- Konfigurierbare Einstellungen über Azure Storage Explorer, Apache Spark und Apache Hive
- Sichere (und kostengünstige) Skalierbarkeit von Speicher auf Dateiebene
- Funktionen für hohe Verfügbarkeit/Notfallwiederherstellung (HA/DR)

Vorteile

Azure Data Lake Storage Gen2 bietet folgende Vorteile:

- Unterstützung für Anwendungen, die den offenen Apache-Hadoop-Distributed-File-System-Standard (HDFS-Standard) implementieren
- Kostengünstige Speicherkapazität und Transaktionen

Azure Databricks

Azure Databricks ist eine auf Apache Spark basierende Analytics-Plattform, mit der Sie Lösungen mit künstlicher Intelligenz implementieren und gemeinsam in einem interaktiven Arbeitsbereich Insights erarbeiten können. Die Plattform unterstützt Sprachen wie Python, Java, R, Scala und SQL sowie eine Vielzahl von Data-Science-Tools wie TensorFlow, scikit-learn und PyTorch.

Features

Azure Databricks bietet folgende Features:

- Interaktiver, gemeinsamer Arbeitsbereich
- Umfassende Apache Spark-Clusterfunktionen mit Komponenten wie Spark SQL und DataFrames, Mlib, GraphX und Spark Core API

Vorteile

Azure Databricks bietet die folgenden Vorteile:

- Einfache Einrichtung von vollständig verwalteten Apache Spark-Clustern auf Azure
- Zero-Management-Cloud-Plattform
- Interaktiver, gemeinsamer Arbeitsbereich für Durchsuchung und Visualisierung
- Schnelle Clustererstellung
- Dynamische automatische Skalierung von Clustern, einschließlich serverloser Cluster
- Betrieb von Clustern mit Code und APIs
- Fähigkeit zur sicheren Integration von Daten auf Basis von Apache Spark
- Sofortiger Zugriff auf die neuesten Apache Spark-Features mit jeder Version
- Native Integration in Azure Synapse Analytics, Azure Data Lake Storage, Azure Cosmos DB, Azure Blob Storage und Power BI

Quick-Start-Leitfaden

In diesem Quick-Start-Leitfaden werden Sie Azure Synapse Analytics (früher SQL DW) erstmals bereitstellen. Sie werden Ihr Data Warehouse mit einer Beispieldatenbank namens **`AdventureWorksDW`** füllen. Anschließend werden Sie Ihre Datenbank verbinden und SQL-Abfragen ausführen, um mithilfe Ihrer Daten Ergebnisse zu erzielen.

Schließlich werden Sie Ihr Azure SQL Data Warehouse (auch als Azure Synapse Analytics bezeichnet) anhalten, um die Computing-Abrechnung zu stoppen, wenn Sie das Data Warehouse nicht mehr nutzen müssen (in diesem Fall wird Ihnen nur der Speicher in Rechnung gestellt).

Wenn Sie eine der Techniken testen möchten, die in diesem Buch vorgestellt werden, erstellen Sie Ihr kostenfreies Azure-Konto, und steigen Sie direkt ein

> **Hinweis**
>
> Zum Zeitpunkt dieser Veröffentlichung ist Microsoft noch dabei, **Azure SQL Data Warehouse** zu **Azure Synapse Analytics** weiterzuentwickeln. Daher wird in einigen Schritten in diesem Leitfaden der neue Name, **Azure Synapse Analytics,** verwendet, während in anderen Schritten der ursprüngliche Name, **Azure SQL Data Warehouse**, angegeben ist. Besuchen Sie Microsoft Azure, um aktuelle Informationen zu erhalten.

Erstes Bereitstellen von Azure Synapse Analytics (früher SQL DW)

Führen Sie die folgenden Schritte aus, um Ihr Data Warehouse bereitzustellen:

1. Melden Sie sich in einem Webbrowser beim Azure-Portal an.
2. Klicken Sie oben links im Azure-Portal auf **Create a resource** (Ressource erstellen):

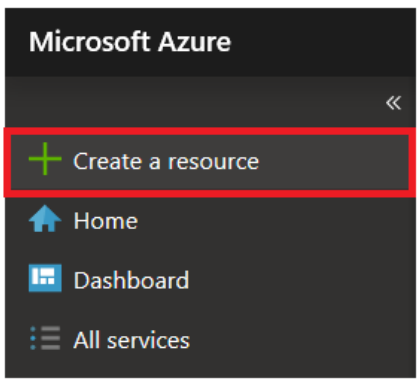

Abbildung 2.4: Erstellen einer Ressource im Azure-Portal

3. Wählen Sie **Databases** (Datenbanken) im Abschnitt **Azure Marketplace** auf der Seite **New** (Neu) und anschließend **Azure Synapse Analytics (formerly SQL DW)** (Azure Synapse Analytics (früher SQL DW)) im Abschnitt **Featured** (Empfohlen) aus:

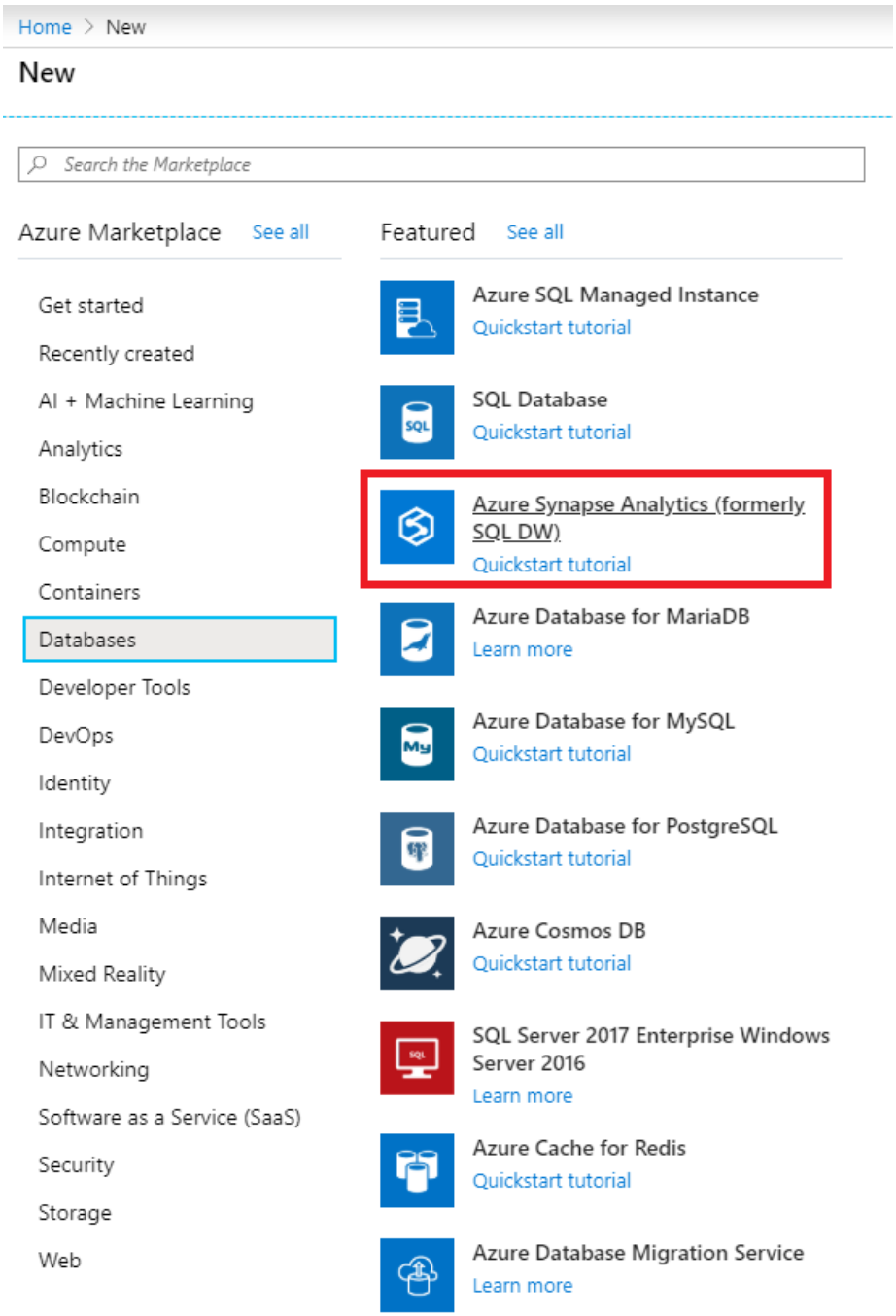

Abbildung 2.5: Erstellen eines neuen Azure Synapse Analytics (früher SQL DW)

4. Füllen Sie alle Projektdetails im Formular **SQL Data Warehouse** aus, wie in der folgenden Abbildung dargestellt:

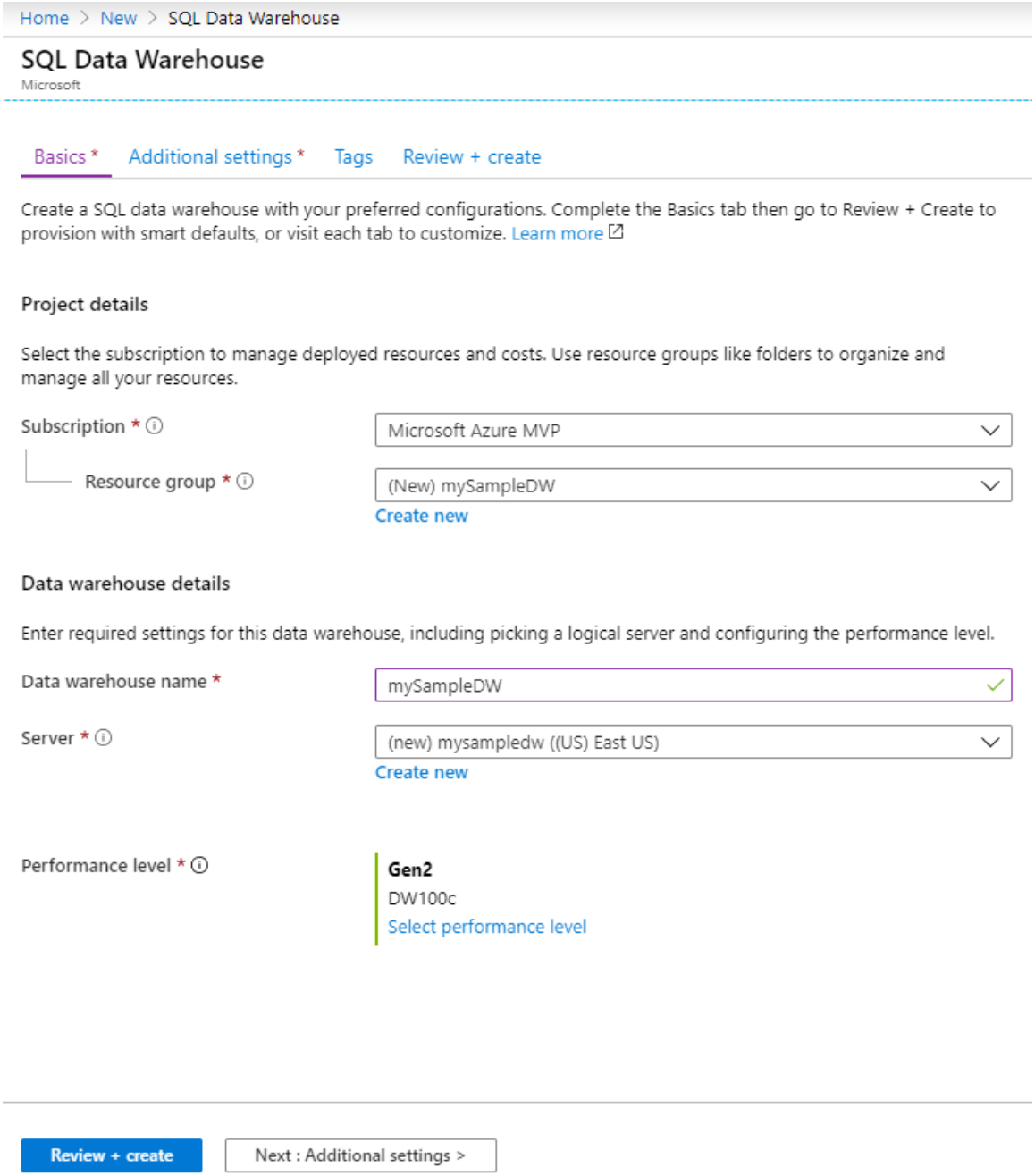

Abbildung 2.6: Hinzufügen von Projektdetails zum Data Warehouse

Der jeweils vorgeschlagene Wert und eine Erläuterung sind in der folgenden Tabelle angegeben:

Setting	Empfohlener Wert	Beschreibung
Abonnement	Ihr Azure-Abonnement	Wählen Sie Ihr Azure-Abonnement.
Ressourcengruppe	mySampleDW	Erstellen Sie eine neue Ressourcengruppe namens mySampleRG.
Name des Data Warehouse	mySampleDW	Verwenden Sie einen beliebigen gültigen Datenbanknamen.
Quelle auswählen	Beispiel	Geben Sie an, dass eine Beispieldatenbank geladen werden soll.
Beispiel auswählen	AdventureWorksDW	Geben Sie an, dass AdventureWorksDW-Beispieldatenbank geladen werden soll.
Leistungsstufe	Gen2, DW100c	Geben Sie die Leistungsstufe des Data Warehouse an.
Ort	USA, Osten	Geben Sie den Speicherort Ihres Data Warehouse und der zugehörigen Ressourcen an.

Abbildung 2.7: Details der Projekteinstellung

5. Klicken Sie unter **Data warehouse details** (Details zu Data Warehouse), **Server**, auf **Create new** (Neu erstellen). Die folgende Ansicht wird aufgerufen. Geben Sie Informationen zu dem Server in den Feldern **Server name** (Servername), **Server admin login** (Serveradministratoranmeldung), **Password** (Kennwort) und **Location** (Standort) an:

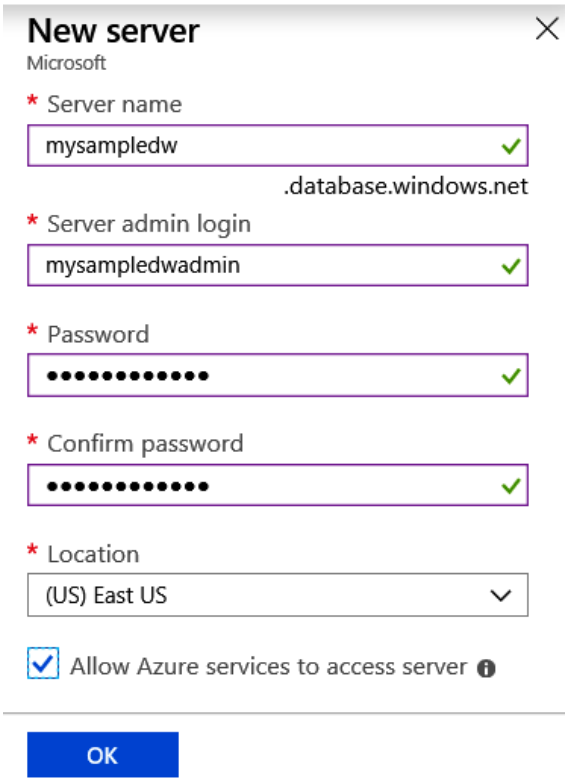

Abbildung 2.8: Erstellen eines neuen Servers

6. Wählen Sie **Performance level** (Leistungsebene) und anschließend **Gen2** aus. In diesem Beispiel werden Sie Ihr System auf **DW100c** skalieren. Klicken Sie auf **Apply** (Anwenden), um die Auswahl zu bestätigen.

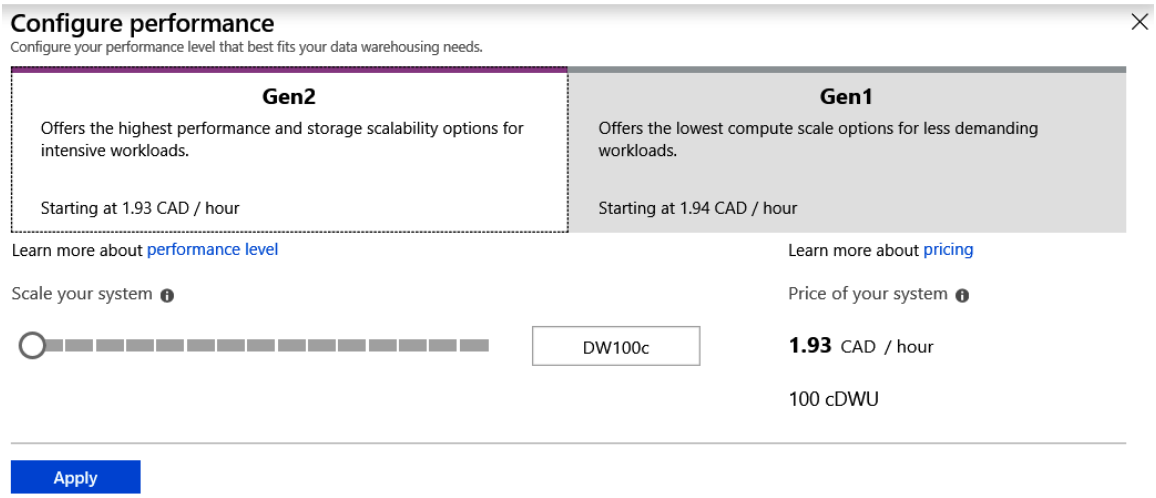

Abbildung 2.9: Auswählen von Leistungsebenen

7. Klicken Sie auf **Next: Additional settings** (Weiter: zusätzliche Einstellungen).

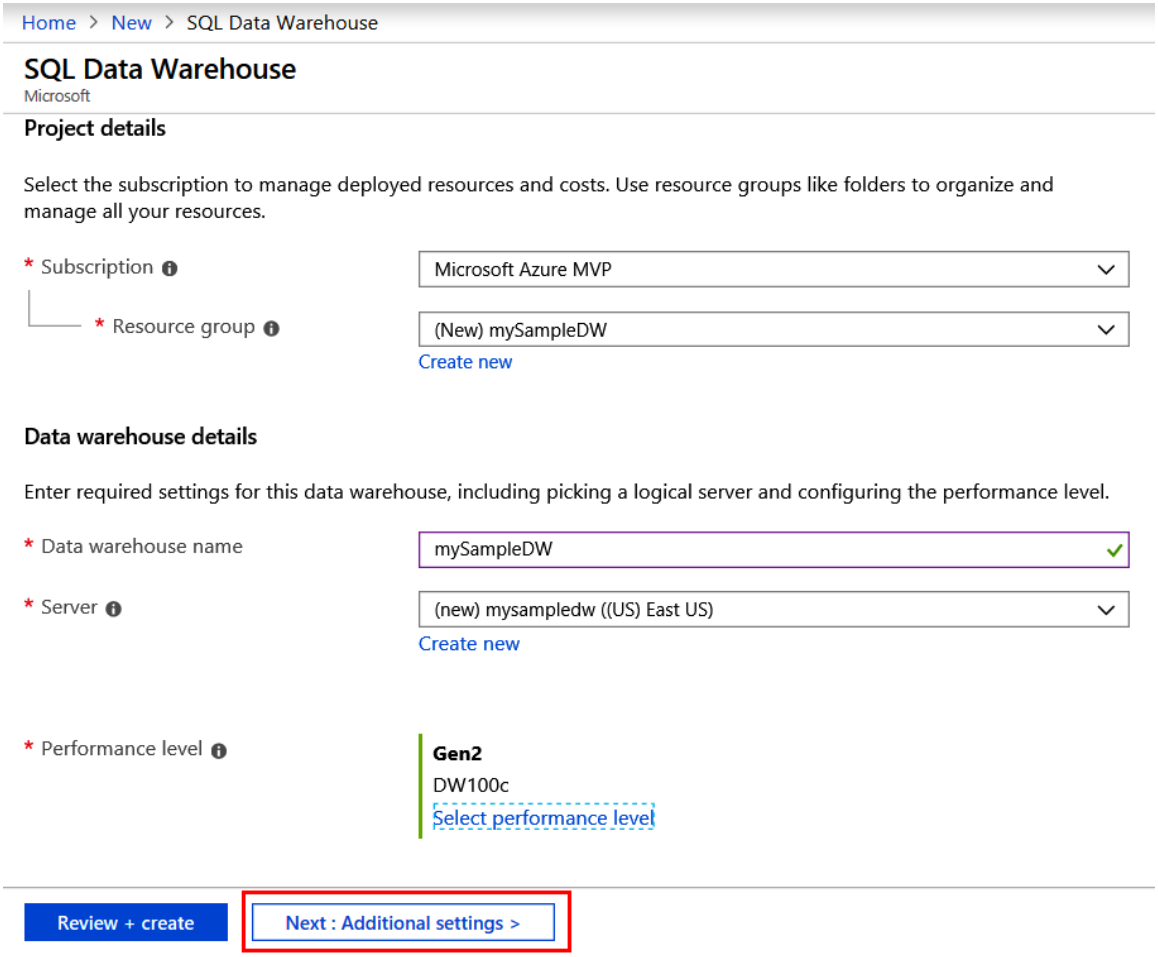

Abbildung 2.10: Konfigurieren zusätzlicher Einstellungen

8. Legen Sie für Ihre Quick-Start-Tour für **Use existing data** (Vorhandene Daten verwenden) **Sample** (Beispiel) fest. Daraufhin wird die Beispieldatenbank **`AdventureWorksDW`** in Ihr neu erstelltes Data Warehouse geladen. Klicken Sie auf **Review + create** (Prüfen + erstellen), um fortzufahren.

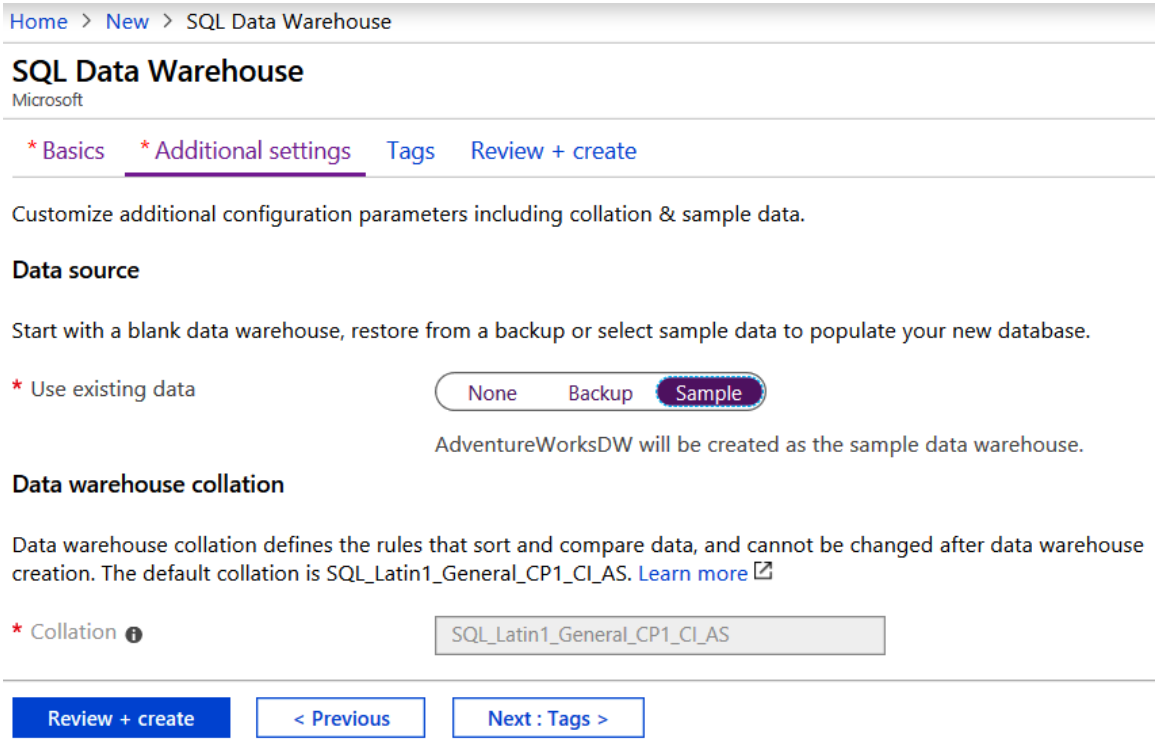

Abbildung 2.11: Verwenden des Beispiel-Datasets für den Quick-Start-Leitfaden

9. Führen Sie eine abschließende Überprüfung durch, und klicken Sie dann auf **Create** (Erstellen), um mit der Bereitstellung Ihres SQL Data Warehouse zu beginnen:

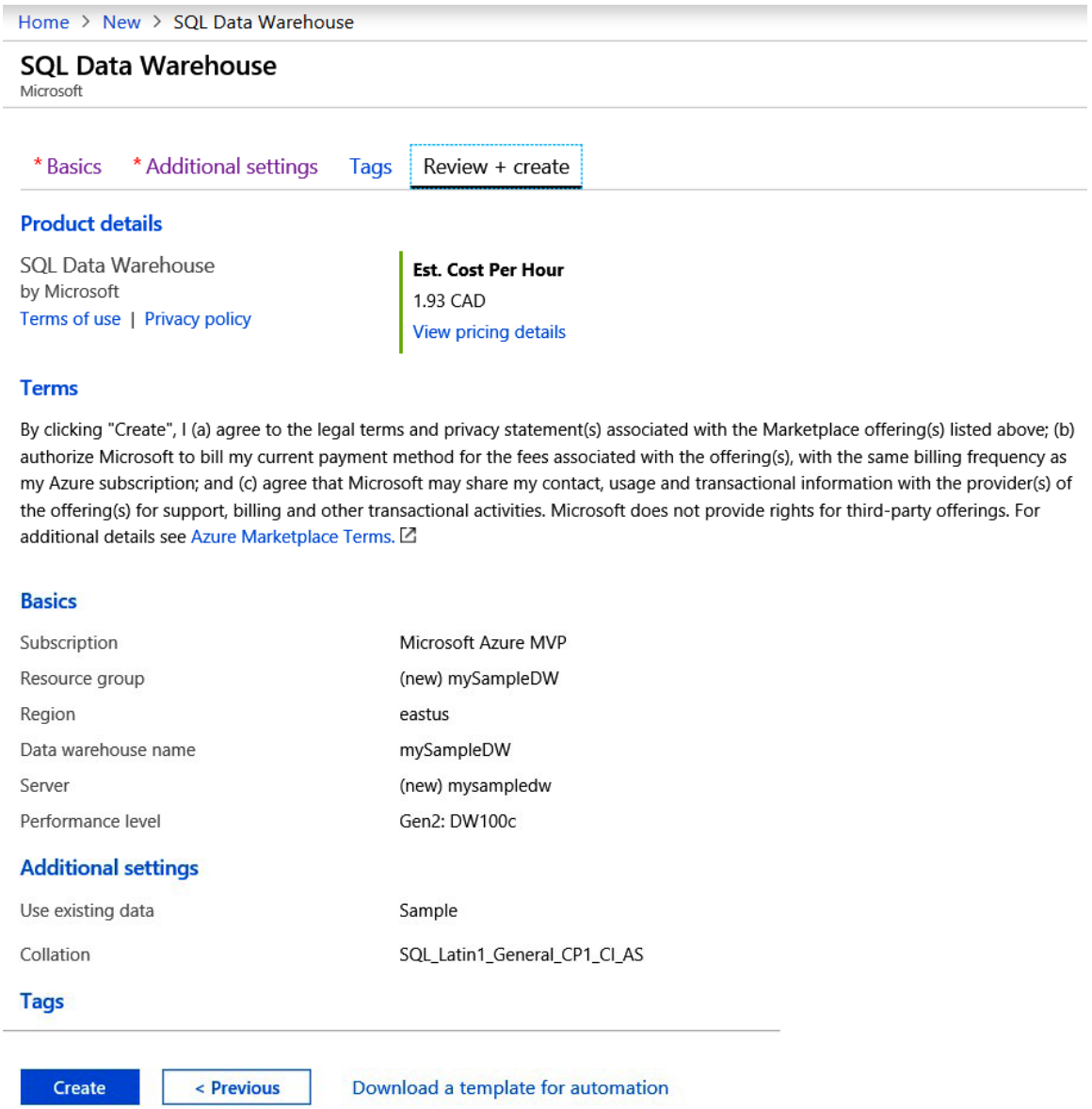

Abbildung 2.12: Erstellen Ihres SQL Data Warehouse

10. Während Ihr neues SQL Data Warehouse bereitgestellt wird, können Sie den Bereitstellungsfortschritt überwachen:

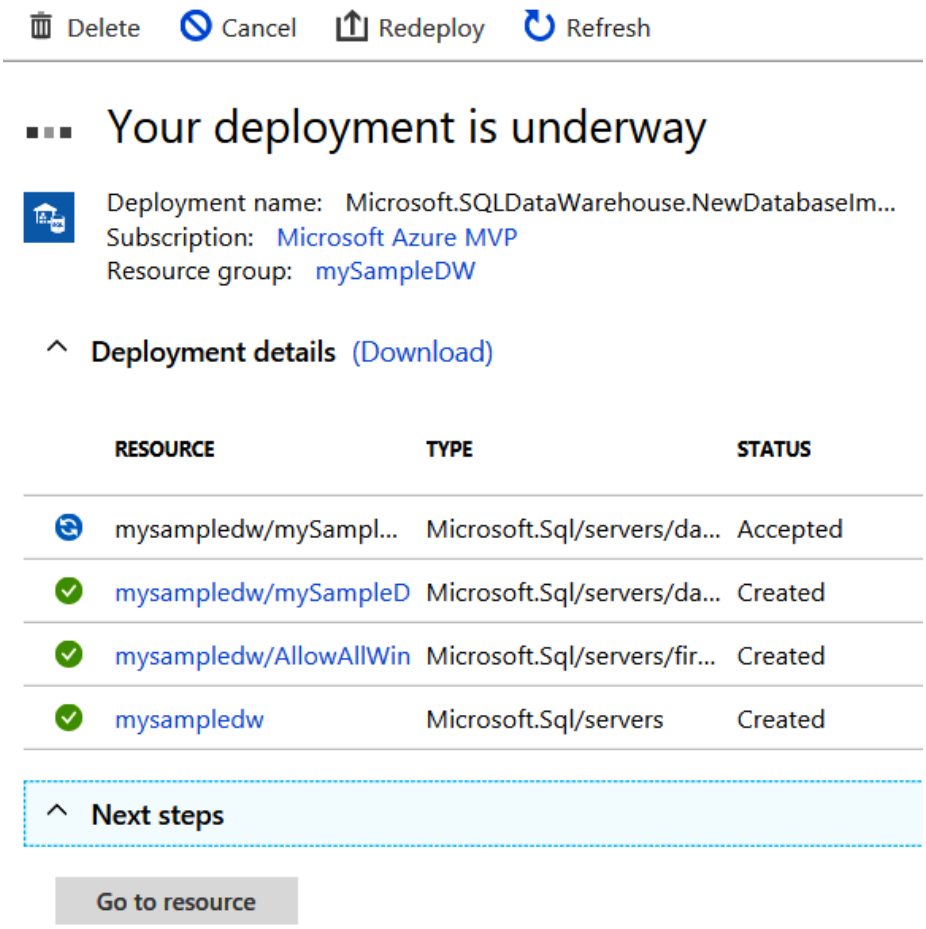

Abbildung 2.13: Überwachen der Data-Warehouse-Bereitstellung

11. Sobald die Bereitstellung abgeschlossen ist, wird der folgende Bildschirm angezeigt:

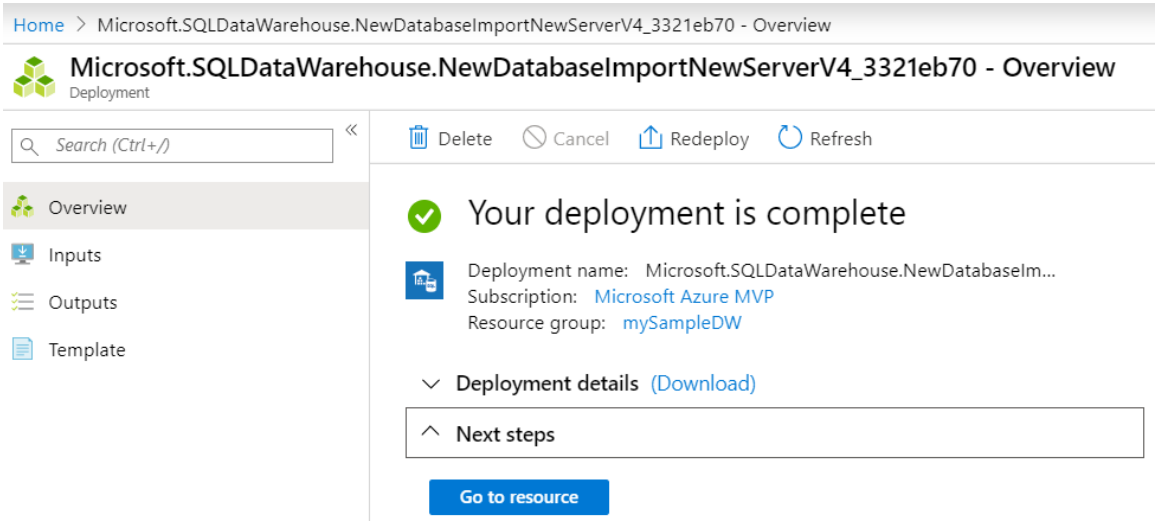

Abbildung 2.14: Bereitstellung des Data Warehouse abgeschlossen

Sie haben Azure Synapse Analytics (früher als Azure SQL Data Warehouse bezeichnet) erfolgreich bereitgestellt.

Wenn Sie zu der Ressourcengruppe mit dem Namen „mySampleDW“ wechseln, sehen Sie die folgenden Ressourcen, die erfolgreich bereitgestellt wurden:

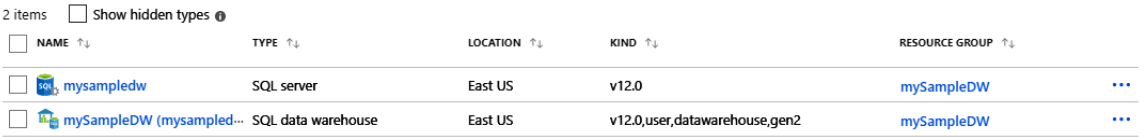

Abbildung 2.15: Liste der bereitgestellten Ressourcen

Abfragen der Daten

Zum Abfragen der Daten können Sie den in das Azure-Portal integrierten **Query editor (preview)** (Abfrage-Editor (Vorschau)) verwenden. Dieser ist äußerst praktisch. Führen Sie die folgenden Schritte aus, um die Daten abzufragen:

1. Klicken Sie auf **Query editor (preview)** (Abfrage-Editor (Vorschau)) in der Datenbank `mySampleDW`. Der folgende Bildschirm wird angezeigt. Geben Sie einfach die Anmelde- und Kennwortinformationen an, die Sie zuvor bei der Bereitstellung von Azure Synapse Analytics (früher Azure SQL Data Warehouse) eingegeben haben:

Abbildung 2.16: SQL Database-Abfrage-Editor

> **Hinweis**
>
> Wenn Sie eine Fehlermeldung wie z. B. „Cannot open server 'mysampledw' requested by the login. Client with IP address 'xx.xx.xx.xx' is not allowed to access the server“ (Der von der Anmeldung angeforderte Server 'mysampledw' kann nicht geöffnet werden. Der Client mit IP-Adresse 'xx.xx.xx.xx' ist nicht zum Zugriff auf den Server berechtigt) erhalten, befolgen Sie die Anweisungen im Abschnitt *Whitelisting Ihrer Client-IP-Adresse für den Zugriff auf Azure Synapse Analytics (früher SQL DW)*, um Ihre Client-IP-Adresse auf die Whitelist zu setzen. Nachdem Ihre Client-IP-Adresse auf die Whitelist gesetzt wurde, können Sie Schritt 1 wiederholen.

2. Wenn Sie sich bei Ihrer Data-Warehouse-Datenbank angemeldet haben, können Sie SQL-Abfragen ausführen, um die benötigten Informationen abzurufen. Geben Sie die folgende Abfrage in den Abfragebereich ein, und klicken Sie auf **Run** (Ausführen):

```
select firstname, lastname, phone
from dimCustomer
order by lastname
```

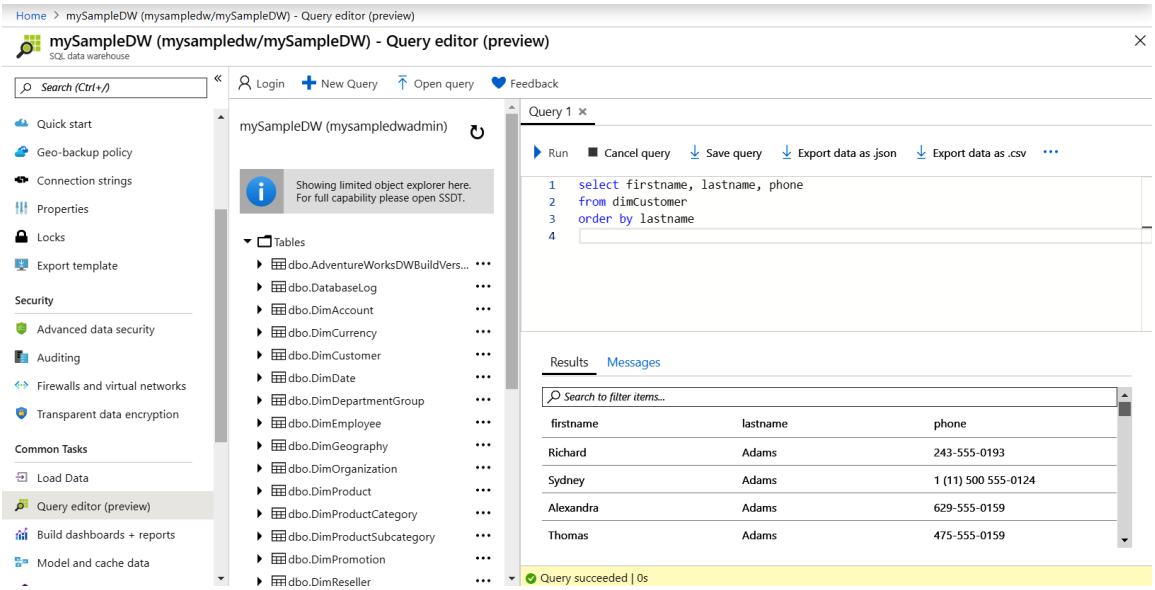

Abbildung 2.17: Ausführen von Abfragen im Abfrage-Editor

3. Versuchen Sie es mit einer anderen Abfrage, und klicken Sie auf **Run** (Ausführen):

```
select firstname, lastname, phone
from dimCustomer
where lastname = 'Lee'
order by firstname
```

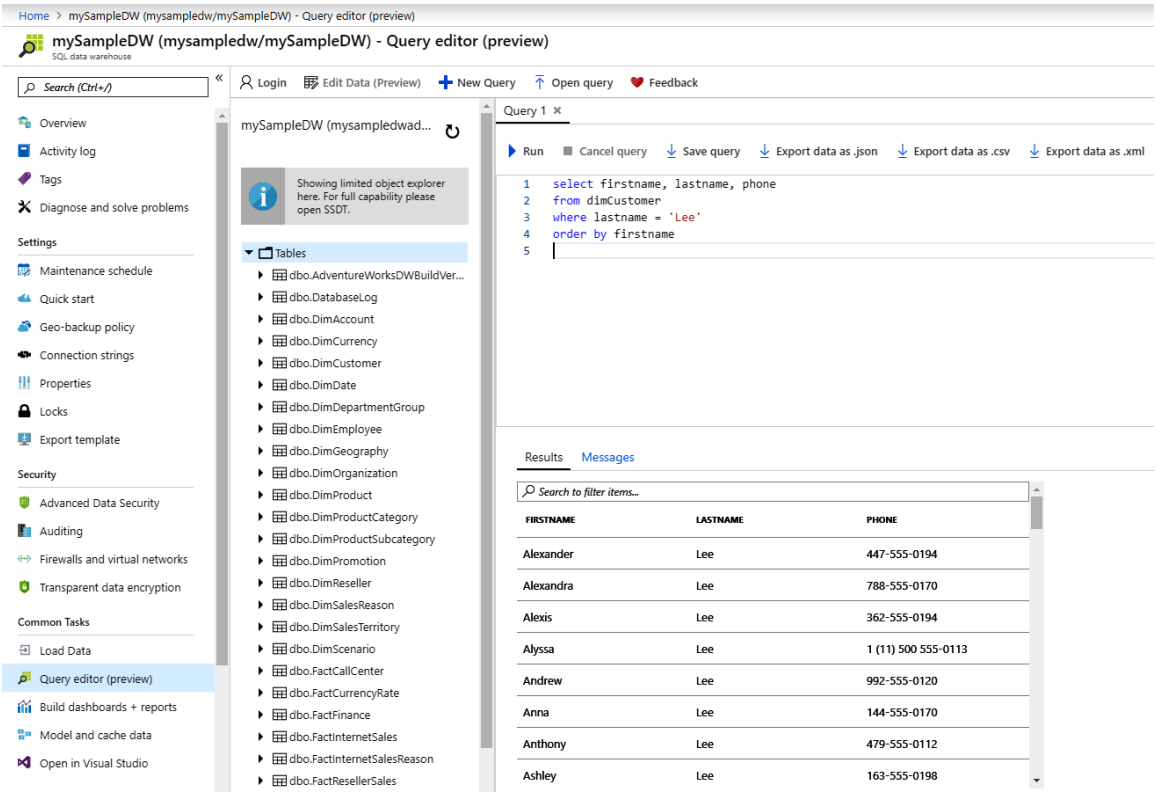

Abbildung 2.18: Ausführen von Abfragen im Abfrage-Editor

Sie können auch gerne selbst weitere Abfragen testen.

Whitelisting Ihrer Client-IP-Adresse für den Zugriff auf Azure Synapse Analytics (früher SQL DW)

Wenn Sie von Ihrem lokalen SQL Server Management Studio (SMS) oder dem Abfrage-Editor (Vorschau) aus auf Azure Synapse Analytics (früher SQL DW) zugreifen möchten, müssen Sie eine serverseitige Firewallregel hinzufügen, die Konnektivität für Ihre Client-IP-Adresse ermöglicht. Führen Sie hierfür die folgenden Schritte aus:

1. Wechseln Sie unter **Security** (Sicherheit) zu **Firewalls and virtual networks** (Firewalls und virtuelle Netzwerke). Klicken Sie auf **Add client IP** (Client-IP hinzufügen), um Ihre aktuelle IP-Adresse auf die Whitelist zu setzen. Klicken Sie dann auf **Save** (Speichern):

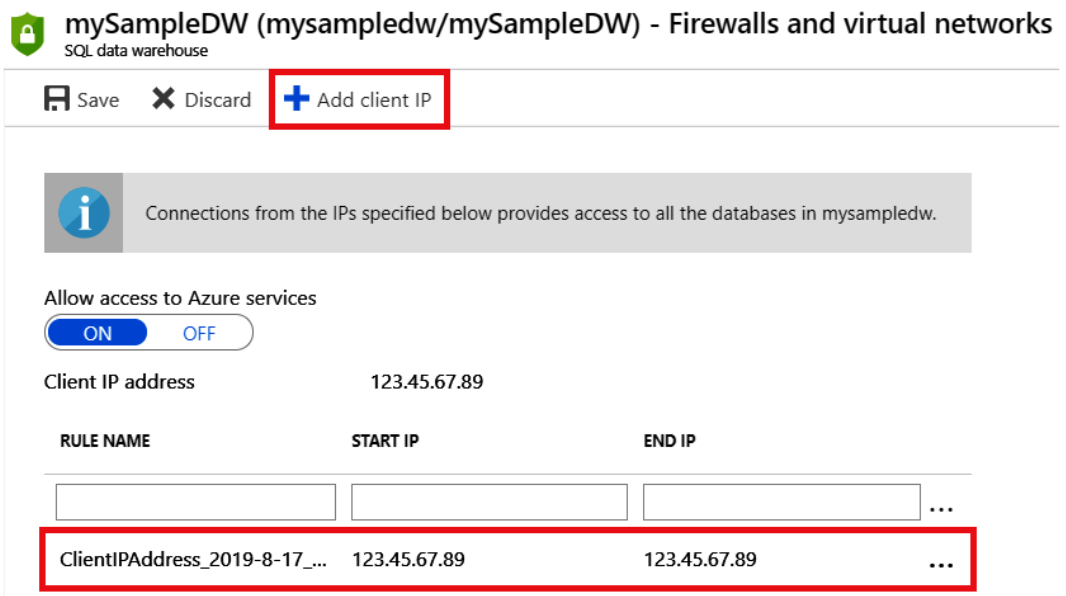

Abbildung 2.19: Hinzufügen von Client-IP-Adressen

2. Wechseln Sie nun zu Ihrem SQL Data Warehouse, **Overview** (Übersicht), und suchen Sie den Namen Ihres Servers unter **Server name** (Servername).

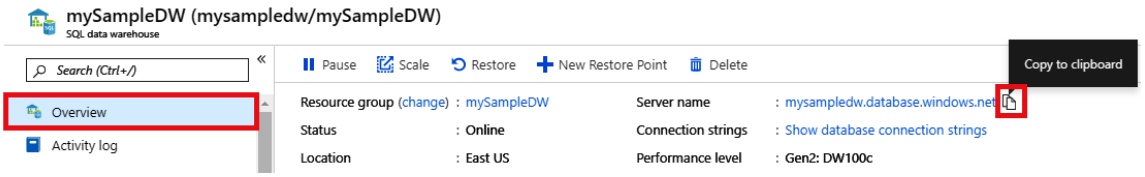

Abbildung 2.20: Kopieren des Servernamens

3. Wenn Ihre Client-IP-Adresse in der Whitelist enthalten ist, können Sie sich jetzt von Ihrem lokalen SSMS oder dem Abfrage-Editor (Vorschau) aus mit Azure Synapse Analytics (früher SQL DW) verbinden, indem Sie einfach mit Ihrem Servernamen, Anwendernamen und Kennwort eine Verbindung herstellen.

Anhalten von Azure Synapse Analytics, wenn es nicht verwendet wird

Wenn Sie Azure Synapse Analytics (früher SQL Data Warehouse) nicht verwenden, empfiehlt es sich, es anzuhalten, damit Ihnen keine zusätzlichen Computing-Gebühren in Rechnung gestellt werden. Wenn Sie Ihr Data Warehouse anhalten und nicht löschen, hat dies den Vorteil, dass Sie Ihr Data Warehouse einfach fortsetzen können, wenn Sie die Arbeit wiederaufnehmen möchten, ohne dass hierfür eine erneute Bereitstellung notwendig ist.

> **Hinweis**
>
> Wenn Ihr Azure SQL Data Warehouse angehalten ist, fallen zwar keine Computing-Gebühren an, Sie müssen jedoch weiterhin die Speichergebühren tragen.

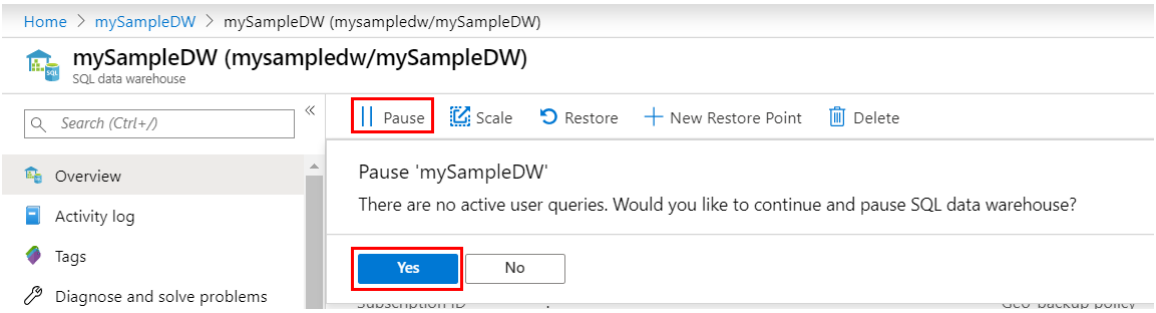

Abbildung 2.21: Anhalten Ihres Data Warehouse

Wenn das Azure SQL Data Warehouse angehalten wurde, wird als Status „Paused“ (Angehalten) angezeigt:

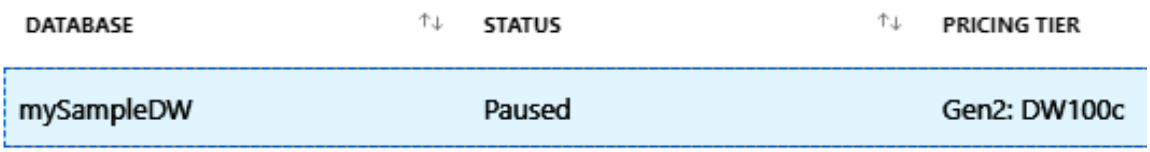

Abbildung 2.22: Status von SQL Data Warehouse

Wenn Sie bereit sind, erneut mit Ihrem Azure SQL Data Warehouse zu arbeiten, klicken Sie einfach auf **Resume** (Fortsetzen):

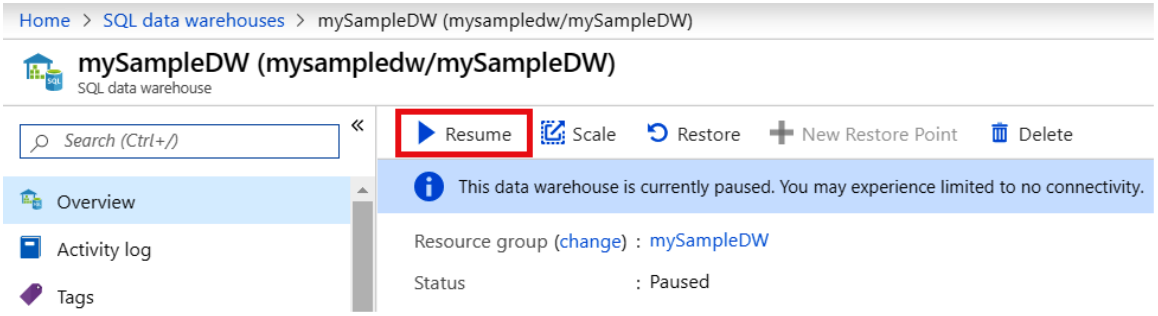

Abbildung 2.23: Fortsetzen Ihres Data Warehouse

Bereitstellen von Azure Data Factory

Nachdem Sie nun Azure Synapse Analytics (früher Azure Data Warehouse) erstmals bereitgestellt haben, können Sie Azure Data Factory wie im Folgenden beschrieben bereitstellen:

1. Klicken Sie oben links im Azure-Portal auf **Create a resource** (Ressource erstellen):

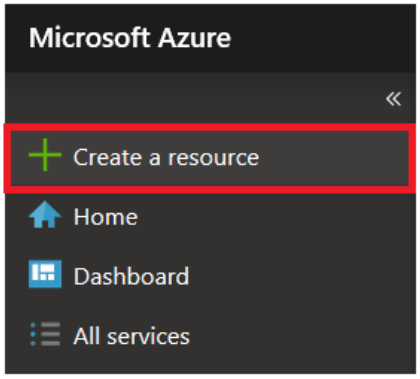

Abbildung 2.24: Erstellen Ihrer Ressource

2. Wählen Sie **Analytics** im Abschnitt „Azure Marketplace“ auf der Seite **New** (Neu) und anschließend **Data Factory** im Abschnitt **Featured** (Empfohlen) aus.

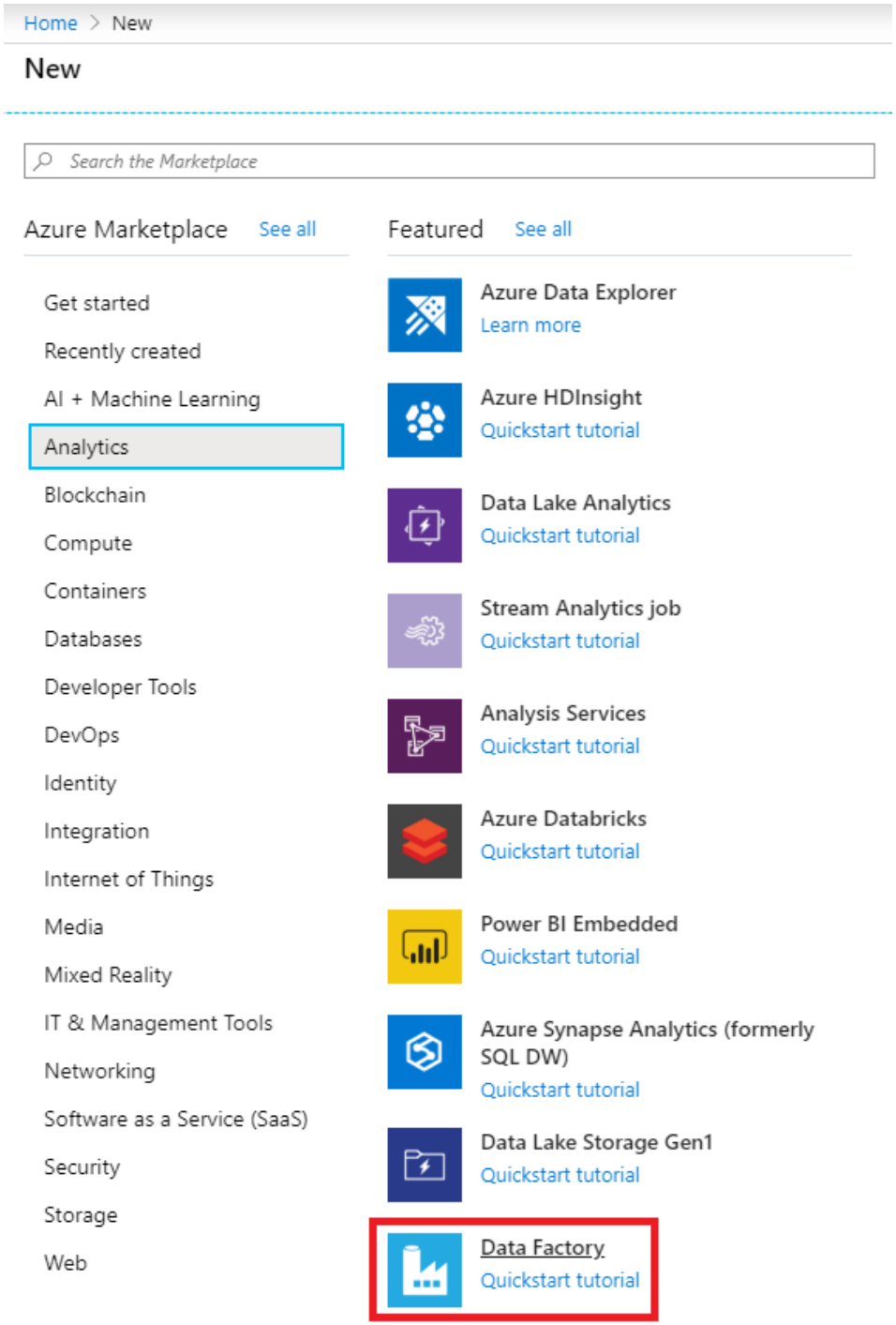

Abbildung 2.25: Auswählen von Azure Data Factory

3. Füllen Sie das Formular **New data factory** (Neue Data Factory) aus, wie in der folgenden Abbildung dargestellt:

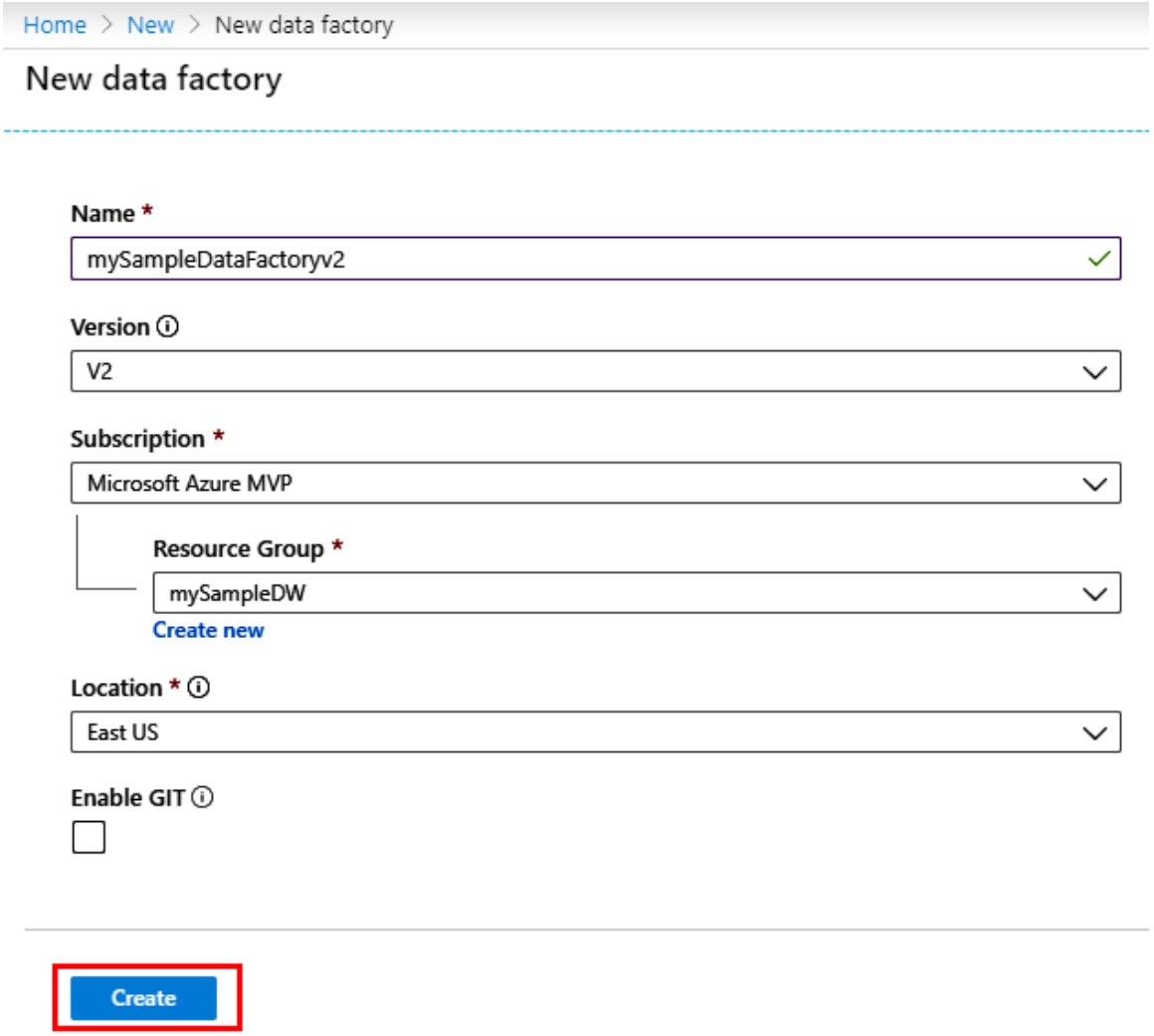

Abbildung 2.26: Hinzufügen von Details zum Erstellen der Data Factory

Im Folgenden sehen Sie die vorgeschlagenen Werte und ihre Erläuterungen:

Setting	Empfohlener Wert	Beschreibung
Name	mySampleDataFactoryv2	Geben Sie einen global eindeutigen Namen für Ihre Azure Data Factory an. Wenn der Name bereits verwendet wird, versuchen Sie es mit einem anderen
Abonnement	Ihr Azure-Abonnement	Wählen Sie Ihr Azure-Abonnement.
Ressourcengruppe	mySampleDW	Wählen Sie die Ressourcengruppe aus, die bei der Bereitstellung des Azure SQL Data Warehouse erstellt wurde.
Version	V2	Wählen Sie Data Factory V2.
Ort	USA, Osten	Wählen Sie den Speicherort der Data Factory aus.

Abbildung 2.27: Vorgeschlagene Werte für die Erstellung der Data Factory

4. Klicken Sie auf **Create** (Erstellen), um die Bereitstellung Ihrer Azure Data Factory zu starten.

Nachdem Sie die Azure Data Factory erfolgreich bereitgestellt haben, ist der nächste Schritt die Bereitstellung von Azure Data Lake Storage Gen2. Sie werden die beiden Technologien später in einer anderen Übung integrieren.

Bereitstellen Ihres Azure Data Lake Storage Gen2

Führen Sie nun die folgenden Schritte aus, um Azure Data Lake Storage Gen2 bereitzustellen:

1. Klicken Sie oben links im Azure-Portal auf **Create a resource** (Ressource erstellen):

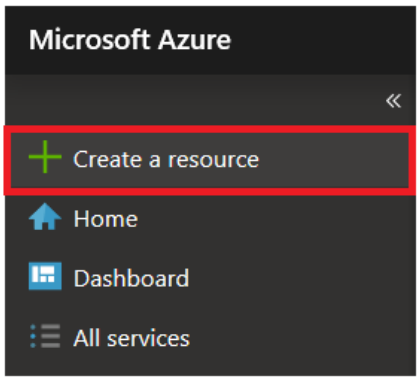

Abbildung 2.28: Erstellen einer Ressource im Azure-Portal

2. Wählen Sie **Storage** (Speicher) im Abschnitt „Azure Marketplace“ auf der Seite **New** (Neu) und anschließend **Storage account** (Speicherkonto) im Abschnitt **Featured** (Empfohlen) aus:

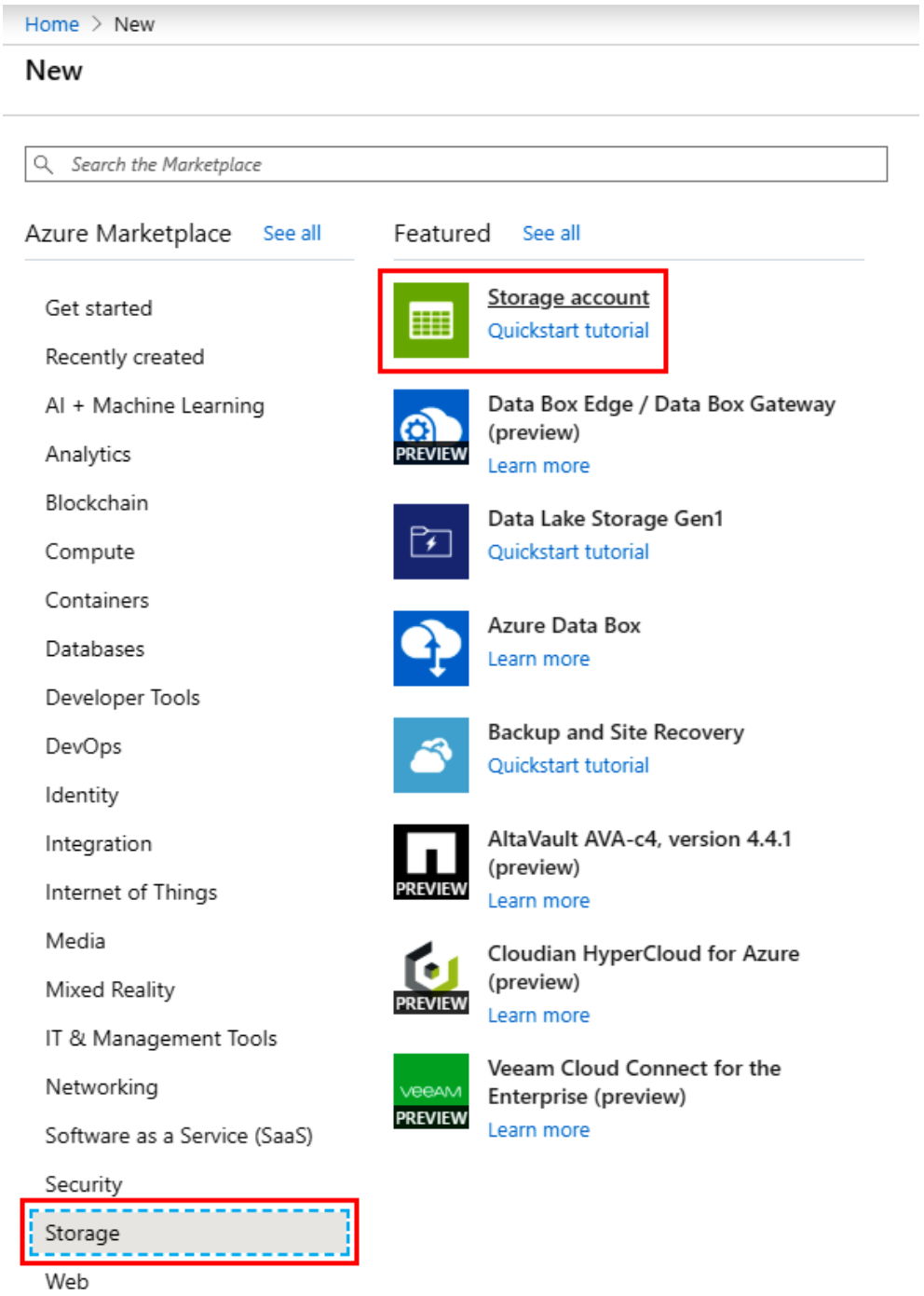

Abbildung 2.29: Erstellen eines Speicherkontos

3. Tragen Sie in das Formular **Create storage account** (Speicherkonto erstellen) die unten angegebenen Informationen ein. Sie müssen Ihr eigenes Abonnement auswählen und einen eindeutigen Namen für das Speicherkonto angeben. Wählen Sie als Ressourcengruppe die Ressourcengruppe aus, die Sie zuvor bei der Bereitstellung von Azure Synapse Analytics (früher Azure SQL Data Warehouse) erstellt haben.

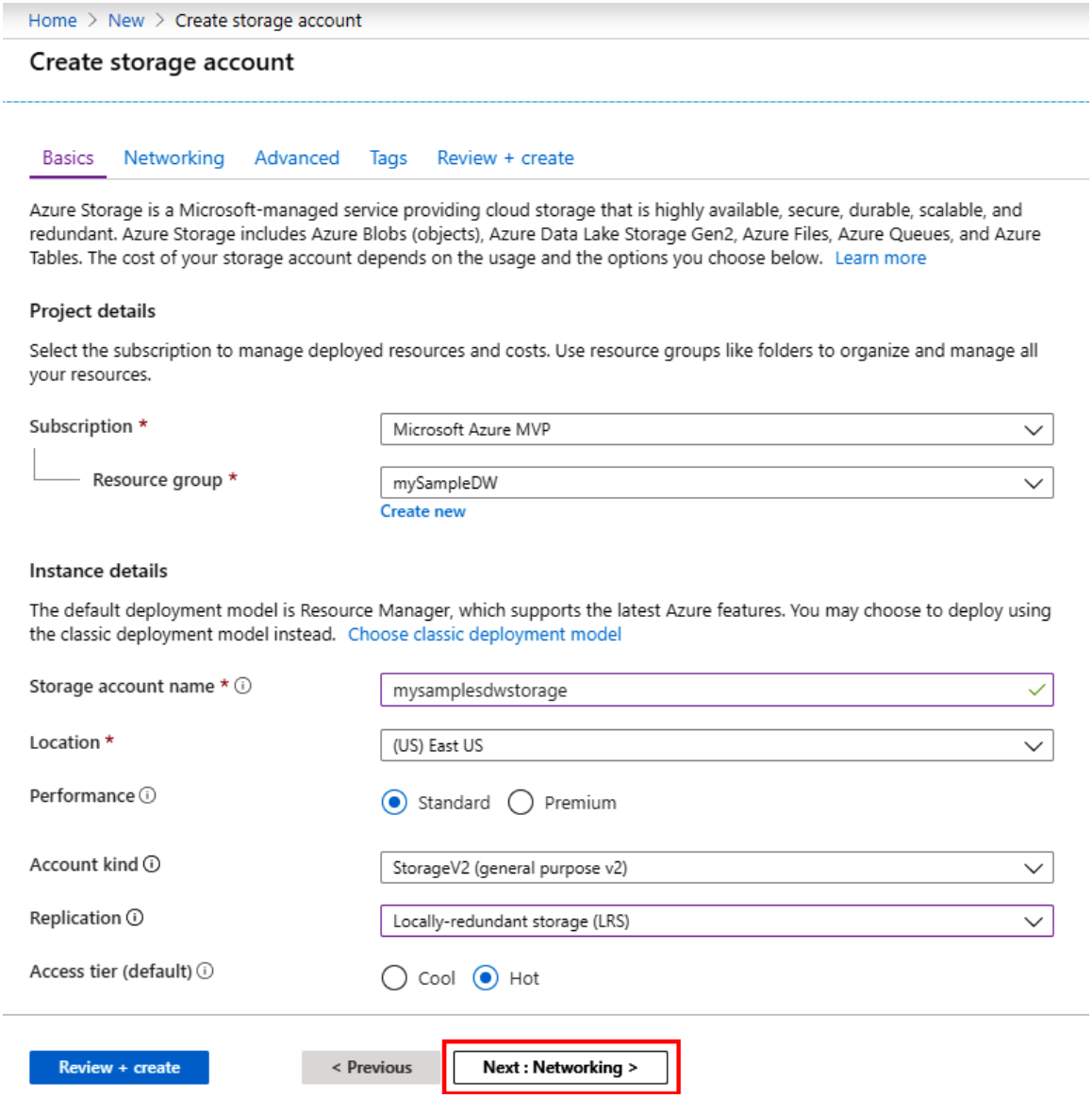

Abbildung 2.30: Speicherkontodetails

4. Legen Sie als Netzwerkverbindungsmethode **Public endpoint (all networks** (Öffentlicher Endpunkt (alle Netzwerke)) fest, und klicken Sie auf die Schaltfläche **Next: Advanced >** (Weiter: Erweitert >).

5. Legen Sie im Abschnitt **Data Lake Storage Gen2** für **Hierarchical namespace** (Hierarchischer Namespace) **Enabled** (Aktiviert) fest. Klicken Sie anschließend auf **Review + create** (Prüfen + erstellen):

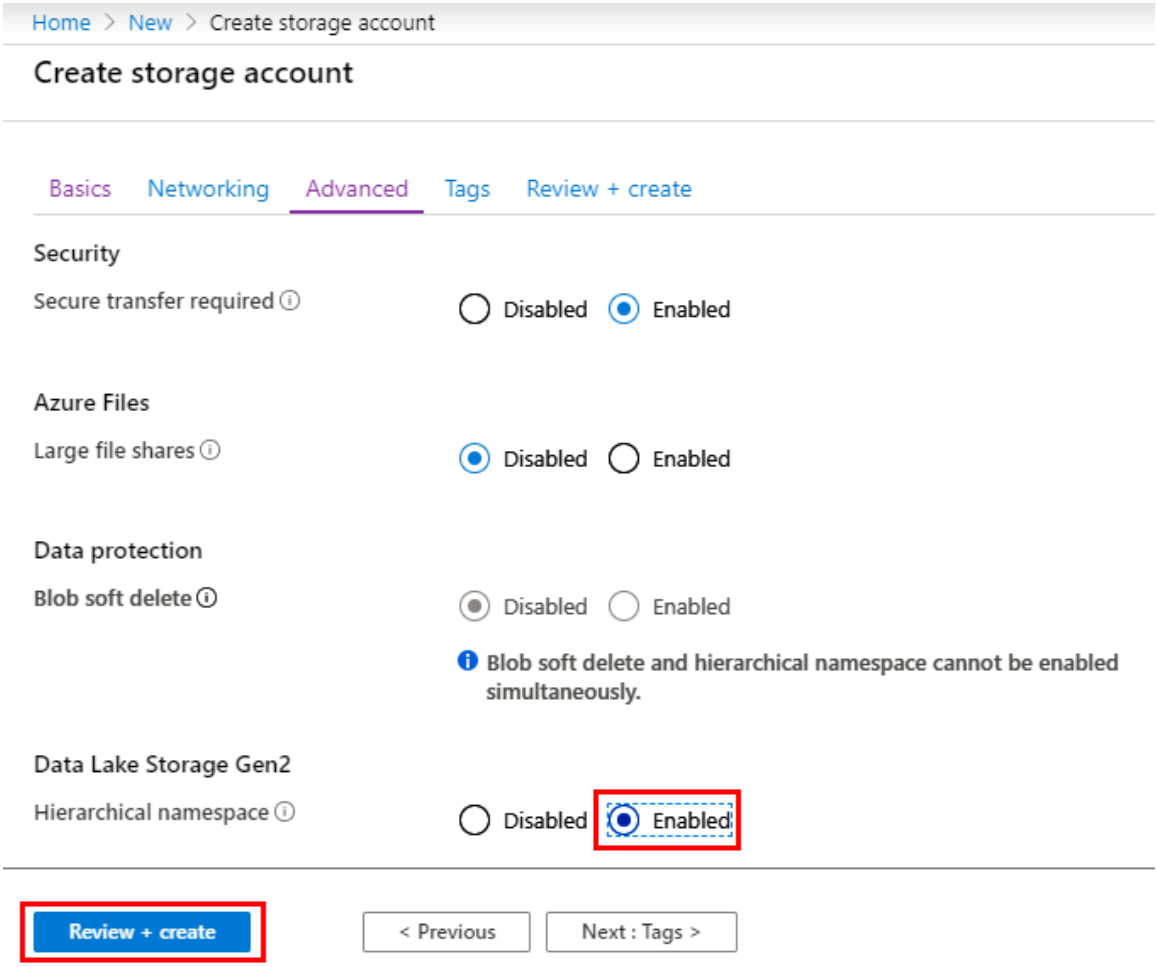

Abbildung 2.31: Aktivieren des hierarchischen Namespace von Data Lake Storage

6. Führen Sie eine abschließende Überprüfung durch, und klicken Sie dann auf **Create** (Erstellen), um mit der Bereitstellung Ihres Azure Data Lake Storage Gen2-Kontos zu beginnen.

Wie Sie in den schrittweisen Anleitungen oben sehen können, ist das Erstellen eines Azure Data Lake Storage Gen2 so unkompliziert, wie ein Azure Storage-Konto zu erstellen.

Integration von Azure Data Factory mit Azure Data Lake Storage Gen2

Nachdem Sie in den vorherigen Übungen die Azure Data Factory und den Azure Data Lake Storage Gen2 bereitgestellt haben, können Sie jetzt die beiden Technologien integrieren.

Zunächst werden Sie eine **`JSON`**-Datei mithilfe der Azure Data Factory erfassen, die Daten extrahieren und in das **`CSV`**-Dateiformat transformieren. Laden Sie dann die so entstandene CSV-formatierte Datei in Azure Data Lake Storage Gen2.

1. Wechseln Sie zu Azure Data Factory, und starten Sie die Datenintegrationsanwendung durch Klicken auf **Author & Monitor** (Erstellen und überwachen).

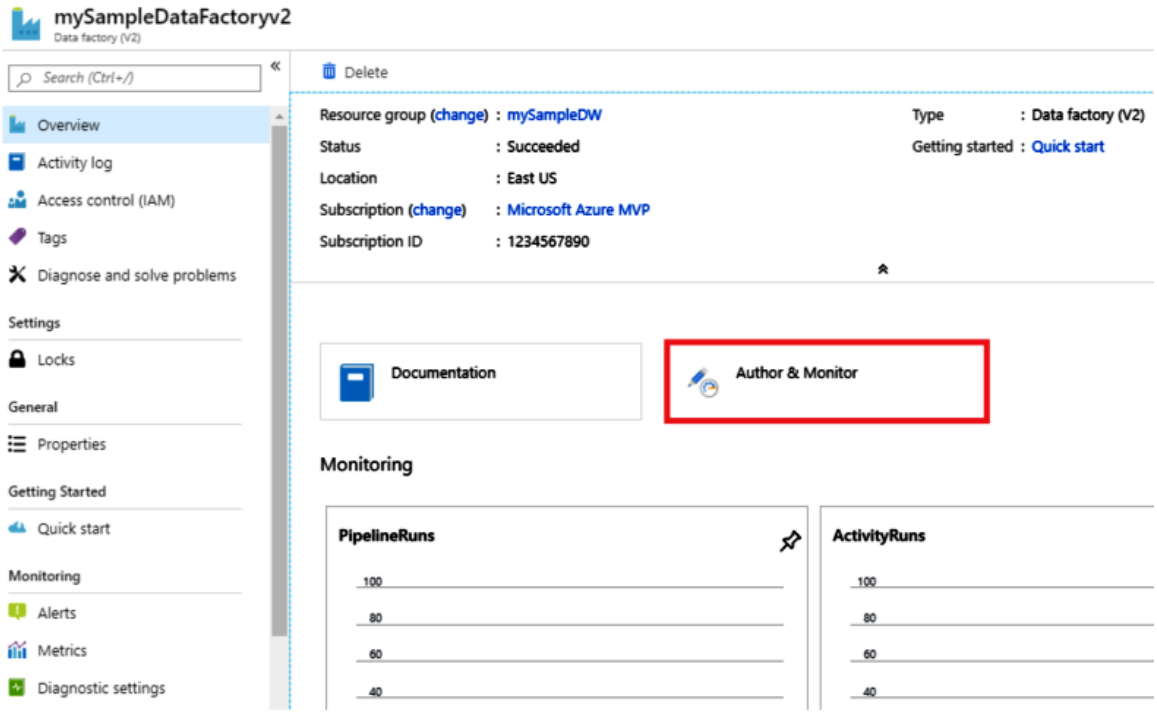

Abbildung 2.32: Starten der Datenintegrationsanwendung

2. Die Datenintegrationsanwendung wird in einer separaten Browser-Registerkarte gestartet. Klicken Sie auf **Copy Data** (Daten kopieren):

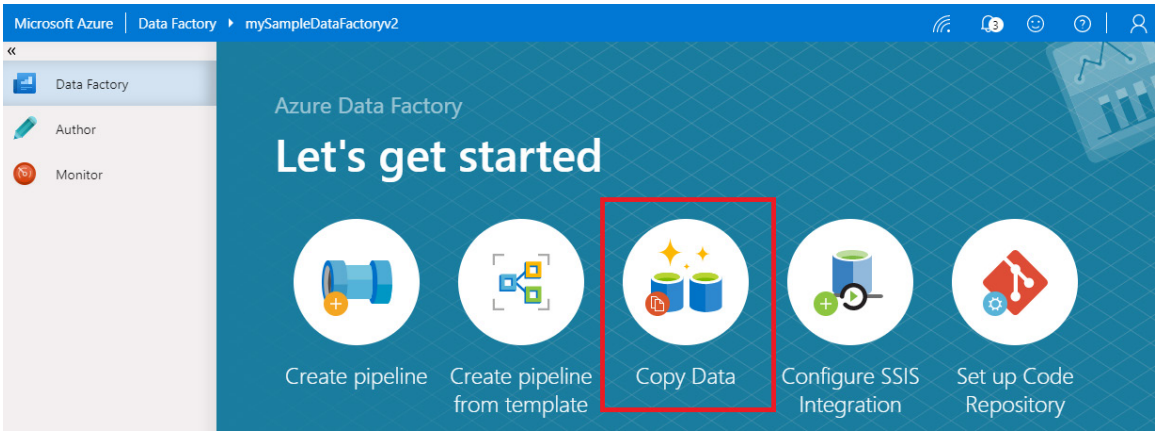

Abbildung 2.33: Initiieren der Aufgabe „Copy data“ (Daten kopieren)

3. Machen Sie für Ihre Kopierpipeline eine Angabe unter **Task name** (Aufgabenname) (und gegebenenfalls unter **Task description** (Aufgabenbeschreibung)), und klicken Sie dann auf **Next** (Weiter):

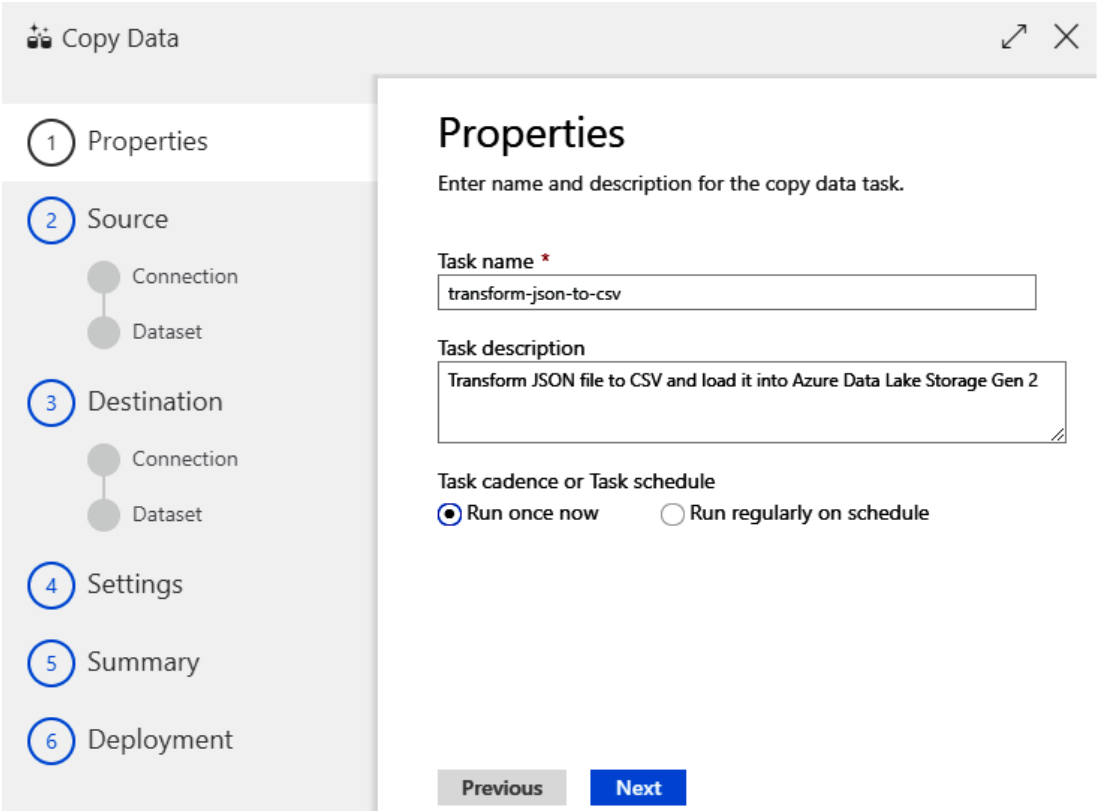

Abbildung 2.34: Hinzufügen von Details zur Aufgabe „Copy data“ (Daten kopieren)

Die Azure Data Factory besteht aus einem Dataset. Hierbei handelt es sich im Grunde um strukturierte Daten im Datenspeicher. Eine Pipeline umfasst Aktivitäten, die zur Ausführung einer Aufgabe logisch verbunden sind. Mit einem verknüpften Dienst können Sie Azure Data Factory mit verschiedenen Datenquellen verbinden. Im nächsten Schritt erstellen Sie einen neuen verknüpften Dienst für die neue Datenquelle.

4. Klicken Sie auf **Create new connection** (Neue Verbindung erstellen):

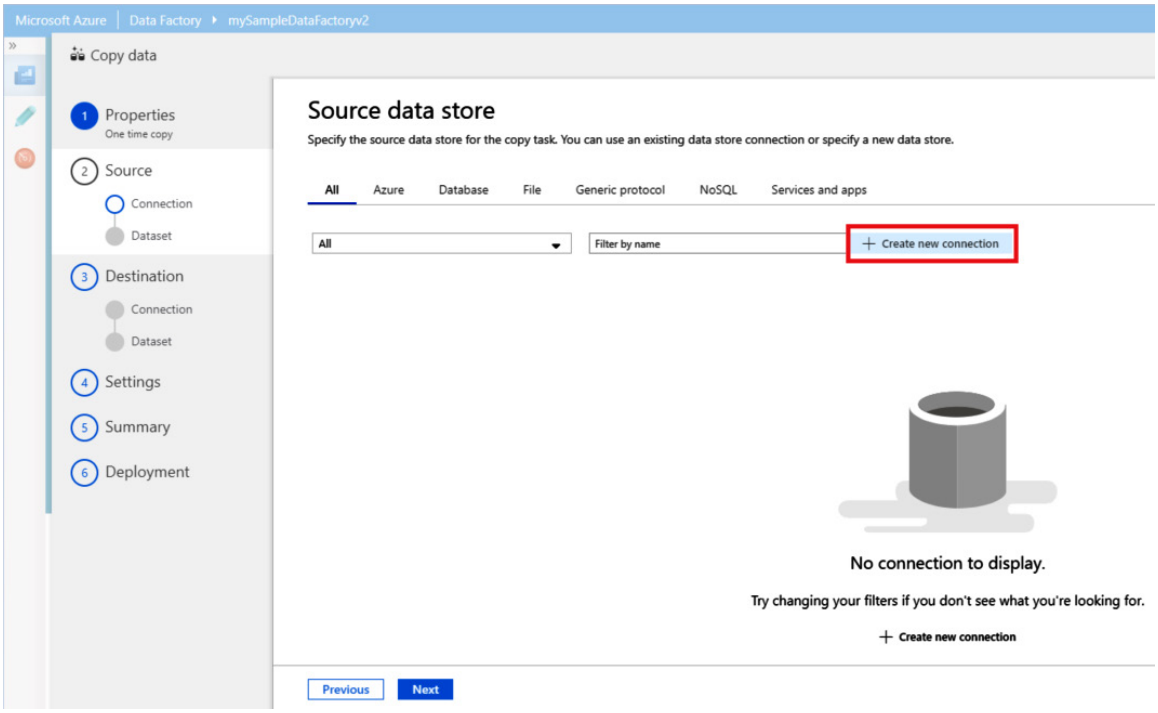

Abbildung 2.35: Erstellen einer neuen Verbindung

5. Geben Sie **HTTP** in das Suchfeld ein, und klicken Sie im Suchergebnis auf die Schaltfläche **HTTP**:

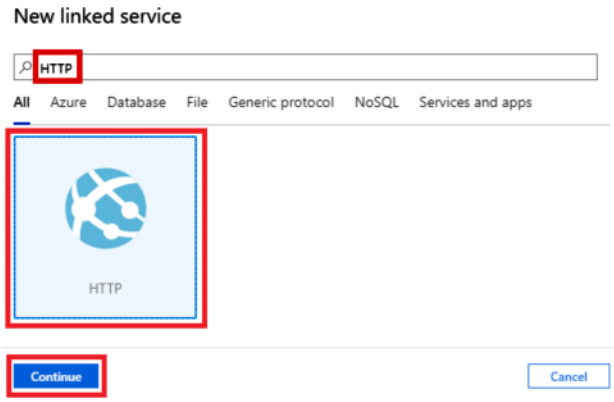

Abbildung 2.36: Auswählen des HTTP-Diensts

6. Füllen Sie das Formular **New Linked Service (HTTP)** (Neuer verknüpfter Dienst (HTTP)) aus, um den Quelldatenspeicher (**Source data store**) zu definieren, wie in der folgenden Abbildung dargestellt. Geben Sie als Basis-URL diese URL ein.

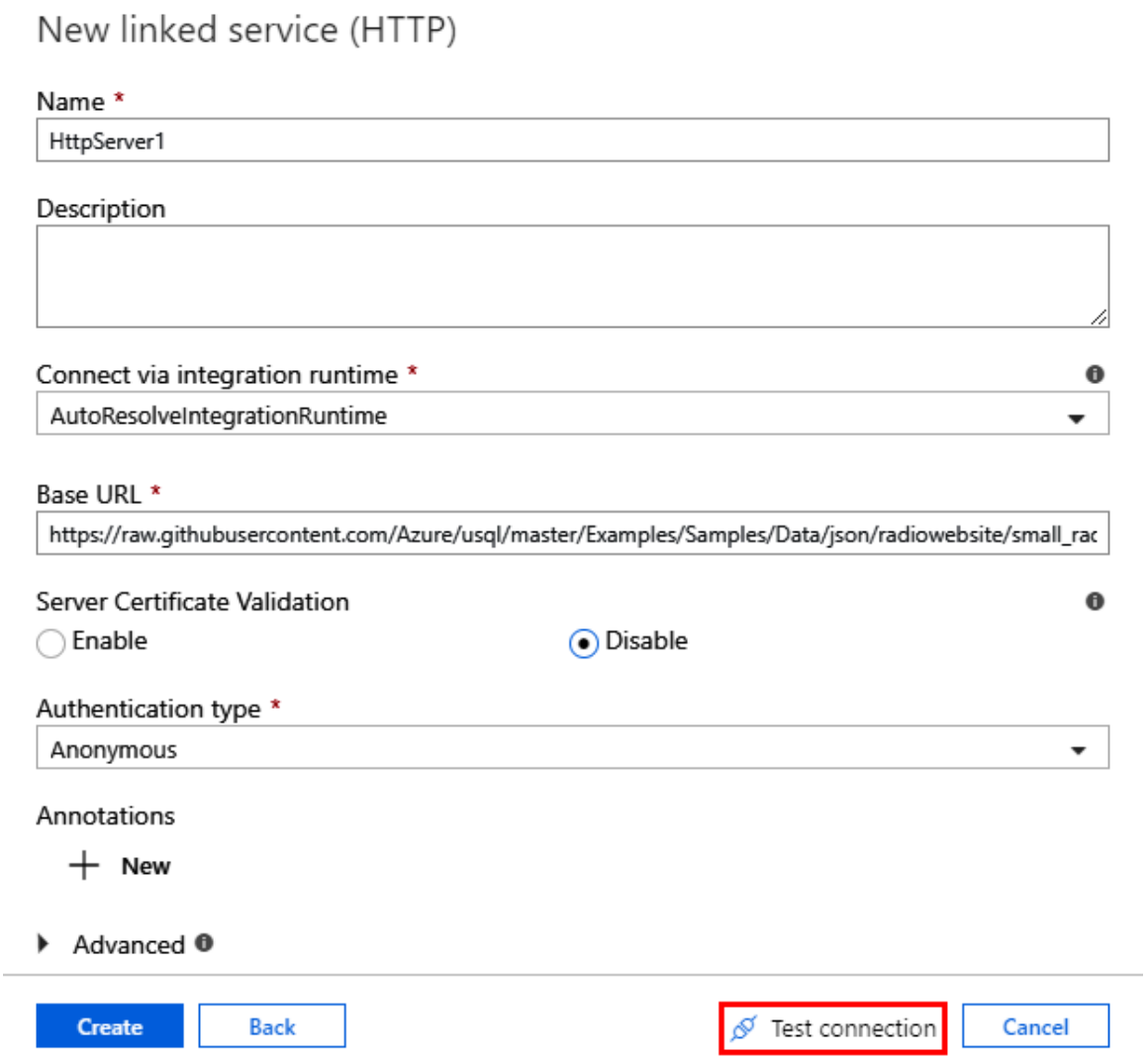

Abbildung 2.37: Testen der Verbindung des neuen verknüpften Diensts

7. Wenn Ihre Testverbindung erfolgreich ist, klicken Sie auf **Create** (Erstellen). Nehmen Sie andernfalls Korrekturen an Ihren Einträgen in diesem Formular vor, und testen Sie die Verbindung erneut:

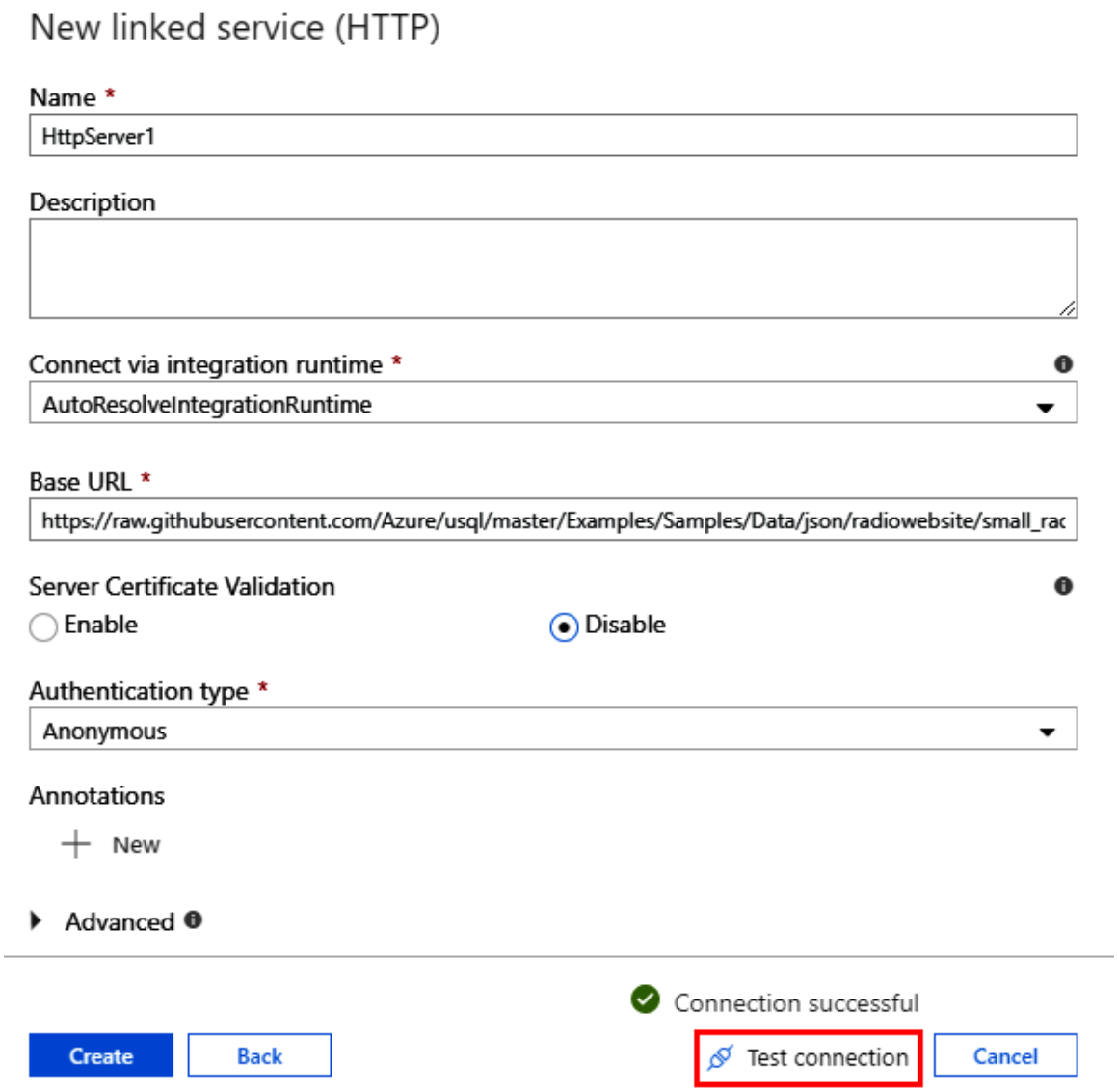

Abbildung 2.38: Erstellen des HTTP-Diensts nach erfolgreicher Testverbindung

8. Klicken Sie auf die Schaltfläche **Next** (Weiter), bis Sie auf die Seite mit den Dateiformateinstellungen gelangen:

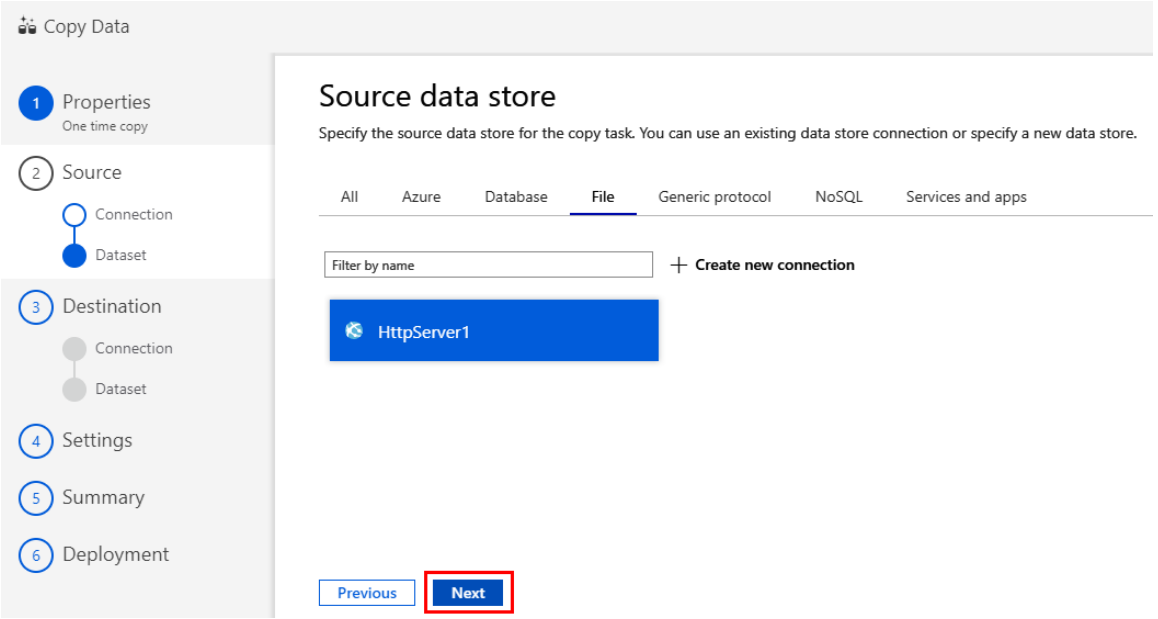

Abbildung 2.39: Navigieren zur Seite mit den Dateiformateinstellungen

9. Bestätigen Sie die Angaben unter **File format settings** (Dateiformateinstellungen), wie in der folgenden Abbildung dargestellt:

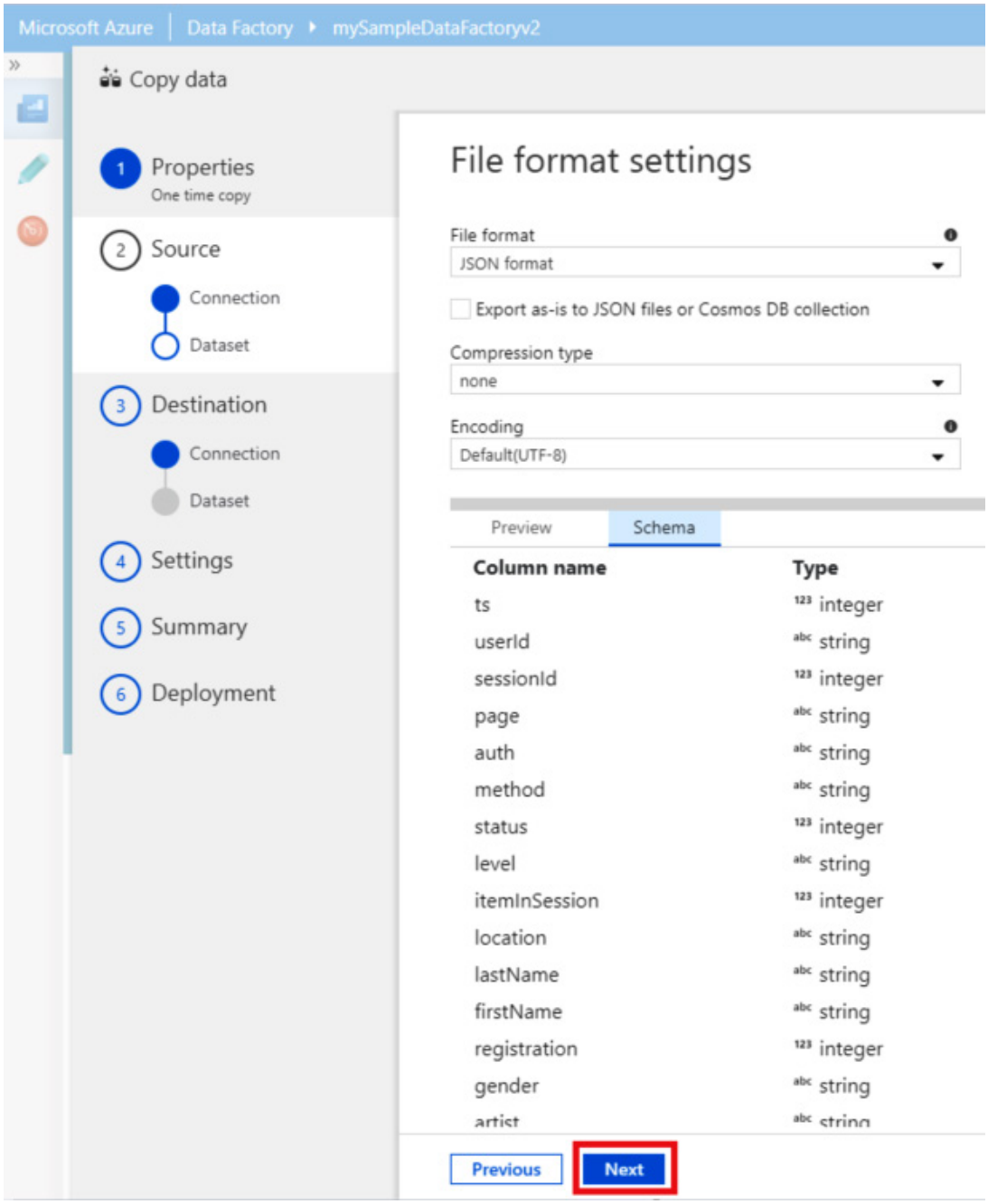

Abbildung 2.40: Seite „File format settings“ (Dateiformateinstellungen)

Die Einstellungen können wie folgt in einer Vorschau angezeigt werden:

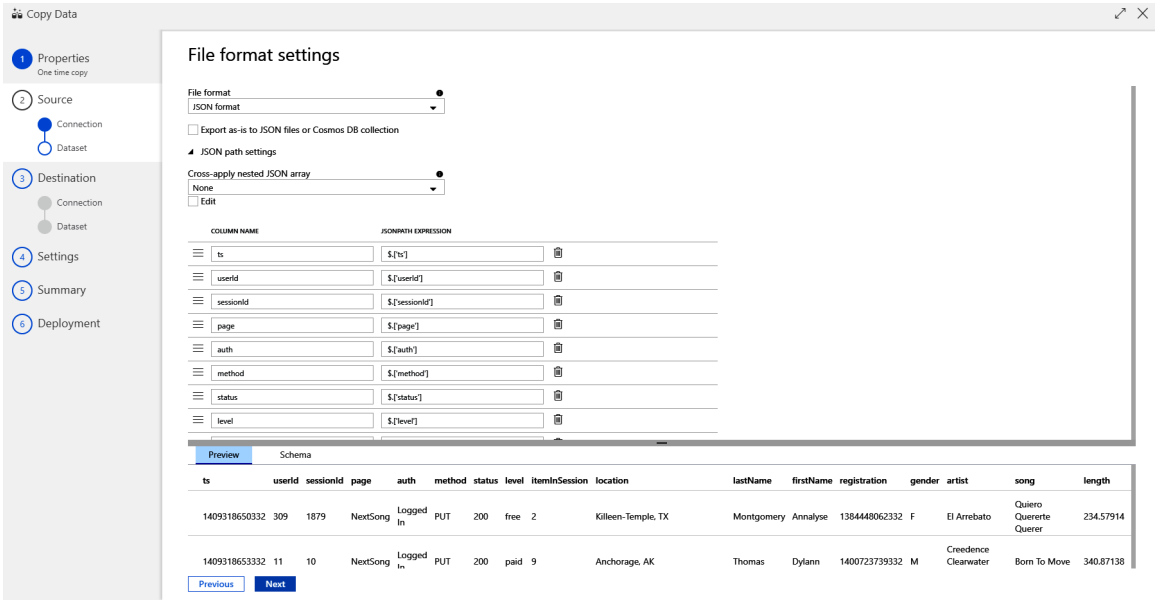

Abbildung 2.41: Vorschau der Dateiformateinstellungen

10. Klicken Sie nun auf **Create new connection** (Neue Verbindung erstellen), um den Zieldatenspeicher (**Destination data store**) zu konfigurieren:

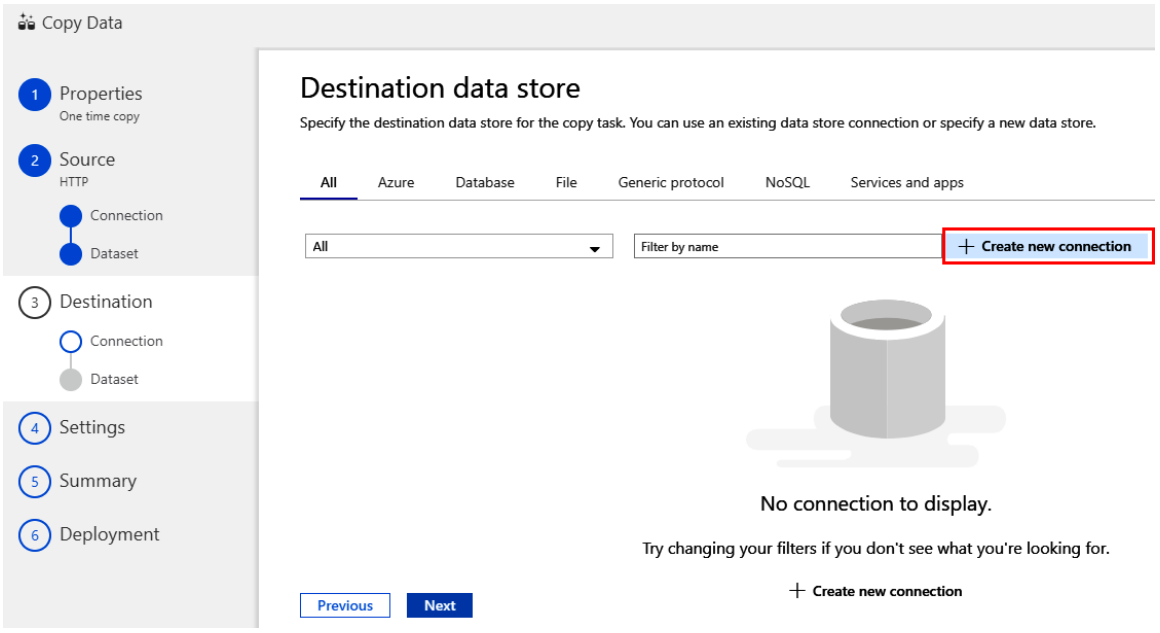

Abbildung 2.42: Konfigurieren des Zieldatenspeichers

11. Wählen Sie **Azure Data Lake Storage Gen2** aus, und klicken Sie auf **Continue** (Weiter):

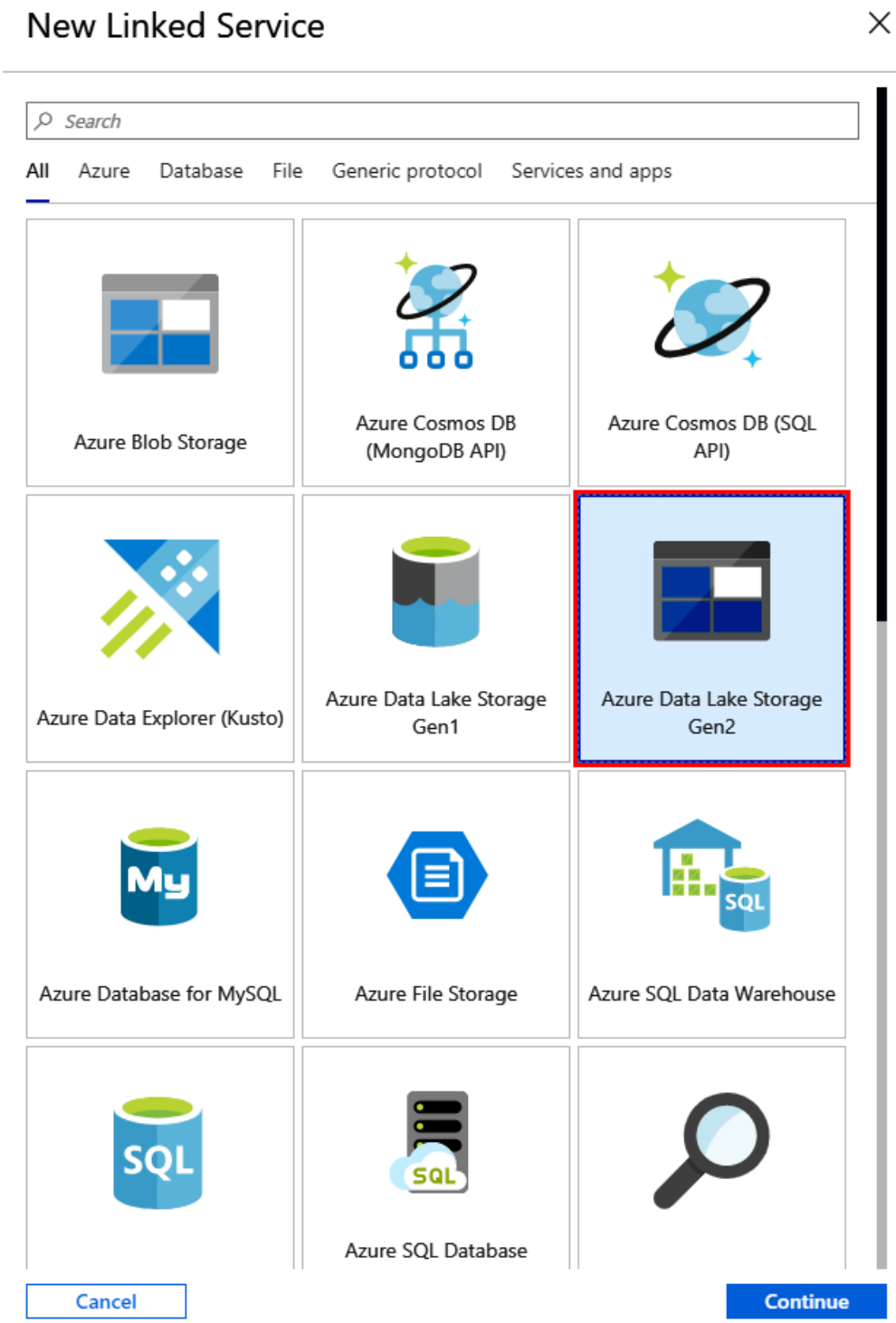

Abbildung 2.43: Auswählen von Data Lake Storage Gen2

12. Wählen Sie **Azure Data Lake Storage Gen2** (in der vorherigen Übung bereitgestellt) im Abschnitt **New Linked Service** (Neuer verknüpfter Dienst) aus:

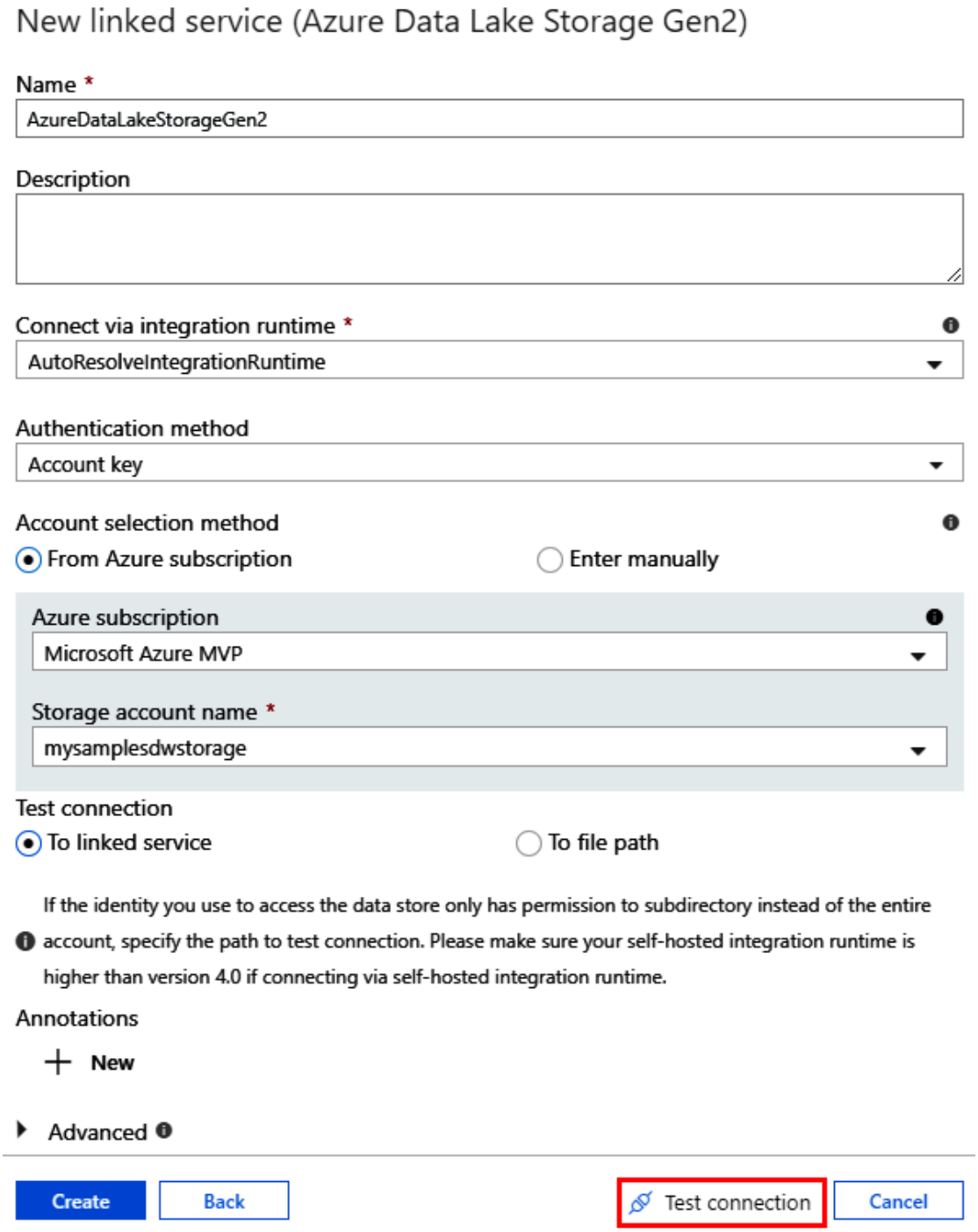

Abbildung 2.44: Testen der Verbindung des bereitgestellten Data Lake Storage

13. Wenn Ihre Testverbindung erfolgreich ist, klicken Sie auf **Create** (Erstellen). Nehmen Sie andernfalls Korrekturen an Ihren Einträgen in dem Formular vor, und testen Sie die Verbindung erneut.

14. Klicken Sie auf **Next** (Weiter):

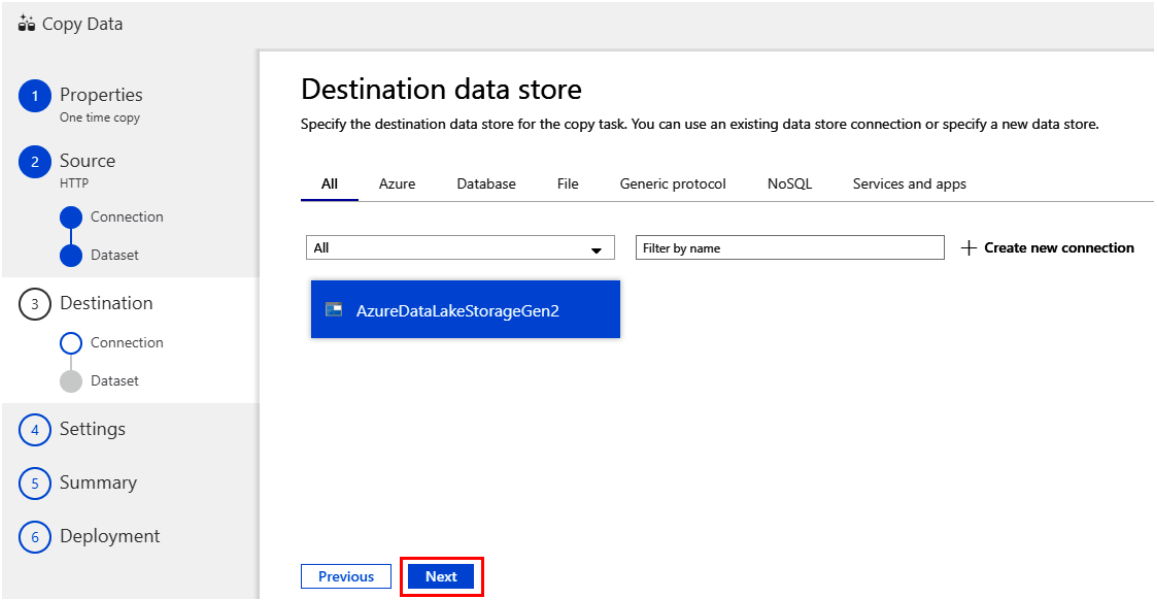

Abbildung 2.45: Angeben des Zieldatenspeichers

15. Füllen Sie das Formular **Choose the output file or folder** (Ausgabedatei oder Ordner auswählen) wie in der folgenden Abbildung dargestellt aus, und klicken Sie auf **Next** (Weiter):

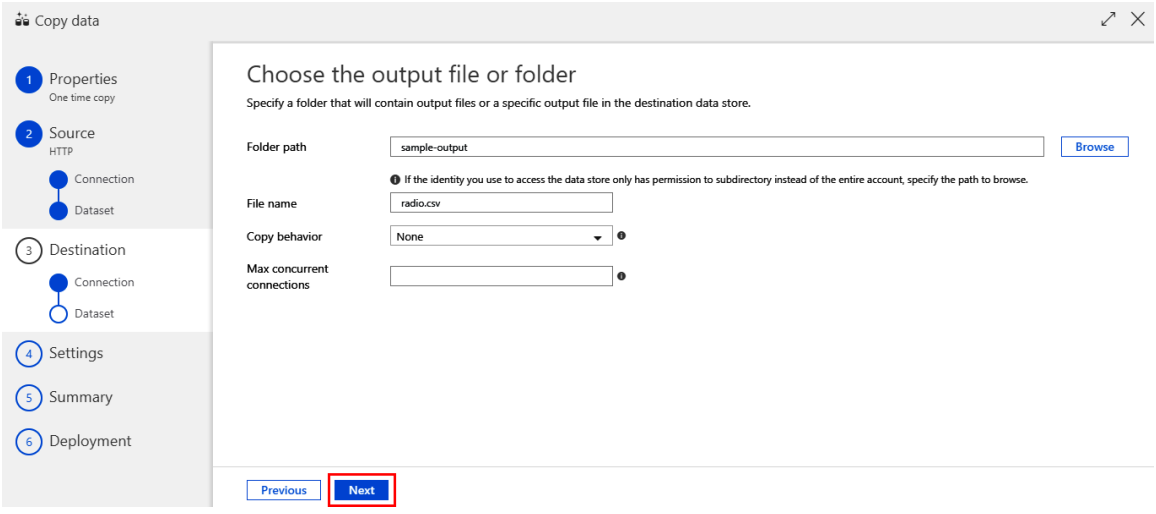

Abbildung 2.46: Angeben des Ordners für Ausgabedateien

16. Füllen Sie das Formular **File format settings** (Dateiformateinstellungen) wie in der folgenden Abbildung dargestellt aus, und klicken Sie auf **Next** (Weiter):

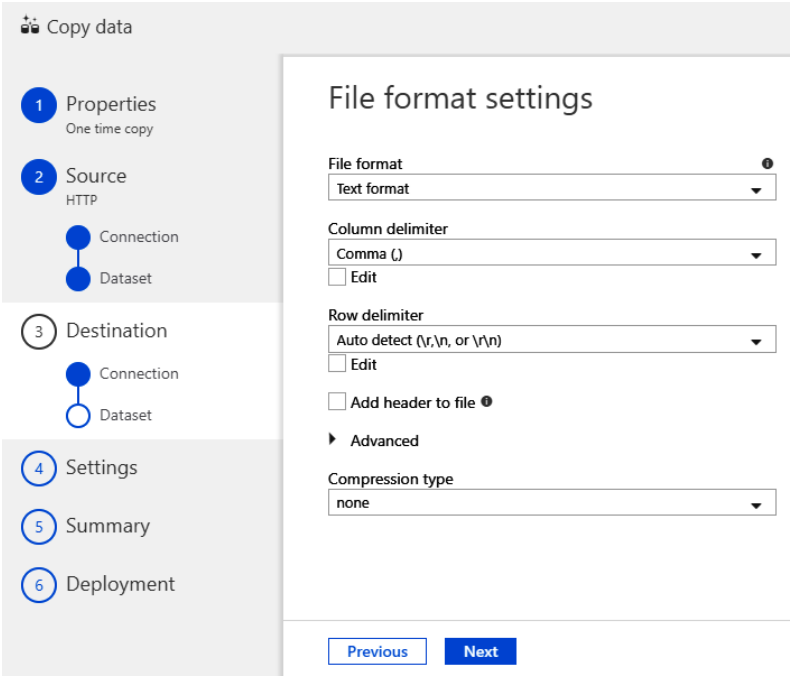

Abbildung 2.47: Angeben der Dateiformateinstellungen

17. Übernehmen Sie die standardmäßige Schemazuordnung (**Schema mapping**) und klicken Sie auf **Next** (Weiter):

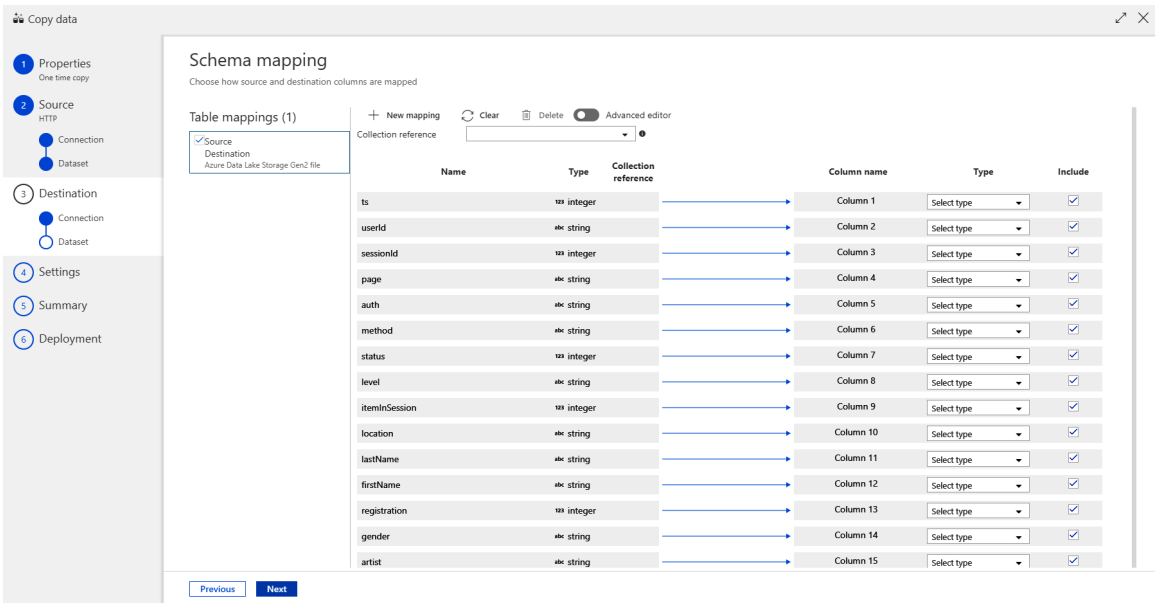

Abbildung 2.48: Standardmäßige Schemazuordnung

18. Bestätigen Sie die Angaben unter **Settings** (Einstellungen), und klicken Sie auf **Next** (Weiter):

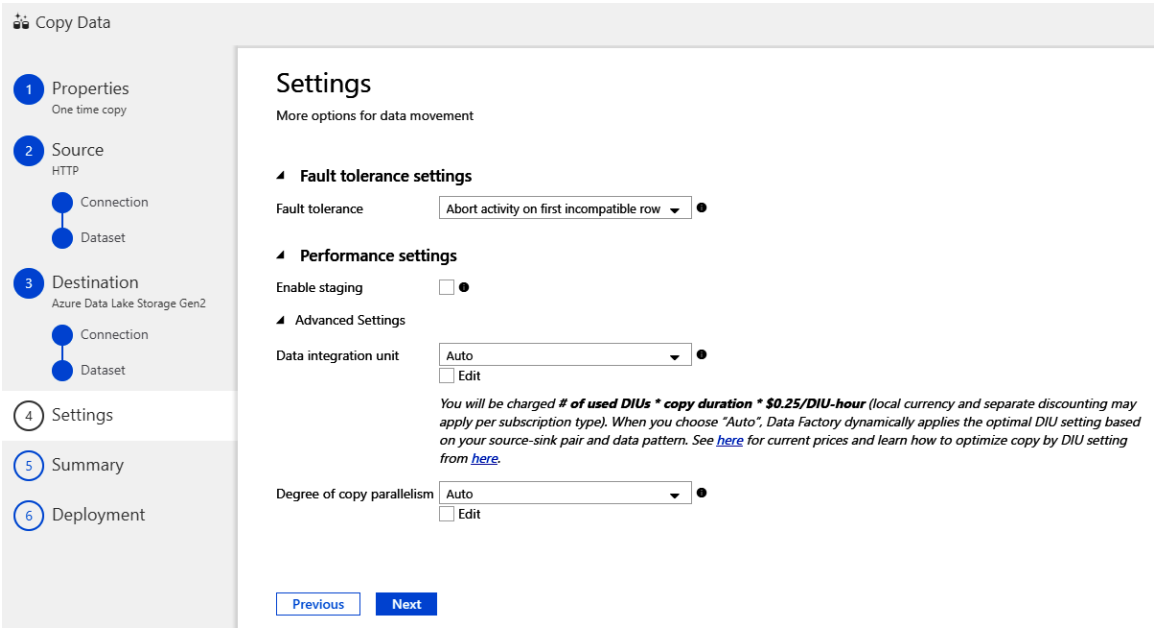

Abbildung 2.49: Bestätigen der Fehlertoleranz und der erweiterten Einstellungen

19. Überprüfen Sie die Angaben unter **Summary** (Übersicht)), und klicken Sie auf **Next** (Weiter):

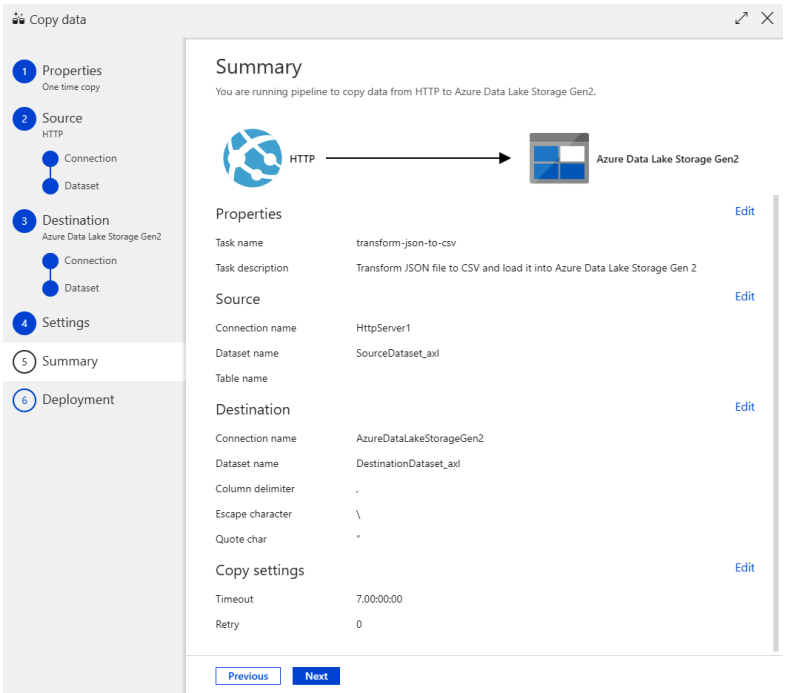

Abbildung 2.50: Kopieren der Zusammenfassung der Datenpipeline

20. Beenden Sie die Bereitstellung, indem Sie auf **Finish** (Fertig stellen) klicken:

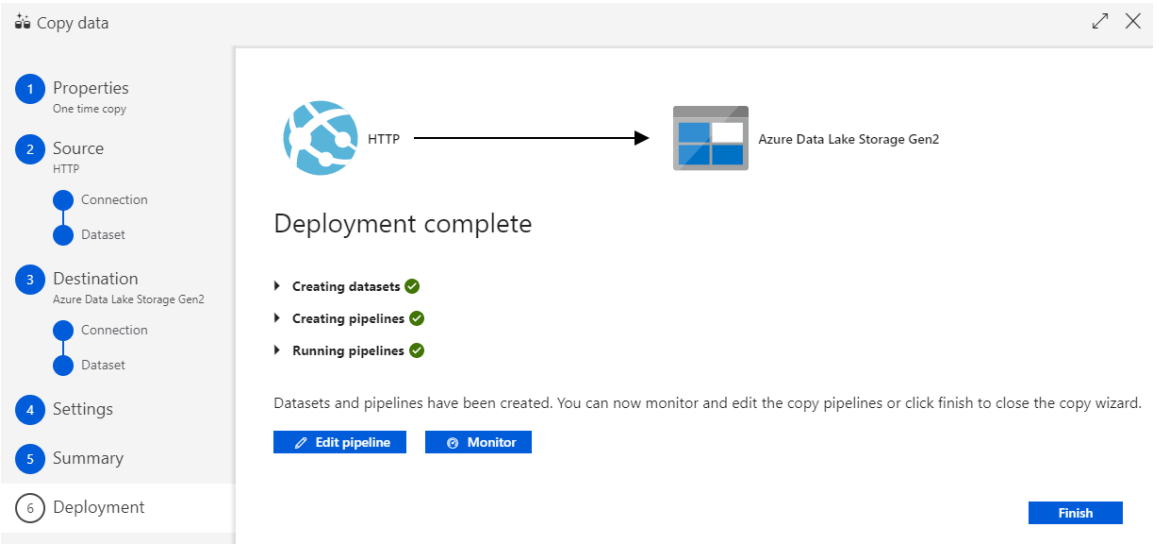

Abbildung 2.51: Abschließen der Bereitstellung

Überprüfen des Ergebnisses in Azure Data Lake Storage Gen2

Zu diesem Zeitpunkt wurde die Kopierpipeline in Azure Data Factory bereits ausgeführt.

1. Wechseln Sie zu Azure Data Lake Storage Gen2, um das Ergebnis zu überprüfen:

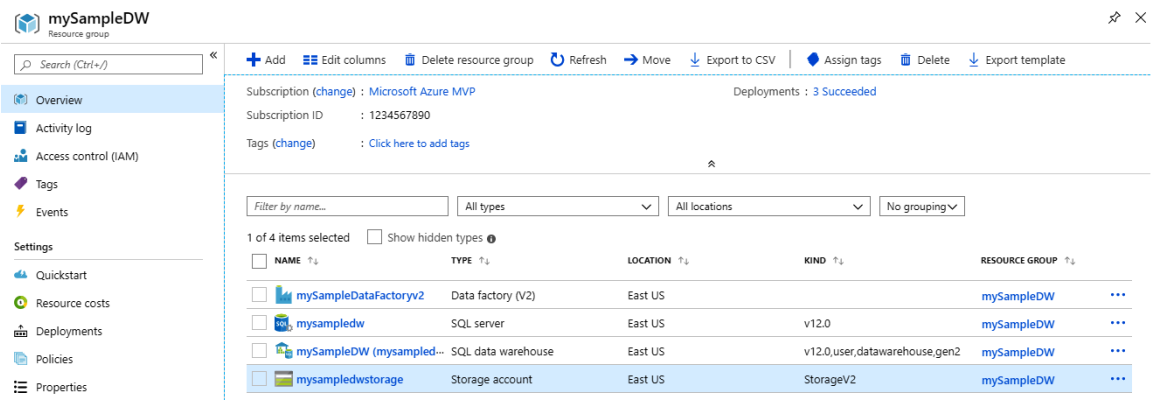

Abbildung 2.52: Überprüfen des Ergebnisses

2. Mit dem integrierten Storage Explorer (Vorschau) können Sie die resultierende Datei aus Azure Data Lake Storage Gen2 anzeigen:

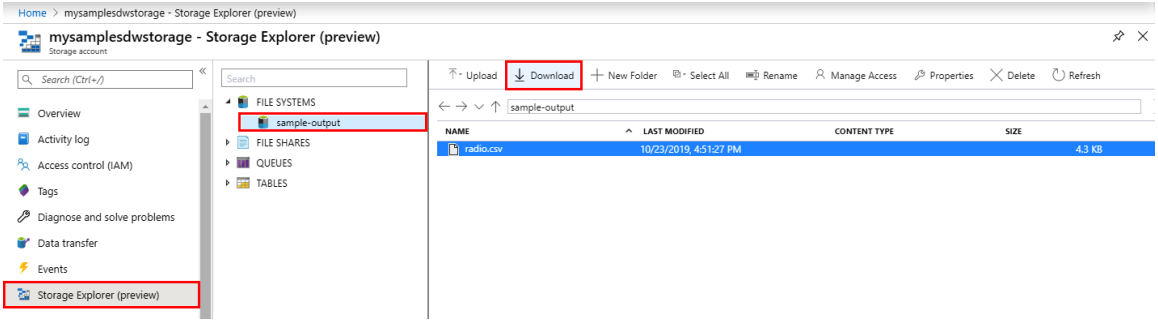

Abbildung 2.53: Herunterladen einer Kopie Ihrer CSV-Datei

3. Klicken Sie auf **Download** (Herunterladen), um eine Kopie von **`radio.csv`** herunterzuladen, und vergleichen Sie die „Vorher“- und „Nachher“-Versionen.

Hier sehen Sie die „Vorher“-Version der Datei im ursprünglichen JSON-Format:

https://raw.githubusercontent.com/Azure/usql/master/Examples/Samples/Data/json/radiowebsite/small_radio_json.json


```
{"ts":1409318650332,"userId":"309","sessionId":1879,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"fr
TX","lastName":"Montgomery","firstName":"Annalyse","registration":1384448062332,"gender":"F","artist":"El Arrebato","song":"Quie
{"ts":1409318653332,"userId":"11","sessionId":10,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"paid"
AK","lastName":"Thomas","firstName":"Dylann","registration":1400723739332,"gender":"M","artist":"Creedence Clearwater Revival","
{"ts":1409318685332,"userId":"201","sessionId":2047,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"pa
PA","lastName":"Watts","firstName":"Liam","registration":1406279422332,"gender":"M","artist":"Gorillaz","song":"DARE","length":2
{"ts":1409318686332,"userId":"779","sessionId":2136,"page":"Home","auth":"Logged In","method":"GET","status":200,"level":"free",
TN","lastName":"Townsend","firstName":"Tess","registration":1406970190332,"gender":"F"}
{"ts":1409318697332,"userId":"401","sessionId":400,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"fre
GA","lastName":"Smith","firstName":"Margaux","registration":1406191211332,"gender":"F","artist":"Otis Redding","song":"Send Me S
{"ts":1409318714332,"userId":"521","sessionId":520,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"pai
WI","lastName":"Morse","firstName":"Alan","registration":1401760632332,"gender":"M","artist":"Slightly Stoopid","song":"Mellow M
{"ts":1409318743332,"userId":"244","sessionId":2261,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"fr
CA","lastName":"Shelton","firstName":"Gabriella","registration":1389460542332,"gender":"F","artist":"NOFX","song":"Linoleum","le
{"ts":1409318804332,"userId":"969","sessionId":968,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"pai
MI","lastName":"Williams","firstName":"Elijah","registration":1388691347332,"gender":"M","artist":"Nirvana","song":"The Man Who
{"ts":1409318832332,"userId":"401","sessionId":400,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"fre
GA","lastName":"Smith","firstName":"Margaux","registration":1406191211332,"gender":"F","artist":"Aventura","song":"La Nina","len
{"ts":1409318891332,"userId":"779","sessionId":2136,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"fr
TN","lastName":"Townsend","firstName":"Tess","registration":1406970190332,"gender":"F","artist":"Harmonia","song":"Sehr kosmisch
{"ts":1409318912332,"userId":"521","sessionId":520,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"pai
WI","lastName":"Morse","firstName":"Alan","registration":1401760632332,"gender":"M","artist":"Spragga Benz","song":"Backshot","l
{"ts":1409318931332,"userId":"201","sessionId":2047,"page":"NextSong","auth":"Logged In","method":"PUT","status":200,"level":"pa
PA","lastName":"Watts","firstName":"Liam","registration":1406279422332,"gender":"M","artist":"Bananarama","song":"Love In The Fi
{"ts":1409318931332,"userId":"201","sessionId":2047,"page":"Home","auth":"Logged In","method":"GET","status":200,"level":"paid",
PA","lastName":"Watts","firstName":"Liam","registration":1406279422332,"gender":"M"}
```


Abbildung 2.54: Originaldatei im JSON-Format

Hier sehen Sie die transformierte Datei im CSV-Format, die jetzt in Azure Data Lake Storage Gen2 gespeichert ist:

```
1409318650332,"309",1879,"NextSong","Logged In","PUT",200,"free",2,"Killeen-Temple, TX","Montgomery",'
1409318653332,"11",10,"NextSong","Logged In","PUT",200,"paid",9,"Anchorage, AK","Thomas","Dylann",140
1409318685332,"201",2047,"NextSong","Logged In","PUT",200,"paid",11,"New York-Newark-Jersey City, NY-
1409318686332,"779",2136,"Home","Logged In","GET",200,"free",0,"Nashville-Davidson--Murfreesboro--Fra
1409318697332,"401",400,"NextSong","Logged In","PUT",200,"free",2,"Atlanta-Sandy Springs-Roswell, GA"
1409318714332,"521",520,"NextSong","Logged In","PUT",200,"paid",39,"Chicago-Naperville-Elgin, IL-IN-W
1409318743332,"244",2261,"NextSong","Logged In","PUT",200,"free",1,"San Jose-Sunnyvale-Santa Clara, C
1409318804332,"969",968,"NextSong","Logged In","PUT",200,"paid",0,"Detroit-Warren-Dearborn, MI","Will
1409318832332,"401",400,"NextSong","Logged In","PUT",200,"free",3,"Atlanta-Sandy Springs-Roswell, GA"
1409318891332,"779",2136,"NextSong","Logged In","PUT",200,"free",1,"Nashville-Davidson--Murfreesboro--
1409318912332,"521",520,"NextSong","Logged In","PUT",200,"paid",40,"Chicago-Naperville-Elgin, IL-IN-W
1409318931332,"201",2047,"NextSong","Logged In","PUT",200,"paid",12,"New York-Newark-Jersey City, NY-
1409318931332,"201",2047,"Home","Logged In","GET",200,"paid",13,"New York-Newark-Jersey City, NY-NJ-P
```

Abbildung 2.55: CSV-Datei, die in Azure Data Lake Storage Gen2 gespeichert ist

Sie haben jetzt Ihre erste Azure Data Factory-Pipeline zum Erfassen einer JSON-Datei, Extrahieren der Daten, Transformieren der Daten in CSV-Format und Laden der Datei in Azure Data Lake Storage Gen2 erstellt.

Bereitstellen Ihres Azure Databricks-Diensts

Wir werden Ihnen nun zeigen, wie Sie den Azure Databricks-Dienst bereitstellen. An späterer Stelle im Quick-Start-Leitfaden werden wir das Muster des modernen Data Warehouse in einer Übung vervollständigen. Dabei werden die Daten in Azure Data Lake Storage Gen2 erfasst und durch Bereinigung und Transformation mithilfe von Azure Databricks vorbereitet. Schließlich werden die bereinigten und transformierten Daten in Azure Data Warehouse geladen.

1. Klicken Sie oben links im Azure-Portal auf **Create a resource** (Ressource erstellen):

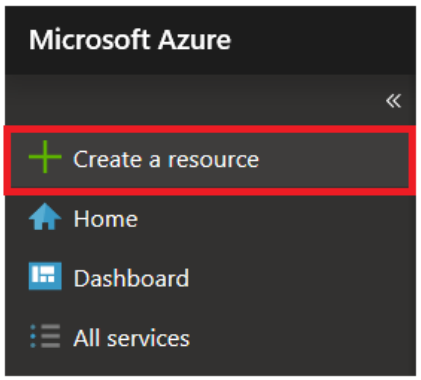

Abbildung 2.56: Erstellen einer Ressource

2. Wählen Sie **Analytics** im Abschnitt „Azure Marketplace“ auf der Seite **New** (Neu) und anschließend **Azure Databricks** im Abschnitt **Featured** (Empfohlen) aus:

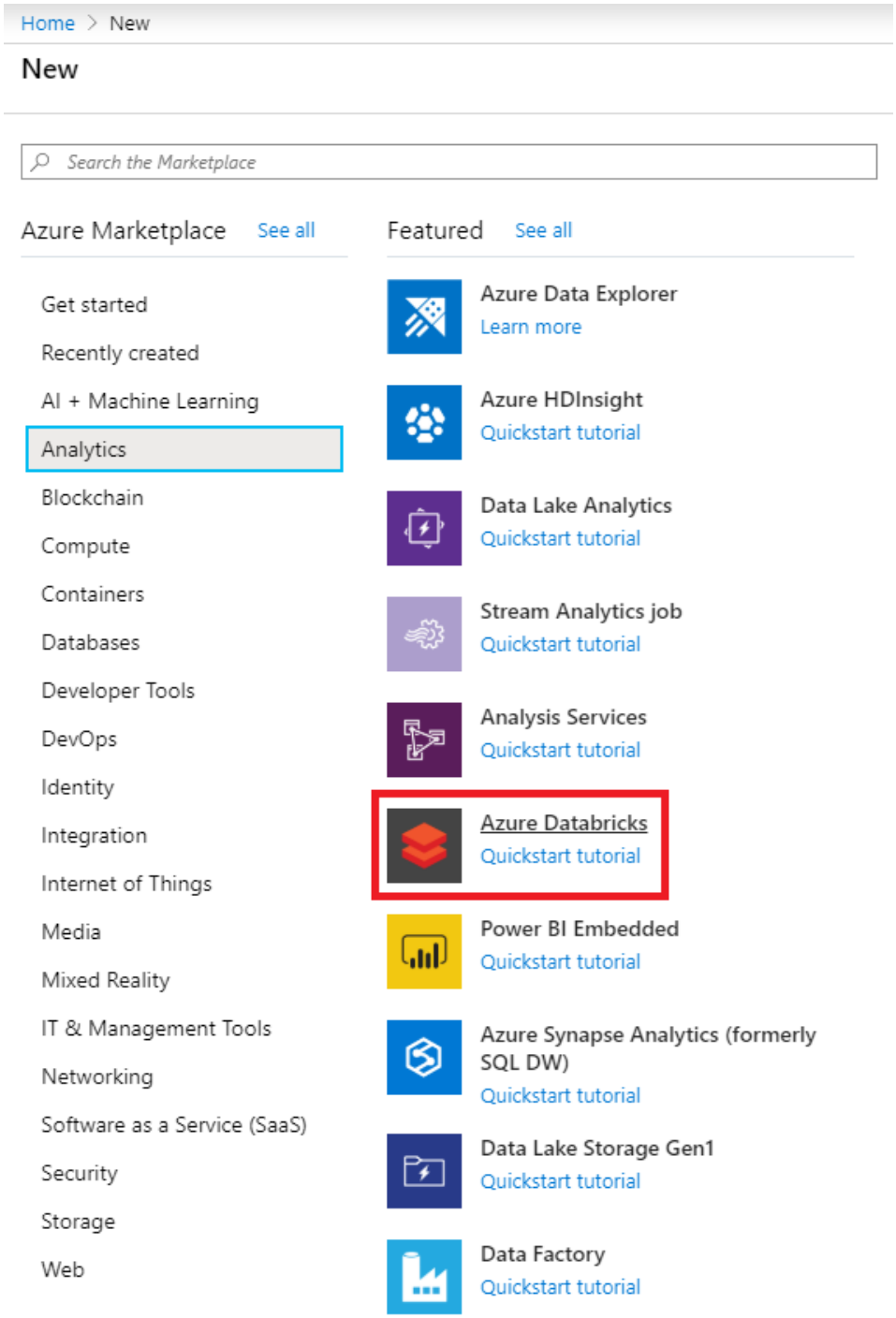

Abbildung 2.57: Auswählen von Databricks für die Bereitstellung

3. Füllen Sie das Formular **Azure Databricks Service** (Azure Databricks-Dienst) wie in der folgenden Abbildung dargestellt aus:

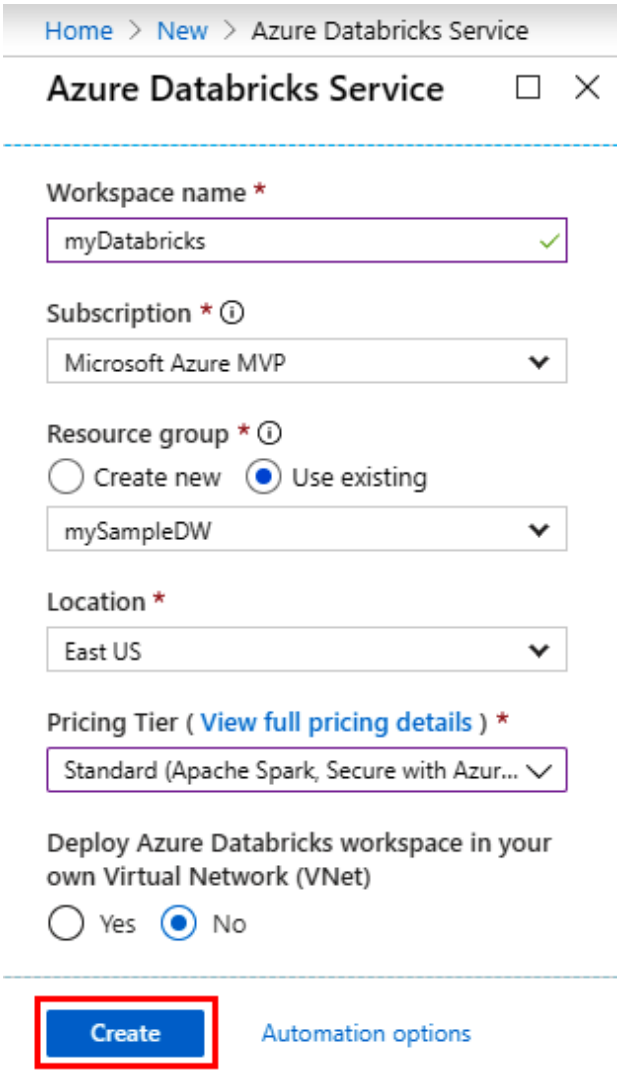

Abbildung 2.58: Hinzufügen von Informationen zum Databricks-Dienst

4. Wählen Sie Ihren Azure Databricks-Dienst nach der Bereitstellung aus, indem Sie seinen Namen anklicken:

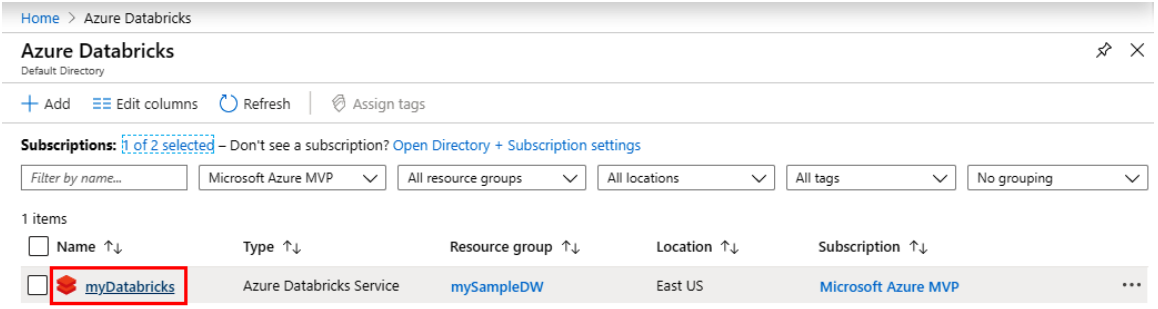

Abbildung 2.59: Auswählen des bereitgestellten Databricks-Diensts

5. Klicken Sie auf **Launch Workspace** (Arbeitsbereich starten), um das Azure Databricks-Portal in einer separaten Browser-Registerkarte zu starten:

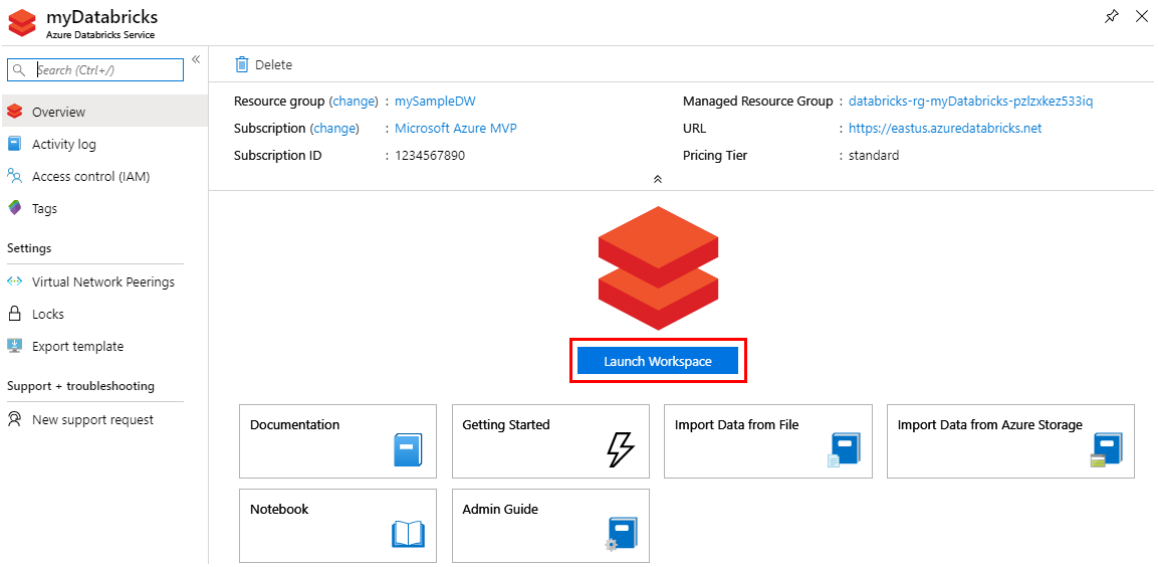

Abbildung 2.60: Starten des Databricks-Arbeitsbereichs

6. Erstellen Sie nun einen neuen Spark-Cluster. Klicken Sie dazu auf **New Cluster** (Neuer Cluster):

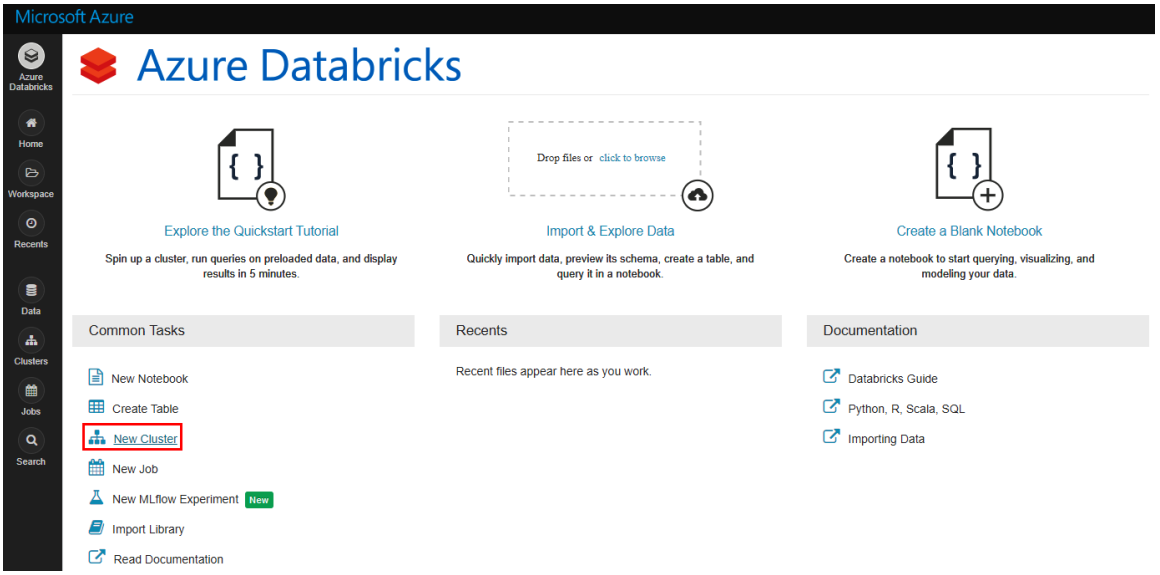

Abbildung 2.61: Erstellen eines neuen Spark-Clusters

7. Füllen Sie die Seite „New Cluster“ (Neuer Cluster) wie in der folgenden Abbildung dargestellt aus, und klicken Sie auf **Create Cluster** (Cluster erstellen).

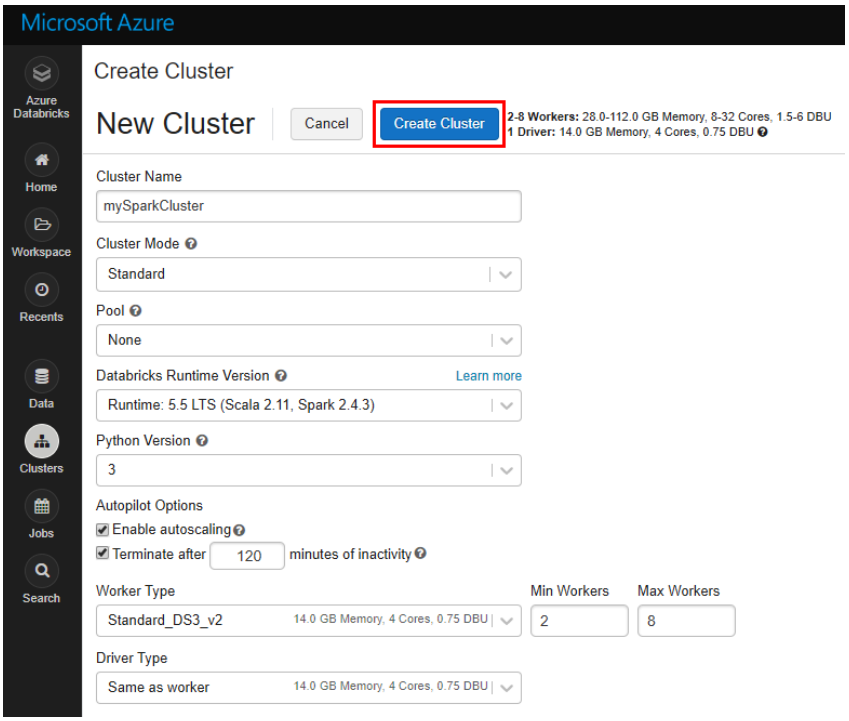

Abbildung 2.62: Hinzufügen von Informationen zum Spark-Cluster

8. Wenn Ihr Spark-Cluster bereitgestellt ist, wird ein etwa wie folgt aussehender Bildschirm angezeigt:

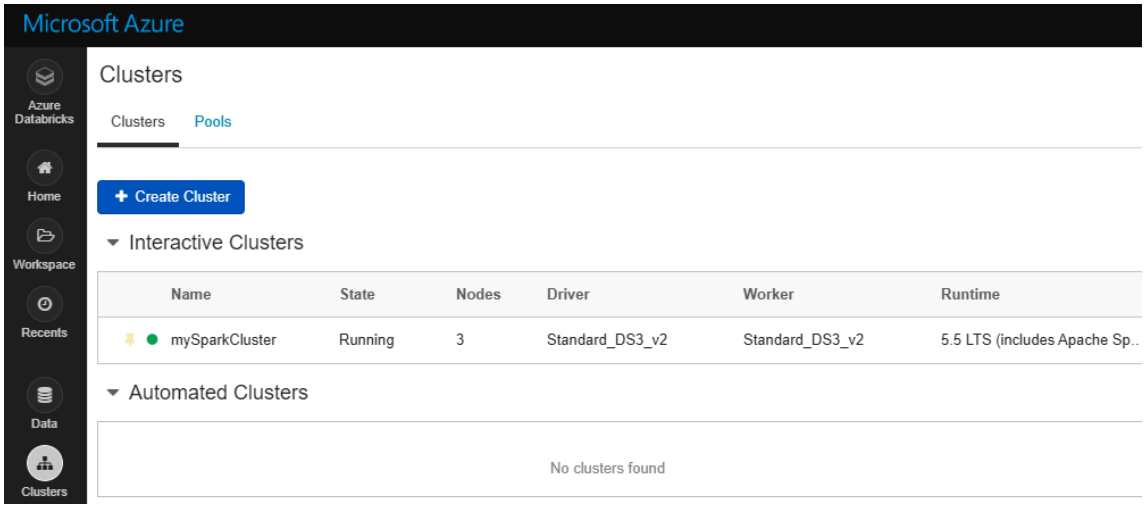

Abbildung 2.63: Erstellter Cluster im aktiven Zustand

Sie haben jetzt erstmals Azure Databricks und einen Spark-Cluster bereitgestellt.

Verwenden von Azure Databricks zum Vorbereiten und Transformieren von Daten

Um den Prozess des modernen Data Warehouse abzuschließen, müssen Sie diese letzte Übung absolvieren. Sie werden Daten in Azure Data Lake Storage Gen2 erfassen, die Daten mit Azure Databricks bereinigen und transformieren und schließlich die bereinigten und transformierten Daten in Azure Synapse Analytics laden.

1. Wählen Sie im Azure Databricks-Arbeitsbereich **Create** (Erstellen) und dann **Notebook** aus:

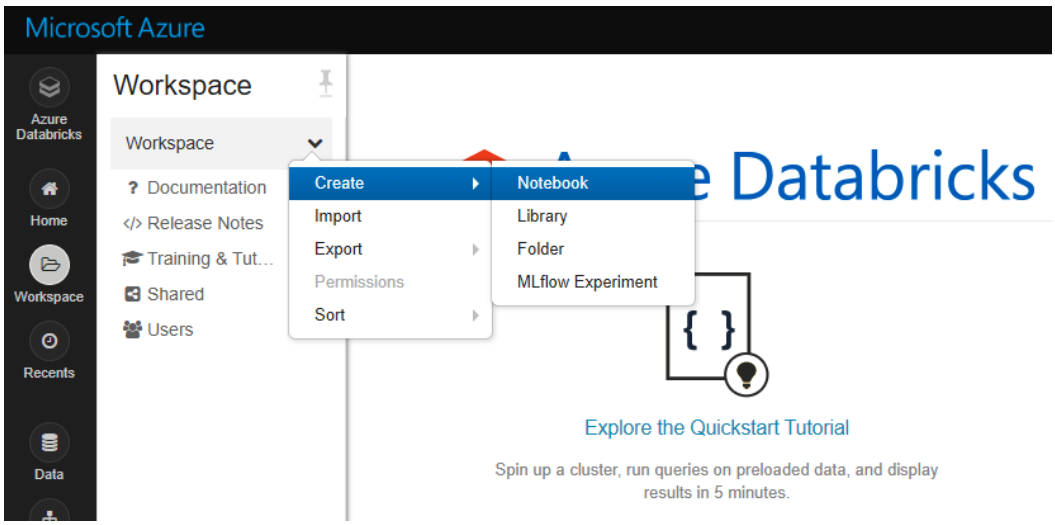

Abbildung 2.64: Erstellen eines Notebooks im Databricks-Arbeitsbereich

2. Erstellen Sie Ihr Notebook, wie in der folgenden Abbildung dargestellt:

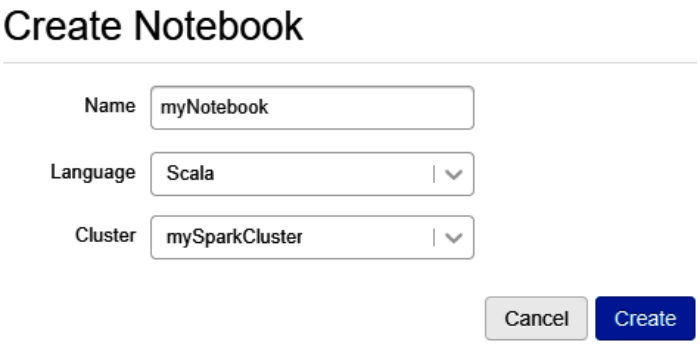

Abbildung 2.65: Hinzufügen von Details zum Erstellen des Notebooks

3. Verwenden Sie in der ersten Zelle des Notebooks das unten aufgeführte Scala-Skript, um Ihre Spark-Sitzung in Ihrer Azure Databricks-Umgebung zu konfigurieren. Ersetzen Sie die Werte wie in dem Skript angegeben Ihrer eigenen Umgebung entsprechend. Wenn Sie alle Werte in das Skript eingetragen haben, drücken Sie die *UMSCHALTTASTE + EINGABETASTE*, um das Skript im Notebook auszuführen.

```
// replace the following values based on your environment
val storageAccountName = "replace with your own Azure Storage Account
name"
val fileSystemName = "replace with your own File System name"
val tenantID = "replace with your own tenant id"
val appID = "replace with your own appID"
val password = "replace with your own password"
// configuration for blob storage
val blobStorageAcctName = "replace with your own Azure blob storage
account name"
val blobContainer = "replace with your own blob container name"
val blobAccessKey = "replace with your own access key"
// configuration for Azure SQL Data Warehouse
val dwDatabase = "replace with your own database name"
val dwServer = "replace with your own database server name"
val dwUser = "replace with your own database user name"
val dwPass = "replace with your own database password"
val newTable = "replace with your own new table name"
```

> **Hinweis**
>
> Geben Sie für das folgende Scala-Skript jeden einzelnen Codeblock in Ihre Azure Databricks Notebook-Zelle ein, und drücken Sie die *UMSCHALTTASTE + EINGABETASTE*, um jeden einzelnen Codeblock auszuführen. Prüfen Sie die Ergebnisse nach jeder Ausführung.

4. Konfigurieren Sie die Spark-Sitzung wie folgt:

```
spark.conf.set("fs.azure.account.auth.type", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type", "org.apache.hadoop.
fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id", "" + appID + "")
spark.conf.set("fs.azure.account.oauth2.client.secret", "" + password +
"")
spark.conf.set("fs.azure.account.oauth2.client.endpoint", "https://login.
microsoftonline.com/" + tenantID + "/oauth2/token")
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization",
"true")
dbutils.fs.ls("abfss://" + fileSystemName + "@" + storageAccountName +
".dfs.core.windows.net/")
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization",
"false")
```

5. Konfigurieren Sie das Azure Data Lake Storage Gen2-Konto wie folgt:

```
spark.conf.set("fs.azure.account.auth.type." + storageAccountName + ".dfs.
core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type." +
storageAccountName + ".dfs.core.windows.net", "org.apache.hadoop.
fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id." + storageAccountName +
".dfs.core.windows.net", "" + appID + "")
spark.conf.set("fs.azure.account.oauth2.client.secret." +
storageAccountName + ".dfs.core.windows.net", "" + password + "")
spark.conf.set("fs.azure.account.oauth2.client.endpoint." +
storageAccountName + ".dfs.core.windows.net", "https://login.
microsoftonline.com/" + tenantID + "/oauth2/token")
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization",
"true")
dbutils.fs.ls("abfss://" + fileSystemName + "@" + storageAccountName +
".dfs.core.windows.net/")
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization",
"false")
```

6. Rufen Sie die JSON-Beispieldatei wie folgt in **`/tmp`** ab:

```
%sh wget -P /tmp https://raw.githubusercontent.com/Azure/usql/master/
Examples/Samples/Data/json/radiowebsite/small_radio_json.json
```

7. Kopieren Sie die Beispieldatei aus **`/tmp`** in Azure Data Lake Storage Gen2:

```
dbutils.fs.cp("file:///tmp/sample.json", "abfss://" + fileSystemName + "@" +
storageAccountName + ".dfs.core.windows.net/")
```

8. Laden Sie die JSON-Beispieldatei in einen DataFrame:

```
val df = spark.read.json("abfss://" + fileSystemName + "@" +
storageAccountName + ".dfs.core.windows.net/sample.json")
```

9. Geben Sie die Inhalte des DataFrame aus, um sicherzustellen, dass alles ordnungsgemäß funktioniert:

```
df.show()
```

10. Bereinigen Sie die Daten, indem Sie nur zwei Spalten auswählen:

```
// called firstName and lastName from the DataFrame
val specificColumnsDf = df.select("firstname", "lastname")
```

11. Ausgabe der Auswahlergebnisse:

```
specificColumnsDf.show()
```

12. Transformieren Sie die Daten, indem Sie die Spalte **`lastName`** (Nachname) in **`surname`** (Zuname) umbenennen:

```
val transformedDF = specificColumnsDf.withColumnRenamed("lastName",
"surname")
```

13. Ausgabe der transformierten Ergebnisse:

```
transformedDF.show()
```

14. Laden Sie die Daten in Azure SQL Data Warehouse:

```
val blobStorage = blobStorageAcctName + ".blob.core.windows.net"
val tempDir = "wasbs://" + blobContainer + "@" + blobStorage + "/tempdir"
val acntInfo = "fs.azure.account.key." + blobStorage
sc.hadoopConfiguration.set(acntInfo, blobAccessKey)
```

15. Laden Sie den transformierten **`DataFrame`** als neue Tabelle mit dem Namen **`NewTable`** (Neue Tabelle) in Azure SQL Data Warehouse:

```
val dwJdbcPort = "1433"
val dwJdbcExtraOptions =
"encrypt=true;trustServerCertificate=true;hostNameInCertificate=*.database.
windows.net;loginTimeout=30;"
val sqlDwUrl = "jdbc:sqlserver://" + dwServer + ":" + dwJdbcPort +
";database=" + dwDatabase + ";user=" + dwUser+";password=" + dwPass +
";$dwJdbcExtraOptions"
val sqlDwUrlSmall = "jdbc:sqlserver://" + dwServer + ":" + dwJdbcPort +
";database=" + dwDatabase + ";user=" + dwUser+";password=" + dwPass
spark.conf.set("spark.sql.parquet.writeLegacyFormat", "true")
transformedDF.write.format("com.databricks.spark.sqldw").option("url",
sqlDwUrlSmall).option("dbtable", newTable).option("forward_spark_azure_
storage_credentials","True").option("tempdir", tempDir).mode("overwrite").
save()
```

Wenn Sie die Ausführung der einzelnen oben genannten Codeblöcke abschließen, haben Sie eine JSON-Beispieldatei in Azure Data Lake Storage Gen2 erfasst, die Daten durch ausschließliche Auswahl der beiden gewünschten Spalten bereinigt, einen Spaltennamen von **`lastName`** (Nachname) in **`surname`** (Zuname) transformiert und den transformierten **`DataFrame`** in Azure Synapse Analytics geladen. Damit ist der gesamte Prozess für das moderne Data Warehouse abgeschlossen.

Bereinigen von Azure Synapse Analytics

Wenn Sie Azure Synapse Analytics (früher SQL DW) nicht mehr benötigen, können Sie Geld sparen, indem Sie die betreffende Ressourcengruppe dauerhaft löschen. Bei diesem Vorgehen werden Azure Synapse Analytics (früher SQL DW) und alle zugehörigen Ressourcen (z. B. Azure Data Factory, Azure Data Lake Storage Gen2, Azure Databricks usw.), die Sie innerhalb derselben Ressourcengruppe bereitgestellt haben, dauerhaft gelöscht. Navigieren Sie hierfür zum Übersichtsbereich, und klicken Sie auf **Delete resource group** (Ressourcengruppe löschen), wie in der folgenden Abbildung dargestellt:

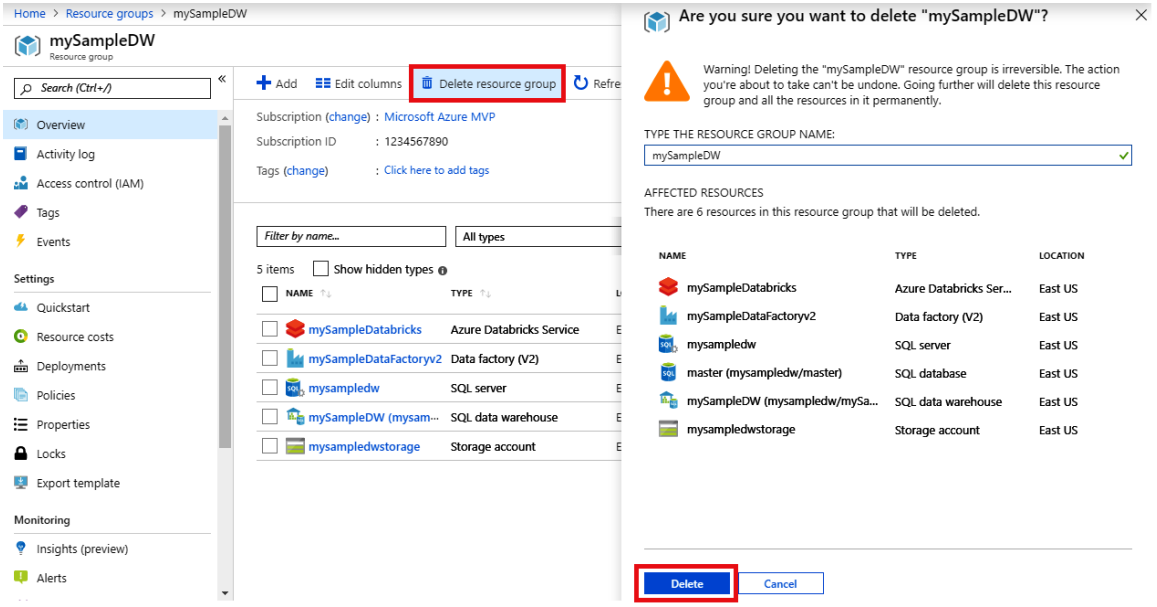

Abbildung 2.66: Bereinigen des Data Warehouse durch Löschen der Ressourcengruppe

Zusammenfassung

In diesem Kapitel haben Sie sich eingehend mit Azure Synapse Analytics, Azure Data Lake Storage Gen2, Azure Data Factory und Azure Databricks befasst. Sie haben die Features und Vorteile dieser Schlüsseltechnologien kennengelernt und erfahren, wie Sie Azure Synapse Analytics (früher SQL DW) bereitstellen können. Im nächsten Kapitel geht es um die analytische Seite des modernen Data Warehouse. Sie erfahren, wie Sie die Daten mithilfe von Power BI verarbeiten und visualisieren und Machine Learning implementieren können. Später in diesem Buch sehen Sie anhand von realen Anwendungsfällen, wie alle diese Technologien miteinander integriert werden können, um vollständige End-to-End-Data-Warehouse-Lösungen bereitzustellen, mit deren Hilfe die Entscheidungsträger im Unternehmen aussagekräftige Insights aus Echtzeitdaten gewinnen können.

3

Verarbeitung und Visualisierung von Daten

Im letzten Kapitel haben wir Azure Data Factory, Azure Data Lake Storage und Azure Synapse Analytics (früher Azure SQL Data Warehouse) für die Erfassung und Speicherung von Daten bereitgestellt. Außerdem haben wir unstrukturierte Daten mithilfe von Azure Databricks in ein strukturiertes Format transformiert.

In diesem Kapitel werden wir die strukturierten Daten analysieren, um aussagekräftige Insights zu gewinnen. Das Kapitel besteht aus zwei Hauptteilen:

- Datenmodellierung und -bereitstellung mit Azure Analysis Services
- Datenvisualisierung mithilfe von Power BI

Azure Analysis Services

Unternehmen generieren ständig riesige Datenmengen in verschiedenen Formaten aus mehreren Quellen. Aufgrund ihrer komplexen Infrastruktur, die einen zeitgerechten Zugriff auf die relevanten Daten nur beschränkt möglich macht, haben die Unternehmen dabei jedoch oft Schwierigkeiten sicherzustellen, dass alle Stakeholder auf die Daten zugreifen können. Echtzeitzugriff auf die Daten ist entscheidend für eine hervorragende datengestützte Kultur. Häufig ist es den Analysten und Datenwissenschaftlern aus folgenden Gründen nicht möglich, die Datensätze in den Unternehmensdatenbanken direkt zu erkunden und zu analysieren:

- Es könnten vertrauliche Informationen durchsickern.
- Wenn die Analysen direkt auf den Produktionsservern ausgeführt werden, kann sich dies auf die Leistung auswirken und mehr Ausfallzeiten zur Folge haben.

In der Phase der semantischen Datenmodellierung geht es darum, ein semantisches Datenmodell zu erstellen, um Analysten einen nahtlosen Zugriff auf die Daten zu ermöglichen. Ein semantisches Datenmodell liegt in strukturierter Tabellenform vor und ist für die meisten Anwender leicht visualisierbar und verständlich. In der Praxis fragen Analysten die Transaktionsdatenbanken nicht direkt ab, da diese Datenbanken für die Endkunden und Anwendungen bestimmt sind. Stattdessen entwickeln sie ein Datenmodell und speichern es im Data Warehouse. Oft handelt es sich bei den Daten des semantischen Modells nur um eine zwischengespeicherte Version der Produktionsdaten, die irgendwann gelöscht oder aktualisiert wird. Azure Analysis Services kann diese Datenlücke schließen.

Azure Analysis Services (AAS) ist ein cloudgehosteter Dienst und ein vollständig verwaltetes Platform-as-a-Service-Tool (PaaS), mit dessen Hilfe Dateningenieure Datenbanktabellen modellieren und die semantischen Datenmodelle Anwendern bereitstellen können. Ein Beispiel hierfür ist das Generieren von Verkaufsberichten, wenn mehrere Datenquellen vorhanden sind und die Stakeholder lediglich wissen möchten, ob das Unternehmen Gewinne oder Verluste macht. Die Datenmodelle bieten Anwendern (insbesondere Analysten) eine komfortable Möglichkeit, riesige Mengen von zwischengespeicherten Daten für On-Demand-Datenanalysen zu erkunden und zu durchsuchen. Somit müssen die Analysten nicht warten, bis die Dateningenieure manuell eine Momentaufnahme der Daten (oft im CSV- oder Excel-Format) erstellen und ihnen per E-Mail zusenden.

SQL Server Analysis Services

Da AAS Features von SSAS (SQL Server Analysis Services) übernommen hat, ist es zum besseren Verständnis der Verbindungspunkte zwischen einer Datenquelle und Power BI sinnvoll, zunächst die Funktionen von SSAS zu erläutern.

Im folgenden Diagramm ist eine eine hybride Architektur dargestellt, in der SSAS zur Datenanalyse und zur Verbindung mit dem Power BI-Dienst über VPN Gateway verwendet wird:

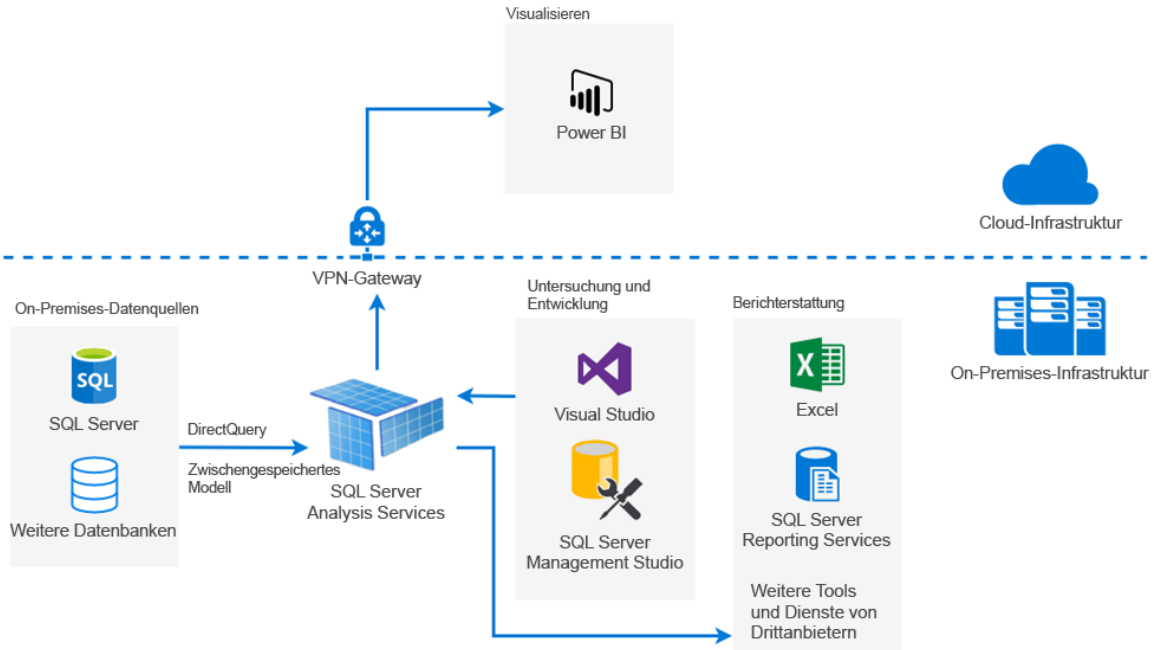

Abbildung 3.1: SQL Server Analysis Services

In einem On-Premises-Szenario wird Ihre Datenbank (oder Ihr Data Warehouse) in Ihrem eigenen Rechenzentrum gehostet. Unternehmen verwenden SQL Server Analysis Services heute in der Regel für ihre On-Premises-BI-Lösung. Dateningenieure haben damit die Möglichkeit, eine Verbindung mit ihren On-Premises-Datenquellen herzustellen und aus komplexen Datasets eine einzelne strukturierte Datenbank zu schaffen, die allgemein als SSOT (Single Source of Truth, einzige Quelle der Wahrheit) bezeichnet wird.

Dateningenieure verwenden Tools wie Visual Studio oder SQL Server Management Studio (SSMS), um semantische Datenmodelle zu erstellen, zu erkunden und zu entwickeln, die sie dann als Bericht oder tabellarische Modelle für Power BI (oder andere Business-Intelligence-Tools) senden können.

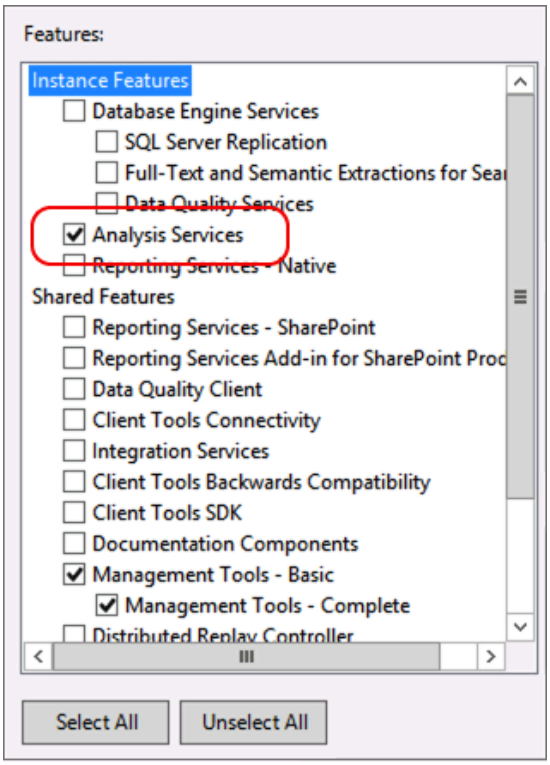

Abbildung 3.2: Im Fenster „Features“ aufgeführte Analysis Services

Der größte Nachteil der semantischen Datenmodellierung bei SQL Server Analysis Services ist die starke Abhängigkeit von der SQL Server-Instanz und -Lizenzierung. Da sich Ihr Rechenzentrum On-Premises befindet, stellt die Skalierung nach Bedarf eine Herausforderung dar. Hier zeigen sich die Vorteile von Azure Analysis Services gegenüber SSAS.

AAS ist ein cloudgehosteter Dienst, der leicht skalierbar und unabhängig von einer SQL Server-Instanz ist. Im Folgenden finden Sie einige der wichtigsten Features und Vorteile der Verwendung von AAS.

Features und Vorteile

Azure Analysis Services beruht auf dem bewährten Analysemodul von SQL Server 2016 Analysis Services. Analysedienste werden nativ auf Cloudressourcen wie Azure Synapse Analytics ausgeführt.

AAS bietet folgende Features:

- Der Anwender kann Daten aus verschiedenen komplexen Quellen kombinieren und eine tabellarische Darstellung der Daten erstellen, die allgemein leicht verständlich ist.
- Die Plattform bietet bedarfsgesteuerte Leistung entsprechend der Datengröße und dem Umfang der Operationen.
- Darüber hinaus bietet AAS eine zusätzliche Sicherheitsebene, die dafür sorgt, dass nur berechtigte Personen über Azure Active Directory Zugriff auf die entsprechenden Daten erhalten (rollenbasierte Zugriffssteuerung).

Azure Analysis Services kann leicht bereitgestellt, skaliert und verwaltet werden, da es sich um eine PaaS-Lösung handelt. Innerhalb von Sekunden können Sie einen Azure Analysis Service bereitstellen. Sie können Ihre Betriebsebenen flexibel hoch- oder herunterskalieren, je nachdem, wie häufig Sie den Dienst benötigen. Schließlich entfällt der mit der Verwaltung der zugrunde liegenden Infrastruktur (z. B. Netzwerk, Festplattenspeicher, Arbeitsspeicher und Festplatten) verbundene Aufwand, da der Dienst vollständig in der Cloud verwaltet wird.

Abbildung 3.3 zeigt die geänderte Architektur nach dem Übergang von SSAS zu AAS:

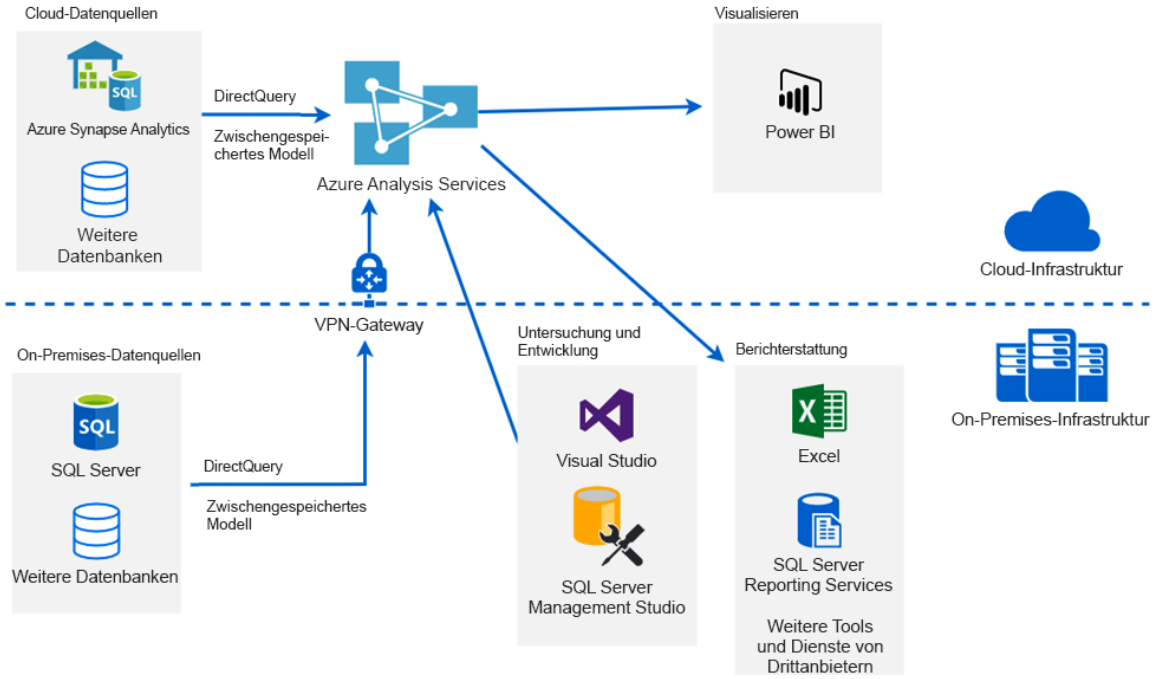

Abbildung 3.3: Azure Analyse Services

Im vorstehenden Diagramm fungiert AAS als Brücke zwischen Azure Synapse Analytics (zum Speichern von Daten verwendet) und Power BI, das als Visualisierungstool dient. Das bedeutet nicht, dass AAS nur auf die Verwendung von Azure Synapse Analytics als Datenquelle beschränkt ist. AAS kann mit mehreren On-Premises- und Cloud-Datenbanklösungen verbunden werden.

Azure Analysis Services bietet dieselben Funktionen und Möglichkeiten wie SQL Server Analysis Services, es handelt sich hierbei jedoch um einen vollständig verwalteten Clouddienst. AAS unterstützt komplexe Modellszenarien wie z. B. bidirektionale Kreuzfilter. Ein Beispiel hierfür ist die Gewinnung aussagekräftiger Insights von Kunden und Produkten, wenn eine Viele-zu-viele-Datenbeziehung besteht, also ein Kunde mehrere Produkte kaufen kann und verschiedene Kunden das gleiche Produkt kaufen können.

AAS ist hochgradig skalierbar. Der Dienst kann linear skaliert werden, indem die Anzahl von Kernen erhöht wird. Außerdem kann der auf Intel TBB (Threading Building Blocks) basierende skalierbare Allocator verwendet werden, der für jeden Kern separate Speicherpools bereitstellt.

Power BI

Power BI ist eine Suite von Tools, mit denen Anwender Daten visualisieren und Insights in Teams und Organisationen teilen oder in ihre Websites oder Anwendungen einbinden können. Die Lösung unterstützt verschiedene Datenquellen (sowohl strukturierte als auch unstrukturierte Datentypen). Analysten und Anwender können damit leichter nach Bedarf Live-Dashboards und Berichte über die Daten des Unternehmens erstellen. Ein Beispiel hierfür ist die Visualisierung der Unternehmensverkäufe in den letzten Monaten und die Ermittlung der Stadt, in der die meisten Artikel verkauft wurden.

Anders als eine Tabellenkalkulationssoftware wie Microsoft Excel ist Power BI als gehostete Benutzeroberfläche (oft ein Live-Dashboard) konzipiert. Dabei entfällt für die Anwender die Notwendigkeit, eine Datei häufig auf ihrem lokalen Computer zu speichern und zu öffnen. Mit Power BI können Sie das Potenzial der Cloud nutzen, um komplexe Daten einzubeziehen und in Form von aussagekräftigen Grafiken oder Diagrammen darzustellen. Anstelle Ihres eigenen Systems kann der Server alle Berechnungen ausführen. Wenn nun die Datengröße von 500 Megabyte auf mehrere Gigabyte anwachsen würde, hätten die meisten universellen Systeme (wie beispielsweise PCs mit begrenztem Arbeitsspeicher) Schwierigkeiten, die Excel-Datei zu laden. Mit Power BI ist dies jedoch so, wie eine Webseite zu öffnen, da es sich um einen gehosteten Dienst handelt.

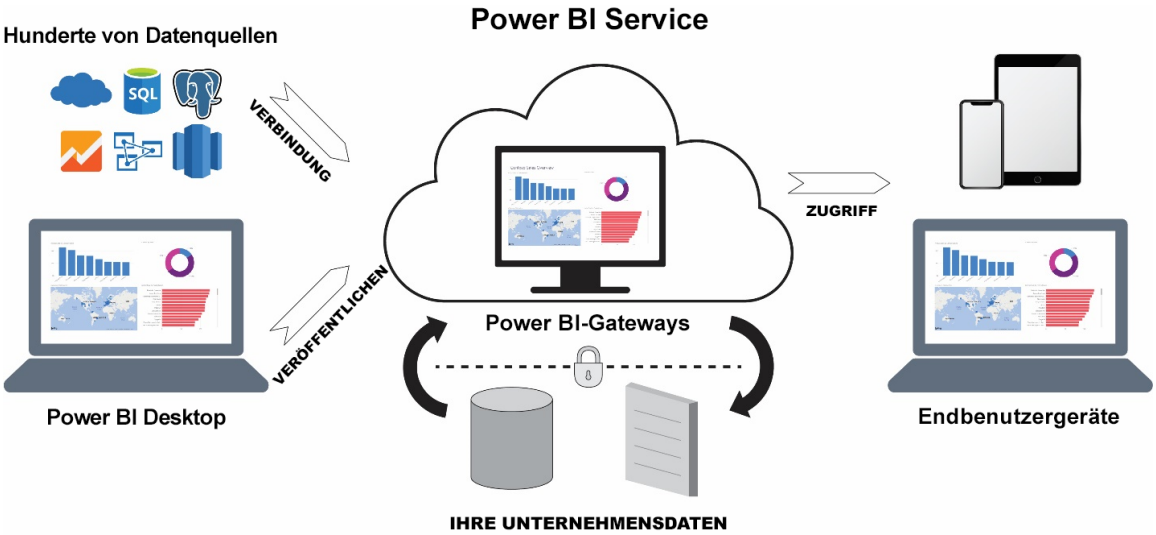

Abbildung 3.4: Power BI-Architektur

Power BI besteht aus verschiedenen Komponenten, die verschiedene Funktionen ausführen können:

Power BI Desktop

Eine desktopbasierte Windows-Anwendung, die häufig als „Erstellungstool“ bezeichnet wird und in der Sie hauptsächlich Berichte für den Dienst entwerfen und veröffentlichen.

Power BI

Die verwaltete Plattform für die Bereitstellung Ihrer Berichte. Es handelt sich hier um eine Software-as-a-Service-(SaaS-)Anwendung, die von „Power BI for Office 365" zu „Power BI“ weiterentwickelt wurde.

Mobile Power BI-Apps

Native mobile Anwendungen, die von einem in Power BI gehosteten Arbeitsbereich aus auf Berichte zugreifen können. Verfügbar im Apple iOS App Store und im Google Play Store.

Power BI Gateway

Ein Mechanismus zur Synchronisierung externer Daten in Power BI. In Unternehmensszenarien mit On-Premises-Speicher fungiert Power BI Gateway als Mechanismus zum Abfragen der Datenquelle, ohne Datenbanken in die Cloud übertragen zu müssen. Die Daten, die in Power BI-Berichten gehostet werden, befinden sich jedoch in Azure Cloud.

Power BI-Berichtsserver

In einem On-Premises-Szenario bietet Ihnen der Power BI-Berichtsserver die Möglichkeit, Power BI-Berichte in Ihrem eigenen Rechenzentrum zu hosten. Diese Berichte können nach wie vor von verschiedenen Mitgliedern gemeinsam genutzt werden, wenn diese über den entsprechenden Netzwerkzugriff verfügen.

Power BI Embedded

Ermöglicht Ihnen ein Whitelabeling von Power BI in Ihren eigenen Anwendungen. Strategisch sind diese häufig in vorhandene Dashboards und Back-Office-Systeme integriert, bei denen dieselbe Anwendergruppe nur auf die Berichte zugreifen kann.

Features und Vorteile

Grundsätzlich bietet Power BI die folgenden Vorteile:

- Personalisierte Dashboards, die Analysten ein Branding des Erscheinungsbilds von Grafiken, Diagrammen und Tabellen ermöglichen
- Zusammenarbeit verschiedener Anwender
- Governance und Sicherheit, sodass nur berechtigte Anwender auf die Dashboards zugreifen können
- Keine Einschränkungen in Bezug auf Speicher oder Geschwindigkeit, da es sich um einen cloudgehosteten Dienst handelt. Nicht anders, als eine Webseite zu laden
- Kein spezieller technischer Support erforderlich, da die Interaktion mit den Berichten einfach ist
- Unterstützung für erweiterte Datendienste, wie z. B. die Fragefunktion, Integration in R, Segmentierung und Kohortenanalyse

Power BI ist ein intuitives Tool, bei dem für einen schnellen Datenzugriff und die Visualisierung von Daten oft nur ein Mausklick oder Drag-and-Drop erforderlich ist. Das Erstellungstool (Power BI Desktop) verfügt über viele integrierte Funktionen zur Ableitung von Analytics. Es ist in der Lage, auf der Grundlage der Felder Ihrer Wahl ein Visualisierungsmodell vorzuschlagen.

Power BI-Dashboards und -Berichte sind hochgradig anpassbar. Sie haben die Möglichkeit, die Oberfläche Ihrer Marke entsprechend zu personalisieren. Sie können Designs auswählen, eigene Diagramme verwenden, Beschriftungen erstellen, Zeichnungen und Bilder einfügen und vieles mehr.

Im Vergleich zum Senden einer E-Mail mit einer angehängten PowerPoint-Datei ermöglicht Power BI eine offene Zusammenarbeit zwischen Analysten und anderen Mitarbeitern des Unternehmens durch gemeinsame Nutzung eines zentralen Dashboards. Für den Zugriff auf die Berichte können Sie gängige Webbrowser oder mobile Anwendungen verwenden, die Sie im Apple App Store und im Google Play Store herunterladen können. Die Mitarbeiter können Kommentare und Anmerkungen zu den Berichten senden. Durch die Verwendung von Warnungen und Benachrichtigungen entsteht so eine schnellere Feedbackschleife.

Power BI ist in vielerlei Hinsicht sicher. Erstens ist beim Erstellen eines Berichts sichergestellt, dass Sie nur auf Datenquellen zugreifen können, für die Sie zugriffsberechtigt sind. Gestützt wird dies durch **Sicherheit auf Zeilenebene** (**Row Level Security, RLS**). Analysten können beispielsweise nur auf lokale Daten ihrer Region zugreifen. Es ist sichergestellt, dass sie keinen Zugriff auf die Daten anderer Städte oder Länder haben. Wenn Sie bereit sind, den Bericht freizugeben, können Sie ihn schnell in Ihrem persönlichen Arbeitsbereich speichern. Sie können auswählen, für wen Sie den Bericht in Ihrem Unternehmen freigeben möchten, oder Anwender von externen Mandanten einladen.

Wenn Sie klein anfangen möchten, während Sie Power BI kennenlernen, können Sie zunächst damit beginnen, nur Excel-Dateien als Datenquelle zu verwenden. Es gibt Fälle, in denen Analysten eine CSV-Datei von Dateningenieuren erhalten, da der Datensatz nicht allzu groß ist.

In diesem Buch werden wir Power BI verwenden, um Berichte aus semantischen Datenmodellen von AAS zu generieren. Zwar unterstützt Power BI mehrere Datenquellen, darunter auch Azure Synapse Analytics und On-Premises-Datenbanken, es empfiehlt sich jedoch, stattdessen einen Analysis Service zu nutzen. Manchmal können Sie auch Daten aus einem zwischengespeicherten Ergebnis von Azure Databricks abfragen.

Es gibt verschiedene Vorgehensweisen hierfür, für diesen Ansatz (AAS + Power BI, wie *in Abbildung* 3.3 dargestellt) sprechen jedoch in erster Linie die damit verbundene Elastizität und Aufgabentrennung. Wenn Sie AAS verwenden, erhalten Sie eine Momentaufnahme des Zustands Ihrer Datenquellen, von IoT-Streams bis hin zu Datenbanken (mithilfe von ADF, Data Lake, SQL DW). Die Daten aus Ihrer Produktions-Transaktionsdatenbank werden zwischengespeichert und hätten selbst dann keinen signifikanten Einfluss auf die Leistung, wenn Ihre Datenbanken Milliarden von Zeilen umfassen würden.

Quick-Start-Leitfaden (Datenmodellierung und -visualisierung)

Nachdem wir uns mit den Begriffen AAS und Power BI befasst haben, fahren wir nun mit der Datenmodellierung und -visualisierung mithilfe dieser Tools/Dienste fort.

Voraussetzungen

Für diese Aktivität benötigen Sie Folgendes:

- Aktives Azure-Abonnement
- Power BI Desktop
- Erstellungstool (optional)
- SQL Server Data Tools (SSDT) in Visual Studio
- SQL Server Management Studio
- Azure Synapse Analytics als AAS-Datenquelle (optional)

Bereitstellen des Azure Analysis Service

Suchen Sie im Azure-Portal nach **Analysis Services**, und klicken Sie auf die Schaltfläche „Create“ (Erstellen).

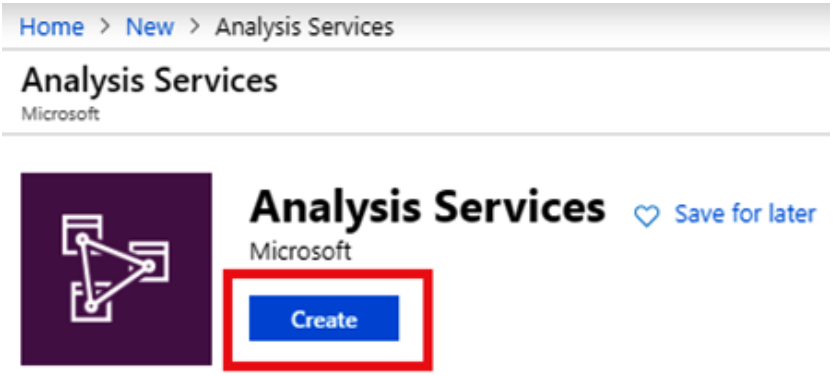

Abbildung 3.5 Erstellen eines Azure Analysis Service

Füllen Sie in Analysis Services die erforderlichen Felder aus, und wählen Sie dann „Create“ (Erstellen) aus:

- **Server name** (Servername): ein eindeutiger Name für die Ressource.
- **Subscription** (Abonnement): ein Abonnement, in dem Sie Ihre Ressource bereitstellen möchten.
- **Resource group** (Ressourcengruppe): eine logische Gruppierung von Ressourcen für die zentrale Verwaltung von Zugriffskontrollen, Sperren und Tags.
- **Location** (Standort): die Rechenzentrumsregion, die den Server hostet.
- **Pricing tier** (Tarif): für Entwicklungszwecke können Sie D1 wählen.
- **Administrator**: ein berechtigter Anwender des Servers. Sie können später weitere hinzufügen.
- **Backup Storage settings** (Backup Storage-Einstellungen): optional. Speicher für die Modelldatenbanksicherung.
- **Storage key expiration** (Speicherschlüsselablauf): optional. Geben Sie den Zeitraum an, in dem der Speicherschlüssel verwendet werden kann.

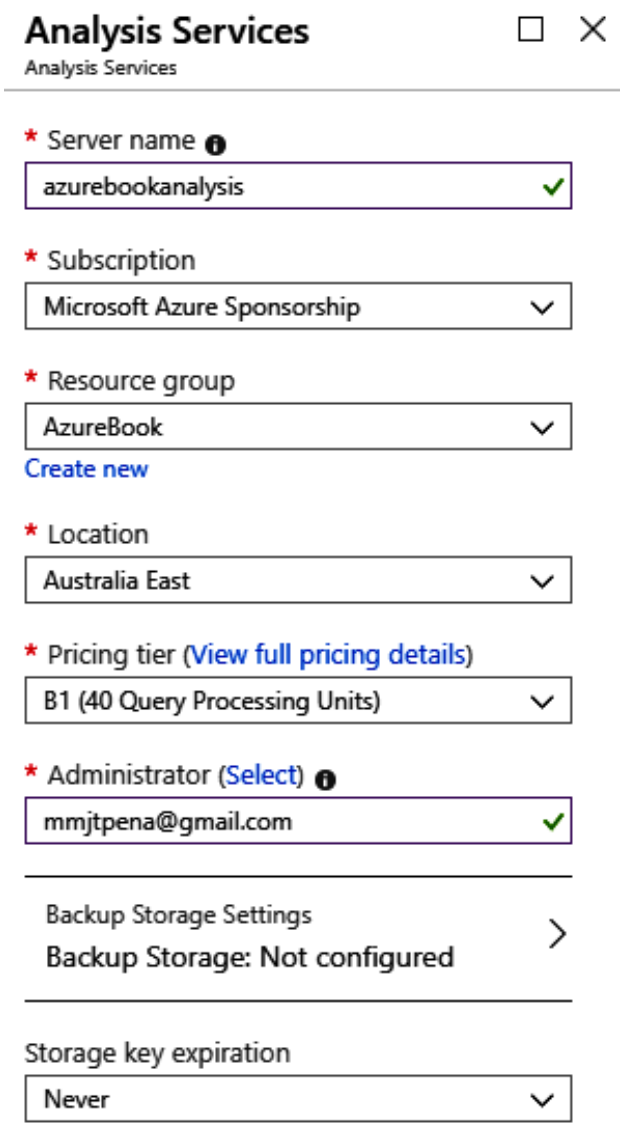

Abbildung 3.6: Hinzufügen von Details zum Analysis Service

Die Bereitstellung des Servers dauert in der Regel nicht länger als eine Minute.

Ermöglichen des Clientzugriffs

Beim Erstellen eines semantischen Datenmodells müssen Sie ein Desktop-Tool wie Power BI Desktop, SQL Server Data Tools (SSDT) für Visual Studio oder SQL Server Management Studio (SSMS) verwenden. Da der Server in der Cloud gehostet wird, müssen Sie Ihre IP-Adresse auf eine Whitelist setzen, damit Ihre Client-App (SSDT oder SSMS) auf den Azure Analysis Services-Server zugreifen kann.

1. Wechseln Sie im bereitgestellten Analysis Services zum Abschnitt **Firewall**. (Siehe *Abbildung* 3.7)
2. Stellen Sie sicher, dass die Firewall aktiviert ist, damit Ihr Server nicht öffentlich zugänglich ist.
3. Klicken Sie auf **Allow access from Power BI** (Zugriff über Power BI zulassen), damit der Power BI-Dienst einen **`DirectQuery`**-Zugriff ausführen kann.
4. Klicken Sie auf **Add Client IP** (Client-IP hinzufügen). Die IP-Adresse Ihres vorhandenen Clients wird dann zu den IPs auf der Whitelist hinzugefügt. Optional können Sie auch einen IP-Adressbereich angeben, dem Sie Zugriff erteilen möchten.
5. Klicken Sie auf **Save** (Speichern).

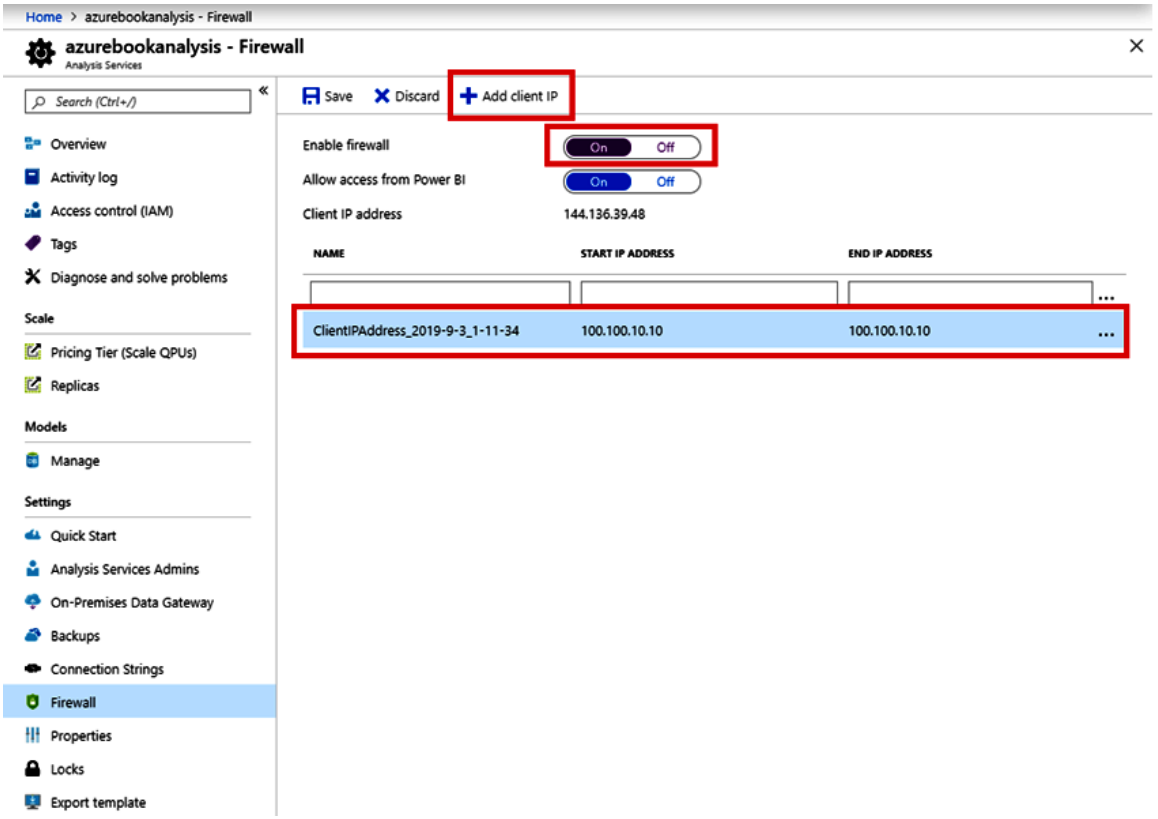

Abbildung 3.7: Aktivieren des Firewall-Zugriffs für Power BI

Durch die Ausführung dieser Aktivität hat Ihr Clientcomputer (z. B. ein Laptop) jetzt Zugriff auf die AAS-Modelle. Dies ist ein sehr leistungsstarkes Feature von Azure, das nur bestimmten IP-Adressen Zugriff auf den Dienst ermöglicht.

Erstellen eines Modells

In diesem Abschnitt werden wir ein Modell erstellen und dazu folgende Schritte ausführen:

1. Wechseln Sie in derselben Analysis Services-Ressource zur Registerkarte **Manage** (Verwalten) im Abschnitt **Models** (Modelle).

Abbildung 3.8: Erstellen eines neuen Modells

2. Klicken Sie auf die Schaltfläche **New model** (Neues Modell). Daraufhin wird eine Ansicht geöffnet. Wählen Sie aus Gründen der Einfachheit in dieser Übung **Sample data** (Beispieldaten) aus, und klicken Sie auf **Add** (Hinzufügen).

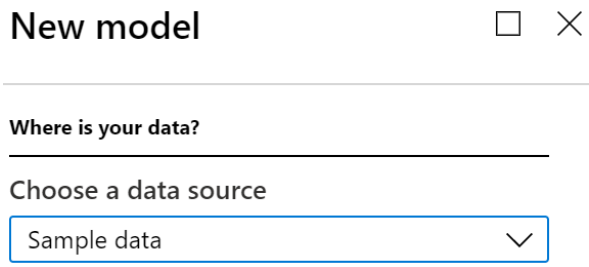

Abbildung 3.9: Auswählen der Beispieldaten als AAS-Datenquelle

> **Hinweis**
>
> Hiermit wird ein neues bereitgestelltes semantisches Modell für AAS, basierend auf der bekannten Datenbank **`AdventureWorks`** mit Daten des Online-Fahrradladens, erstellt.

In der Praxis werden Sie ein Erstellungstool wie SQL Server Development Tools in Visual Studio verwenden und ein tabellarisches Modell auf Azure Synapse Analytics entwickeln. Hierfür erstellen Sie ein Analysis Services-Projekt für tabellarische Modelle über Visual Studio.

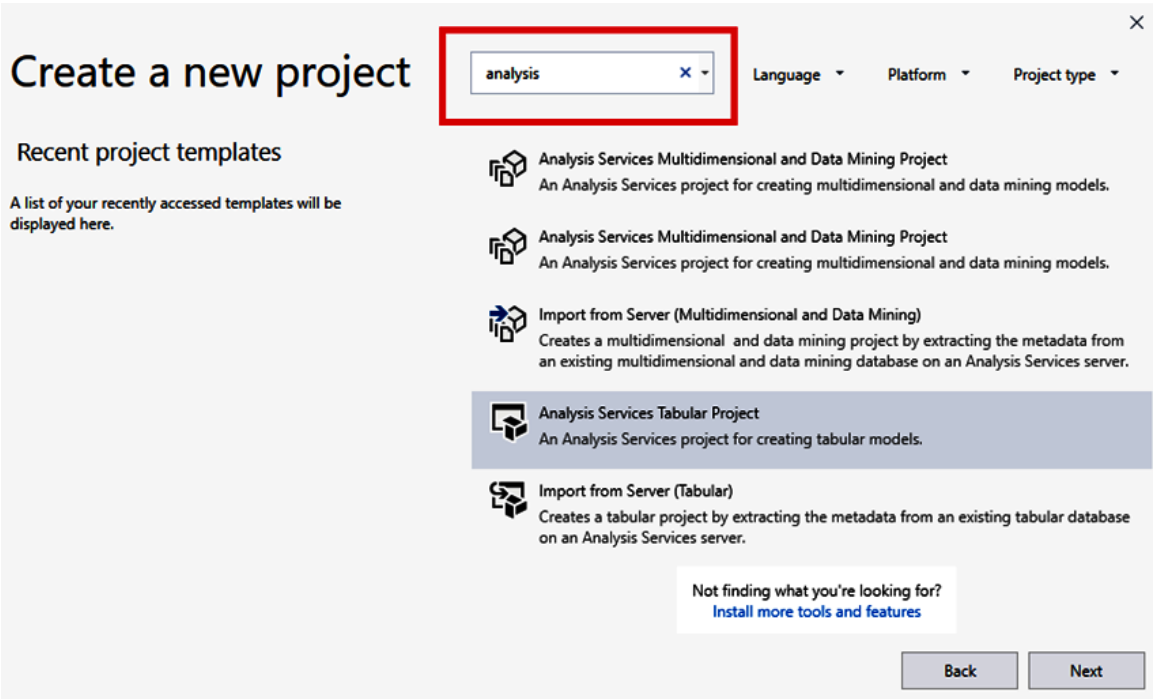

Abbildung 3.10: SQL Server-Entwicklungstools in Visual Studio

Wir haben uns in diesem Buch nicht eingehend mit dem Erstellen eines Datenmodells befasst. Für alle, die jedoch mehr erfahren möchten, hat Microsoft eine praktische Ressource zum Erstellen eines Analysis Services-Datenmodells mit Kompatibilitätsgrad 1400 verfasst. 1400 bedeutet, dass das Modell sowohl für Azure Analysis Services als auch für SQL Server Analysis Services verwendet werden kann.

Weitere Informationen hierzu finden Sie hier.

Ressourcen herunterladen:

- Visual Studio
- SSDT
- SSMS

Öffnen des erstellten Modells mit Power BI

In diesem Abschnitt werden wir das erstellte Modell in Power BI Desktop öffnen.

Achten Sie nach dem Erstellen eines Modells darauf, dass Sie auf der Registerkarte „Manage model“ (Modell verwalten) bleiben. Klicken Sie auf die Auslassungszeichen rechts in dem Modell und anschließend auf **Open in Power BI Desktop** (In Power BI Desktop öffnen).

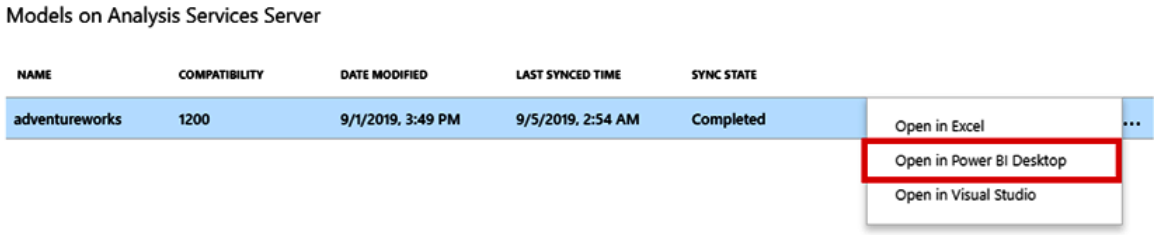

Abbildung 3.11: Öffnen von Power BI Desktop

AAS wird dann eine **`PBIX`**-Datei herunterladen, eine Power BI Desktop-Datei, in der eine Verbindung zwischen dem **`AdventureWorks`**-Datenmodell und Power BI hergestellt wurde.

In den folgenden Teilen der Aktivität müssen Sie Power BI Desktop verwenden, um ein Live-Dashboard zu erstellen.

Zum Zeitpunkt des Schreibens dieses Dokuments gelten für Power BI Desktop folgende Einschränkungen:

- Funktioniert nur unter Windows 7/Windows Server 2008 R2 oder höher
- Sie müssen eine geschäftliche E-Mail-Adresse verwenden, um Berichte und Dashboards mit Power BI Desktop zu erstellen. Einer privaten E-Mail-Adresse (wie z. B. **`@outlook.com`** oder **`@gmail. com`**) kann jedoch ein Gastzugriff zum Anzeigen der Berichte und Dashboards über den Power BI-Dienst erteilt werden.

> **Hinweis**
>
> Wenn Sie Power BI Desktop noch nicht heruntergeladen haben, können Sie es kostenfrei über die Microsoft Store-App für Windows 10 herunterladen oder aber den eigenständigen Installer herunterladen. Wenn Sie in Power BI Desktop einsteigen möchten, verwenden Sie für die Registrierung eine geschäftliche E-Mail-Adresse. Rufen Sie den folgenden Link auf, um herauszufinden, ob Sie zum Erstellen eines Kontos berechtigt sind.

In einem Unternehmensszenario werden Sie beim Zugriff auf das Azure-Portal dasselbe Arbeitskonto verwenden. Einige Zugriffssteuerungsebenen werden übernommen, da sie sich auf demselben Azure Active Directory-Mandanten befinden. Wenn Sie für Ihr Azure-Portal ein persönliches Konto und für Ihr Power BI-Konto eine separate geschäftliche E-Mail-Adresse verwendet haben, müssen Sie das Power BI zugeordnete Konto dem Azure Analysis Server als Administrator hinzufügen.

Führen Sie die folgenden Schritte aus, um das Modell in Power BI zu öffnen:

1. Kehren Sie zum Azure-Portal zurück. Wechseln Sie in derselben Azure Analysis Services-Ressource zu **Analysis Services Admins** (Analysis Services-Administratoren).

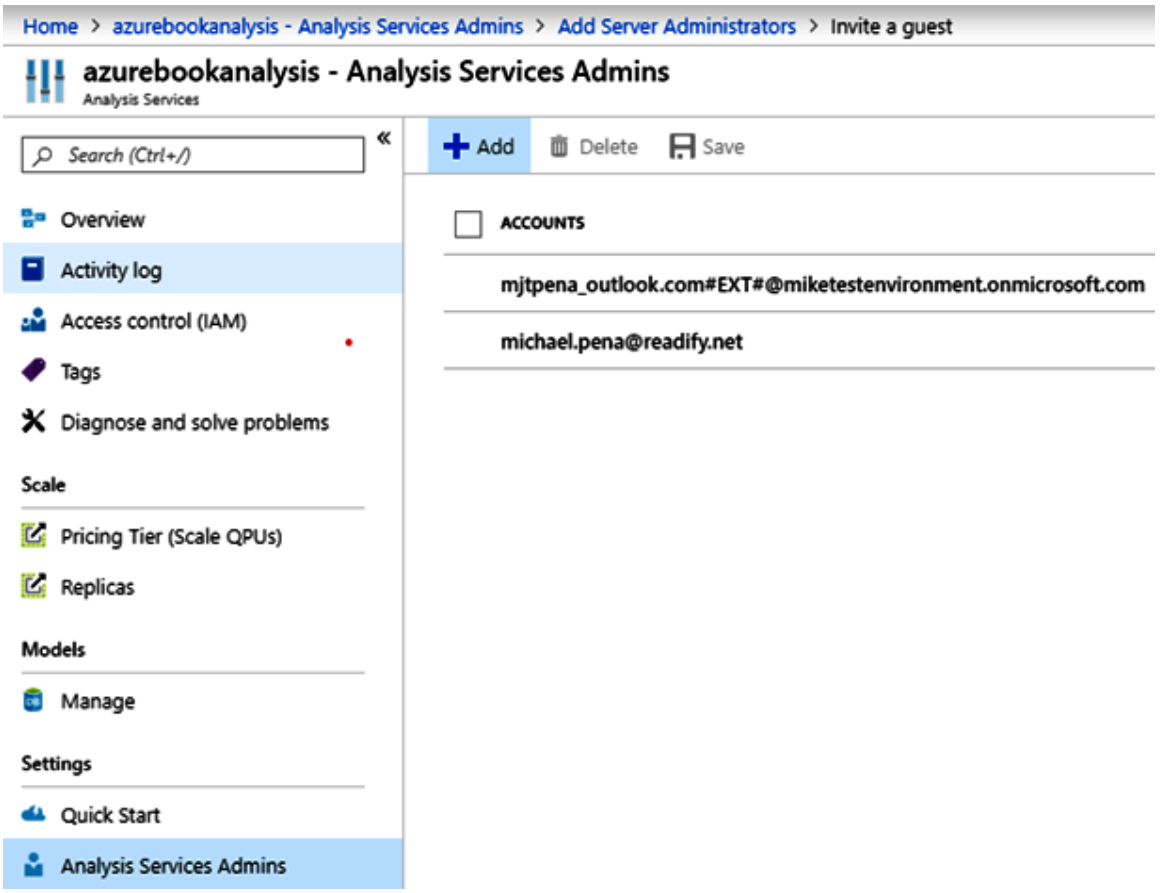

Abbildung 3.12: Hinzufügen eines Gastadministrators

2. Klicken Sie auf **Add** (Hinzufügen).

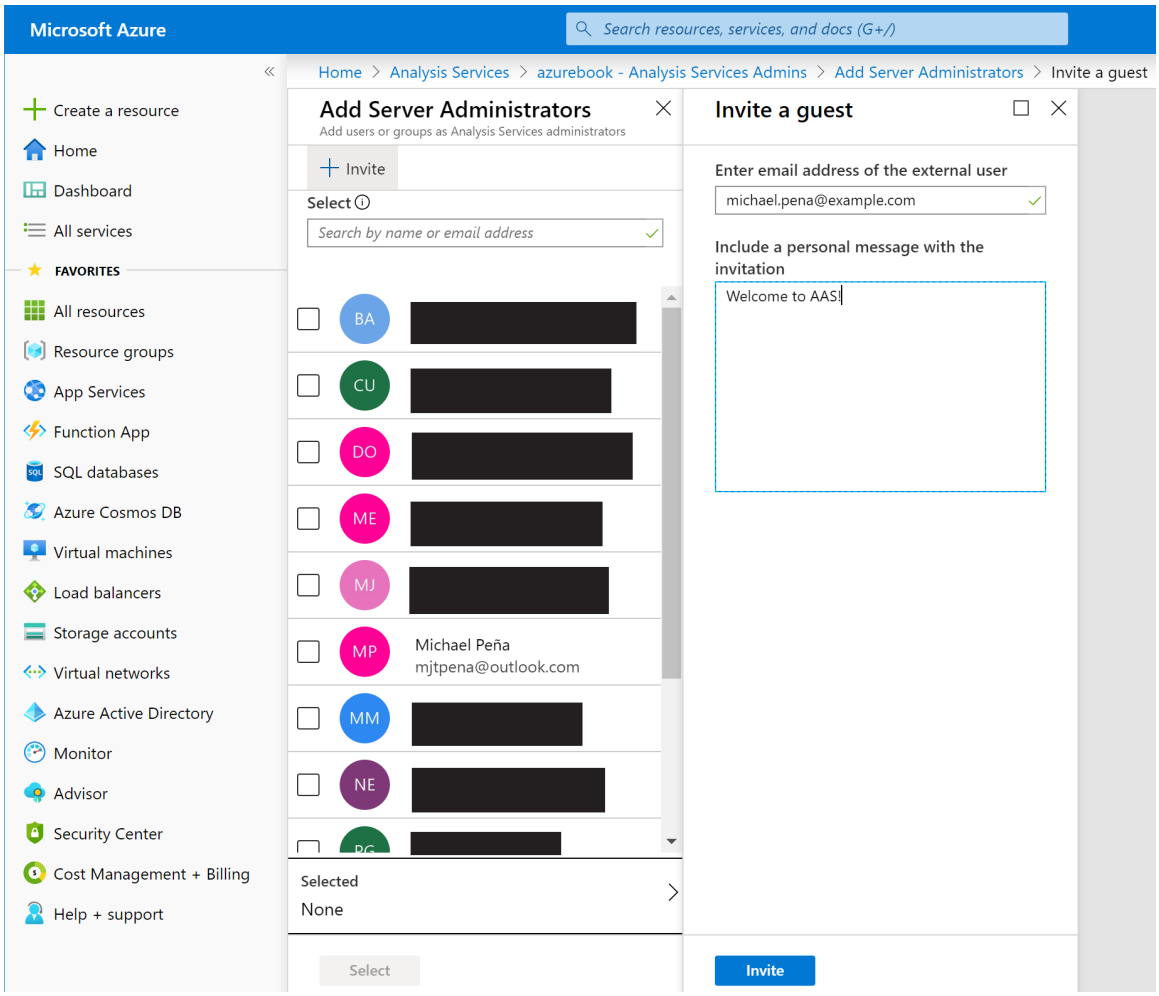

Abbildung 3.13: Einladen des Anwenders als Serveradministrator

Fügen Sie einen Anwender aus demselben Active Directory/Mandanten hinzu, oder laden Sie einen Gastanwender von einem anderen Mandanten ein. Fügen Sie die E-Mail-Adresse hinzu, die beim Erstellen eines Power BI-Kontos zugeordnet wurde.

3. Wenn Sie Power BI Desktop heruntergeladen, ein Power BI-Konto erstellt und Administratorzugriff auf den Azure Analysis Services-Server erteilt haben, klicken Sie auf die **PBIX**-Datei, die Sie zuvor von AAS heruntergeladen haben.

4. Daraufhin wird die Power BI Desktop-App geöffnet, und Sie werden möglicherweise aufgefordert, sich anzumelden. Verwenden Sie die Anmeldeinformationen, die Sie in den vorherigen Schritten zum Erstellen eines Power BI-Kontos verwendet haben. Wenn beim Öffnen der Datei ein Fehler aufgetreten ist, kann dies daran liegen, dass das Konto, das Sie zur Anmeldung bei Power BI verwendet haben, nicht Administrator von Azure Analysis Services ist. Daher kann keine Verbindung hergestellt werden.

> **Hinweis**
>
> Zum Erstellen eines Power BI-Berichts brauchen Sie Azure Analysis Services nicht. Power BI unterstützt verschiedene Datenquellen wie z. B. Azure Synapse Analytics direkt, dies wird allerdings nicht für die Produktion und den Einsatz in der Praxis empfohlen.

5. Stellen Sie nach dem Öffnen der Datei sicher, dass auf der rechten Seite der Desktop-App Felder zu sehen sind.

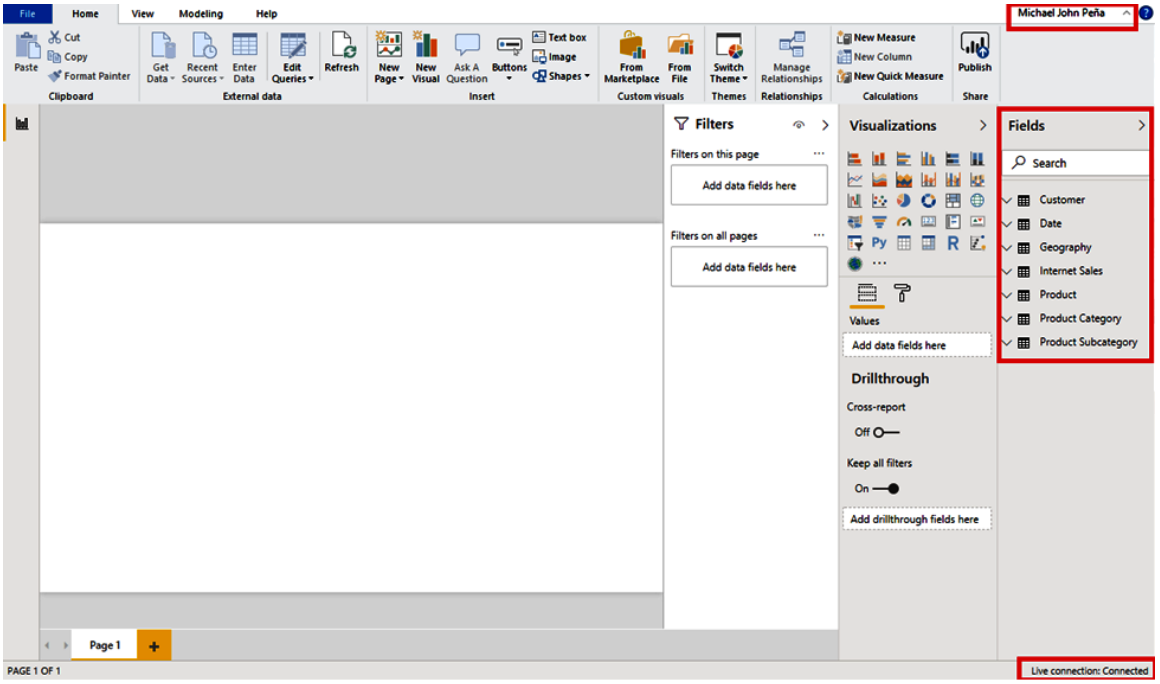

Abbildung 3.14: Feldabschnitt der ausgewählten Datenbank

Diese Felder geben an, dass Sie über eine aktive Verbindung mit Azure Analysis Services verfügen, da Sie jetzt das semantische Datenmodell abfragen können.

Sie können nun mit dem nächsten Teil der Aktivität, dem Erstellen des Live-Dashboards, fortfahren.

Visualisieren von Daten

Nachdem Ihnen nun Felder zum Experimentieren zur Verfügung stehen, ist es an der Zeit, Insights aus den Daten zu generieren. In einem Unternehmensszenario bietet Ihnen die Verwendung von AAS als Power BI-Datenquelle die Flexibilität, ein zwischengespeichertes Modell von Azure Synapse Analytics abzufragen, für das Sie über den erforderlichen Datenzugriff als Analyst verfügen. Sie müssten nicht warten, bis Ihnen ein Datenbankadministrator oder -entwickler eine CSV-Datei sendet, um mit der Visualisierung der Daten zu beginnen. Die Daten, zu denen Sie Berichte entwickeln, spiegeln im Prinzip eine Momentaufnahme Ihrer vorhandenen Datenbanken wider. In diesem speziellen Szenario kann `AdventureWorks` diese Daten aus mehreren Datenquellen ableiten, beispielsweise aus Datenbanken, Speicher, IoT-Sensoren und sozialen Medien. Es empfiehlt sich nicht, mit Power BI diese Datenquellen direkt abzufragen.

Power BI ist ein sehr intuitives Tool, und die meisten Schritte erfordern nur Drag-and-Drop und Klicks. Wenn die `AdventureWorks`-Datenbank unsere Datenquelle für AAS ist, sehen wir Felder, die sich auf ein Online-Einzelhandelsgeschäft beziehen, das Fahrräder und Zubehör verkauft. Die Felder sind bereits vorab für Sie ausgefüllt, im Allgemeinen haben die Felder jedoch die folgenden Bedeutungen:

- **Internet Sales** (Internetverkäufe): Verkaufsdaten in Zusammenhang mit den Produkten
- **Product** (Produkt): Metadaten zu dem Produkt, z. B. der Name des Produkts
- **Product Category** (Produktkategorie): eine Zuordnung der Produkte zu einer allgemeinen Kategorie
- **Product Subcategory** (Produktunterkategorie): eine untergeordnete Kategorisierung der Produkte
- **Customer** (Kunde): kundenbezogene Informationen über den Verkauf eines Produkts
- **Date** (Datum): ein Zeitmaß für den Verkauf der Produkte
- **Geography** (Geografie): eine Kennzahl für den Ort der Verkäufe der Produkte

Wir führen nun die folgenden Schritte aus, um die Daten in Power BI zu visualisieren:

1. Beginnen wir mit einer einfachen Tabelle **`Product Category`** (Produktkategorie), in der Produktnamen aufgeführt sind. Klicken Sie im Abschnitt **fields** (Felder) der App auf die Tabelle **Product Category** (Produktkategorie), und wählen Sie **Product Category Name** (Name der Produktkategorie) aus:

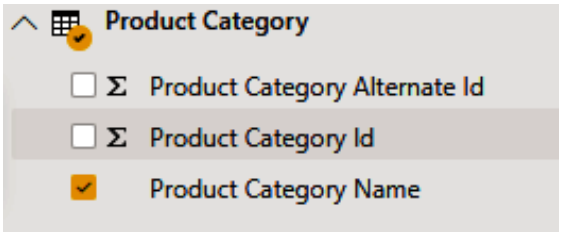

Abbildung 3.15: Tabelle „Product Category“ (Produktkategorie)

   Es wird eine Tabelle erstellt, die wie folgt aussieht:

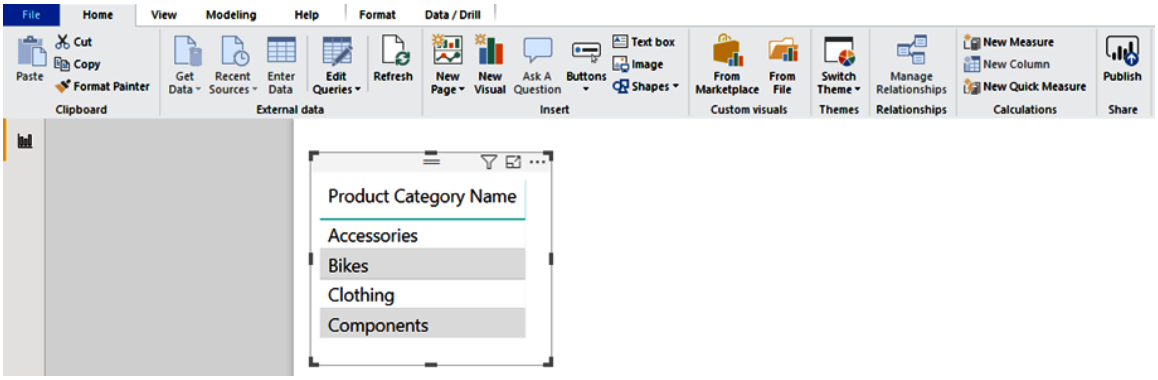

Abbildung 3.16: Tabelle „Product Category Name“ (Name der Produktkategorie)

   Wie Sie sehen, ist Power BI interaktiv und in der Lage, eine Tabelle mit der Liste der Kategorienamen anzuzeigen.

2. Als Nächstes werden wir eine ähnliche Tabelle für Produktunterkategorien anzeigen. Gehen Sie ähnlich wie im Feld **Categories** (Kategorien) vor. Wechseln Sie zum Feld **Product Subcategory** (Produktunterkategorie), und klicken Sie auf **Product Subcategory Name** (Name der Produktunterkategorie). Es wird dann eine weitere Tabelle erstellt, in der alle Unterkategorien aufgeführt sind.

   Interessant hierbei ist, dass Power BI angesichts der beiden Tabellen in der Lage ist, Korrelationen zwischen den beiden hervorzuheben.

3. Klicken Sie in der Tabelle **Product Category Name** (Name der Produktkategorie) auf die Zeile **Bikes** (Fahrräder):

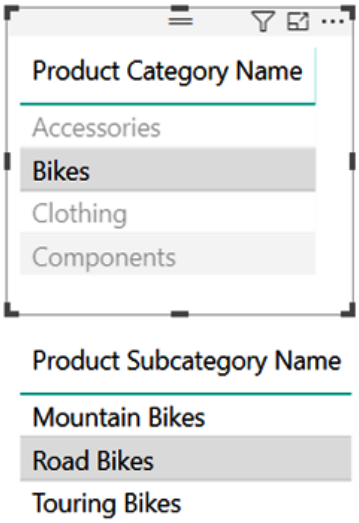

Abbildung 3.17: Power BI-Filterung von Produktunterkategorien

Wie Sie sehen, filtert Power BI automatisch alle Produktunterkategorien, die zur übergeordneten Kategorie gehören. Angezeigt werden die Optionen **Mountain Bikes**, **Road Bikes** (Rennräder) und **Touring Bikes** (Tourenräder), die alle der Produktkategorie **Bikes** (Fahrräder) zugeordnet sind. Idealerweise legen Sie (oder der Datenbankentwickler/-administrator) dies in Ihrem semantischen Datenmodell in Analysis Services fest.

4. Wenn Sie in **Product Category Name** (Name der Produktkategorie) erneut auf die Zeile **Bikes** (Fahrräder) klicken, wird der Filter der Unterkategorien entfernt.

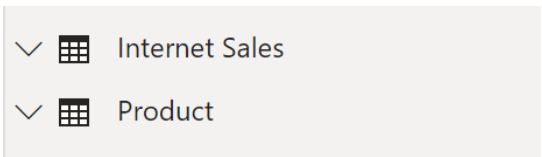

Abbildung 3.18: Felder „Internet Sales“ (Internetverkäufe) und „Product“ (Produkt)

5. Wir wollen nun eine weitere Tabelle erstellen, die **Internet Sales** (Internetverkäufe) und **Product** (Produkt) zuordnet. Klicken Sie im Feld **Internet Sales** (Internetverkäufe) auf **Internet Total Sales** (Gesamte Internetverkäufe) und im Feld **Product** (Produkt) auf **Category** (Kategorie). Power BI erstellt automatisch ein gruppiertes Säulendiagramm, wie in *Abbildung* 3.19 dargestellt:

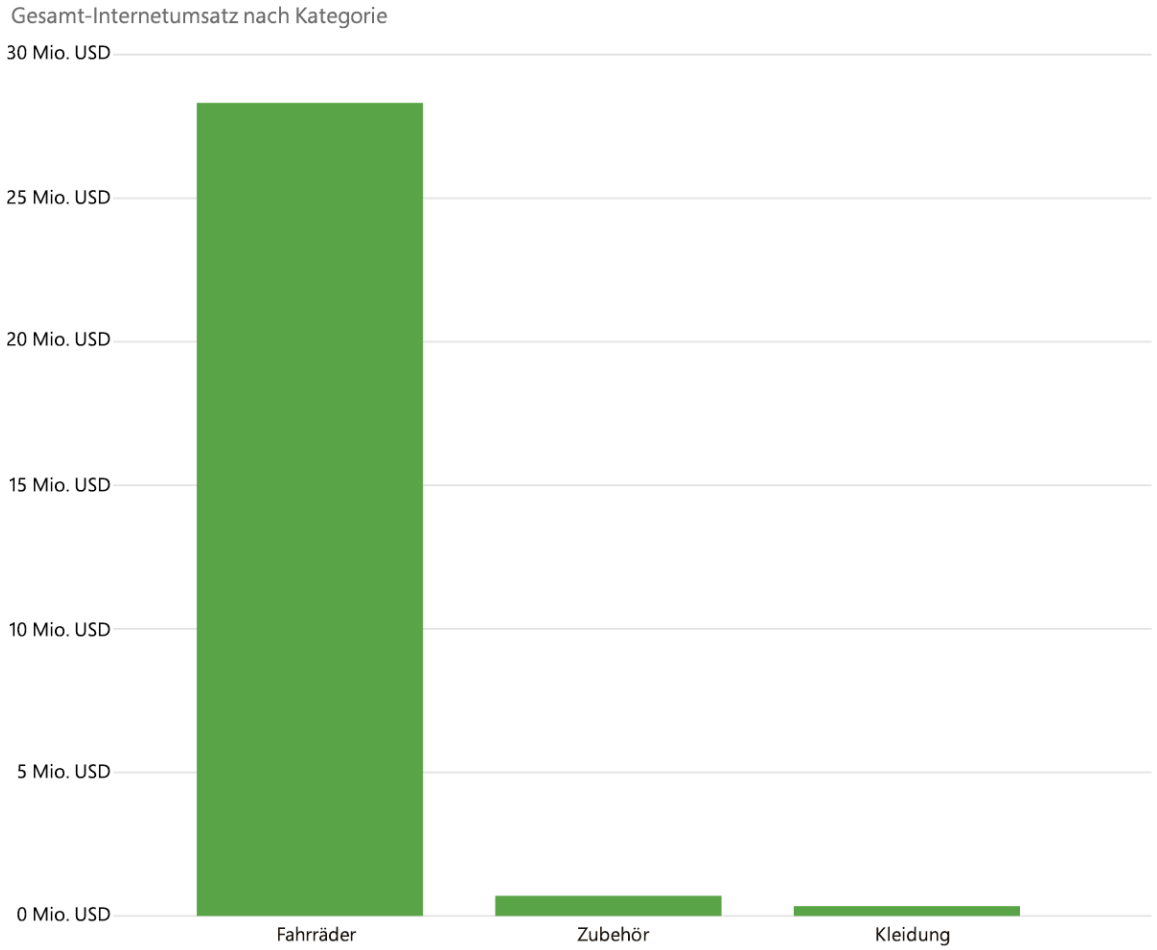

Abbildung 3.19: Gruppiertes Säulendiagramm

Power BI bietet Flexibilität durch die Bereitstellung verschiedener Visualisierungsoptionen für dasselbe Dataset.

6. Klicken Sie auf dasselbe Diagramm (**Internet Total Sales by Category** (Gesamte Internetverkäufe nach Kategorie)), und ändern Sie die Angabe auf der Registerkarte **Visualizations** (Visualisierungen) in **Matrix**:

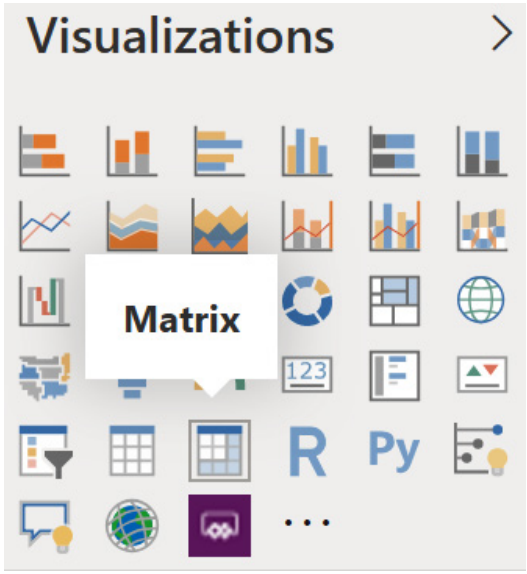

Abbildung 3.20: Auswählen des Matrixformats für die Visualisierung

Power BI erstellt dann eine Tabelle für die gesamten Internetverkäufe nach Kategorien.

Category	Internet Total Sales
Bikes	$28,318,144.65
Accessories	$700,759.96
Clothing	$339,772.61
Total	**$29,358,677.22**

Abbildung 3.21: Matrix-Visualisierung der gesamten Internetverkäufe nach Kategorie

Als Nächstes geht es darum, herauszufinden, in welchen Regionen des Lands die meisten Verkäufen verzeichnet wurden.

7. Navigieren Sie zum Feld **Internet Sales** (Internetverkäufe), und klicken Sie auf **Internet Total Sales** (Gesamte Internetverkäufe). Wechseln Sie zum Feld **Geography** (Geografie), und wählen Sie **Country Region Name** (Name des Lands/der Region) aus. Power BI erstellt dann ein gruppiertes Säulendiagramm. Ändern Sie die Visualisierung in ein Ringdiagramm (**Donut chart**) oder ein Kreisdiagramm (**Pie chart**), um einen schnellen Überblick über die Leistungskorrelationen für alle Regionen zu gewinnen. Wenn Sie auf ein Segment des Rings oder Kreises zeigen, werden Ihnen indikative Kennzahlen zu der betreffenden Region angezeigt:

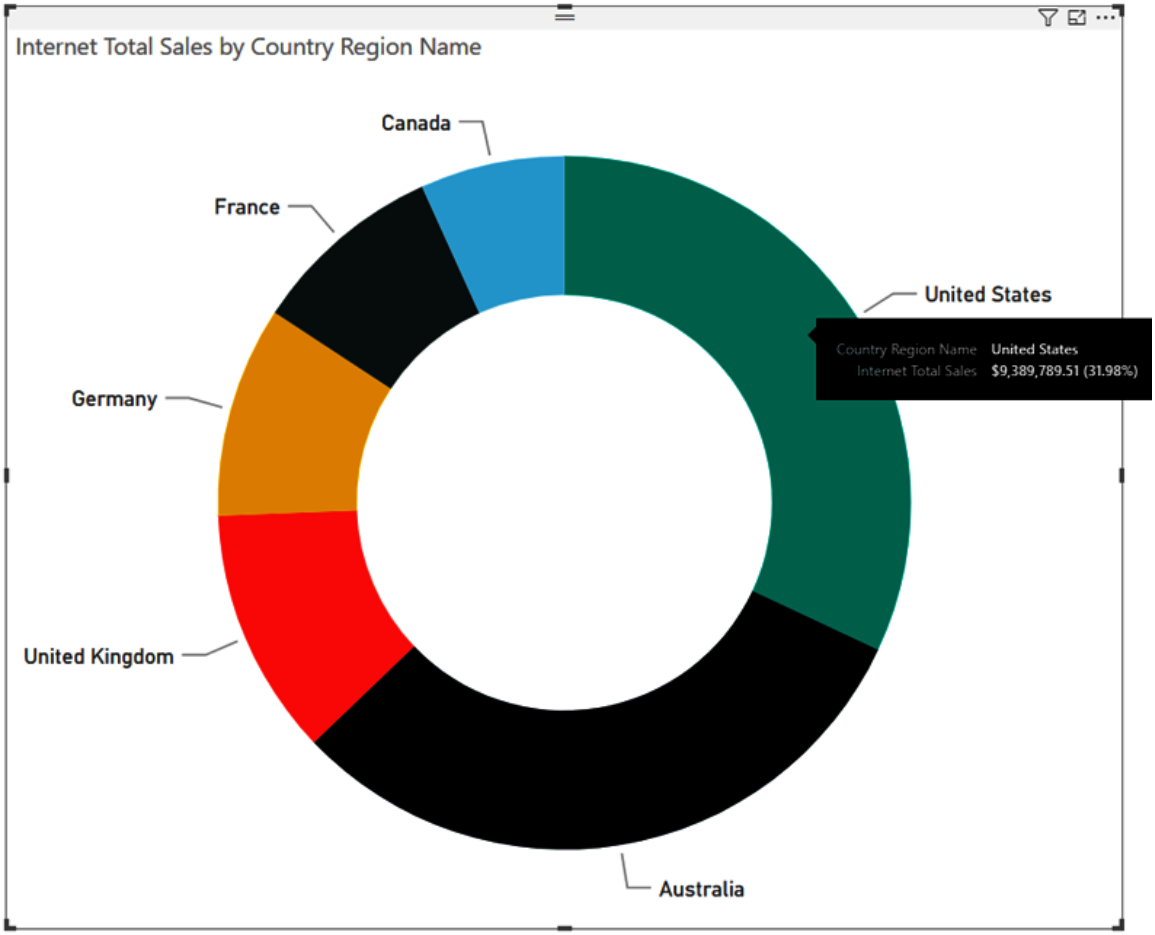

Abbildung 3.22: Gesamte Internetverkäufe nach Region – Ringdiagramm

8. Verwenden Sie neben dem Diagramm dasselbe Datenfeld (**Country Region Name** (Name des Lands/der Region) und **Internet Total Sales** (Gesamte Internetverkäufe)), und verwenden Sie stattdessen ein Visualisierungselement **Filled Map** (Flächenkartogramm). Power BI zeigt eine Kartenansicht Ihrer weltweiten Verkäufe an. Wenn Sie auf ein Element klicken, reagieren alle anderen Diagramme den zugehörigen Werten entsprechend.

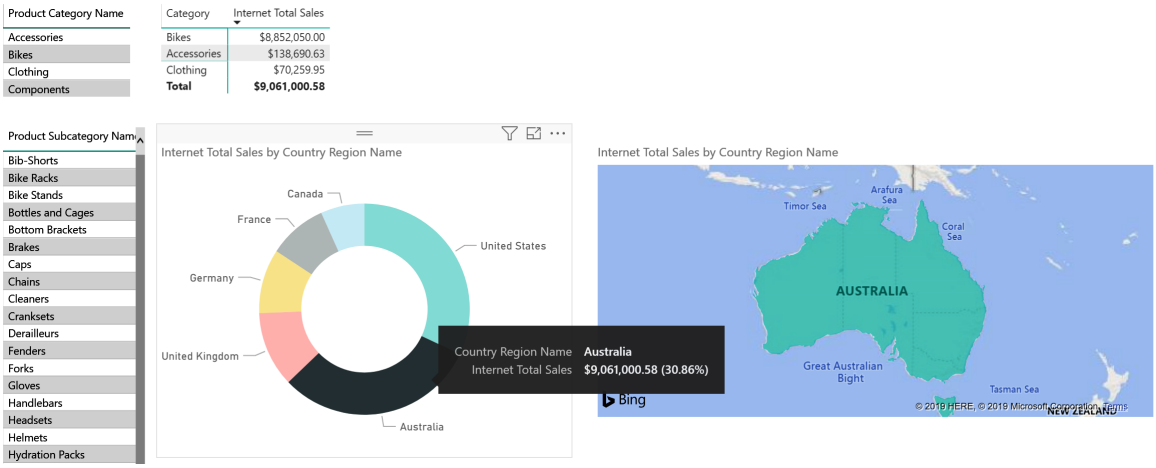

Abbildung 3.23: Flächenkartogramm-Visualisierung der Gesamtverkäufe nach Region

9. Fügen Sie nun ein Diagramm **Internet Total Sales by Year** (Gesamte Internetverkäufe nach Jahr) hinzu. Klicken Sie im Feld **Internet Sales** (Internetverkäufe) auf **Internet Total Sales** (Gesamte Internetverkäufe), und klicken Sie zur Nachverfolgung im Feld **Date** (Datum) auf **Fiscal** (Steuerlich). Es wird dann ein gruppiertes Säulendiagramm erstellt, wie im folgenden Diagramm dargestellt.

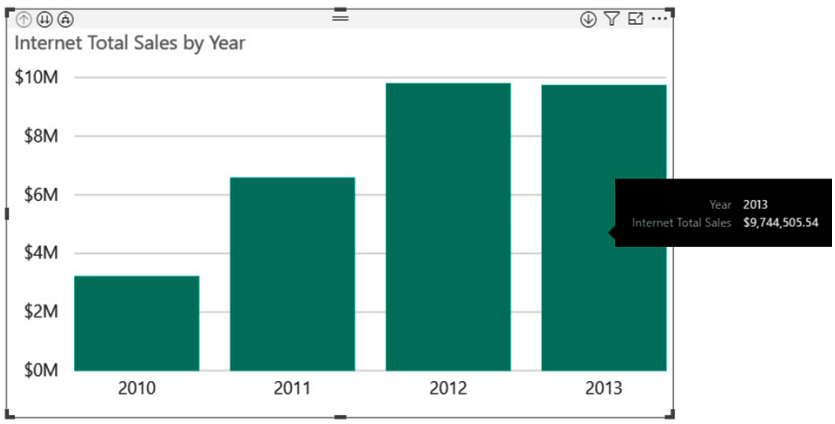

Abbildung 3.24: Gesamte Internetverkäufe nach Jahr: gruppiertes Säulendiagramm

10. Bei dem letzten Diagramm, das wir erstellen werden, handelt es sich lediglich um eine Zuordnung zwischen Kunden und Internetverkäufen. Klicken Sie unter **Internet Sales** (Internetverkäufe) erneut auf **Internet Total Sales** (Gesamte Internetverkäufe), und stellen Sie in diesem Fall eine Verknüpfung mit den Feldern **Commute Distance** (Entfernung zur Arbeit) und **Gender** (Geschlecht) des Kunden her. Ändern Sie die Visualisierung in ein gruppiertes Balkendiagramm.

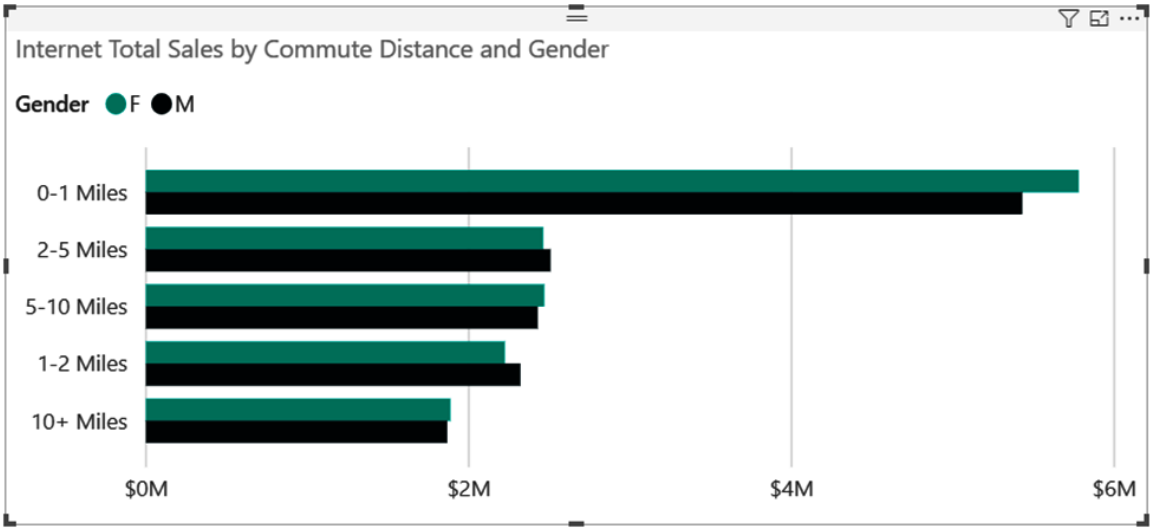

Abbildung 3.25: Gesamtverkäufe nach Entfernung zur Arbeit und Geschlecht: Balkendiagramm

So können Sie einige Insights über Ihre Kunden gewinnen. Ihre Produkte werden am meisten von Kunden gekauft, deren Fahrtweg zur Arbeit maximal 1 Meile (1,6 km) beträgt. Dies ist verständlich, da `AdventureWorks` Fahrräder und Zubehör verkauft. Bei denen, die Fahrräder kaufen, wird sich der Arbeitsplatz höchstwahrscheinlich in Fahrradnähe befinden. Es ist auch ersichtlich, dass in Bezug auf die Verkäufe kein großer Unterschied zwischen den Geschlechtern besteht. Die indikative Kennzahl ist die Entfernung zur Arbeit.

11. Wenn Sie Ihrem Dashboard einige Formatvorlagen hinzufügen möchten, klicken Sie oben in der App auf das Menüband **Switch Theme** (Design wechseln). Wählen Sie die Farbe Ihrer Wahl.

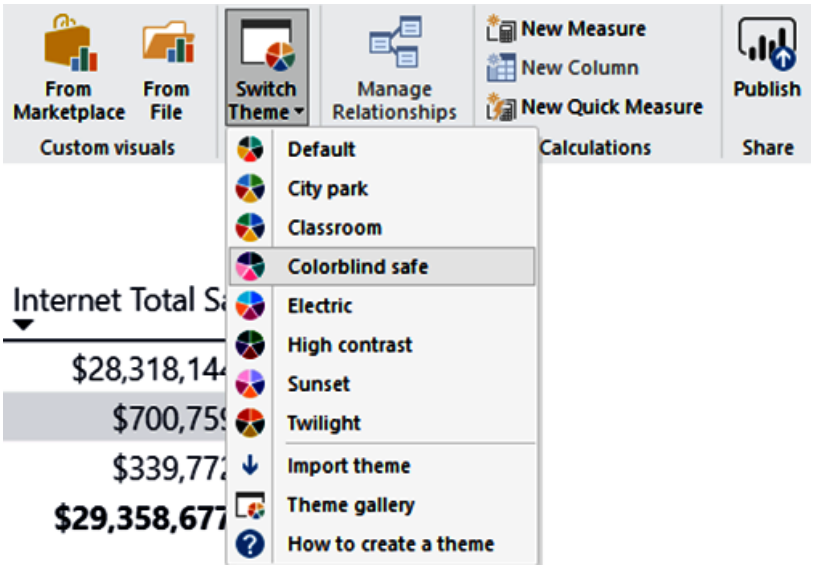

Abbildung 3.26: Hinzufügen von Formatvorlagen zu Ihrem Dashboard

Sie können nun das Dashboard in Ihrem Arbeitsbereich speichern und dieses Diagramm für Ihre Kollegen freigeben.

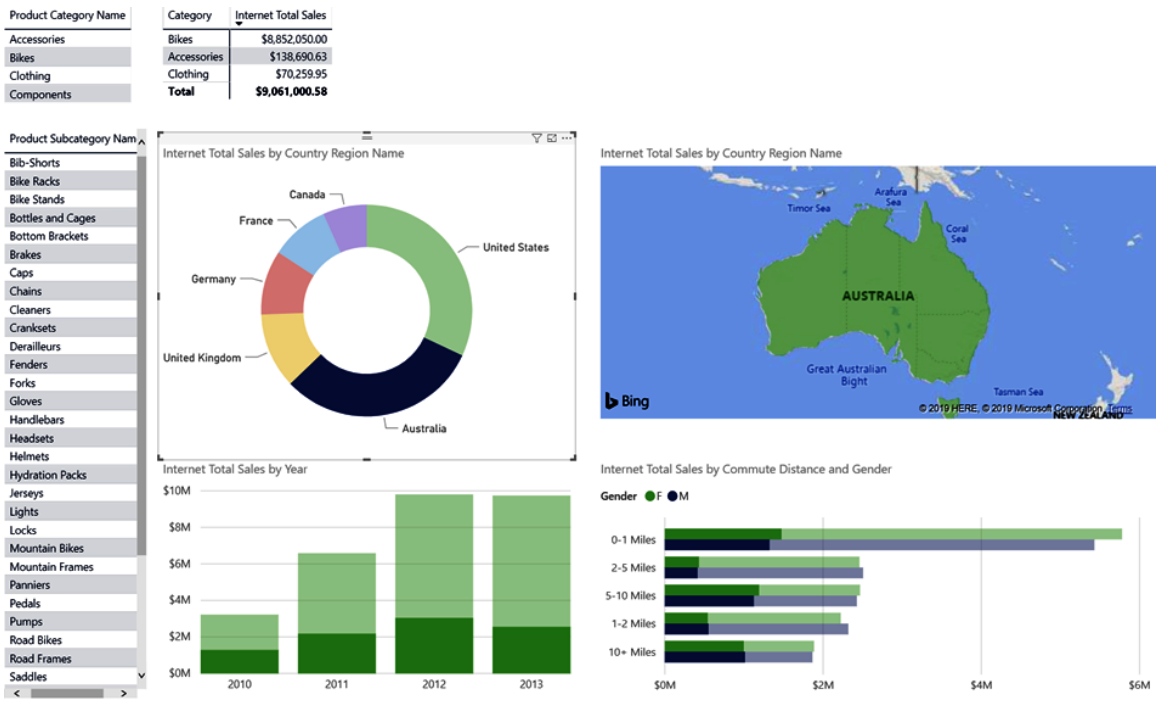

Abbildung 3.27: Dashboard „Final Sales“ (Endgültige Verkäufe)

In diesem Abschnitt haben wir gesehen, welche Möglichkeiten Power BI bietet, um mit wenigen Klicks und Drag-and-Drop über Power BI Desktop umfangreiche Diagramme zu erstellen. Von semantischen Datenmodellen ausgehend können wir aussagekräftige Daten darstellen, die für die Anwender leicht verständlich sind.

Veröffentlichen des Dashboards

Nachdem Sie Ihre Diagramme und Ihr Dashboard erstellt haben, ist es nun an der Zeit, diese in Ihrem Arbeitsbereich zu veröffentlichen. Gehen Sie hierfür wie folgt vor:

1. Klicken Sie auf das Symbol **Publish** (Veröffentlichen) in dem Menüband im oberen Bereich der App. Wählen Sie einen Zielarbeitsbereich aus. Für diese Aktivität können Sie „My workspace“ (Mein Arbeitsbereich) als Ziel auswählen.

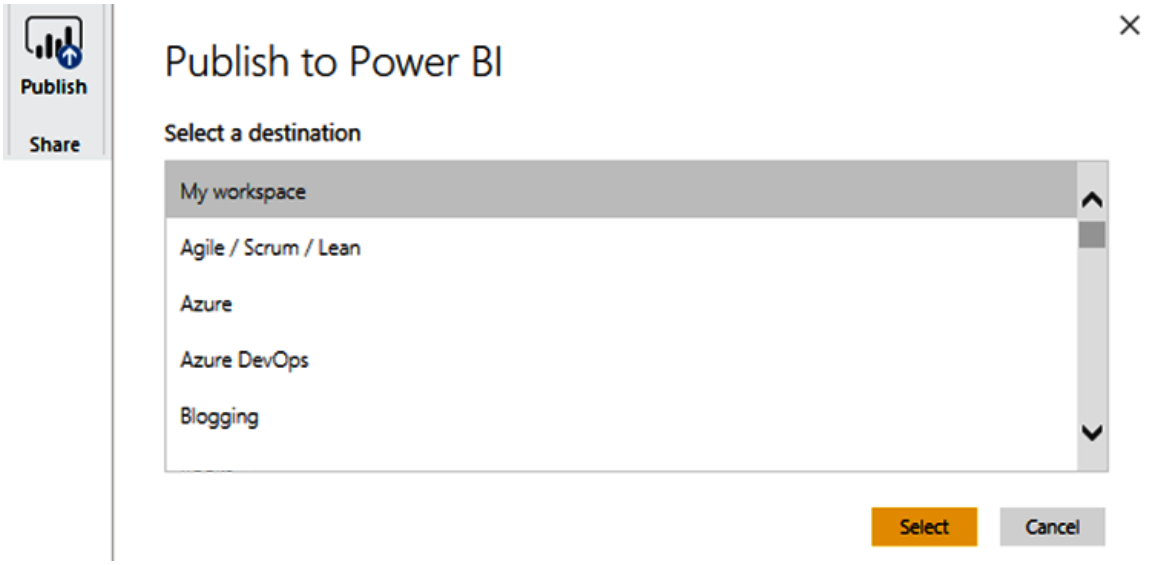

Abbildung 3.28: Veröffentlichen des Dashboards in Power BI

Warten Sie, bis Power BI die Veröffentlichung der Berichte abgeschlossen hat. Nach der Veröffentlichung wird ein Link angezeigt, über den Sie von einer Webseite aus auf das Dashboard zugreifen können.

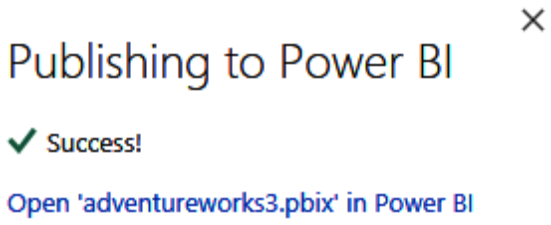

Abbildung 3.29: Veröffentlichen des Dashboards in Power BI 2

2. Möglicherweise müssen Sie sich anmelden, um auf das Dashboard zugreifen zu können.

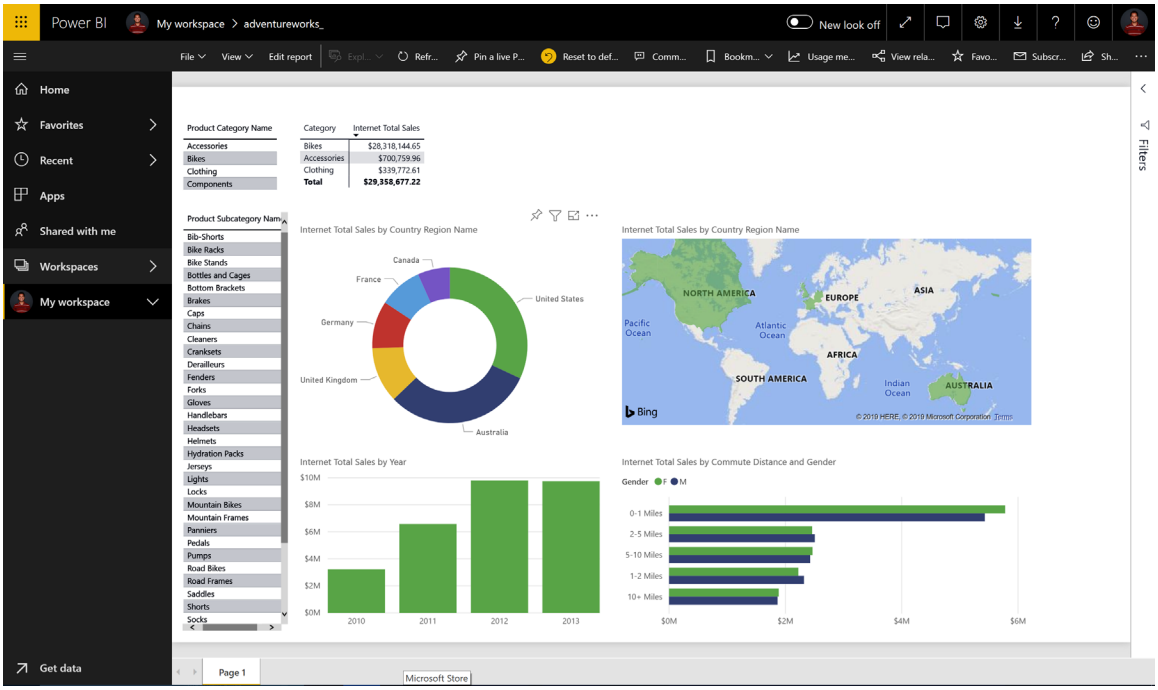

Abbildung 3.30: Zugriff auf den Bericht über eine Webseite

3. Klicken Sie oben rechts auf der Webseite auf **Share** (Freigeben).

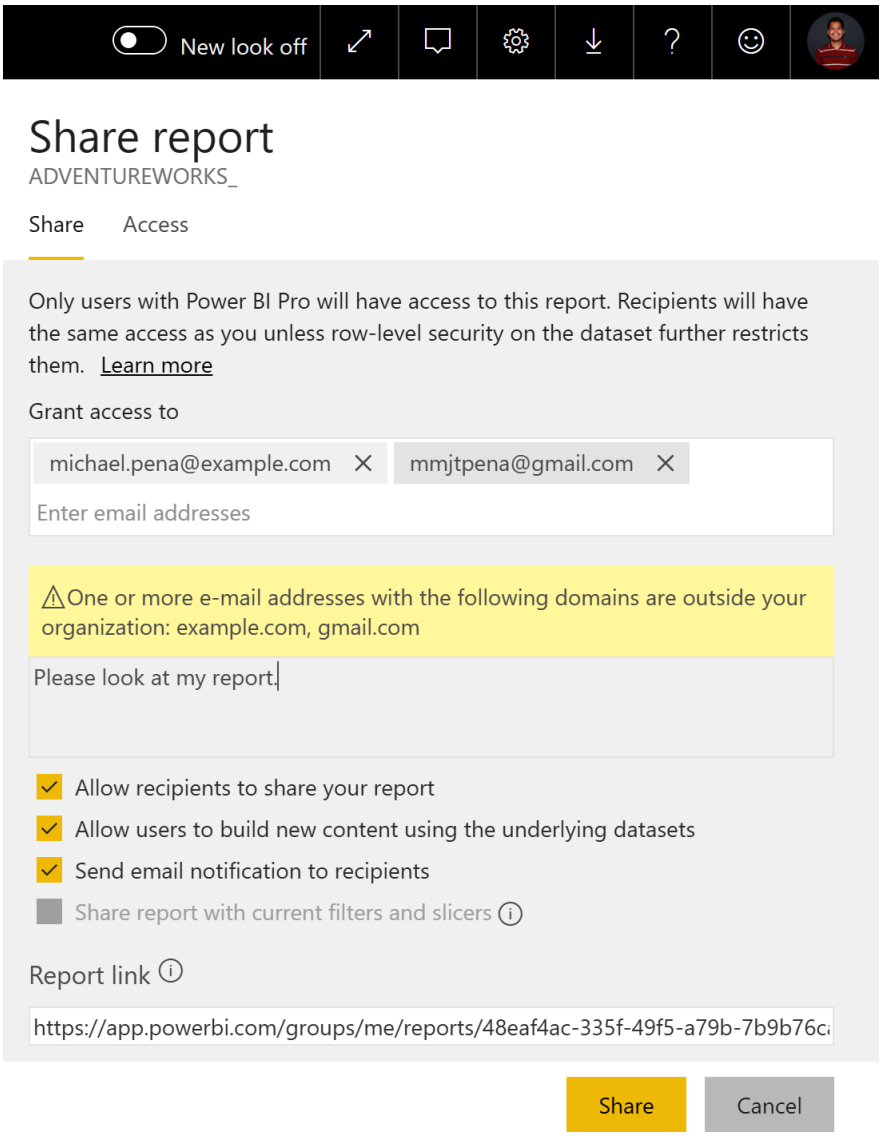

Abbildung 3.31: Freigeben des Berichts für mehrere Anwender

Sie können den Bericht für Ihre Kollegen (in derselben Domäne) oder für Gastanwender freigeben.

Die Anwender können auch eine mobile Version des Berichts anzeigen, wenn sie die mobile Power BI-App aus dem Apple iOS App Store oder Google Play Store herunterladen.

In der mobilen App können Sie Ihre eigenen Berichte sowie die für Ihr Konto freigegebenen Berichte anzeigen. Die folgende Abbildung zeigt eine mobile Ansicht des zuvor erstellten **AdventureWorks**-Berichts:

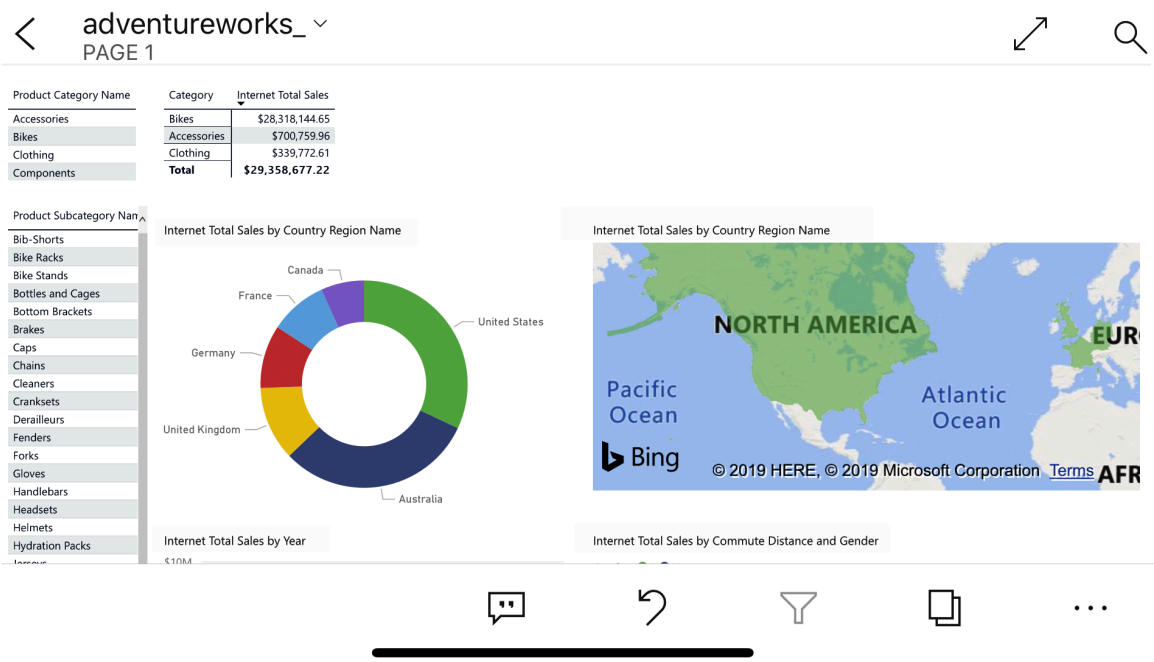

Abbildung 3.32: Zugriff auf Power BI-Bericht auf mobiler Power BI-App

Sie können dann mit Ihrem Team zusammenarbeiten und dem Bericht Kommentare und Anmerkungen hinzufügen. Dies bietet eine sehr intuitive Erfahrung im Vergleich zum Teilen einer PowerPoint-Präsentation mit Ihrem Vorgesetzten und Ihren Kollegen.

Im nächsten Abschnitt werden wir unter Verwendung von Azure Machine Learning Services Machine Learning auf Azure implementieren. Wir werden uns auch kurz Azure Databricks bei der Durchführung einer leistungsstarken Analyse von Datensätzen ansehen.

Machine Learning auf Azure

Es gibt mehrere Möglichkeiten des Machine Learning auf Azure. Microsoft macht Data Science allen Anwendern leichter zugänglich und verhilft Datenwissenschaftlern zu mehr Produktivität. Microsoft bietet Entwicklern, Datenbankentwicklern und Datenwissenschaftlern eine Reihe von Technologien zum Erstellen von Machine-Learning-Algorithmen. Unabhängig von Ihrem Kenntnisstand und Ihren Kompetenzen im Bereich Data Science steht ein hilfreicher Dienst, ein Tool oder Framework von Microsoft zur Verfügung, um Ihren Weg des Machine Learning zu beschleunigen.

Die folgende Abbildung zeigt eine Machine-Learning-Landschaft innerhalb der Microsoft Azure-Infrastruktur. Mithilfe von Azure Cognitive Services können Sie vortrainierte Modelle verwenden und diese direkt mit Ihren Anwendungen integrieren, ohne eine Datenpipeline einrichten zu müssen. Sie können gängige Frameworks wie **TensorFlow** und **Keras** in Azure verwenden, unabhängig davon, ob die Installation in einer virtuellen Maschine oder über einen Machine Learning-Arbeitsbereich erfolgt. Sie können verschiedene Plattformen wie Azure Machine Learning Services und Azure Databricks auswählen, um Ihre ML-Experimente vorzubereiten und auszuführen.

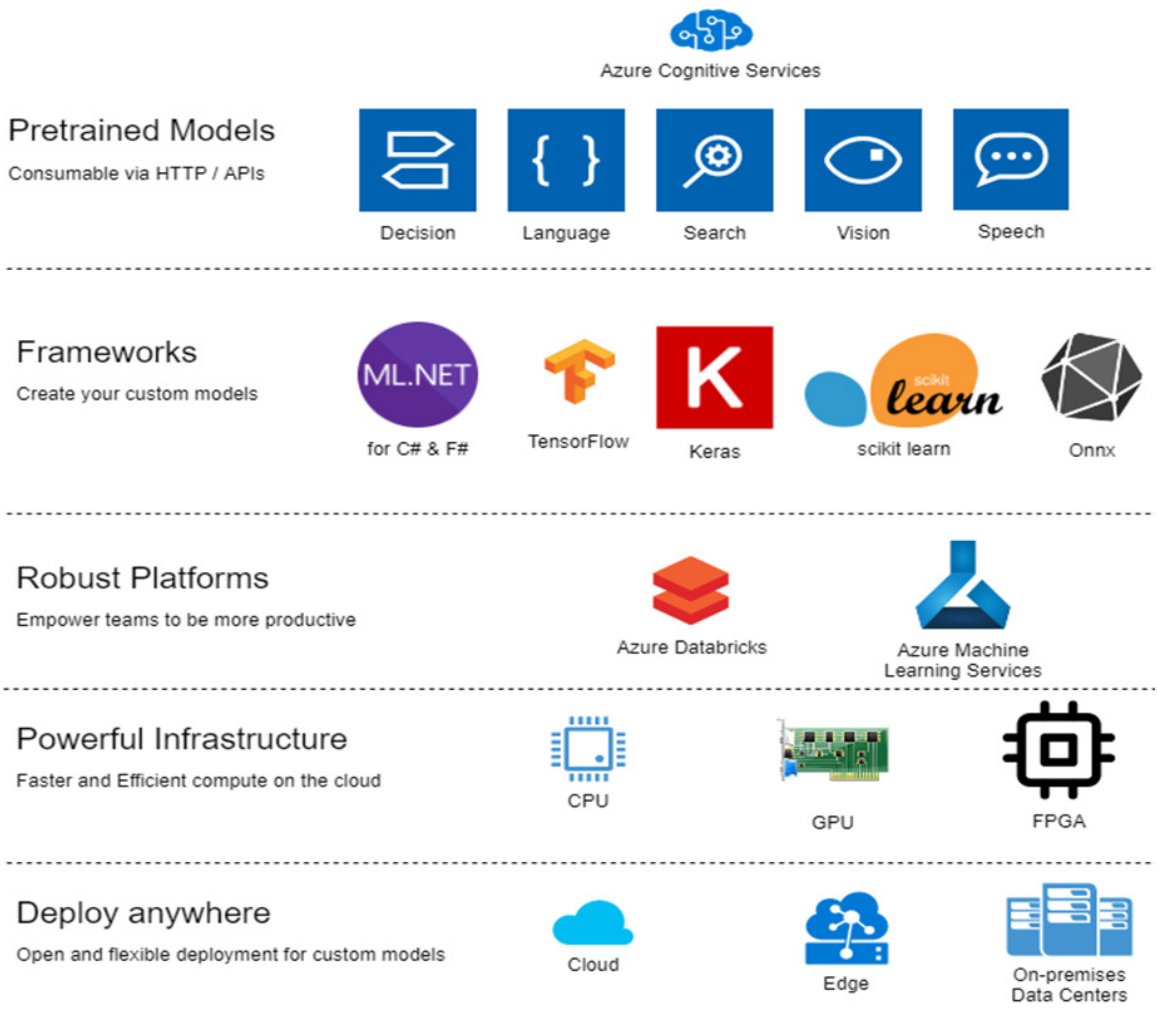

Abbildung 3.33: Microsoft Azure-Features und -Dienste für Machine Learning

Wenn Sie Azure zur Unterstützung Ihrer Berechnungen für die Machine-Learning-Analyse verwenden, erhalten Sie spezielle Hardware, die Ihre Experimente beschleunigen kann. Mit beschleunigter Hardware wie z. B. schnellen **Grafikprozessoren** (**Graphics Processing Units, GPUs**) und **Field Programmable Gate Arrays** (**FPGAs**) können Sie Milliarden von Datensätzen lesen und verschiedene Modelle gleichzeitig testen, um schnellere Ergebnisse bei Ihren ML-Experimenten zu erzielen.

In den folgenden Abschnitten erhalten Sie eine Übersicht der wichtigsten Technologien und Plattformen zur Implementierung von Machine Learning und KI in der Microsoft Azure-Infrastruktur.

ML.NET

ML.NET ist ein plattformübergreifendes Open-Source-Framework für **.NET**-Entwickler. Mit ML.NET können Sie und Ihr Team bereits in der .NET-Infrastruktur vorhandene Fähigkeiten, Bibliotheken und Frameworks nutzen. Mit ASP.NET können Sie Webanwendungen, mit **Xamarin** mobile Anwendungen, mit WPF Desktop-Anwendungen und mit Windows IoT sogar IoT-Anwendungen erstellen. Darüber hinaus können Sie mit TensorFlow und **ONNX** die Erstellung von ML-Modellen erweitern. ML.NET bietet sofort verfügbare Unterstützung für Algorithmen, die Stimmungsanalysen, Produktempfehlungen, Objekterkennung, Umsatzprognosen und viele weitere reale Geschäftsszenarien ermöglichen.

Für Aufgaben wie Regression und Klassifizierung können Training und Nutzung mithilfe von ML.NET erfolgen. Abgesehen hiervon werden Kerndatentypen, erweiterbare Pipelines, Datenstrukturen, Toolsupport, fortgeschrittene Mathematik und mehr unterstützt.

ML.NET kann über nuget.org installiert werden. Nuget.org ist ein öffentliches Repository mit .NET-Paketen, Bibliotheken und Frameworks zum Herunterladen, die Sie ganz einfach Ihrem .NET-Projekt hinzufügen können.

AutoML

Automatisiertes Machine Learning (**AutoML**) ist ein Microsoft-Forschungsprojekt, das Machine Learning für alle einfacher machen soll. AutoML ist dafür konzipiert, automatisch die besten ML-Modelle für Sie zu ermitteln. Zum Zeitpunkt des Schreibens dieses Dokuments ist AutoML in der Lage, automatisch den richtigen Algorithmus auszuwählen und die Optimierung von Hyperparametern für Prognosen, Klassifizierung und Regression zu unterstützen. Dies ist sehr hilfreich, wenn Sie keinen Datenwissenschaftler in Ihrem Team haben.

AutoML hilft Anwendern (Entwicklern, Analysten oder sogar Datenwissenschaftlern), Machine Learning ohne große Zugangsbarriere in Bezug auf Programmiersprachen, Bibliotheken, Frameworks und Data-Science-Konzepte zu implementieren. Kürzere Markteinführungszeiten aufgrund eines iterativen Prozesses bieten Unternehmen Möglichkeiten zur Innovation. Außerdem können die Unternehmen bei der Durchführung von Experimenten Best Practices auf dem Gebiet der Data Science nutzen.

Der Anwendungsbereich ist derzeit jedoch noch begrenzt. Zum jetzigen Zeitpunkt können nur bestimmte Algorithmen ausgeführt werden, und es kann nur eine begrenzte Anzahl von Datenmodellen erstellt werden.

Azure Machine Learning Studio

Azure Machine Learning Studio ist ein visuelles Drag-and-Drop-Tool, mit dem die Anwender ihre Machine-Learning-Experimente ohne Codierung intuitiv durchführen können. Dies beinhaltet die Verbindung von Datenquellen, die Durchführung von Analysen und die Bereitstellung des trainierten Modells als Webdienst unter Verwendung eines API-Schlüssels. Azure Machine Learning Studio unterstützt grundlegende Machine-Learning-Algorithmen, die Klassifizierung, Regression und Clustering umfassen.

Azure Databricks

Im vorigen Kapitel haben wir uns bereits ausführlich mit Azure Databricks befasst. Azure Databricks bietet Machine-Learning-Möglichkeiten durch die Verwendung von Databricks Runtime for Machine Learning (Databricks Runtime ML) auf Ihren virtuellen Knoten. Databricks Runtime ML enthält gängige Bibliotheken wie TensorFlow, **PyTorch**, Keras und **XGBoost** für ML-Analysen im großen Maßstab. Databricks übernimmt auch die Installation dieser Frameworks für Sie. Azure Databricks kann auch **Apache Spark MLlib** verwenden und eine Hyperparameter-Optimierung mit **MLFlow** durchführen.

Cognitive Services

Microsoft Cognitive Services ist eine Suite von cloudbasierten, universellen, vortrainierten Modellen und APIs, die für ein weiteres Training für spezifische Anwendungsfälle genutzt und erweitert werden können. Wenn Sie beispielsweise KI für die Objekterkennung erstellen möchten, die weiß, was eine Banane ist, müssen Sie möglicherweise weitere Daten einspeisen, damit die KI erkennen kann, dass auf dem Bild eine Banane zu sehen ist. Die Nutzung von Cognitive Services erfolgt über HTTP und ist **plattformunabhängig**, Sie können also jede Programmiersprache und jedes Betriebssystem verwenden. Es gibt fünf Hauptkategorien von Cognitive Services: Entscheidungen, Bildanalyse, Spracheingabe, Suche und Sprache. Mit Cognitive Services können Sie KI und ML problemlos mit Ihren mobilen Apps, Ihren Web-, Desktop- oder sogar IoT-Anwendungen integrieren.

Die Sprach- und Sprechererkennungsfunktionen der Speech Services-API sind gute Beispiele für Cognitive Services. Mithilfe dieser Funktionen können Sie Sprachdaten in Text umwandeln, in andere Sprachen übersetzen und die Identität des Referenten erkennen, ohne einen Machine Learning-Arbeitsbereich mit Millionen von Datensätzen und einer Reihe von ML-Modellexperimenten einzurichten.

Die Verwendung von Cognitive Services ist optimal geeignet für alle, die nach einer einfachen Möglichkeit suchen, KI und ML mit nur minimalen Data-Science-Kenntnissen in ihre Anwendungen zu integrieren. Microsoft bietet sehr flexible Preisoptionen, bei denen Sie nur für die tatsächliche Nutzung zahlen. In den meisten Fällen gibt es kostenfreie Tarife zum Entdecken der Dienste.

Mehr über Cognitive Services erfahren Sie hier.

Bot Framework

Mit Microsoft Bot Framework können Anwendungen intelligente Bots entwickeln (oft für Chatbots verwendet), um Workflows zu automatisieren. Bot Framework ist eng mit Microsoft Cognitive Services wie dem **Language Understanding Intelligence Service** (**LUIS**) und **QnA Maker** verknüpft. QnA Maker ist ein Dienst zur Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), mit dem es schneller möglich ist, konversationsbasierte KI wie Chatbots zu erstellen. Mit Bot Framework können Entwickler ganz einfach Konversations-KI erstellen, die durch Training aus Äußerungen und Absichten lernt. Darüber hinaus können die Entwickler den Bot problemlos auf verschiedenen Kanälen wie Microsoft Teams, Cortana und Slack veröffentlichen.

Bot Framework wird heute von großen Unternehmen wie Banken und Einzelhandelskonzernen vielfach für den **First-Level**-Support genutzt. Die Bank von Beirut beispielsweise hat mithilfe von Azure Bot Framework den Chatbot „Digi Bob“ geschaffen, der Anwender bei der Beantragung von Darlehen und der Inanspruchnahme anderer Bankdienstleistungen unterstützt. Mehr über diesen Anwendungsfall erfahren Sie hier.

Mit Bot Framework können Entwickler intelligente Bots auf Unternehmensniveau bereitstellen, die Anfragen und Nachrichten (Absichten) von Anwendern leicht übersetzen und mit sinnvollen Maßnahmen reagieren können. Zu diesen Maßnahmen können die Abfrage einer Datenquelle oder die Orchestrierung eines Befehls in einem System gehören. Mehr über Bot Framework erfahren Sie hier.

In der Microsoft-Infrastruktur gibt es noch weitere Machine-Learning-Tools und -Produkte, wie z. B.:

- SQL Server Machine Learning Services
- Microsoft Machine Learning Server
- Azure Data Science Virtual Machine
- Windows ML
- MMLSpark
- Azure Notebooks
- Azure Batch
- ML Services in HDInsight
- ML in Power BI
- Azure Machine Learning für VS Code
- Ausführung eigener ML-Frameworks auf einem Linux-Container oder Server-Image

Alle oben genannten Technologien zu behandeln, würde den Rahmen dieses Buchs sprengen. Daher konzentrieren wir uns auf Azure Machine Learning Services. Mehr über die oben genannten Dienste erfahren Sie unter diesem Link.

Features und Vorteile von Azure Machine Learning Services

AMLS bietet eine Vielzahl von Features sowie Flexibilität für Anwender mit unterschiedlichem Hintergrund und Wissen. AMLS kann in Ihre vorhandene Datenpipeline integriert werden, um beispielsweise Daten aus Azure Data Lake oder Azure Synapse Analytics zu nutzen und die Modelle direkt in Power BI bereitzustellen. Darüber hinaus können Sie mit Azure Databricks die Hardwarecluster, in denen Sie Ihre Machine-Learning-Experimente ausführen, weiter automatisieren.

AMLS bietet einen End-to-End-Arbeitsbereich zur Ausführung von Machine-Learning-Operationen. Mit AMLS können Sie unter Verwendung von AutoML, der grafischen Benutzeroberfläche oder dem **Software Development Kit** (**SDK**) in Ihrem ML-Notebook Experimente erstellen. Sie können auch ein portierbares Datenmodell erstellen, das in einem Container ausgeführt werden kann. Dieses Modell kann dann in ACI (Azure Container Instances) veröffentlicht werden.

Software Development Kit (SDK)

Azure Machine Learning bietet ein Python-SDK, das ausgereifte Frameworks wie **MXNet**, TensorFlow, PyTorch und **Scikit-learn** vollumfänglich unterstützt. Sie können die SDKs mit Jupyter Notebooks, Azure Notebooks oder sogar Visual Studio Code in Ihre Experimente importieren.

Grafische Benutzeroberfläche

Zum Erstellen und Ausführen von Experimenten können Sie auch eine grafische Benutzeroberfläche (mit minimaler erforderlicher Codierung) verwenden. Die Oberfläche ist ähnlich wie bei Azure Machine Learning Studio, wo Sie viele Drag-and-Drop-Tools und Verbindungsentitäten verwenden. Es handelt sich hier um eine intuitive Möglichkeit, Datenquellen zu verbinden und ein ML-Modell zu erstellen, das trainiert und bereitgestellt werden soll.

AutoML

AutoML ist ein Mechanismus, um den besten Algorithmus für Ihre Experimente vorzuschlagen. Dieses Feature ist in AMLS integriert. Sie können zeitintensive Aufgaben wie die Datenbereinigung und die Auswahl der richtigen Algorithmen für Ihr Modell automatisieren. Mit AutoML können Sie schnell viele Kombinationen von Algorithmen und Hyperparametern durchlaufen, um das beste Modell für Ihr gewünschtes Ergebnis zu finden.

Flexible Bereitstellungsziele

Microsoft und Azure beschränken Ihre Optionen für die Bereitstellung von Modellen nicht. Selbst wenn Sie Ihren Arbeitsbereich verwalten und die Analyse in der Cloud durchführen, sind Sie nicht gezwungen, das Ergebnis Ihrer Experimente nur auf Azure bereitzustellen. Sie haben die Möglichkeit, die Bereitstellung mithilfe von Containern in On-Premises- und Edge-Umgebungen vorzunehmen.

Schnellere ML-Operationen (MLOps)

In einem modernen Data Warehouse sind mit der Kombination aus Azure Databricks und Azure Machine Learning Services schnellere Machine-Learning-Operationen möglich. Azure MLS bietet Ihnen einen End-to-End-Arbeitsbereich, in dem Sie Daten aus verschiedenen Quellen mit Azure Synapse Analytics verbinden und Datenmodelle vorbereiten und trainieren, für Consumer wie Power BI bereitstellen und anschließend überwachen und erneut trainieren können, um ihre Genauigkeit zu verbessern.

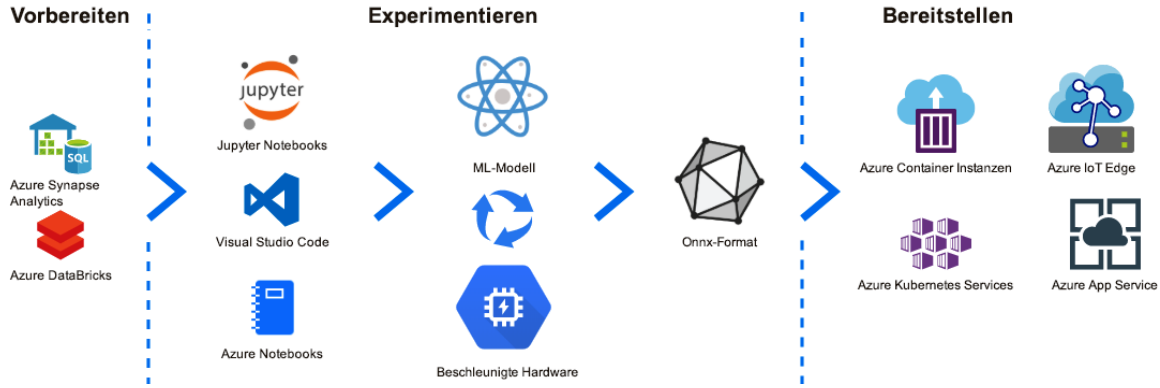

Abbildung 3.34: Vorbereiten, Experimentieren und Bereitstellen in Azure Machine Learning Services

Mit Azure MLS können Sie Azure Databricks verwenden, um die Daten für Ihre Experimente vorzubereiten. Sie können dann zum Erstellen Ihrer Experimente entweder Jupyter Notebooks oder Visual Studio Code verwenden. Alternativ können Sie auch die integrierte Azure Notebooks-Funktion von MLS verwenden. Anschließend führen Sie Ihre Experimente aus, um Ihr ML-Modell zu trainieren und zu testen. Dabei nutzen Sie Computer, um komplexe Data-Science-Algorithmen auszuführen. Es wird dann ein ML-Modell in einem hoch portierbaren ONNX-Format erstellt, das leicht in einem Container wie Azure Container Instance bereitgestellt werden kann. Auch eine Ausführung auf AKS (Azure Kubernetes Services) oder sogar auf Edgegeräten, die Docker unterstützen, ist möglich.

Die Verwendung von Azure Databricks als Computecluster von Azure Machine Learning Services wird in diesem Buch nicht behandelt. Diese Kombination bringt jedoch Vorteile mit sich. Wenn Sie Azure Databricks bereits verwenden, um Echtzeit-Analytics in Ihrem modernen Data Warehouse abzuleiten, könnten Sie sich auch überlegen, Databricks zur Ausführung Ihrer ML-Experimente in Azure MLS zu nutzen.

Weitere Informationen hierzu finden Sie hier.

Quick-Start-Leitfaden (Machine Learning)

Nachdem wir uns die Möglichkeiten angesehen haben, die Microsoft im Machine-Learning-Bereich bietet, folgt nun ein einfacher Quick-Start-Leitfaden zur Verwendung von Azure Machine Learning Services.

Für dieses konkrete Beispiel verwenden wir den Datensatz **`Credit Card Fraud Detection`** (Erkennung von Kreditkartenbetrug), der auf Kaggle öffentlich verfügbar ist. In einer modernen Data-Warehouse-Pipeline erhalten Sie den Datensatz idealerweise aus Ihrem Warehouse.

Führen Sie die folgenden Schritte aus, um Ihr ML-Modell zu entwickeln:

1. Stellen Sie zunächst sicher, dass Ihr Azure-Portal geöffnet ist, und erstellen Sie dann einen Machine Learning Service-Arbeitsbereich in Ihrer Azure-Ressourcengruppe. Suchen Sie in der Suchleiste nach **Machine Learning Service Workspace** (Machine Learning Service-Arbeitsbereich), und klicken Sie auf **Create** (Erstellen).

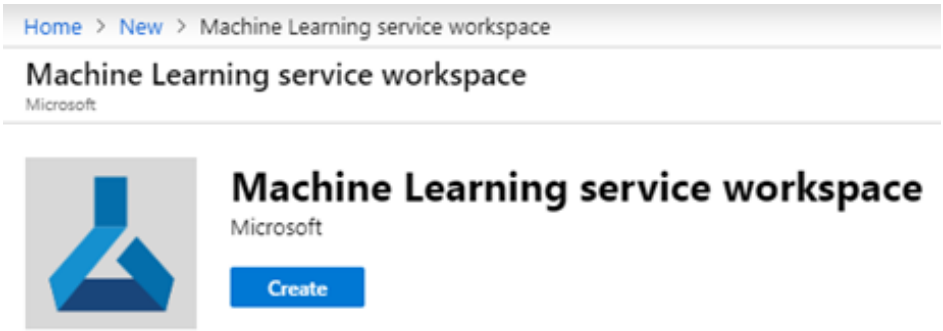

Abbildung 3.35: Erstellen eines ML Service-Arbeitsbereichs

2. Geben Sie den Namen des Arbeitsbereichs, das Abonnement, die Ressourcengruppe und den Standort ein. Klicken Sie nach der Überprüfung auf **Create** (Erstellen). Warten Sie, bis die Ressource erstellt wurde. Dies kann einige Minuten dauern.
3. Wenn der Arbeitsbereich erstellt wurde, navigieren Sie hierhin. Azure ML ist ein zentrales universelles Dashboard für Ihre ML-Experimente.
4. Beginnen wir mit dem Erstellen Ihres ersten Computeclusters. Wechseln Sie zur Registerkarte **Compute**.

Abbildung 3.36: Registerkarte „Compute“

5. Klicken Sie auf die Schaltfläche **+ Add** (Hinzufügen). Geben Sie den Computernamen ein. Wählen Sie **Machine Learning Compute** als **Compute type** (Berechnungstyp) aus. Die **Region** ist anhand der Arbeitsbereichsregion voreingestellt. Wählen Sie eine Größe für Ihren Computer aus. Für den Einstieg können Sie die Option **Standard_DS12_v2** verwenden. Wählen Sie **Dedicated** (Dediziert) für die VM aus. Legen Sie für **Minimum number of nodes** (Mindestanzahl von Knoten) **1** und für **Maximum number of nodes** (Maximale Knotenanzahl) **6** fest:

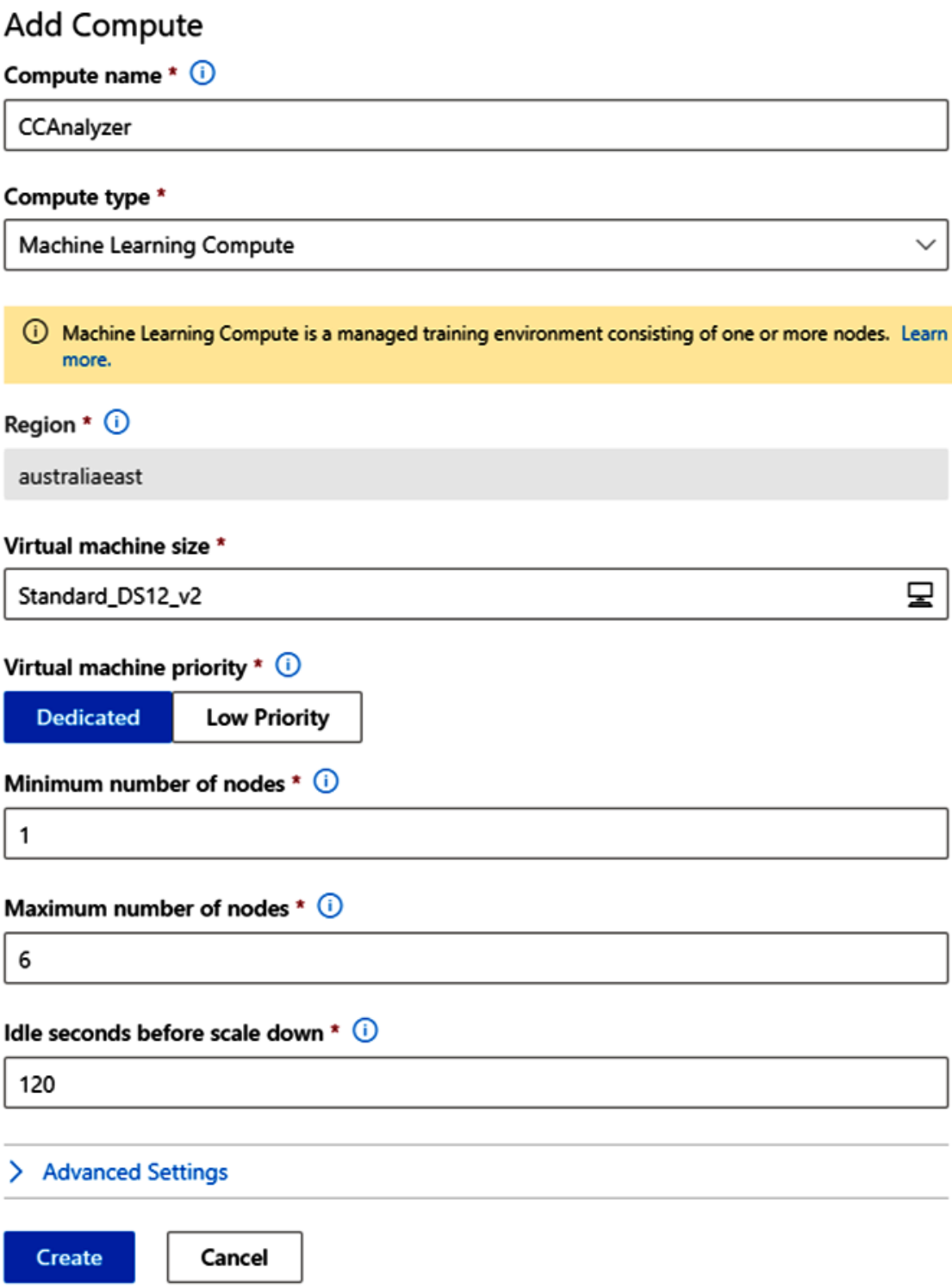

Abbildung 3.37: Hinzufügen von Compute-Details

6. Klicken Sie auf **Create** (Erstellen). Warten Sie, bis der Cluster bereitgestellt ist. Wenn der Bereitstellungsstatus in **Succeeded** (Erfolgreich) geändert wurde, können Sie die Knoten verwenden, um Ihre Experimente auszuführen.

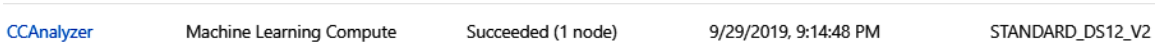

Abbildung 3.38: Bereitstellungsstatus von Machine Learning Compute

7. Besuchen Sie kaggle, und laden Sie die **CSV**-Datei herunter. (Klicken Sie auf „Download“ (Herunterladen), und entpacken Sie den Ordner.)

8. Kehren Sie zu Azure ML zurück, und klicken Sie auf die Registerkarte „Datasets“ im Navigationsbereich. Klicken Sie auf **+ Create dataset** (Dataset erstellen), und wählen Sie **From local files** (Aus lokalen Dateien) aus. Laden Sie die **CSV**-Datei hoch, die Sie von Kaggle heruntergeladen haben. Verwenden Sie einen kreativen Namen für Ihr Dataset, und legen Sie als Typ **Tabular** (Tabellarisch) fest.

   Unter **Advanced Settings** (Erweiterte Einstellungen) werden Sie sehen, dass die Datei in den ML-Arbeitsbereichsspeicher hochgeladen wird. Sie haben die Möglichkeit, den Speicherort für den Datenspeicher zu ändern.

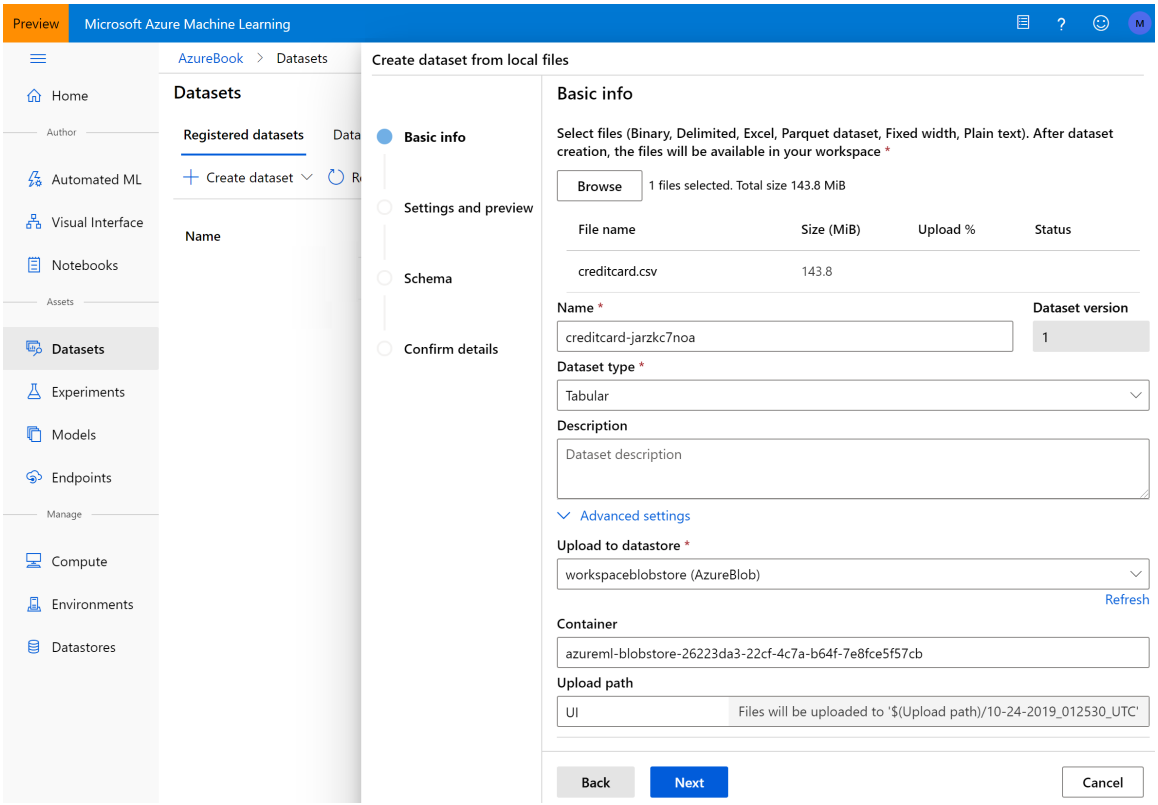

Abbildung 3.39: Erstellen eines Datasets aus einer CSV-Datei

9. Klicken Sie auf „Done“ (Fertig). Warten Sie, bis die Datei hochgeladen wurde.
10. Wechseln Sie nun zur Registerkarte **Automated ML** (Automatisiertes ML) in der Navigationsleiste links. Erstellen Sie ein neues Experiment, und klicken Sie dazu auf **+ Create experiment** (Experiment erstellen).
11. Legen Sie den Namen des Experiments fest. Wählen Sie eine Trainingscompute-Option aus, indem Sie den Computecluster auswählen, den Sie zuvor in *Schritt* 5 erstellt haben. Wählen Sie als Dataset die **`CSV`**-Datei aus, die Sie in den Datenspeicher hochgeladen haben (die **`CSV`**-Datei **`creditcard`** von Kaggle).

    AML zeigt dann eine Vorschau des Datensatzes an. Viele Datenfelder werden aus Sicherheitsgründen anonymisiert. Dies beinhaltet eine PCA-Dimensionsreduktion.

Abbildung 3.40: Dataset-Vorschau, wie von AML angezeigt

12. Wählen Sie bei der Vorhersageaufgabe **Classification** (Klassifizierung) aus, um Transaktionen als betrügerisch oder nicht betrügerisch zu klassifizieren.
13. Wählen Sie für **`Target column`** (Zielspalte) die Option **`Class`** (Klasse) aus, da dies das Ergebnis der Klassifizierung der Transaktion als betrügerisch oder nicht betrügerisch ist.

14. Fügen Sie unter „Advanced Settings“ (Erweiterte Einstellungen) die folgenden Werte hinzu:

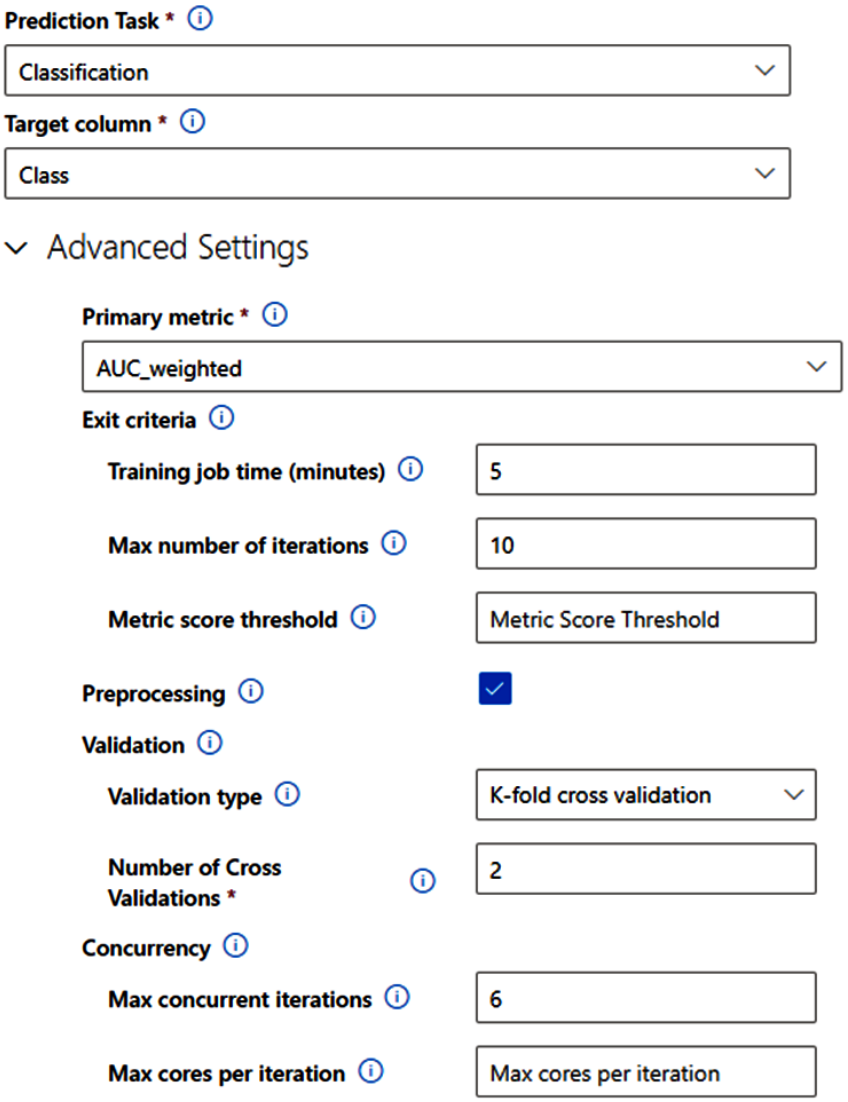

Abbildung 3.41: Hinzufügen von Details zum Klassifizierungsalgorithmus

> Hinweis
>
> **AUC_weighted** ist das arithmetische Mittel der Ergebnisse für jede Klasse, gewichtet nach der Anzahl der „True“-Instanzen in jeder Klasse.
>
> **K-fache Kreuzvalidierung** ist ein Verfahren zur Neuberechnung, bei dem das Dataset in K Gruppen aufteilt wird, um Machine-Learning-Modelle anhand einer begrenzten Datenprobe zu evaluieren.

Sie können die Werte der oben genannten Parameter erhöhen, wenn Sie mehr gleichzeitige Tests ausführen oder die Trainingszeiten erhöhen möchten. Dies wird sich auf den Computeverbrauch und die Ausführungszeiten auswirken. Sie können auch Optionen festlegen, um Algorithmen zu blockieren, die nicht in Ihr Szenario passen.

15. Klicken Sie auf **Start**.

    Daraufhin wird Azure MLS ausgeführt und testet verschiedene Algorithmen, um das gewünschte Ergebnis zu erzielen. Entsprechend der in den vorherigen Schritten festgelegten Konfiguration können zum Trainieren Ihres Modells gleichzeitige und simultane Algorithmen ausgeführt werden. Die Ausführung wird einige Zeit in Anspruch nehmen.

    Nach Abschluss des Vorgangs wird die Formel angezeigt, die die höchste Punktzahl erreicht hat. Mit dieser Methode können Sie verschiedene Formeln testen, um herauszufinden, ob eine Transaktion betrügerisch ist oder nicht.

    Sie können dieses Modell zunächst in einer **Azure Container Instance** (**ACI**) und später auf anderen von Containern unterstützten Geräten wie z. B. IoT-Geräten bereitstellen.

    Die Ergebnisse können nie hundertprozentig präzise sein. Sobald wir ein Modell entwickelt haben, können wir jedoch neue Transaktionen filtern und ermitteln, ob sie betrügerisch sein könnten.

    Die folgende Abbildung zeigt, dass Sie mit **VotingEnsemble** bei der Vorhersage, ob eine Transaktion betrügerisch ist, eine Genauigkeit von 98 % erzielen können.

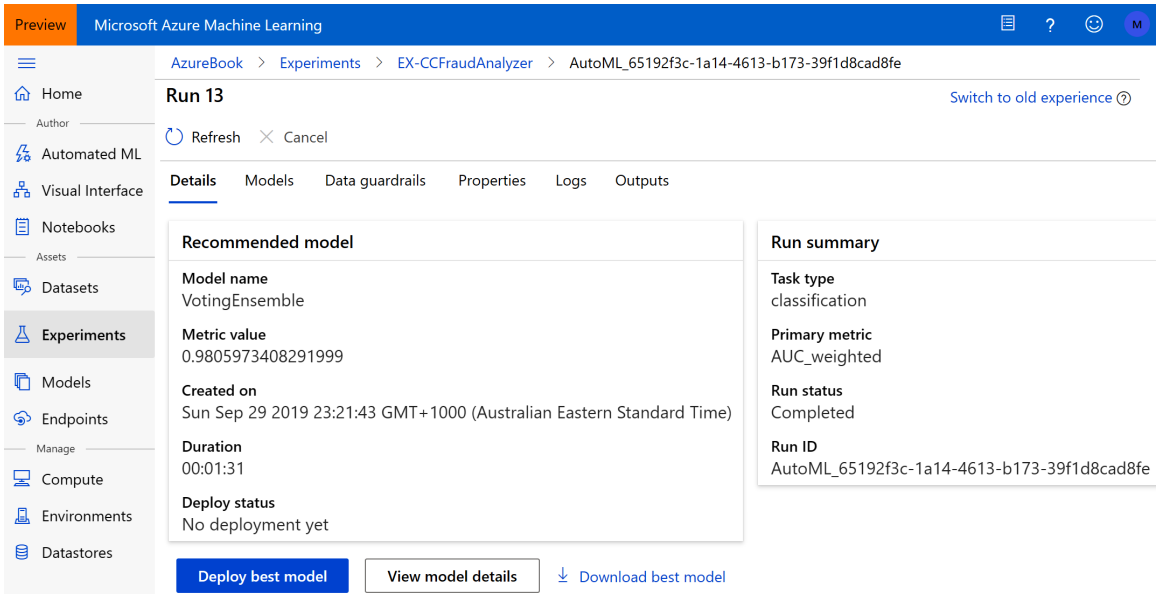

Abbildung 3.42: Iterationsdiagramm für das Modell zum Kreditkartenbetrug

Diese Übung kann Ihnen den Einstieg in Machine Learning auf Azure erleichtern. Sie können ML-Algorithmen ausführen und innerhalb weniger Minuten einen komplexen Datensatz ohne Codierung klassifizieren. Sie können Ihre Experimente weiter verfeinern, indem Sie die Anzahl der Iterationen erhöhen oder die Trainingszeit verlängern.

Wenn Sie mehr über die Bedeutung der Diagramme erfahren möchten, finden Sie Informationen hierzu in dieser Ressource.

Nach diesen Übungen können Sie die Ressourcen aus dem Azure-Portal löschen, damit Ihnen Ihr Abonnement nicht durchgehend berechnet wird.

Zusammenfassung

In diesem Kapitel haben wir die semantische Modellierung mit Azure Analysis Services behandelt. Mit AAS haben wir eine Brücke zwischen der Datenquelle und der Datenvisualisierung geschaffen. Wir haben die Verwendung von Power BI Desktop zum Erstellen von Berichten erkundet. Mit Power BI können Sie umfangreiche, aussagekräftige Diagramme erstellen, aus denen Business Insights abgeleitet werden können. Anschließend haben wir den Bericht veröffentlicht, um in verschiedenen Medien daran zusammenzuarbeiten.

Wir haben gesehen, dass viele Tools und Technologien zur Implementierung von ML und KI in Azure verfügbar sind. Wir haben Machine Learning Services erkundet, über die Features und Vorteile hiervon gesprochen und Machine Learning auf Azure ausgeführt, um die Erkennung von Kreditkartenbetrug vorherzusagen. Im nächsten Kapitel werden wir über die neuen Features und Funktionen sprechen, die im modernen Data Warehouse hinzugekommen sind.

4

Einführung in Azure Synapse Analytics

In den vorherigen Kapiteln haben Sie mehr über die Muster eines modernen Data Warehouse erfahren und gelernt, mithilfe von Azure-Diensten Ihr eigenes durchgängiges modernes Data Warehouse zu implementieren. In diesem Kapitel finden Sie eine Vorschau auf Azure Synapse Analytics, eine spannende neue Suite von Funktionen, die dem Microsoft Data Warehouse hinzugefügt wurde.

Was ist Azure Synapse Analytics?

Azure Synapse ist ein unbegrenzter Analytics-Dienst, der Enterprise Data Warehousing und Big Data Analytics vereint. Sie haben damit die Möglichkeit, Datenabfragen den Anforderungen Ihres Unternehmens entsprechend durchzuführen. Sie können serverlose bedarfsgesteuerte oder bereitgestellte Ressourcen im gewünschten Umfang nutzen. Azure Synapse vereint beide Komponenten in einer einheitlichen Benutzeroberfläche, auf der Sie Daten für unmittelbare Business-Intelligence- und Machine-Learning-Anforderungen erfassen, vorbereiten, verwalten und bereitstellen können.

Azure Synapse ist die Weiterentwicklung von Azure SQL Data Warehouse. Microsoft hat die branchenführenden Data-Warehouse-Lösungen auf ein neues Level in puncto Leistung und Funktionalität gebracht. Unternehmen können ihre vorhandenen Data-Warehouse-Workloads in der Produktion mit Azure Synapse weiter ausführen und profitieren automatisch von den neuen Funktionen. Sie können ihre Daten viel schneller, produktiver und sicherer nutzbar machen, da Insights aus allen Datenquellen, Data Warehouses und Big Data Analytics-Systemen zusammengeführt werden.

Mit Azure Synapse können verschiedenste Datenexperten problemlos zusammenarbeiten und ihre wichtigsten Daten verwalten und analysieren – all das mit demselben Dienst. Von der Apache Spark-Integration in der leistungsstarken und vertrauenswürdigen SQL-Engine bis hin zur codefreien Datenintegration und -verwaltung bietet Azure Synapse für jeden Datenexperten Vorteile.

Die folgende Abbildung gibt eine Übersicht der neuen Dienste und Features in Azure Synapse Analytics:

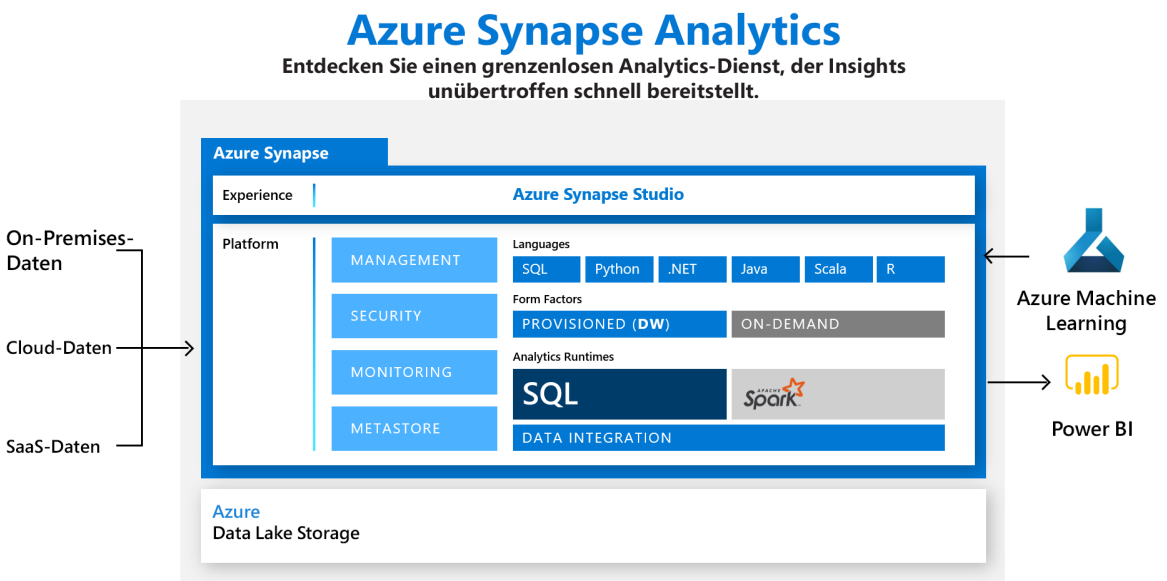

Abbildung 4.1: Azure Synapse Analytics

Warum brauchen wir Azure Synapse Analytics?

Heutzutage sind die Unternehmen mit der Herausforderung konfrontiert, zwei Arten von Analytics-Systemen verwalten zu müssen:

- **Data Warehouse**, das wichtige Insights über das Unternehmen bietet
- **Data Lakes**, die mithilfe von verschiedenen Analytics-Methoden aussagekräftige Insights über Kunden, Produkte, Mitarbeiter und Prozesse liefern

Beide Analytics-Systeme sind für die Unternehmen von entscheidender Bedeutung und laufen unabhängig voneinander. Dies kann zu uninformierten Entscheidungen führen. Gleichzeitig müssen Unternehmen Insights aus allen ihren Daten gewinnen, um wettbewerbsfähig zu bleiben und Prozessinnovationen einzuführen, mit denen bessere Ergebnisse erzielt werden können.

Wenn Kunden eine eigene End-to-End-Datenpipeline entwickeln möchten, sind folgende Schritte erforderlich:

1. Datenerfassung aus verschiedenen Datenquellen
2. Laden all dieser Datenquellen in einen Data Lake zur weiteren Verarbeitung
3. Datenbereinigung für eine Reihe unterschiedlicher Datenstrukturen und -typen
4. Vorbereiten, Transformieren und Modellieren der Daten
5. Bereitstellung der bereinigten Daten für Tausende von Anwendern über BI-Tools und -Anwendungen

Bislang wurde für jeden dieser Schritte ein anderes Tool benötigt. Angesichts des enormen Angebots an verschiedenen Tools, Diensten und Anwendungen kann es eine Herausforderung darstellen, die richtige Wahl zu treffen.

Es gibt zahlreiche Dienste, die Daten erfassen, laden, vorbereiten und bereitstellen. Je nachdem, für welche Programmiersprache sich der Entwickler entschieden hat, gibt es auch mehrere Dienste für die Datenbereinigung. Manche Entwickler bevorzugen Spark, andere möchten SQL verwenden, wieder andere bevorzugen codefreie Umgebungen für die Datentransformation.

Auch nach Auswahl der richtigen Tools muss eine steile Lernkurve bewältigt werden, um mit den Tools zurechtzukommen. Hinzu kommen die logistischen Schwierigkeiten der Unterhaltung einer Datenpipeline über verschiedene Plattformen und Sprachen hinweg. Angesichts so vieler Probleme kann es ein schwieriges Unterfangen sein, eine Cloud-Analytics-Plattform zu implementieren und zu unterhalten.

Das Muster des modernen Data Warehouse

Wie in Kapitel 2 dargelegt, bietet Azure mit den folgenden Diensten eines der besten und verständlichsten Muster eines modernen Data Warehouse:

- **Azure Data Factory**, ein Dienst, der Ihnen bei der Erfassung von Daten aus Datenquellen in **Azure Data Lake Storage** und bei der Orchestrierung der Datenpipeline hilft.
- **Azure Databricks** und **HDInsight**, die Ihnen Spark-Funktionen bieten, sodass Sie Python, Scala, .NET, Java und R verwenden können, um die Daten vorzubereiten und zu analysieren und Machine-Learning-Modelle zur Datenverarbeitung zu entwickeln.
- **Azure Synapse Analytics**, ein Dienst, der hauptsächlich für Analytics-Zwecke verwendet wird.
- Weitere Dienste, die stark optimiert sind, um gleichzeitige Abfragen für **Power BI**-Dashboards oder Anwendungen für die Visualisierung bereitzustellen.

In diesem Muster eines modernen Data Warehouse werden alle diese Dienste nahtlos vereint, um die End-to-End-Pipeline zu erstellen.

Kundenherausforderungen

Man könnte meinen, die größte Herausforderung für ein effizientes Data Warehouse sei es, herauszufinden, wie die Pipeline entwickelt werden muss, um die Daten zu integrieren, oder das Warehouse zu optimieren, um eine bessere Leistung zu erzielen. Eine Kundenstudie von Microsoft ergab jedoch, dass die Verwaltung verschiedener Funktionalitäten, die Überwachung Hunderter Pipelines über verschiedene Compute-Engines hinweg, die Sicherung verschiedener Ressourcen (Compute, Speicher, Artefakte) und die Bereitstellung von Code ohne fehlerverursachende Änderungen die größten Herausforderungen für die Kunden darstellten. Inmitten von organisatorischen Silos, Datensilos und Tooling-Silos wird es fast unmöglich, eine Cloud Analytics-Plattform zu implementieren und zu unterhalten.

Stellen Sie sich beispielsweise vor, Ihr Unternehmen benötigt ein einziges Sicherheitsmodell, um alle seine Dienste zu schützen und damit die neuesten internen Compliance-Richtlinien zu erfüllen. Eine solche Aufgabe mag zunächst einfach klingen, ist aber tatsächlich ziemlich komplex. Sie müssen schnell herausfinden, welches dieses „einzelne Sicherheitsmodell“ ist, und sich dann über das Bereitstellungsmodell für diese Dienste klar werden. Sie müssen sich damit befassen, wie Sie Hochverfügbarkeit und Notfallwiederherstellung für jeden dieser Dienste implementieren können. Und schließlich müssen Sie sich um alle zugehörigen Aufgaben in Zusammenhang mit der Lebenszyklusverwaltung kümmern, einschließlich der Überwachung dieser Dienste, um ihre ordnungsgemäße Ausführung sicherzustellen. Alle diese Dienste zusammenzubringen, ist kein einfaches Unterfangen. In der Vergangenheit erforderte dies eine komplexe Planung.

Azure Synapse Analytics ist die Lösung

Azure Synapse Analytics ist die Lösung für diese Probleme. Der Dienst vereinfacht das gesamte Muster des modernen Data Warehouse. Kunden haben damit die Möglichkeit, End-to-End-Analytics-Lösungen in einer einheitlichen Benutzeroberfläche zu entwickeln.

Azure Synapse Analytics ist eine Weiterentwicklung von Azure SQL Data Warehouse, das mehr Funktionen für Datenexperten bietet und serverlose On-Demand-Abfragen und Machine Learning-Unterstützung hinzufügt. Zu den weiteren Features von Azure Synapse Analytics gehören die native Einbettung von Spark, die Bereitstellung kollaborativer Notebooks und die Möglichkeit der Datenintegration – alles in einem einzigen Dienst. Über verschiedene Module werden verschiedene Sprachen (wie C#, SQL, Scala, Python) unterstützt.

Zu den wichtigsten Funktionen von Azure Synapse Analytics zählen:

- SQL Analytics mit Pools und On-Demand (serverlos)
- Spark mit vollständigem Support für C#, SQL, Scala und Python
- Data Flow mit codefreier Big Data Transformation-Erfahrung
- Datenintegration und Orchestrierung zur Integration Ihrer Daten und zum Operationalisieren Ihrer Codeentwicklung
- Azure Synapse Studio erlaubt Ihnen den Zugriff auf sämtliche oben genannte Fähigkeiten innerhalb einer einheitlichen Web-UI

Das Azure Synapse-Studio bietet eine einheitliche Benutzeroberfläche für Datenvorbereitung, Datenverwaltung, Data Warehousing, Big Data Analytics und KI-Aufgaben. Folgende Features sind verfügbar:

- Codefreie visuelle Umgebungen für die Verwaltung von Datenpipelines
- Automatisierte Abfrageoptimierung
- Funktionalität zum Entwickeln von Proofs of Concept innerhalb von Minuten
- Serverlose On-Demand-Abfragen
- Option für einen sicheren Zugriff auf Datasets und Möglichkeit, mit Power BI-Dashboards in wenigen Minuten zu entwickeln – all dies mit demselben Analytics-Dienst

Azure Synapse Analytics kann blitzschnell Insights aus allen Ihren Daten in Data Warehouse und Big Data Analytics-Systemen gewinnen und bereitstellen. Datenexperten können damit die vertraute Programmiersprache SQL verwenden, um relationale und nicht relationale Datenbanken in einer Größenordnung von Petabyte abzufragen. Mithilfe erweiterter Funktionen wie intelligentem Workload-Management, Workload-Isolierung und grenzenloser Nebenläufigkeit kann die Leistung aller Abfragen für unternehmenskritische Workloads optimiert werden.

Im folgenden Diagramm sehen Sie die Architektur von Azure Synapse Analytics:

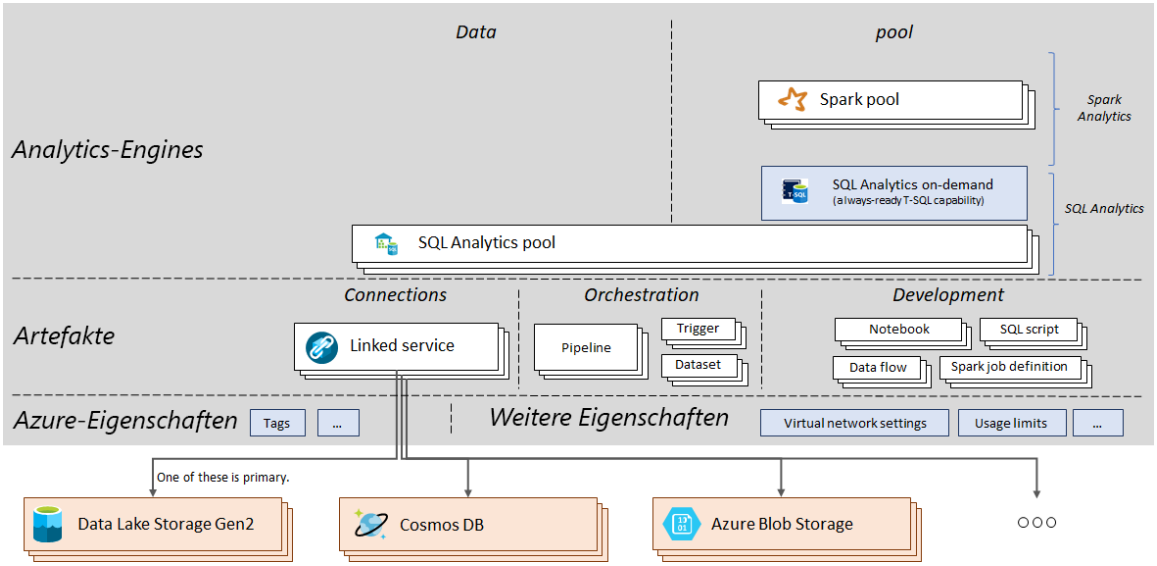

Abbildung 4.2: Architektur von Azure Synapse Analytics

Azure Synapse Analytics ermöglicht Kunden eine problemlose Durchführung von Business-Intelligence-Projekten und Machine Learning. Azure Synapse Analytics ist tief mit Power BI und Azure Machine Learning integriert, um deutlich mehr Insights aus allen Ihren Daten zu ermöglichen und Machine-Learning-Modelle auf alle Ihre intelligenten Apps anzuwenden. Anwender können die Projektentwicklungszeit für BI- und Machine-Learning-Projekte mit diesem unbegrenzten Analytics-Dienst erheblich reduzieren, der die nahtlose Anwendung von Intelligenz auf Ihre wichtigsten Daten möglich macht – von Dynamics 365 und Office 365 bis hin zu Software-as-a-Service-Implementierungen (SaaS), die die Open Data Initiative unterstützen. Mit nur wenigen Klicks kann zudem eine Datenfreigabe erfolgen.

Dies alles wird in einer einzelnen Umgebung bereitgestellt, die Abfrage-Editoren und Notebooks für die Freigabe und gemeinsame Bearbeitung von Daten sowie Ressourcen und Code für SQL- und Spark-Entwickler umfasst.

Im Prinzip erledigt Azure Synapse Analytics alles.

Ausführliche Informationen zu Azure Synapse Analytics

Nachdem Sie nun wissen, warum Azure Synapse Analytics entwickelt wurde, können Sie sich die Azure Synapse Analytics-Dienste etwas genauer ansehen.

Azure Synapse Analytics ist ein vollständig verwalteter, integrierter Daten-Analytics-Dienst, der Data Warehousing, Datenintegration und Big Data-Verarbeitung sowie schnellere Insights bietet.

Der Vorteil eines einzelnen integrierten Datendiensts besteht darin, dass die Unternehmen damit schneller Business Intelligence, künstliche Intelligenz, Machine Learning, das Internet der Dinge und intelligente Anwendungen bereitstellen können. Die folgende Abbildung zeigt den einzelnen integrierten Datendienst von Azure:

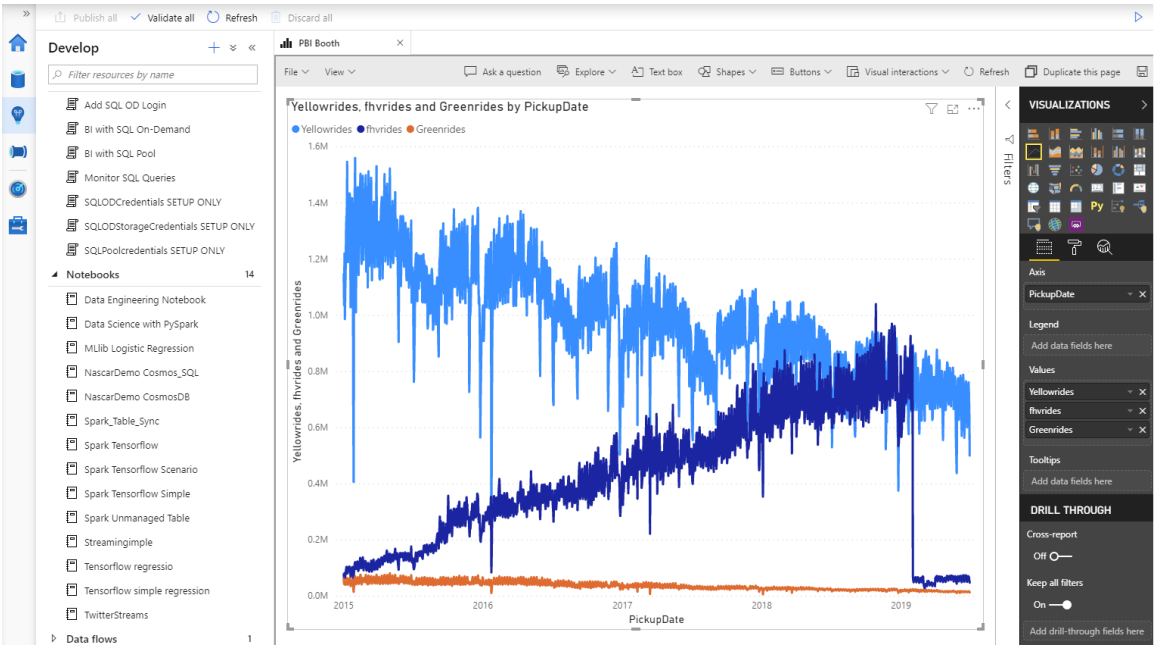

Abbildung 4.3: Analysieren und Visualisieren von Daten in Azure Synapse Analytics

Um sich ein klares Bild von den Vorteilen von Azure Synapse zu verschaffen, sollten Sie sich zunächst die zentralen Dienste ansehen.

Azure Synapse Analytics-Arbeitsbereiche

Im Zentrum von Azure Synapse steht der Arbeitsbereich. Ein Arbeitsbereich ist die Ressource der obersten Ebene, die Ihre Analytics-Lösung im Data Warehouse umfasst. Der Azure Synapse-Arbeitsbereich unterstützt sowohl die Verarbeitung von relationalen Daten als auch von Big Data. Die Umgebung für Zusammenarbeit ist ideal für die Datenfreigabe und das gemeinsame Arbeiten der Dateningenieure und Datenwissenschaftler an ihren Analytics-Lösungen, wie in der folgenden Abbildung dargestellt:

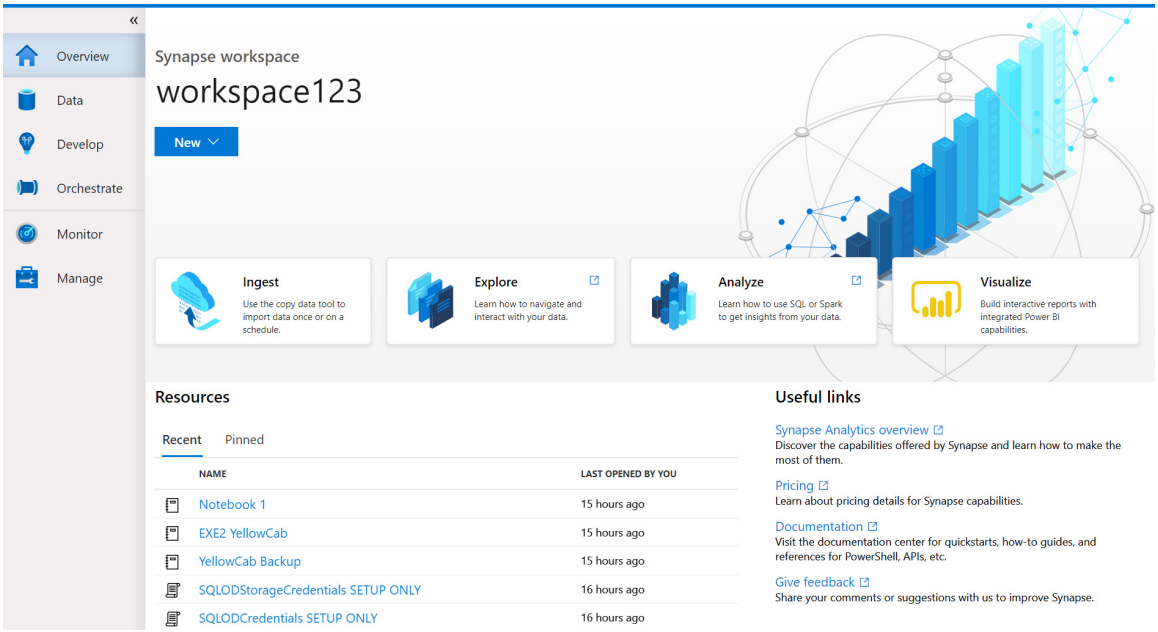

Abbildung 4.4: Azure Synapse Analytics-Studio und -Arbeitsbereich

Wichtige Features

Der Azure Synapse Analytics-Arbeitsbereich bietet folgende Features:

- Schnelles, hochelastisches, sicheres Data Warehouse
- Möglichkeit zur Ausführung gleichzeitiger T-SQL-Abfragen über SQL-Pools für Petabyte von Daten, um BI-Tools und -Anwendungen zu unterstützen
- SQL On-Demand bietet serverlose SQL-Abfragen zur bequemen Erkundung und Analyse von Daten in Azure Data Lake Storage, ohne eine Infrastruktur einzurichten oder zu unterhalten
- Erfüllt das gesamte Spektrum der Analytics-Anforderungen – von Data Engineering bis Data Science – mit einer Vielzahl von Sprachen wie Python, Scala, C# und Spark SQL
- Spark-Pools, die die komplexe Einrichtung und Unterhaltung von Clustern erleichtern und die Entwicklung von Spark-Anwendungen sowie die Nutzung von Spark-Notebooks vereinfachen

- Tiefe Integration von Spark und SQL, sodass Dateningenieure Daten in Spark vorbereiten, die verarbeiteten Ergebnisse in den SQL-Pool schreiben und jede Kombination von Spark mit SQL für Data Engineering und Datenanalyse verwenden können, mit integrierter Unterstützung für Azure Machine Learning
- Hochgradig skalierbare Möglichkeit zur Integration von Hybriddaten, dadurch beschleunigte Datenerfassung und Operationalisierung über automatisierte Datenpipelines
- Reibungsloser integrierter Dienst mit einheitlicher Sicherheit, Bereitstellung, Überwachung und Abrechnung

Azure Synapse Analytics-Studio

Das Azure Synapse-Studio verfügt über eine anwenderfreundliche, webbasierte Oberfläche, die einen End-to-End-Arbeitsbereich und eine durchgängige Entwicklungsumgebung bietet.

Die folgende Abbildung zeigt ein Beispiel einer modernen Datenpipeline mit Azure Synapse. In diesem Beispiel beginnt die Datenerfassung aus einer BLOB-Speicherquelle in einen Azure Data Lake Storage im Azure Synapse Analytics-Arbeitsbereich. Mit dem Spark-Pool können Sie mehrere Datenquellen aus dem Azure Data Lake und der SQL Database lesen und jegliche Transformation und Datenbereinigung durchführen. Schließlich können Sie die zusammengestellten Ergebnisse in den SQL-Pool schreiben, um BI-Tools und -Anwendungen zu unterstützen.

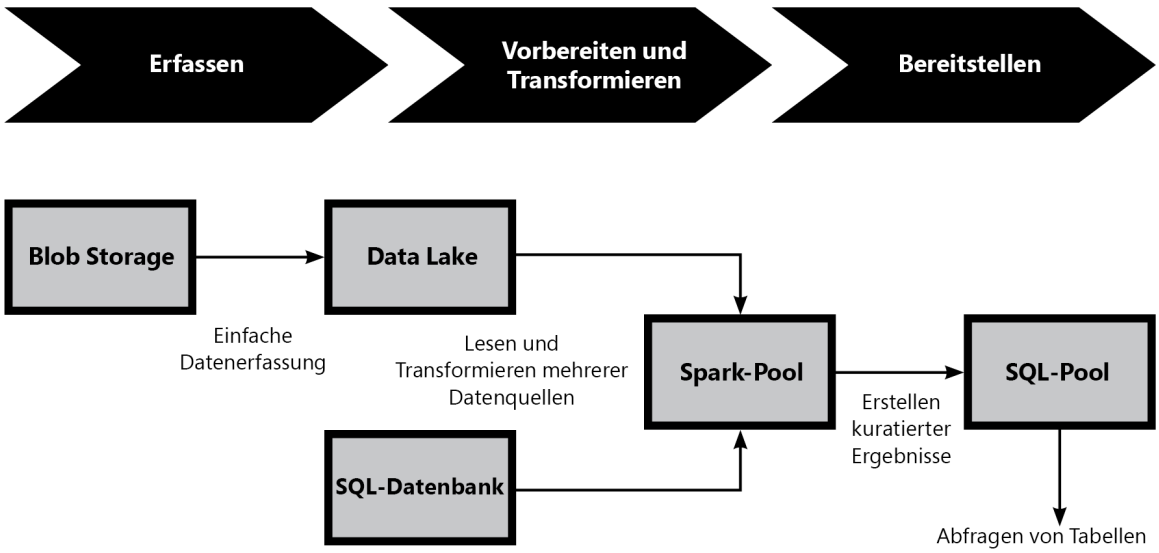

Abbildung 4.5: Moderne Datenpipeline mit Azure Synapse Analytics

In den folgenden Abschnitten sind die Funktionen, wichtige Features, die Plattformdetails und die Anwenderdienste aufgeführt:

Funktionen

- Schnelles, hochelastisches, sicheres Data Warehouse mit branchenführender Leistung und Sicherheit
- Möglichkeit, Azure Data Lake Storage und Data Warehouse mit vertrauter T-SQL-Syntax in (serverlosen) SQL-Abfragen von SQL On-Demand zu erkunden
- Apache Spark mit Azure Machine Learning integriert
- Integration von Hybriddaten zur Beschleunigung der Datenerfassung und der Operationalisierung des Analytics-Prozesses (Erfassung, Vorbereitung, Transformation und Bereitstellung)
- Generierung und Bereitstellung von Geschäftsberichten mit Power BI-Integration

Wichtige Features

- Direkte Erkundung von Daten in Azure Data Lake Storage, Data Warehouse sowie externen Verbindungen mit dem Arbeitsbereich mit Azure Synapse Analytics-Studio
- Erstellen und Operationalisieren von Pipelines für die Datenerfassung und -orchestrierung
- Schreiben von Code mit Notebooks und T-SQL-Abfrage-Editoren
- Codefreies Datentransformationstool, wenn Sie es vorziehen, keinen eigenen Code zu schreiben
- Überwachen, Sichern und Verwalten Ihrer Arbeitsbereiche, ohne die Umgebung zu verlassen
- Webbasierte Entwicklungsumgebung für die gesamten Analytics-Lösungen
- Möglichkeit, Azure Data Lake Storage, die Datenbanken und externen Verbindungen mit dem Arbeitsbereich zu erkunden

Plattform

- Unterstützt sowohl bereitgestellte als auch serverlose Computes. Beispiele für bereitgestellte Computes sind SQL Computes und Spark Computes. Mit diesen bereitgestellten Computes können Teams ihre Computeressourcen segmentieren, um Kosten und Nutzung zu kontrollieren und damit besser auf die Organisationsstruktur auszurichten. Serverlose Computes bieten den Teams andererseits die Möglichkeit, den Dienst nach Bedarf zu nutzen, ohne eine zugrunde liegende Infrastruktur bereitzustellen oder zu verwalten.
- Tiefe Integration von Spark und SQL-Engines

Apache Spark

Für Kunden, die mit Apache Spark arbeiten möchten, ist über Azure Databricks First-Party-Support verfügbar und die Verwaltung erfolgt vollständig durch Azure. Die neueste Version von Apache Spark wird den Anwendern automatisch zur Verfügung gestellt, zusammen mit allen Sicherheitspatches.

Wenn Sie Spark in Azure Synapse Analytics verwenden, wird es als SaaS bereitgestellt. So können Sie Spark beispielsweise verwenden, ohne eigene Dienste einzurichten oder zu verwalten, z. B. ein virtuelles Netzwerk. Azure Synapse Analytics kümmert sich um die zugrunde liegende Infrastruktur. So können Sie Spark sofort in Ihrer Azure Synapse Analytics-Umgebung verwenden.

SQL On-Demand

SQL On-Demand bietet serverlose SQL-Abfragen. Dies ermöglicht eine einfache Datenerkundung und -analyse in Azure Data Lake Storage, ohne eine Infrastruktur einzurichten oder zu unterhalten:

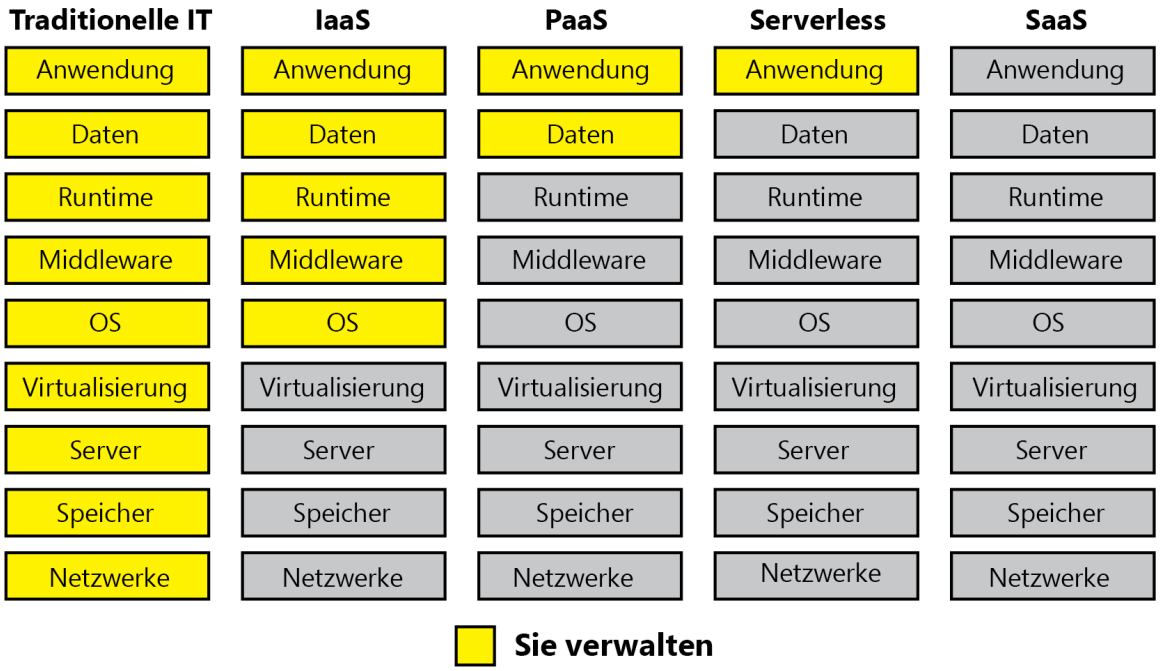

Abbildung 4.6: Vergleich zwischen verschiedenen Infrastrukturen

Wichtige Features

- Analysten können sich auf die Analyse der Daten konzentrieren, ohne sich um die Verwaltung einer Infrastruktur kümmern zu müssen.
- Kunden profitieren von dem einfachen und flexiblen Preismodell, da sie nur für die tatsächliche Nutzung zahlen.
- Verwendung der vertrauten Syntax von T-SQL und des besten SQL-Abfrageoptimierers auf dem Markt. Der SQL-Abfrageoptimierer ist die Intelligenz hinter dem Abfragemodul.
- Sie können Ihre Compute- und Speicherkapazitäten problemlos unabhängig voneinander skalieren, wenn Ihre Anforderungen steigen.
- Nahtlose Integration in SQL Analytics Pool und Spark über Metadatensynchronisierung und native Connectors.

Beispielsweise können Sie Azure Data Lake Storage mithilfe der vertrauten T-SQL-Syntax abfragen. Führen Sie dazu die folgenden Schritte aus:

1. Wählen Sie eine Datei aus Azure Data Lake Storage aus.
2. Klicken Sie auf die rechte Maustaste, und führen Sie Ihre (serverlose) SQL-Abfrage von SQL On-Demand unter Verwendung der T-SQL-Syntax aus (z. B. GROUP BY, ORDER BY usw.).

Datenintegration

Azure Synapse Analytics verwendet Azure Data Factory-(ADF-)Technologie, um Datenintegrationsfunktionen bereitzustellen. Die wichtigsten Features von ADF, die für die moderne Data-Warehouse-Pipeline unverzichtbar sind, stehen in Azure Synapse Analytics zur Verfügung. Für alle diese Features gilt ein gemeinsames Sicherheitsmodell, rollenbasierte Zugriffssteuerung (Role-Based Access Control, RBAC) im Azure Synapse Analytics-Arbeitsbereich.

Die folgende Abbildung zeigt ein Beispiel der Datenpipeline und der Aktivitäten von ADF, die direkt in die Azure Synapse Analytics-Umgebung integriert sind:

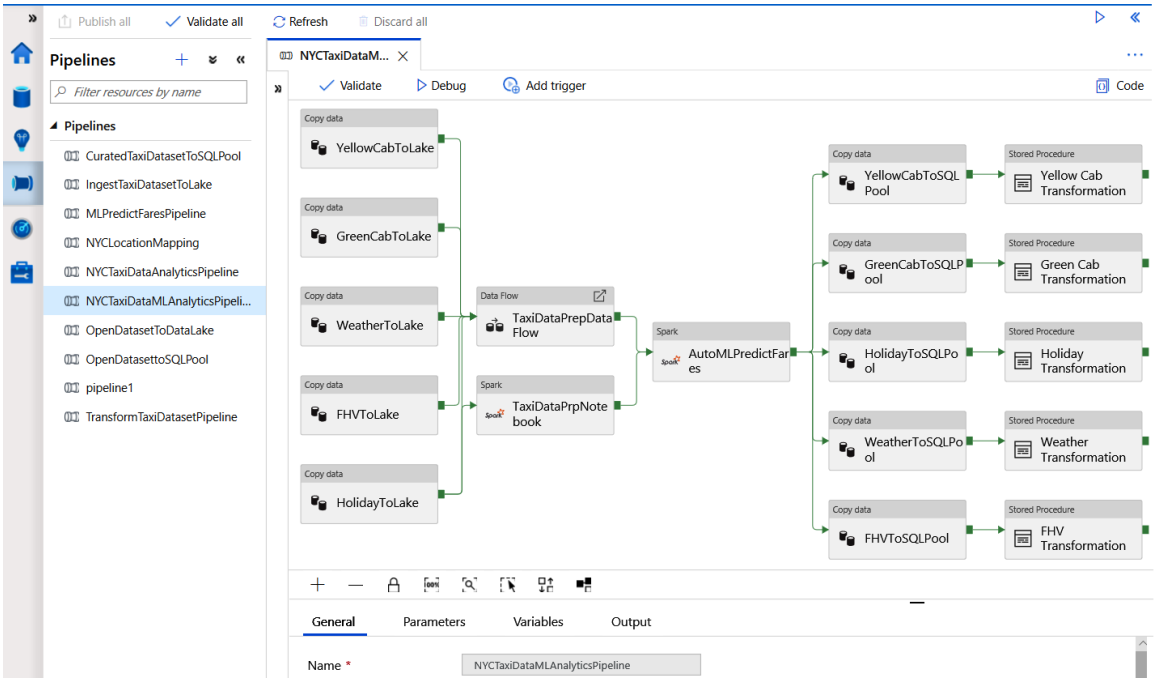

Abbildung 4.7: Datenpipelines in Azure Synapse Analytics

Wichtige Features

- Integrierte Plattformdienste für Verwaltung, Sicherheit, Überwachung und Metadatenverwaltung
- Native Integration von Spark und SQL. Einzelne Codezeile zum Lesen und Schreiben mit Spark aus/in SQL Analytics
- Möglichkeit, eine Spark-Tabelle zu erstellen und sofort mit SQL Analytics abzufragen, ohne ein Schema zu definieren
- Common Data Model-kompatibel
- „Schlüsselfreie" Umgebung. Mit Single Sign-On und AAD-Pass-Through ist kein Schlüssel und keine Anmeldung erforderlich, um mit ADLS/Datenbanken zu interagieren

Unterstützung mehrerer Sprachen

Azure Synapse Analytics unterstützt mehrere Sprachen, die für verschiedene Analytics-Workloads geeignet sind:

- SQL
- Python
- C#
- Java
- Scala
- R

Bevorstehende Änderungen

Wenn sich ein Produkt weiterentwickelt, ist es manchmal notwendig, das Produkt neu auszurichten und seinen Namen zu aktualisieren. In der folgenden Tabelle sehen Sie die Namensänderungen/-aktualisierungen, die künftig gelten.

Dienst/Funktion/Ressourcen	Vorher	Nachher
Service	Azure SQL Data Warehouse	Azure Synapse Analytics
Funktion		SQL-Analytics-Pool
Datenverarbeitungsressource	SQL Data Warehouse	SQL-Pool
Speicherressource		Datenbank

Abbildung 4.8: Bevorstehende Änderungen

Zusammenfassung

Azure Synapse Analytics ist eine bahnbrechende Weiterentwicklung von Azure SQL Data Warehouse. Mit diesem Dienst können Datenexperten End-to-End-Analytics-Lösungen in einer einheitlichen Benutzeroberfläche entwickeln.

Azure Synapse nutzt die besten Merkmale von Azure SQL Data Warehouse und modernisiert den Dienst durch die Bereitstellung von mehr Funktionen für SQL-Entwickler, das Hinzukommen einer serverlosen On-Demand-Abfrage, das Hinzufügen von Machine-Learning-Support, die native Einbettung von Spark, die Bereitstellung kollaborativer Notebooks und die Möglichkeit der Datenintegration – alles in einem einzigen Dienst.

Die im Muster des modernen Data Warehouse beschriebenen bestehenden Azure-Dienste (die Sie in den vorherigen Kapiteln kennengelernt haben) werden weiterhin erfolgreich genutzt werden können. Mit den neuen Features bietet Microsoft einen vereinfachten, optimierten Ansatz für Kunden, die eigene Analytics-Lösungen entwickeln möchten.

Im nächsten Kapitel sehen Sie anhand von realen Anwendungsfällen, wie alle diese Technologien integriert werden, um vollständige End-to-End-Data-Warehouse-Lösungen bereitzustellen, mit deren Hilfe die Entscheidungsträger im Unternehmen aussagekräftige Insights aus Echtzeitdaten gewinnen können.

5

Geschäftliche Anwendungsfälle

In den vorherigen Kapiteln haben Sie mehr über Cloud Scale Analytics und die Dienste von Microsoft Azure erfahren, mit deren Hilfe Unternehmen Insights gewinnen können. Darüber hinaus haben Sie die neuen Features und Funktionen kennengelernt, die dem modernen Data Warehouse hinzugefügt wurden. In diesem Kapitel werden Sie sich zwei geschäftliche Anwendungsfälle aus der Praxis ansehen, bei denen mithilfe von Microsoft Azure hochwertige Lösungen ermöglicht werden. Diese Anwendungsfälle sollen verdeutlichen, wie Echtzeitdaten in Azure analysiert werden können, um aussagekräftige Insights zu gewinnen und geschäftliche Entscheidungen zu treffen.

Die hier verwendeten Firmennamen sind fiktiv, und für die Implementierungsdemos verwenden wir Beispiel-Datasets. Die geschäftlichen Anwendungsfälle, die Herausforderungen und die tatsächlichen Probleme sind jedoch real. Sie veranschaulichen die Arten von Datenproblemen, denen Sie im Alltag begegnen könnten.

Im ersten Business Case geht es darum, einem Unternehmen zu helfen, nahezu in Echtzeit umsetzbare Insights aus seinen Daten zu gewinnen. Der zweite Business Case bezieht sich auf die Verwendung von Daten-Analytics auf Azure, um durch eine verbesserte Nutzung der Infrastruktur des verkehrsreichsten Flughafens Ägyptens betriebliche Probleme zu lösen und den Passagieren bessere Dienstleistungen zu bieten. In beiden Fällen werden wir zunächst kurz das Problem und die Herausforderungen ansprechen und uns dann ein potenzielles Lösungsdesign und die hierfür notwendigen Azure-Dienste ansehen.

Anwendungsfall 1: Customer Insights in Echtzeit mit Azure Synapse Analytics

Contoso ist ein großes multinationales Einzelhandelsunternehmen mit Geschäften in Australien, Neuseeland und Japan. Das Unternehmen verkauft Konsumgüter, Elektronik und Körperpflegeprodukte über seine Ladengeschäfte und digitalen Onlinekanäle (mobile Apps und Webanwendungen).

Contoso hat eine neue CEO (Chief Executive Officer) ernannt, die sich für Daten begeistert. Sie hat ein neues Daten-Analytics-Team aufgestellt und dieses beauftragt, nahezu in Echtzeit Customer Insights zu generieren und zu pflegen, um geschäftliche Entscheidungen zu unterstützen.

Das Problem

Wie viele andere Unternehmen versucht auch Contoso, sich als datengesteuertes Unternehmen neu zu erfinden. Vieles deutet darauf hin, dass dies die richtige Strategie ist. Damit Contoso jedoch erfolgreich sein kann, müssen viele Probleme gelöst werden. Einige dieser Probleme sind technischer Natur, andere kultureller und organisatorischer Art.

Die CEO von Contoso möchte, dass das neue Daten-Analytics-Team dem Unternehmen hilft, Fragen zu beantworten, die für operative Entscheidungen relevant sind. Das Führungsteam erhofft sich, mit Daten viele Fragen beantworten zu können. Um den Umfang dieses Projekts jedoch besser zu formulieren, hat sich das Datenteam von Contoso (in Abstimmung mit der CEO) auf folgende Definition der Problemstellung verständigt:

Wie kann Contoso seine Gewinne steigern?

Genauer gesagt wird dem Datenteam von Contoso eine 20-tägige Aufgabe gestellt: Es soll ein Pilotprogramm für Daten-Analytics durchführen, um Möglichkeiten aufzuzeigen, wie Contoso seine Gewinnmarge um 10 % steigern kann.

Das Team konzentriert sich zunächst auf zwei zentrale Bereiche:

- Verständnis des Kaufverhaltens der Kunden, um die Produktverkäufe vorherzusagen. Dies beinhaltet eine Optimierung von Logistikprozessen, eine bessere Nutzung von Regalflächen und eine Verringerung der Verschwendung von abgelaufenen Produkten.
- Verwenden von Kunden-, Verkaufs- und Marketingdaten zur Optimierung der Ausgaben von Contoso für Werbeaktionen und Marketing, um mit der richtigen Werbung die richtigen Kunden zu erreichen.

Begeistert von dieser Aufgabe begann das Datenteam von Contoso mit einem Workshop, um die Anforderungen und technischen Herausforderungen zu präzisieren. Wie dies bei den meisten Unternehmen der Fall ist, sind die derzeitigen Datenpraktiken von Contoso auf die Geschehnisse in der Vergangenheit ausgerichtet. Die aktuellen Berichte beantworten Fragen wie „Wie viele Produkte wurden verkauft?“ und „Welcher Umsatz wurde mit Produkt A generiert?“. Contoso möchte jedoch etwas ganz anderes versuchen, nämlich Muster zu finden, um vorherzusagen, welche Produkte Contoso in welchen Mengen verkaufen sollte – und das nahezu in Echtzeit. Um dies umzusetzen, kam das Datenteam von Contoso zu dem Schluss, dass es sich den folgenden Herausforderungen stellen muss.

Erfassen und Verarbeiten neuer Daten

Contoso interagiert mit seinen Kunden über mehrere physische und digitale Kanäle. Bei allen diesen Interaktionen werden Daten generiert, die für Contoso wertvoll sein können. Denken Sie nur an alle Transaktionen an der Kasse, die Reaktionen der Kunden auf unterschiedliche Werbung, Anpassungen der Gänge im Geschäft sowie die Treuekartenpunkte, die Kunden möglicherweise gesammelt haben. Bei jeder Kundeninteraktion wird eine Vielzahl von Daten generiert, die Contoso erfassen muss.

Darüber hinaus werden im Onlineshop von Contoso mithilfe von Trackern und Beacons die Aktivitäten der Kunden und ihre Reaktionen auf Produkte aus der Werbung aufgezeichnet. Die mobile Anwendung von Contoso bietet eine ähnliche Funktionalität. Contoso kann sich damit ein sehr gutes Bild davon verschaffen, was Kunden mögen und was nicht. Contoso verwendet eine Kombination aus Azure Application Insights und Splunk sowie interne Tools, um die Click-Ereignisse und die Navigation der Anwender, die auf den einzelnen Seiten verbrachte Zeit, die dem Warenkorb hinzugefügten Produkte und die Zahl der abgeschlossenen Bestellungen zu erfassen. Mit diesen Daten in Verbindung mit Anwendungs-Log-Dateien, Netzwerküberwachungsereignissen und den bereits bekannten Informationen über die Kunden hat Contoso ein leistungsstarkes Tool zur Hand, um Nutzungstrends und -muster vorherzusagen.

Contoso muss alle diese Daten aus physischen und digitalen Kundeninteraktionen nicht nur erfassen und speichern, sondern auch bereinigen, validieren, vorbereiten und aggregieren. Eine enorme Aufgabe, mit der das Team noch keine Erfahrung hat. Bislang hat das Team eine Stapelverarbeitung der Daten durchgeführt und dazu historische Daten in das Data Warehouse geladen, um Tages- und Wochenberichte zu generieren. Dies ist eine ziemlich große Herausforderung für das Team, eine spannende, aber auch gewaltige Aufgabe.

Zusammenbringen aller Daten

Daten liegen gewöhnlich in verschiedenen Formaten und Formen vor. Kauftransaktionen beispielsweise beinhalten hoch strukturierte Tabellendaten, die einfach zu bearbeiten sind. Anwendungs-Log-Dateien andererseits sind semistrukturierte Dateien, in denen Millionen von Ereignissen und Ablaufverfolgungsmeldungen zu den Aktivitäten auf den Anwendungsservern aufgeführt sind. Contoso muss diese beiden Arten von Daten erfassen: strukturierte und unstrukturierte Daten.

Um das Ganze noch interessanter zu machen, sind Social-Media-Feeds unstrukturiert und in einer natürlichen Sprache verfasst, die Kunden beim Schreiben im Web verwenden. Diese Feeds können für Contoso sehr wertvoll sein, da das Unternehmen hier etwas über das tatsächliche Feedback seines Kundenstamms erfährt. Für die Datenbearbeiter (Data Practitioner) ist es jedoch schwierig, diese losen Feeds mit Beiträgen in natürlicher Sprache in demselben Format und derselben Form wie die hoch strukturierten Transaktionsdaten zu erfassen und zu organisieren.

Das Datenteam von Contoso muss sich den Herausforderungen stellen, nicht nur Daten in unterschiedlichen Formen (strukturiert, semistrukturiert und unstrukturiert) zu erfassen, sondern auch eine Möglichkeit zu finden, all diese Daten an einem zentralen Ort zu bereinigen und zu speichern, damit sie mit anderen Formen von Daten aus anderen Quellen verbunden und korreliert werden können und so neue Insights möglich sind.

Finden von Insights und Mustern in Daten

Nachdem alle Daten erfasst, bereinigt, validiert und gespeichert wurden, muss das Datenteam von Contoso mit der schwierigen Aufgabe beginnen, aussagekräftige Insights und Muster in den Daten zu finden. Dies kann sich als kompliziert erweisen. Wie sollen Sie angesichts verschiedener Datasets in einer Größenordnung von Gigabyte (oder gar Terabyte) Muster finden? Wo fangen Sie an?

Traditionelle Techniken für die Berichterstellung und Statistiken sind nicht skalierbar und reichen nicht aus, um diese Herausforderung zu bewältigen. Herkömmliche Formen der Programmierung sind nicht sehr hilfreich, da die Programmierer und die Datenbearbeiter selbst nicht wissen, wonach sie suchen oder wie sie diese Insights gewinnen können.

Ermittlung in Echtzeit

Contoso muss schnell aussagekräftige Insights gewinnen und die Erkenntnisse umgehend umsetzen. Daten verlieren in der Regel mit der Zeit an Wert, manche Daten verlieren ihren Wert sogar direkt nach der Erfassung. Stellen Sie sich etwa vor, dass Contoso eine größere Werbeaktion für ein bestimmtes Produkt, beispielsweise einen Softdrink der Marke ABC, durchführt. Das Getränk verkauft sich heute in den Filialen X, Y und Z sehr gut. Es hätte wenig Sinn, wenn Contoso diesen Trend morgen feststellen würde, da die Regale der Filialen X, Y und Z zu diesem Zeitpunkt leer wären und die Kunden enttäuscht wären, wenn sie das gewünschte Produkt nicht bekommen könnten. Damit würde Contoso gute Chancen verpassen, mehr zu verkaufen.

Aus diesem Grund ist Contoso bestrebt, Insights und Trends in Echtzeit oder nahezu in Echtzeit zu gewinnen. „Nahezu in Echtzeit“ wird bei Contoso als 5 – 10 Sekunden nach Echtzeit definiert. Den Datenpipelines bleibt damit gerade genug Zeit, um Livedaten zu verarbeiten und zu analysieren, die bei Contoso generiert werden.

Die CEO von Contoso hat dem Team klar gemacht, dass das Unternehmen nahezu in Echtzeit wissen muss, wie der Betrieb läuft und wie die Kunden über seine Marke und Dienstleistungen denken. Sie hat eine Situation erwähnt, in der Contoso gerade beschlossen hatte, den Verkauf von Produkt A einzustellen. Nach dieser Entscheidung gab es viele Diskussionen von Kunden in Social-Media-Plattformen. Dann hat die CEO von Contoso folgende Frage gestellt: Was wäre, wenn viele Kunden von Contoso online über einen möglichen Wechsel zu einem Mitbewerber von Contoso nur aufgrund dieser Entscheidung sprechen würden? Die Antwort auf diese Frage liefert ein Stück Informationen, das für Contoso von entscheidender Bedeutung ist und den Erfolg des Unternehmens beeinträchtigen könnte. Die CEO macht deutlich, dass die Möglichkeit, in Echtzeit Analysen durchzuführen, Insights zu gewinnen und umzusetzen, ein enormer Wettbewerbsvorteil für Contoso sein kann.

Zusammenfassend ist Contoso mit den folgenden Herausforderungen konfrontiert:

- Contoso möchte große Datasets aus unterschiedlichen Datenquellen mit potenziell hohem Durchsatz erfassen und speichern. Zu diesen Datenquellen gehören Transaktionsdatenspeicher, **IoT**-Sensoren (**Internet of Things**), die Onlineshops von Contoso und Anwendungs-Log-Dateien.
- Das Unternehmen möchte auch strukturierte, semistrukturierte und unstrukturierte Daten kombinieren und daraus durch Verknüpfung und Korrelation von Daten aus mehreren Quellen ein einzelnes Dataset erstellen.
- Contoso muss die unterschiedlichen Granularitätsebenen und Qualitätsstufen der verschiedenen Datenpunkte bewältigen. Das Team muss diese verschiedenen Datasets bereinigen, vorbereiten, transformieren und verknüpfen.
- Contoso möchte nahezu in Echtzeit aussagekräftige Insights und Muster aus den Daten gewinnen.
- Schließlich möchte das Unternehmen den Datenermittlungsprozess skalieren, um den Anforderungen des Unternehmens gerecht zu werden.

Design-Brainstorming

In den folgenden Abschnitten geht es darum, die Anforderungen besser zu formulieren und eine technische Lösung zu finden, die diesen Anforderungen gerecht wird.

Datenerfassung

Die erste Aufgabe für jeden Datenbearbeiter besteht darin, nach Daten zu suchen, sie zu erfassen, zu bereinigen, zu validieren und dann mit der spannenden Datenermittlung und -erkundung zu beginnen. Für das aktuelle Szenario müssen Sie die Datenquellen definieren, aus denen Sie Daten abrufen müssen. Sie müssen auch darüber nachdenken, wie Sie Daten aus verschiedenen Quellen laden können, um ein einzelnes Dataset zu erstellen, das von den Datenanalysten leicht erkundet und abgefragt werden kann. Einige Quellsysteme, die Sie für diesen Anwendungsfall benötigen sind:

- **Verkaufstransaktionen**: Aus den Verkaufstransaktionen geht nicht nur hervor, welche und wie viele Produkte in einem bestimmten Geschäft verkauft wurden, sondern auch, welche Kunden welche Produkte gekauft haben. Dies liegt daran, dass Contoso bereits über ein Treueprogramm verfügt, für das Kunden an der Kasse ihre Treuekarte scannen. Contoso verfügt über zwei verschiedene Datenspeicher für Verkaufstransaktionen: einen Datenspeicher für physische Geschäfte und einen anderen für die Onlineshops von Contoso.
- **Kundendaten**: Contoso verfügt über ein **CRM**-System (**Customer Relationship Manager, Kundenbeziehungsmanager**), das Kundendaten enthält. Zu den Kundendaten gehören (unter anderem) Vornamen, Nachnamen, Privatadressen, Telefonnummern, E-Mail-Adressen und Informationen zur Altersgruppe.
- **Treueprogrammdaten**: Die Daten des Treueprogramms werden in einem anderen Quellsystem gespeichert. Sie helfen Contoso bei der Verknüpfung von Kundendaten mit Verkaufstransaktionen.
- **Clickstreams und Nutzungsdaten zu digitalen Anwendungen**: Diese zeigen, wie die Kunden von Contoso auf das Design und die Inhalte der Contoso-Anwendungen reagieren.
- **Sensoren und IoT-Daten**: Einige Filialen von Contoso sind mit digitalen Sensoren ausgestattet, um das Verhalten der Kunden im Geschäft zu verstehen. In einigen Filialen sind IoT-Sensoren installiert, die zählen, wie viele Kunden zu welchem Zeitpunkt an den einzelnen Gängen vorbeigehen. Es gibt auch Sensoren, die die Temperatur und Feuchtigkeit in den Filialen von Contoso messen. Mithilfe dieser Sensoren stellt Contoso sicher, dass frische Produkte wie Milch unter den richtigen Bedingungen aufbewahrt werden. Darüber hinaus verfügt Contoso auch über Sensoren, um die Kundenzahl nahezu in Echtzeit zu zählen. So kann Contoso in Stoßzeiten/Hauptgeschäftszeiten mehr Mitarbeiter einsetzen, um einen schnelleren Service zu gewährleisten, damit die Kunden nicht lange warten müssen.

- **Andere Datasets**: Um die Daten von Contoso anzureichern und eine weitere Dimension zu eröffnen, erwägt das Datenteam von Contoso, andere Daten wie z. B. Wetterdaten, Daten von **Geo-Informationssystemen** (**GIS**)/Kartendaten, Vorort- und Stadtprofildaten und andere ähnliche Daten aus öffentlichen Datasets abzurufen. Diese Datasets können die Daten von Contoso bereichern und die Trends und Muster im Kundenverhalten und bei den Verkaufszahlen in einen größeren Zusammenhang stellen. Nehmen wir zum Beispiel das Wetter. Contoso könnte feststellen, dass der Verkauf bestimmter Produkte mit bestimmten Wetterbedingungen korreliert. So könnte etwa ein verstärkter Verkauf von Speiseeis im Sommer beobachtet werden. Ebenso kann ein Stadtprofil mit gewissen Merkmalen in Bezug auf Altersgruppen und Durchschnittseinkommen in starker Korrelation zu den Verkaufszahlen bestimmter Produkte stehen. In Vororten, in denen das Durchschnittsalter bei 25 Jahren liegt, könnten beispielsweise mehr Hairstylingprodukte verkauft werden, während der Verkauf dieser Produkte in den Vororten, in denen das Durchschnittsalter bei 45 Jahren liegt, weitaus geringer ausfällt.

Datenspeicher

Wie bereits erwähnt, muss Contoso Daten aus einer Vielzahl von Quellen erfassen. Contoso schätzt die Größe seiner aktuellen Datasets auf über 400 TB, pro Tag kommen durchschnittlich 10 – 15 GB dazu. Das Format dieser Datasets ist ziemlich unterschiedlich. Manche sind hoch strukturiert, andere dagegen völlig unstrukturiert. Eines haben alle diese verschiedenen Datasets gemeinsam: Sie wachsen schnell und kommen mit hoher Durchsatzrate an. Um den Anforderungen von Contoso gerecht zu werden, benötigen wir einen Datenspeicher, für den Folgendes gilt:

1. Der Speicher muss skalierbar und elastisch sein, um mit den Anforderungen des Datenteams von Contoso wachsen zu können.
2. Es muss sich um eine sichere und kontrollierte Plattform handeln, damit die Ressourcen und das geistige Eigentum von Contoso gut geschützt sind.
3. Der Speicher muss mit anderen vorhandenen Systemen und Tools kompatibel sein.
4. Der Preis muss angemessen sein.
5. Operationen mit hohem Durchsatz und parallele Verarbeitung müssen unterstützt werden.

Data Science

Nachdem alle Daten erfasst und in einem zentralen Datenspeicher gespeichert wurden, benötigt das Datenteam von Contoso eine Plattform für folgende Aufgaben:

- Bereinigen, Transformieren und Erkunden der Datasets
- Zusammenarbeit mit anderen Stakeholdern aus dem geschäftlichen und technischen Bereich, um Muster und Trends zu entdecken
- Integration in Frameworks und Runtimes für künstliche Intelligenz, um Machine-Learning-Algorithmen auf die Datasets anzuwenden und alle Muster aufzudecken
- Training und Operationalisierung der neuen Machine-Learning-Modelle, die aus der Arbeit der *vorherigen Integration* hervorgehen könnten
- Bereitstellen einer Möglichkeit für das Team, Datenpipelines zu planen, auszuführen und zu überwachen, um die Datentransformation,-bereinigung und -integration zu ermöglichen

Dashboards und Berichte

Die Entwicklung von Daten-Analytics-Lösungen kann als kontinuierlicher Dialog zwischen den Datenbearbeitern, den Stakeholdern des Unternehmens und den Daten selbst angesehen werden. Sie erfordert eine kontinuierliche Präzisierung und Hypothesentests. Daher muss das Datenteam von Contoso interaktive Berichte und Dashboards entwickeln und unterhalten, um dem Unternehmen seine Arbeit und die Ergebnisse seiner Datenermittlungsprozesse zu kommunizieren.

Der erste Schritt beim Erstellen solcher Berichte und Dashboards besteht darin, ein konsistentes gemeinsames Datenmodell zu entwickeln, das eine allgemeine Verständigung im gesamten Unternehmen ermöglicht.

Die Lösung

Das Datenteam von Contoso entschied sich, Microsoft Azure zur Implementierung der Analytics-Lösung des Unternehmens zu verwenden. Als Hauptfaktoren für diese Entscheidung nannte das Contoso-Team unter anderem die Skalierbarkeit, die Compliance und die regionale Verfügbarkeit von Azure in den Geschäftsregionen von Contoso (Australien und Japan). Das Team erläuterte auch, warum die einzelnen ausgewählten Azure-Dienste für seine Zwecke geeignet sind, wie wir in den nächsten Abschnitten sehen werden.

Contoso nutzte die präzisierten Anforderungen und Brainstorming-Ideen (aus den vorherigen Abschnitten), um ein Design für die Lösungsarchitektur zu entwickeln. Das Datenteam von Contoso erarbeitete folgende Architektur (wie in *Abbildung* 5.1dargestellt):

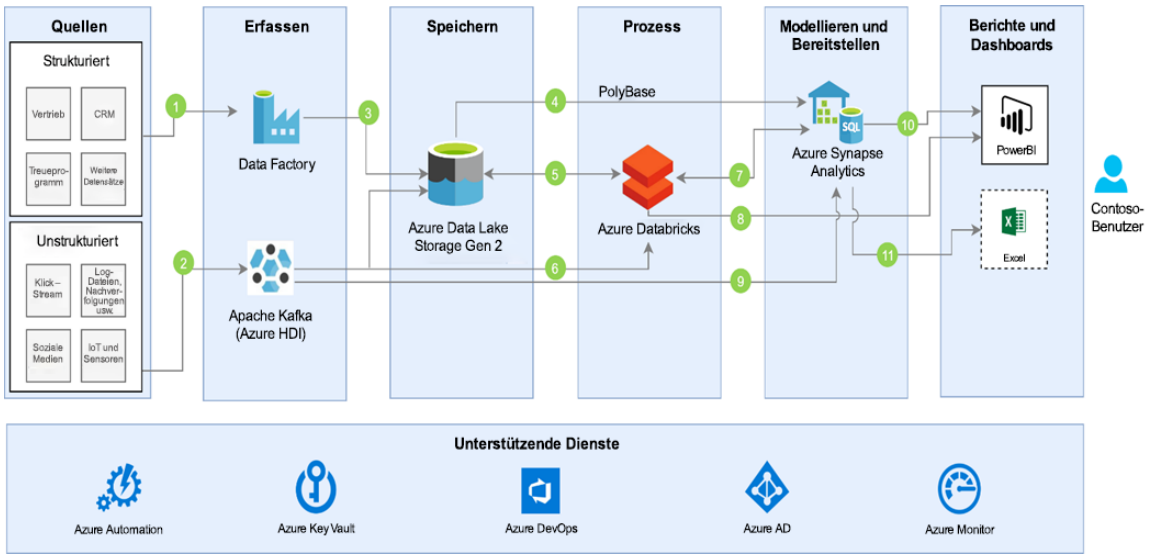

Abbildung 5.1: Die Lösungsarchitektur von Contoso

Datenfluss

Abbildung 5.1 zeigt die Lösungsarchitektur und den Datenfluss zwischen den einzelnen Komponenten. Im Folgenden werden die einzelnen Workflow-Segmente erläutert, wie im vorstehenden Diagramm markiert (nummeriert):

1. Contoso muss viele verschiedene Datasets erfassen. Einige davon enthalten strukturierte Daten, andere unstrukturierte Daten. Für die strukturierten Daten verwendet Contoso **Azure Data Factory**, um diese Daten mithilfe periodischer (5-Minuten-)Batchaktivitäten zu erfassen und **in Azure Data Lake Storage** zu übertragen.

2. Für die unstrukturierten Daten nutzt Contoso **Apache Kafka** in einem **Azure HDInsight**-Cluster, um diese Daten nahezu in Echtzeit zu erfassen und in **Azure Data Lake Storage** und **Databricks** (Spark Structured Streaming) zu übertragen. Auf diese Weise stehen alle neuen Daten für die Verarbeitung durch die Analytics-Lösung von Contoso zur Verfügung, und Contoso kann nahezu in Echtzeit (5 – 10 Sekunden später) eine Aktion für die Daten auslösen. Zu den unstrukturierten Daten gehören Daten aus Clickstream-Analytics (Berichte über das Verhalten der Anwender in den digitalen Kanälen von Contoso), Social-Media-Feeds (von Twitter, Facebook usw.), Log-Dateien und Ablaufverfolgungsinformationen von den Servern von Contoso sowie alle Daten von IoT-Sensoren.

3. Nach der Datenerfassung überträgt **Azure Data Factory** die Daten in **Azure Data Lake Storage**. **Azure Data Factory** verfügt über einen nativen Datenkonnektor zu **Azure Data Lake Storage Gen2**.

4. Daten, die in **Azure Data Lake Storage** gespeichert sind, können über **PolyBase** direkt in **Azure Synapse Analytics** geladen werden. PolyBase ist eine Technologie, die von Microsoft entwickelt wurde, um Abfragen und die Arbeit mit Daten in SQL-Servern (und Synapse Analytics) sowie Hadoop-basierten Dateisystemen (wie Azure Data Lake Storage) zu vereinheitlichen.

5. Die Daten von Contoso kommen in **Azure Data Lake Storage Gen2** an. Es handelt sich hier um Daten aus verschiedenen Quellen mit unterschiedlichen Qualitätsstufen und Granularitäten. Daher muss das Datenteam von Contoso die Datasets bereinigen, vorbereiten, validieren und anreichern. Hierfür wird **Azure Databricks** verwendet. Azure Databricks kann direkt mit Azure Data Lake Storage verbunden werden, um Daten abzurufen und die Ergebnisse verarbeiteter Daten zu speichern.

6. Azure Databricks bietet die Möglichkeit, Datenströme mithilfe von Apache Spark Streaming zu verarbeiten. Das Datenteam von Contoso nutzt dies für die Verarbeitung unstrukturierter Datenströme, die von **Apache Kafka** erfasst werden. Darüber hinaus werden mit dem nativen Support von Azure HDInsight für **Azure Data Lake Storage Gen2** eingehende Daten aus Anwendungs-Log-Dateien und sozialen Netzwerken sowie alle anderen unstrukturierten Daten von **Apache Kafka** in **Azure Data Lake Storage Gen2** übertragen. Andere strukturierte Datenströme können direkt in **Azure Synapse Analytics** übertragen werden (siehe *Punkt* 9 unten).

7. Während der Bereinigung und Vorbereitung der Daten in **Azure Databricks** muss das Team von Contoso einige Daten aus dem **Data Warehouse** abrufen, um historische Daten mit den neu ankommenden Daten zu kombinieren. Nach Abschluss der Datenvorbereitung werden die Daten in **Azure Synapse Analytics** übertragen, wo dann alle Daten kombiniert, modelliert und zur Nutzung vorbereitet werden. Darüber hinaus kann Azure Synapse Analytics als Ausgabesenke für Apache Spark Structured Streaming verwendet werden. Damit stehen nachgelagerten Anwendern und Datenanalysten von Contoso die Tools zur Verfügung, um mithilfe von Synapse Analytics nahezu in Echtzeit auf die Datenströme zuzugreifen. Auf diese Weise können die Systeme der Anwender von Contoso nicht nur Abfragen ausführen und Fragen zu den neu ankommenden Daten beantworten, sondern diese neuen Daten auch mit den historischen Daten zu kombinieren, die sich bereits in Azure Synapse Analytics befinden, um zu einer Übereinstimmung über das Geschäftsergebnis und das Kundenverhalten zu gelangen.

8. Mithilfe von **Power BI** können Ad-hoc-Abfragen für Daten ausgeführt werden, die in **Azure Databricks** bereinigt und verarbeitet werden. Power BI unterstützt diese Integration auf verschiedene Arten, einschließlich **Datenflüssen**, **direkter Abfrage** und **Datenimport**.

9. Mit den neuen Funktionen von Azure Synapse Analytics zur Verarbeitung von Live-Datenströmen können semistrukturierte Streamingdaten direkt von **Apache Kafka** (HDInsight-Cluster) in **Azure Synapse Analytics** übertragen werden.
10. Power BI bietet Contoso nicht nur die Möglichkeit, Berichte und Dashboards für Contoso-Anwender zu veröffentlichen. Mit Power BI kann darüber hinaus jeder Anwender zum Datenanalysten für seine Domäne werden – mit einem Self-Service-Ansatz und durch Erkundung der veröffentlichten Datenmodelle. Für große Datensätze kann Contoso zusammengesetzte Datenmodelle verwenden. Dies ist ein Feature von Power BI Premium.
11. Contoso hat in komplexe Modelle investiert, die **Microsoft Excel** verwenden. Einige der Datenanalysten von Contoso würden gerne über **Microsoft Excel** auf Daten von Azure Synapse Analytics zugreifen. Dies wird in Microsoft Excel und Azure Synapse Analytics unmittelbar unterstützt.

Azure-Dienste

In den folgenden Abschnitten werden wir uns eingehender mit den Azure-Diensten befassen, die in dem Lösungsdesign in *Abbildung* 5.1dargestellt sind. Für jeden Dienst erläutern wir zunächst, warum die betreffende Komponente benötigt wird und warum die Azure-Dienste für Contoso geeignet sind. Abschließend zeigen wir ein kurzes praktisches Beispiel für den Kernteil der Implementierung.

Azure Data Factory

Rolle im Design: Wie die meisten anderen Unternehmen verfügt Contoso über eine große Anzahl von Datenquellen. Einige dieser Datenquellen befinden sich On-Premises, andere in der Cloud. Wie bereits erwähnt, muss Contoso alle diese Daten an einem Ort zusammenbringen, um die Datasets kombinieren, korrelieren, modellieren und transformieren zu können und so Trends zu erkennen und Insights zu gewinnen. Hierfür müssen viele Datenkonnektoren entwickelt und verwaltet werden, um die Daten aus den Quellsystemen von Contoso in den zentralen Datenspeicher (den Data Lake) zu verschieben. Genau an diesem Punkt zeigen sich die Vorteile von Azure Data Factory, da es sich hier um einen verwalteten Dienst handelt, der die Datenintegration für Anwender jedes Kenntnisstands vereinfachen soll.

Vorteile von Azure Data Factory

1. Azure Data Factory bietet mehr als 80 vorgefertigte Datenkonnektoren. So kann Contoso Quellsysteme ohne zusätzliche Kosten schnell und einfach mit dem neuen modernen Data Warehouse verbinden. Diese Datenkonnektoren werden von Microsoft entwickelt. Sie bieten eine effiziente und robuste Integration und nutzen das Microsoft Azure-Netzwerk, das einen Durchsatz von bis zu 1,5 GB/s ermöglicht. Damit kann Contoso nicht nur eine schnelle Markteinführung erreichen, sondern erhält auch eine Plattform für die Orchestrierung der Datenverschiebung mit minimalem Aufwand.

2. Contoso verfügt über eine Reihe vorhandener SQL-Server. Diese Server hosten mehrere **SSIS**-Pakete (**SQL Server Integration Services)** für vorhandene Berichte und Dashboards. Azure Data Factory bietet eine Integrationslaufzeit, die für die Verarbeitung von SSIS-Paketen konzipiert ist. Dadurch ist Azure Data Factory eine perfekte Plattform für Contoso, da das Unternehmen seine vorhandenen Investitionen in diese SSIS-Pakete weiter nutzen kann.

3. Neben den vorgefertigten Datenkonnektoren bietet Azure Data Factory eine grafische Oberfläche, über die jeder mit wenig oder ganz ohne Code umfassende Pipelines für die Datenverschiebung entwickeln kann. Darüber hinaus ermöglicht der Visual Editor von Azure Data Factory eine Integration in Git-Repositorys für die Quellcodeverwaltung, um Flexibilität und Wartbarkeit zu verbessern. Dies kommt dem Contoso-Team zugute, da dadurch eine Verbesserung der Produktivität und Entwicklungsgeschwindigkeit des Teams bei geringerem Aufwand möglich wird. Mit diesem Feature kann das Contoso-Team die leistungsstarken Möglichkeiten der visuellen Datentransformation von Azure Data Factory sowie Data Wrangling in dem visuellen Portal nutzen und gleichzeitig die Versionskontrolle der gesamten Arbeit aufrechterhalten.

4. Azure Data Factory ist ein vollständig verwaltetes Tool, das es dem Contoso-Team ermöglicht, klein anzufangen und dabei wenig oder gar nicht zu investieren und dann nach Bedarf zu skalieren. Das bedeutet auch, dass keine Infrastruktur verwaltet werden muss und das Team von Contoso nur für die tatsächliche Nutzung bezahlt.

5. Neben anderen Zertifizierungen verfügt Azure Data Factory über Zertifizierungen nach **ISO/IEC 27001** und **27018** und ist in 25 Ländern/Regionen verfügbar, darunter auch Australien und Japan, wo Contoso tätig ist. Das macht Azure Data Factory für Contoso sehr interessant, da der Dienst alle Kriterien der Checkliste des Unternehmens in Bezug auf Sicherheit und Compliance erfüllt.

6. Azure Data Factory bietet die Tools zum Entwickeln von Datenpipelines, die Schemaabweichungen gegenüber robust sind. Wenn das Contoso-Team Pipelines entwickelt, um Daten von Quelle A nach B zu verschieben, kann es somit sicher sein, dass die Pipelines auch dann noch funktionsfähig bleiben, wenn sich das Schema der Daten aus Quelle A geändert hat. Dies bedeutet eine erhebliche Verbesserung der Zuverlässigkeit und Resilienz der Datenpipelines von Contoso.

7. Schließlich hat Contoso mit Azure Data Factory die Möglichkeit, alle Aktivitäten der Datenverschiebung und -verarbeitung über eine einzelne Steuerungsebene zu verwalten.

Beispielimplementierung

Hier sehen Sie ein Beispiel dafür, wie Contoso Azure Data Factory konfiguriert, um Daten aus seiner Verkaufstransaktionsdatenbank (die sich auf einem Azure SQL Server befindet) in Azure Data Lake Storage Gen2 zu übertragen:

1. Wie in *Kapitel 2*, *Entwicklung Ihres modernen Data Warehouse*, erläutert, bietet Azure Data Factory native Integration in Azure Data Lake Storage Gen2. Contoso kann eine Verbindung mit Azure Data Lake Storage Gen2 herstellen, indem es wie folgt einen verknüpften Dienst in Azure Data Factory konfiguriert:

```
{
    "name": "ContosoAzureDLStorageLS",
    "properties": {
        "type": "AzureBlobFS",
        "typeProperties": {
            "url": "https://{accountname}.dfs.core.windows.net",
            "accountkey": {
                "type": "SecureString",
                "value": "{accountkey}"
            }
        },
        "connectVia": {
            "referenceName": "{name of Integration Runtime}",
            "type": "IntegrationRuntimeReference"
        }
    }
}
```

   Zu beachten ist, dass dieses Beispiel Platzhalter für die wichtigsten Konfigurationswerte enthält, wie z. B. den **`Namen`** des Azure Storage-Kontos, **`accountKey`** (Kontoschlüssel) und den Namen der Integrationslaufzeit.

2. Nach dem Erstellen eines verknüpften Diensts in Azure Data Factory benötigen wir ein Azure-Dataset, das auf diesen verknüpften Dienst verweist. Dieses kann wie folgt erstellt werden:

```
{
    "name": "ContosoAzureDataLakeSalesDataset",
    "properties": {
        "type": "DelimitedText",
        "linkedServiceName": {
            "referenceName": "ContosoAzureDLStorageLS",
            "type": "LinkedServiceReference"
        },
        "schema": [ { optional } ],
        "typeProperties": {
            "location": {
                "type": "AzureBlobFSLocation",
                "fileSystem": "{filesystemname}",
                "folderPath": "contoso/sales"
            },
            "columnDelimiter": ",",
            "quoteChar": "\"",
            "firstRowAsHeader": true,
            "compressionCodec": "gzip"
        }
    }
}
```

   Im vorherigen Codeausschnitt wird der verknüpfte Dienst von Azure Data Lake Storage zum Erstellen eines Datasets verwendet. Dieses Dataset erstellt **`CSV`**-Dateien (Comma-Separated Values – durch Trennzeichen getrennte Werte) und speichert diese als komprimierte Dateien (**`gzip`**).

3. Konfigurieren Sie die Azure SQL Database als verknüpften Dienst:

```
{
    "name": "ContosoSalesAzureSqlDbLS",
    "properties": {
        "type": "AzureSqlDatabase",
        "typeProperties": {
            "connectionString": {
                "type": "SecureString",
```

```
                "value": "Server=tcp:{servername}.
database.windows.net,1433;Database={databasename};User
ID={username}@{servername};Password={password};Trusted_
Connection=False;Encrypt=True;Connection Timeout=30"
            }
        },
        "connectVia": {
            "referenceName": "{name of Integration Runtime}",
            "type": "IntegrationRuntimeReference"
        }
    }
}
```

Im vorherigen Codeausschnitt werden Platzhalter für die folgenden Parameter verwendet: für den Azure SQL Server-Namen, den Namen der SQL Database, den Benutzernamen und das Kennwort für SQL Server und den Namen der Integrationslaufzeit. Auch dient das Beispiel nur zu Demonstrationszwecken: Der Code sollten nie Kennwörter enthalten, und die Kennwörter sollten in Azure Key Vault gespeichert werden, um die Sicherheit zu gewährleisten.

4. Ähnlich wie in *Schritt 2* müssen Sie ein Dataset in Azure Data Factory für die Verkaufsdatenbank von Contoso konfigurieren. Im folgenden Codeausschnitt wird mithilfe des verknüpften Diensts der Azure SQL Database ein Dataset erstellt, das auf **`sales_table`** in der SQL Database von Contoso verweist:

```
{
    "name": "ContosoSalesDataset",
    "properties":
    {
        "type": "AzureSqlTable",
        "linkedServiceName": {
            "referenceName": "ContosoSalesAzureSqlDbLS",
            "type": "LinkedServiceReference"
        },
        "schema": [ {optional} ],
        "typeProperties": {
            "tableName": "sales_table"
        }
    }
}
```

5. Mit dem folgenden Codeausschnitt wird die Aktivität der Datenverschiebung von der SQL-Verkaufsdatenbank in Azure Data Lake konfiguriert. Dadurch wird eine Aktivität in Azure Data Factory erstellt, die auf die beiden in *Schritt 2* und *Schritt 4* erstellten Datasets verweist. Die Aktivität legt die Azure SQL-Verkaufsdatenbank als Quelle der Datenverschiebung und Azure Data Lake Storage als Ziel der Datenverschiebungsaktivität fest:

```
{
    "name": "CopyFromAzureSQLSalesDatabaseToAzureDataLake",
    "type": "Copy",
    "inputs": [
        {
            "referenceName": "ContosoSalesDataset",
            "type": "DatasetReference"
        }
    ],
    "outputs": [
        {
            "referenceName": "ContosoAzureDataLakeSalesDataset",
            "type": "DatasetReference"
        }
    ],
    "typeProperties": {
        "source": {
            "type": "AzureSqlSource",
            "sqlReaderQuery": "SELECT * FROM SALES_TABLE"
        },
        "sink": {
            "type": "ParquetSink",
            "storeSettings":{
                "type": "AzureBlobFSWriteSetting",
                "copyBehavior": "PreserveHierarchy"
            }
        }
    }
}
```

Apache Kafka auf Azure HDInsight

Rolle im Design

Contoso muss unstrukturierte Datenströme erfassen, die aus Clickstreams, Social-Media-Feeds und IoT-Geräten stammen könnten. Hierfür wird ein robustes, skalierbares Modul für die Erfassung und Verarbeitung dieser Streams bei ihrer Ankunft benötigt.

Apache Kafka ist eine verteilte Streaming-Plattform, die die Erfassung von Datenströmen aus verschiedenen Quellen ermöglicht und ein Publish-Subscribe-Modell zur Aufzeichnung und Übertragung von Datenströmen bietet. Apache Kafka ist ein sehr beliebtes Open-Source-Projekt, das für die Entwicklung von Echtzeit-Streaming-Anwendungen verwendet wird. So möchte auch Contoso dies nutzen. Apache Kafka wird als Cluster ausgeführt, Contoso benötigt für die Verwendung also einen Cluster von Computern, auf denen Apache Kafka installiert ist. Dieser Ansatz ist kostspielig, nicht skalierbar und mit einem großen Wartungsaufwand verbunden.

Glücklicherweise bietet Microsoft Azure einen verwalteten HDInsight-Cluster, um Apache Kafka einfach und kostengünstig auszuführen und gleichzeitig die **Service Level Agreements** (**SLAs**) auf Unternehmensniveau von Azure zu nutzen.

Vorteile von Apache Kafka auf Azure HDInsight

Wie bereits erwähnt, kann Contoso Apache Kafka mit Azure HDInsight einfach und schnell ausführen, ohne Cluster verwalten zu müssen. Contoso hat bei der Entscheidung bezüglich Azure HDInsight für Apache Kafka folgende Aspekte berücksichtigt:

- Azure HDInsight ist sehr einfach einzurichten, und der Einstieg ist sehr leicht möglich. Allgemein sind Cluster schwer einzurichten und zu verwalten. Wenn Microsoft Azure die Verwaltung des Clusters übernimmt, entfällt dieses Problem, und das Team von Contoso kann sich auf die Lösung von Geschäftsproblemen konzentrieren, anstatt sich Gedanken um die Infrastruktur machen zu müssen.
- Contoso achtet sehr auf Sicherheit und Compliance. Azure HDInsight bietet Sicherheitskontrollen auf Unternehmensniveau und hat mehr als 30 Compliance-Zertifizierungen erhalten.
- Azure HDInsight basiert auf Open-Source-Technologien und ist für Hadoop und Spark optimiert. Contoso kann somit andere gängige Big-Data-Lösungen auf demselben Cluster (neben Apache Kafka) ausführen. Dies macht die Investition von Contoso lohnender, da derselbe Cluster zur Ausführung mehrerer Open-Source-Frameworks verwendet werden kann.
- Azure HDInsight-Cluster sind nicht nur einfach einzurichten und zu verwalten, sondern auch sehr kostengünstig, da Contoso Cluster auf Anforderung hochfahren und nach Bedarf hoch- oder herunterskalieren kann. Contoso muss sich also nicht um die Verwaltung oder Bezahlung von nicht genutzten Clustern kümmern und nur für die tatsächliche Nutzung bezahlen.
- Darüber hinaus kann Azure HDInsight für Analysen und für die Meldung von Statistiken zur Verfügbarkeit und Nutzung von Big Data verwendet werden. Mithilfe von Azure HDInsight kann Contoso produktive Tools für Hadoop und Sparks mit jeder bevorzugten Entwicklungsumgebung verwenden.

Beispielimplementierung

Hier sehen Sie ein Beispiel für die Erfassung eines Datenstroms in Azure HDInsight (Apache Kafka) und das Schreiben der Daten in eine komprimierte Parquet-Datei.

```
// Reading a kafka stream to a dataframe
val kafkaStreamDF = spark.readStream.format("kafka")
                .option("kafka.bootstrap.servers", kafkaBrokers)
                .option("subscribe", kafkaTopic)
                .option("startingOffsets", "earliest")
                .load()
// Writing streaming data to a parquet file
kafkaStreamDF.select(from_json(col("value")
            .cast("string"), schema) as "tweet")
            .writeStream
            .format("parquet")
            .option("path","/contoso/socialmedia/twitterfeed")
            .start.awaitTermination(10000)
```

Der vorherige Codeausschnitt liest den Datenstrom von Apache Kafka in einen DataFrame in Scala ein, basierend auf der Konfiguration des Apache Kafka-Servers und -Topics. Der DataFrame wird dann unter Verwendung des angegebenen Pfads in eine komprimierte (Parquet-)Datei geschrieben. Dabei wird der Auftrag nach **`10.000`** Sekunden automatisch beendet. In dem Codeausschnitt wird zudem davon ausgegangen, dass die Spark-Bibliotheken für Kafka ordnungsgemäß installiert und referenziert sind und mit der Spark-Version im HDInsight-Cluster übereinstimmen.

Azure Data Lake Storage Gen2

Rolle im Design

Azure Data Lake Storage fungiert als zentraler Datenspeicher von Contoso. So kann Contoso riesige Datenmengen aus verschiedenen Quellen zusammenbringen. Darüber hinaus unterscheiden sich die Datasets von Contoso in ihrer Art und ihrem Format erheblich (strukturiert, semistrukturiert und unstrukturiert). Daher wird ein leistungsfähigerer Datenspeicher als ein reiner Tabellenspeicher benötigt – Azure Data Lake Storage. Azure Data Lake Storage kann schemalose Daten als BLOBs speichern und verschiedene Formate verarbeiten (z. B. Textdateien, Bilder, Videos, Social-Media-Feeds, gezippte Dateien usw.). Dank der Möglichkeit, schemalose Datenformate zu verarbeiten, kann Contoso problemlos Daten im Rohformat erfassen. Dies ist für Advanced Analytics sehr wichtig, da die Analyse so an den ursprünglichen Daten ohne Verzerrung durch Datenaggregation erfolgen kann.

Darüber hinaus benötigt das Team von Contoso elastischen Speicher für eine Sandbox-Umgebung zum Erkunden und Transformieren der Daten. Auch Azure Data Lake Storage kann hierfür genutzt werden.

Vorteile von Azure Data Lake Storage Gen2

- Contoso verwendet Azure Active Directory für die Zugriffsverwaltung. Azure Data Lake Storage Gen2 bietet native, sofortige Integration in Azure Active Directory, um den Zugriff auf Daten unter Verwendung von Azure Active Directory als Steuermechanismus des Unternehmens zu verwalten. Dies verringert die Komplexität des Designs und verbessert Sicherheit und Compliance.
- Azure Data Lake Storage Gen2 erleichtert die Verwaltung und Organisation von Daten mithilfe von Verzeichnissen und Unterverzeichnissen, die mit hierarchischen Namespaces integriert sind. Auf diese Weise wird eine bessere Leistung erzielt, da Daten nicht mehr so oft über mehrere Ordner hinweg kopiert oder transformiert werden müssen. Gleichzeitig wird dadurch die Datenverwaltung vereinfacht.
- Azure Data Lake Storage Gen2 basiert auf Azure Blob Storage, der zum kostengünstigen Speichern entwickelt wurde. Azure Data Lake Storage Gen2 bietet eine Reihe von wertschöpfenden Features, wie z. B. hierarchische Namespaces, die die Gesamtbetriebskosten weiter reduzieren.
- Azure Data Lake Storage ist nicht nur kostengünstig, sondern legt auch keine Grenzen in Bezug auf die Menge der zu speichernden Daten fest. Das Team von Contoso kann also mit minimalen Kosten klein beginnen und dann nach Bedarf skalieren, ohne sich Gedanken über irgendwelche Höchstgrenzen machen zu müssen.
- Azure Data Lake Storage kann nativ mit Azure Synapse Analytics, Data Factory, Power BI und vielen anderen Microsoft Azure-Diensten integriert werden. Dies ist ein überzeugendes Argument für das Contoso-Team, da es bereits Power BI und Data Factory verwendet.
- Neben der Integration in Azure Active Directory bietet Azure Data Lake Storage Gen2 die Sicherheitsfunktionen, die das Sicherheitsteam von Contoso verlangt. Hierzu gehören Datenverschlüsselung – für ruhende und in Übertragung begriffene Daten – Single Sign-On, mehrstufige Authentifizierung, differenzierte Zugriffssteuerung für Anwender und Gruppen sowie vollständige Compliance mit Überwachungsrichtlinien durch Überwachung jedes Zugriffs und jeder Konfigurationsänderung im Data Lake.

Beispielimplementierung

Bei der Erfassung von Daten in Azure Data Lake Storage empfiehlt es sich, Namespaces und Container zu verwenden, um die Daten im Data Lake zu organisieren. Dies erleichtert nicht nur das Auffinden der Daten, sondern hilft auch bei der Zugriffssteuerungsverwaltung. *Abbildung* 5.2 zeigt ein Beispiel einer einfachen Data Lake-Zonenzuweisung. Dabei ist der Data Lake in vier Zonen unterteilt: **Landing Zone** (Zielzone, Erfassung), **Staging Zone** (Bereitstellungszone), **Secure Zone** (Sichere Zone) und **Analytics Sandbox**:

- **Landing Zone** (Zielzone): Hier treffen alle in den Data Lake gelangenden Daten (außer sensiblen Daten) ein, bevor sie verarbeitet, bereinigt, aggregiert werden usw.
- **Staging Zone** (Bereitstellungszone): Hier werden die Daten bereinigt/bereitgestellt, bevor sie für die Nutzung vorbereitet werden.
- **Analytics Sandbox**: Diese Zone wird von Datenwissenschaftlern und Dateningenieuren als Sandbox zum Speichern von Daten während der Verarbeitung und Erkundung verwendet.
- **Secure Zone** (Sichere Zone): Hier werden hoch sensible Daten gespeichert und verarbeitet. Durch die Trennung der sicheren Zone von den anderen Zonen ist eine bessere Zugriffssteuerungsverwaltung möglich. Diese Zone enthält sensible Daten, beispielsweise Daten zu Fusionen und Übernahmen, Finanzdaten und andere Kundendaten, die möglicherweise nur schwer zu maskieren sind, wie z. B. Angaben zum Geschlecht, zum Alter und zur ethnischen Zugehörigkeit der Kunden, sofern bekannt:

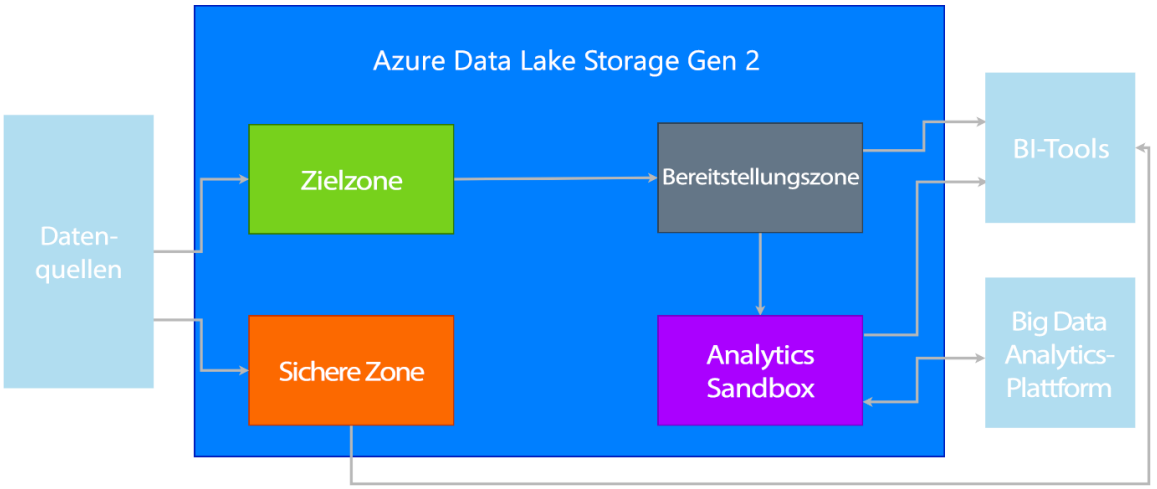

Abbildung 5.2: Beispiel für die Data Lake Storage-Zonenzuweisung

Azure Databricks

Rolle im Design

Innovation erfordert eine fruchtbare Zusammenarbeit zwischen Datenbearbeitern, Entwicklern und dem Unternehmen. Außerdem müssen große Datasets erfasst, bereinigt, kombiniert und transformiert werden. Databricks wurde entwickelt, um diese Zusammenarbeit zu ermöglichen und eine einheitliche Daten-Analytics-Plattform bereitzustellen. Contoso wird Azure Databricks als einheitliche Plattform für Data Science und Data Engineering verwenden. Azure Databricks bietet die Rechenleistung zur Bereinigung und Analyse von Daten und unterstützt mehrere Programmiersprachen und Frameworks für verschiedene Workloads.

Vorteile von Azure Databricks

- Azure Databricks bietet einen vollständig verwalteten Spark-Cluster. Genau das benötigt das Team von Contoso für die Vorbereitung, Transformation und Analyse von Daten. Azure Databricks verringert die Komplexität der Bereitstellung und Verwaltung des Spark-Clusters und ermöglicht Contoso einen schnellen und einfachen Einstieg.
- Azure Databricks lässt sich nativ mit anderen Azure-Diensten integrieren. Dies ist besonders wichtig für Sicherheit und Leistung. Azure Databricks kann mit Azure Active Directory integriert werden, es müssen also keine neuen Benutzerkonten hinzugefügt oder verwaltet werden. Auf ähnliche Weise unterstützt Azure Databricks Azure Synapse Analytics als Ausgabesenke für Spark Structured Streaming. So kann Contoso neue Daten nahezu in Echtzeit in Azure Synapse Analytics übertragen.
- Azure Databricks bietet eine umfassende Umgebung für die Zusammenarbeit. Hier können mehrere Stakeholder von Contoso gleichzeitig an demselben Notebook arbeiten. Dies verbessert die Produktivität erheblich und fördert Innovationen, da das Wissen aller Teammitglieder zusammenkommt.
- Azure Databricks ermöglicht die automatische Skalierung und Beendigung von Clustern. Contoso muss also nur für den Cluster bezahlen, wenn er verwendet wird. Dies ermöglicht eine deutliche Senkung der Kosten, da die Clusterknoten nur ausgeführt werden, wenn dies tatsächlich notwendig ist. Gleichzeitig bietet dies größere Flexibilität, da der Cluster automatisch skaliert werden kann.
- Vielfalt fördert Innovation. Daher unterstützt Azure Databricks mehrere Programmiersprachen und Frameworks. Datenwissenschaftler und Dateningenieure bei Contoso können R, Python, SQL, Scala, Java und C# verwenden, um Code in Azure Databricks zu schreiben. Dies ist besonders wichtig für Contoso, das Schwierigkeiten gehabt hatte, kompetente Mitarbeiter auf diesem Gebiet zu gewinnen und zu halten.

Beispielimplementierung

Hier sehen Sie ein kurzes Beispiel für Python-Code zum Übertragen von Daten aus dem Data Lake in einen DataFrame:

```
configs = {"fs.azure.account.auth.type": "OAuth",
      "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.
oauth2.ClientCredsTokenProvider",
      "fs.azure.account.oauth2.client.id": "{appId}",
      "fs.azure.account.oauth2.client.secret": "{password}",
      "fs.azure.account.oauth2.client.endpoint": "https://login.
microsoftonline.com/{tenant}/oauth2/token",
      "fs.azure.createRemoteFileSystemDuringInitialization": "true"}


dbutils.fs.mount(
source = "abfss://{file-system-name}@{storage-account-name}.dfs.core.windows.
net/folder1",
mount_point = "/mnt/contoso/sales",
extra_configs = configs)


salesDF = spark.read.format('csv').options(
   header='true', inferschema='true').load("/mnt/contoso/sales/*.csv")
```

In diesem Codeausschnitt wird davon ausgegangen, dass ein Dienstprinzipal vorhanden ist, der mit der richtigen Zugriffsebene in Azure Active Directory konfiguriert ist. Es sind auch Platzhalter für den Dienstprinzipal **`appId`**, das Kennwort des Dienstprinzipals, den Azure AD-Mandanten, den Namen des Azure Storage-Kontos und den Azure Data Lake-Dateisystemnamen enthalten. Darüber hinaus stellt der Code das Azure Data Lake Storage-Dateisystem in **`/mnt/contoso/sales`** bereit und liest dann alle Dateien mit der Dateierweiterung **`CSV`** in einen Spark-DataFrame ein.

Azure Synapse Analytics

Rolle im Design

Contoso benötigt eine Single Source of Truth (einzige Quelle der Wahrheit) für alle seine Daten. Vor irgendwelchen weiteren Schritten müssen diese Daten bereinigt, validiert, transformiert und aggregiert werden. Sobald dies geschehen ist, müssen die neu verfügbaren Daten mit historischen Datasets zusammengeführt werden, damit sich Contoso ein umfassendes Bild vom Geschäftsergebnis und den Betriebsprozessen verschaffen kann.

Azure Synapse Analytics fungiert als Speicher für diese Single Source of Truth. Contoso kann so ein mehrdimensionales Modell der Daten erstellen, und es wird ein gut strukturiertes Datenformat bereitgestellt, das für die Abfrage umfassender Datasets optimiert ist. Darüber hinaus benötigt Contoso auch eine Plattform, auf der gut zusammengestellte Daten auf prognostizierbarer Leistungsebene verwendet werden können. Auch diese Aufgabe erfüllt Azure Synapse Analytics, da die Datenbereinigung und -validierung hier von der eigentlichen Datenbereitstellung getrennt ist.

Vorteile von Azure Synapse Analytics

- Azure Synapse Analytics ist ein vollständig verwalteter Dienst, der dynamisch skalierbar ist. Dies ist für Contoso sehr interessant, da weniger Infrastruktur verwaltet werden muss, der Aufwand geringer ist und ein skalierbarer Dienst mit dem Unternehmen wachsen kann.
- Azure Synapse Analytics ist laut einem unabhängigen Benchmark-Bericht von GigaOm 14-mal schneller als andere Cloudanbieter. Azure SQL bietet blitzschnelle Leistung bei großen Datenmengen. Darüber hinaus erreicht Azure Synapse Analytics dies mithilfe von Massively Parallel Processing.
- Azure Synapse Analytics ist viel billiger als andere Cloudprodukte. Laut der Studie von GigaOm ist Azure Synapse Analytics rund 94 % günstiger als andere Cloudanbieter.
- Azure Synapse Analytics bietet sofortige Integration in Apache Spark Structured Streaming. So kann Contoso durch Kombination neuer Streamingdaten mit historischen Daten Berichte und Dashboards erstellen, um Customer Insights nahezu in Echtzeit zu gewinnen.
- Das Sicherheitsteam von Contoso hat eine klare Auflage, die den Schutz den Datenbestände vorschreibt. Das Sicherheitsteam hat verlangt, dass das Data Warehouse nicht öffentlich im Web zugänglich sein darf. Dies wird von Azure Synapse Analytics über die Azure Virtual Network-Integration nativ unterstützt. Dabei wird Azure Synapse Analytics als Teil des Contoso-Netzwerks (ein virtuelles privates Netzwerk) bereitgestellt.
- Auch die anderen Sicherheitsfeatures in Azure Synapse Analytics gefielen Contoso. Hierzu gehören die Integration in Azure Active Directory, die Aktivitätsüberwachung, die native Sicherheit auf Zeilen- und Spaltenebene, die ExpressRoute-Integration und die sofort einsatzbereite Bedrohungserkennung sowie Datenverschlüsselungsfunktionen.
- Einige Bereiche von Contoso verwenden andere BI-Tools wie Power BI, Tableau und Qlik und wollten daher sicherstellen, dass das neue Data Warehouse diese Integration unterstützt. Azure Synapse Analytics ist mit vielen BI-Tools kompatibel, einschließlich der vorhandenen BI-Tools von Contoso.

- Azure Synapse Analytics kann bei Bedarf auch Apache Spark einsetzen. Dies kann für das Datenteam von Contoso erhebliche Vorteile bringen, da das Team dann dieselben Open Source Tools verwenden kann, um in Azure Synapse Analytics mit seinen Daten zu arbeiten. Dies ermöglicht eine größere Produktivität, da Spark-Cluster mehrere Sprachen und Frameworks direkt unterstützen (Python, R, Scala usw.). So können die Teammitglieder von Contoso produktiv und komfortabel mit den Tools arbeiten, die ihnen am besten vertraut sind.
- Ein Data Warehouse zu entwickeln, bereitzustellen und zu verwalten, kann eine sehr komplexe Aufgabe sein. Hier zeigen sich die Vorteile von Azure Synapse Analytics, da dieser Dienst auf der langjährigen Erfahrung von Microsoft als Entwicklungsunternehmen gründet. Contoso zeigt unter anderem aufgrund des optimierten Workload-Managements und der hervorragenden Entwicklerproduktivität so großes Interesse an Azure Synapse Analytics. Azure Synapse Analytics bietet als einziges Cloud Data Warehouse native SSMS- und SSDT-Unterstützung, einschließlich Visual Studio-Projekten für die Code- und Schemaverwaltung, die unerlässlich sind, um einen optimierten Entwicklungszyklus zu gewährleisten und die Gesamtbetriebskosten zu reduzieren.

Beispielimplementierung

Azure Synapse Analytics kann als Ausgabesenke für Spark Structured Streaming in Azure Databricks verwendet werden. Dies bietet die Möglichkeit, schnell und einfach Live-Streamingdaten in Azure Synapse Analytics einzubringen. Hier ein kurzes Beispiel für die wichtigsten Schritte:

1. Mit dem folgenden Codeausschnitt wird eine Tabelle namens **`TweetsStream`** in Azure Synapse Analytics erstellt, die den Stream empfangen soll. Diese Tabelle umfasst zwei einfache Spalten: eine für den **`Zeitstempel`** und eine für den **`Wert`**, der aus dem Datenstrom empfangen wird. Im folgenden Beispiel wird **`ROUND_ROBIN`** dieser Tabelle als Verteilungsrichtlinie zugewiesen. Die Auswahl einer Verteilungsrichtlinie kann erhebliche Auswirkungen auf die Leistung haben. Wählen Sie im Allgemeinen Tabellen mit Hashverteilung (**`HASH-DISTRIBUTED`**), um die Abfrageleistung für große Faktentabellen zu verbessern, und **`ROUND_ROBIN`**, um die Ladegeschwindigkeit zu verbessern:

```
CREATE TABLE [dbo].[TweetsStream]
(
  [timestamp] DATETIME NULL,
  [Value] BIGINT NULL
)
WITH
(
  DISTRIBUTION = ROUND_ROBIN,
  CLUSTERED INDEX ([timestamp])
)
```

2. Im nächsten Schritt wird das Python-Skript auf Azure Databricks eingerichtet. Im folgenden Codeausschnitt werden Platzhalter für die Verbindungsdetails von Azure Synapse Analytics verwendet. Der Code stellt mit **Java Database Connectivity** (**JDBC**) eine Verbindung mit Azure Synapse Analytics her. Außerdem wird ein Azure Blob Storage-Konto eingerichtet, das für den temporären Speicher verwendet werden soll:

```
# Setup the connection to the Azure SQL DW
dwDatabase = {databaseName}
dwServer = {servername}
dwUser = {sqlUser}
dwPass = {password}
dwJdbcPort =  "1433"
dwJdbcExtraOptions =
"encrypt=true;trustServerCertificate=true;loginTimeout=30;"
sqlDwUrl = "jdbc:sqlserver://" + dwServer + ".database.windows.net:" +
dwJdbcPort + ";database=" + dwDatabase + ";user=" + dwUser+";password=" +
dwPass + ";"+dwJdbcExtraOptions

# Setup Blob Storage for temporary storage of the data stream
stAccount = {StorageAccount}
container = {container}
blobAccessKey =  {accessKey}
spark.conf.set("fs.azure.account.key."+blobStorage , blobAccessKey)
```

Der folgende Python-Codeausschnitt liest den Spark-Datenstrom. Der Code konfiguriert die Anzahl der pro Sekunde zu lesenden Zeilen und legt hierfür **4000** Zeilen pro Sekunde fest, für die Anzahl der Partitionen wird **4** festgelegt:

```
df = spark.readStream \
  .format("rate") \
  .option("rowsPerSecond", "4000") \
  .option("numPartitions", "4") \
  .load()
```

3. Sie schreiben nun die Streamdaten kontinuierlich in die Data-Warehouse-Tabelle:

```
df.writeStream \
 .format("com.databricks.spark.sqldw") \
 .option("url", sqlDwUrl) \
 .option("tempDir", "wasbs://"+container+"@"+stAccount+"/tmpdir/twts") \
 .option("forwardSparkAzureStorageCredentials", "true") \
 .option("dbTable", "TweetsStream")  \
 .trigger(processingTime="5 seconds") \
 .start()
```

   Der vorherige Codeausschnitt speichert den Datenstrom vorübergehend in dem angegebenen Azure Storage-Container, anschließend wird PolyBase die Daten aus dem temporären Container in die Azure Synapse Analytics-Zieltabelle (**`TweetsStream`**) aufnehmen. Für den Code ist festgelegt, dass er alle **`5 Sekunden`**ausgelöst werden soll.

Power BI

Rolle im Design

Das Team von Contoso muss seine Erkenntnisse sowie einige der Rohdaten visualisieren und dem Unternehmen kommunizieren. Dies ist wichtig, um eine Einbeziehung der Stakeholder des Unternehmens zu gewährleisten und schnell und einfach Feedback zu erhalten. Contoso benötigt auch eine Plattform, die es den Anwendern ermöglicht, Self-Service-Berichte und Dashboards zu verwenden und Daten selbst zu erkunden.

Diese Aufgabe erfüllt Power BI. Contoso kann mit Power BI Daten unter Verwendung einer Vielzahl von Visuals und Formen visualisieren, außerdem können geschäftliche und nicht technische Anwender den Berichterstellungs- und/oder Datenanforderungen selbst nachkommen.

Vorteile von Power BI

Power BI ist ein Business-Intelligence-**SaaS-Angebot** (**Software-as-a-Service**), mit dem Contoso seine Daten schnell und einfach in interaktive Visuals und Dashboards transformieren kann. Contoso wählt Power BI nicht nur aufgrund seiner Visualisierungsfunktionen, sondern auch in dem Bemühen, die Zusammenarbeit und den Self-Service für Daten und Berichte zu verbessern – Kernfeatures des Power BI-Diensts. Das Datenteam von Contoso hat die Gründe dafür zusammengefasst, warum Power BI für diesen Zweck geeignet ist:

- Power BI ist ein vollständig verwaltetes SaaS-Angebot, somit muss das Team von Contoso weniger Infrastruktur verwalten.
- Power BI hat die Datenvisualisierung und Berichterstellung vereinfacht. Mit Power BI kann jeder Anwender bei Contoso zum Datenanalysten werden. Die User Experience in Power BI ist ein großer Vorteil der Plattform. Anwender können Daten und interaktive Dashboards selbst erkunden. Contoso hofft, dass sich damit der Verwaltungsaufwand für sein Datenteam verringern lässt, weniger Datenanfragen eingehen und die Zusammenarbeit sowie die Interaktion mit geschäftlichen Anwendern verbessert werden können.
- Power BI bietet eine Desktopanwendung, die von den Anwendern bei Contoso zum Erkunden und Bereinigen von Daten sowie zum Erstellen von Visuals verwendet werden kann. Dies ist für Contoso sehr interessant, da die Nutzung der Power BI-Desktop-App kostenfrei ist und keine kommerzielle Lizenz erfordert. Darüber hinaus ist die User Experience in der Power BI-Desktop-App und dem cloudbasierten Power BI-Dienst sehr ähnlich. Dies bedeutet weniger Schulungen und einen einfacheren Wissenstransfer.
- Power BI verfügt über eine native Integration in Azure Active Directory, sodass Anwender bei Contoso ihre vorhandenen Identitäten verwenden können. Dies vereinfacht die Bereitstellung, verbessert die Governance und erhöht die Sicherheit. Darüber hinaus hat Power BI viele Compliance-Zertifizierungen erhalten und ist in vielen Regionen weltweit verfügbar, darunter auch Australien, wo Contoso seinen Sitz hat.
- Contoso verfügt über klar definierte Richtlinien für Branding und Werbung. Das bedeutet, dass alle Visualisierungen und Dashboards der Farbgestaltung von Contoso, Best Practices usw. entsprechen müssen. Contoso ist der Ansicht, dass dies das Branding und die Verständlichkeit der Informationen verbessert, da die Berichte konsistent und den Anwendern vertraut sind. Power BI unterstützt diese Anforderung mit einer Reihe von Features, wie z. B. anpassbaren Designs, anpassbaren Layouts und anderen benutzerdefinierten Visuals.
- Power BI bietet eine sofortige Integration in Azure. So kann Contoso direkt lokal mit der Datenvorbereitung und -transformation beginnen und bei Bedarf auf Azure skalieren. Darüber hinaus verfügt Power BI über eine native Integration in Azure AI Services. Somit kann Contoso, KI- und Machine-Learning-Funktionen einbinden, um schneller und leichter Mehrwert zu bieten – all das innerhalb von Power BI.

- Power BI Composite Models bieten Contoso die Möglichkeit, umfangreiche Berichte zu erstellen, die Daten aus mehreren Quellen abrufen. Mithilfe von Composite Models kann Contoso Datenverbindungen aus mehr als einer DirectQuery nahtlos einbeziehen oder die Datenverbindung in einer beliebigen Kombination importieren. Dies vereinfacht Datenverbindungen von Berichten mit den Datenquellen und unterstützt Contoso bei der Entwicklung komplexer Datenmodelle durch Kombination mehrerer Quellsysteme sowie bei der Verknüpfung von Tabellen aus verschiedenen Datasets. Darüber hinaus kann die Verwendung der Speichermodusfunktion von Composite Models in Power BI Premium zu einer Leistungsverbesserung beitragen und die Back-End-Auslastung verringern. Dies liegt daran, dass Power BI dem Autor eines Berichts die Möglichkeit gibt, festzulegen, welche Visuals Back-End-Datenquellen erfordern (oder nicht erfordern). Die Visuals, die keine kontinuierlichen Aktualisierungen von Back-End-Datenquellen erfordern, werden dann von Power BI (zwischen)gespeichert. Dies wiederum verbessert die Leistung und verringert die Auslastung von Back-End-Systemen.

Beispielimplementierung

Hier sehen Sie ein Beispiel für die Dashboards von Contoso. Die folgenden Berichte sollen die Leistung von Contoso in Bezug auf Produktverkäufe in den jeweiligen Kategorien, den aktuellen Bestand der meistverkauften Produkte, die Umsatzzahlen pro Jahr und die regionale Verteilung der Kunden von Contoso kommunizieren. Der Bericht wurde anhand von Beispieldaten erstellt, die von Microsoft bereitgestellt werden:

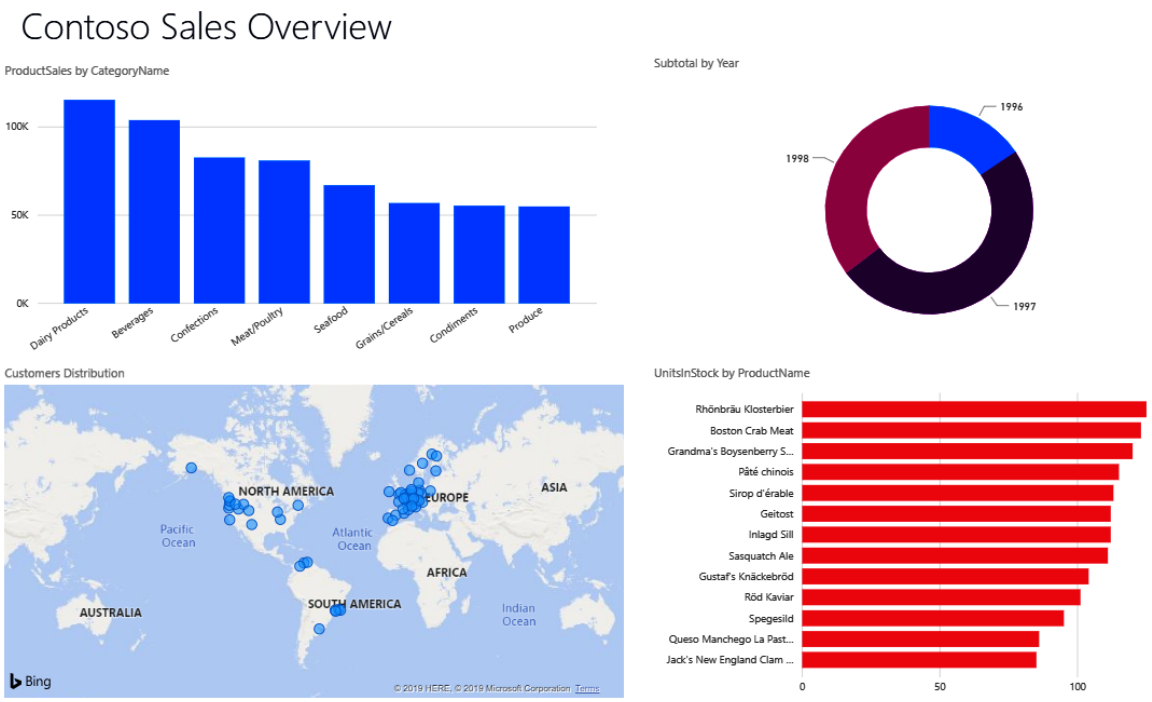

Abbildung 5.3: Beispiel für Dashboards von Contoso zum Geschäftsergebnis

Azure-Unterstützungsdienste

Neben allen bislang behandelten Azure-Diensten benötigt Contoso eine Reihe weiterer Dienste, die diese Lösungsarchitektur unterstützen und möglich machen. Diese Dienste sind im Abschnitt **Supporting Services** (Unterstützende Dienste) des Lösungsdesigns in *Abbildung 5.1* dargestellt. In diesem Abschnitt wollen wir diese Azure-Dienste kurz beschreiben.

Azure Automation

Contoso verfügt über eine Reihe von Datenbankservern, Testcomputern und anderen unterstützenden Servern. Mithilfe von Azure Automation kann Contoso die Konfiguration und Installation von Updates auf diesen Computern automatisieren. Azure Automation bietet Contoso die Möglichkeit einer konsistenten Verwaltung dieser Server und gewährleistet mithilfe von serverlosen Runbooks die Sicherheit und Compliance. Dies vereinfacht die Abläufe und den Verwaltungsaufwand. Dadurch hat das Team von Contoso wieder mehr Zeit, sich auf den wichtigeren Punkt zu konzentrieren, nämlich Insights zu gewinnen und Mehrwert für das Unternehmen zu schaffen.

Azure Key Vault

In jedem Unternehmen gibt es viele Verschlüsselungsschlüssel, Kennwörter, Zertifikate, Verbindungszeichenfolgen und andere sensible Daten, die gut gesichert werden müssen. Contoso ist sich bewusst, dass diese sensiblen Informationen geschützt und auf sichere und gut organisierte Weise verwaltet werden müssen. Azure Key Vault wurde entwickelt, um genau dieses Problem zu lösen. Alle sensiblen Daten werden an einem zentralen Ort geschützt, an dem eine sichere Zugriffsverwaltung möglich ist und Schlüssel organisiert werden können. Azure Key Vault verbessert nicht nur die Sicherheitskontrollen, sondern vereinfacht auch operative Aufgaben wie Zertifikatswechsel und Schlüsselrotation.

Darüber hinaus möchte Contoso Azure Key Vault verwenden, damit einzelne Anwender und Anwendungen keinen direkten Zugriff auf Schlüssel mehr benötigen. Contoso kann mithilfe der verwalteten Dienstidentität von Azure dafür sorgen, dass Anwender und Anwendungen Schlüssel und Kennwörter verwenden können, ohne lokale Kopien auf ihren Computern speichern zu müssen. Dies verbessert den Sicherheitsstatus von Contoso insgesamt und optimiert gleichzeitig die Verwaltung von Geheimnissen.

Azure DevOps

Azure DevOps bietet Contoso Tools, Frameworks und Dienste, um ein flexibles Verfahren zur Bereitstellung Ihrer Lösung durchzuführen. Azure DevOps ist ein vollständig verwalteter Dienst, der dem Team von Contoso folgende Möglichkeiten bietet:

- Planen, Nachverfolgen, Diskutieren und Überwachen von Arbeitsaufgaben mit Azure Boards. Contoso setzt bereits flexible Verfahren ein, bei denen derzeit physische Barrieren zum Nachverfolgen von Arbeitsaufgaben eingesetzt werden. Contoso stellt jedoch fest, dass physische Barrieren nicht für größere Teams skaliert werden können und in ihrer Funktionalität eingeschränkt sind. So kann Contoso beispielsweise Azure Boards verwenden, um Fehler und Arbeitsaufgaben mit Codeänderungen zu verknüpfen, um die Codequalität zu überwachen und zu verbessern.

- Kontinuierliches Entwickeln, Testen und Bereitstellen von Codeänderungen mithilfe von Azure Pipelines. Mit Azure Pipelines werden flexible Verfahren wie kontinuierliche Integration und kontinuierliche Bereitstellung möglich. Dies kann die Bereitstellungsqualität und -geschwindigkeit erheblich verbessern. Mit Azure Pipelines kann Contoso zudem alle Bereitstellungsschritte automatisieren, die für die Push-Übertragung von Codeänderungen erforderlich sind. Dies verringert den Aufwand und stärkt das Vertrauen in die neuen Bereitstellungen.
- Contoso benötigt ein Quellcodeverwaltungssystem zum Hosten von Code und Skripts. Contoso möchte hierfür Azure Repos verwenden, da dies Support auf Unternehmensniveau, eine unbegrenzte Anzahl von Repositorys und eine Umgebung für die Zusammenarbeit bietet, in der das Entwicklungsteam Codeänderungen vor dem Zusammenführen diskutieren und prüfen kann.
- Azure Test Plans kann Contoso bei der Validierung und Überprüfung von Code- und Datenänderungen unterstützen, um Contoso mehr Sicherheit bei Änderungen zu geben, bevor diese zusammengeführt werden. Contoso kann Azure Test Plans für manuelle und explorative Tests verwenden. Da Azure Test Plans Teil von Azure DevOps ist, bietet sich Contoso eine hervorragende Möglichkeit der End-to-End-Ablaufverfolgung von Geschichten, Features und Fehlern.

Azure Active Directory

Contoso verwendet Microsoft Office 365 für die Office-Zusammenarbeit, folglich nutzt Contoso Azure Active Directory bereits. Contoso möchte nicht mehrere Identitätsserver verwalten müssen und ist sich bewusst, dass die Verwaltung von Anwendernamen und Kennwörtern eine enorme Aufgabe ist, die besser einem gut ausgestatteten Team wie dem Active Directory-Team überlassen wird. Azure Active Directory verfügt über eine Integration in vielen der Dienste, die Contoso verwenden möchte, wie z. B. SQL Server, Azure Synapse Analytics, Azure Data Lake Storage und Power BI. Daher entscheidet sich Contoso ganz klar für Azure Active Directory, um eine einfache und problemlose Anmeldung bei allen diesen Diensten zu ermöglichen und gleichzeitig die Sicherheitskontrollen für die Daten und Anwendungen von Contoso zu verbessern.

Contoso kann auch von dem umfassenden Identitätsschutz von Azure Active Directory profitieren, der Bedrohungserkennung und -reaktion umfasst. Insgesamt kann Contoso mithilfe von Azure Active Directory den Aufwand erheblich reduzieren und die Sicherheit verbessern.

Azure Monitor

Contoso erkennt, dass die Verfügbarkeit und Leistung seiner Datenplattform von größter Bedeutung ist, um das Vertrauen aller Stakeholder zu gewinnen. Um dies zu erreichen, muss Contoso nicht nur Telemetriedaten aus allen Lösungsbereichen erfassen und speichern, sondern auch Daten analysieren und entsprechende Maßnahmen ergreifen. Zu diesem Zweck wird ein dedizierter Dienst benötigt, da die Implementierung eine große Herausforderung darstellt. Genau hierfür wurde Azure Monitor entwickelt.

Azure Monitor ist ein vollständig verwalteter Dienst, mit dem Contoso Daten aus allen Komponenten der Datenplattform (einschließlich Azure-Diensten, virtuellen Maschinen, Netzwerkleistung und anderen Quellen) einfach und schnell erfassen und analysieren und entsprechende Maßnahmen ergreifen kann. Azure Monitor bietet zwei grundlegende Arten von Daten: Log-Dateien und Metriken. Contoso kann Metriken verwenden, um Informationen über den Zustand seiner Dienste zu einem bestimmten Zeitpunkt zu erhalten, wogegen Log-Dateien dem Contoso-Team helfen, Ablaufverfolgungsmeldungen von den einzelnen Lösungskomponenten zu analysieren und zu visualisieren. Azure Monitor bietet zudem eine große Anzahl von Diagrammen und Visualisierungen, mit deren Hilfe Contoso den Zustand des Systems auf einen Blick visualisieren kann. Schließlich kann Contoso mit Azure Monitor Aktionen wie Warnungen auslösen, wenn bestimmte Bedingungen erfüllt sind. (So kann das Contoso-Team beispielsweise per E-Mail oder SMS benachrichtigt werden, wenn die Anzahl der Fehler einen bestimmten Schwellenwert überschreitet.)

Insights und Aktionen

Mit Microsoft Azure war das Datenteam von Contoso in der Lage, die Lösung schnell und einfach zu entwerfen, zu entwickeln und bereitzustellen. Innerhalb von zwei Wochen konnte das Team eine Reihe wichtiger Insights gewinnen, mit deren Hilfe Contoso seine Gewinnmarge steigern kann. Drei dieser Insights sind unten aufgeführt.

Verringerung der Abfälle um 18 %

Beschreibung: Mit der ursprünglichen Modellierung konnte das Datenteam von Contoso die Abfälle um 18 % reduzieren. Derzeit verliert das Unternehmen fast 46 Millionen USD pro Jahr durch den Überbestand an Produkten mit kurzer Haltbarkeitsdauer. Hierzu gehören Produkte wie Brot und Milch. Das Team kombinierte historische Verkaufsdaten mit anderen Quellen wie Wetterdaten und Schulkalendern und konnte so die Nachfrage nach diesen Produkten mit höherer Genauigkeit prognostizieren. Auf diese Weise konnten die Abfälle deutlich reduziert werden.

Geschätzter Geschäftswert: 8,28 Mio. USD/Jahr

Wichtige Datenquellen: Verkaufstransaktionen (Onlineshop und Ladengeschäft), Lagerdaten (Filialstandorte und Lagerbestände im Laufe der Zeit), Wetterdaten, Vorortprofildaten, Schulkalender und Feiertagskalender

Aktionen: Die Stakeholder von Contoso waren sehr beeindruckt und wollten eine schnelle Bereitstellung. Mit Azure Synapse Analytics und Power BI konnte das Datenteam von Contoso die Lösung schnell bereitstellen, damit sie von den Filialleitern verwendet werden konnte. Somit verfügen die Filialleiter von Contoso nun über ein interaktives Dashboard, das Verkäufe präzise prognostizieren und Empfehlungen in Bezug auf den für jedes Produkt notwendigen Lagerbestand geben kann.

Datenpipeline: Hier sehen Sie die vereinfachte Datenpipeline für diese Initiative:

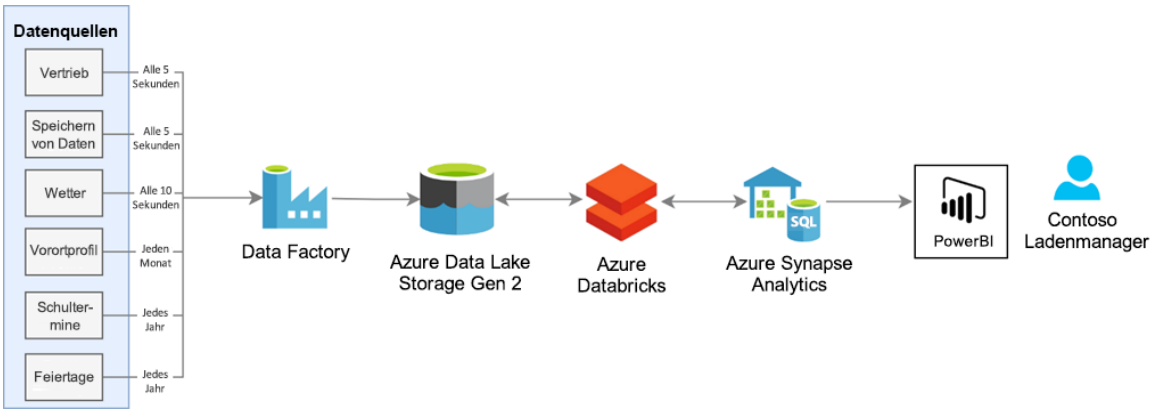

Abbildung 5.4: Datenpipeline für Initiative 1 (Abfallreduzierung)

Social-Media-Trends sorgen für eine Erhöhung der Verkäufe um 14 %

Beschreibung: Das Datenteam von Contoso stellte eine Hypothese auf, wonach Social-Media-Trends die Verkäufe erhöhen können. Das Team führte eine erste Datenermittlung durch, um festzustellen, dass ein solches Muster tatsächlich vorhanden ist. Interessanterweise fand das Team heraus, dass eine zeitliche Abstimmung der Marketingaktivitäten von Contoso auf Social-Media-Trends zu einer Erhöhung der Verkäufe um 14 % beitragen kann.

Ein klares Beispiel hierfür fand das Team in den Daten für den Sommer 2017 in Australien. Im Januar 2017 gab es einen sehr starken Social-Media-Trend im Zusammenhang mit gesundem Essen. Dies wurde nicht von Contoso organisiert. Über 4,5 Millionen Australier twitterten, likten, teilten oder kommentierten auf Twitter und Facebook Beiträge mit dem Hashtag **#BeHealthy**. Zufällig gab es bei Contoso eine Marketingkampagne für Obstsalatprodukte. Das Team stellte fest, dass diese Marketingkampagne außergewöhnlich erfolgreich war und den Umsatz um mehr als 25 % erhöhte, was deutlich über dem erwarteten Durchschnitt von 5–10 % lag.

Geschätzter Geschäftswert: 15,4 Mio. USD/Jahr (basierend auf einer Steigerung um 14 % bei den betreffenden Produkten)

Wichtige Datenquellen: Social-Media-Feeds (Twitter, Facebook und Instagram), Verkaufstransaktionen (Onlineshop und Ladengeschäft), Lagerdaten (Filialstandorte und Lagerbestände im Laufe der Zeit) sowie Daten aus Marketingkampagnen

Aktionen: Nach Besprechung der Ergebnisse mit dem Marketingteam von Contoso war man sich einig, dass Contoso den Erfolg der Werbekampagne vom Januar 2017 durch eine Überwachung und Ausrichtung seiner Werbeaktionen auf Social-Media-Trends wiederholen könnte. Das Datenteam von Contoso implementierte die Datenpipeline, wie im folgenden Screenshot dargestellt, und stellte sie als interaktives Echtzeit-Dashboard bereit, um sowohl das Marketingteam von Contoso als auch die Filialleiter zu informieren.

Datenpipeline: Hier sehen Sie die vereinfachte Datenpipeline für diese Initiative:

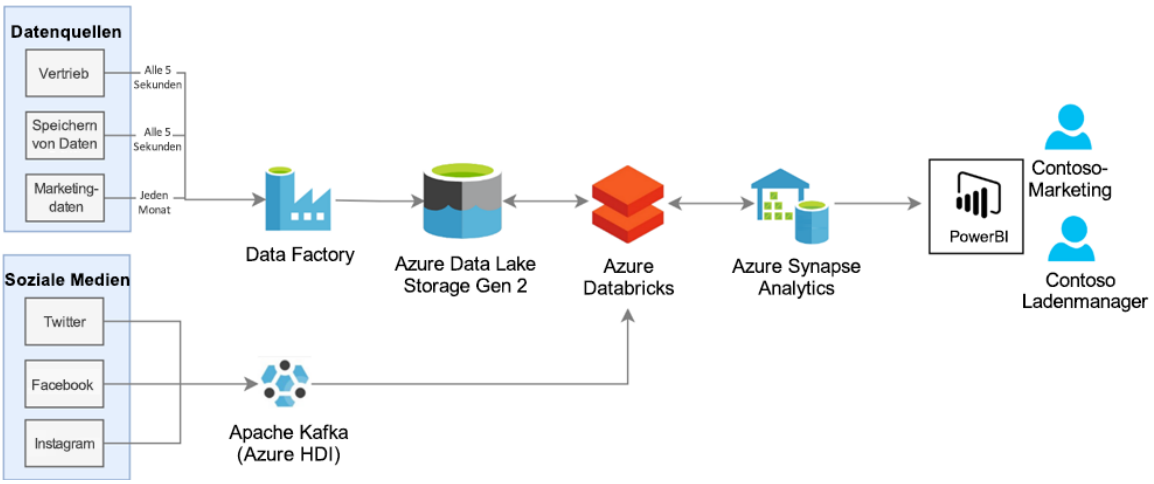

Abbildung 5.5: Datenpipeline für Initiative 2 (Erhöhung der Verkäufe)

Fazit

Sie haben gesehen, wie Contoso (oder jedes andere Unternehmen) von Microsoft Azure profitieren kann, um nahezu in Echtzeit Customer Insights zu gewinnen und Mehrwert zu schaffen. Microsoft Azure bietet eine breite Palette von Diensten für die Datenverwaltung und Analytics und zielt darauf ab, den Entwicklungsprozess zu optimieren und gleichzeitig die Messlatte für Qualität und Leistung anzuheben.

Darüber hinaus bietet Microsoft Azure viele Daten- und Analytics-Dienste als vollständig verwaltete Angebote. Dies bedeutet für Contoso (und jedes andere Unternehmen) einen geringeren Aufwand. Der andere Vorteil der Verwendung von serverlosem Azure besteht darin, dass Unternehmen und Teams ohne größere Investitionen klein anfangen und mit steigendem Bedarf skalieren können. Es handelt sich hier um einen großartigen Geschäftsansatz, da die mit Vorabinvestitionen verbundenen Risiken verringert werden und gleichzeitig der Verwaltungsaufwand in Zusammenhang mit einer bürokratischen Genehmigung hoher Ausgaben zu Beginn eines Datenprojekts reduziert wird.

Schließlich stellt Microsoft Azure eine Vielzahl von Dokumentationen und Lernmaterialien online zur Verfügung. Mit einer kostenfreien Gutschrift, die die Teams oder Anwender einsetzen können, um mit dem Lernen und der Entwicklung mit Azure zu beginnen, soll die Einstiegsbarriere beseitigt werden.

Anwendungsfalll 2: Verwenden von Advanced Analytics auf Azure zur Schaffung eines intelligenten Flughafens

Najad ist eine Großstadt im nördlichen Teil Ägyptens. Am Hauptflughafen der Stadt, dem Najad International Airport (NIA), werden jährlich 25 Millionen Passagiere abgefertigt, dies sind bis zu 70.000 Passagiere täglich. Es ist der verkehrsreichste Flughafen Ägyptens mit durchschnittlich 200.000 Flügen pro Jahr.

Das Management des NIA hofft, Daten-Analytics auf Azure einführen zu können, um die Kapazitätsplanung und Servicequalität zu verbessern. Ziel ist es, mithilfe von Daten betriebliche Probleme zu lösen, die den Flughafen zurzeit daran hindern, seine Infrastruktur und Ressourcen komplett auszuschöpfen. Dies wird wiederum die Kundenzufriedenheit verbessern und NIA in die Lage versetzen, den Betrieb durch Abfertigung von mehr Passagieren und Flugzeugen zu skalieren.

In den folgenden Abschnitten geht es darum, die Probleme zu definieren, mit denen NIA konfrontiert ist, und nach Ideen für Lösungsdesigns zu suchen. Schließlich werden Sie eine mögliche Lösungsarchitektur auf Microsoft Azure erstellen, die dieses Problem lösen kann und zeigt, warum Azure die perfekte Plattform für solche Lösungen ist.

Das Problem

Um das geschäftliche Problem angemessen zu definieren, müssen Sie zunächst die Herausforderungen aus der Geschäftsperspektive betrachten. Anschließend werden Sie sich die technischen Probleme ansehen, die Verbesserungen des Flughafens behindern.

Geschäftliche Herausforderungen

Wie bereits erwähnt, fertigt NIA jährlich mehrere zehn Millionen Passagiere ab. Die Zahl der Fluggäste soll in den nächsten 3 – 5 Jahren voraussichtlich um etwa 20 % steigen. Im vergangenen Jahr erlitt der Flughafen aufgrund betrieblicher Ineffizienzen einen Verlust von mehr als 370 Millionen USD. Hierzu gehörten Kosten durch Flugverspätungen aufgrund von Überlastung und langen Warteschlangen, entgangene Chancen im Einzelhandel aufgrund schlechter Erfahrungen der Fluggäste, Kosten durch schlechte Personalplanung und unzureichende Auslastung der Flughafenressourcen.

Die CIO von NIA, Zara Hassan, ist relativ neu (sie ist erst seit 6 Monaten bei NIA). Sie kommt aus dem Daten- und Business-Intelligence-Bereich. Zara hat eine Vision, wie die enorme Ineffizienz bei NIA in eine Geschäftschance umgewandelt werden kann. Sie stellt dem Vorstand von NIA einen Business Case vor. Es geht darum, kleine inkrementelle Investitionen in Advanced Analytics zu tätigen, um die Gesamtbetriebskosten des Flughafens zu reduzieren und gleichzeitig die Customer Experience zu verbessern.

Als Visionärin versteht Zara, dass der Flughafen von der Beobachtung historischer Berichte zu Zukunftsprognosen übergehen muss, um erfolgreich zu sein. Sie möchte, dass ihr Team dem Flughafenmanagement hilft, Flugverspätungen zu prognostizieren und solche Vorfälle zu reduzieren. Wenn das Flughafenmanagement Zugriff auf die richtigen Tools hat, können Kapazitätsplanung, Ressourcenzuweisung und Sicherheit ihrer Ansicht verbessert werden.

Der Vorschlag besteht darin, Daten und künstliche Intelligenz zu verwenden, um Passagiere, Flüge, Gepäck, Ressourcen und andere Datasets zu modellieren, um das Passagieraufkommen und die Bewegung der Menschenmassen zuverlässig vorhersagen zu können. Dadurch wiederum kann der Flughafen seine Abläufe verbessern und die Kosten senken.

Die geschäftlichen Herausforderungen für das Daten-Analytics-Team von NIA lassen sich wie folgt zusammenfassen:

- Die erste große Herausforderung für das Flughafenmanagement besteht darin, die Kapazitätsplanung zu verbessern. Derzeit trifft das Führungsteam von NIA diese Entscheidungen auf der Grundlage von Annahmen und früheren Erfahrungen, die nicht unbedingt die Realität widerspiegeln. Bislang hatte NIA keinen einheitlichen, datengestützten Ansatz, um die Zahl der an einem bestimmten Tag zu erwartenden Fluggäste zu prognostizieren. Eine genaue Vorhersage der erwarteten Fluggastzahl ist entscheidend für die Kapazitätsplanung, wie z. B. die Verwaltung des Personalbestands und den Kauf von Equipment sowie für die Planung von Upgrades der Infrastruktur. Darüber hinaus steht NIA keine Lösung zur Verfügung, um vorherzusagen, bei welchen Fluggesellschaften es Verspätungen geben könnte oder wie viele Sicherheitsmitarbeiter der Flughafen an einem bestimmten Tag zur Abfertigung der Passagiere benötigen würde. Die Folgen: Überfüllung, langen Warteschlangen und ineffiziente Nutzung der Infrastruktur. Allein die schlechte Kapazitätsplanung soll NIA im letzten Jahr Schätzungen zufolge fast 160 Millionen USD gekostet haben. Hinzu kamen neue Anschaffungen wie Fahrzeuge und Gepäckwagen, die der Flughafen aufgrund einer vermeintlichen Notwendigkeit tätigte, wobei es ausgereicht hätte, die vorhandenen Mittel besser zu nutzen.

- Die Ressourcenzuweisung ist ein weiteres wichtiges Anliegen des Managements von NIA. Die Passagiere müssen am Flughafen sehr lange warten, ob am Zoll oder an den Check-in-Schaltern der Fluggesellschaften. Die langen Wartezeiten sind zumeist auf die schlechte Zuteilung von NIA-Mitarbeitern auf die verschiedenen Bereiche des Flughafens zurückzuführen. Das Management von NIA möchte die Ressourcenzuweisung verbessern und damit eine bessere Kundenzufriedenheit erzielen.

- Die Einzelhandelsgeschäfte und Duty-Free-Shops am Flughafen haben einen ordentlichen Anteil am Umsatz des Flughafens. NIA verfügt über eine Reihe großer Werbeflächen und verwendet Kundeninformationen für gelegentliche Werbeaktionen. Das Management von NIA möchte die Interaktion mit Kunden und letztendlich die Geschäftschancen in den Einzelhandelsgeschäften am Flughafen verbessern.

- Ein Großteil des Kundenservice besteht darin, Kunden die benötigten Informationen zum richtigen Zeitpunkt zur Verfügung zu stellen. Auf einem Flughafen unterwegs zu sein, kann sehr ermüdend und auch stressig sein, wenn sich Passagiere verspäten oder ihr Flug verspätet ist oder annulliert wurde. Daher muss NIA den Flugstatus/Verspätungen nahezu in Echtzeit aktualisieren. Um dies zu ermöglichen, muss das NIA-Management kreative innovative Möglichkeiten finden, den Kunden relevante Informationen zur Verfügung zu stellen, wenn diese sie benötigen. Für die Kunden bedeutet dies weniger Verwirrung und Stress und insgesamt einen besseren Kundenservice.
- NIA benötigt langfristig eine Überholung der Infrastruktur. Dies würde das Problem der Überlastung lösen, das in der Vergangenheit zu kleineren Unfällen geführt und den Flughafen Geld gekostet hat und sich zudem negativ auf die Customer Experience auswirkt. Als kurzfristige Lösung für die nahe Zukunft möchte NIA jedoch durch angemessene Nutzung von Ressourcen den Passagierfluss verbessern und die Überlastung verringern. Überlastungssituationen behindern den Passagierfluss und schaffen Sicherheitsrisiken, wenn zu viele Menschen gezwungen sind, kleine Hallen und/oder enge Gänge zu durchqueren. Dies gilt insbesondere für alte Menschen, Babys und Fluggäste mit Behinderungen. Es kommt zu Sicherheitsvorfällen. Jeder dieser Sicherheitsalarme und -vorfälle kostet den Flughafen Geld, bringt das Leben der Fluggäste in Gefahr und wirkt sich negativ auf die Customer Experience aus. Der Flughafen möchte den Passagierfluss verbessern, um die Überlastung zu reduzieren und die Sicherheit zu verbessern.

Nachdem Sie nun die wichtigsten Probleme kennen, die die Geschäftsseite von NIA angehen möchte, müssen Sie auch die technischen Herausforderungen berücksichtigen, damit Sie mit der Entwicklung einer Lösung beginnen können.

Technische Herausforderungen

Keine Single Source of Truth (einzige Quelle der Wahrheit): Ein Hauptproblem, das die CIO von NIA zu lösen versucht, ist die Tatsache, dass NIA in Bezug auf seine Datenquellen derzeit nicht über eine Single Source of Truth verfügt. Heute verlässt sich der Flughafen auf Berichte aus verschiedenen alten internen Systemen sowie auf Berichte von Partnern. Diese Berichte beziehen sich in der Regel auf betriebliche Aspekte des vorherigen Tags und der letzten Woche und enthalten widersprüchliche Zahlen. So werden Flugdaten zum Beispiel derzeit von den einzelnen Lufttransportgesellschaften gespeichert. Am Flughafen NIA gibt es mehr als 35 Fluggesellschaften, die jeweils über eigene Systeme verfügen und verschiedene Terminologien verwenden. Dies macht es für das NIA-Management extrem schwierig, rechtzeitig zuverlässige Berichte zu erhalten oder gar datengesteuerte Abläufe zu haben.

Latenz beim Erhalt von Daten und Berichten: Da NIA keine Kontrolle über Flug- und Frachtdaten hat, verlässt sich der Flughafen auf Partner, die Betriebsberichte generieren, aggregieren und senden. Diese Berichte sind in der Regel um Tage oder Wochen verspätet. Dies reduziert die Fähigkeit des Unternehmens, Insights aus diesen Berichten umzusetzen, und zwingt NIA dazu, in seinem Betrieb immer nachträglich zu reagieren, anstatt vorauszuplanen. Wenn dem Flughafenmanagement beispielsweise ein Bericht vorgelegt wird, aus dem hervorgeht, dass gestern lange Warteschlangen zu Flugverspätungen geführt haben, kann das Flughafenmanagement an dieser Situation nichts ändern, da sie in der Vergangenheit passiert ist. Der rechtzeitige Zugriff auf diese Daten ist für NIA und fast alle anderen Unternehmen von entscheidender Bedeutung.

Datenverfügbarkeit und -zugriff: Innovation erfordert die Erkundung von Möglichkeiten und das Experimentieren mit Optionen. In Bezug auf Daten bedeutet dies, dass NIA die Flug- und Passagierdaten kontinuierlich erkunden, anreichern und mit externen Datenquellen korrelieren muss. Leider ist NIA heute zu all dem nicht in der Lage, da sich die Daten in vielen Silosystemen befinden, die der Flughafen nicht kontrolliert.

Skalierbarkeit: NIA verfügt derzeit über ein SQL Data Warehouse, das im virtuellen Rechenzentrum des Flughafens gehostet wird. Das Managementteam war nicht gern bereit, in dieses Data Warehouse zu investieren, da es nicht alle Daten enthält. Dadurch ist das aktuelle Data Warehouse veraltet, da es dem Flughafen nicht hilft, die notwendigen Insights zu gewinnen. Darüber hinaus verfügt das aktuelle SQL Data Warehouse nicht über die Möglichkeit, alle Daten zu erfassen und/oder zu speichern, die NIA sammeln kann.

Sicherheit: NIA hat klare und strenge Richtlinien zum Schutz seiner Daten und aller Kundendaten. Der Flughafen muss die Zertifizierungen ISO/IEC 27001 und ISO/IEC 27018 erhalten, um sicherzustellen, dass die Sicherheitsmaßnahmen ordnungsgemäß zum Schutz des Flughafens, seiner Zulieferer, Kunden und aller Stakeholder angewendet werden. NIA muss all diese Sicherheitsanforderungen in einer potenziellen Lösung erfüllen.

Brauchbarkeit der Daten: Damit Daten nützlich sind, müssen sie den relevanten Anwendern zur richtigen Zeit zur Verfügung gestellt werden. NIA bietet derzeit Benachrichtigungen und Erinnerungen für Passagiere über Audioankündigungen und einige große Monitore, die an einigen wenigen Stellen am Flughafen platziert sind. Dies ist sehr ineffizient, da es Lärm verursacht und nicht berücksichtigt, wer der Anwender ist oder was der Anwender wissen möchte. NIA erkennt nun an, dass weitere Anstrengungen nötig sind, um nicht nur die Daten- und Berichtseffizienz, sondern auch die Art der Bereitstellung dieser Berichte für die Anwender zu verbessern.

Angesichts dieser Anforderungen kommt das Business-Intelligence-Team von NIA mit der CIO überein, die Problemstellung wie folgt zu definieren:

NIA macht Verluste in Höhe von mehr als 350 Mio. USD pro Jahr aufgrund betrieblicher Ineffizienzen. Hierzu gehören lange Warteschlangen, ein unzureichender Personalbestand und eine unzureichende Auslastung der Flughafenressourcen. Das Business-Intelligence-Team von NIA wird an der Bereitstellung von Data-Analytics-Tools (Dashboards, Berichte und Apps) arbeiten, die dem Unternehmen helfen, Prozesse zu optimieren und Ineffizienzen zu beseitigen.

Design-Brainstorming

Nachdem wir das Problem definiert und die geschäftlichen und technischen Herausforderungen formuliert haben, sollen Ihnen die nächsten Abschnitte helfen, einige Ideen für Lösungskonzepte zu entwickeln, um ein Lösungsdesign für NIA zu schaffen.

Datenquellen

Daten stehen im Mittelpunkt jeder Analytics-Lösung. Daher müssen Sie zunächst an die verschiedenen Arten von Daten denken, die NIA benötigt. Dann müssen Sie sich Gedanken über ein Design machen, um diese Daten zusammenzubringen. NIA muss Daten aus folgenden Quellen erfassen:

- **Zolldaten**: Zolldaten enthalten Informationen über Passagiere und ihre Zollerklärungen bei der Einreise oder dem Verlassen des Lands. Zurzeit werden die Zolldaten in externen Systemen gespeichert. Der Flughafen kann diese Daten jedoch abrufen und in seine Systeme integrieren. Der aktuelle Mechanismus der Integration im Zolldatensystem verwendet eine geplante Dateisicherung auf einem Dateiserver. NIA kann daraus Zolldaten auf seine neue Plattform übertragen.
- **Airline-/Flugdaten**: Derzeit speichern die einzelnen Airline-Systeme die Passagierdaten, ihre Reisen, Check-in-Zeiten und andere zugehörige Details. Obwohl diese Daten von den Airline-Systemen gespeichert werden, kann der Flughafen mithilfe von Integrations-APIs eine Integration in diese Systeme vornehmen. Die konkrete Implementierung dieser Integration hängt von den einzelnen Fluggesellschaften ab, der Flughafen muss diese Daten jedoch nahezu in Echtzeit abrufen.
- **Parksystemdaten**: Der Flughafen verfügt über Sensoren an allen Parkanlagen, die die ein- und ausfahrenden Autos zählen. Die Parksysteme verfügen auch über eine Anzeige, die zu jedem Zeitpunkt angibt, wie viele Parkplätze wo verfügbar sind. Diese Daten müssen mit anderen Quellen erfasst werden.

- **IoT und Videostreams**: NIA verfügt über eine Reihe von Verkehrsüberwachungskameras, die auf dem gesamten Gelände installiert sind. Diese Kameras senden Live-Video-Streaming und werden von der Leitstelle verwendet, um Ressourcen einzusetzen und betriebliche Abläufe anzupassen und so den Verkehr zu bewältigen. Auch am Flughafen sind in der Nähe der Gates IoT-Sensoren installiert, um den Status jedes Gates anzugeben. Es gibt auch Sensoren zur Überwachung der Verteilung der Menschenmassen am Flughafen. Daten aus allen diesen Quellen (IoT und Kameras) können für Echtzeitanalysen gestreamt werden, um dem Management von NIA umsetzbare Insights zu bieten, wenn Verkehrsprobleme auftreten.
- **Gepäcksystemdaten**: Der Flughafen verfügt über ein internes System zur Verwaltung aller Gepäckdaten. Dies beinhaltet Informationen darüber, welches Gepäck mit welchem Flug angekommen ist und wo sich das Gepäck jetzt befindet. Der Flughafen bedient auch Logistikunternehmen, jeden Tag kommen mehrere Frachtflüge an. Es ist wichtig, alle relevanten Daten zu erfassen und zu analysieren, die diesen Logistikunternehmen für das Frachtmanagement bereitgestellt werden.
- **Social-Media-Feeds**: Um einen guten Kundenservice zu bieten, ist es für NIA unerlässlich, die Stimmung und das Feedback der Fluggäste zu analysieren, da diese ihre Erfahrungen vermutlich auf Social-Media-Plattformen teilen werden. Auf diese Weise kann NIA seine Dienstleistungen verbessern und Probleme sofort beheben.
- **Andere Datenquellen**: Wie in *Anwendungsfall 1* erörtert, ist es bei Daten-Analytics durchaus üblich, vorhandene Datasets mit anderen externen Datenquellen anzureichern, um gefundene Trends oder Muster in einen größeren Kontext zu stellen. Dies gilt insbesondere für den Flughafenbetrieb, wo Wetterdaten, Urlaubszeiten und andere Faktoren einen starken Einfluss haben können. NIA muss viele dieser externen Datenquellen erfassen, um seine eigenen operativen Daten zu ergänzen.

Datenspeicher

Der Flughafen schätzt sein aktuelles Datenvolumen auf annähernd 310 TB. Dabei sind die Daten aller Partner, die erfasst und gespeichert werden müssen, nicht einbezogen. Hinzu kommt, dass der Flughafen auch Kamera-Streamingdaten und Social-Media-Feeds abrufen möchte. Damit könnten den Zahlen aus der Vergangenheit zufolge weitere 15 GB Daten pro Tag hinzukommen. Dies erfordert einen hochgradig skalierbaren Datenspeicherdienst, der elastisch an die schnell wachsenden Mengen angepasst werden kann.

Um diese Anforderung zu erfüllen, ist es sinnvoll, einen cloudbasierten Dienst wie Azure Data Lake Storage zu verwenden. Auf diese Weise lassen sich eine elastische Skalierbarkeit und die Möglichkeit zum Speichern von Daten in verschiedenen Formaten sicherstellen.

Datenerfassung

Damit die Daten den Mitarbeitern und Kunden des Flughafens zeitnah zur Verfügung gestellt werden können, müssen sie schnell und effizient aus internen und externen Quellen erfasst werden. Entsprechend den bereits erörterten Datenquellen muss die Lösung mehrere Formen der Datenerfassung ermöglichen. Hierzu gehören das Laden von Dateisicherungen, die Verarbeitung von Echtzeitdatenströmen aus Social Media und Überwachungskameras sowie das Abrufen von Daten durch den Aufruf externer APIs. Das Datenteam bei NIA kann entweder eine eigene Lösung für die Integration und Erfassung entwickeln, was sehr teuer wäre und sehr viel Entwicklungszeit erfordern würde, oder ein cloudbasiertes Datenerfassungs- und Orchestrierungstool wie Azure Data Factory (ADF) verwenden. ADF optimiert den Datenerfassungsprozess durch die Bereitstellung von mehr als 80 vorgefertigten Datenkonnektoren, die für die Integration in eine Vielzahl von Quellsystemen wie SQL Databases, Blob-Speicher und Flatfiles verwendet werden können.

Sicherheit und Zugriffssteuerung

Die Lösung muss die richtigen Sicherheitskontrollen für das Management von NIA bieten, damit die Daten gesichert und geschützt werden können. Verständlicherweise gibt es am Flughafen eine lange Liste von Stakeholdern, die Zugriff auf die Daten benötigen. Hierzu gehören Flughafenmitarbeiter, Sicherheitsunternehmen, das Flugpersonal, Passagiere und Partner. Daher muss NIA mit der Lösung Sicherheit auf Zeilenebene bieten können, um sicherzustellen, dass die Anwender nur Zugriff auf ihre jeweiligen Daten haben. Hierfür wird ein Verwaltungssystem mit differenzierter Zugriffssteuerung benötigt, das in die ausgewählte Plattform integriert ist, sodass sich das Business-Intelligence-Team von NIA nicht zu lange mit Sicherheitsfragen befassen muss. Das Business-Intelligence-Team von NIA sollte sich darauf konzentrieren, hilfreiche Insights für das Flughafenmanagement zu gewinnen.

Erkennen von Mustern und Insights

Die Strategie von Zara besteht zu einem wesentlichen Teil darin, dem Unternehmen intelligente Entscheidungen zu ermöglichen. Die Informationen hierfür sollen durch Erkunden und Erkennen von Trends und Mustern in den Daten erhalten werden. Die größte Herausforderung stellt dabei die Frage dar, wo und wie diese Machine-Learning-Modelle entwickelt werden sollen. Um solche Modelle zu entwickeln, ist es notwendig, mit vielen großen Datasets und einem elastischen Pool von Compute-Ressourcen zu arbeiten. Das Team sieht die bevorstehende Herausforderung und möchte Azure Databricks, eine hochmoderne Analytics-Plattform, verwenden. Das Team war von der Skalierbarkeit und Sicherheit sowie der breiten Unterstützung von Tools und Frameworks auf Azure Databricks beeindruckt.

Die Lösung

Die CIO von NIA ist mit dem Business-Intelligence-Team übereingekommen, Microsoft Azure als Cloudanbieter zu verwenden, um die neue Lösung zu entwickeln. Die Begründung hierfür wurde wie folgt zusammengefasst.

Vorteile von Azure für NIA

- NIA nutzt bereits Microsoft-Technologien wie Windows 10, Office 365 und andere Tools. Azure bietet eine bessere native Integration in allen diesen Diensten als jeder andere Cloudanbieter. Daher ist es durchaus sinnvoll, Azure zu verwenden. Darüber hinaus ist NIA sehr daran interessiert, die **Open-Data-Initiative** (**ODI**) zu nutzen, die es Unternehmen ermöglicht, durch die Kombination von Verhaltens-, Transaktions-, Finanz- und Betriebsdaten in einem Datenspeicher außergewöhnliche Insights zu gewinnen. Die Initiative vereinfacht das Erstellen gemeinsamer Datenmodelle im gesamten Unternehmen und wurde von Adobe, Microsoft und SAP gemeinsam entwickelt.
- Durch die Verwendung von Azure kann NIA weiterhin denselben zentralisierten Identitätsserver verwenden, der mit Azure Active Directory für Office 365 verwaltet wird. Dies bedeutet eine bessere Sicherheit für NIA und einen geringeren Aufwand beim Erstellen und Verwalten neuer Benutzerkonten.
- Azure verfügt über mehr regionale Rechenzentren als andere große Cloudanbieter. Damit kann Azure eine höhere Resilienz und Serviceverfügbarkeit für NIA bieten. Darüber hinaus ist Azure der einzige Cloudanbieter mit einem regionalen Rechenzentrum in Afrika, wo sich NIA befindet. Dies macht Azure zur perfekten Wahl, da die Lösung alle Kriterien erfüllt.
- Das Business-Intelligence-Team von NIA fand heraus, dass Azure kostengünstiger als andere Cloudanbieter ist. Azure Synapse Analytics (früher als SQL Data Warehouse bezeichnet) ist 14-mal günstiger als AWS oder Google-Dienste, wie in *Anwendungsfall 1*erläutert. Darüber hinaus bietet Azure die Möglichkeit, reservierte Instanzen für virtuelle Maschinen sowie Compute-Instanzen zu verwenden, wodurch noch größere Rabatte möglich werden. Außerdem kann NIA mit seinem bestehenden Enterprise Agreement mit Microsoft sogar noch weitere Rabatte auf alle Einzelhandelspreise für Azure-Dienste erhalten. Somit ließe sich die Auswahl eines anderen Cloudanbieters nur schwer rechtfertigen.
- Zudem hat NIA die Erfolgsbilanz von Microsoft in Bezug auf Entwicklertechnologien und Entwickleroberflächen als großen Vorteil von Azure angesehen. Als Softwareentwicklungsunternehmen bietet Microsoft unter Nutzung seines großes geistigen Eigentums auf diesem Gebiet die beste Entwickleroberfläche. Mit Azure erhält NIA somit eine gute Entwicklungs- und Bereitstellungsumgebung.

- Azure hat mehr als 30 Compliance-Zertifizierungen erhalten, darunter auch **ISO/IEC 27001** und **ISO/IEC 27018**, die NIA ebenfalls benötigt. Hinzu kommt die Tatsache, dass das Geschäftsmodell von Azure nicht auf der Nutzung oder dem Verkauf von Kundendaten basiert, was bei anderen Cloudanbietern Teil des Geschäftsmodells ist. Dies gibt NIA und seinem Vorstand mehr Sicherheit, dass die Daten des Flughafens und seiner Kunden gut geschützt sind.
- Zudem hofft NIA, Azure Stack zu verwenden, eine Plattform, die dem Flughafen die Möglichkeit bietet, Anwendungen und Dienste problemlos unter Verwendung derselben Basisinfrastruktur On-Premises und in der Cloud zu hosten. Dies wird durch Azure und Azure Stack unterstützt.
- Schließlich möchte NIA in der Lage sein, eine Mischung aus **PaaS**, **SaaS** und Open Source Tools zu wählen. Azure macht genau das möglich: NIA kann damit großartige PaaS- und SaaS-Dienste wie Azure Data Lake, Azure Data Factory und mehr nutzen, gleichzeitig wird die native Integration in Open-Source-Dienste wie Databricks und Kafka unterstützt.

Lösungsarchitektur

Nachdem das BI-Team die Anforderungen präzisiert hat und eine Cloud Platform ausgewählt wurde, ist es nun an der Zeit, ein sicheres und skalierbares Design zu entwickeln. Das Business-Intelligence-Team von NIA entschied sich für die folgende Lösungsarchitektur:

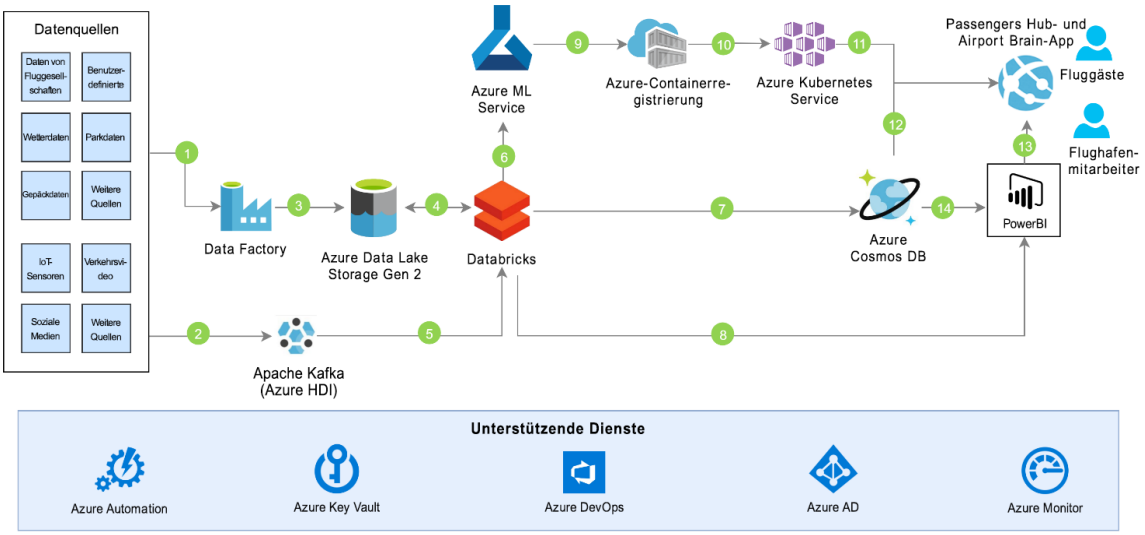

Abbildung 5.6: Lösungsarchitektur für NIA

Abbildung 5.6 zeigt die Lösungsarchitektur und den Datenfluss zwischen den einzelnen Komponenten. Im Folgenden werden die einzelnen Workflow-Segmente erläutert, wie in dem Diagramm markiert (nummeriert):

1. Strukturierte Daten wie beispielsweise Daten der Fluggesellschaften (**Airlines Data**), Zolldaten (**Custom Data**) und Gepäckdaten (**Baggage Data**) werden mithilfe von **Azure Data Factory** erfasst. Hierzu gehören auch andere Datenquellen, wie z. B. Daten aus den Parksystemen sowie Wetterdaten. Azure Data Factory bietet NIA die Möglichkeit, eine Integrationslaufzeit zu konfigurieren, die als Gateway verwendet werden kann, um von Azure aus eine Verbindung mit den On-Premises-Datenquellen von NIA herzustellen.

2. Alle unstrukturierten Daten, wie beispielsweise Daten von IoT-Sensoren (**IoT Sensors**), Verkehrsvideo-Streamingdaten (**Traffic Video**) und **Social-Media**-Feeds, werden mithilfe von **Apache Kafka** in einem **Azure HDInsight**-Cluster erfasst. Durch die Verwendung von Apache Kafka auf Azure HDInsight kann NIA eingehende Datenströme bei ihrer Ankunft filtern und transformieren, bevor sie erfasst werden.

3. **Azure Data Factory** überträgt die erfassten strukturierten Daten zum Speichern in **Azure Data Lake Storage Gen2**.

4. **Azure Databricks** wird als einheitliche Daten-Analytics-Plattform verwendet, damit die Datenwissenschaftler und Dateningenieure von NIA an der Bereinigung, Transformation und Ermittlung von Daten arbeiten können. Azure Databricks liest die erfassten Datasets aus Azure Data Lake Storage und verwendet Data Lake Storage auch zum Speichern von Ergebnissen und zusammengestellten Daten.

5. Der eingehende Datenstrom (von **Social Media** und **IoT-Sensoren**) wird von Kafka zu Apache Structured Streaming in Azure Databricks übertragen. So kann das Business-Intelligence-Team die eingehenden Daten in Azure Data Lake Storage speichern.

6. **Azure Machine Learning Services** wird verwendet, um die Machine-Learning-Modelle, Datasets, Experimente und neuen Modellbilder zu verwalten. Azure Machine Learning Services verfügt über eine native Integration in Azure Databricks.

7. Bereinigte Daten, die nutzungsbereit sind, werden dann von Azure Databricks in **Azure Cosmos DB** übertragen. Hierzu gehören die neuesten Flug- und Gepäckdaten, die zur Nutzung durch die Passagiere und Mitarbeiter des Flughafens verfügbar gemacht werden sollen.

8. **Power BI** ist zudem so konfiguriert, dass umfangreichere Datasets direkt von **Azure Databricks** gelesen werden. Zu den Daten, die mit diesem Mechanismus übertragen werden, gehören beispielsweise die Ergebnisse der Decodierung der Verkehrsvideo-Streamdaten, um Heatmaps der Menschenmengen zu erstellen.

9. Azure Machine Learning Services wird zum Trainieren und Entwickeln von Machine-Learning-Modellen verwendet. Die daraus resultierenden Modelle werden als Docker-Images in der **Azure Container Registry** gespeichert. Docker-Images werden in der Regel als Möglichkeit zum Verpacken von Machine-Learning-Modellen mit allen ihren Abhängigkeiten (Bibliotheken, Quellcode und Konfigurationsdateien) als einzelnes, bereitstellbares Paket verwendet. Dies verbessert den Entwicklungszyklus und reduziert Bereitstellungsfehler.

10. Mit der Bereitstellung ist **Azure Kubernetes Service** konfiguriert, um die neuen Machine-Learning-Modellbilder aus der **Azure Container Registry** zu verwenden und diese Modelle als Kubernetes-Pods auszuführen. Damit können die Machine-Learning-Modelle zum Generieren von Vorhersagen durch einfache HTTP-Aufrufe verwendet werden. Beispiele für Machine-Learning-Modelle sind ein Empfehlungsmodul für Gate-Zuweisungen und ein Machine-Learning-Modell für die Prognose der Parkplatznachfrage.

11. Das Business-Intelligence-Team von NIA kann die neuen Machine-Learning-Modelle über Webanwendungen bereitstellen, die auf **Azure Kubernetes Service** gehostet werden können. Diese Webanwendungen können dann mit **Azure Cosmos DB** interagieren, um Machine-Learning-Rückschlussdaten zu speichern (z. B. empfohlene Maßnahmen für das Flughafenpersonal) und zusammengestellte Daten wie Flugplan und Gepäckdaten bereitzustellen. Beispiele für diese Webanwendungen sind **Passengers Hub** und **Airport Brain**. Passagiere Hub ist als zentrales Portal zur Bereitstellung aller Passagierdaten konzipiert. Hierzu gehören beispielsweise Flugdetails, Gate-Nummern, Check-in-Schalter und Empfehlungen für die Passagiere. Die Passagiere können alle diese Informationen auf ihrem mobilen Gerät sehen, wenn sie die mobile App des Flughafens herunterladen.

12. Airport Brain ist der Name des neuen zentralen Portals für die Mitarbeiter des Flughafenmanagements. Ziel ist es, dem Management von NIA die Tools an die Hand zu geben, um einen effizienten Betrieb zu ermöglichen. Das Portal verwendet Daten, um Empfehlungen zu Gate-Zuweisungen, zum Personalbestand und zur Verteilung der Flughafenressourcen abzugeben.

13. Sowohl **Passengers Hub** als auch **Airport Brain** nutzen die zusammengestellten Daten, die in Azure Cosmos DB gespeichert sind. Azure Cosmos DB wird auch verwendet, um anwendungsspezifische Datentypen wie Sitzungen der Anwender und Benachrichtigungen zu speichern. Dies alles wird durch das extrem schnelle Abfragemodul von Azure Cosmos DB und eine hohe Reaktionsfähigkeit ermöglicht.

14. Sowohl **Passengers Hub** als auch **Airport Brain** erfordern eine Datenvisualisierung. Diese Berichte werden mithilfe von Power BI entwickelt. Anschließend wird das Power BI-Feature zum Einbetten in Webseiten verwendet, um diese Power BI-Berichte in den neuen Webanwendungen zu präsentieren. Die zusammengestellten Daten umfassen fluggastbezogene Informationen wie Flugdetails, prognostizierte Verspätungen und Informationen zum Gepäck der Passagiere.
15. Das Power BI-Dashboard stellt Daten aus Azure Cosmos DB bereit. Hierzu gehören zusammengestellte fluggastbezogene Daten, wie unter *Punkt 13* aufgeführt.

Azure-Dienste

Wie in *Anwendungsfall 1* werden wir uns in den folgenden Abschnitten eingehender mit den Azure-Diensten befassen, die in dem Lösungskonzept in *Abbildung 5.6* dargestellt sind. Zunächst erläutern wir, warum der betreffende Dienst benötigt wird und warum er für NIA geeignet ist. Abschließend zeigen wir ein kurzes praktisches Beispiel für den Kernteil der Implementierung. Um Wiederholungen zu vermeiden, werden die bereits in *Anwendungsfall 1* behandelten Azure-Dienste nicht mehr erörtert, es sei denn, NIA stellt spezifische Anforderungen an den betreffenden Dienst.

Azure Databricks

Rolle im Design

Azure Databricks dient als einheitliche Plattform zum Bereinigen, Transformieren, Zusammenführen und Erkunden von Daten. Azure Databricks wird benötigt, um die Rechenleistung bereitzustellen, die für die Verarbeitung von Daten und zur Förderung einer besseren Zusammenarbeit zwischen den vielen Stakeholdern benötigt wird.

Vorteile von Azure Databricks

Neben allen Vorteilen von Azure Databricks, die in *Anwendungsfall 1* erörtert wurden, bietet Azure Databricks Unterstützung für mehrere Sprachen und Machine-Learning-Frameworks (wie TensorFlow und PyTorch) und lässt sich mit vielen Open Source Tools integrieren. Das Business-Intelligence-Team von NIA benötigt eine Plattform, die sowohl Data-Engineering- als auch Data-Science-Workloads bewältigen kann. Azure Databricks ist für diesen Zweck konzipiert: Die Dateningenieure können Daten bereinigen, zusammenführen, transformieren und überprüfen, während die Datenwissenschaftler gleichzeitig eines der gängigen Machine-Learning-Frameworks wie TensorFlow oder PyTorch verwenden können.

Beispielimplementierung

Mit dem folgenden Codeausschnitt wird eine Verbindung mit Azure Cosmos DB von Azure Databricks unter Verwendung des Azure Cosmos DB-Connectors konfiguriert. Der folgende Python-Code enthält einen Platzhalter für den Hauptschlüssel der Azure Cosmos DB-Instanz und geht davon aus, dass NIA über eine Azure Cosmos DB-Instanz namens **`NIAAnalytics`** mit einer Sammlung namens **`flights_data`** verfügt. Der Code speichert einen DataFrame **`flights`** (Spark-DataFrame) in Azure Cosmos DB:

```
# Config to connect to Cosmos db
config = {
    "Endpoint": "https://NIAairport.documents.azure.com:443/",
    "Masterkey": "{masterKey}",
    "Database": "NIAAnalytics",
    "Collection": "flights_data",
    "Upsert": "true"

}

# Writing flights data from DataFrame to Azure Cosmos db
flightsDf.write.format("com.microsoft.azure.cosmosdb.spark").options(**config).
save()
```

Azure Cosmos DB

Rolle im Design

Azure Cosmos DB dient zwei Hauptzwecken: dem Speichern aller Anwendungsdaten für Anwendungen wie „Passengers Hub“ und „Airport Brain“ und der Bereitstellung zusammengestellter Daten, die vom Flughafenpersonal und externen Stakeholdern (wie z. B. Passagieren) genutzt werden können.

Vorteile von Azure Cosmos DB

Es gibt viele Möglichkeiten, die zusammengestellten Daten und Anwendungsdaten von NIA zu speichern. Das Business-Intelligence-Team von NIA hat sich jedoch aus folgenden Gründen für Azure Cosmos DB entschieden:

- Azure Cosmos DB bietet eine sofort einsatzbereite globale Verteilung. Dies ist ideal, um die Verfügbarkeit und Resilienz der NIA-Plattform zu gewährleisten. Verständlicherweise kann sich NIA keine Ausfallzeiten leisten, da das ganze Jahr Millionen von Passagieren abgefertigt werden müssen. Daher muss die neue Plattform hochverfügbar sein. Dies kann von Azure Cosmos DB unterstützt werden.
- Die NIA-Plattform muss Daten nahezu in Echtzeit bereitstellen. Daher ist es wichtig, die Latenz zu reduzieren. Mit Azure Cosmos DB kann NIA eine Latenz im einstelligen Millisekundenbereich erzielen. Dies wird zudem durch das beeindruckende SLA von 99,999 % ergänzt, das für Azure Cosmos DB gilt.
- Wie bereits erwähnt, schätzt NIA seine aktuelle Datenmenge auf über 310 TB, wobei die Wachstumsrate bei 15 GB pro Tag liegt. Hierbei sind die Daten von Airline-Partnern und externen Datenquellen wie Wetter- und Verkehrsdaten noch nicht mit einbezogen. Aus diesem Grund hat sich das Team wegen der Elastizität und unbegrenzten Skalierbarkeit für Azure Cosmos DB entschieden. Azure Cosmos DB bietet NIA die erforderliche Skalierbarkeit mit der Option, nur für den tatsächlich genutzten Speicher und Durchsatz zu bezahlen.
- Der Flughafen verfügt derzeit über mehrere interne Systeme, um seine aktuellen Daten zu speichern. Hierzu gehören SQL-Server und MongoDB-Server. Das Team wünscht eine bessere Kompatibilität mit allen diesen vorhandenen Quellsystemen und möchte, dass vorhandene Anwendungen mit der neuen Datenbank verwendet werden können, ohne Änderungen vornehmen zu müssen. Azure Cosmos DB wird dieser Anforderung perfekt gerecht, da dieser Dienst ein Multi-Modell-Modul mit einem Wire Protocol-kompatiblen API-Endpunkt bietet. NIA-Anwendungen können somit mit mehreren Treibern wie MongoDB, SQL und Gremlin eine Verbindung mit derselben Azure Cosmos DB-Instanz herstellen. Dies vereinfacht die Entwicklung und Bereitstellung, da die APIs derselben Treiber verwendet werden. Außerdem verringern sich die Gesamtbetriebskosten durch die Möglichkeiten für einen Wissenstransfer und die geringere Notwendigkeit von Nacharbeiten.
- Ein weiteres Feature von Azure Cosmos DB, das bei dem Business-Intelligence-Team von NIA Anklang fand, war die Möglichkeit, operative Analytics in Echtzeit und KI auf Cosmos DB auszuführen. Azure Cosmos DB verfügt über eine sofort einsatzbereite Integration in Apache Spark und ermöglicht die Ausführung von Jupyter-Notebooks, um direkt ohne weitere Integration oder Entwicklungsarbeiten mit Daten in Cosmos DB zu arbeiten.

- Kommerziell betrachtet ist Azure Cosmos DB eine kostengünstige Option, da das Business-Intelligence-Team damit die erforderliche Flexibilität und Kontrolle erhält. Das Gute an Azure Cosmos DB ist, dass Sie damit globale Funktionalität erhalten und die Möglichkeit haben, das Kostenmodell dem benötigten Speicher und Durchsatz entsprechend zu steuern. Wenn ein Update für einen Datensatz in Azure Cosmos DB ausgeführt wird, ist dieses Update somit für jeden Anwender weltweit innerhalb von Millisekunden sichtbar.
- Schließlich ist Azure Cosmos DB ein vollständig verwalteter Dienst, das NIA-Team muss sich also nur um die Daten kümmern, die es in Cosmos DB speichert, ohne sich Gedanken um die Infrastruktur machen zu müssen. Darüber hinaus kann das Team schnell und kostengünstig einsteigen und später skalieren, wenn mehr Datasets hinzukommen und der geschäftliche Nutzen steigt.

Beispielimplementierung

Einer der Vorteile von Azure Cosmos DB ist die Kompatibilität mit vielen Abfragemodellen und Treibern. Die folgenden Codeausschnitte zeigen, wie Cosmos DB mithilfe von SQL oder MongoDB abgefragt werden kann. Beide Beispiele sind in C# geschrieben:

1. Im ersten Codeausschnitt werden Datensätze aus der Tabelle **`passengers`** (Passagiere) abgefragt, wobei nach Passagieren mit dem Namen **`Bob`** gesucht wird. Anschließend werden alle zurückgegebenen Ergebnisse durchlaufen, und der Name des Passagiers wird in der Konsole ausgegeben:

```
var sqlQuery = "SELECT * FROM P WHERE P.FirstName = 'Bob'";
Console.WriteLine("Running query: {0}\n", sqlQueryText);
var queryDefinition = new QueryDefinition(sqlQueryText);
var queryResultSetIterator = this.container
                        .GetItemQueryIterator<Passenger>(queryDefinition);

List<Passenger> passengers = new List<Passenger>();
while (queryResultSetIterator.HasMoreResults)
{
    var currentResultSet = await queryResultSetIterator.ReadNextAsync();
    foreach (Passenger p in currentResultSet)
    {
        passengers.Add(p);
        Console.WriteLine("\tRead {0}\n", p);
    }
}
```

2. Im zweiten Codeausschnitt wird eine ähnliche Abfrage ausgeführt, dabei wird jedoch die MongoDB-API verwendet. Zunächst wird **`MongoClientSettings`** und dann **`MongoClient`** erstellt. Dies wird dann verwendet, um einen Verweis auf Azure Cosmos DB zu erhalten. In dem Code wird davon ausgegangen, dass die Konfigurationseinstellungen an dieser Stelle bereits konfiguriert wurden. Der Code erstellt einen Verweis auf die Azure Cosmos DB von NIA (**`NIAAnalytics`**) und fragt **`passengerCollection`** ab:

```
var settings = new MongoClientSettings();
MongoClient client = new MongoClient(settings);

var dbName = "NIAAnalytics";
var collectionName = "Passengers";

var database = client.GetDatabase(dbName);
var passengerCollection = database.
GetCollection<Passenger>(collectionName);

passengers = passengerCollection.Find(new BsonDocument()).ToList();
```

Azure Machine Learning Services

Rolle im Design

Azure Machine Learning Services wird vom Business- Intelligence-Team von NIA zur Operationalisierung seiner Machine-Learning-Modelle verwendet. Zur Optimierung der Ressourcenzuweisung muss das Team eine Reihe von Machine-Learning-Modellen entwickeln, um die Anzahl der Fluggäste vorherzusagen und eine Empfehlung für die Gate-Zuweisung zu erstellen. Azure Machine Learning Services bietet dem Business-Intelligence-Team eine konsistente und reproduzierbare Möglichkeit, Machine-Learning-Modelle zu generieren und gleichzeitig alle Machine-Learning-Experimente, Datasets und Machine-Learning-Trainingsumgebungen nachzuverfolgen. Dies ist von entscheidender Bedeutung für die Implementierung von Machine-Learning-Modellen, bei denen Kunden und Stakeholder grundsätzlich Erklärbarkeit voraussetzen.

Vorteile von Azure Machine Learning Services

- Mithilfe von Azure Machine Learning Services kann NIA den gesamten Machine-Learning-Lebenszyklus von der Datenbereinigung und dem Feature-Engineering bis hin zur Modellerstellung und -validierung optimieren und beschleunigen. Mit Azure Machine Learning Services ist es leicht möglich, viele Teile der Pipeline zu automatisieren. Dies wiederum bedeutet einen geringeren Aufwand, führt zu Qualitätsverbesserungen und ermöglicht dem NIA-Team schnellere Innovationen.
- Beim Erstellen von Machine-Learning-Modellen und Experimentieren mit diesen Modellen sind eine Versionsverwaltung und die Unterhaltung mehrerer Momentaufnahmen von Datasets allgemein üblich. Es kann sehr mühsam und verwirrend sein, mehrere Versionen derselben Datasets zu verwalten. Azure Machine Learning Services bietet eine umfassende Reihe von Features, die Kunden wie NIA helfen sollen, diese Herausforderung zu meistern. Mithilfe von Azure Machine Learning-Datensätzen kann NIA mühelos Datasets nachverfolgen, versionieren und validieren, wie im Abschnitt *Beispielimplementierung* zu sehen ist.
- Eine Herausforderung für jedes Advanced-Analytics-Team besteht darin, den richtigen Algorithmus zum Erstellen eines Machine-Learning-Modells zu finden. Das Business-Intelligence-Team von NIA muss nicht nur den richtigen Algorithmus auswählen, sondern auch alle Hyperparameter feinabstimmen. Azure Machine Learning Services automatisiert diesen gesamten Prozess, sodass jeder Datenanalyst zum Datenwissenschaftler werden kann. Mit Azure AutoML kann das Business-Intelligence-Team von NIA den Prozess des Erstellens von Machine-Learning-Modellen schnell, einfach und kostengünstig automatisieren.
- Ein weiteres großes Plus von Azure Machine Learning Services ist Kompatibilität. Azure Machine Learning Services lässt sich gut mit Open Source Tools wie Databricks integrieren. Darüber hinaus kann NIA alle Machine-Learning-Frameworks (z. B. TensorFlow und PyTorch) verwenden und gleichzeitig alle Vorteile von Azure Machine Learning Services nutzen.
- Mit Azure stehen Unternehmen wie NIA die neuesten bahnbrechenden Innovationen in den Bereichen Daten und KI zur Verfügung. Eine dieser bahnbrechenden Neuerungen ist das Konzept, Computing von den eigentlichen Daten und der Pipeline zu trennen. So kann das Business-Intelligence-Team von NIA seinen Code einmal schreiben und bei jedem Computing ausführen. Dazu gehören Datentransformationscode und Machine-Learning-Modellcode. Das Team von NIA kann sein Machine-Learning-Modell entwickeln, es lokal auf seinen Entwicklungscomputern ausführen und den Code, wenn er bereit ist, in die Cloud verschieben. Dies bietet Entwicklern und Unternehmen eine hohe Flexibilität in Bezug auf Entwicklungs- und Betriebskosten. NIA muss nur für die benötigte Rechenleistung bezahlen und muss die Machine-Learning-Modelle in der Cloud nur trainieren, wenn große Computing-Ressourcen benötigt werden.

- Die CIO von NIA ist starke Verfechterin von DevOps und den Vorteilen, die dies einem Unternehmen bringt. Support für DevOps-Prozesse war ein wichtiger Aspekt bei der Entscheidung für Azure Machine Learning Services. Azure Machine Learning Services verfügt über eine native Integration in Azure DevOps, NIA ist damit problemlos in der Lage, Machine-Learning-Modelle zu erstellen und bereitzustellen.
- Sicherheit, Reproduzierbarkeit und Governance sind für jedes Advanced-Analytics-Team von großer Bedeutung. Microsoft Azure bietet eine schöne, elegante Lösung hierfür durch die native Integration in anderen bewährten Azure-Diensten für Unternehmen. Azure Machine Learning Services ermöglicht eine sofortige Integration in Azure AD und Azure Monitor. Darüber hinaus können Unternehmen wie NIA mithilfe von Azure Resource Manager-Vorlagen und Azure Blueprints ordnungsgemäße Governance und Standards durchsetzen.

Beispielimplementierung

Mit Azure Machine Learning Services ist es einfach, mehrere Versionen eines Datasets für Machine-Learning-Zwecke zu versionieren, nachzuverfolgen und zu bearbeiten. Mit dem folgenden Codeausschnitt wird zunächst ein Datenspeicher erstellt, damit Azure Machine Learning Services weiß, wo die Daten gespeichert werden sollen:

```
# creating a ref to Azure ML Service Workspace

import azureml.core

from azureml.core import Workspace, Datastore

ws = Workspace.from_config()


# Registering Azure Blob container as the datastore

datastore = Datastore.register_azure_blob_container(workspace=ws, datastore_
name='NIA_airport_datastore',

container_name='NIA_Analytics',

account_name={storageAccount},

account_key={storageAccountKey},

create_if_not_exists=True)


# get named datastore (if exist)

datastore = Datastore.get(ws, datastore_name='NIA_airport_datastore')
```

Mit dem vorherigen, in Python geschriebenen Codeausschnitt wird zunächst ein Azure Machine Learning Services-Arbeitsbereich aus einer vorhandenen Konfigurationsdatei erstellt. Der Code erstellt dann einen Datenspeicher, indem ein Azure BLOB-Container als Datenspeicher registriert wird. Der Datenspeicher in dem Beispiel heißt **`NIA_airport_datastore`**, für den Namen und Schlüssel des Azure Storage-Kontos werden hier Platzhalter verwendet. Schließlich wird in dem Beispiel durch Verwendung des entsprechenden Namens ein Verweis auf einen bereits vorhandenen Datenspeicher erstellt.

Mit dem folgenden Codeausschnitt wird ein neues Dataset registriert. Um die Suche nach diesem Dataset in Zukunft zu vereinfachen, sind ein Name, eine Beschreibung und ein Tag angegeben.

```
passengers_ds = passengers_ds.register(workspace =ws,name='passengers_
dataset',description = 'passengers personal data and address',tags =
{'year': '2019'})
```

Mit dem folgenden Codeausschnitt wird ein vorhandenes Dataset anhand des Namens und/oder der Versions-ID abgerufen. Dies ist sehr nützlich, wenn es mehrere Versionen desselben Datasets gibt:

```
#get Passengers dataset by name
passengers_ds = ws.datasets['passengers_dataset']


# get specific version of the passengers dataset
passengers_ds = ws.datasets['passengers_dataset']
passengers_ds_v3 = passengers_ds.get_definition(version_id = 3)
```

Azure Container Registry

Rolle im Design

Mithilfe von Azure Machine Learning Services kann das Business-Intelligence-Team von NIA seine Machine-Learning-Modelle als Standardcontainer erstellen, die auf jedem Containermodul, wie z. B. Docker und Kubernetes, ausgeführt werden können. Das Team verwendet Azure Container Registry zum sicheren Hosten und Freigeben dieser Docker-Container, die die Machine-Learning-Modelle enthalten.

Vorteile von Azure Container Registry

- Mit Azure Container Registry kann NIA Images für alle Arten von Containern zu speichern. Azure Container Registry trennt das Hosting der Images von der Bereitstellung dieser Images auf den verschiedenen Bereitstellungszielen wie Docker Swarm und Kubernetes. Dadurch kann NIA eine Container Registry (ACR) verwenden, um Images für alle Arten von Containern zu hosten.
- Azure Container Registry baut auf der Funktionalität der standardmäßigen Container Registrys auf. So kann Azure Container Registry beispielsweise mit Azure AD integriert werden, um die Sicherheit zu verbessern. Darüber hinaus bietet Azure Container Registry eine einfache Möglichkeit der Integration in Containeraktionen unter Verwendung von Triggern. NIA kann beispielsweise einen Webhook konfigurieren, um Azure DevOps Services auszulösen, wenn ein neues Image zu Azure Container Registry hinzugefügt wird.
- Azure Container Registry ist vollständig mit der standardmäßigen Docker Registry V2 kompatibel. Somit kann das Team von NIA dieselben Open-Source-Docker-**CLI**-Tools (**Command-line Interface, Befehlszeilenschnittstelle**) für die Interaktion mit beiden Registrys (Azure Container Registry und Docker Registry V2) verwenden.
- Azure Container Registry unterstützt die Replikation in mehreren Regionen. Dies ist für NIA interessant, da es in zweierlei Hinsicht hilfreich ist. Erstens verringert es die Netzwerklatenz und die Kosten, da die Container Registry in der Nähe der Bereitstellungsziele bleibt. Zweitens verbessert es Business Continuity und Notfallwiederherstellung, da dieselbe Container Registry in mehreren Regionen repliziert wird.

Beispielimplementierung

Der folgende Code ist Teil der **Azure Resource Manager**-(**ARM-**)Vorlage, die NIA zum Erstellen der Azure Container Registry-Instanz verwendet. Mit diesem Code wird eine Azure Container Registry-Instanz (Standardebene) im Azure-Rechenzentrum in Südafrika, Norden, erstellt. Mit der Vorlage kann das Administratorbenutzerkonto zudem die Registry verwalten. Die ARM-Vorlage enthält zwei Parameter: einen Parameter für den Namen der Registry und einen anderen Parameter für die ARM-API-Version:

```
{
    "resources": [
        {
            "name": "[parameters('registryName')]",
            "type": "Microsoft.ContainerRegistry/registries",
            "location": "South Africa North",
            "apiVersion": "[parameters('registryApiVersion')]",
            "sku": {
                "name": "Standard"
            },
            "properties": {
                "adminUserEnabled": "True"
            }
        }
    ]
}
```

Azure Kubernetes Service (AKS)

Rolle im Design

Mithilfe von Azure Kubernetes Service (AKS) können Machine-Learning-Modelle als verwendbare APIs bereitgestellt werden. Ein Beispiel für diese Machine-Learning-Modelle ist ein Modell, das die Bewegung der Menschenmassen durch den Flughafen prognostiziert. Solche Modelle werden vom Team unter Verwendung von historischen Daten in Azure Databricks trainiert. Anschließend wird das Modell mithilfe von Azure Machine Learning Services als Docker-Image übertragen. AKS führt diese Modelle und andere Apps wie z. B. Passenger Hub aus. Darüber hinaus hilft AKS, die Dienstermittlung dieser Apps zu verwalten, bietet automatische Skalierungsmechanismen und unterstützt Selbstreparatur-Richtlinien für die Behandlung von Fehlern.

Vorteile von AKS

- Die Verwaltung eines Computerclusters ist eine schwierige Aufgabe, und einen Kubernetes-Cluster zu verwalten und zu konfigurieren, ist noch schwieriger. Das liegt daran, dass Kubernetes viele bewegliche Teile umfasst und viel Konfiguration erfordert. AKS vereinfacht dies mit dem Angebot eines verwalteten Clusters. Dabei verwaltet Microsoft Azure die Masterknoten, und das Team von NIA muss nur die Slave-Knoten für die Bereitstellung seiner Workloads konfigurieren und verwenden. Dies reduziert den Aufwand für NIA erheblich.
- Mithilfe von Konzepten wie **virtuellen Knoten** und **virtuellen Kubelets** bietet AKS NIA die Möglichkeit, jederzeit elastisch zusätzliche Kapazitäten bereitzustellen. Dies ist für NIA entscheidend, da es sehr schwierig ist, die Auslastung und die benötigte Kapazität zu prognostizieren. Daher ist diese elastische Bereitstellung bei Bedarf wichtig.
- Die native Integration und der Support für AKS in Azure DevOps ist ein weiterer Vorteil von AKS. Dies vereinfacht die Konfiguration und Automatisierung von Bereitstellungen von NIA-Workloads in AKS. Darüber hinaus verfügt AKS über eine native Integration in Diensten wie Azure Monitor und Azure Key Vault.
- Das Team von NIA kann die End-to-End-Entwicklungsumgebung dank des Visual Studio Code-Supports für AKS verbessern und beschleunigen.
- Neben der nativen Integration in anderen Azure-Diensten kann AKS gut mit Azure Active Directory integriert werden. Durch die Nutzung dieser Integration kann NIA die Sicherheit erhöhen. Darüber hinaus kann NIA mithilfe von Azure Policy Governance im gesamten Unternehmen durchsetzen.
- Azure bietet erstklassige Unterstützung für Open Source Tools wie Kubernetes, nicht nur in der Cloud, sondern auch am Edge. Das Team von NIA ist sich bewusst, dass es Situationen gibt, in denen die Verlagerung von Computing zum Edge die beste Option sein kann. Ein Beispiel hierfür ist der Plan des Teams, Machine-Learning-Modelle in die Nähe von Verkehrsüberwachungskameras zu bringen, um Warnungen auszulösen, wenn ein Sicherheitsereignis eintritt. Microsoft Azure bietet gute Unterstützung für die Ausführung von Kubernetes auf Azure IoT Edge in solchen Situationen. Daher ist die Verwendung von AKS eine gute Option für zukünftige Pläne, Machine-Learning-Modelle durch Verwendung von Kubernetes mit Azure IoT Edge zum Edge zu verlagern.

Beispielimplementierung

Der folgende Codeausschnitt ist Teil der Azure DevOps Services-Pipeline von NIA, die die neue Webanwendung „Airport Brain“ bereitstellt. Der Code nutzt die Unterstützung von Azure DevOps für Kubernetes durch Verwendung des Aufgabentyps **`KubernetesManifest`**. Mit der Aufgabe wird das Docker-Image unter **`nia/aiportbrain: lastest`** unter Verwendung von **`NIAairport_AksServcieConnection`** in dem vorkonfigurierten AKS bereitgestellt. Der folgende Code enthält einen Platzhalter für **`imagePullSecret`**, das als Authentifizierungsmechanismus zum Übertragen von Images aus Azure Container Registry zum Bereitstellungsziel (AKS) verwendet wird:

```
steps:
- task: "KubernetesManifest@0"
  displayName: "Deploy AirportBrain to K8s"
  inputs:
    action: deploy
    kubernetesServiceConnection: "NIAairport_AksServiceConnection"
    namespace: "airportbrain"
    manifests: "manifests/deployment.yml"
    containers: 'nia/airportBrain:latest'
    imagePullSecrets: |
      $(imagePullSecret)
```

Power BI

Rolle im Design

Die Strategie von Zara für die Berichterstellung von NIA besteht zu einem Teil darin, Power BI als Visualisierungstool bei NIA zu verwenden. Power BI kann zum Generieren von Berichten und Dashboards und zu Self-Service-Zwecken verwendet werden. Das BI-Team hofft, auch die Power BI-Möglichkeit zur Einbettung in andere Web-Apps nutzen zu können, um Power BI-Visualisierungen in anderen neuen Apps, wie z. B. Passenger Hub, wiederzuverwenden.

Vorteile von Power BI

Neben allen Vorteilen von Power BI, die bereits in *Anwendungsfall 1* erörtert wurden, besteht die Möglichkeit, Power BI-Berichte und Dashboards in andere Webanwendungen einzubetten. Das Business-Intelligence-Team von NIA möchte die Einfachheit und Leistungsfähigkeit der Visuals von Power BI nutzen, um die Dashboards für die Apps „Passenger Hub“ und „Airport Brain“ zu entwickeln. Hierfür kann das Team das Feature des Power BI-Diensts zur Einbettung in Webanwendungen nutzen. Mit diesem Feature kann das Business-Intelligence-Team von NIA Berichte schnell und einfach entwickeln und versenden und die Berichte gleichzeitig auf sichere Weise über die neuen Webanwendungen von NIA bereitstellen.

Beispielimplementierung

In Power BI können Sie jeden Power BI-Bericht in eine Webseite einbetten. Wählen Sie dazu im Menü **File** (Datei) die Option **Web Publish** (Webveröffentlichung) aus. Daraufhin wird ein Dialogfeld aufgerufen, in dem Sie einen Einbindungscode generieren können:

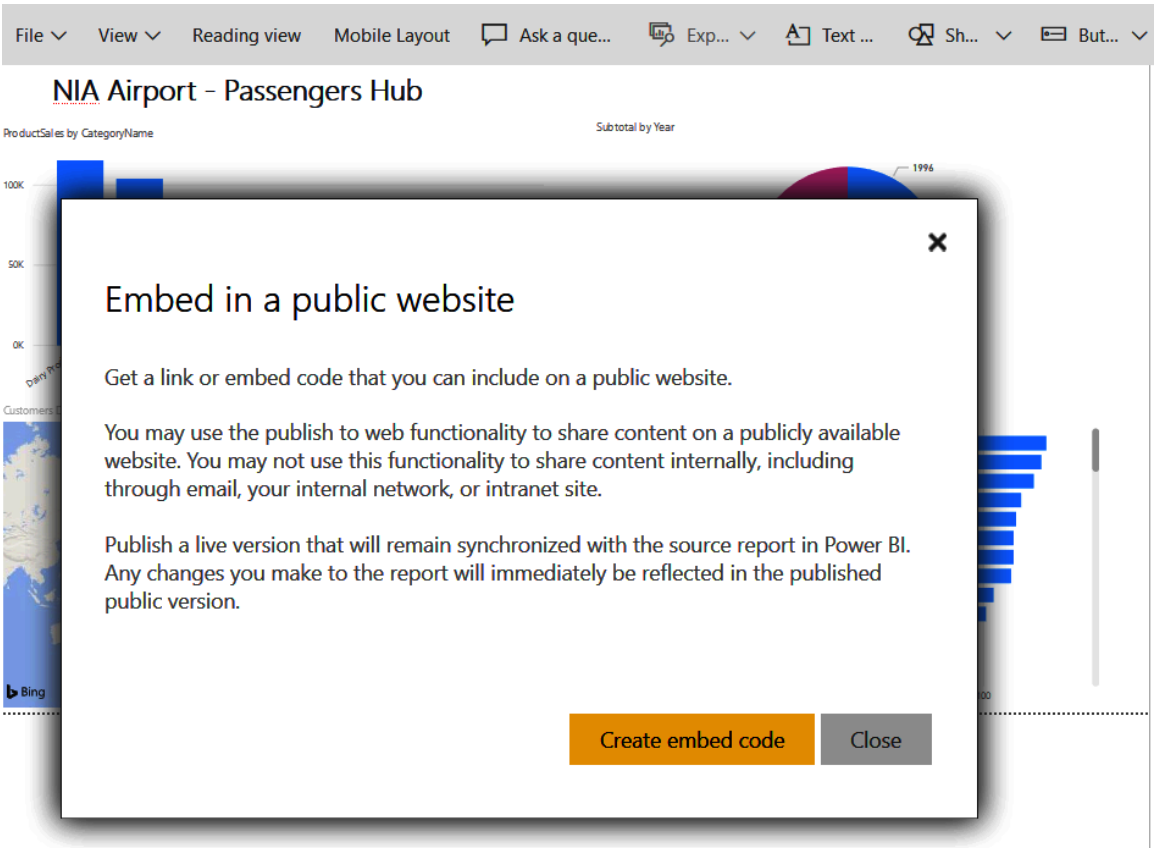

Abbildung 5.7: Erstellen eines Einbindungscodes

Wenn Sie auf die Schaltfläche **Create embed code** (Einbindungscode erstellen) klicken, wird ein `iFrame`-Code generiert, der auf jeder HTML-Webseite verwendet werden kann. Im nächsten Dialogfeld können die Mitglieder des NIA-Teams die Eigenschaften des `iFrame`-Codes, wie z. B. **width** (Breite) und **height** (Höhe), konfigurieren. Anschließend kann der Bericht mit dem `iFrame`-Code in jede Webanwendung eingebettet werden:

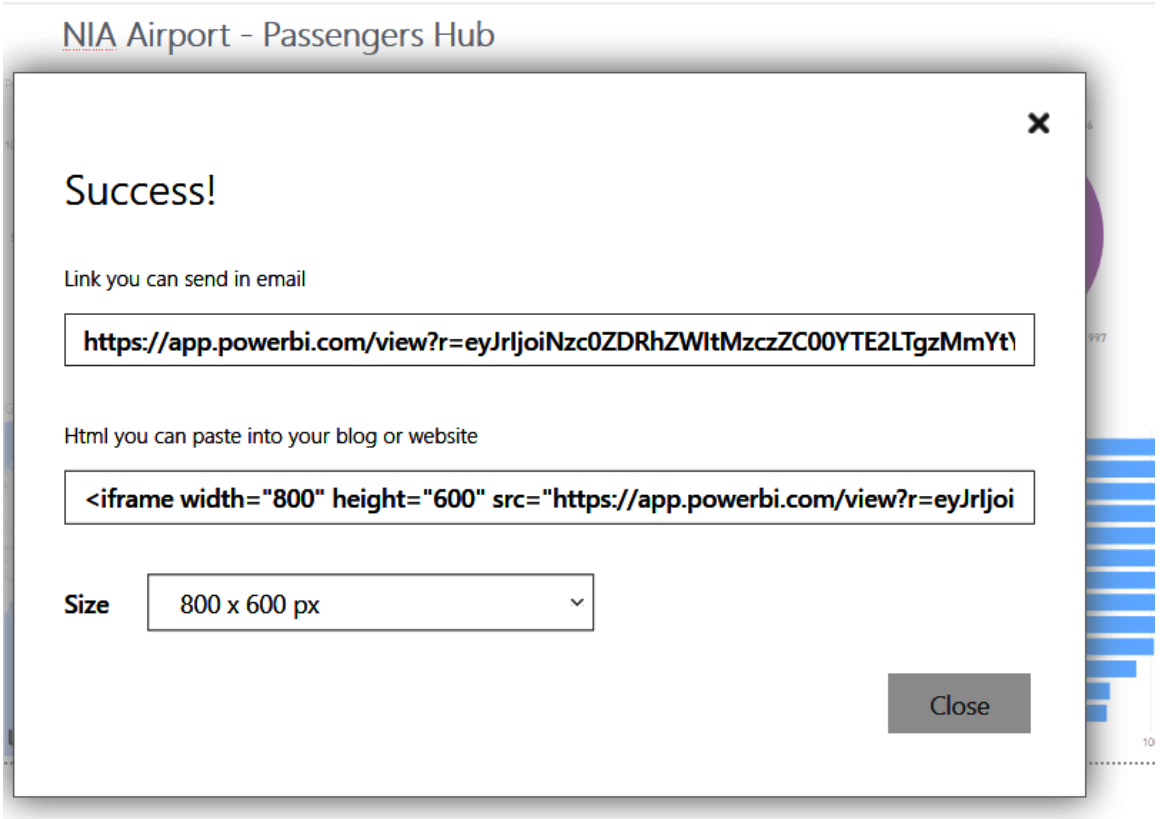

Abbildung 5.8: Konfigurieren der iframe-Eigenschaften

Unterstützende Dienste

NIA möchte sicherstellen, dass die neue Lösung sicher und skalierbar ist und ein hohes Maß an Überwachung und Unterstützung bietet. Azure verfügt über viele Dienste, mit denen Unternehmen wie NIA ihre Lösungen schützen, skalieren und überwachen können. Hierzu gehören alle in *Anwendungsfall 1*aufgeführten Dienste wie Azure DevOps, Azure Key Vault und Azure Monitor.

Insights und Aktionen

Mithilfe von Azure konnte NIA nach der Analyse aussagekräftige Insights gewinnen und die erforderlichen Maßnahmen bereitstellen, wie in den folgenden Abschnitten beschrieben.

Verringerung der Flugverspätungen um 17 % mit Predictive Analytics

Beschreibung: Während der ersten Datenermittlung und -erkundung stellte das Business-Intelligence-Team von NIA fest, dass ineffiziente Gate-Zuweisungen wesentlich zu Flugverspätungen beitrugen. Flugverspätungen haben einen Schneeballeffekt, da sich die Verzögerung eines Fluges auf den nächsten und dies wiederum auf den übernächsten Flug auswirken kann. Für die Passagiere bedeutet dies zudem eine negative Erfahrung. Derzeit basiert die Zuweisung von Gates bei NIA auf der Kapazität des Wartebereichs und der maximalen Kapazität der Flugzeuge. Dabei wird davon ausgegangen, dass alle Flüge ausgebucht sind, was nicht unbedingt stimmt.

Durch die Kombination von Wetterdaten, Daten zum Stadtverkehr, historischen Daten zu Flugverspätungen sowie Daten aus anderen Quellen konnte das Business-Intelligence-Team ein besseres Empfehlungsmodul für die Gate-Zuweisung erstellen. Das neue Empfehlungsmodul, das mithilfe von Machine Learning entwickelt wurde, prüft kontextbezogene Daten (Wetter und Verkehr) sowie historische Daten, um die Anzahl der Passagiere eines bestimmten Flugs zu schätzen und entsprechend ein Gate zuzuweisen. Bei der ersten Modellierung und Validierung stellte das Team fest, dass die Bereitstellung eines solchen Empfehlungsmoduls in der App „Airport Brain“ die Flugverspätungen um 17 % reduzieren kann.

Geschätzter Geschäftswert: 14,7 Mio. USD/Jahr

Wichtige Datenquellen: Flugdaten der Fluglinien, Flughafendaten (Layout und Gates), Wetterdaten, Daten zum Stadtverkehr, Schulkalender und Feiertagskalender

Aktionen: Das Business-Intelligence-Team von NIA hat die Lösung unter Verwendung der in *Abbildung* 5.6 dargestellten Architektur bereitgestellt. Infolge dieser Lösung verfügt das Flughafenmanagement nun über ein neues Tool im Rahmen des neuen Portals (**Airport Brain**), um Empfehlungen in Echtzeit für die Zuweisung von Gates bereitzustellen. Dies verbessert die Effizienz und senkt den betrieblichen Aufwand, da bei der Planung auf Vermutungen verzichtet wird und betriebliche Entscheidungen eingeführt werden, die auf Fakten und Wissenschaft basieren.

Datenpipeline: Die vereinfachte Datenpipeline für diese Initiative ist in *Abbildung* 5.9 dargestellt:

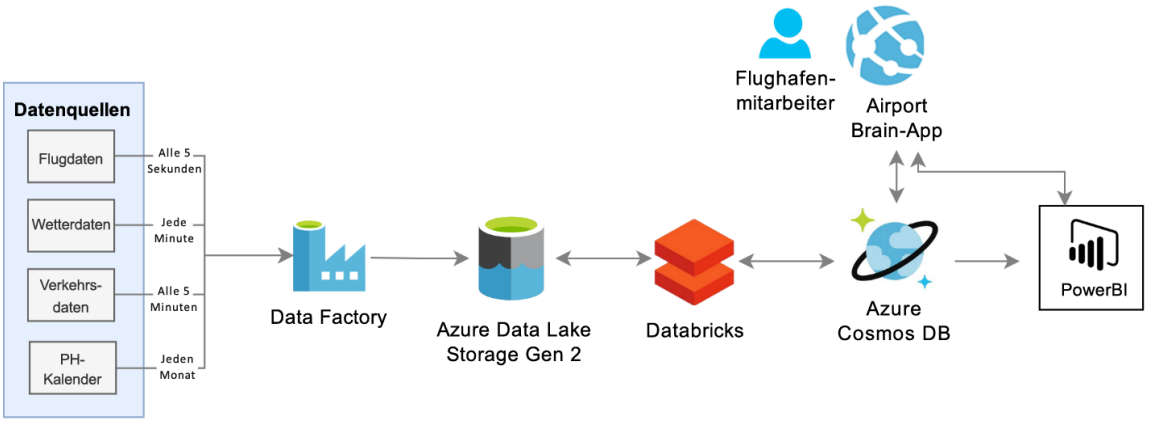

Abbildung 5.9: Datenpipeline für Initiative 1

Verringerung von Überlastung und Verbesserung des Einzelhandels mit intelligenter Visualisierung

Beschreibung: Das Business-Intelligence-Team von NIA machte eine weitere interessante Entdeckung: die Korrelation zwischen der Ankunftszeit der Fluggäste mit dem Auto und langen Warteschlangen. Das Team stellte fest, dass lange Warteschlangen und Überfüllung zu einem Problem wurden, wenn viele Passagiere mehr als 4 Stunden vor der Abflugzeit am Flughafen ankamen. Dies ist darauf zurückzuführen, dass das Flughafenmanagement nicht mit der Ankunft der Passagiere zu diesem Zeitpunkt gerechnet hatte/dies nicht eingeplant hatte. Dadurch kam es zu den langen Warteschlangen und der Überlastung. Einer der Geschäftsführer am Flughafen fand noch eine weitere Erklärung, nämlich, dass sich diese früh ankommenden Passagiere direkt zum Gate begaben und keine anderen Bereiche am Flughafen aufsuchten.

Zur Lösung dieses Problems beschloss das Team daher, die früh ankommenden Passagiere in andere Bereiche des Flughafens, wie z. B. die Duty-Free-Zone, das Kino und die Ruhebereiche zu leiten. Nach ersten Tests schätzt das Team, dass damit die Verkaufschancen des Einzelhandels um 11 % steigen und gleichzeitig die Überfüllung an den Gates um ca. 15 % reduziert werden kann.

Geschätzter Geschäftswert: 9,3 Mio. USD/Jahr

Wichtige Datenquellen: Flugdaten der Fluglinien, Flughafendaten (Layout und Gates), Wetterdaten, Fluggastinformationen, Daten des Einzelhandels am Flughafen und Feiertagskalender

Aktionen: Aufgrund dieser Erkenntnisse hat das Team neue Dashboards in der App „Passenger Hub“ erstellt. Wenn Passagiere früh ankommen und ihren Ausweis scannen, zeigt ihnen die App „Passenger Hub“ ihre Flugdaten an und führt sie zu Ruhebereichen, Duty-Free-Shops und zum Flughafenkino. Das Team nutzte auch Echtzeit-Verkehrsüberwachungsdaten, um große Hinweistafeln zu erstellen, die auf großen Bildschirmen am Flughafen angezeigt werden, damit sie sogar ohne Scannen von den Anwendern gesehen werden können.

Datenpipeline: Die vereinfachte Datenpipeline für diese Initiative ist in *Abbildung* 5.10 dargestellt:

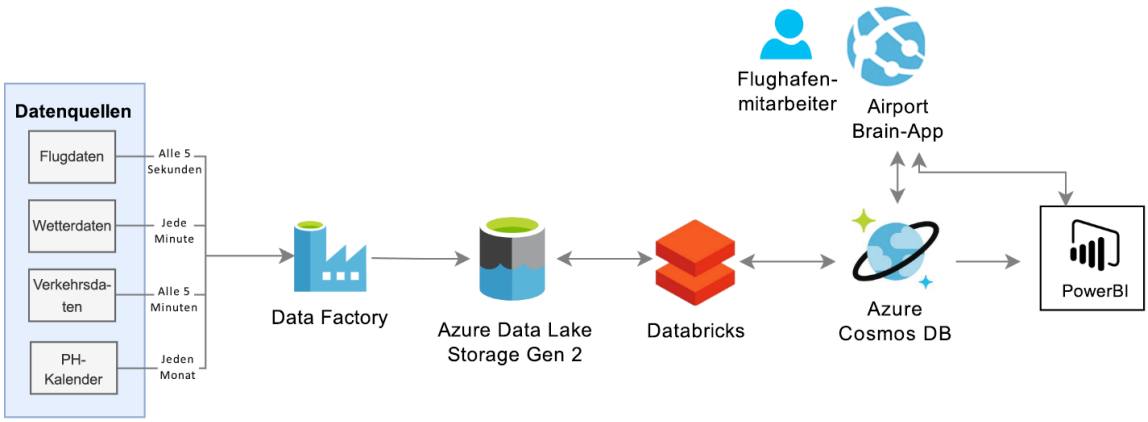

Abbildung 5.10: Datenpipeline für Initiative 2

Fazit

Die Abläufe und Verfahren an Flughäfen sind komplex, und es herrscht rund um die Uhr Betrieb. Daher können selbst kleine Verbesserungen für den Flughafen große Einsparungen bedeuten und die Sicherheit und Kundenzufriedenheit erhöhen.

Auf den vorherigen Seiten haben Sie sich mit einem praktischen Beispiel eines großen Flughafens befasst. Zwar sind die Namen fiktiv, viele hier erörterte Zahlen basieren jedoch auf einem tatsächlichen Anwendungsfall, in den der Autor einbezogen war. Sie haben gesehen, wie mithilfe von Advanced Analytics Effizienzverbesserungen erzielt und Millionen gespart werden können. Mithilfe von Daten können Flughäfen nicht nur Betriebskosten sparen, sondern auch einen Wettbewerbsvorteil erzielen.

Sie haben auch gesehen, wie eine datengesteuerte Lösung mit Azure implementiert werden kann und warum Azure die perfekte Plattform für die Ausführung solcher Workloads ist. Azure ist kostengünstig, sicher und bietet Unternehmen Flexibilität und Skalierbarkeit.

6

Schlussbemerkungen

In diesem Buch haben Sie eine Reihe von Technologien kennengelernt, die Azure für Cloud-Analytics bereitstellt. Sie haben den umfassenden Satz von Azure-Diensten entdeckt, mit denen Sie Daten aus verschiedenen Datenquellen wie Datenbanktabellen, Dateien, Streaming und anderen Arten von Quellen erfassen, speichern und analysieren können.

In diesem letzten Kapitel konzentrieren wir uns auf den Lebenszyklus eines modernen Data Warehouse, um diese Azure-Dienste und -Konzepte zusammenzuführen und ihre jeweiligen Vorteile hervorzuheben. Darüber hinaus beschreiben wir, wie einige dieser Dienste von Unternehmen weltweit eingeführt werden.

Azure-Lebenszyklus eines modernen Data Warehouse

Der Lebenszyklus eines modernen Data Warehouse stellt Ihnen eine solide Grundlage für alle Arten von Unternehmens-Analytics bereit. Dieser Lebenszyklus ermöglicht Ihnen die Durchführung aller Arten von Analytics (von SQL-Abfragen bis zu erweitertem Machine Learning), ohne die Genauigkeit Ihrer Daten kompromittieren oder eine übermäßige Anzahl von Datenverarbeitungsressourcen bereitstellen zu müssen, die anschließend möglicherweise nicht ausgelastet werden.

Der Lebenszyklus eines modernen Data Warehouse besteht aus fünf Schritten:

1. Aufnehmen der Daten
2. Speichern der Daten
3. Vorbereiten und Trainieren der Daten
4. Modellieren und Bereitstellen der Ergebnisse
5. Visualisierung

Da Daten in Bezug auf ihre Struktur und Komplexität immer vielfältiger werden, die Menge der Daten weiter zunimmt und sich die Art der gewünschten Analytics weiterentwickelt, benötigen Datenwissenschaftler und Entwickler Funktionen, die diese geänderten Anforderungen erfüllen können. An dieser Stelle kommt der Azure-Lebenszyklus eines modernen Data Warehouse zum Tragen. Azure verfügt über die nötigen Technologien zur Erfüllung der Anforderungen – unabhängig davon, ob Sie mit strukturierten tabellarischen Daten oder mit umfangreichen oder komplexen unstrukturierten Big Data aus Geräten, Diensten und Anwendungen arbeiten.

Um die Azure-Dienste und -Konzepte zusammenzuführen, werden in diesem Kapitel die fünf Prozesse im Lebenszyklus eines modernen Data Warehouse erneut und abschließend betrachtet.

Aufnehmen der Daten

Sie haben an früherer Stelle gelernt, dass die Implementierung einer modernen Warehouse-Lösung mit der Aufnahme der Daten in Azure beginnt. Zunächst verbinden Sie verschiedene Datensätze aus verschiedenen Quellen und nehmen die Daten in Azure auf. Die Daten können aus RDBMS-Tabellen stammen, die Kundeninformationen wie Namen, Adressen, Telefonnummern und Kreditkartendetails enthalten. Es kann sich auch um semistrukturierte Daten aus Social-Media-Plattformen, z. B. Twitter-Feeds, oder um unstrukturierte Daten aus IoT-Sensoren handeln. Ihre Daten können aus Ihrem On-Premises-Rechenzentrum, aus Clouddiensten oder aus beidem stammen.

Mit Azure können Sie Daten mithilfe robuster Dienste für die Batch- oder Echtzeitaufnahme aufnehmen. Daher können Sie Ereignisse erfassen, noch während sie von Ihren Diensten und Geräten generiert werden. In diesem Abschnitt werden die Tools und Dienste beschrieben, die für diesen Schritt verwendet werden.

Azure Data Factory

Azure Data Factory ist der primäre Dienst für die Batchaufnahme von Daten. Es handelt sich um einen vollständig verwalteten Dienst für Aufnahme, Orchestrierung und Planung, der eine skalierbare Datenintegration ermöglicht. Azure Data Factory unterstützt Sie beim Erstellen durchgängiger Daten-Pipelines, die Daten aus verschiedenen Quellen aufnehmen, diese Datensätze verarbeiten und transformieren, ereignisgesteuerte Pipelines planen und auslösen und Datenvisualisierung bereitstellen können.

Der Vorteil von Azure Data Factory besteht darin, dass der Dienst sofort verwendbare Connectors für Verbindungen mit anderen Anwendungen bereitstellt – von anderen Clouds (z. B. Amazon S3) bis zu SaaS-Anwendungen (z. B. Salesforce oder Google AdWords) und Hybridkonnektivität (z. B. SQL, MongoDb oder Oracle), die in Ihren On-Premises-Rechenzentren vorhanden sind. Zurzeit kann Azure Data Factory mit mehr als 80 nativ entwickelten und wartungsfreien Connectors ohne zusätzliche Kosten integriert werden.

Beispielsweise bietet das kalifornische Unternehmen LUMEDX Kardiologen kardiovaskuläre Informationssysteme für die Patientenversorgung an. Die Systeme nutzen Azure Data Factory, um strukturierte und unstrukturierte Daten aus mehreren Quellen aufzunehmen. Anschließend werden diese Daten verarbeitet, um Insights abzuleiten und schnell Datenpipelines bereitzustellen. Auf diese Weise können Gesundheitsdienstleister ihren Patienten frühzeitig bessere Lösungen anbieten.

Azure Import/Export Service

Wenn Sie massive Datenmengen besitzen, die Sie in Azure aufnehmen möchten, können Sie Azure Import/Export verwenden. Azure Import/Export verwendet das Befehlszeilentool **WAImportExport**, das die Daten auf der Festplatte, die Sie an ein Azure-Rechenzentrum senden möchten, mithilfe von BitLocker verschlüsselt. Unternehmen können mit Azure Import/Export in den folgenden Fällen Vorteile erzielen:

- Wenn große Datenmengen schnell und kostengünstig zu Azure migriert werden sollen.
- Wenn große Datenmengen aus der Cloud wiederhergestellt und an einem On-Premises-Standort bereitgestellt werden sollen.
- Wenn On-Premises-Daten gesichert und in der Cloud gespeichert werden müssen.

Azure Data Migration Service

Wenn Sie strukturierte Daten besitzen, ermöglicht Ihnen der Azure Data Migration Service die Aufnahme dieser Daten durch die Migration aus On-Premises-Strukturdatenbanken zu Azure. Gleichzeitig werden die relationalen Strukturen beibehalten, die Ihre aktuellen Anwendungen verwenden. Der Vorteil dieses Ansatzes besteht darin, dass Sie die Änderungen an der vorhandenen Datenstruktur minimieren können, während Sie die Daten zu Azure migrieren.

Azure Event Hubs

Azure Event Hubs bietet Big-Data-Streaming und die Echtzeitaufnahme von Ereignissen mit dynamischen Skalierungsmöglichkeiten. Der Vorteil von Azure Event Hubs besteht in der Fähigkeit, Millionen von Ereignissen pro Sekunde zu verarbeiten. Dies ermöglicht die Aufnahme umfangreicher Telemetrie- und Ereignisdaten aus Millionen von Geräten und Ereignissen bei langlebiger Pufferung und niedriger Latenz.

Beispielsweise verwendet das britische Unternehmen Pizza Express Azure Event Hubs zusammen mit Azure Data Factory, um Kundenaktivitäten nachzuverfolgen. Hierzu werden Daten aus verschiedenen Quellen und in verschiedenen Formaten aufgenommen. Dies hilft dem Unternehmen, Kundenpräferenzen zu verstehen und Strategien für die Kundenbindung zu entwickeln.

Azure IoT Hub

Azure IoT Hub ist ein Dienst für die Übertragung von Telemetriedaten aus Geräten zur Cloud und ermöglicht die Nachverfolgung und Interpretation des Zustands von Geräten und Ressourcen. Es handelt sich um einen verwalteten und sicheren Dienst für die bidirektionale Kommunikation zwischen IoT-Geräten und Azure.

Beispielsweise entwickelte Bridgestone, einer der weltweit führenden Reifenhersteller, mithilfe von Azure-Diensten eine Lösung namens Tirematics. Tirematics sendet Daten aus Reifensensoren an Azure. Dies hilft Technikern, frühe Anzeichen von Reifenproblemen wie Temperatur- und Druckanomalien zu erkennen. Die Tirematics-Lösung nutzt Azure IoT Hub, um Daten aus Reifensensoren zu empfangen und im Data Lake zu speichern. Anschließend wird Stream Analytics verwendet, um diese Daten zu analysieren und Anomalien zu entdecken.

Azure CLI

Die **Azure-Befehlszeilenschnittstelle** (**CLI**) ist eine plattformübergreifende Befehlszeilenumgebung für die Verwaltung von Azure-Ressourcen. Der Vorteil von Azure CLI besteht darin, dass Sie Skripte schreiben können, um Datenformate programmgesteuert auszuwählen und in Azure aufzunehmen.

Azure SDK

Eine weitere Möglichkeit, Daten in Azure aufzunehmen, besteht in der Verwendung des **Azure Software Development Kit** (**SDK**). Der Vorteil des Azure SDK besteht darin, dass Entwickler benutzerdefinierte Anwendungen schreiben können, um verschiedene Datenformate in Azure aufzunehmen.

All diese Tools und Dienste unterstützen Sie bei der Aufnahme Ihrer Daten. Als Nächstes müssen Sie entscheiden, wie die aufgenommenen Daten gespeichert werden sollen.

Speichern der Daten

Das Speichern der aufgenommenen Daten ist der zweite Schritt im Lebenszyklus eines modernen Data Warehouse. Im folgenden Abschnitt werden die Tools und Dienste beschrieben, die für diesen Schritt verwendet werden.

Azure Blob Storage

Azure Blob Storage kann massive Datensätze (einschließlich Videos, Bildern usw.) unabhängig von ihrer Struktur speichern und für die Analyse bereithalten. Der Vorteil von Azure Blob Storage besteht darin, dass Daten in verschiedenen Formaten und Strukturen einfach bereitgestellt und verarbeitet werden können.

Azure Data Lake Storage Gen2

Sie haben gesehen, dass Azure Data Lake Storage Gen2 eine kostengünstige, Analytics-optimierte Speicherplattform darstellt. Sie basiert auf Azure Blob Storage und ist vollständig unabhängig von Datenverarbeitungs-Engines. Daher können auf dieser Plattform beliebige Arten von Daten akkumuliert und organisiert werden, um anschließend von Analytics-Engines verarbeitet zu werden. Der Vorteil der Verwendung von Azure Data Lake Storage Gen2 besteht darin, dass die Speicherung unabhängig von der Datenverarbeitung erfolgt und die Datenverarbeitung unabhängig von der Speicherung gestaltet werden kann, um die Kosten zu minimieren.

Die Piraeus Bank, eines der größten Finanzinstitute Griechenlands, nutzt Microsoft Azure, um Insights abzuleiten, Status und Nutzung von Kunden zu verstehen und neue Management-KPIs zu erstellen. Hierzu werden die Daten aus Azure Data Lake und Azure Data Factory extrahiert.

Azure SQL Database

Sie können Azure SQL Database für operative und transaktionale Daten in strukturierter oder relationaler Form verwenden. Azure SQL Database funktioniert wie die On-Premises-Version von Microsoft SQL Server, jedoch in Form eines Azure-Diensts. Der Vorteil von Azure SQL Database besteht darin, dass Sie sich keine Gedanken über die Verwaltung oder Skalierung Ihrer Host-Infrastruktur machen müssen. Wenn Sie möchten, können Sie vorhandene Datenbankanwendungen auch auf Windows- oder Linux-basierten virtuellen Maschinen auf Azure hosten.

Das Barcelona Smart City-Projekt erfasst und analysiert verschiedene Daten aus verschiedenen Systemen und öffentlichen Quellen, z. B. GPS-Signalen, Software-Log-Dateien und sozialen Medien. Dies ermöglicht ihnen Insights in die Effektivität von Behörden und die Bereitstellung besserer Services für Bürger, indem die Zusammenarbeit zwischen Behörden, Menschen und Unternehmen verbessert wird. Die Stadt verwendet Azure SQL Database, um eine Vielzahl von Daten zu speichern, z. B. Wahlergebnisse, Daten aus öffentlichen Einrichtungen, Bevölkerungsdaten, Termine, Auftragnehmerprofile usw. Sie verwendet diese Daten auch, um ein Dashboard zu erstellen, das Bürgern nahezu in Echtzeit-anhand von ungefähr 120 KPIs Insights bereitstellt. Diese KPIs stellen Informationen zu verschiedenen Themen bereit (z. B. öffentliche Fahrradnutzung, Wirtschaft, Demografie, Buslinien, von Bürgern genutzte Buslinien usw.).

Azure Synapse Analytics

Azure Synapse Analytics stellt für Analysedaten, die über mehrere Jahre aggregiert wurden, einen elastischen Dienst bereit, der auf mehrere Petabyte skaliert werden kann. Auf diese Weise können Sie Ihre Daten dynamisch skalieren, entweder in Form von On-Premises-Daten oder in Azure. Der Dienst ist mit Azure Data Factory, Azure Machine Learning, HDInsight und Power BI nahtlos kompatibel. Dank der parallelen und verteilten Verarbeitungsarchitektur kann Azure Synapse Analytics eine immense Datenverarbeitungsleistung bereitstellen. Die skalierbare Natur von Azure Synapse Analytics kann auch komplexeste analytische Workload-Anforderungen bewältigen.

Komatsu ist einer der weltweit führenden Hersteller von Baumaschinen. Das Unternehmen entwickelte eine intelligente Fabrikplattform namens KOM-MICS, um die betrieblichen Bedingungen in Fabriken durch die Erfassung von Daten mittels Tools und Robotern zu visualisieren und zu optimieren. Komatsu nutzt seit Januar 2017 Azure, um die von KOM-MICS generierten Daten zu erfassen und zu speichern. Komatsu verwendet Azure SQL Database, um die von den einzelnen KOM-MICS-Systemen gesammelten Stammdaten zu verwalten und anschließend mithilfe von Azure Synapse Analytics zu aggregieren. Diese Daten werden dann mittels Power BI visualisiert.

Azure Cosmos DB

Wie in *Kapitel* 5 „*Geschäftliche Anwendungsfälle*“ beschrieben, ist Azure Cosmos DB ein schlüsselfertiger, global verteilter NoSQL DB-Dienst, der für die Verarbeitung von schemaunabhängigen Daten verwendet werden kann. Der Dienst ermöglicht Ihnen die Verwendung von Schlüssel-Wert-Tabellen, Grafiken und Dokumentdaten mit mehreren Konsistenzebenen, um die Anforderungen Ihrer Anwendungen zu erfüllen. Er bietet eine nur minimale Latenz, eine hohe Verfügbarkeit und unbegrenzte Skalierbarkeit. Dank der Datenreplikation auf globaler Ebene sind die Anwender so nahe wie möglich an den Daten und können so vom weltweit schnellsten Zugriff auf Daten profitieren. „Core to Azure Cosmos DB“ ist ein Konsistenzmodellkonzept, bei dem Sie bis zu fünf Konsistenzmodelle für Ihre Anwendung auswählen können, um eine Datenkonsistenz und zuverlässige Leistung zu gewährleisten, die sich durch Folgendes auszeichnet:

- Starke Konsistenz
- Begrenzte Veraltung
- Sitzung
- Konsistentes Präfix
- Letztendliche Datenkonsistenz

Azure Cosmos DB indiziert automatisch alle Daten für Sie, unabhängig davon, welches Modell Sie wählen. Sie können Speicher und Durchsatz für verschiedene Azure-Regionen global, unabhängig und elastisch skalieren. Damit werden Durchsatz, Konsistenz und hohe Verfügbarkeit planbar. Der Dienst ist ideal für Szenarien geeignet, in denen Sie mit global verteilten unternehmenskritischen Anwendungen arbeiten, sowie für E-Commerce-, IoT-, Mobil- und Gaming-Anwendungen.

Archive 360 ist ein Unternehmen, das Kunden bei der Migration großer Datenmengen aus älteren On-Premises-Systemen zur Cloud und deren Verwaltung unterstützt. Das Unternehmen nutzt Cosmos DB, um unterschiedliche Datentypen zu verarbeiten und Kunden zu helfen, ihre Daten zu organisieren, zu klassifizieren und zu verwalten.

Vorbereiten und Trainieren der Daten

Nachdem Sie die Daten aufgenommen und gespeichert haben, benötigen Sie eine Methode, um die Daten für Ihre Zwecke zu transformieren. An dieser Stelle kommen skalierbare Compute-Engines zum Tragen.

Sie können Ihre Daten und Datenspeicher vorbereiten und trainieren, um Insights abzuleiten und mittels Machine Learning- und Deep Learning-Techniken prädiktive und präskriptive Modelle anhand Ihrer Daten zu erstellen.

Der nächste Abschnitt bietet eine kurze Zusammenfassung der in diesem Buch behandelten Dienste.

Azure Databricks

Azure Databricks ist ein Analytics-Clusterdienst, der auf Apache Spark basiert und Ihnen mit kollaborativen Notebooks und Unternehmensfunktionen alle Vorteile von Spark bereitstellt. Dieses leistungsstarke Tool kann auf skalierbare Weise komplexe Transformationen ausführen. Der Dienst kann mit Azure Active Directory und nativen Connectors integriert werden, um weitere Azure-Datendienste einzuführen. Im Folgenden finden Sie einige Beispiele für Azure Databricks-Daten-Connectors:

- Azure SQL Database
- Azure Synapse Analytics
- Azure HDInsight (HDFS)
- SQL Server-Datenbank
- SQL Server Analysis Services-Datenbank
- MySQL-Datenbank
- Oracle-Datenbank
- Access-Datenbank
- Excel
- Text/CSV
- XML
- JSON
- Dateiordner
- PostgreSQL-Datenbank
- Sybase-Datenbank
- Tera-Datenbank

Im Fall einer Big-Data-Pipeline können Sie die Daten (strukturiert oder unstrukturiert) über Azure Data Factory in Batches zu Azure übertragen oder beinahe in Echtzeit mittels IoT Hub, Event Hub oder Kafta zu Azure streamen. Wenn Sie die Daten langfristig und persistent speichern möchten, können Sie diese in Azure Data Lake Storage oder Azure Blob Storage speichern. Azure Databricks liest anschließend die Daten und erzeugt mit Spark bahnbrechende Insights. Azure Databricks stellt eine ausgezeichnete Plattform für die Erzeugung prädiktiver Modelle mit datenwissenschaftlichen Methoden dar.

Zeiss, ein führender Hersteller von High-End-Optiksystemen, hatte Schwierigkeiten mit der Skalierung seiner Infrastruktur, um massive Mengen unstrukturierter Daten analysieren zu können. Azure Databricks stellte dem Unternehmen eine einheitliche Analytics-Plattform bereit, die dessen Skalierungsprobleme löste und einen vollständig verwalteten, hoch skalierbaren und konsistenten Service bereitstellte. Dank Azure Databricks kann Zeiss Batchdaten und unstrukturierte IoT-Daten kombinieren und die Datenverarbeitung vereinfachen.

Azure Stream Analytics

Azure Stream Analytics verwendet eine leistungsstarke Ereignisverarbeitungs-Engine, die Muster und Beziehungen aus Echtzeitdaten analysiert, die aus Geräten, Sensoren und mehr erfasst werden. Zusammen mit Azure Event Hubs können Sie mit Azure Stream Analytics Millionen von Ereignissen erfassen und Anomalien entdecken, Dashboards erstellen, Muster erkennen oder ereignisgesteuerte Aufgaben mithilfe einer SQL-ähnlichen Sprache in Echtzeit automatisieren.

Wie bereits erwähnt, kann die Piraeus Bank, ein griechisches Finanzinstitut, die User Experience und den Onlinepfad von Kunden mithilfe von Azure Stream Analytics und Power BI überwachen.

Azure Cognitive Services

Azure Cognitive Services ermöglichen Entwicklern, mithilfe einer Reihe vorab entwickelter KI-Dienste für Bilderkennung, Sprechen, Text, Sprache, Wissen und Suche kognitive Funktionen auf einfache Weise in ihre Anwendungen zu integrieren. Der Vorteil von Azure Cognitive Services besteht darin, dass Bilderkennung und Textübersetzung als Teil des Daten-Analytics-Prozesses integriert werden können.

Beispielsweise nutzt IndiaLends, ein digitaler Marktplatz für Kredite und Darlehen, Analytics-Algorithmen, um Indiens führende Banken mit Millionen von Kreditnehmern in Kontakt zu bringen. Das Unternehmen nutzt Azure Cognitive Services für Aufgaben wie Textverarbeitung, Bildverarbeitung und Stimmungsanalyse, um Kundenabfragen besser auflösen zu können.

Azure Machine Learning

Im Fall erweiterter Analytics stellen Ihnen Azure Machine Learning und Microsoft Machine Learning Server die nötigen Infrastrukturen und Tools bereit, um Daten zu analysieren, hochwertige Datenmodelle zu erstellen und Machine-Learning-Systeme zu trainieren und zu orchestrieren, während Sie intelligente Apps und Dienste entwickeln. Azure Machine Learning stellt die prädiktive Intelligenz bereit, die Unternehmen benötigen, um wettbewerbsfähig zu bleiben.

British Petroleum nutzt KI- und Microsoft Machine Learning-Algorithmen, um die Menge der Kohlenwasserstoffe zu prognostizieren, die aus potenziellen Öl- und Erdgasreservoirs extrahiert werden können. Solche Prognosen erforderten ursprünglich eine manuelle Analyse von beinahe 200 verschiedenen Eigenschaften der betreffenden Reservoirs. Ein dreistufiger Prozess bestehend aus der Auswahl der richtigen Variablen, der Entwicklung des Algorithmus und der Verbesserung der Modellleistung durch die Anpassung von Modellparametern hilft BP, den Prozess zu vereinfachen und ein Modell zu entwickeln, das alle möglichen Szenarien berücksichtigt.

Modellieren und Bereitstellen der Ergebnisse

Im Anschluss an das Trainieren und Vorbereiten analysieren Sie Ihre Daten und leiten Insights aus diesen ab. Der nächste Schritt besteht darin, Ihren Anwendern diese erweiterten Daten bereitzustellen.

Das beste Ziel für all diese analysierten Daten stellt Azure Synapse Analytics dar. Mit Azure Synapse Analytics können Sie historische Trends mit neuen Insights zu einer einzigen Version der Daten für Ihr Unternehmen kombinieren. Der Dienst ist darüber hinaus dank seiner Fähigkeit, nahtlos Verbindungen mit Analytics-Tools und -Diensten herzustellen, erweiterbar und flexibel. Darüber hinaus kann er mit Business Intelligence-Tools integriert werden.

Visualisierung und mehr

Azure Analysis Services und Power BI stellen leistungsstarke Optionen bereit, um Data Insights zu finden und zu teilen. Azure Synapse Analytics ist die Engine, die diese Insights bereitstellt. Power BI ist hingegen ein Visualisierungstool, das Anwendern das Analysieren von Daten ermöglicht.

Darüber hinaus können Sie Daten aus Ihren Insights in operative Datenspeicher wie Azure SQL Database und Azure Cosmos DB übertragen, um benutzerdefinierte Web- und Anwendungserfahrungen zu optimieren.

Mit Azure-Plattformtools für Entwickler wie Visual Studio, Azure Machine Learning Studio oder benutzerdefinierten serverlosen Apps und Diensten, die Azure Functions verwenden, können Sie Daten sogar direkt zu Ihren Apps übertragen.

Um den Datenzugriff zu schützen, können Sie sicherstellen, dass Benutzer von Azure Active Directory authentifiziert werden und nur bestimmte Benutzergruppen die Daten basierend auf Ihren Spezifikationen nutzen dürfen.

Zusammenfassung

In diesem Buch haben Sie die verschiedenen Phasen kennengelernt, aus denen der Daten-Analytics-Vorgang in der Cloud besteht. Die beispielhaften Implementierungen und Anwendungsfälle zeigen, wie reale Unternehmen Azure-Technologien in verschiedenen Bereichen nutzen, um Daten optimal zu nutzen. So erhalten Sie eine Vorstellung davon, wie Sie diese leistungsstarke Technologie für Ihr eigenes Unternehmen verwenden können.

Das Cloud-Modell für das moderne Data Warehousing ist nicht nur flexibel und skalierbar, sondern aufgrund seiner einzigartigen elastischen Eigenschaften auch kostengünstig. Analytics-Workloads sind ein Szenario, in dem Elastizität wirkliche Vorteile bietet.

Nachdem Sie nun das Ende dieses Buchs erreicht haben, besitzen Sie die nötigen Kenntnisse in Bezug auf das moderne Data Warehouse und die Dienste und Tools, um eine eigene vollständige Datenanalyselösung entwickeln zu können. Sie sollten klein beginnen, indem Sie einige der in diesem Buch vorgestellten Technologien in Ihren Workflow integrieren und dann in der Zukunft schrittweise weitere Funktionen hinzufügen, wenn sich Ihre Anforderungen weiterentwickeln.

Erfahren Sie mehr über den Lebenszyklus eines modernen Data Warehouse.

Erstellen Sie ein kostenfreies Azure-Abonnement.

Viel Erfolg mit Ihrem modernen Warehouse und Ihrem Weg zu Cloud-Analytics!