Cloud Analytics with Microsoft Azure

Second Edition

Transform your business with the power of analytics in Azure

Has Altaiar, Jack Lee and Michael Peña

Cloud Analytics with Microsoft Azure, Second Edition

Transform your business with the power of analytics in Azure

Has Altair, Jack Lee and Michael Peña

Packt›

Cloud Analytics with Microsoft Azure, Second Edition

Authors: Has Altair, Jack Lee and Michael Peña

Technical Reviewer: Aaditya Pokkunuri

Managing Editors: Aditya Datar and Neha Pande

Acquisitions Editor: Ben Renow-Clarke

Production Editor: Deepak Chavan

Editorial Board: Vishal Bodwani, Ben Renow-Clarke, Arijit Sarkar, Dominic Shakeshaft and Lucy Wan

First Published: October 2019

Production Reference: 2221220

ISBN: 978-1-80020-243-6

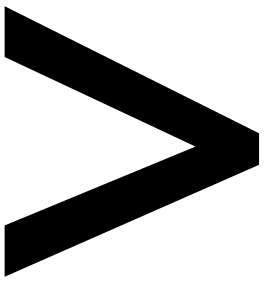
Published by Packt Publishing Ltd.

Livery Place, 35 Livery Street

Birmingham B3 2PB, UK

Table of Contents

Chapter 4: Business use cases 105



Preface

About

This section briefly introduces the authors and reviewer, the coverage of this book, the technical skills you'll need to get started and the hardware and software requirements needed to complete all of the activities and exercises.

About Cloud Analytics with Microsoft Azure, Second Edition

Cloud Analytics with Microsoft Azure serves as a comprehensive guide to processing and analysing big data using a range of Microsoft Azure features. This book covers everything you need to build your own data warehouse and learn numerous techniques to gain useful insights by analysing big data.

The book begins by introducing you to the power of data with big data analytics, the **Internet of Things** (**IoT**), machine learning, artificial intelligence and DataOps. You will learn about cloud-scale analytics and the services Microsoft Azure offers to empower businesses to discover insights. You will also be introduced to the new unified experience in the Azure Synapse workspace and Synapse Studio with a practical hands-on guide.

Finally, you will look at two real-world business use cases to demonstrate high-level solutions using Microsoft Azure. The aim of these use cases will be to illustrate how real-time data can be analysed in Azure to derive meaningful insights and make business decisions. You will learn to build an end-to-end analytics pipeline on the cloud with machine learning and deep learning concepts.

By the end of this book, you will be proficient in analysing vast data pools with Azure and using it effectively to benefit your organisation.

About the authors

Has Altaiar is a software engineer at heart and a consultant by trade. Has lives in Melbourne, Australia and is the Executive Director at vNext Solutions. His work focuses on data, IoT and AI on Microsoft Azure and two of his latest IoT projects won multiple awards. Has is a Microsoft Azure MVP and a regular organiser and speaker at local and international conferences, including Microsoft Ignite, NDC and ServerlessDays. He's also a board member of the Global AI Community. You can follow him on Twitter at `@hasaltaiar`.

Jack Lee is a senior Azure certified consultant and an Azure practice lead with a passion for software development, cloud and DevOps innovations. He is an active Microsoft tech community contributor and has presented at various user groups and conferences, including the Global Azure Bootcamp at Microsoft Canada. Jack is an experienced mentor and judge at hackathons and is also the president of a user group that focuses on Azure, DevOps and software development. He is the co-author of *Azure for Architects* and *Cloud Analytics with Microsoft Azure*, published by Packt Publishing. He has been recognised as a Microsoft MVP for his contributions to the tech community. You can follow Jack on Twitter at `@jlee_consulting`.

Michael Peña is an experienced technical consultant based in Sydney, Australia. He is a Microsoft MVP and a certified professional with over 10 years of experience in data, mobile, cloud, web and DevOps. Throughout these years, he wore various hats, but considered himself a developer at heart. He is also an international speaker, having spoken at numerous events, including Microsoft Ignite, NDC, DDD, Cross-Platform Summit and various in-person and virtual meet-ups. Michael has interned with Microsoft and is also a Microsoft student partner alumnus. You can follow him on Twitter at **`@mjtpena`**.

About the reviewer

Aaditya Pokkunuri is an experienced senior database engineer with a history of working in the information technology and services industry; he has a total of 11 years of experience. He is skilled in performance tuning, MS SQL Database server administration, SSIS, SSRS, Power BI and SQL development.

He possesses strong knowledge about replication, clustering, SQL Server high availability options and ITIL processes, as well as expertise in Windows administration tasks, Active Directory and Microsoft Azure technologies.

He also has expertise in AWS Cloud and is an AWS Solution Architect Associate. Aaditya is a strong information technology professional with a Bachelor of Technology degree focused on computer science and engineering from Sastra University, Tamil Nadu.

Learning objectives

- Explore the concepts of modern data warehouses and data pipelines
- Discover unique design considerations while applying a cloud analytics solution
- Design an end-to-end analytics pipeline on the cloud
- Differentiate between structured, semi-structured and unstructured data
- Choose a cloud-based service for your data analytics solutions
- Use Azure services to ingest, store and analyse data of any scale

Audience

This book is designed to benefit software engineers, developers, cloud consultants and anyone who is keen to learn the process of deriving business insights from big data using Azure.

Though not necessary, a basic understanding of data analytics concepts such as data streaming, data types, the machine learning life cycle and Docker containers will help you get the most out of the book.

Approach

Cloud Analytics with Microsoft Azure introduces complex concepts with real-world examples so that you get hands-on experience while also understanding the fundamentals. The book contains numerous quick-start guides that enable you to learn faster.

Hardware and software requirements

Hardware requirements

For the optimal student experience, we recommend the following hardware configuration:

- Memory: Minimum 4 GB RAM
- Display: Minimum 1440 × 900 or 1600 × 900 (16:9) recommended
- CPU: 1 gigahertz (GHz) or faster x86- or x64-bit processor recommended

Software requirements

We also recommend that you have the following software configuration in advance:

- Windows 10 latest version or Windows Server latest version
- Azure subscription. You can set up a free Azure account at https://azure.microsoft.com/free/synapse-analytics/
- Microsoft Edge latest version

Conventions

Code words in the text, database names, folder names, filenames and file extensions are shown as follows.

The following code snippet makes use of the Azure SQL Database linked service to create a dataset that references **sales_table** in **Coolies'** SQL Database:

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

Installation and set-up

You can install Power BI desktop (https://packt.live/37hUTmK) and start creating interactive reports.

1

Introducing analytics on Azure

According to a survey by Dresner Advisory Service in 2019, an all-time high of 48% of organisations say business intelligence in the cloud is either critical or very important in conducting their business operations. The *Cloud Computing and Business Intelligence Market Study* also showed that sales and marketing teams get the most value out of analytics.

As businesses grow, they generate massive amounts of data every day. This data comes from different sources, such as mobile phones, the **Internet of Things** (**IoT**) sensors and various **Software-as-a-Service** (**SaaS**) products such as **Customer Relationship Management** (**CRM**) systems. Enterprises and businesses need to scale and modernise their data architecture and infrastructure in order to cope with the demand to stay competitive in their respective industries.

Having cloud-scale analytics capabilities is the go-to strategy for achieving this growth. Instead of managing your own data centre, harnessing the power of the cloud allows your businesses to be more accessible to your users. With the help of a cloud service provider such as Microsoft Azure, you can accelerate your data analytics practice without the limitations of your IT infrastructure. The game has changed in terms of maintaining IT infrastructures, as **data lakes** and cloud data warehouses are capable of storing and maintaining massive amounts of data.

Simply gathering data does not add value to your business; you need to derive insights from it and help your business grow using data analytics, or it will just be a **data swamp**. Azure is more than just a hub for gathering data; it is an invaluable resource for data analytics. Data analytics provides you with the ability to understand your business and customers better. By applying various data science concepts, such as ML, regression analysis, classification algorithms and time series forecasting, you can test your hypotheses and make data-driven decisions for the future. However, one of the challenges that organisations continuously face is how to derive these analytical modeling capabilities quickly when processing billions of data rows. This is where having a modern data warehouse and data pipeline can help (more on this in the next sections).

There are a number of ways in which data analytics can help your business thrive. In the case of retail, if you understand your customers better, you will have a better idea of what products you should sell, where to sell them, when to sell them and how to sell them. In the financial sector, data analytics is helping authorities fight crime by detecting fraudulent transactions and providing more informed risk assessments based on historical criminal intelligence.

This chapter will cover fundamental topics on the power of data with:

- Big data analytics
- IoT
- Machine Learning (ML)
- Artificial Intelligence (AI)
- DataOps

You will also learn why Microsoft Azure is the platform of choice for performing analytics on the cloud. Lastly, you will study the fundamental concepts of a modern data warehouse and data pipelines.

The power of data

As a consumer, you have seen how the advent of data has influenced our activities in the daily grind. Most popular entertainment applications, such as YouTube, now provide a customised user experience with features such as video recommendations based on our interests and search history logging information. It is now child play to discover new content that similar to our preferred content, and also to find new and popular trending content.

Due to the major shift in wearable technology, it has also become possible to keep track of our health statistics by monitoring heart rates, blood pressure, and so on. These devices then formulate a tailored recommendation based on the averages of these statistics. But these personalised health stats are only a sample of the massive data collection happening every day on a global scale, to which we actively contribute.

Millions of people all over the world use social networking platforms and search engines every day. Internet giants such as Facebook, Instagram and Google use clickstream data to come up with innovations and improve their services. Data collection is also carried out extensively under projects such as *The Great Elephant Census* and *eBird* that aim to boost wildlife conservation. Data-driven techniques have been adopted for tiger conservation projects in India. It even plays an invaluable role in global efforts to compile evidence, causes and possible responses to climate change – to understand sea surface temperature, analyse natural calamities such as coastal flooding and highlight global warming patterns in a collective effort to save the ecosystem.

Organisations such as **Global Open Data for Agriculture and Nutrition** (**GODAN**), which can be used by farmers, ranchers and consumers alike, contribute to this tireless data collection as well.

Furthermore (as with the advent of wearable technology), data analysis is contributing to pioneering advancements in the healthcare sector. Patient datasets are analysed to identify patterns and early symptoms of diseases in order to divine better solutions to known problems.

The scale of data being talked about here is massive – hence, the popular term **big data** is used to describe the harnessing power of this data at scale.

> **Note**
>
> You can read more about open data here.

Big data analytics

The term 'big data' is often used to describe massive volumes of data that traditional tools cannot handle. It can be characterised by the five 'V's:

- **Volume:** This indicates the volume of data that needs to be analysed for big data analytics. We are now dealing with larger datasets than ever before. This has been made possible because of the availability of electronic products such as mobile devices and IoT sensors that have been widely adopted all over the globe for commercial purposes.
- **Velocity:** This refers to the rate at which data is being generated. Devices and platforms, such as those just mentioned, constantly produce data on a large scale and at rapid speed. This makes collecting, processing, analysing and serving data at rapid speeds necessary.
- **Variety:** This refers to the structure of data being produced. Data sources are inconsistent, having a mix of structured, unstructured and some semi-structured data (you will learn more about this in the *Bringing your data together* section).
- **Value:** This refers to the value of the data being extracted. Accessible data may not always be valuable. With the right tools, you can derive value from the data in a cost-effective and scalable way.
- **Veracity:** This is the quality or trustworthiness of data. A raw dataset will usually contain a lot of noise (or data that needs cleaning) and bias and will need cleaning. Having a large dataset is not useful if most of the data is not accurate.

Big data analytics is the process of finding patterns, trends and correlations in unstructured data to derive meaningful insights that shape business decisions. This unstructured data is usually large in file size (images, video and social graphs, for instance).

This does not mean that relational databases are not relevant for big data. In fact, modern data warehouse platforms such as Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse) support structured and semi-structured data (such as JSON) and can infinitely scale to support terabytes to petabytes of data. Using Microsoft Azure, you have the flexibility to choose any platform. These technologies can complement each other to achieve a robust data analytics pipeline.

Here are some of the best use cases of big data analytics:

- **Social media analysis:** Through social media sites such as Twitter, Facebook and Instagram, companies can learn what customers are saying about their products and services. Social media analysis helps companies to target their audiences by utilising user preferences and market trends. The challenges here are the massive amount of data and the unstructured nature of tweets and posts.

- **Fraud prevention:** This is one of the most familiar use cases of big data. One of the prominent features of big data analytics when used for fraud prevention is the ability to detect anomalies in a dataset. Validating credit card transactions by understanding transaction patterns such as location data and categories of purchased items is an example of this. The biggest challenge here is ensuring that the AI/ML models are clean and unbiased. There might be a chance that the model was trained just for a specific parameter, such as a user's country of origin, hence the model will focus on determining patterns on just the user's location and might miss out on other parameters.
- **Price optimisation:** Using big data analytics, you can predict what price points will yield the best results based on historical market data. This allows companies to ensure that they do not price their items too high or too low. The challenge here is that many factors can affect prices. Focusing on just a specific factor, such as a competitor's price, might eventually train your model to just focus on that area, and may disregard other factors such as weather and traffic data.

Big data for businesses and enterprises is usually accompanied by the concept of having an IoT infrastructure, where hundreds, thousands or even millions of devices are connected to a network that constantly sends data to a server.

Internet of Things (IoT)

IoT plays a vital role in scaling your application to go beyond your current data sources. IoT is simply an interconnection of devices that are embedded to serve a single purpose in objects around us to send and receive data. IoT allows us to constantly gather data about 'things' without manually encoding them into a database.

A smartwatch is a good example of an IoT device that constantly measures your body vital signs. Instead of getting a measuring device and encoding it to a system, a smartwatch allows you to record your data automatically. Another good example is a device tracker for an asset that captures location, temperature and humidity information. This allows logistics companies to monitor their items in transit, ensuring the quality and efficiency of their services.

At scale, these IoT devices generate anywhere from gigabytes to terabytes of data. This data is usually stored in a data lake in a raw, unstructured format, and is later analysed to derive business insights. A data lake is a centralised repository of all structured, semi-structured and unstructured data. In the example of the logistic company mentioned previously, patterns (such as the best delivery routes) could be generated. The data could also be used to understand anomalies such as data leakage or suspected fraudulent activities.

Machine learning

As your data grows in size, it opens a lot of opportunities for businesses to go beyond understanding business trends and patterns. Machine learning and artificial intelligence are examples of innovations that you can exploit with your data. Building your artificial intelligence and ML capabilities is relatively easy now because of the availability of the requisite technologies and the ability to scale your storage and compute on the cloud.

Machine learning and artificial intelligence are terms that are often mixed up. In a nutshell, machine learning is a subset (or application) of artificial intelligence. Machine learning aims to allow systems to learn from past datasets and adapt automatically without human assistance. This is made possible by a series of algorithms being applied to the dataset; the algorithm analyses the data in near-real-time and then comes up with possible actions based on accuracy or confidence derived from previous experience.

The word 'learning' indicates that the program is constantly learning from data fed to it. The aim of machine learning is to strive for accuracy rather than success. There are three main categories of machine learning algorithms: **supervised**, **unsupervised** and **reinforcement**.

Supervised machine learning algorithms create a mapping function to map input variables with an output variable. The algorithm uses existing datasets to train itself to predict the output. Classification is a form of supervised ML that can be used in applications such as image categorisation or customer segmentation, which is used for targeted marketing campaigns.

Unsupervised machine learning, on the other hand, is when you let a program find a pattern of its own without any labels. A good example is understanding customer purchase patterns when buying products. You get inherent groupings (**clustering**) according to purchasing behaviours, and the program can associate customers and products according to patterns of purchase. For instance, you may discern that customers who buy Product A tend to buy Product B too. This is an example of a user-based recommendation algorithm and market-based analysis. What it would eventually mean for users is that when they buy a particular item, such as a book, the user is also encouraged to buy other books that belong to the same series, genre or category.

Reinforcement Learning (**RL**) provides meaningful insights and actions based on rewards and punishment. The main difference between this and supervised learning is that it does not need labelled input and output as part of the algorithm. An excellent example of this is the new financial trend for 'robo-advisers'. Robo-advisers run using agents that get rewarded and punished based on their stock performance (that is, gains and losses). In time, the agent can recognise whether to hold, buy or sell stocks. This has been a game-changer because, in the past, analysts had to make every single decision; now most of the complicated data trends are already analysed for you and analysts can choose to listen to the agent or not. However, financial trading is very complex given the nature of parameters present in the world, and so not all robo-advisers' predictions are accurate.

Artificial intelligence (AI)

Artificial intelligence extends beyond what machine learning can do. It is about making decisions and aiming for success rather than accuracy. One way to think of it is that machine learning aims to gain knowledge while artificial intelligence aims for wisdom or intelligence. An example of AI in action would be Boston Dynamic's Atlas robot, which can navigate freely in the open world and avoid obstacles without the aid of human control. The robot does not fully depend on the historical map data to navigate. However, for machine learning, it's about creating or predicting a pattern from historical data analysis. Similar to the robot navigation, it is about understanding the most optimal route by creating patterns based on historical and crowd-sourced traffic data.

Setting up a modern data warehouse with cloud analytics is the key factor in preparing to execute ML/AI. Without migrating the workloads to the cloud, deriving ML/AI models will mean encountering various roadblocks in order to maximise the business value of these emerging technologies. A modern data warehouse and analytics pipeline form the backbone that enables you to pass these roadblocks.

Microsoft is a leader in machine learning and artificial intelligence as they have been driving a lot of innovation throughout their products and tools – for instance, Windows' digital assistant, Cortana and Office 365's live captions and subtitles. They offer a range of products, tools and services such as Microsoft Cognitive Services, Azure Machine Learning studio, the Azure Machine Learning service and ML.NET.

Microsoft is setting an example with their *AI for Good* initiative, which aims to make the world more sustainable and accessible through AI. One particularly interesting project is *AI for Snow Leopards*, in which Microsoft uses AI technology to detect snow leopards (who are almost invisible in snow) in order to protect the endangered species. Exploring artificial intelligence and deep learning (the ability to learn without human supervision), specifically the data science and formula aspects, is not the focus of this book, but you will tackle some concepts in later chapters (see more on this in *Chapter 3, Processing and visualising data*).

DataOps

In order to be efficient and agile with implementing data analytics in your company, you need the right culture and processes. This is where the concept of **DataOps** comes in. DataOps removes the co-ordination barrier between data (analysts, engineers and scientists) and operations (administrators and operations managers) teams in order to achieve speed and accuracy in data analytics.

DataOps is about a culture of collaboration between different roles and functions. Data scientists have access to real-time data to explore, prepare and serve results. Automated processes and flows prove invaluable to this collaborative effort between analysts and developers, as they provide easy access to data through visualisation tools. Relevant data should be served to end users via web or mobile applications; this is usually possible with an **Application Programming Interface** (**API**). For CEOs, DataOps means faster decision-making, as it allows them to monitor their business at a high level without waiting for team leaders to report. *Figure 1.1* tries to explain the idea of a collaborative DataOps culture:

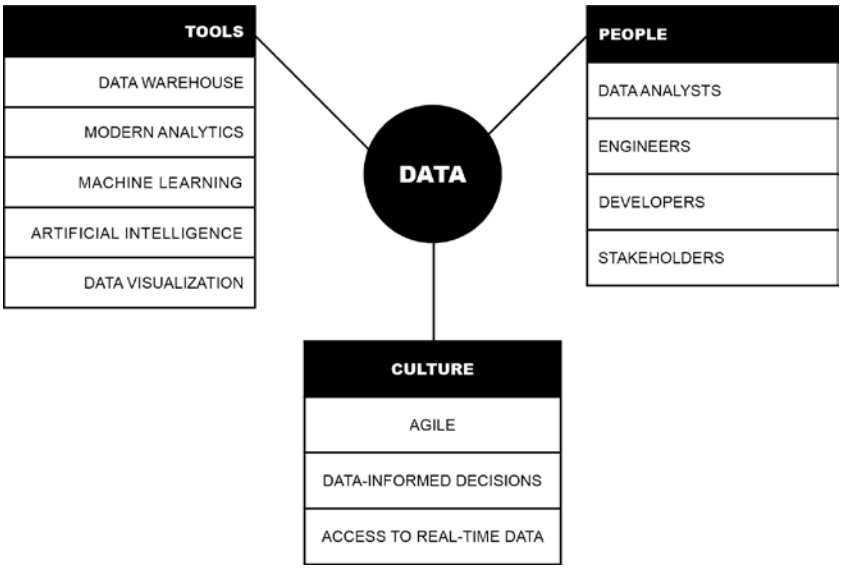


Figure 1.1: The DataOps process

Once a team attains the desired speed and accuracy in testing their hypotheses (such as the likelihood of someone buying a product based on their characteristics and behaviour), they are able to derive better insights. Once there are better insights, there are more actionable and reasonable decision points for business stakeholders that minimise risks and maximise profits.

Why Microsoft Azure?

Microsoft Azure is an enterprise-grade set of cloud computing services created by Microsoft using their own managed data centres. Azure is the only cloud with a true end-to-end analytics solution. With Azure, analysts can derive insights in seconds from all enterprise data. Azure provides a mature and robust data flow without limitations on concurrency.

Azure supports **Infrastructure-as-a-Service** (**IaaS**), **Platform-as-a-Service** (**PaaS**) and **SaaS**. Many government institutions across the world, as well as 95% of Fortune 500 companies, use Azure, ranging from industries such as healthcare and financial services to retail and manufacturing.

Microsoft is a technology conglomerate that has empowered many people to achieve more with less for decades with their software, tools and platforms. Azure provides flexibility. Familiar Microsoft tools and infrastructures (such as SQL Server, Windows Server, **Internet Information Services** (**IIS**) and .NET) or tools such as MySQL, Linux, PHP, Python, Java or any other open source technologies can all run on the Azure cloud. Gone are the days when you could only work on a walled-garden set of tools and technologies.

Azure provides you with various products and services, depending on your needs. You have the option of doing everything in a bespoke way, from managing your IaaS by spinning up Windows Server virtual machines with Enterprise SQL Server installed, to using a managed PaaS offering such as Azure Synapse Analytics.

Figure 1.2 shows the wide range of data-specific Azure tools and services that can be used to create end-to-end data pipelines:

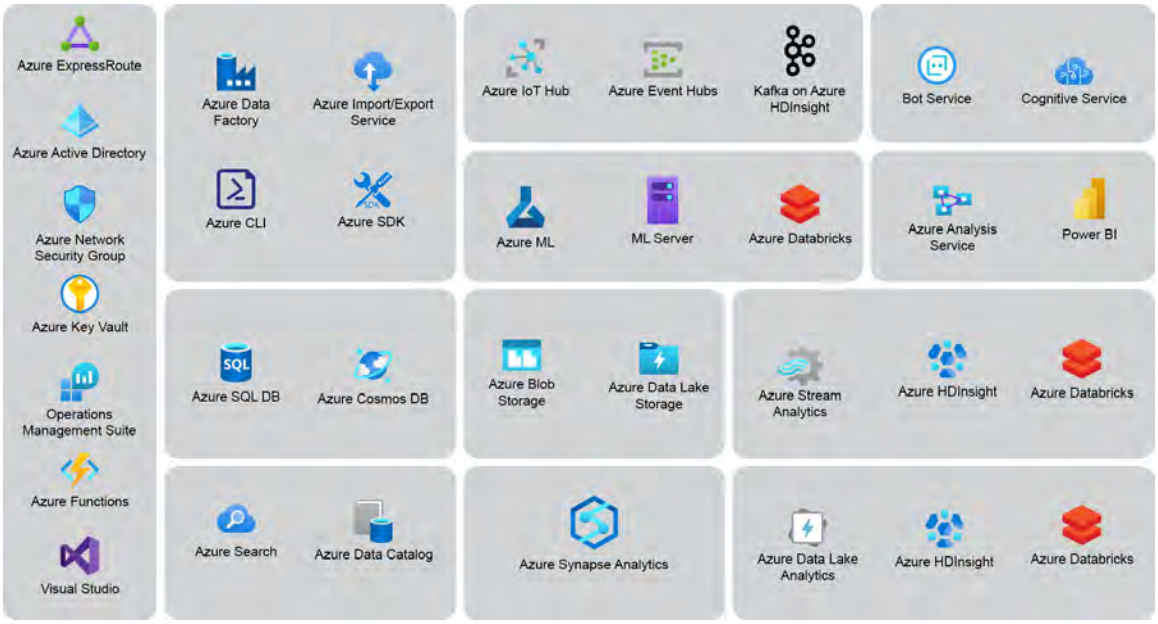


Figure 1.2: Microsoft Azure data-related services

Azure grants you the flexibility to choose the best approach to solve a problem for yourself, rather than being forced to bend a less adaptable product to perform an unnatural function. You're not just limited to SQL Server, either. You also have the flexibility to choose other types of databases or storage, whether through a service installed on a Linux server or containerised solution, or a managed platform (such as Azure Cosmos DB for your Cassandra and MongoDB instances). This is very important because, in the real world, different scenarios require different solutions, tools and products.

Microsoft Azure provides you with an end-to-end platform, from Azure Active Directory for managing your user identity and access to Azure IoT offerings (such as IoT Hub) for gathering data from hundreds and thousands of IoT devices. It also provides services such as development tools and cloud hosting options for getting your developers up to speed, as well as various analytics and machine learning tools that enable data scientists, data engineers and data analysts to be more productive (more on this in *Chapter* 3, *Processing and visualising data*).

The full spectrum of Azure services is too wide to cover here, so instead, this book will focus on the key data warehousing and business intelligence suite of products: Azure Data Lake, Azure Synapse Analytics, Power BI and Azure Machine Learning.

Security

Microsoft views security as the top priority. When it comes to data, privacy and security are non-negotiable; there will always be threats. Azure has the most advanced security and privacy features in the analytics space. Azure services support data protection through **Virtual Networks** (**VNets**) so that, even though they are in the cloud, data points cannot be accessed by the public internet. Only the users in the same VNet can communicate with each other. For web applications, you get a **Web Application Firewall** (**WAF**) provided by Azure Application Gateway, which ensures that only valid requests can get into your network.

With role-based access control (**authorisation**), you can ensure that only those with the right roles, such as administrators, have access to specific components and the capabilities of different resources. **Authentication**, on the other hand, ensures that if you don't have the right credentials (such as passwords), you will not be able to access a resource. Authorisation and authentication are built into various services and components of Microsoft Azure with the help of Azure Active Directory.

Azure also provides a service called **Azure Key Vault**. Key Vault allows you to safely store and manage secrets and passwords, create encryption keys and manage certificates so that applications do not have direct access to private keys. By following this pattern with Key Vault, you do not have to hardcode your secrets and passwords in your source code and script repository.

Azure Synapse Analytics uses ML and AI to protect your data. In Azure SQL, Microsoft provides advanced data security to ensure that your data is protected. This includes understanding if your database has vulnerabilities, such as port numbers that are publicly available. These capabilities also allow you to be more compliant with various standards, such as **General Data Protection Regulation** (**GDPR**), by ensuring that customer data that are considered sensitive are classified. Azure SQL also recently announced their new features, **row-level security** (**RLS**) and **column-level security** (**CLS**), to control access to rows and columns in a database table, based on the user characteristics.

Microsoft invests at least USD 1 billion each year in the cybersecurity space, including the Azure platform. Azure holds various credentials and awards from independent assessment bodies, which means that you can trust Azure in all security aspects, from physical security (such that no unauthorised users can get physical access to data centres) to application-level security.

These are a few security features that you need to consider if you are maintaining your own data centre.

Cloud scale

Azure changed the industry by making data analytics cost-efficient. Before the mass adoption of cloud computing, in order to plan for data analytics with terabytes, or even petabytes, of data, you needed to properly plan things and ensure that you had the capital expenditure to do it. This would mean very high upfront infrastructure and professional services costs, just to get started. But with Azure, you can start small (many of the services have free tiers). You can scale your cloud resources effortlessly up or down, in or out, in minutes. Azure has democratised scaling capability by making it economically viable and accessible for everyone, as shown in *Figure 1.3*:

Figure 1.3: Microsoft Azure regions

Microsoft Azure currently has over 60 data centre regions supporting over 140 countries. Some enterprises and business industries require that your data is hosted within the same country as business operations. With the availability of different data centres worldwide, it is easy for you to expand to other regions. This multi-region approach is also beneficial in terms of making your applications highly available.

The true power of the cloud is its elasticity. This allows you to not only scale resources up, but also scale them down when necessary. In data science, this is very useful because data science entails variable workloads. When data scientists and engineers are analysing a dataset, for instance, there is a need for more computation. Azure, through services such as Azure Machine Learning (more on this in *Chapter 3*, *Processing and visualising data*), allows you to scale according to demand. Then, during off-peak times (such as weekends, and 7 PM to 7 AM on weekdays), when the scientists and engineers don't need the processing power to analyse data, you can scale down your resources so that you don't have to pay for running resources 24/7. Azure basically offers a pay-as-you-go or pay-for-what-you-use service.

Azure also provides a **Service Level Agreement** (**SLA**) for their services as their commitments to ensure uptime and connectivity for their production customers. If downtime or an incident occurs, they will apply service credits (rebates) to the resources that were affected. This will give you peace of mind as your application will always be available with a minimal amount of downtime.

There are different scaling approaches and patterns that Microsoft Azure provides:

- **Vertical scaling:** This is when more resources are added to the same instance (server or service). An example of this is when a virtual machine is scaled up from 4 GB of RAM to 16 GB of RAM. This is a simple and straightforward approach to take when your application needs to scale. However, there is a technical maximum limit on how much an instance can be scaled up, and it is the most expensive scaling approach.
- **Horizontal scaling:** This is when you deploy your application to multiple instances. This would logically mean that you can scale your application infinitely because you don't use a single machine to perform your operations. This flexibility also introduces some complexities. These complexities are usually addressed by using various patterns and different orchestration technologies, such as Docker and Kubernetes.
- **Geographical scaling:** This is when you scale your applications to different geographical locations for two major reasons: resilience and reduced latency. Resilience allows your application to freely operate in that region without all resources being connected to a master region. Reduced latency would mean users of that region can get their web requests faster because of their proximity to the data centre.
- **Sharding:** This is one of the techniques for distributing huge volumes of related, structured data onto multiple independent databases.

- **Development, Testing, Acceptance and Production (DTAP):** This is the approach of having multiple instances living in different logical environments. This is usually done to separate development and test servers from staging and production servers. Azure DevTest Labs offers a development and testing environment that can be configured with group policies.

Another advantage of your business being in the cloud is the availability of your services. With Azure, it is easier to make your infrastructure and resources **geo-redundant** – that is, available to multiple regions and data centres across the world. Say you want to expand your business from Australia to Canada. You can achieve that by making your SQL Server geo-redundant so that Canadian users do not need to query against the application and database instance in Australia.

Azure, despite being a collective suite of products and service offerings, does not force you to go 'all in'. This means that you can start by implementing a hybrid architecture of combined on-premises data centres and cloud (Azure). There are different approaches and technologies involved in a hybrid solution, such as using **Virtual Private Networks (VPNs)** and Azure ExpressRoute, if you need dedicated access.

With Azure Synapse Analytics, through data integrations, Azure allows you to get a snapshot of data sources from your on-premises SQL Server. The same concept applies when you have other data sources from other cloud providers or SaaS products; you have the flexibility to get a copy of that data to your Azure data lake. This flexibility is highly convenient because it does not put you in a **vendor lock-in** position where you need to do a full migration.

Top business drivers for adopting data analytics in the cloud

Different companies have different reasons for adopting data analytics using a public cloud such as Microsoft Azure. But more often than not, it boils down to three major reasons: rapid growth and scale, reducing costs and driving innovation.

Rapid growth and scale

Enterprises and businesses need to rapidly expand their digital footprint. With the rapid growth of mobile applications – particularly media types (such as images and videos), IoT sensors and social media data – there is just so much data to capture. This means enterprises and businesses need to scale their infrastructure to support these massive demands. Company database sizes continuously grow from gigabytes of data to terabytes, or even petabytes, of data.

End users are more demanding now than ever. If your application does not respond within seconds, the user is more likely to disengage with your service or product.

Scaling does not only apply to the consumers of the applications; it is also important for data scientists, data engineers and data analysts in order to analyse a company's data. Scaling an infrastructure is vital, as you cannot expect your data engineers to handle massive chunks of data (gigabytes to terabytes) and run scripts to test your data models on a single machine. Even if you do serve this in a single high-performance server instance, it going to take weeks or days for it to finish the test, not to mention the fact that it going to cause performance bottlenecks for the end users who are consuming the same database.

With a modern data warehouse like Azure Synapse Analytics, you have some managed capabilities to scale, such as a dedicated caching layer. Caching will allow analysts, engineers and scientists to query faster.

Reducing costs

Due to scaling demands, enterprises and businesses need to have a mechanism to expand their data infrastructure in a cost-effective and financially viable way. It is too expensive to set up an on-premises data warehouse. The following are just a few of the cost considerations:

- The waiting time for server delivery and associated internal procurement processes
- Networking and other physical infrastructure costs, such as hardware cooling and data center real estate
- Professional services costs associated with setting up and maintaining these servers
- Licensing costs (if any)
- The productivity lost from people and teams who cannot ship their products faster

With a modern data warehouse, you can spin up new high-performance servers with high-performance graphics cards on demand. And with the use of a cloud provider such as Microsoft Azure, you will only need to pay for the time that you use these servers. You can shut them down if you don't need them anymore. Not only can you turn them off on demand, but if it turns out that a particular service is not suitable for your requirements, you can delete these resources and just provision a different service.

Azure also provides a discount for 'reserved' instances that you are committing to use for a specific amount of time. These are very helpful for those databases, storage solutions and applications that need to be up 24/7 with minimal downtime.

Driving innovation

Companies need to constantly innovate in this very competitive market, otherwise someone else will rise up and take the market share. But obviously, no one can predict the future with 100% accuracy; hence, companies need to have a mechanism to explore new things based on what they know.

One good example of this is the **Business Process Outsourcing** (**BPO**) and **telecommunications** (**telco**) industries, where there are petabytes of data that may not have been explored yet. With Microsoft Azure's modern data warehouse, actors in such industries can have the infrastructure to do data exploration. With Azure Synapse Analytics, Power BI and Azure Machine Learning, they can explore their data to drive business decisions. Maybe they can come up with a data model that can detect fraudulent actions or better understand their customer preferences and expectations to improve satisfaction ratings. With advanced analytics, these companies can come up with decisions that are relevant today (and possibly in the future) and are not just restricted to analysing historical data.

What if you want to create an autonomous vehicle? You will need a robust data warehouse to store your datasets and a tremendous amount of data processing. You need to capture massive amounts of data – whether through pictures or videos that the car is continuously capturing – and need to come up with a response almost instantly based on your dataset and algorithms.

Using a cloud provider such as Microsoft Azure would allow you to test and validate your ideas early on, without a massive investment. With various Azure services and related tools such as GitHub and Visual Studio, you can rapidly prototype your ideas and explore possibilities. What if it turns out that the product or service that you or your team is working on does not really gain traction? If you are doing this on-premises, you will still have high liability and operations costs since you physically own the infrastructure, in addition to any associated licensing and services costs.

Why do you need a modern data warehouse?

A data warehouse is a centralised repository that aggregates different (often disparate) data sources. The main difference between a data warehouse and a database is that data warehouses are meant for **Online Analytical Processing** (**OLAP**) and databases, on the other hand, are intended for **Online Transaction Processing** (**OLTP**). OLAP means that data warehouses are primarily used to generate analytics, business intelligence and even machine learning models. OLTP means that databases are primarily used for transactions. These transactions are the day-to-day operations of applications, which concurrently read and write data to databases.

A data warehouse is essential if you want to analyse your big data as it also contains historical data (often called **cold data**). Most data that stored has legacy information, such as data stored five years ago, 10 years ago or even 15 years ago. You probably don't want the same database instance that your end users are querying against to also contain that historical data, as it might affect your performance when at scale.

Here are some of the advantages of having a modern data warehouse:

- Supports any data source
- Highly scalable and available
- Provides insights from analytical dashboards in real-time
- Supports a machine learning environment

Microsoft offers the following tools and services that collectively create a modern data warehouse:

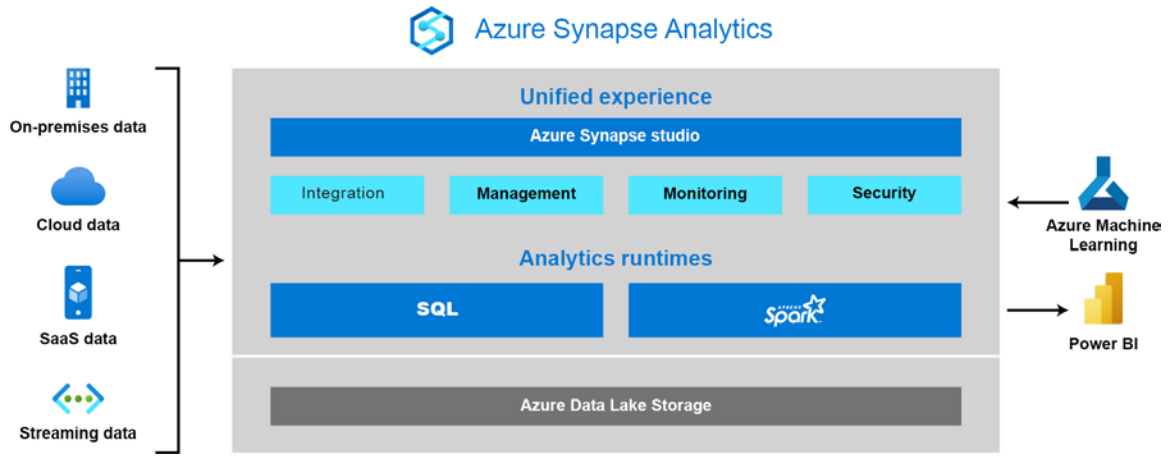


Figure 1.4: Modern data warehouse using Azure Synapse Analytics

There are a lot of emerging patterns and architectures for data warehousing, but the most popular ones are those that support the separation of duties and responsibilities in different phases of the data pipeline (more on this in the *Creating a data pipeline* section).

In order to understand what it means for a data warehouse to be modern, you first need to understand how you create and manage a traditional one. It boils down to two major concepts:

- **Compute:** This refers to the ability to process the data and make sense out of it. It can be in the form of a database query to make the results accessible to another interface, such as web applications.
- **Storage:** This refers to the ability to keep data in order for it to be accessible at any time in the future.

A modern data warehouse separates compute and storage in cost-effective ways. Unlike the case traditionally with SQL Server and **SQL Server Integration Services** (**SSIS**), the pricing model involves both the storage capacity and computing power to analyse data. Azure is the first cloud provider to offer a data warehouse that separates compute and storage.

Another change in pattern is that the traditional **Extract-Transform-Load** (**ETL**) model of data warehousing has now changed to **Extract-Load-Transform** (**ELT**). In the traditional ETL model, analysts are accustomed to waiting for the data to be transformed first, since they don't have direct access to all data sources. In a modern data warehouse, massive amounts of data can be stored in either a data lake or data warehouse, and can be transformed anytime by analysts without the need to wait for data engineers or database admins to serve the data.

Of course, there are more factors to consider in order to modernise your data warehouse, such as extensibility, disaster recovery and availability. However, this section will focus on compute for the time being.

Bringing your data together

In the past, databases were often the only source of data for your applications. But nowadays, you have hundreds and thousands of different data sources. The data coming from these different sources is of different data types – some structured, some unstructured, some semi-structured.

Structured data: The word 'structured' suggests that there is a pattern that can be easily interpreted. This usually comes with a predefined set of models and a schema. A **relational database management system** (**RDBMS**) such as Microsoft SQL Server is a common example of a data storage solution that is structured. This is because it comes with a database schema and table columns that define the data that you are storing.

Here are some examples of structured data types:

- Customer names
- Addresses
- Geolocation
- Date and time
- Mobile and phone numbers
- Credit card numbers
- Product names and **Stock Keeping Units** (**SKUs**)
- General transaction information such as 'From' and 'To' with time stamps and amount values

A good example of structured data is the information provided by the users when signing up to an application for the first time. They are presented with a form that needs to be filled in. Once that person clicks the submit button, it sends the data to a database and inserts it into a user table with predefined columns: names, addresses and other details. This will then allow the user to log in to the application since the system can now look up the existing record for the registered user in the database.

From there, a user can access the application and perform transactions, such as transferring money and assets. In time, users will generate a series of transactions that will eventually make your database larger. Your database schema will also expand to support different business requirements.

Once you have enough data, you can perform data exploration. This is where you start looking for patterns in data. You may identify fraudulent transactions and test hypotheses by analysing large and repeated transaction amounts from the same user.

Your data exploration is limited because you can only base it on a dataset that is structured and with a semantic form. What if you also want to consider other data sources that are unstructured, such as free-form text? An example is a transaction description, which may state the nature or the recipient of the transaction. You don't want to manually read each transaction description and insert it in the right column of a database table. You probably want to extract only the relevant information and transform it into a structured format. This is where unstructured data comes in.

Unstructured data: This data type, more or less, is the rest – that is, everything that isn't structured data. This is mainly because you are not limited to any storage and data type.

Unstructured data types usually don't have a predefined data model that can fit directly into a database. Unstructured data can be text-heavy and is usually read per line or is space-separated.

Here are some examples of unstructured data sources:

- Image files
- Videos
- Email messages and documents
- Log files
- IoT devices and sensors
- NoSQL databases such as MongoDB
- Social media and Microsoft Graph

Image files and videos are classified as unstructured data because of their dynamic nature. Although their metadata is something you can consider as structured (such as title, artist, filename, and so on), the content itself is unstructured. With modern tools and data analytics technology, you can now examine this data and make sense of it. The usual example is face recognition in either images or videos.

Emails, documents and log files all have metadata, but what you're actually more interested in is the content of those files. Usually, in emails, documents and log files, data is separated per line and the messages are unstructured. You would want to describe the content without manually reading everything (which could be hundreds or even millions of files). An example is doing sentiment analysis on content to determine whether the prevailing emotion is happy, sad or angry. For log files, you probably want to separate the error messages, time stamps (dates) and measurements (traces) between messages.

IoT devices and sensors, like log files, are used to capture measurements and errors about a certain item. The main difference is that these devices usually work on a large number of clusters (hundreds to thousands of devices) and continuously stream data. Data generated from these devices is semi-structured or unstructured since it is in JSON or XML format. Modern technologies, such as Azure IoT services, already solve these complexities with services such as Azure IoT Hub, which aggregates all this data from various sensors and continuously exports it to a data source. Sometimes you can classify this data as semi-structured since these traces and logs are things that a system can easily understand.

Social media platforms and Microsoft Graph both provide semi-structured data. It is classified this way because just querying all of Twitter's tweets about a topic is not enough. The results don't really make a lot of sense until you do some analysis on them. The primary focus is to discern patterns and anomalies. For example, you may want to identify trends about news and topics, but also want to remove data that is irrelevant, such as tweets coming from fake accounts.

Interestingly, some **line-of-business** (**LOB**) applications provide both structured and unstructured data. For example, both Microsoft Dynamics CRM and Salesforce provide structured data that can easily be interpreted and exported to your SQL database tables, such as data for products and their amounts and value. However, they also support unstructured data such as images, videos and text notes. Note that even though text notes are considered as the string data type, they can still be considered as unstructured data because they are designed to be free text. They don't have a proper format to follow, but they are still worth exploring. A common scenario for unstructured data its use is to understand why sales were not successful.

Creating a data pipeline

Once you have identified your data sources, the next step is to create a data pipeline (sometimes also referred to as a data flow). At a high level, the steps involved are data ingestion, data storage, data preparation and training, data modelling and serving and data visualisation.

With this approach, you will build a highly scalable architecture that serves all the users of the system: from end users, data engineers and scientists who are doing the data exploration and analysts who interpret the data for the business, to even the CEO if they want to see what happening with the business in real-time:

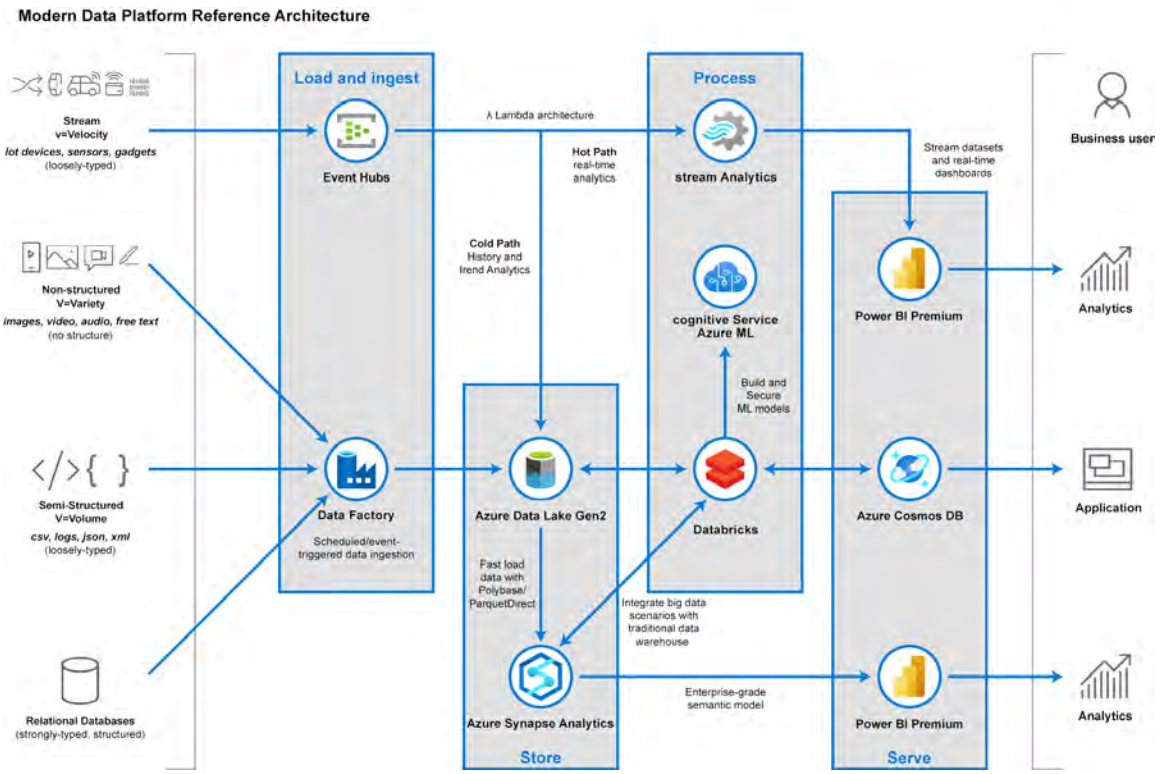


Figure 1.5: Modern data pipeline

Data ingestion

Data ingestion means transferring data (structured, semi-structured or unstructured) from the source to your storage, data lake or data warehouse.

This would involve something such as Azure Synapse Analytics using data integration to transfer data from various sources such as on-premises databases and SaaS products to a data lake. This step allows you to manage your ETL and ELT workflows without the need for manual reconciliation.

This is not a one-time process. Ideally, this is something you schedule or set to be triggered so that your data lake gets a historical snapshot from time to time. An example of this is a connection from your CRM tools, such as Microsoft Dynamics CRM, to Azure Data Lake by means of Azure Synapse Analytics with data integration. This will allow data scientists and data engineers to explore this data at different time intervals without interrupting the actual CRM application.

Data storage

Once data has been ingested from various data sources, all the data is stored in a data lake. The data residing within the lake will still be in a raw format and includes both structured and unstructured data formats. At this point, the data won't bring much value to drive business insights.

Data pipeline orchestration and monitoring

In a modern data warehouse scenario, it is very important that data sources and services efficiently transfer data from source to destination. Azure Synapse Analytics with data integration is an orchestrator that allows services to perform data migrations or transfers. It is not the thing performing the actual transfer, but rather instructs a service to perform it – for example, it can tell a Hadoop cluster to perform a Hive query.

Azure Synapse Analytics with data integration also allows you to create alerts and metrics to notify you when the service orchestration is working. You can create an alert via email for when a data transfer from source to destination was not successful.

Data sharing

In a modern data warehouse pattern, sharing data should be both seamless and secure. Often, this can be done via **File Transport Protocol** (**FTP**), emails or APIs, just to name a few. There is a big management overhead if you want to share data at scale. Azure Data Share allows you to securely manage and share your big data to other parties and organisations. The data provider will have full control of who can access the datasets and the permissions each can perform. This makes it easier for dependent companies to derive insights and explore AI scenarios.

Data preparation

Once data is ingested, the next step is data preparation. This is a phase where the data from different data sources is pre-processed for data analytics purposes. An example of this is querying data from an API and inserting them into a database table. Azure Synapse Analytics with data integration allows you to orchestrate this data preparation. Azure Synapse Analytics through a hosted Apache Spark instance can also help with data preparation, as it can run clusters concurrently to process massive amounts of data in just a matter of seconds or minutes.

Data transform, predict and enrich

Sometimes, data preparation requires further changes beyond a simple copy-and-paste scenario. This is where data transformation comes in. There are instances wherein you want to apply custom logic in the raw data first – applying filters, for instance – before you decide to transfer it to a data warehouse. Azure Synapse Analytics (through data integration), Apache Spark and SQL Analytics can also help in this scenario. If data in a data lake is not properly transformed into meaningful insights, it will eventually become a data swamp.

Furthermore, you can enrich the batch data at scale by invoking Azure Machine Learning, which makes real-time predictions about data. This can be an added feature in your data pipeline in Azure Synapse Analytics. To learn more about Azure Machine Learning, see *Chapter 3*, *Processing and visualising data*.

Data serve

After preparing and training your data, you'll be ready to model and serve it to the consumers. Basically, in this phase, you are modelling the data to be easily understood by systems. This usually involves performing the complex queries you generated from the data preparation and training phase and inserting these records into a database so that the data is structured in a defined table and schema.

All of your company's analytical data is stored in a data warehouse. You potentially have hundreds to thousands of concurrent users, reports and dashboards running off a single data warehouse.

You usually perform data modelling and service integrations with a data warehouse platform such as Azure Synapse Analytics. Completing complex queries can take hours or days. But with the power of the cloud, you can scale your Azure Synapse Analytics to perform these queries faster, making days into hours and hours into minutes.

Data visualisation

Data visualisation is an efficient way of analysing performance through graphs and charts. This is called business intelligence. Tools such as Power BI help analysts to get the most out of data. Working with Azure, you're not just limited to Power BI, but can also use other visualisation services such as Tableau. Data visualisation provides a rich and meaningful representation of your data that adds business value for you and your customers. The team can see trends, outliers and patterns that help in making data-driven decisions.

Various stakeholders within the organisation can collaborate after analysing the different performance parameters. Is your company selling products well? In what regions do you get most of your sales? With rich data backing up your assumptions, business stakeholders such as CEOs can make reasonable data-driven decisions to minimise risks. What product lines should you expand? Where should you expand further? These are some of the common questions that you can answer once you have richer data analytics.

Analysts can use desktop or web application tools to create meaningful representations of their data. Here is an example of a desktop view of Power BI where a user can analyse their company data and visualise it in graphs:

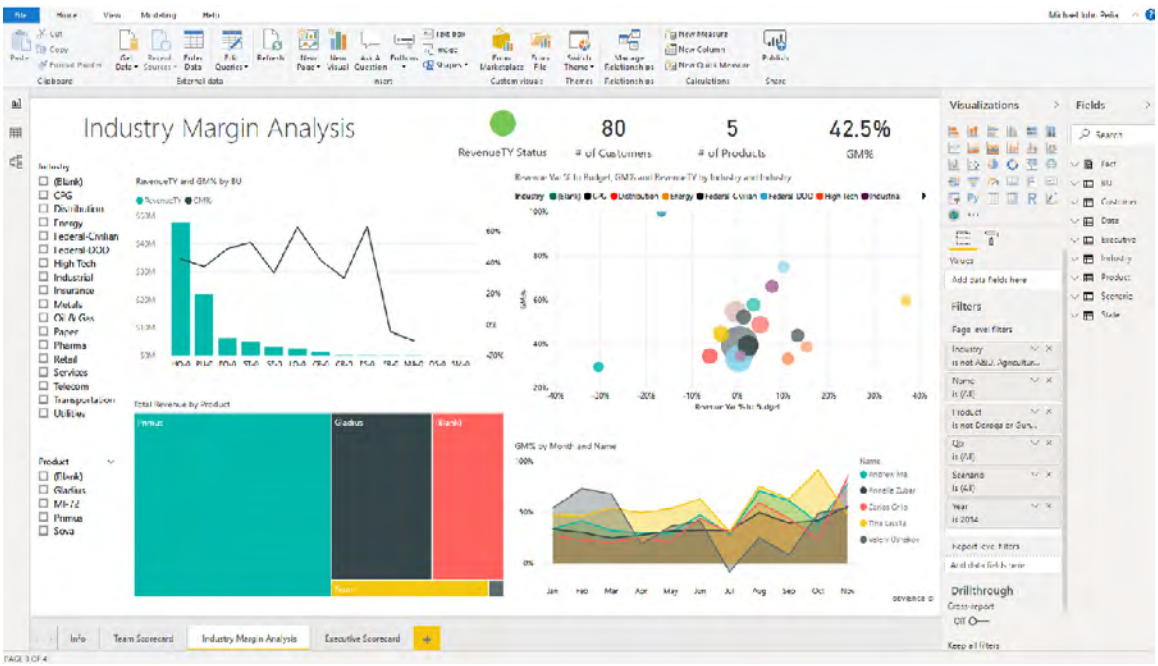


Figure 1.6: Power BI Desktop-dashboard

Once the reports are generated, they can be exported to a workspace where people can work together to improve the reports. Here is an example view of the same report in a mobile application. Users can add comments and annotations to the report, allowing a faster feedback loop for analysts:

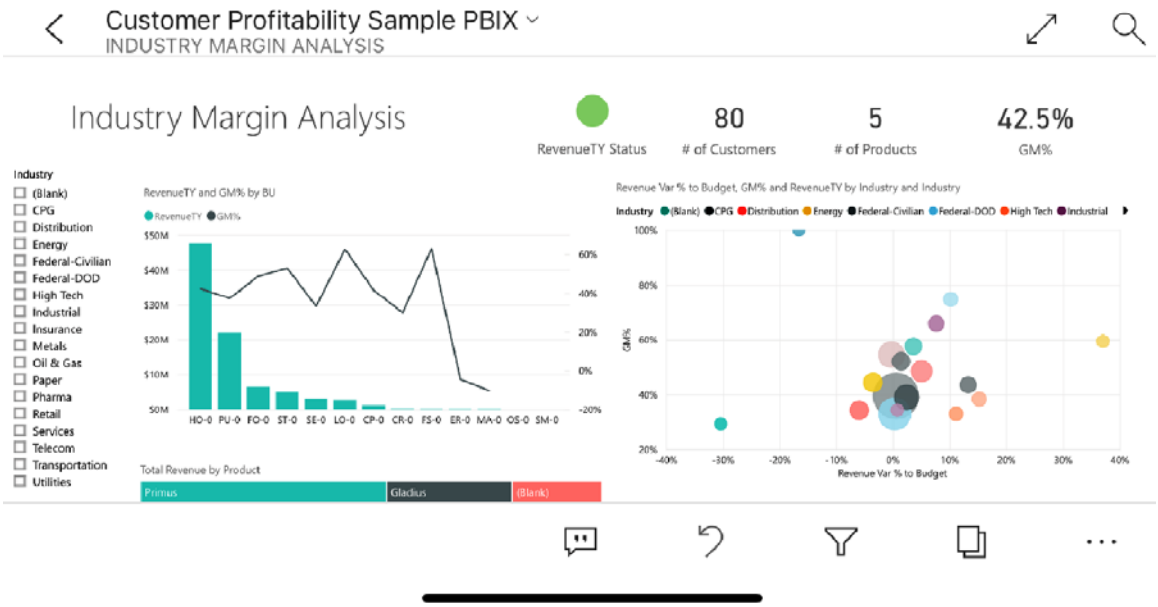


Figure 1.7: Power BI dashboard for mobile

Power BI allows you to create rich personalised dashboards that best suit your requirements and brand. Compared to having presentations with a snapshot of a graph from last week or last month, this mechanism allows you to refresh the same report over and over again.

Smarter applications

Machine learning has helped companies build applications and products such as chatbots that perform specific tasks for end users without the need for human intervention. Some common examples are voice assistants, such as Cortana, which actively learn to help us become more productive with our day-to-day tasks.

Other examples are online games in which you can easily track your performance against everyone in the world. You can see how you rank against other players, what areas you excel in and how you can improve.

The amount of tasks you can perform with rich data is virtually limitless, but in order to perform them, you need to have the right approach and infrastructure to handle a high level of scaling.

Summary

This chapter established the importance of data analytics. It also highlighted several reasons why Microsoft Azure is an ideal platform for achieving business intelligence capabilities in the cloud. It touched on some fundamental concepts around big data, ML and DataOps. You also learned about some of the business drivers for adopting data analytics on the cloud. Lastly, you gained a high-level view of what it takes to have a modern data warehouse.

In the next chapter, you will see how to start building a modern data warehouse with Azure Synapse Analytics and related technologies.

2

Introducing the Azure Synapse Analytics workspace and Synapse Studio

In the previous chapter, we introduced you to Azure and the types of platforms, tools and resources that Azure provides to facilitate the creation of data warehouse solutions.

In this chapter, we will focus on the new unified experience in the Azure Synapse Analytics workspace and Synapse Studio. We will cover the following topics:

- Azure Synapse Analytics and why we need it
- Deep dive into Azure Synapse Analytics
- Introduction to the Azure Synapse Analytics workspace and a step-by-step quick-start guide

- Introduction to Synapse Studio
- Two ways of launching Synapse Studio
- Provisioning an SQL pool, ingesting data and analysing the data in the SQL pool
- Creating an Apache Spark pool, ingesting data and exploring data using Spark
- Copying data to/from SQL pools and Spark pools
- Linked data sources
- Analysing data using serverless SQL pools
- Integrating with pipelines

What is Azure Synapse Analytics?

Azure Synapse Analytics is a limitless analytics service that brings together enterprise data warehousing and big data analytics with a unified experience to ingest, prepare, manage and serve data for immediate business intelligence and machine learning needs. In a nutshell, Azure Synapse Analytics is the next evolution of Azure SQL Data Warehouse. Microsoft has taken the industry-leading data warehouse to a new level of performance and capability.

Azure Synapse Analytics gives you the freedom to choose whether to use dedicated or serverless resources to explore and analyse your data at scale based on your business requirements. Businesses can put their data to work much more quickly, productively and securely, pulling together insights from many data sources, data warehouses and big data analytics systems.

With Azure Synapse Analytics, data professionals of all types can collaborate, manage and analyse their most important data efficiently – all within the same service. From Apache Spark integration to the powerful and trusted SQL engine, to code-free data integration and management, Azure Synapse Analytics is built for every data professional.

Furthermore, enabling BI and machine learning through Azure Synapse Analytics is a snap. Azure Synapse Analytics deeply integrates with Power BI and Azure Machine Learning to greatly expand the discovery of insights from all your data and enable practitioners to easily apply machine learning models to intelligent apps without any data movement. This significantly reduces development time for BI and machine learning projects. With Azure Synapse Analytics, you can seamlessly apply intelligence over all your most important data – from Dynamics 365 and Office 365 to **Software-as-a-Service** (**SaaS**) services that support the Open Data Initiative – and then easily share data with just a few clicks.

In the next section, we will explain why we need Azure Synapse Analytics and the business challenges that it resolves.

Why do we need Azure Synapse Analytics?

One of the many challenges that businesses face today is the need to manage two types of analytics systems:

- **Data warehouse**, which provides critical insights about the business
- **Data lakes**, which provide meaningful insights about customers, products, employees and processes through various analytics methodologies

Both of these analytics systems are critical to businesses and operate independently of one another. This can lead to uninformed decisions. At the same time, businesses need to unlock insights from all of their data to stay competitive and to innovate processes to obtain better results.

For customers wanting to build their own end-to-end data pipeline, they must go through the following steps:

1. Ingest data from various data sources.
2. Load all these data sources into a data lake for further processing.
3. Perform data cleaning over a range of different data structures and types.
4. Prepare, transform and model the data.
5. Serve the cleansed data to thousands of users through BI tools and applications.

Until now, each of these steps has required a different tool, and with so many different tools, services and applications on offer, choosing the right one can be daunting.

There are numerous services that ingest, load, prepare and serve data. There are also multiple services to use for data cleaning, based on the developer's language of choice. Some developers prefer to use Spark, and some want to use SQL, while others want code-free environments to transform the data.

Even once the right collection of tools has been chosen, there is a steep learning curve to get to grips with them, as well as the logistical difficulties of maintaining a data pipeline over different platforms and languages. With such a range of issues, implementing and maintaining a cloud analytics platform can be a difficult task.

Customer challenges

You might think that the biggest challenge for an efficient data warehouse is learning how to build the pipeline to bring the data in, or optimising the warehouse to get better performance. However, in a customer study conducted by Microsoft, it was concluded that the biggest customer challenge was managing different capabilities, monitoring hundreds of pipelines spanning across various compute engines, securing different resources (compute, storage, artifacts) and deploying code without breaking changes. Between organisational silos, data silos and tooling silos, it becomes nearly impossible to implement and maintain a cloud analytics platform.

For example, imagine your company needed to come up with a single security model to protect all of its services in order to meet the latest internal compliance guidelines. A task like this might at first sound straightforward but, in fact, it is quite involved. You need to quickly identify what that 'single security model' is, and then figure out what the deployment model is across your services. You need to work out how to implement high availability and disaster recovery for each of these services. And finally, you need to look after all of the related lifecycle management responsibilities, including monitoring these services to ensure that they are performing well. Bringing all these services together is no small endeavour, and in the past has required complex planning.

Azure Synapse Analytics to the rescue

Azure Synapse Analytics solves the aforementioned problems. As shown in *Figure 2.1*, Azure Synapse Analytics allows customers to build end-to-end analytics solutions and perform data ingestion, data exploration, data warehousing, big data analytics and machine learning tasks from a single, unified environment:

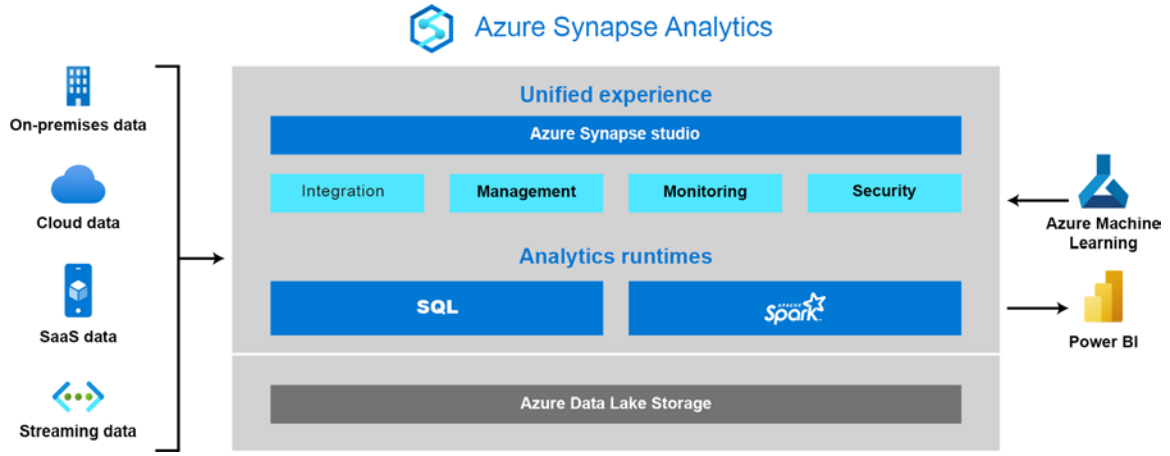


Figure 2.1: Azure Synapse Analytics

Some of the core features offered by Azure Synapse Analytics are listed in *Table 2.1*:

Features	Benefits
Serverless and dedicated options	Serverless and dedicated options support both data lake and data warehouse use cases, allowing you to choose the most cost-effective pricing option for each workload.
Enterprise data warehousing	Build your mission-critical data warehouse on the proven foundation of the industry's top-performing SQL engine.
Data lake exploration	Bring together relational and non-relational data and easily query files in the data lake with the same service you use to build data warehousing solutions.
Code-free hybrid data integration	Build ETL/ELT processes in a code-free visual environment to easily ingest data from 90+ native connectors.
Deeply integrated Apache Spark and SQL engines	Enhance collaboration among data professionals working on advanced analytics solutions. Easily use TSQL queries on both your data warehouse and Spark engine. Azure Synapse Analytics seamlessly integrates Apache Spark for data preparation, data engineering, ETL and machine learning
Cloud-native Hybrid Transactional and Analytical Processing (HTAP)	Get insights from real-time transactional data stored in operational databases, such as Azure Cosmos DB, with a single click.
Choice of language	Use your preferred language, including T-SQL, Python, Scala, Spark SQL and C#.
Integrated AI and BI	Complete your end-to-end analytics solution with deep integration of Azure Machine Learning, Azure Cognitive Services and Power BI.

Table 2.1: Features and benefits of Azure Synapse Analytics

Azure Synapse Analytics can derive and deliver insights from all the data lying in your data warehouse and big data analytics systems at lightning-fast speeds. It enables data professionals to use familiar SQL language to query both relational and non-relational databases at petabyte scale. Advanced features such as intelligent workload management, workload isolation and limitless concurrency help optimise the performance of all queries for mission-critical workloads.

Azure Synapse Analytics takes the best of Azure SQL Data Warehouse and modernises it by providing more functionalities for the SQL developers, adding querying with serverless SQL pool, adding machine learning support, embedding Apache Spark natively, providing collaborative notebooks and offering data integration within a single service. It supports different languages (such as C#, SQL, Scala and Python), all through different engines.

By using Azure Synapse Analytics, customers can carry out business intelligence projects and machine learning with ease. Azure Synapse Analytics is deeply integrated with Power BI and Azure Machine Learning to greatly expand the discovery of insights from all your data and apply machine learning models to all your intelligent apps. The user can significantly reduce project development time for BI and machine learning projects with a limitless analytics service that enables you to seamlessly apply intelligence over all your most important data – from Dynamics 365 and Office 365 to SaaS implementations that support Open Data Initiative – and easily share data with just a few clicks.

This is all provided in a single experience that features query editors and notebooks for sharing and collaborating data, as well as assets and code for both SQL and Spark developers.

Essentially, Azure Synapse Analytics does it all.

Deep dive into Azure Synapse Analytics

Now that you understand why Azure Synapse Analytics was invented, we will take a deeper look at the services offered by Azure Synapse Analytics.

Azure Synapse Analytics is a fully managed, integrated data analytics service that blends data warehousing, data integration and big data processing with accelerating time to insight into a single service.

The advantage of having a single integrated data service is that, for enterprises, it accelerates the delivery of BI, AI, machine learning, Internet of Things and intelligent applications.

Figure 2.2 illustrates how a modern data pipeline can be built using Azure Synapse Analytics. In this example, the ingestion process starts from a blob storage source through to Azure Data Lake Storage Gen2 in the Azure Synapse Analytics workspace. Using a Spark pool, you can read from multiple data sources via Azure Data Lake Storage Gen2 and Azure SQL Database and perform any transformations and data cleansing needed. Finally, the curated results are written in the SQL pool, which can then be used to serve BI tools and applications:

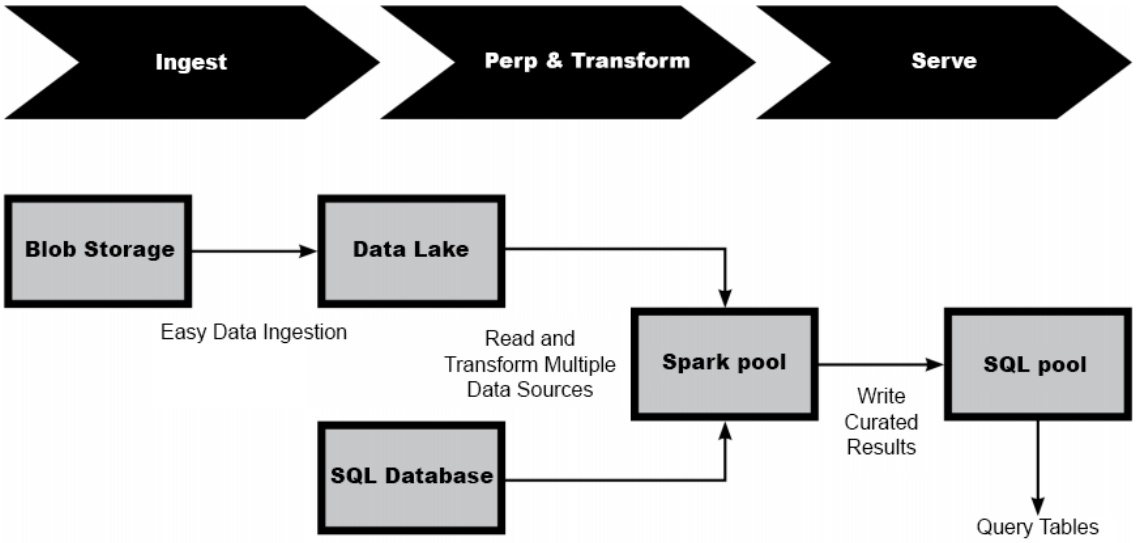


Figure 2.2: Modern data pipeline using Azure Synapse Analytics

To fully appreciate the benefits of Azure Synapse Analytics, we will take you on a tour of the Azure Synapse Analytics workspace and introduce you to the core services that are inside Synapse Studio.

Introducing the Azure Synapse Analytics workspace

At the heart of Azure Synapse Analytics is its workspace. A workspace is the top-level resource and comprises your analytics solution in the data warehouse. The Azure Synapse Analytics workspace can be used in a collaborative environment and supports both relational and big data processing. In essence, the Azure Synapse Analytics workspace is the fuel that jump-starts your entire Azure Synapse Analytics experience.

In the next section, we have provided you with a quick-start guide on how to provision your first Azure Synapse Analytics workspace. Feel free to follow along.

Free Azure account

If you'd like to try out any of the techniques shown in this book, simply create your free Azure account and get started.

Quick-start guide

1. In a web browser, sign in to the Azure portal.
2. In the search box, type in **synapse**. Then, from the search results, select **Azure Synapse Analytics** under **Services:**

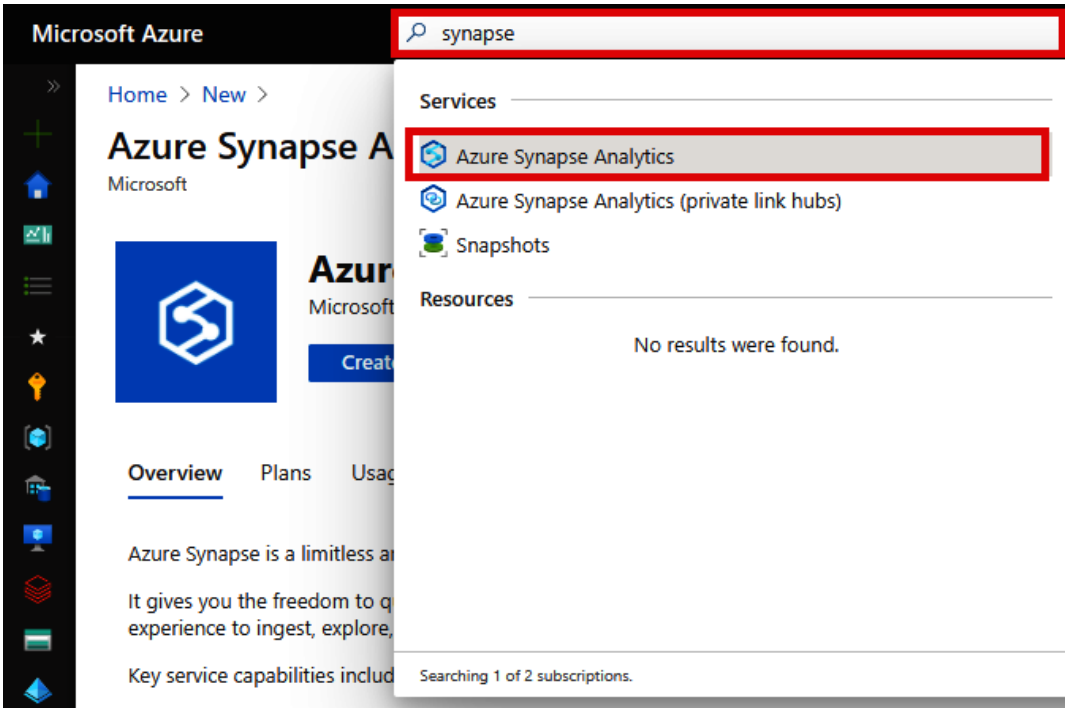


Figure 2.3: Navigating to Azure Synapse Analytics through the Azure portal

Click **Add** to create a new Azure Synapse Analytics workspace as shown in *Figure 2.4*:

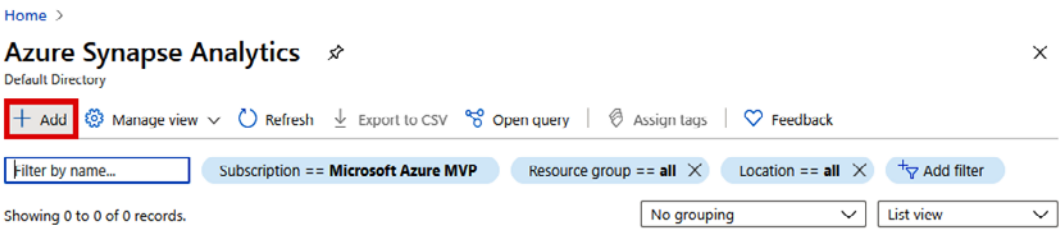
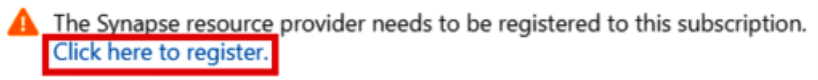


Figure 2.4: Creating a new Azure Synapse Analytics workspace

3. In the **Basics** tab, create a new resource group and provide a unique name for the workspace as shown in *Figure* 2.5. In our example, we will name our resource group `my-synapse-rg` and our workspace `mysynws001`. We have also picked `East US` as our region:

> **Note**
>
> If you see a message stating `The Synapse resource provider needs to be registered to this subscription`, under **Subscription** on the **Create Synapse workspace** page, simply click the link to register as instructed:
>
> 
>

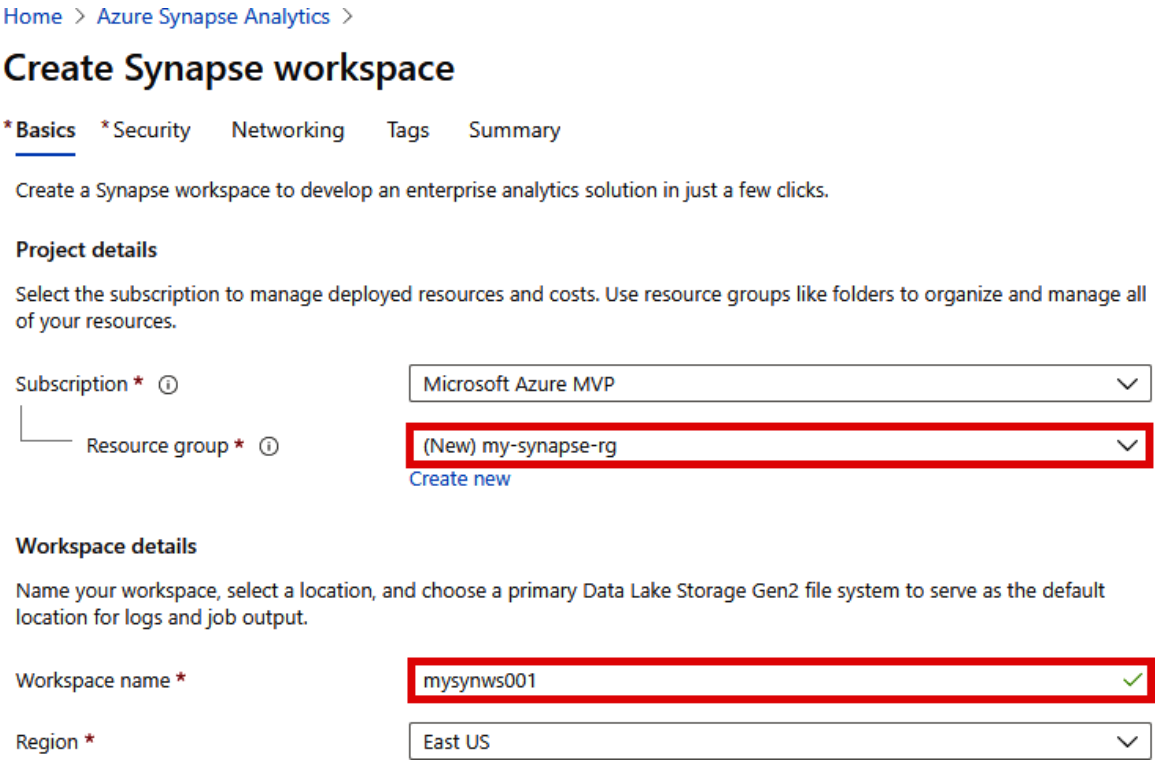


Figure 2.5: Creating a Synapse workspace

4. Next, you need an Azure **Data Lake Storage Gen2** (**mydatalakestorageg2**) **account** to create an Azure Synapse Analytics workspace. In this quick-start guide, we will create a new **mydatalakestorageg2** account as shown in *Figure* 2.6:

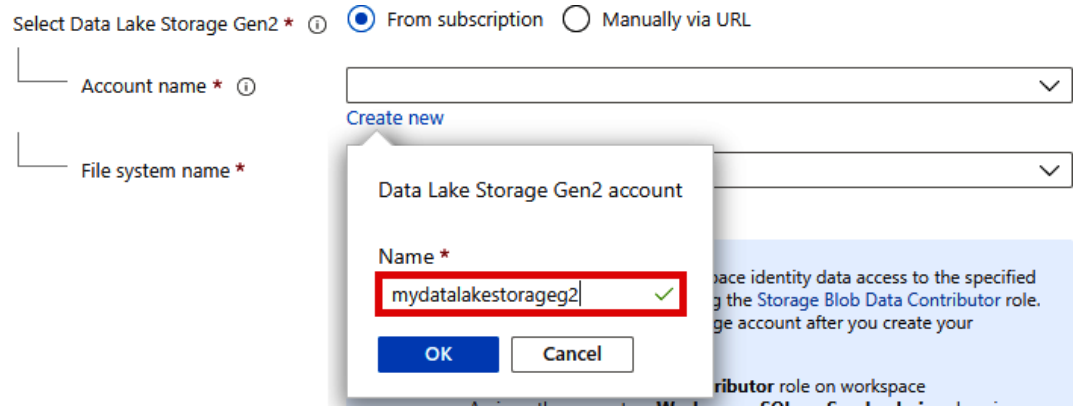


Figure 2.6: Creating a new Azure Data Lake Storage Gen2 account

5. For **File system name**, click **Create new** and give it a name as shown in *Figure* 2.7. In our example, we will name it **`users`**:

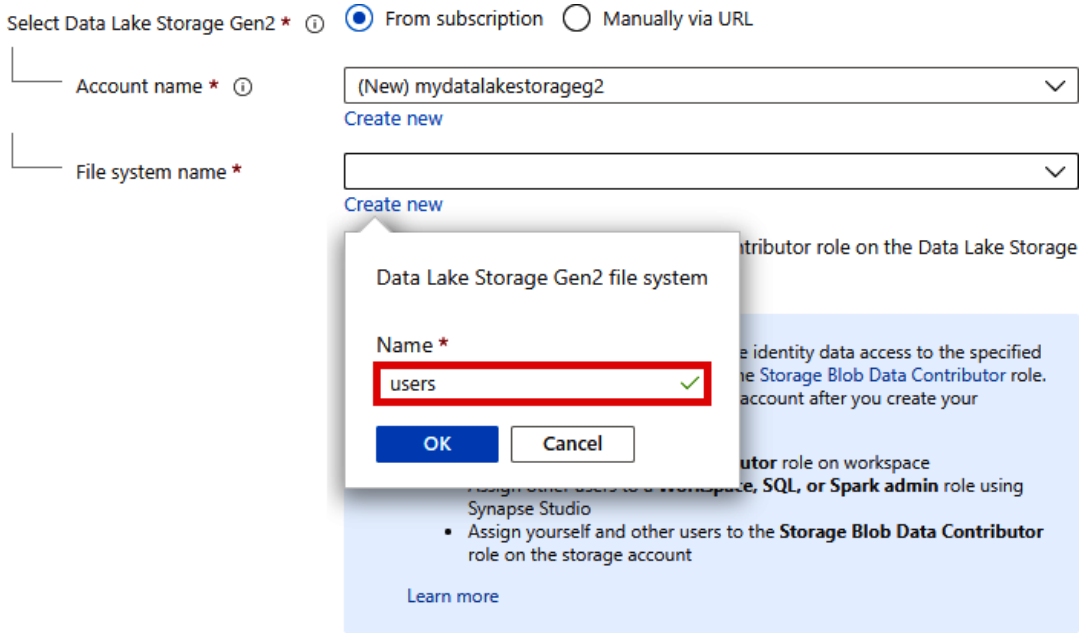


Figure 2.7: Providing a file system name

6. The completed form should resemble *Figure 2.8*. Be sure to check the **Assign myself the Storage Blob Contributor role on the Data Lake Storage Gen2 account** tick box. Click **Next: Security >** to continue:

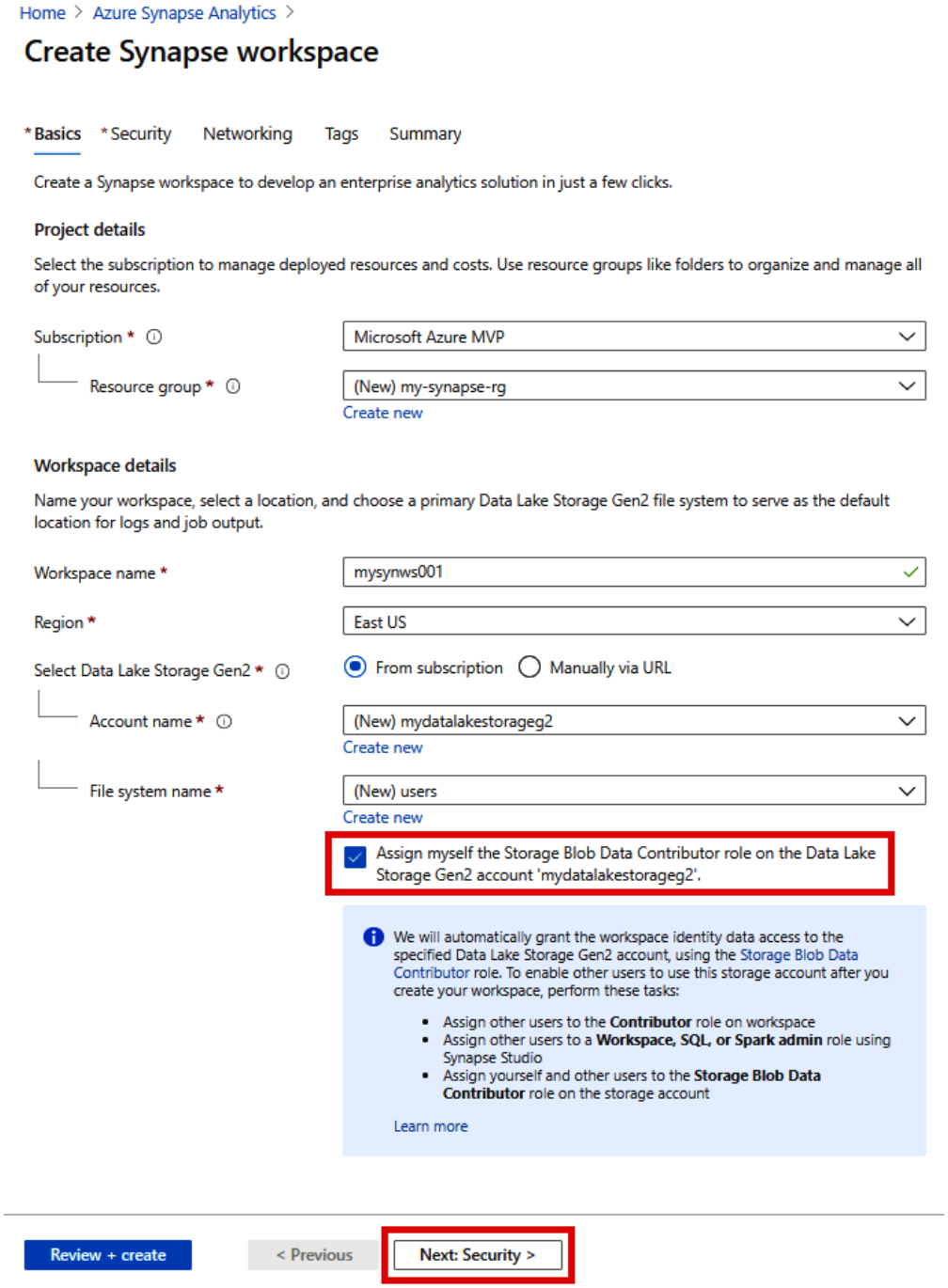


Figure 2.8: Create Synapse workspace – Basics tab

7. In the **Security** tab, you may optionally provide credentials that can be used for administrator access to the workspace's SQL pools. If you don't provide a password, one will be automatically generated. You can change the password later. For now, let's accept the default and click **Next: Networking >**:

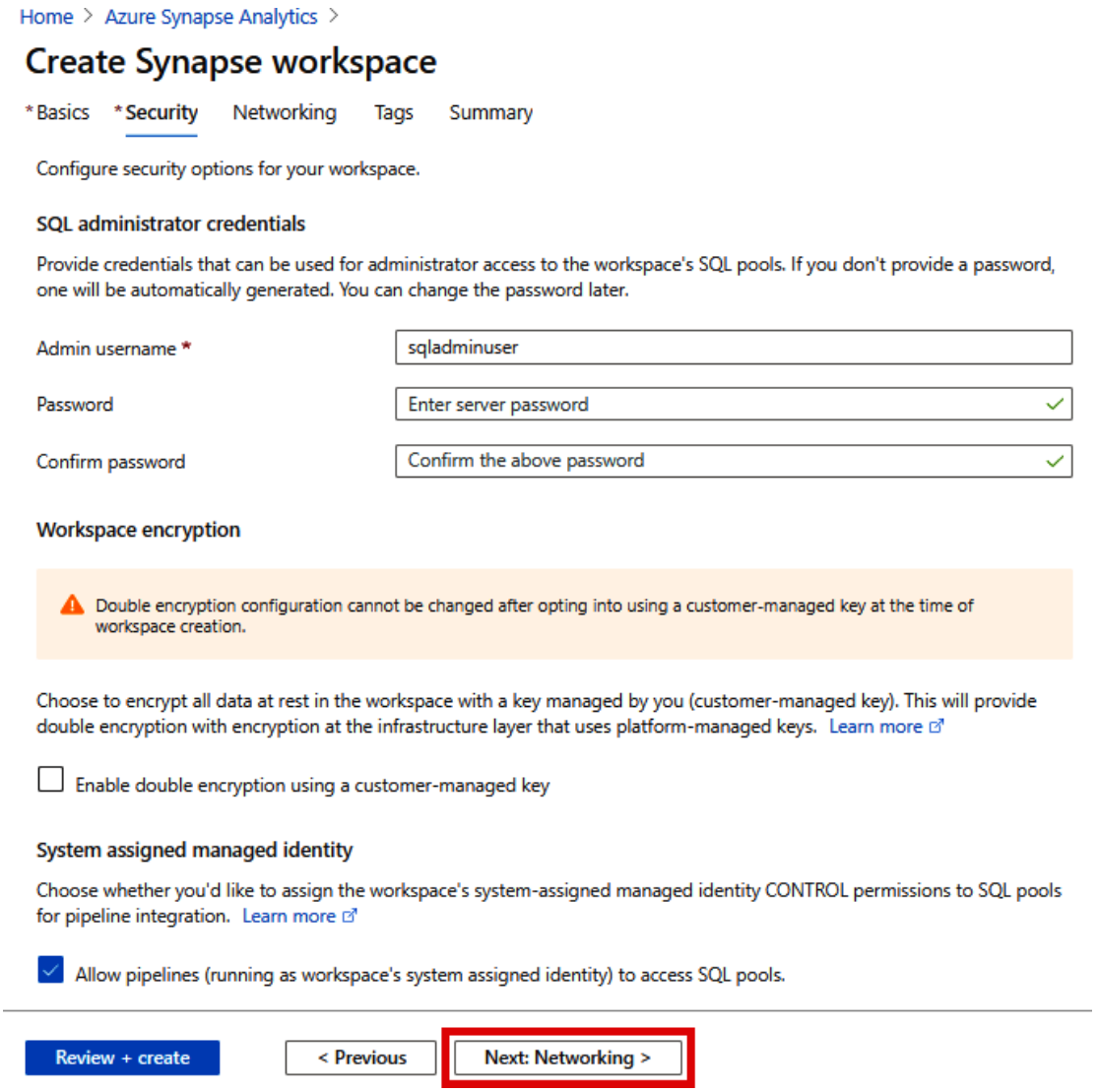


Figure 2.9: Create Synapse workspace – Security tab

8. In the **Networking** tab, accept the default and click **Review + create**:

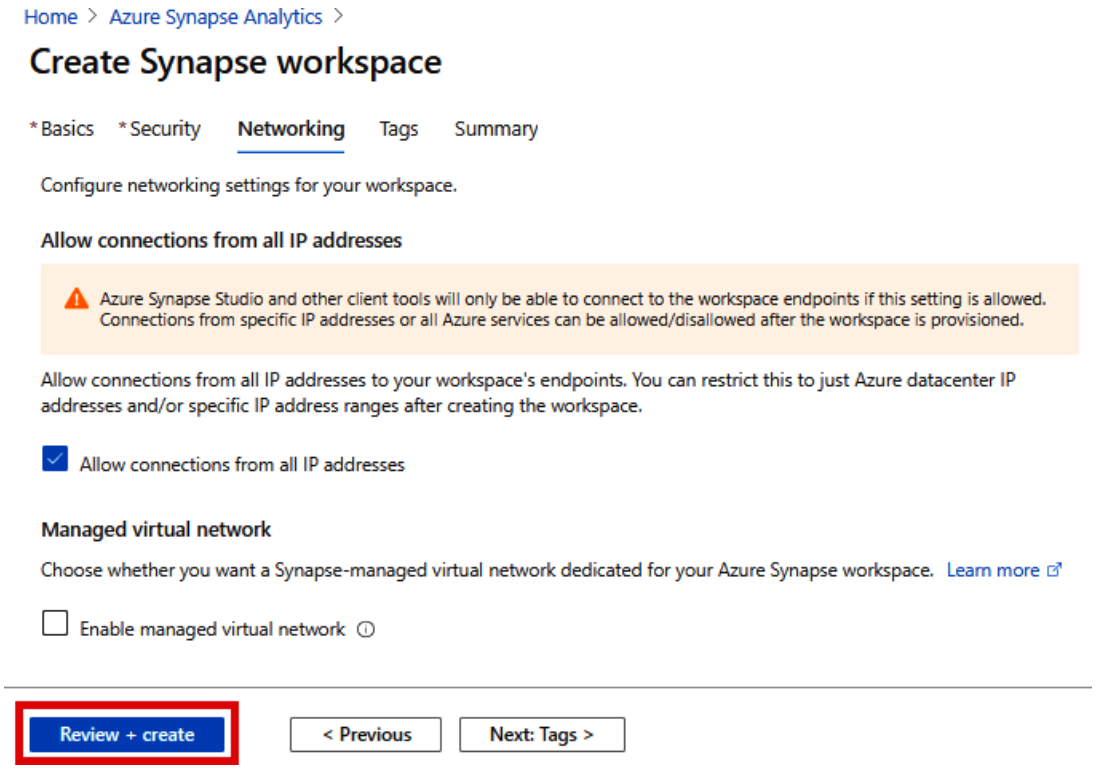


Figure 2.10: Create Synapse workspace – Networking tab

9. In the **Summary** tab, do a final review of your configurations then click **Create**:

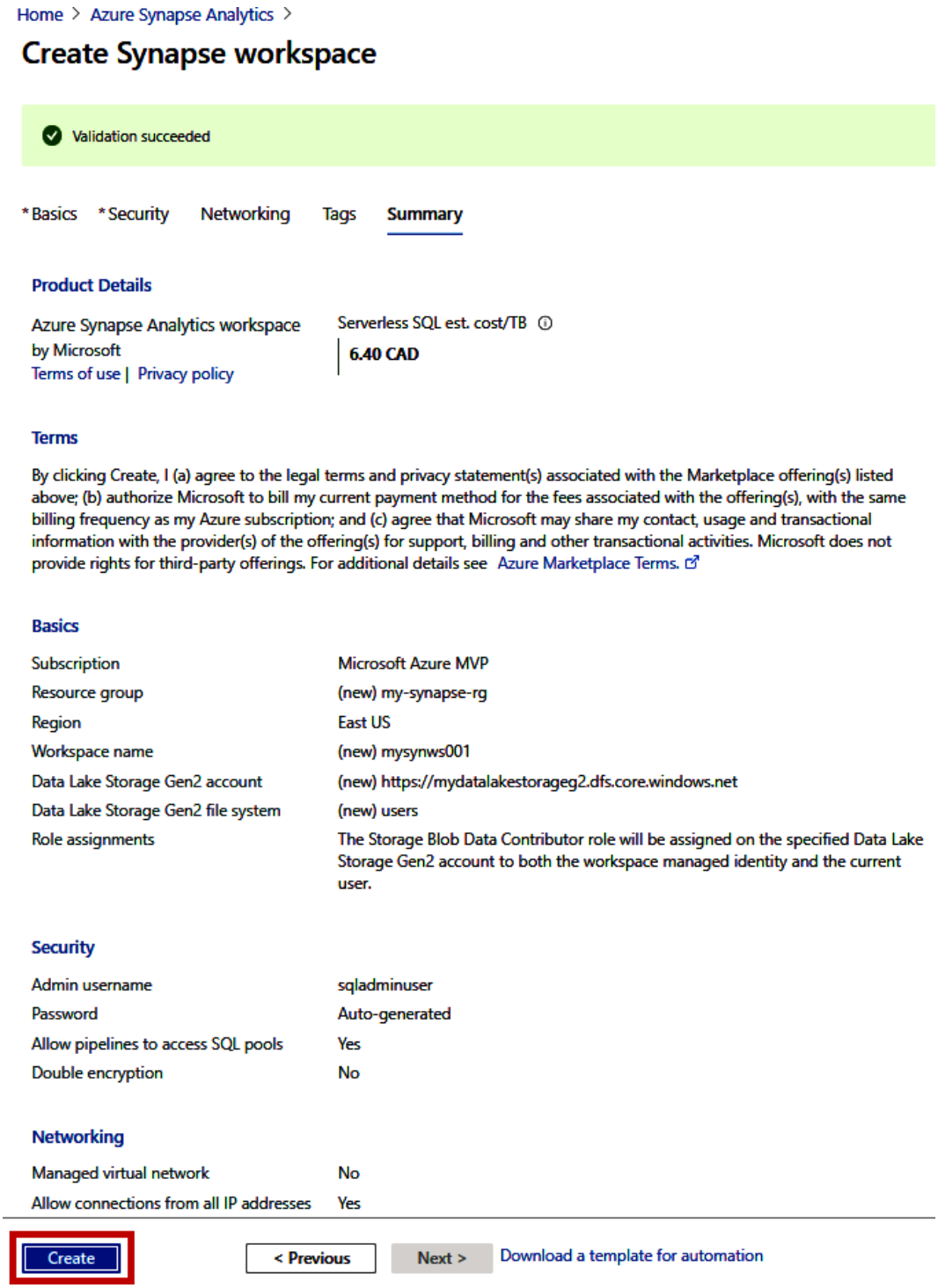


Figure 2.11: The Summary tab

In just a matter of minutes, your new Azure Synapse Analytics workspace will be ready.

In this section, we have shown you how you can get started with Azure Synapse Analytics by creating your Azure Synapse Analytics workspace. Next, we will continue our tour by visiting Synapse Studio.

Introducing Synapse Studio

Synapse Studio features a user-friendly, web-based interface that provides an integrated workspace and development experience. This allows data engineers to build end-to-end analytics solutions (ingest, explore, prepare, orchestrate, visualise) by performing everything they need within a single environment. Furthermore, data engineers can write and debug code in SQL or Spark. Synapse Studio also integrates with enterprise CI/CD processes. Synapse Studio is an ideal environment for data engineers and data scientists to share and collaborate on their analytics solutions.

To continue our tour, we will use the Azure Synapse Analytics workspace that we created in the previous section to launch Synapse Studio.

Launching Synapse Studio

With the Azure Synapse Analytics workspace you created in the previous section, we are now ready to make full use of it in Synapse Studio. Launching Synapse Studio can be done in two ways:

Method 1: Launching Synapse Studio via the Azure portal

1. In the Azure portal, go to the resource group that contains your Azure Synapse Analytics workspace. In our example, our resource group is **`my-synapse-rg`** and our Azure Synapse Analytics workspace is **`mysynws001`**.
2. Click the Azure Synapse Analytics workspace as shown:

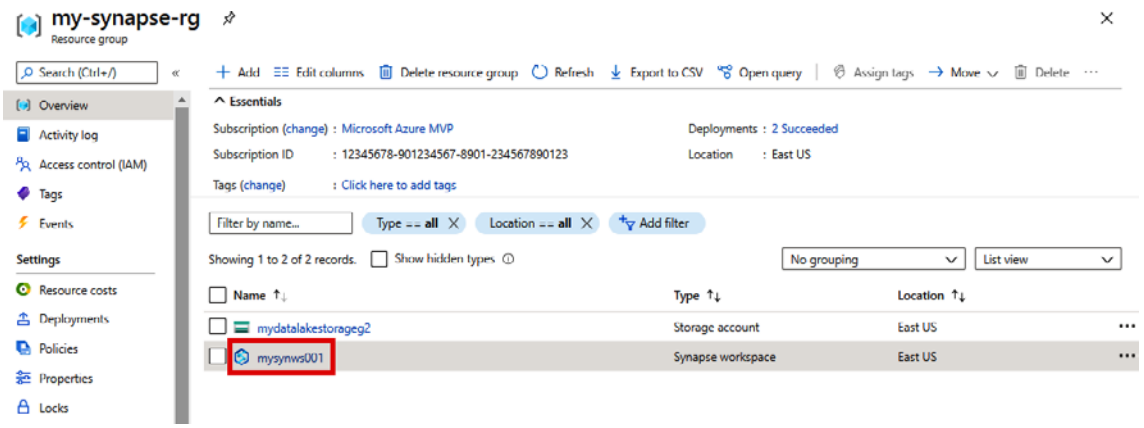


Figure 2.12: The my-synapse-rg resource group

3. From your Azure Synapse Analytics workspace, click **Open** to launch Synapse Studio:

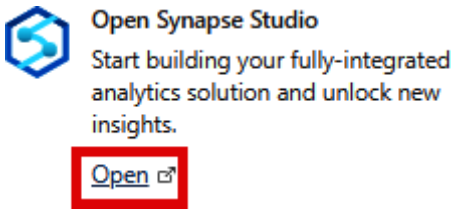


Figure 2.13: Launching the Synapse Studio

At this point, Synapse Studio will be launched (see *Figure* 2.13).

Method 2: Launching Synapse Studio via its URL

You can also launch Synapse Studio via https://web.azuresynapse.net/, then sign into your Azure Synapse Analytics workspace as shown in *Figure* 2.14:

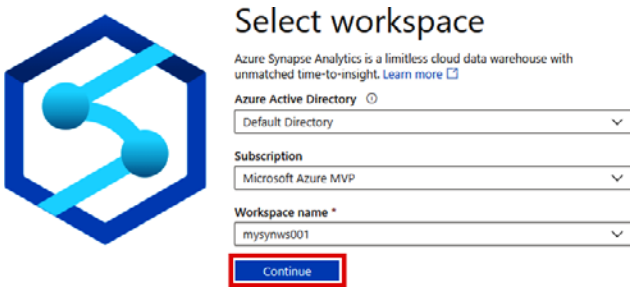


Figure 2.14: Selecting your Azure Synapse Analytics workspace to continue

The Synapse Studio home page is shown in *Figure* 2.15:

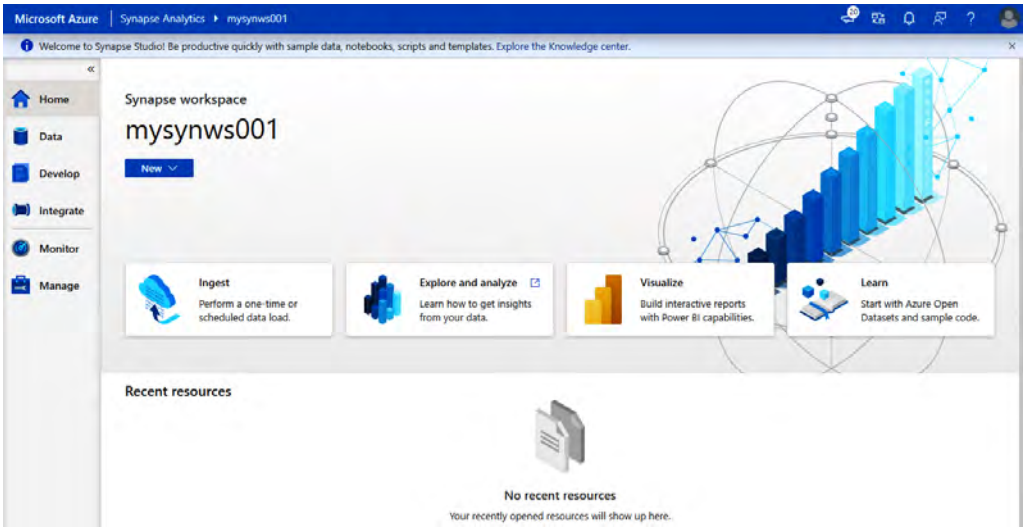


Figure 2.15: Synapse Studio home page

In the next sections, we will show you how easy it is to provision a dedicated SQL pool, ingest data and start exploring the data in the SQL pool.

Provisioning a dedicated SQL pool

In this section, we will show you how you can provision a dedicated SQL pool in our newly created Azure Synapse Analytics workspace through Synapse Studio:

1. In Synapse Studio, select the **Manage** hub on the left-hand pane, then **SQL pools** and then click **+ New**:

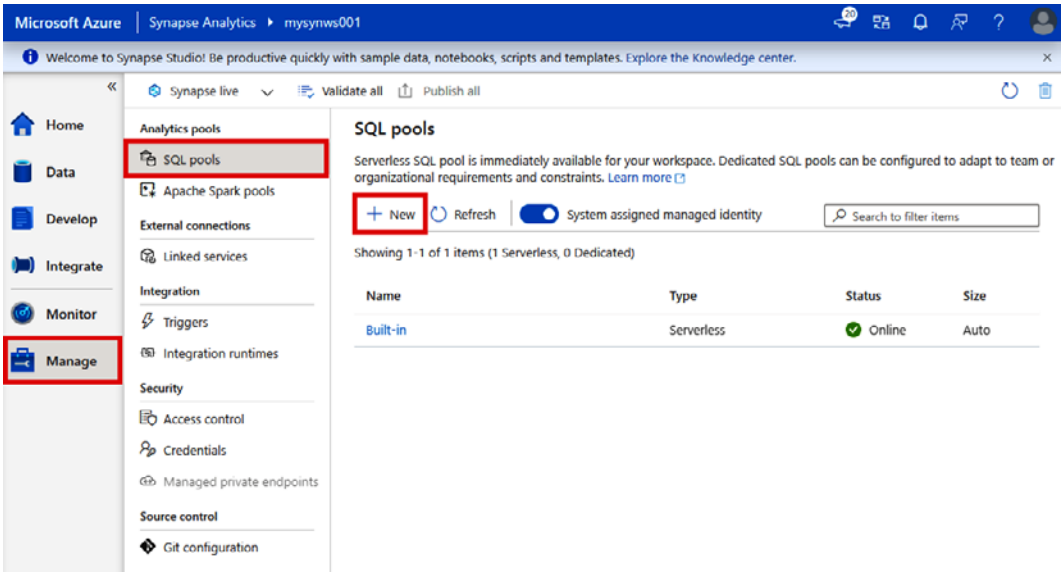


Figure 2.16: Creating a new SQL pool

2. Enter the SQL pool details as shown. Then, click **Review + create**:

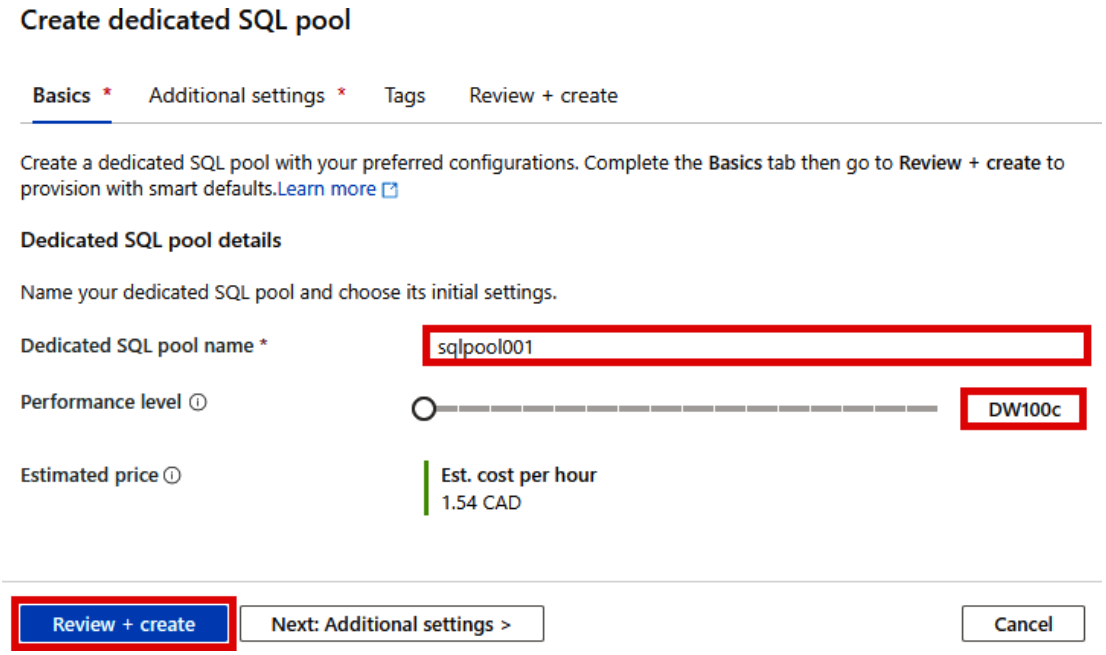


Figure 2.17: Create dedicated SQL pool – Basics tab

3. In the **Review + create** tab, click **Create**.

In just a matter of minutes, your new dedicated SQL pool will be ready for use. In our example, our dedicated SQL pool is associated with an SQL pool database that is also named **`sqlpool001`**.

> **Tip**
>
> An SQL pool, as long as it remains active, will consume billable resources. To minimise costs, you can pause the pool when you are not using it.

Next, we will ingest the NYC Taxi data into our dedicated SQL pool and explore its capabilities.

4. In Synapse Studio, select the **Develop** hub on the left-hand pane, then **+** and **SQL script**:

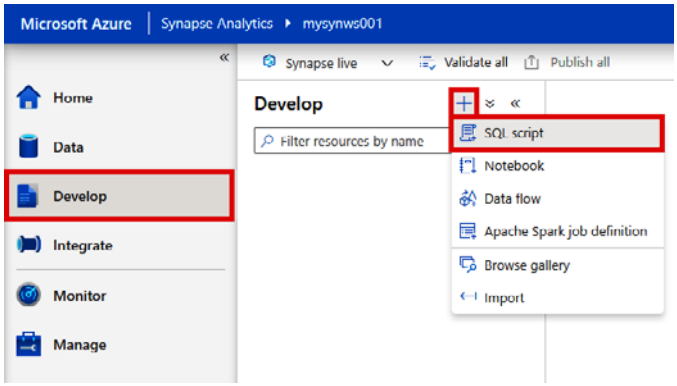


Figure 2.18: Developing a new SQL script

5. Connect to **`sqlpool001`** as shown:

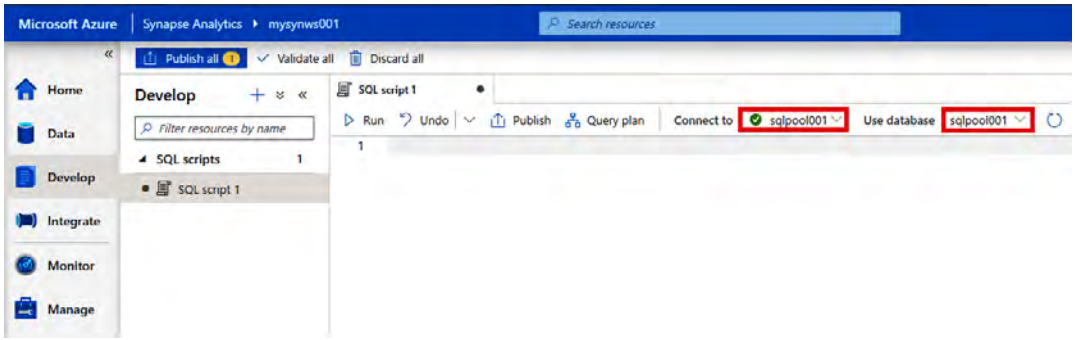


Figure 2.19: Connecting to sqlpool001

6. In the SQL script pane, enter the full SQL script (as shown in *Figure 2.20*), then click **Run**:

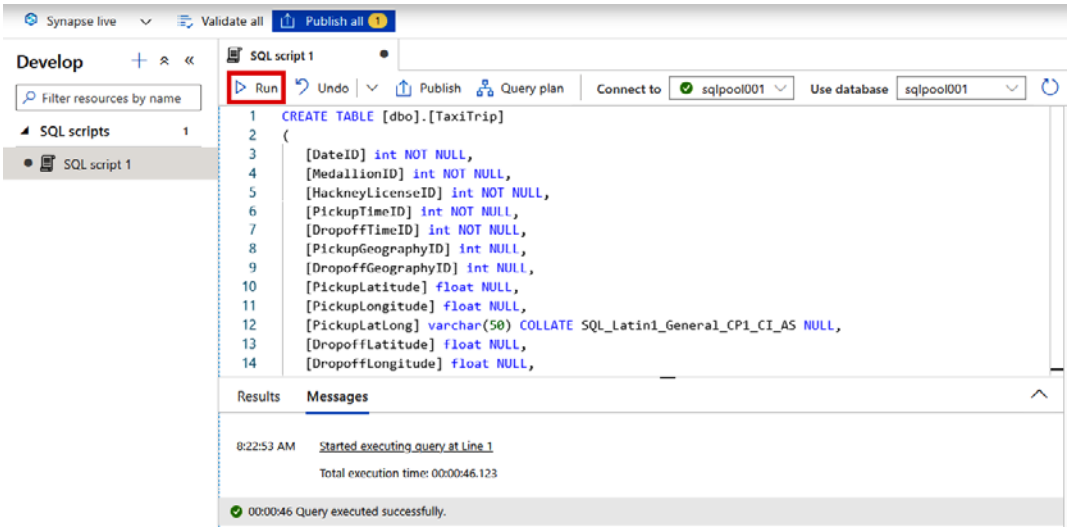


Figure 2.20: Executing the SQL script

The following SQL script creates a table named **dbo.TaxiTrip** in our dedicated SQL pool and ingests more than 2.8 million rows of the NYC Taxi data into the **dbo.TaxiTrip** table:

```
CREATE TABLE [dbo].[TaxiTrip]
(
    [DateID] int NOT NULL,
    [MedallionID] int NOT NULL,
    [HackneyLicenseID] int NOT NULL,
    [PickupTimeID] int NOT NULL,
    [DropoffTimeID] int NOT NULL,
    [PickupGeographyID] int NULL,
    [DropoffGeographyID] int NULL,
    [PickupLatitude] float NULL,
    [PickupLongitude] float NULL,
    [PickupLatLong] varchar(50) COLLATE SQL_Latin1_General_CP1_CI_AS NULL,
    [DropoffLatitude] float NULL,
    [DropoffLongitude] float NULL,
    [DropoffLatLong] varchar(50) COLLATE SQL_Latin1_General_CP1_CI_AS NULL,
    [PassengerCount] int NULL,
```

```
    [TripDurationSeconds] int NULL,
    [TripDistanceMiles] float NULL,
    [PaymentType] varchar(50) COLLATE SQL_Latin1_General_CP1_CI_AS NULL,
    [FareAmount] money NULL,
    [SurchargeAmount] money NULL,
    [TaxAmount] money NULL,
    [TipAmount] money NULL,
    [TollsAmount] money NULL,
    [TotalAmount] money NULL
)
WITH
(
    DISTRIBUTION = ROUND_ROBIN,
    CLUSTERED COLUMNSTORE INDEX
);

COPY INTO [dbo].[TaxiTrip]
FROM 'https://nytaxiblob.blob.core.windows.net/2013/Trip2013/
QID6392_20171107_05910_0.txt.gz'
WITH
(
    FILE_TYPE = 'CSV',
    FIELDTERMINATOR = '|',
    FIELDQUOTE = '',
    ROWTERMINATOR='0X0A',
    COMPRESSION = 'GZIP'
)
OPTION (LABEL = 'COPY: Load taxi dataset');
```

In this SQL script, we use the flexible **`COPY`** statement to bulk load data from an Azure Blob Storage source location into our dedicated SQL pool.

Now that we have our data ingested into our dedicated SQL pool, let's explore the data.

Exploring data in the dedicated SQL pool

Follow the steps below for exploring the data in the dedicated SQL pool:

1. In Synapse Studio, select the **Data** hub on the left-hand pane.
2. Select **`sqlpool001 (SQL)`**, then expand **Tables**.
3. Right-click on the **`dbo.TaxiTrip`** table.
4. Select **New SQL script | Select TOP 100 rows**:

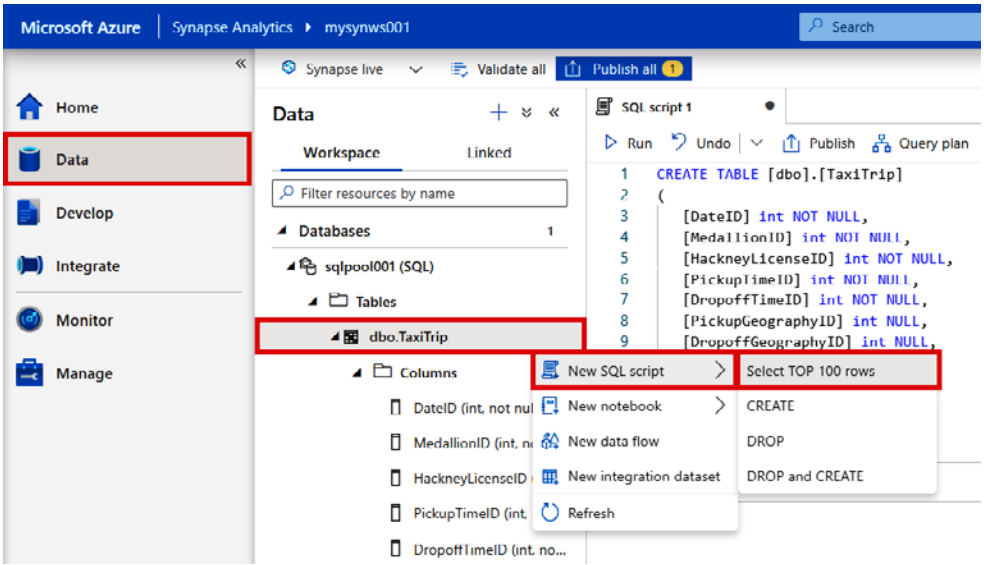


Figure 2.21: Generating a new SQL script

5. You will see the output shown in *Figure 2.22*:

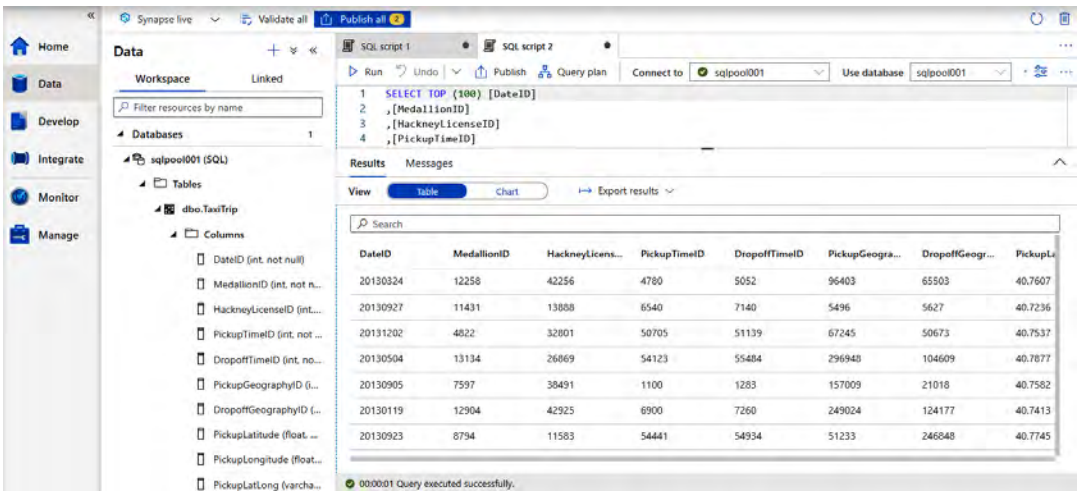


Figure 2.22: Query result

6. Try replacing the query with the following and run it. This query shows how the number of passengers relates to the total trip distance and average trip distance:

```
SELECT PassengerCount,
   SUM(TripDistanceMiles) as TotalTripDistance,
   AVG(TripDistanceMiles) as AverageTripDistance
FROM dbo.TaxiTrip
WHERE PassengerCount > 0 AND TripDistanceMiles > 0
GROUP BY PassengerCount
ORDER BY PassengerCount
```

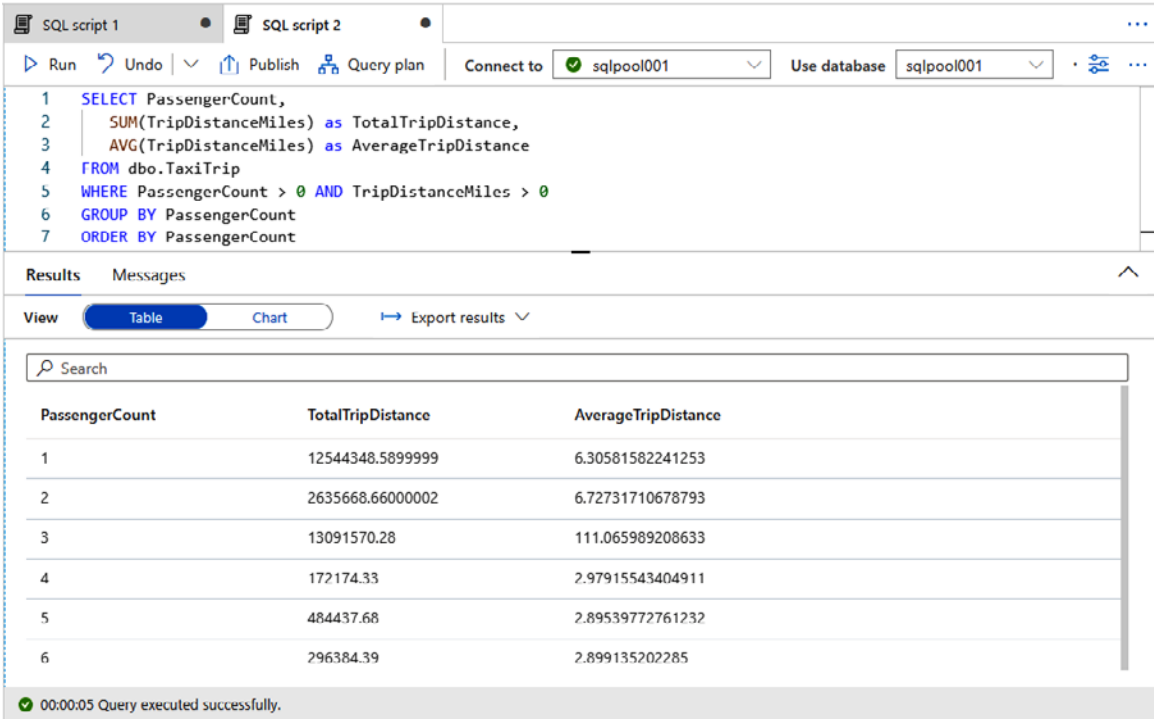


Figure 2.23: Query result

7. You can quickly change the view to **Chart** to see a visualisation of the results as a line chart:

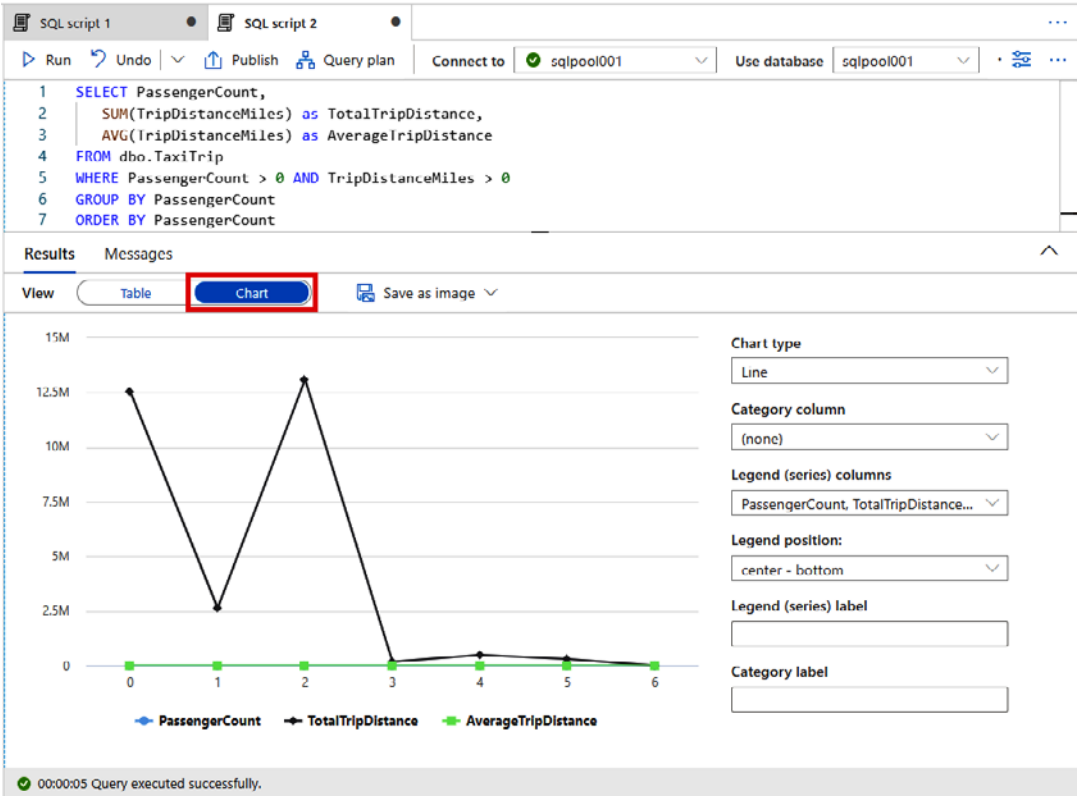


Figure 2.24: Line chart visualisation

Now that we have shown you how easy it is to provision a dedicated SQL pool, ingest data and explore the data, we will turn our attention to another key capability of Azure Synapse Analytics: the Apache Spark pool.

In the next sections, we will show you how to create an Apache Spark pool in Azure Synapse Analytics, load the NYC Taxi data into the Spark database and analyse the NYC Taxi data using Spark and notebooks.

Creating an Apache Spark pool

Next, we will create a serverless Apache Spark pool:

1. In Synapse Studio, select the **Manage** hub on the left-hand pane, then click **Apache Spark pools** under **Analytics pools** and click **+ New**:

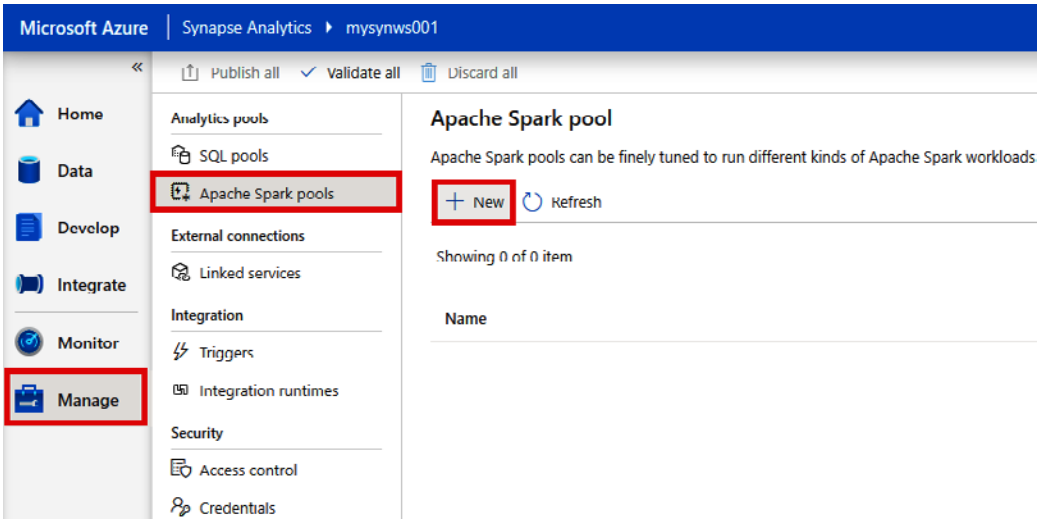


Figure 2.25: Creating a new Apache Spark pool

2. Enter the **Apache Spark pool** details as shown. Then, click **Review + create**:

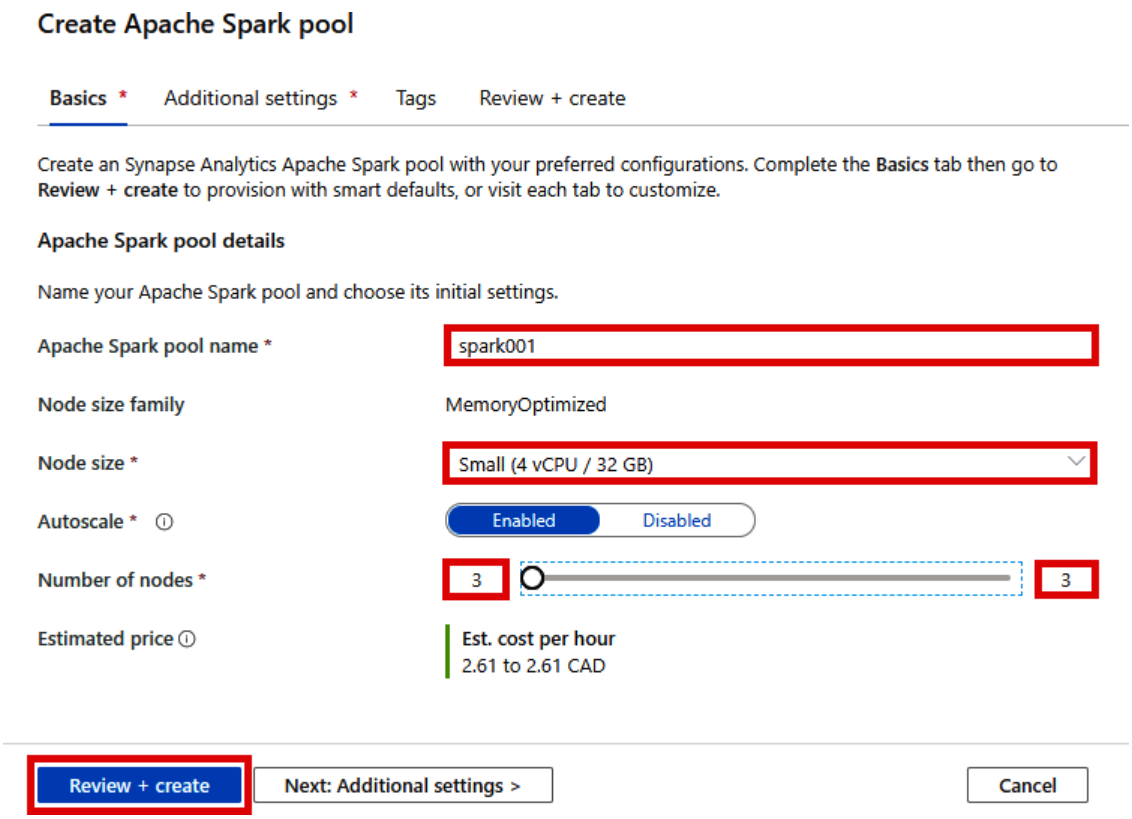


Figure 2.26: Create Apache Spark pool – Basics tab

3. Under the **Review + create** tab, click on **Create**.

In just a matter of minutes, your new Apache Spark pool will be ready. We will look at linked data sources next.

Linked data sources

1. In Synapse Studio, select the **Data** hub on the left-hand pane, then click **+** and **Browse gallery**:

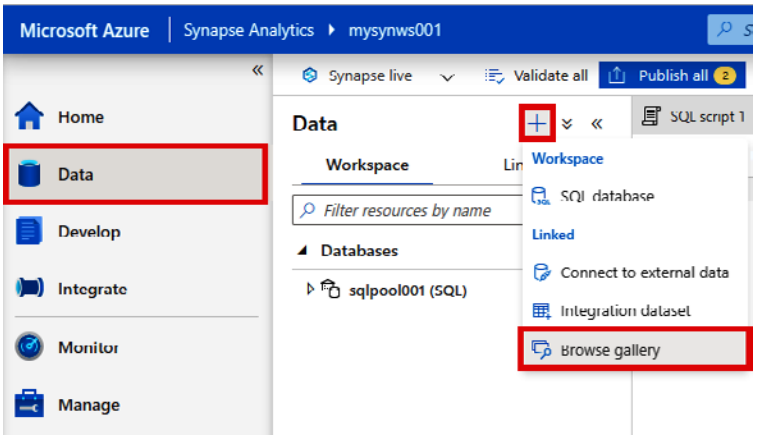


Figure 2.27: The 'Browse gallery' option

2. Select **NYC Taxi & Limousine Commission – yellow taxi trip records** and click on **Continue**:

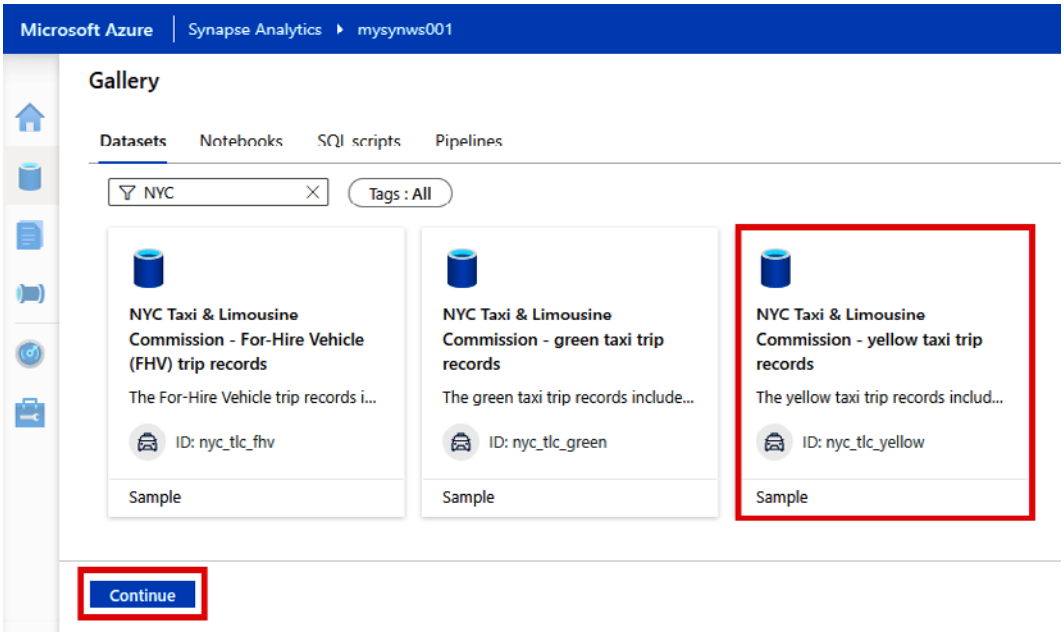


Figure 2.28: Selecting a dataset from the gallery

3. Click on **Add dataset**:

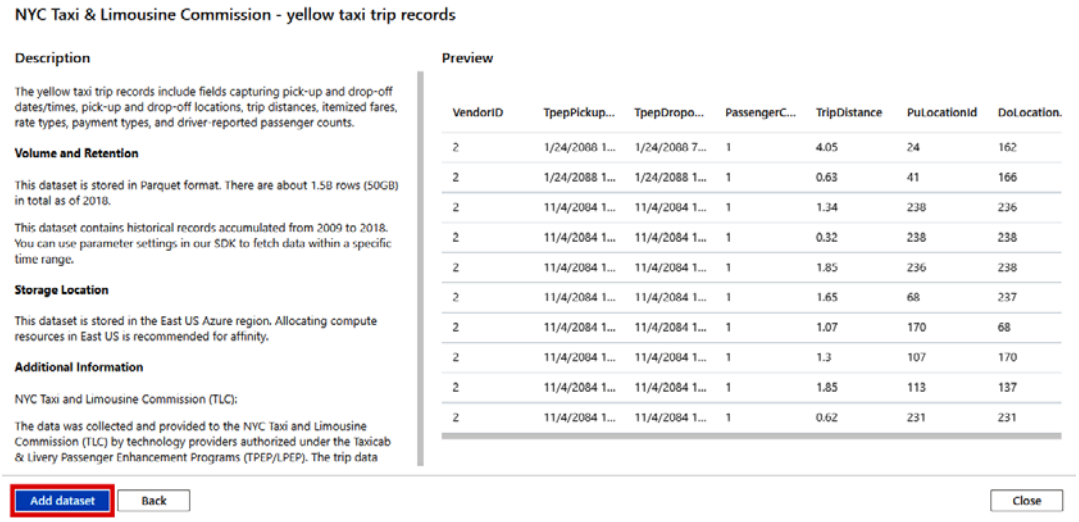


Figure 2.29: Adding a dataset

4. In the **Data** hub, under **Linked**, right-click on **`nyc_tlc_yellow`** and select **New notebook** then **Load to DataFrame**:

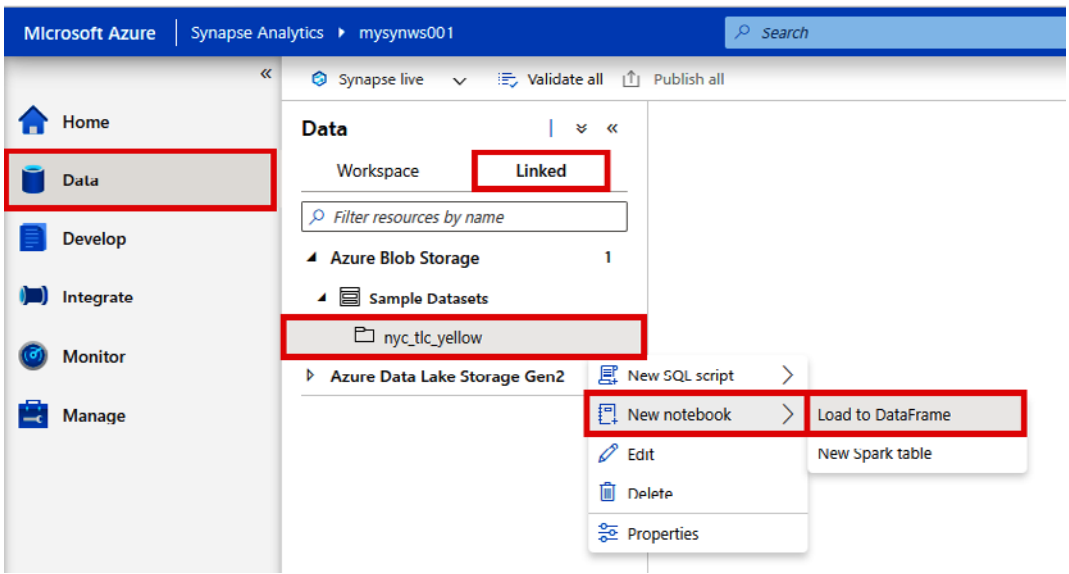


Figure 2.30: Creating a new notebook

5. A new notebook is created with the following code auto-generated. Click **Run all**:

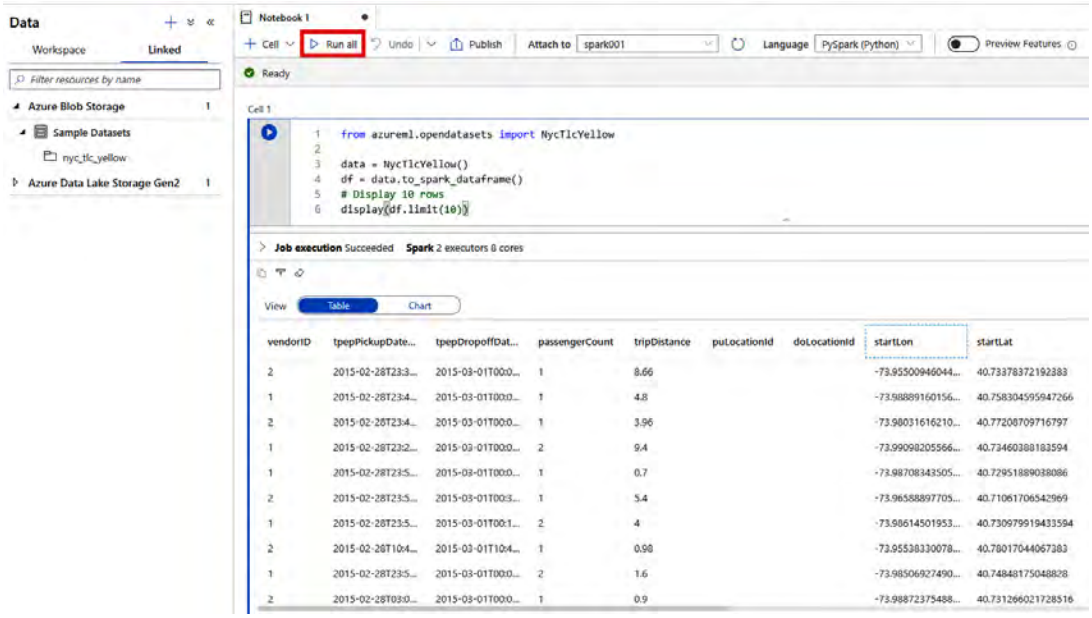


Figure 2.31: Running the notebook

Feel free to experiment by linking other datasets and modifying the code to pull in different results. In the next section, you will learn how to ingest SQL pool data into a Spark database.

Ingesting SQL pool data into a Spark database

Earlier in this chapter, we loaded the NYC Taxi data into our SQL pool named **`sqlpool001`**. Let's use it to demonstrate how we can ingest data from our SQL pool into our Spark database named **`sparknyc`**:

1. In Synapse Studio, select the **Develop** hub on the left-hand pane, then **+** and **Notebook**:

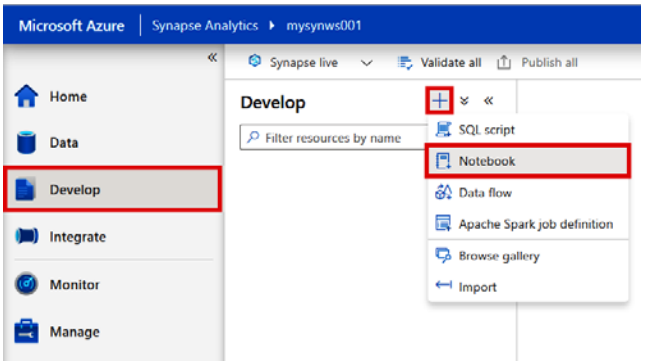


Figure 2.32: Creating a new notebook

2. Click **Add code**:

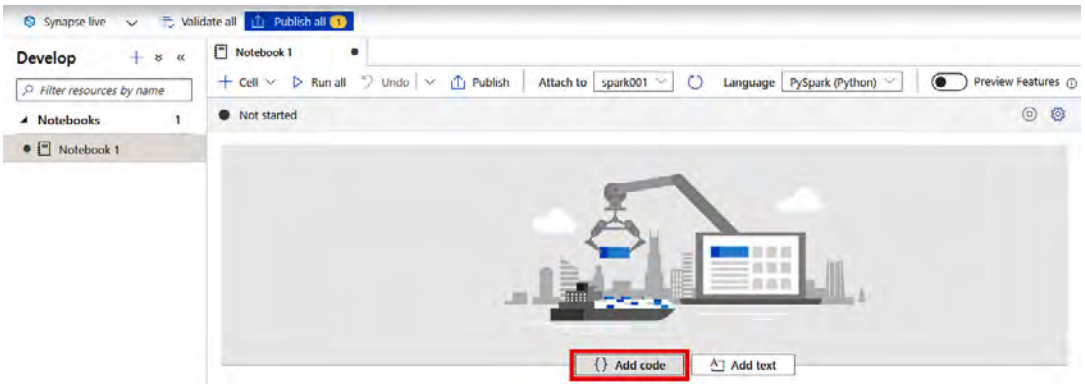


Figure 2.33: Adding code in the notebook cell

3. Enter the following code in the new notebook code cell. Then click the **Run** button:

```
%%spark
spark.sql("CREATE DATABASE IF NOT EXISTS sparknyc")
val df = spark.read.sqlanalytics("sqlpool001.dbo.TaxiTrip")
df.write.mode("overwrite").saveAsTable("sparknyc.taxitrip")
```

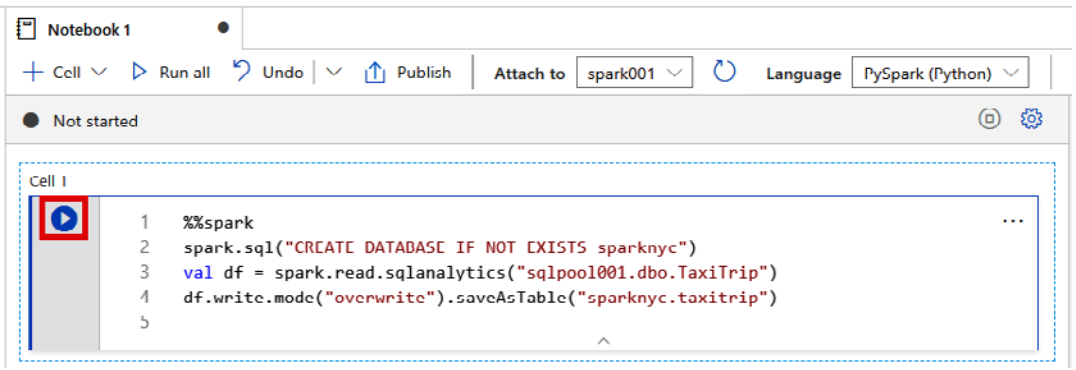


Figure 2.34: Running Spark (Scala) code to ingest SQL pool data into a Spark database

> **Note**
>
> The default language for the notebook is PySpark (Python). By using the **%%spark** magic in line one of the code, we are able to quickly switch this cell language to Spark (Scala).

4. If you go to the **Data** hub and refresh the databases, you will see our newly created **`sparknyc (Spark)`** database:

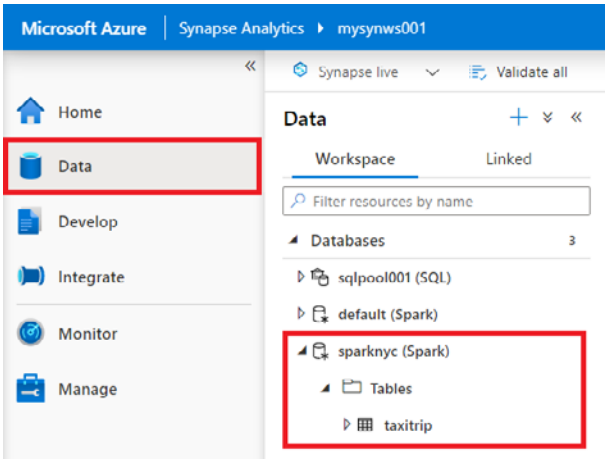


Figure 2.35: The newly created Spark database

Now that we have successfully ingested the data from the SQL pool into our Spark table, we will look at how to analyse data using Spark and notebook in the following section.

Analysing data using Spark and notebook

Now that we have ingested data into our new **`sparknyc`** Spark database, let's use Spark and notebook to perform some data analysis:

1. In Synapse Studio, select the **Develop** hub on the left-hand pane, then **+** and **Notebook**.

2. We are now going to run the same analysis in the Spark as we did in our earlier SQL pool example. We will save the results to a new **`passengerstats`** Spark table. Enter the following code in the new notebook code cell. Then click the **Run** button to run the code in the default notebook language of **PySpark** (Python):

```
df = spark.sql("""SELECT PassengerCount,
        SUM(TripDistanceMiles) as TotalTripDistance,
        AVG(TripDistanceMiles) as AverageTripDistance
   FROM sparknyc.taxitrip
   WHERE PassengerCount > 0 AND TripDistanceMiles > 0
   GROUP BY PassengerCount
   ORDER BY PassengerCount""")

display(df)
df.write.saveAsTable("sparknyc.passengerstats")
```

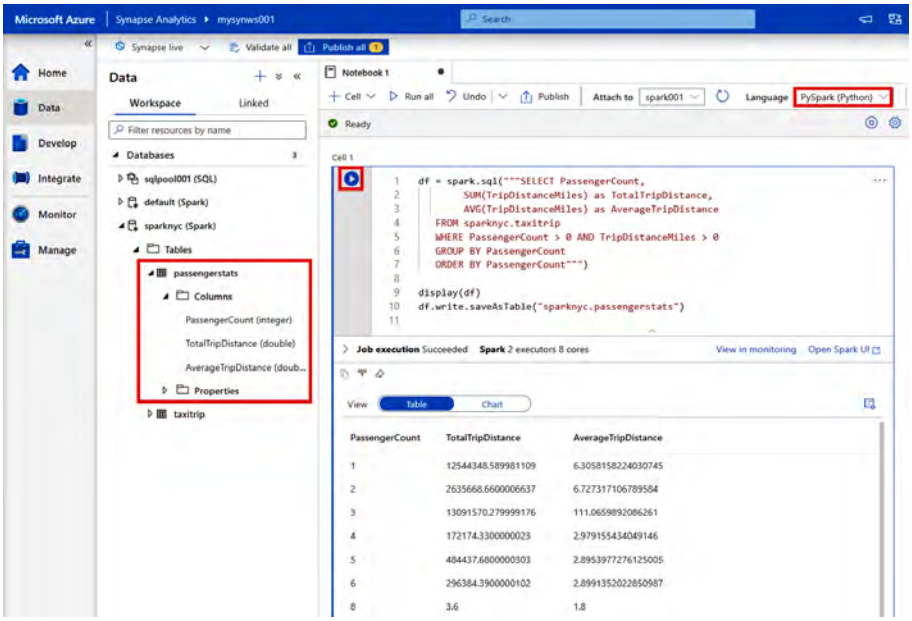


Figure 2.36: Code results

3. As in our SQL pool example, we can quickly change our view to **Chart** to visualise the results:

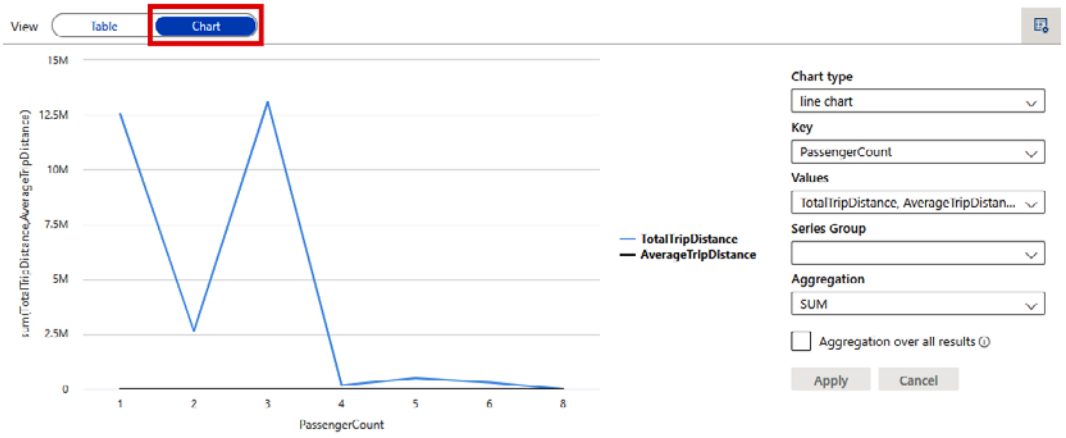


Figure 2.37: Line chart visualisation

Earlier, we showed you how to ingest data from an SQL pool to a Spark table. Now, to complete the circle, we will show you how you can ingest data from a Spark table back into an SQL pool.

Ingesting Spark table data into an SQL pool table

To complete our tour of SQL pools and Spark, we will now demonstrate how to load data from the **passengerstats** Spark table back to an SQL pool table called **sqlpool001.dbo.PassengerStats**:

1. In Synapse Studio, select the **Develop** hub on the left-hand pane, then click **+** and **Notebook**.
2. Add the following code in a new notebook code cell:

```
%%spark
val df = spark.sql("SELECT * FROM sparknyc.passengerstats")
df.write.sqlanalytics("sqlpool001.dbo.PassengerStats",Constants.INTERNAL)
```

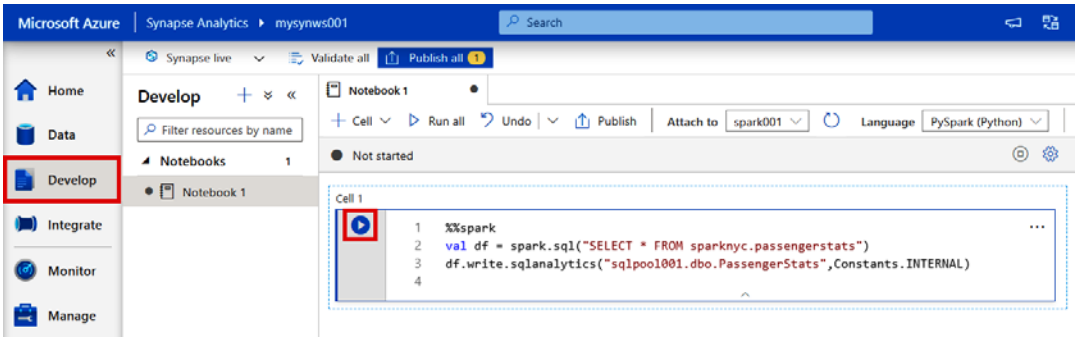


Figure 2.38: Adding a code in a notebook code cell

> **Note**
>
> The default language for the notebook is PySpark (Python). By using the **%%spark** magic in line one of the code, we are able to quickly switch this cell language to Spark (Scala).

3. And just like that, the new **`dbo.PassengerStats`** table is created in the SQL pool along with all the data from the Spark table:

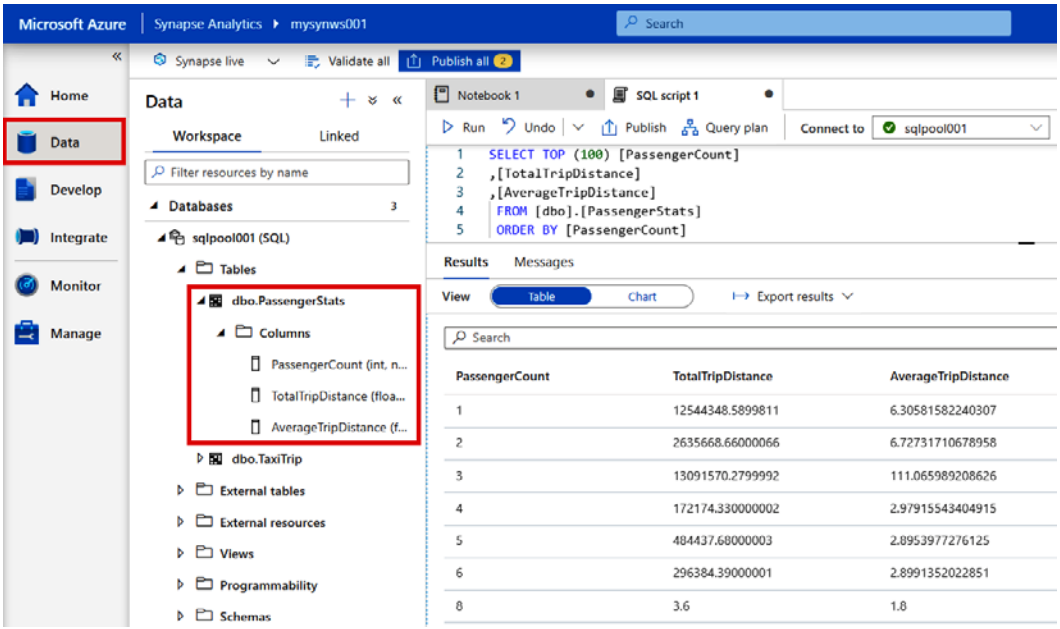


Figure 2.39: New table created in the SQL pool

Now that we have successfully ingested the data from our Spark table into an SQL pool table, we will look at how to analyse data using serverless SQL pools.

Analysing data using serverless SQL pools

Another powerful capability of Azure Synapse Analytics is the ability to analyse data with serverless SQL pools. Serverless SQL pools allow you to run SQL queries without provisioning resources. This allows ease of exploration and data analysis in Azure Data Lake Storage Gen2 without any set-up or infrastructure maintenance.

We will now demonstrate how to use serverless SQL pools to analyse data in Azure Blob Storage:

1. In Synapse Studio, select the **Data** hub on the left-hand pane.
2. Under **Linked**, right-click on **nyc_tlc_yellow | New SQL script | Select TOP 100 rows**:

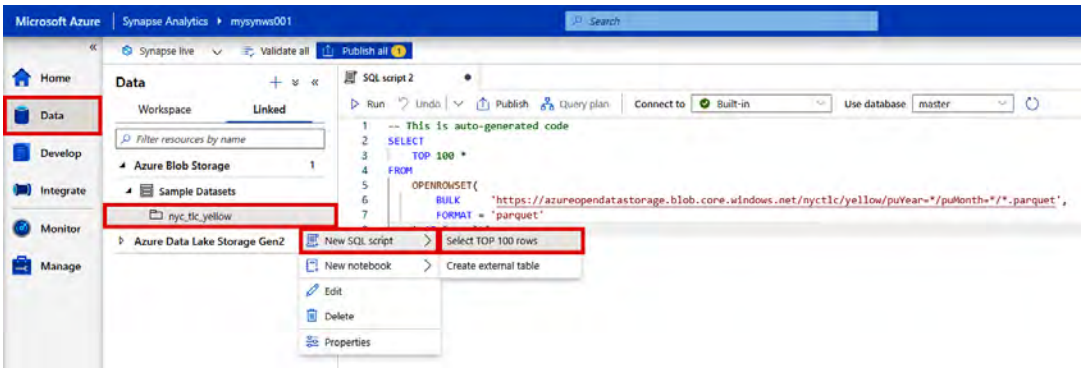


Figure 2.40: Using serverless SQL to analyse data in Azure Blob Storage

3. Click **Run** to see the results.

From this demonstration, you can see that serverless SQL pools allow you to instantly execute queries without having to provision any resources. We will look at how to build data pipelines and perform code-free data transformations next.

Integrating with pipelines

The **Integrate** hub allows you to build data pipelines and perform code-free data transformations. An activity defines the actions to perform on data such as copying data, running a notebook or running an SQL script. A pipeline is a logical grouping of activities that perform a task together. In this section, we will demonstrate how to integrate pipelines and activities using Synapse Studio:

1. In Synapse Studio, select the **Integrate** hub on the left-hand pane.
2. Click **+** and **Pipeline** to create a new pipeline:

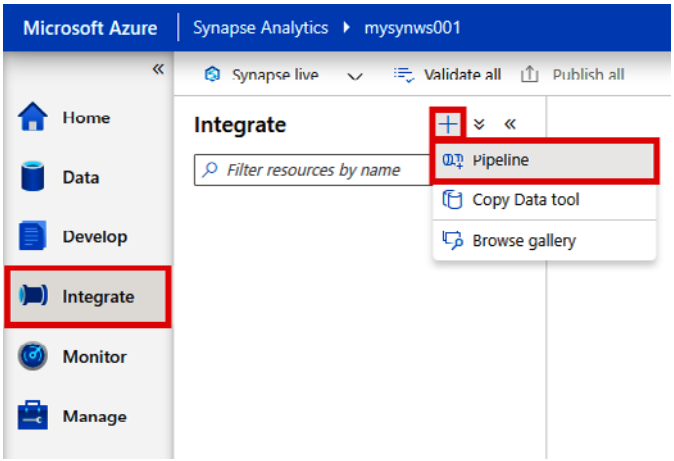


Figure 2.41: Creating a new pipeline

3. Go to the **Develop** hub and drag an existing notebook into the pipeline:

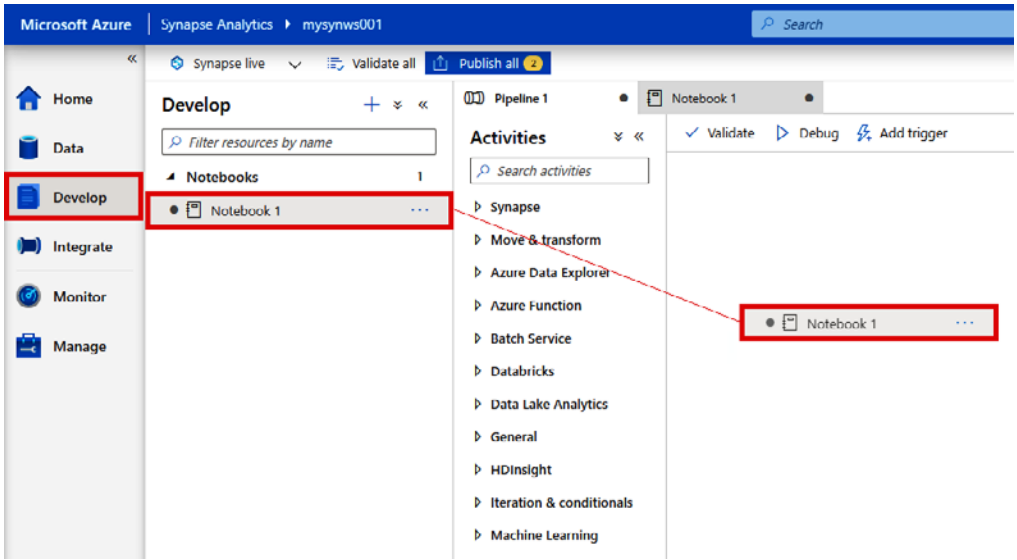


Figure 2.42: Drag an existing notebook into the pipeline

4. Click **Add trigger** and **New/Edit**:

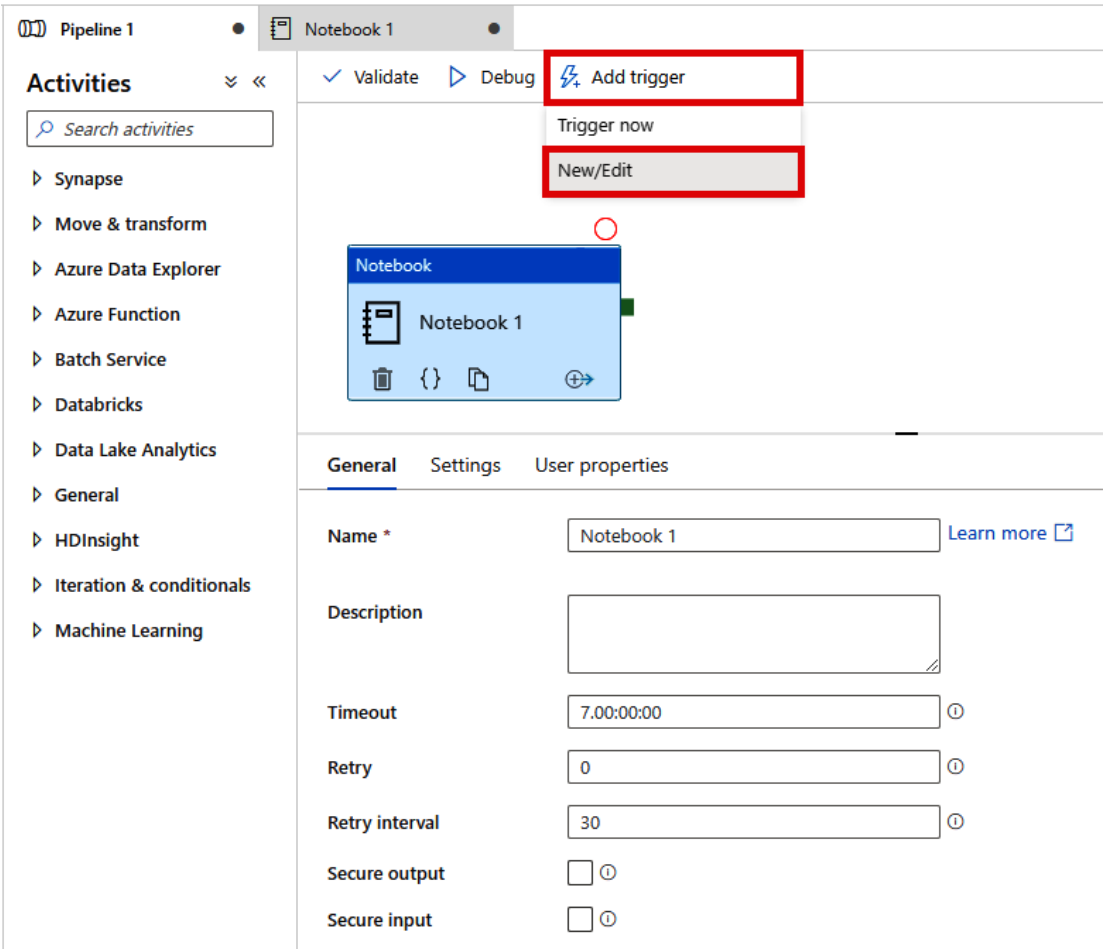


Figure 2.43: Adding a trigger

5. In **Choose trigger...**, select **New**:

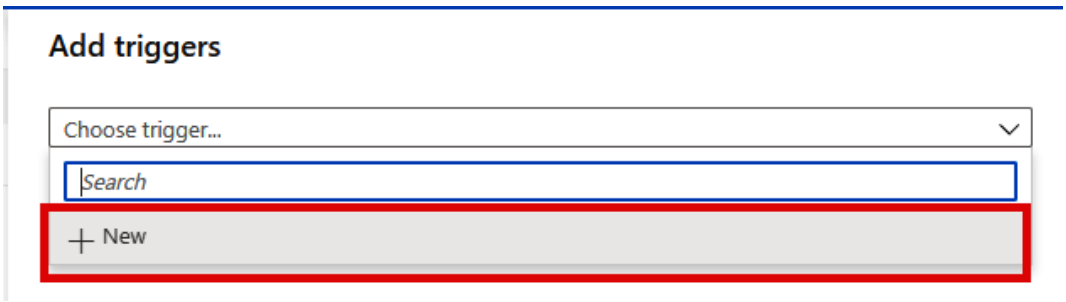


Figure 2.44: Choosing a trigger

6. Configure the trigger as shown, then click **OK**:

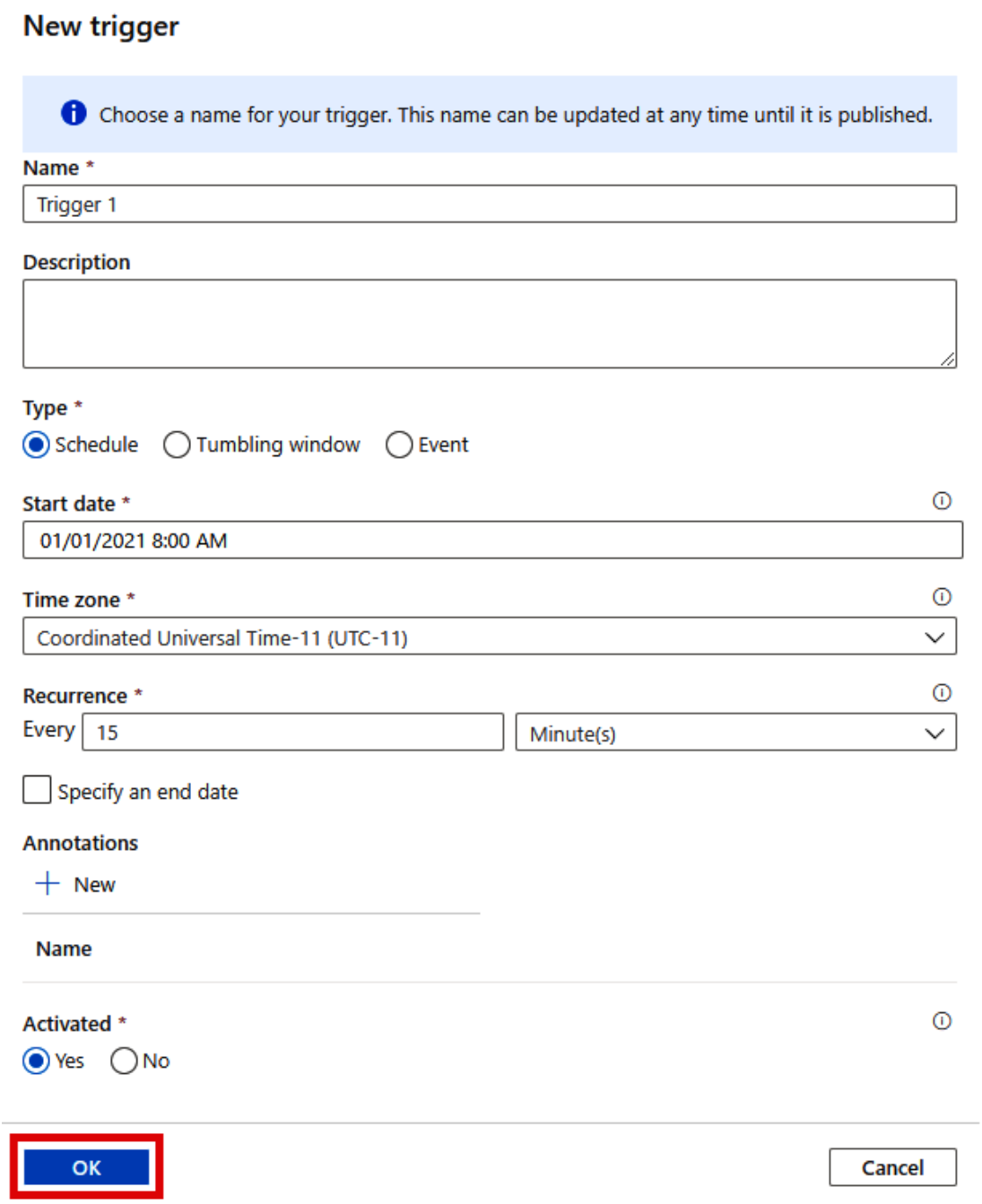


Figure 2.45: Configuring the trigger

7. Commit the changes and activate the trigger by clicking **Publish All**.
8. To run the pipeline immediately, simply go to **Add trigger** and **Trigger now**:

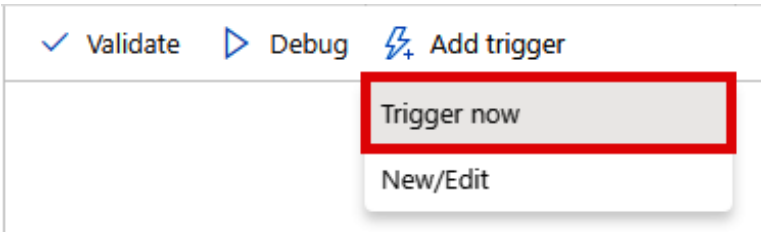


Figure 2.46: Choosing the 'Trigger now' option

As you can see, you can create a pipeline in Azure Synapse Analytics very easily and perform powerful data transformations without writing a single line of code.

The Monitor hub

The final feature that we will look at is the **Monitor** hub. This is where we can monitor currently running jobs. The best way to understand how it works is by seeing it in action:

1. Run the pipeline that we created in the previous section by using **Trigger now**.
2. In Synapse Studio, select the **Monitor** hub on the left-hand pane, and click **Pipeline runs**:

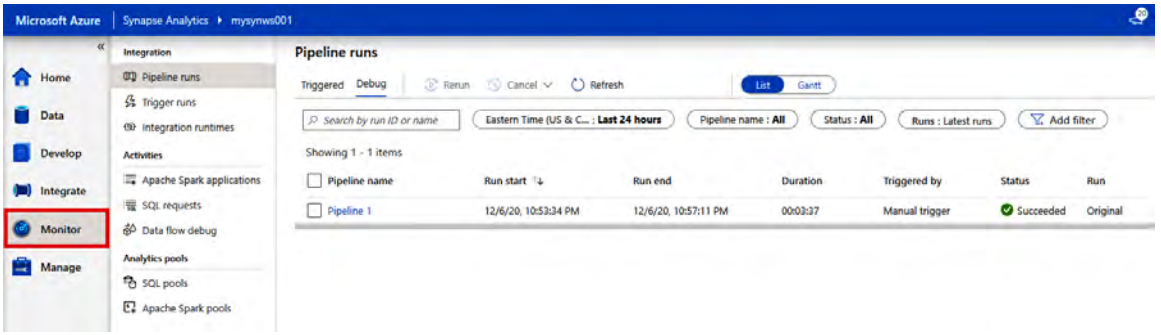


Figure 2.47: Monitor hub

In the **Pipeline runs** section, you can monitor the progress of your pipeline execution. Furthermore, if you encounter an error with your pipeline execution, you can use the **Monitor** hub to troubleshoot the issue.

In addition to pipeline monitoring, you can also monitor triggers, integration runtimes, Apache Spark applications, SQL requests and data flow debug.

We hope that you enjoyed our introductory tour of the Azure Synapse Analytics workspace and Synapse Studio. Before we conclude this chapter, we'd like to highlight some of the advanced features in Azure Synapse Analytics that you can explore on your own:

- Optimisation using materialised views and resultset caching
- Workload importance and workload isolation
- Row-level security
- Dynamic data masking
- Integration with AutoML
- SQL predict
- CI/CD integration

Summary

Azure Synapse Analytics is a groundbreaking evolution of Azure SQL Data Warehouse. It takes the best of the Azure SQL Data Warehouse and modernises it by providing more functionalities for SQL developers, adding querying with serverless SQL pools, machine learning support, embedding Spark natively, collaborative notebooks and data integration – all within a single service.

As you have learned in this chapter, data engineers can provision an Azure Synapse Analytics workspace in a matter of minutes and start building their end-to-end analytics solutions using a unified, simplified and streamlined approach inside Synapse Studio. This remarkable and innovative all-in-one environment is a dream come true for many data professionals.

In the next chapter, you will look at Power BI and Azure Machine Learning. Later, we will see real use cases for how all of these technologies are integrated to provide the complete end-to-end data warehouse solutions that business decision-makers can use to derive meaningful insights from real-time data.

3

Processing and visualising data

In the previous chapters, you were introduced to Azure Synapse Analytics and Synapse Studio. Azure Synapse Analytics offers a lot of features that are flexible and highly scalable. Some of the extensibility features of Azure Synapse Analytics are the ability to use the modern data warehouse approach to serve meaningful reports and a data source for machine learning.

There are many ways to process and visualise data using Azure and the wider set of tools offered by Microsoft. This book will focus on using Power BI as a tool to create reports and dashboards. Power BI offers a wide range of products and services. This chapter will take a look at using Power BI Desktop on Windows to create meaningful dashboards as well as using Power BI workspaces to host and share published reports.

This book will also give you an introduction to the artificial intelligence and machine learning ecosystem within Azure and Microsoft. We will tackle a specific machine learning platform in later sections using Azure Machine Learning.

These technologies work seamlessly with Azure Synapse Analytics as a default approach to visualising and gaining insights from a unified data warehouse.

Power BI

Power BI is a suite of tools that enables users to visualise data and share insights across teams and organisations, or embed dashboard analytics in their websites or applications. It supports different data sources (both structured and unstructured data types) and helps analysts and end users create live dashboards and reports about business data on-demand. An example of this is visualising company sales for recent months and determining the city that sold the most items.

What makes Power BI different from spreadsheet software such as Microsoft Excel is that it is designed to be a hosted user interface, often a live dashboard, where users don't need to frequently store a file in their local machine and open it. With Power BI, you can leverage the power of the cloud to harness complex data and present it through rich graphs or charts, letting the server run all the computations rather than your own machine. Imagine a scenario where your data size was to grow from 500 megabytes to several gigabytes. Most general-purpose machines, such as personal computers with a limited amount of memory, would struggle to load such an Excel file; however, with Power BI, it is just like opening a web page as it is a hosted service:

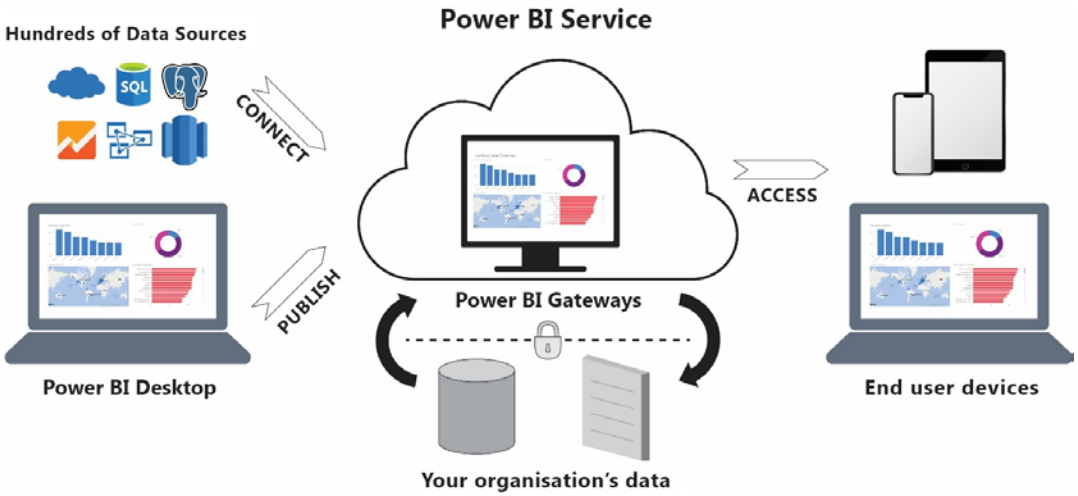


Figure 3.1: Power BI architecture

As seen in *Figure* 3.1, Power BI is composed of different components that can perform different functions:

- **Power BI Desktop:** This is a Windows desktop-based application that is often referred to as an authoring tool, where you primarily design and publish reports to the service.
- **Power BI Service:** This is the managed platform to deploy your reports on the web for your organisation to access. It is a **Software-as-a-Service** (**SaaS**) application and has evolved from being Power BI for Office 365 to just Power BI.

- **Power BI Mobile Apps:** These are native mobile applications that can access reports from a workspace that is hosted in Power BI. It is available on the Apple iOS App Store and Google Play Store.
- **Power BI Gateways:** A Gateway is a mechanism to sync external data into Power BI. For enterprise scenarios with on-premises storage, a Power BI Gateway acts as a mechanism to query against the data source without the need to transfer databases to the cloud. However, the data that is hosted in Power BI reports lives within the Azure Cloud.
- **Power BI Report Server:** In an on-premises scenario, Power BI Report Server allows you to host Power BI reports within your own data centre. These reports are still shareable across different members, as long as they have the right network access.
- **Power BI Embedded:** Embedded allows you to white label Power BI in your own custom applications. It is often strategically integrated into existing dashboards and back-office systems where only a single set of users can access the reports.

Features and benefits

At a high level, Power BI offers the following benefits:

- Personalised dashboards that allow analysts to brand the look and feel of the graphs, charts and tables
- Collaboration across different users
- Governance and security that ensures that only authorised users can access dashboards
- No memory or speed constraints, as it is a cloud-hosted service. It is as if the user is just loading a web page
- Seamless integration with existing applications to create rich and bespoke dashboard analytics
- No specialised technical support is required, as reports are meant to be easy to interact with
- Support for advanced data services, such as the 'Ask a Question' feature, integration with R, segmentation and cohort analysis

Power BI is an intuitive tool that often just requires clicking or dragging-and-dropping in order to quickly access and visualise data. The authoring tool, Power BI Desktop, is equipped with many built-in features to derive analytics. It is smart enough to suggest a visualisation model based on the fields of your choice.

Power BI dashboards and reports are highly customisable and allow you to personalise your experience depending on your branding. You can select themes, use custom charts, create labels, insert drawings and images and a lot more.

Compared to sending an email with a PowerPoint file attached, Power BI allows more open collaboration between analysts and other members of the company just by sharing a centralised dashboard. You can access reports using major web browsers or by means of mobile applications that you can download from the Apple App Store and Google Play Store. People can send comments and annotations about the reports, creating a faster feedback loop with the use of alerts and notifications.

Power BI is secure in different facets and areas. For one, when authoring a report, it is ensured that you can only access data sources that you have been granted access to. This is backed by **Row-Level Security** (**RLS**). For example, analysts can only access data that is local to their region, making sure they don't have access to another city or country's data. Once you are ready to share the report, you can quickly save it to your personal workspace. You can select whomever you want to share the report with across your organisation, or invite users from external tenants.

If you wish to start small while learning Power BI, you can start by just using Excel files as your data source. There are scenarios where analysts receive a CSV file from data engineers because the size of the dataset is not too large.

In an enterprise scenario, engineers and analysts have to deal with various data sources and platforms to create a unified dashboard. This is often a challenge, especially when data is not available for analysts to perform their exploration. This is where a combination of Power BI and Azure Synapse Analytics can be very useful.

Power BI and Azure Synapse Analytics

Power BI is a very flexible platform that can work with multiple data sources. It could be a source as simple as an Excel file, a database on-premises or in the cloud or a data warehouse. One of the best practices for having a modern data warehouse is separating the duty of processing data from data visualisation. That is where Azure Synapse Analytics comes in.

Azure Synapse Analytics, with the use of a dedicated SQL Pool, can perform complex queries for all data sources and create a semantic model for Power BI. A semantic data model is a structured data format that Power BI can easily interpret for data visualisation. This approach is scalable and eases a lot of the complexities of managing data across various teams.

Features and benefits

There are a lot of benefits to combining Power BI and Azure Synapse Analytics. Although there is no one-size that fits all, nor one tool to solve all of your data problems, the combination of Azure Synapse Analytics and Power BI can bring robust, elastic and more maintainable data operations and lifecycles to your team:

- **Unified data model:** Treating Azure Synapse Analytics as the single source of truth will bring a unified data model for all consumer services of data, including Power BI. With this approach, everyone will be confident that data that is stored in the data warehouse (the Synapse workspace) has already been validated and pre-processed. This is very useful, especially for Power BI developers and analysts, meaning that they can focus on generating meaningful reports instead of wrangling disparate data sources.
- **Scalable queries and data transformations:** Power BI offers data querying capabilities by means of **DirectQuery**. This is very useful for small datasets. However, when datasets become massive, you will encounter some scalability problems and it will affect your reporting performance if you perform these operations directly in Power BI. Azure Synapse Analytics helps in this scenario by doing performance optimisations on large-scale queries by means of materialised views and result set caching. These sets of heavy compute operations are performed in an elastic cluster of SQL Pools to cater to scaling demand. With this approach, Power BI can focus on performing data visualisation-related functions instead of worrying about the data preparation to serve for the reports.
- **Centralised governance and security:** Power BI comes with security features to protect the files and reports generated, such as the use of Azure Active Directory (Azure AD) to make sure that only specific workspaces and reports are shared to a group or users. With the help of Azure Synapse Analytics, data access can be further secured by implementing row-level and column-level security on a data tier rather than in data models. In a practical scenario, the database operations team can easily provision and restrict access to various data sources for Power BI consumption.
- **Collaborative workspace:** One of the challenges in running a business intelligence capability, especially in a large company, is the disparity between tools, teams and policies. In a traditional scenario, these teams have a communication barrier but want to achieve the same goal, which is to leverage the power of data. Azure Synapse Analytics creates a collaborative experience for these teams wherein they all access the same web portal – Azure Synapse Studio. In a large organisation, some teams might only need access to the Develop section of Synapse Studio, while others only access the Monitor section. Synapse Studio provides a unified look and feel to all members of the organisation regardless of roles and responsibilities.

In this book, we will use Power BI to generate reports from the queries performed by the SQL Pool in a Synapse workspace. This chapter will focus on how you can leverage Power BI datasets in Synapse Studio.

There are different approaches to performing robust data visualisation, but the main reason for using this approach (Power BI and Azure Synapse Analytics) is elasticity and the separation of duties. This combination facilitates a streamlined reporting capability at scale, even if you have billions of rows in your databases.

To understand further how Azure Synapse Analytics and Power BI work hand in hand, let's start with a simple quick-start guide.

Quick-start guide (Data modelling and visualisation)

In this quick-start guide, we will explore how we can use Azure Synapse Analytics with the use of an SQL Pool to capture COVID-19 data from Bing and display relevant reports to Power BI.

Prerequisites

In order to perform this activity, you need the following:

- An active Azure subscription
- Power BI Desktop (Windows only)
- Power BI Pro Licence
- An Azure Synapse Analytics workspace

Let's get started.

Integrating Power BI with Azure Synapse Analytics

This activity will act as a step-by-step guide to integrating Azure Synapse Analytics with Power BI. We will be using the power of Azure Synapse Analytics to perform queries that will allow Power BI to visualise data without the need to perform complex transformations within your Power BI environment:

1. The first thing we will do is to establish a connection between your Azure Synapse Analytics workspace and your Power BI workspace. To do this, go to Synapse Studio and go to the **Manage** section. Click on **Linked services** and then click on the **+ New** button. Search for Power BI just like in *Figure 3.2*. Click on the **Power BI** option and click on **Continue**:

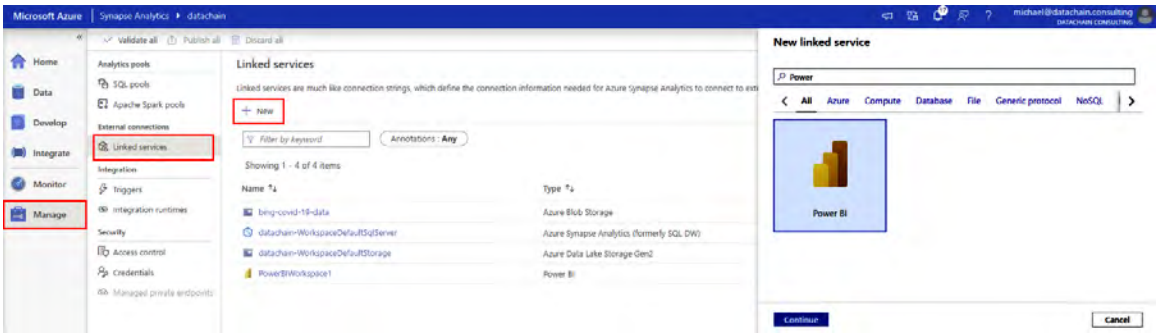


Figure 3.2: Integrating Power BI with Azure Synapse Analytics

2. Fill in **Name**, **Description**, **Tenant** and **Workspace name** for this linked service. Click on **Create**:

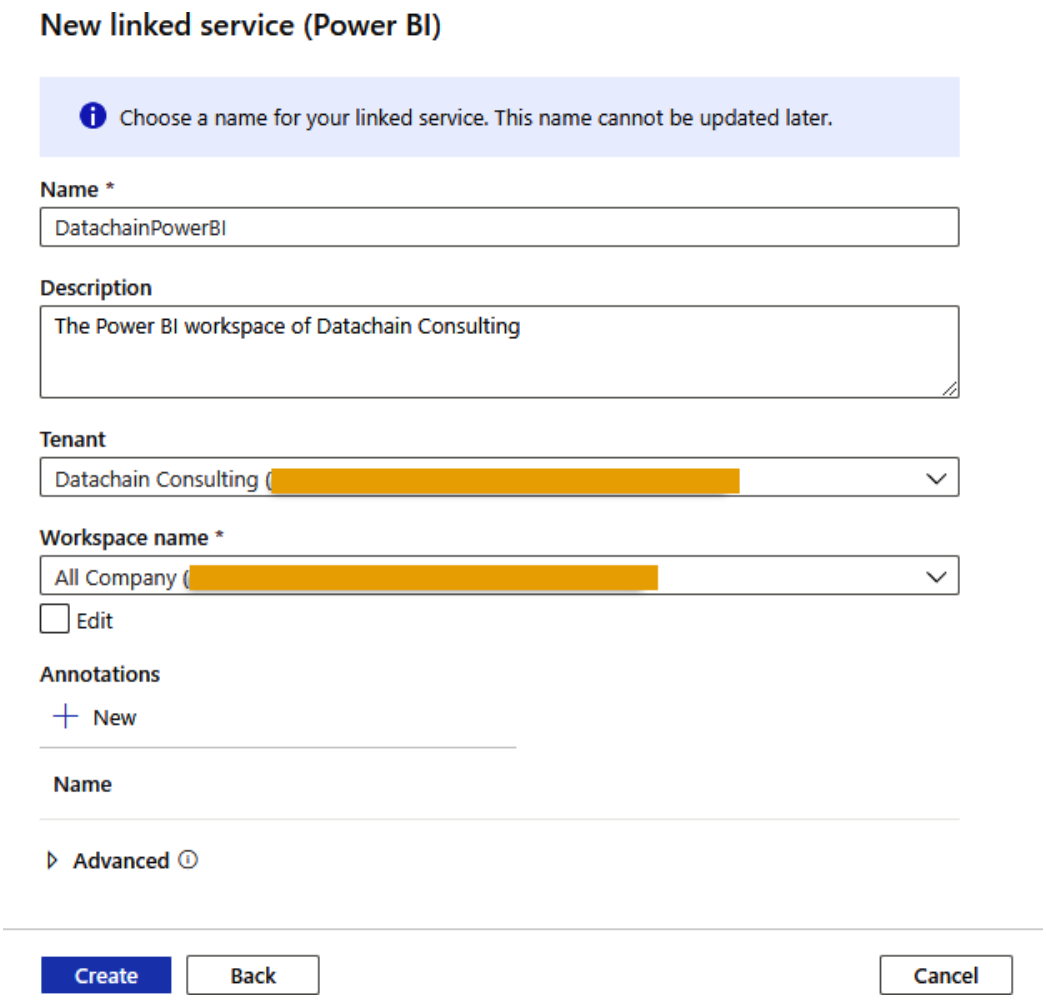


Figure 3.3: Creating a linked service

Creating a database

Follow these steps to create a database for storing data:

1. After setting up the linked service with Power BI, it's time to create a database in order to store the sample data that we will capture in the next section. Go to the **Data** section of Synapse Studio and choose the **Workspace** tab. Click on the **+** button and then choose **SQL database**:

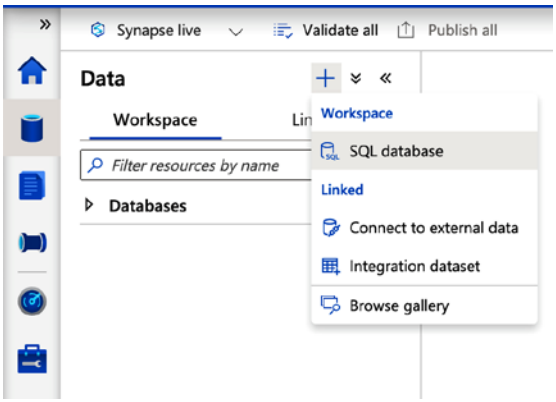


Figure 3.4: Creating an SQL database in Synapse Studio

2. Choose the **Serverless** option and enter a database name. Click on **Create**:

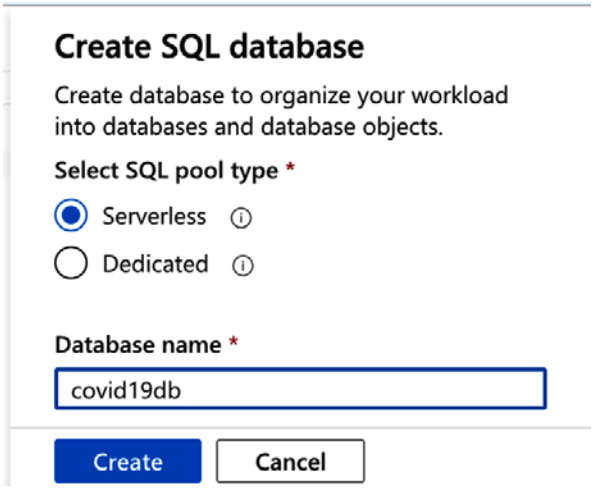


Figure 3.5: Choosing an SQL pool type

> **Note**
>
> When setting up your Azure Synapse Analytics workspace, it comes with a built-in serverless SQL pool for you.

Upon the successful creation of the database, it should now appear in your workspace's list of databases.

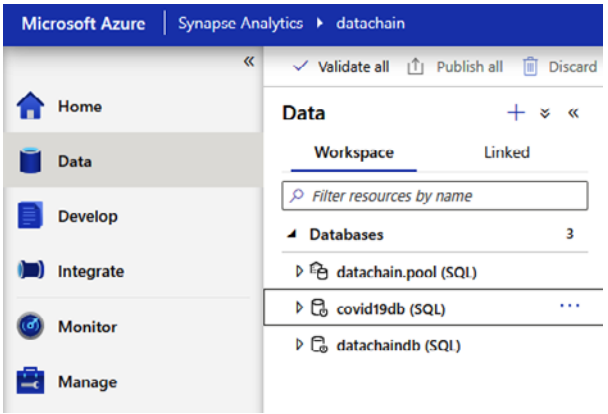


Figure 3.6: List of created databases in the Synapse workspace

Linking the dataset

Follow these steps for linking a dataset:

1. The next part is capturing the COVID-19 data from Bing. On the same page (in the **Data** pane) where you created a database, click again on **+** and choose **Browse gallery**.
2. Choose **Bing COVID-19 Data** from the **Datasets** tab and click on **Continue**:

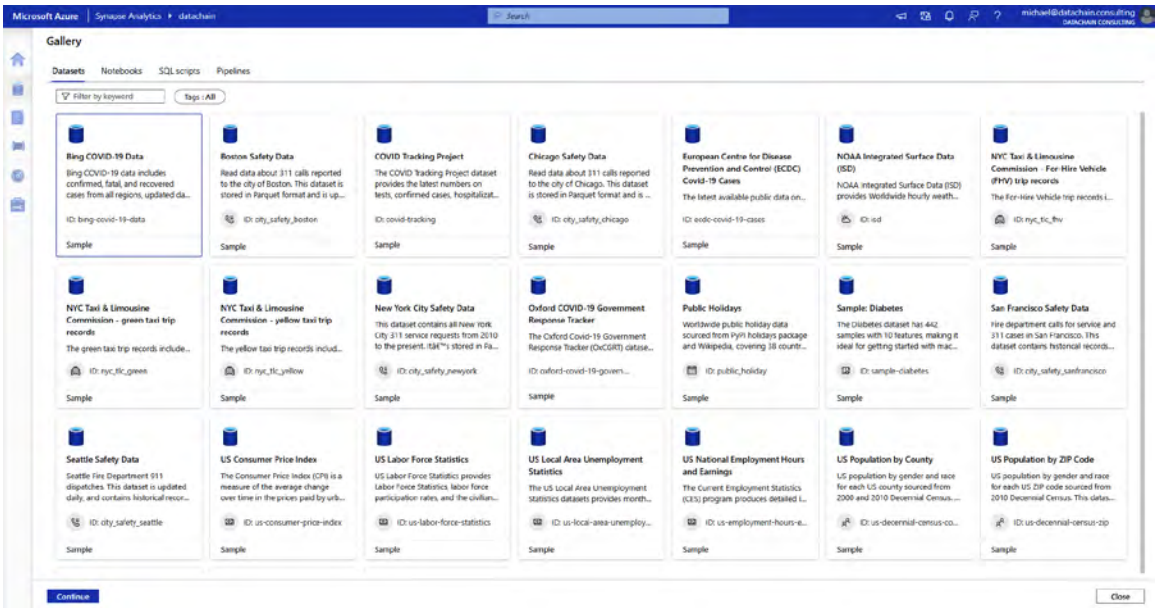


Figure 3.7: Datasets gallery

You will then be brought to the terms and conditions page. Please read the licensing agreements to make sure you follow the fair use of these datasets. It will also show you a preview of some of the rows of that file. Click on **Add dataset**.

The dataset should now appear in your **Linked** data tab under **Azure Blob Storage | Sample Datasets**.

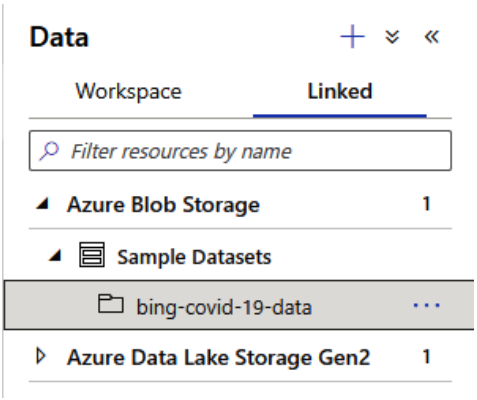


Figure 3.8: Adding the dataset to Azure Synapse Analytics using Azure Blob Storage

Transferring data to your database and running a query

After getting the sample dataset, it's now time to transfer that data to the SQL database that you created earlier:

1. Right-click on the **bing-covid-19-data** folder followed by **New SQL script** and then choose **Select TOP 100 rows**:

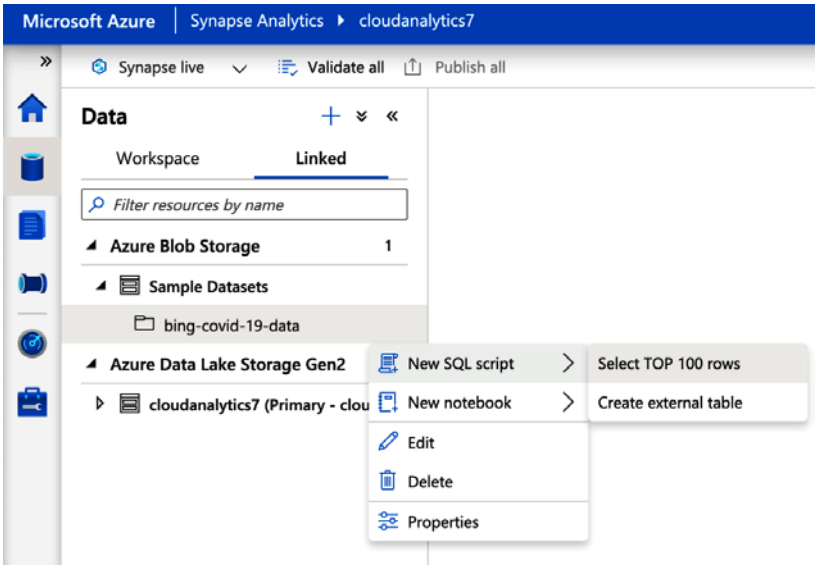


Figure 3.9: Creating a new query for selecting top100 rows

This will prompt you to a pre-filled SQL script development workspace. There are a couple of things that we will modify in this query.

2. Make sure an SQL pool is selected as **Built-in** and the proper database is chosen in the Use database drop-down menu. The Azure Synapse Analytics workspace comes with a built-in SQL pool upon resource creation:

Figure 3.10: Choosing an SQL pool and database

We will modify the query to create a view based on a select statement from the Bing COVID-19 dataset:

```
-- Drop existing "Covid19View"
DROP VIEW IF EXISTS Covid19View;
GO


-- Create a View named "Covid19View"
CREATE VIEW Covid19View
  AS
-- Select ALL data from the dataset however for this specific report.
-- we only want country level reports and excluding "Worldwide" and
country regions.
SELECT
    *
FROM
    OPENROWSET(
        BULK     'https://pandemicdatalake.blob.core.windows.net/public/
curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.parquet',
        FORMAT = 'parquet'
    )
    AS [result]
WHERE [country_region] != 'Worldwide'
    And ( [admin_region_1] is NULL
        Or [admin_region_1] = '' )
```

> **Note**
>
> In the actual report later in Power BI, you might see that other countries are missing. This is intended to simplify the exercise.

3. In the **Properties** section, choose a good query name with an optional description. It's important that you choose the **All rows** option:

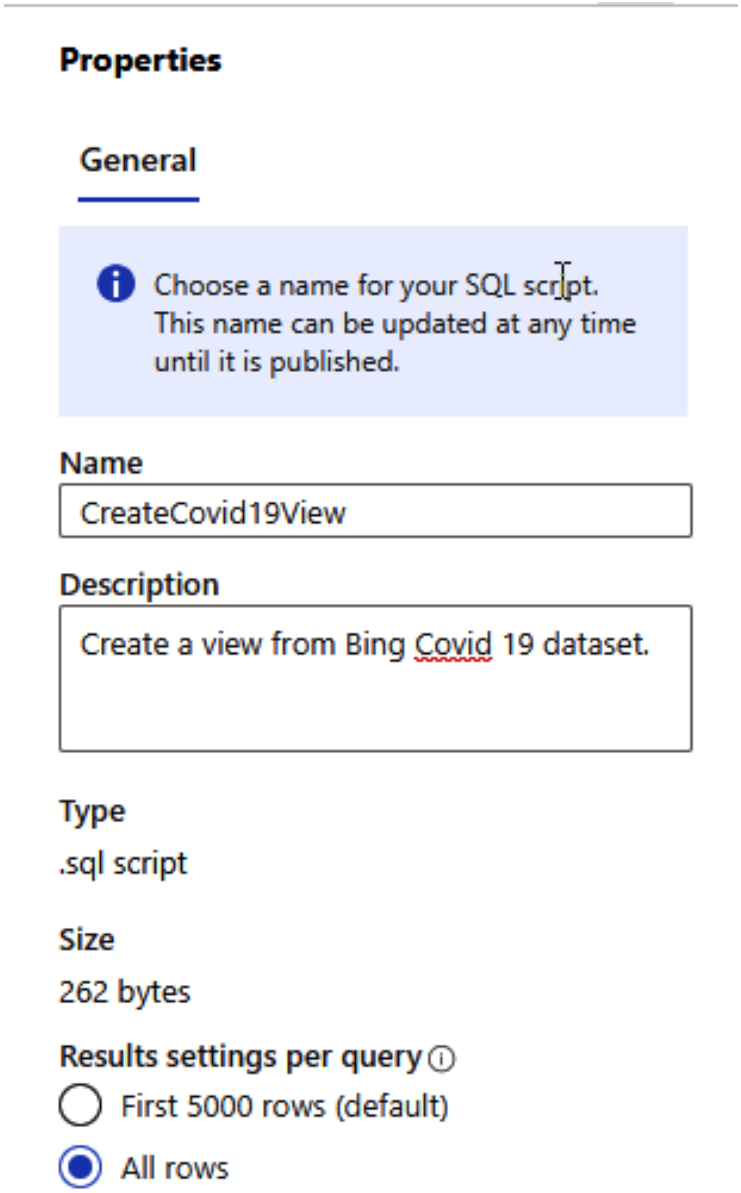


Figure 3.11: SQL script properties

4. Save your script by clicking on the **Publish** button. On Windows, you can also achieve this by pressing *Ctrl* + S.

5. It's now time to run your query. Click on **Run**.

Once the query is finished, it should display the following message:

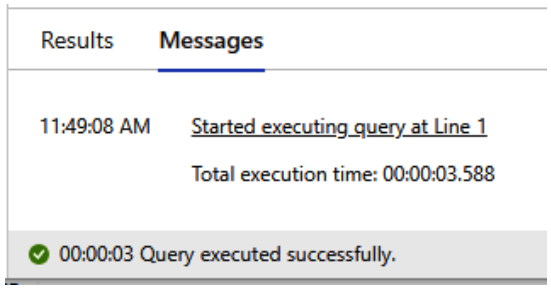


Figure 3.12: Query execution status

Checking the view

Let's check whether the view has been created in the database:

1. Go to the **Workspace** tab of the **Data** section of Synapse Studio. Drill down on your database and views and check whether the view was created:

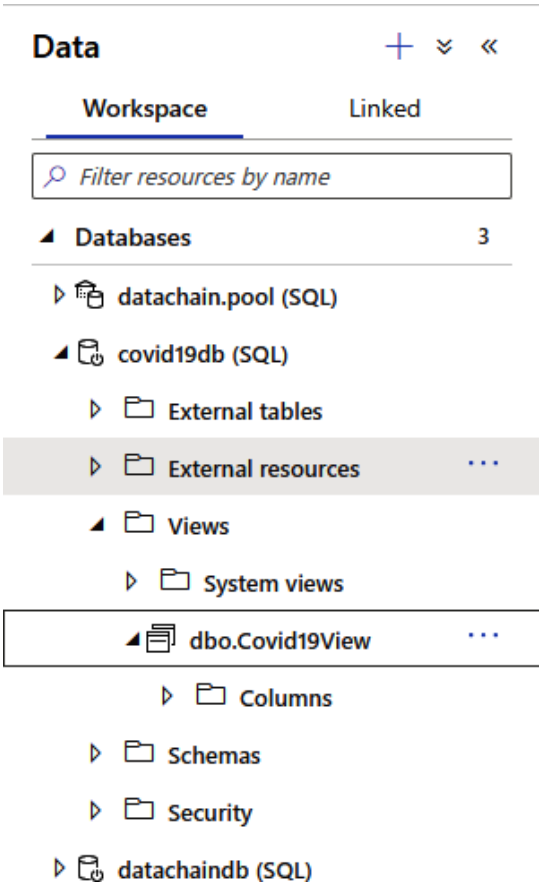


Figure 3.13: View created for the COVID-19 dataset

2. Right-click on the view that was created, click on **New SQL script** and click **Select TOP 100 rows**:

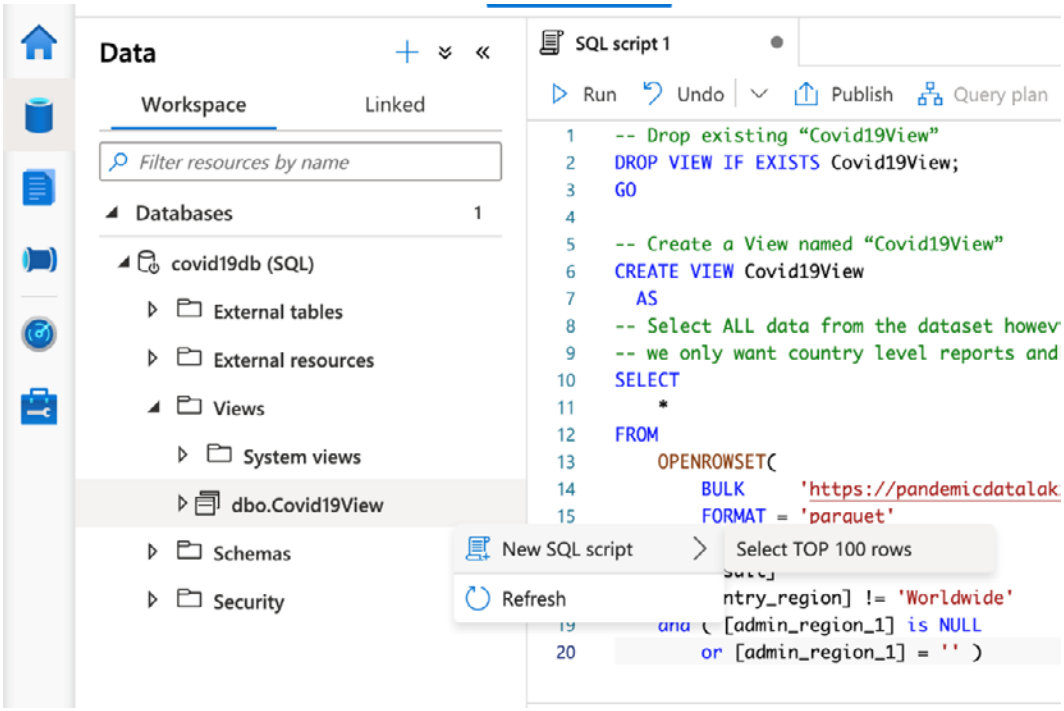


Figure 3.14: Querying the top 100 rows of Covid19View

3. This should open a new window. Click on **Run** to execute the select query. It should display rows of data in the **Results** tab:

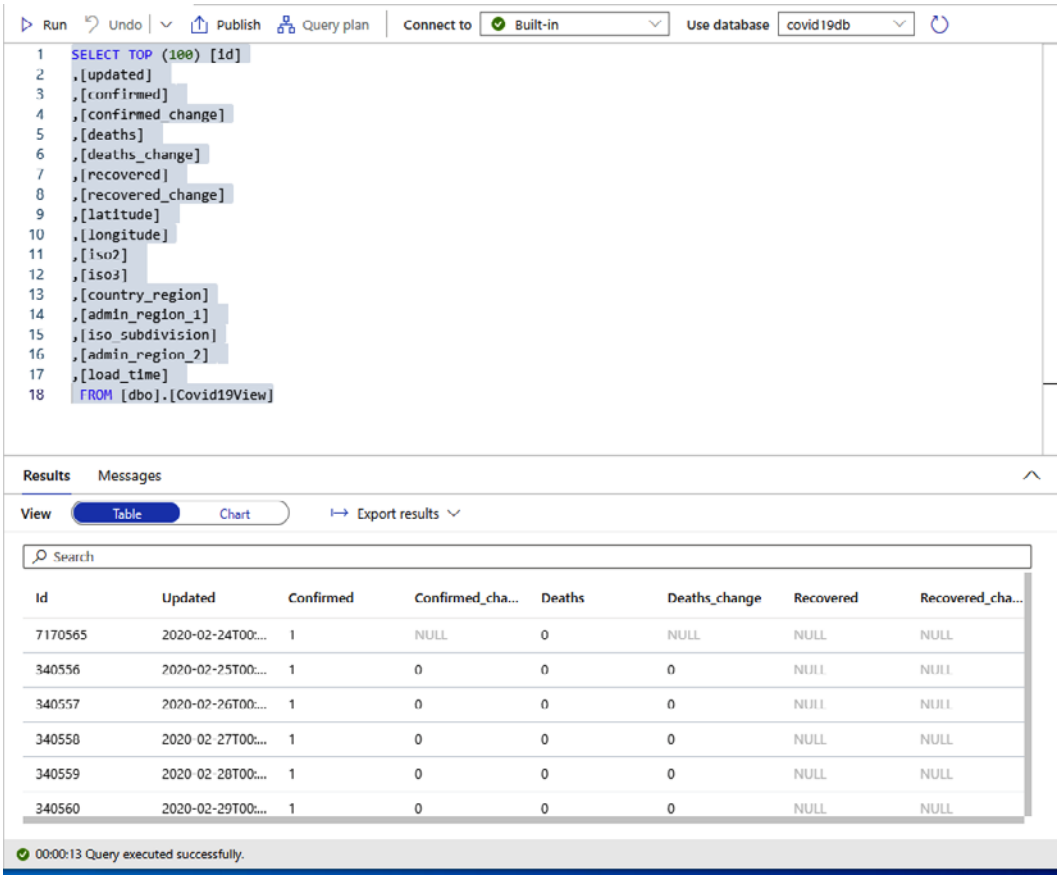


Figure 3.15: Executing the query to get the top 100 rows

Creating and publishing the Power BI report

Now that we have populated the Azure Synapse Analytics database with data, it's time to create a report in Power BI:

1. Go to the **Develop** section of Synapse Studio and drill down to **Power BI | Your Workspace | Power BI datasets**. Click on **+ New Power BI dataset**:

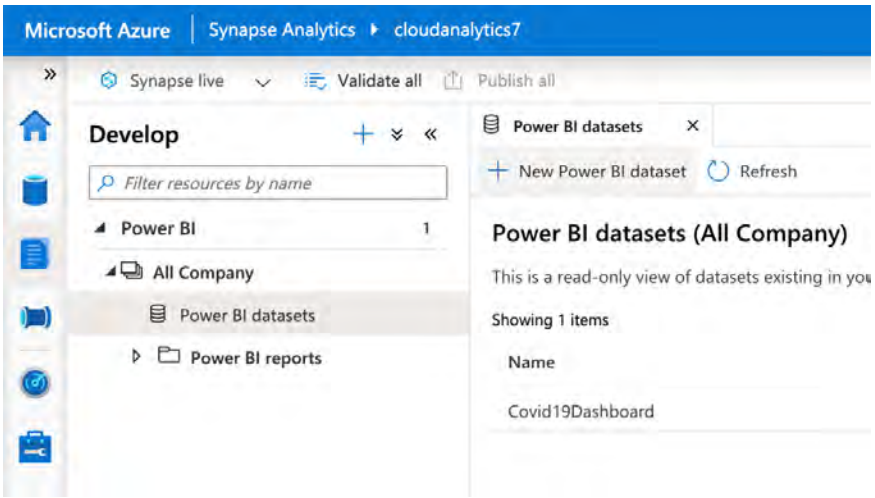


Figure 3.16: Adding a new Power BI dataset

2. This will open a wizard that will allow you to choose a database. Click on **Continue**:

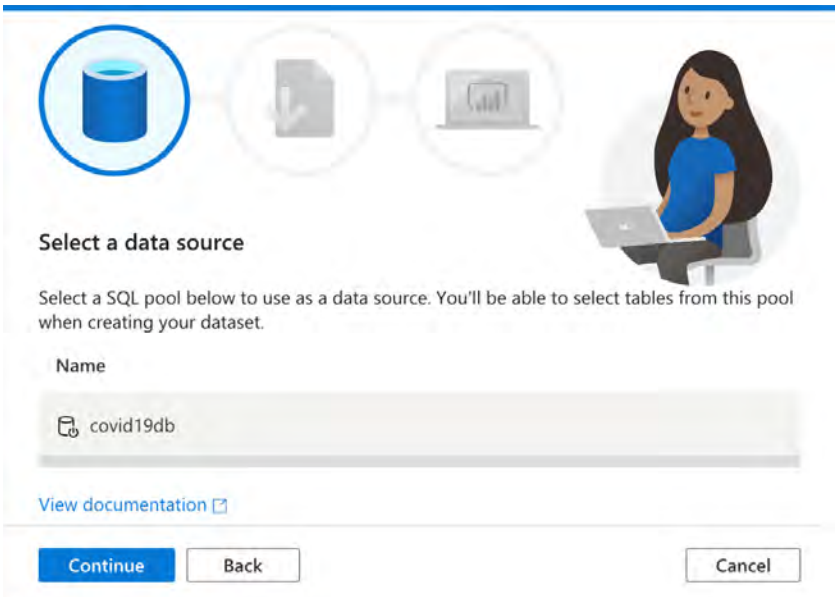


Figure 3.17: Selecting a data source for the Power BI dataset

3. After choosing a database, the next part of the wizard enables you to download a Power BI desktop file. Click on **Download**:

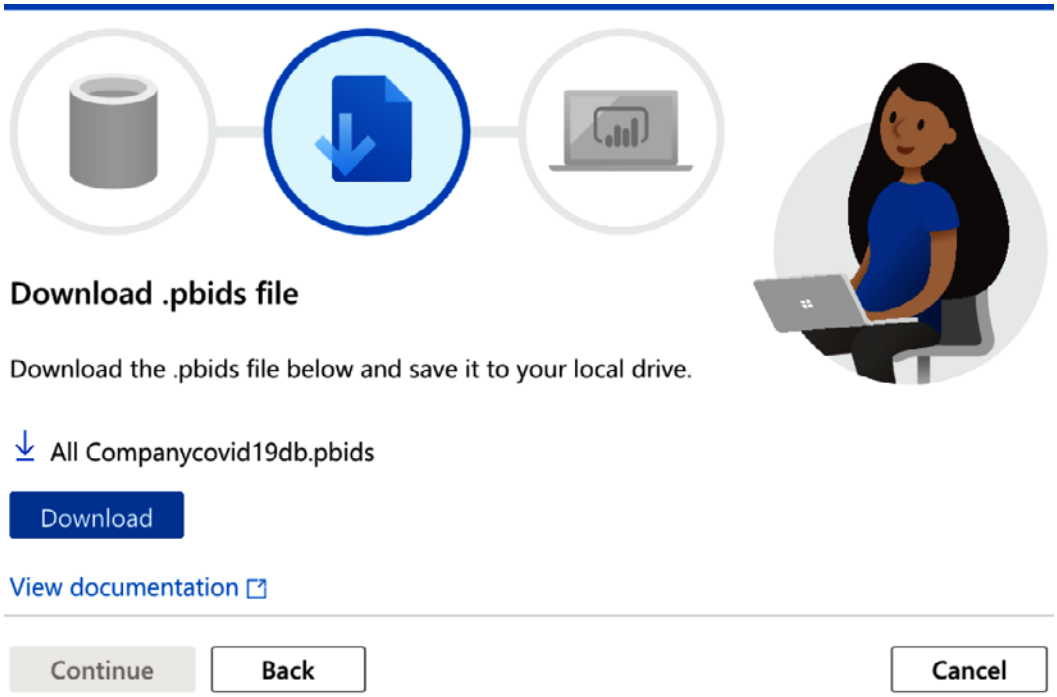


Figure 3.18: Downloading a linked Power BI dataset

After downloading the file, open the `.pbids` file. The file should open Power BI Desktop on your Windows machine. If you're not yet authenticated, it will prompt you to choose a method to connect to the database:

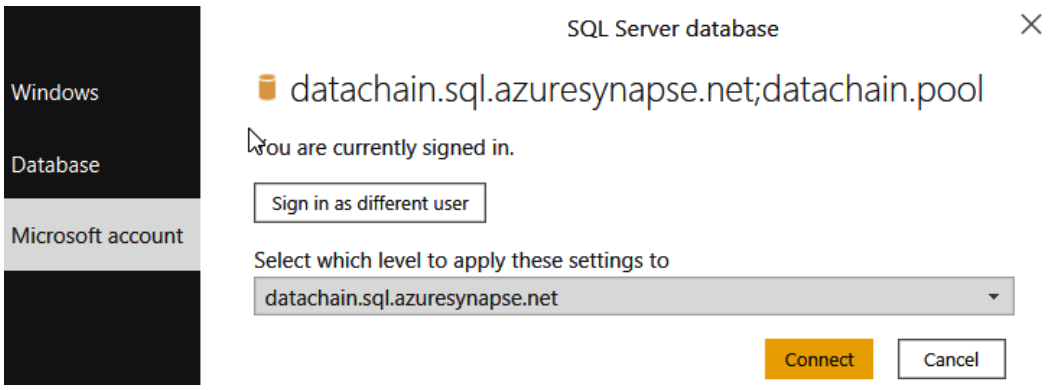


Figure 3.19: Adding an Azure Synapse Analytics SQL database as a data source in Power BI

Once you are authenticated, it will then display a **Navigator** screen for you to select data sources you want to capture. Make sure that the view where we stored our data is ticked. Then hit **Load**:

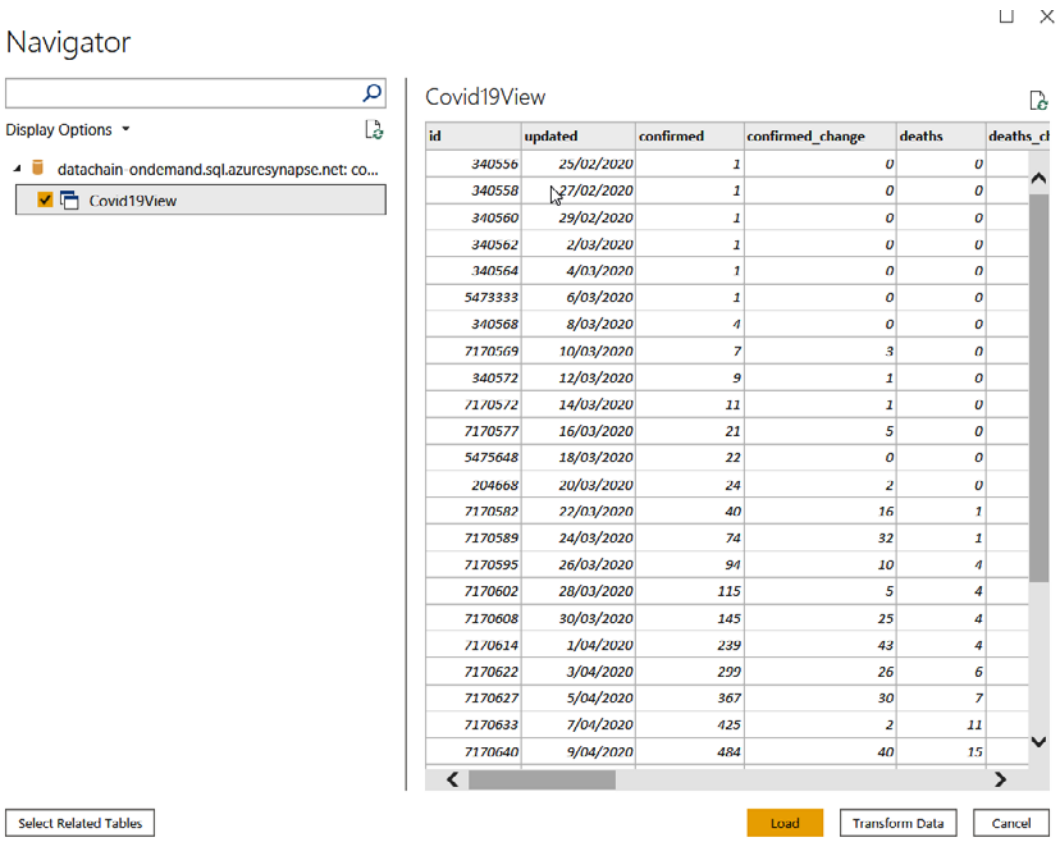


Figure 3.20: Loading the COVID-19 dataset into Power BI

You will then be prompted with an option to choose either **Import** or **DirectQuery**:

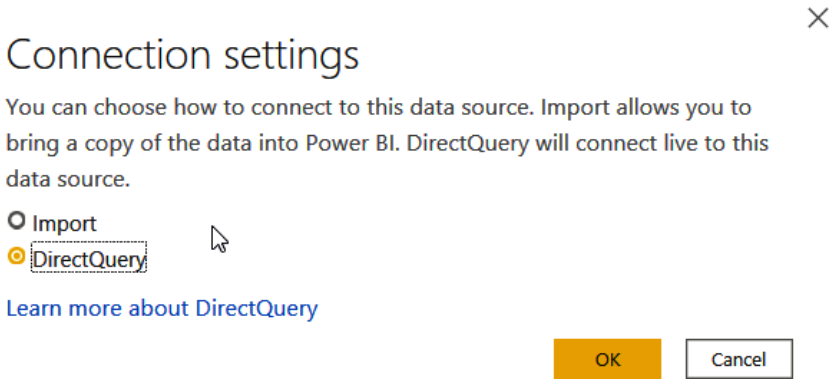


Figure 3.21: Selecting the Power BI connection settings

By choosing the **Import** option, the data will be stored in Power BI Desktop. You will need to perform a refresh in order to get an update of that data. **Import** is the preferred approach if the data size is not that big and if your machine has enough memory (RAM).

If you choose **DirectQuery**, no data will be stored in Power BI Desktop. **DirectQuery** will only create a reference from the data source and will query that data each time you update your workspace. **DirectQuery** is preferred if you have complex external data sources.

After selecting an option, it will then load the data to add as your Power BI fields.

4. To get started creating a report, drag the **updated field** from the **Fields** tab (on the right of Power BI) to the workspace. Then, on the **Visualisations** tab, choose **Matrix**:

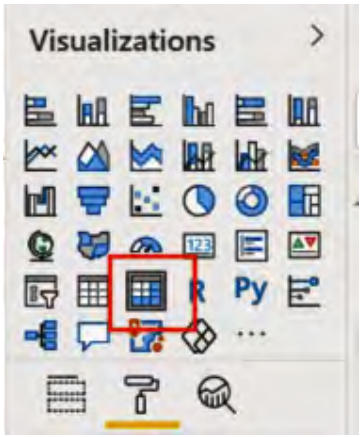


Figure 3.22: The Visualisations tab

5. In the **Format** section, go to **Subtotals** and toggle off **Row subtotals**.

   This should give you a date selector like the one here:

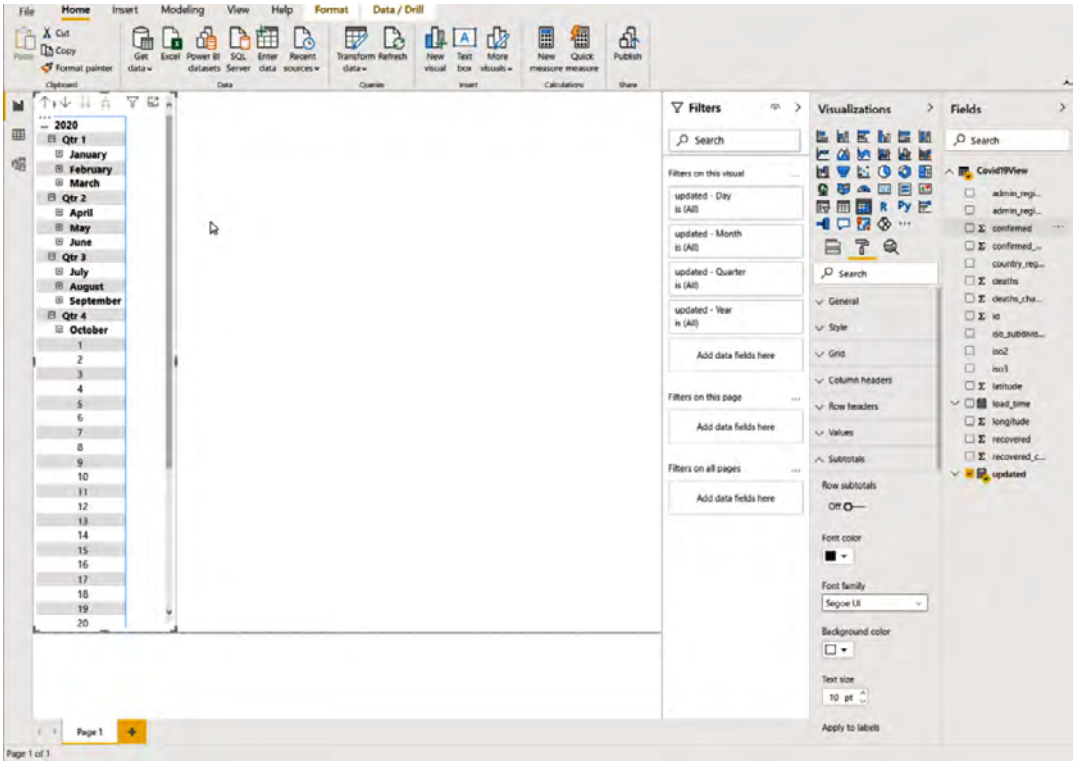


Figure 3.23: Power BI Desktop with a date matrix

6. It's time to display much more meaningful insights. Go to the **Visualisations** section and click on **Filled Map**. This will create an empty map in the workspace. Expand it to capture the rest of the workspace area.

   Drag and drop the following fields onto the map:

   - **country_region**
   - **confirmed**
   - **recovered**
   - **deaths**

   This should create a world map that is filled.

7. Go to **Date selection** at the left and choose a day by drilling down from the matrix. For example, October 21.

8. Go to the **map** section and hover to a **country**. It should show you the number of confirmed, recovered and death cases for that country on that date:

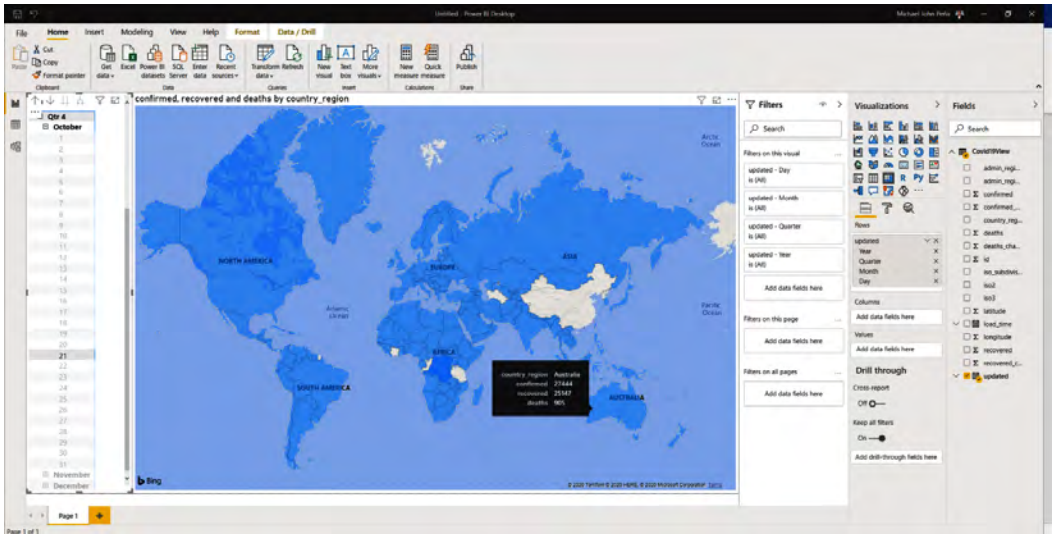


Figure 3.24: Displaying country-wise COVID-19 statistics

Feel free to explore the customisations you might want to implement.

After further customisation, let's publish the report so that other people can see it.

1. Click on the **Publish** button at the top part of the **Home** ribbon. Choose a destination for this report:

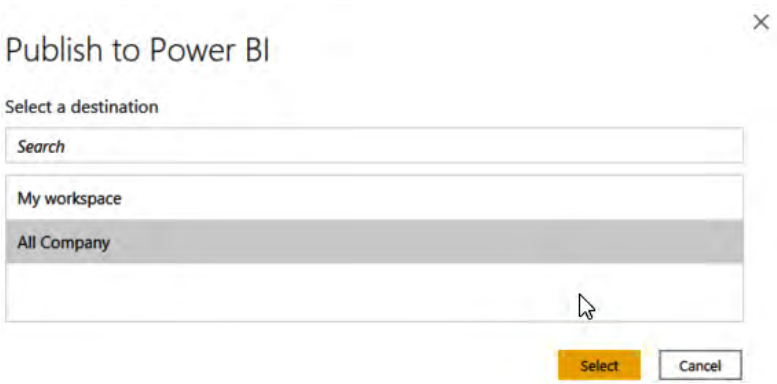


Figure 3.25: Choosing a destination for the report

2. Once it's published, it will give you a link to open the report at PowerBI.com. Open the link:

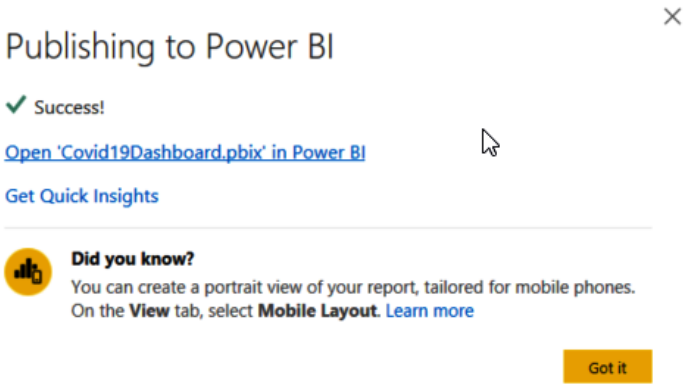


Figure 3.26: Opening the published report in Power BI

Once it's published, it will be available to your Power BI workspace. In the workspace, you can share this report with any of your team members within your tenant by clicking on the **Share** button:

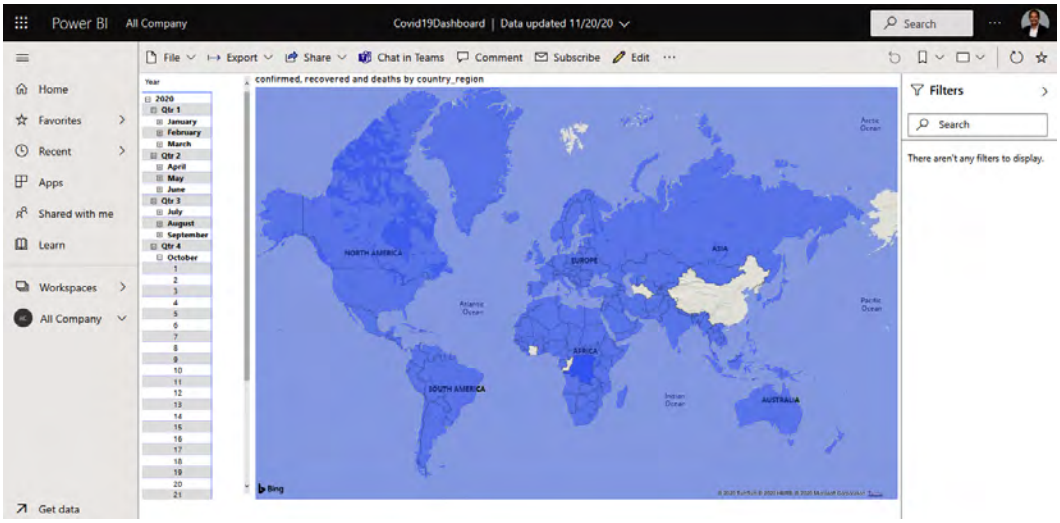


Figure 3.27: Viewing a report in the Power BI workspace

This report can also be viewed on tablet and mobile devices:

Figure 3.28: Viewing the Power BI report on a tablet/mobile device

Congratulations on creating your Power BI dashboard! In this activity, we covered how to use Synapse Studio to transform data into a meaningful semantic model that Power BI can visualise. In the next section, we will explore how we can use machine learning with Azure Synapse Analytics.

Machine learning on Azure

There are multiple ways to perform machine learning on Azure. Microsoft enables data science to be more accessible to all types of users and empowers data scientists to be more productive. Microsoft provides a suite of technologies for developers, database engineers and data scientists to create machine learning algorithms. Whatever your level of proficiency and expertise in data science, there is a useful Microsoft service, tool or framework that can accelerate your machine learning journey.

Figure 3.29 depicts a machine learning landscape within the Microsoft Azure ecosystem. You can use pre-trained models with Azure Cognitive Services and directly integrate them with your applications without the need to set up a data pipeline. You can use popular frameworks such as **TensorFlow** and **Keras** in Azure, whether that's by installing them on a virtual machine or using a machine learning workspace. You can choose different platforms such as Azure Machine Learning or Azure Databricks to prepare and run your machine learning experiments:

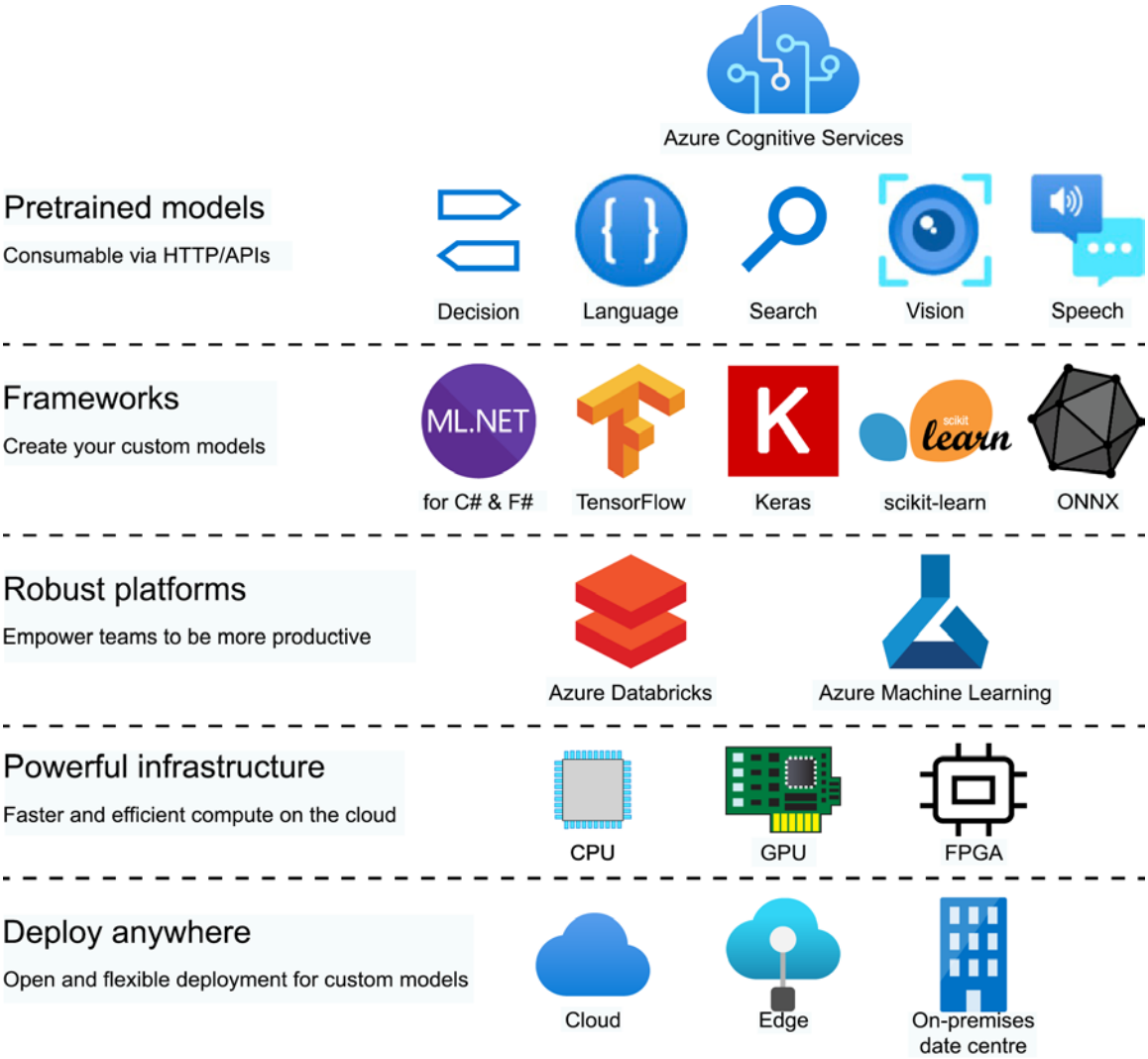


Figure 3.29: Microsoft Azure features and services for machine learning

By using Azure to drive your computation for machine learning analysis, you are provided with specialised hardware that can speed up your experiments. With accelerated hardware such as fast **graphics processing units** (**GPUs**) and **field-programmable gate arrays** (**FPGAs**), you can read billions of database table row entries and try various models concurrently to achieve faster results for your machine learning experiments.

The following sections will give an overview of the major technologies and platforms that implement machine learning and artificial intelligence within the Microsoft Azure ecosystem.

ML.NET

ML.NET is an open-source cross-platform framework for **.NET** developers. By using ML.NET, you and your team can leverage the skills, libraries and frameworks that are already present within the .NET ecosystem. You can create web applications using ASP.NET, mobile applications using **Xamarin**, Desktop applications using WPF and even the **Internet of Things** (**IoT**) using Windows IoT. You can also extend your machine learning model creation capability by using TensorFlow and **ONNX**. ML.NET provides out-of-the-box support for algorithms that enable sentiment analysis, product recommendation, object detection, sales forecasting and many more real-world business scenarios.

For tasks such as regression and classification, both training and consumption can be carried out using ML.NET. Other than this, it also supports core data types, extensible pipelines, data structures, tooling support, advanced performance mathematics and more.

ML.NET can be installed from nuget.org. nuget.org is a public repository of downloadable `.NET` packages, libraries and frameworks that you can easily add to your .NET project.

Automated machine learning

Automated Machine Learning (**Automated ML, or AutoML**) enables data professionals to build models code free within Azure Synapse Analytics powered by Azure Machine Learning. AutoML is designed to automatically detect the best machine learning models for you by selecting the right algorithm and helping tune hyperparameters for forecasting, classification and regression. This is very useful if you do not have a data scientist on your team.

AutoML helps users (developers, analysts and even data scientists) to implement machine learning without a high barrier of entry relating to programming languages, libraries, frameworks and data science concepts. It allows companies to innovate, thanks to faster time-to-market by means of an iterative process, and to leverage data science best practices when running experiments.

AutoML runs are executed on Azure Synapse Analytics serverless Apache Spark pools and are tracked in the Azure Machine Learning service.

Cognitive services

Azure Cognitive Services is a suite of cloud-based, general-purpose, pre-trained models and APIs that can be consumed and extended for further training for specific use cases. If, for example, you want to create an object detection AI that understands what a banana is, you might need to feed in more data to help the AI understand that an image contains a banana. Consuming cognitive services is done via HTTP and is **platform-agnostic**, meaning you can use any programming language and operating system. There are five main categories of cognitive services: decision, vision, speech, search and language. You can readily integrate AI and ML with your mobile, web, desktop or even IoT applications using Cognitive Services.

The speech-to-text and speaker-recognition capabilities of the Speech Services API are good examples of cognitive services. These capabilities allow you to transform speech data to text, translate it to other languages and recognise the identity of the speaker without setting up a machine learning workspace that involves millions of datasets and a series of machine learning model experiments.

Using Cognitive Services is the best approach for those who want to easily integrate AI and machine learning in their applications with minimum data science knowledge. Microsoft offers very flexible pricing options where you only pay for what you use, and most of the services have free tiers for you to explore. There are currently pre-trained models available for text analytics (sentiment analysis) and anomaly detection, but more models will be available in the future.

You can learn more about Cognitive Services here.

Bot framework

Microsoft Bot Framework enables applications to build intelligent bots (often used for chatbots) to automate workflows. The Bot Framework is closely associated with Cognitive Services such as **Language Understanding Intelligence Service** (**LUIS**) and **QnA Maker**. QnA Maker is a **natural language processing** (**NLP**) service that accelerates the creation of conversation-based AI such as chatbots. With the Bot Framework, developers can easily create a conversational AI that learns through training from utterances and intents. This framework also allows developers to easily publish bots to channels such as Microsoft Teams, Cortana and Slack.

The Bot Framework is now widely adopted by large corporations such as banks and retail conglomerates for their **first-level support**. For example, the Bank of Beirut used the Bot Framework to create the Digi Bob chatbot, which assists users in applying for loans and availing themselves of other banking services. To learn more, read about this use case here.

Using the Bot Framework, developers can deploy intelligent enterprise-grade bots that can easily translate enquiries and messages (intents) from users and respond with meaningful actions. These actions can include querying a data source or orchestrating a command to a system. You can learn more about the Bot Framework here.

There are more machine learning tools and products within the Microsoft ecosystem, such as:

- SQL Server Machine Learning Services
- Microsoft Machine Learning Server
- Azure Data Science Virtual Machine
- Windows ML
- Apache Spark MLlib
- Azure Notebooks
- Azure Batch
- ML Services on HDInsight
- ML on Power BI
- Azure Machine Learning for Visual Studio Code
- Running your own ML frameworks to a Linux container or server image

This book is not able to cover all of the technologies mentioned here, so we instead focus on Azure Machine Learning. For more information on these services, visit this link.

Azure Machine Learning features and benefits

Azure Machine Learning offers a variety of features and lots of flexibility to users of various backgrounds and expertise. Azure Machine Learning can integrate into your existing data pipeline to perform tasks such as leveraging data from Azure Data Lake or Azure Synapse Analytics and serving the models directly to Power BI. You can also use Azure Databricks to further automate the hardware clusters where you are running your machine learning experiments.

Azure Machine Learning provides an end-to-end workspace to run your machine learning operations. With Azure Machine Learning, you can create experiments using AutoML, a visual interface or the **Software Development Kit** (**SDK**) in your machine learning notebook. You can also create a portable data model that can run in a container. This model can then be published to **Azure Container Instances** (**ACI**).

Software Development Kit (SDK)

Azure Machine Learning serves a Python SDK that fully supports mature frameworks such as **MXNet**, **TensorFlow**, **PyTorch** and **scikit-learn**. You can import the SDK into your experiments using Jupyter Notebooks, Azure Notebooks or even Visual Studio Code.

Designer

You can also use a visual interface (with minimal coding being required) within Azure Machine Learning called Designer to create and run experiments. The experience uses a low-code/no-code approach with drag-and-drop tools to create and connect entities. This is an intuitive way to connect data sources and create a machine learning model to train and serve.

AutoML

As discussed earlier, AutoML is a mechanism that suggests the best algorithm to use in your experiments. It is a baked-in feature of Azure Machine Learning. You can automate away time-intensive tasks such as data cleaning and choosing the right algorithms for your model. With AutoML, you can rapidly iterate over many combinations of algorithms and hyperparameters to find the best model for your desired outcome.

Flexible deployment targets

Microsoft and Azure do not limit your model deployment options. Even if you are managing your workspace and performing your analysis in the cloud, you are not locked into just deploying the outcome of your experiments to Azure. You have the option to deploy them on-premises and in edge environments by using containers.

Accelerated Machine Learning Operations (MLOps)

In a modern data warehouse, the combination of Azure Databricks and Azure Machine Learning can accelerate your machine learning operations. Azure Machine Learning can provide you with an end-to-end workspace where you can connect data from various sources with Azure Synapse Analytics, prepare and train data models, deploy them to consumers such as Power BI and then monitor and retrain them to improve accuracy:

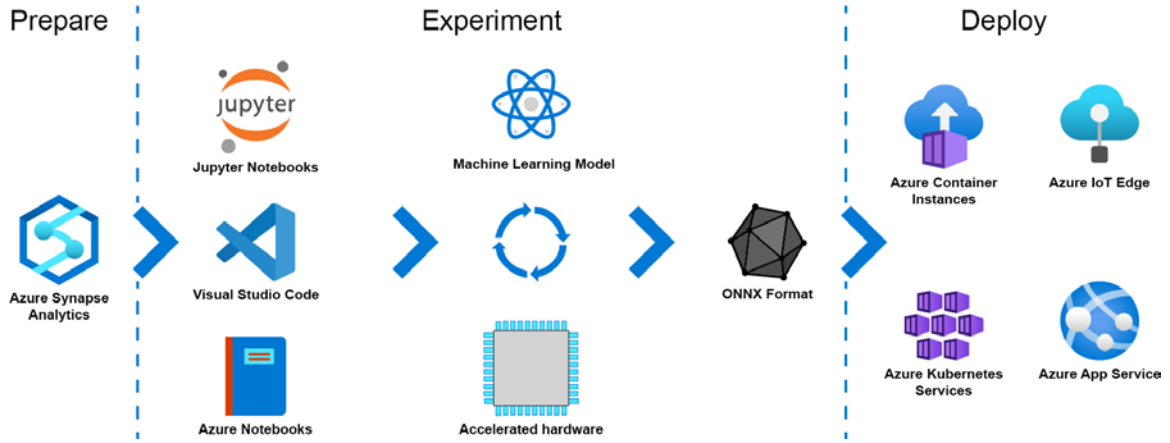


Figure 3.30: Preparation, experimentation and deployment in Azure Machine Learning

With Azure Machine Learning you can use Azure Databricks to prepare the data for your experiments. You can then use either Jupyter Notebooks or Visual Studio Code to author your experiments; alternatively, you can also use the built-in Azure Notebooks feature of Azure Machine Learning. You will then run your experiments to train and test your machine learning model by leveraging computers to run complex data science algorithms. Azure Machine Learning will create a machine learning model with the ONNX format, which is highly portable and can easily be deployed to a container such as Azure Container Instance. You also have the option to run it on **Azure Kubernetes Services** (**AKS**) or even on edge devices that support Docker.

> **Note**
>
> The ONNX model format enables data professionals working in Azure Synapse Analytics to bring a variety of models into Synapse securely without the need for data movement outside of the Azure Synapse Analytics security boundaries.

This book will not cover the use of Azure Databricks as the compute cluster of Azure Machine Learning, but there are advantages of having this combination. If you are already using Azure Databricks to derive real-time analytics on your modern data warehouse, you might also consider using it to run your machine learning experiments in Azure Machine Learning. You can read more about this here.

Azure Machine Learning and Azure Synapse Analytics

Azure Machine Learning and Azure Synapse Analytics can work hand in hand as they solve different problems. You use Azure Synapse Analytics for your modern data warehouse to make a unified data pipeline for your disparate data sources and eventually model and serve that data for a client consumer. Azure Machine Learning on the other hand, is used to create a machine learning model that can eventually be used for your applications to create meaningful inferences (assumptions).

A practical example of using Azure Machine Learning and Azure Synapse Analytics together could be in a physical retail store. Azure Synapse Analytics can aggregate multiple data sources such as beacon and CCTV data (unstructured), NoSQL databases and SQL databases, then query them all to serve meaningful reports for Power BI such as 'number of goods sold versus audience traffic'. Using Azure Machine Learning on the other hand, you can run through experiments from various data sources, including the ones generated by Azure Synapse Analytics, to create a recommendation engine for people coming to the store based on their previous activities and correlation among the rest of the customers.

It is also possible to combine Azure Synapse Analytics and Azure Machine Learning workspaces as a linked service. An Azure Machine Learning linked service can be created from within an Azure Synapse Analytics workspace and enables much simpler collaboration between the two technologies.

Quick-start guide (Azure Machine Learning)

In this quick-start guide on Azure Machine Learning, we will look at how to get started with the platform without any coding required. We will be using a sample called **Image Classification using DenseNet** using the Designer feature of Azure Machine Learning to create a pipeline for your machine learning model.

Prerequisites

In order to perform this activity, you need the following:

- An active Azure subscription
- Creating an Azure Machine Learning workspace

1. In the Azure portal, search for **`Machine Learning`** and click on **Create**:

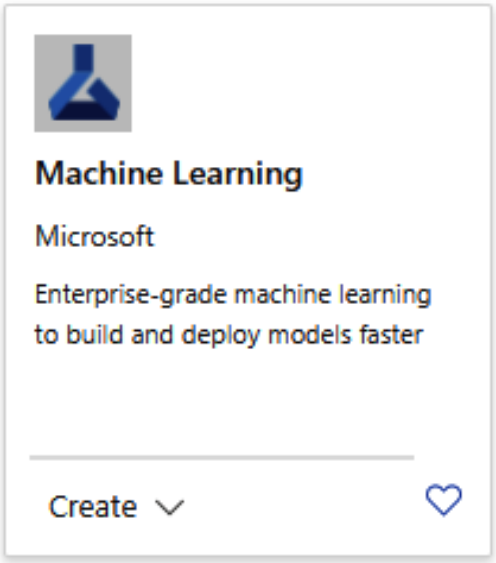


Figure 3.31: Creating an Azure Machine Learning workspace

2. Complete the subscription and workspace details.

> **Note**
>
> For this specific exercise, you can create a new instance or use an existing one for:

- Azure Storage
- Azure Key Vault
- Application Insights
- Azure Container Registry

3. Create the resource by clicking on **Review + create**:

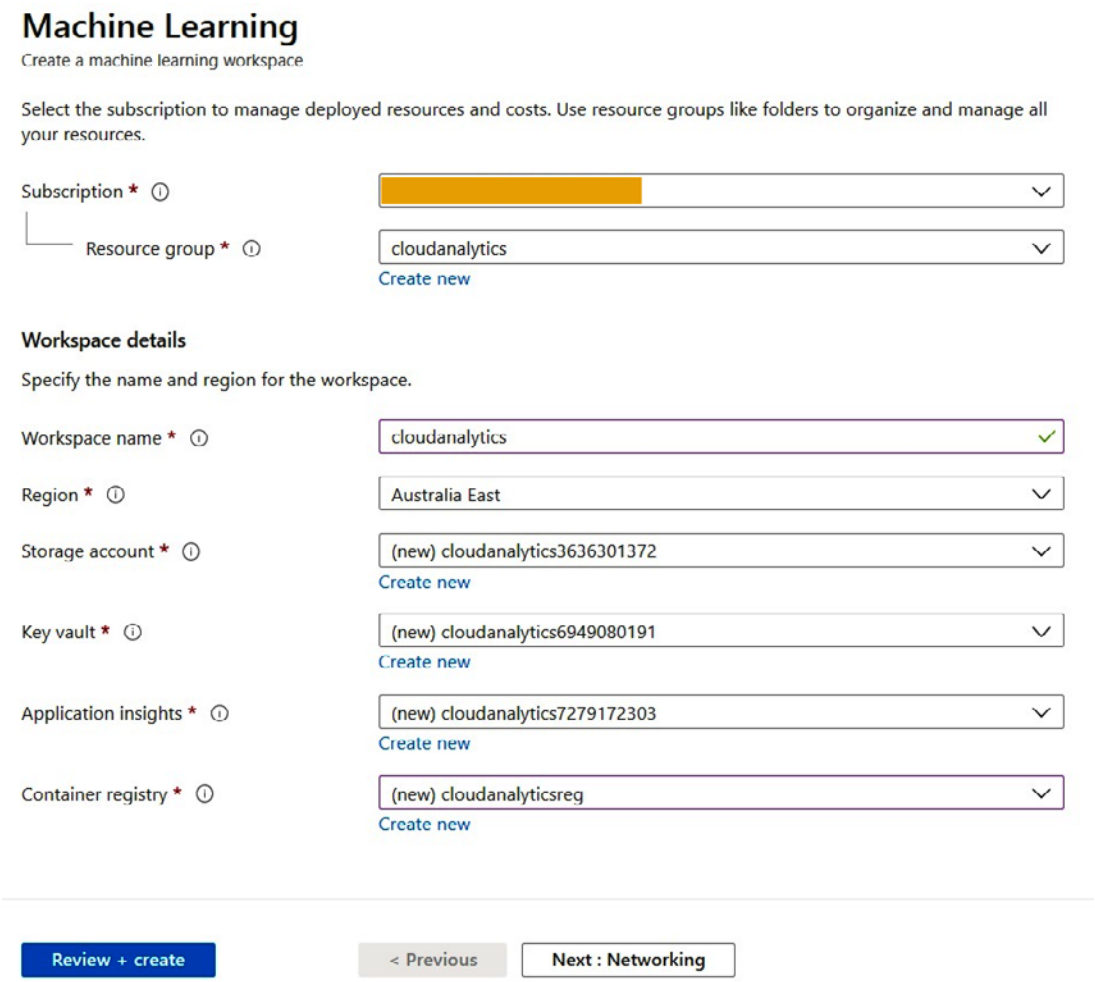


Figure 3.32: Azure Machine Learning workspace details

4. Once the Azure Machine Learning resource has been provisioned, go to its **Overview** section and click on **Launch studio**:

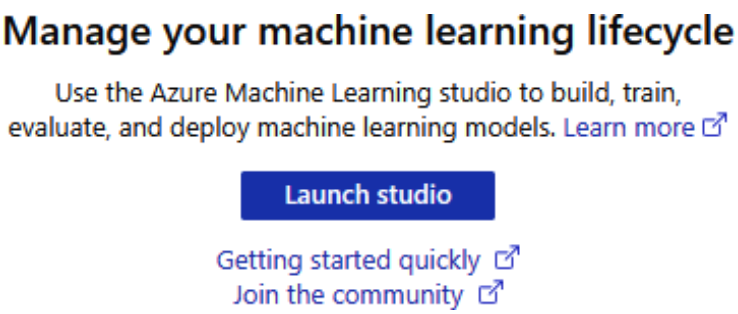


Figure 3.33: Managing the machine learning model through Launch studio

This will redirect you to the Azure Machine Learning studio, where we will create our machine learning model.

Creating a machine learning model using Designer

1. In the Azure Machine Learning studio, click on the **Designer** blade. We will be using a sample pipeline called **Image Classification using DenseNet**. Click on the icon of **Image Classification using DenseNet** to get started:

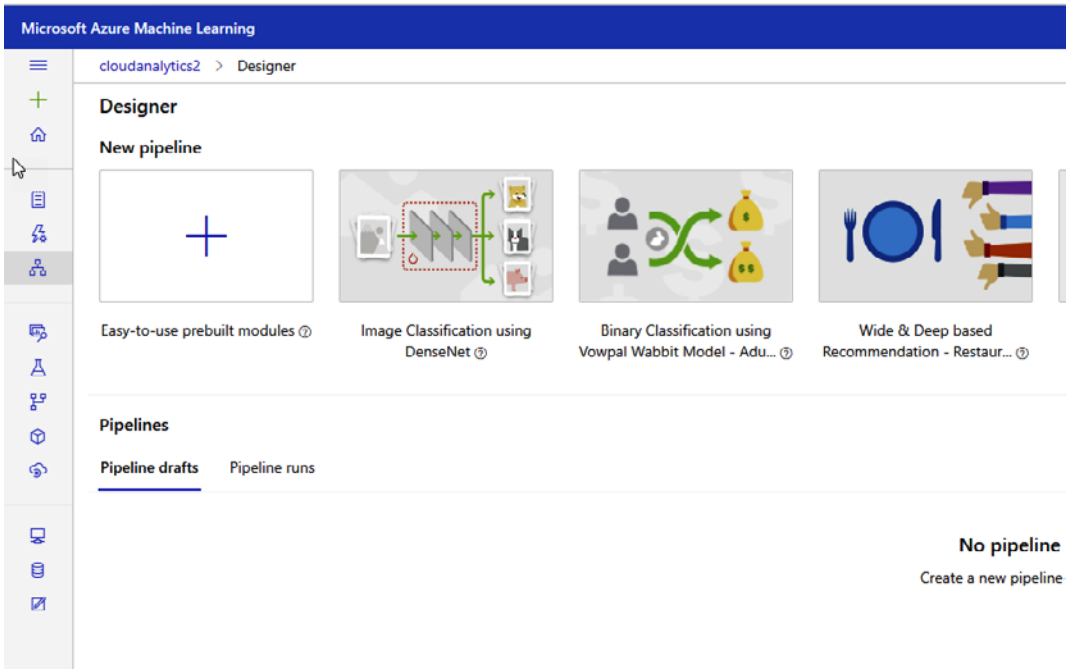


Figure 3.34: Creating a machine learning pipeline using Designer

This sample will create a series of steps in the pipeline as shown in *Figure* 3.35:

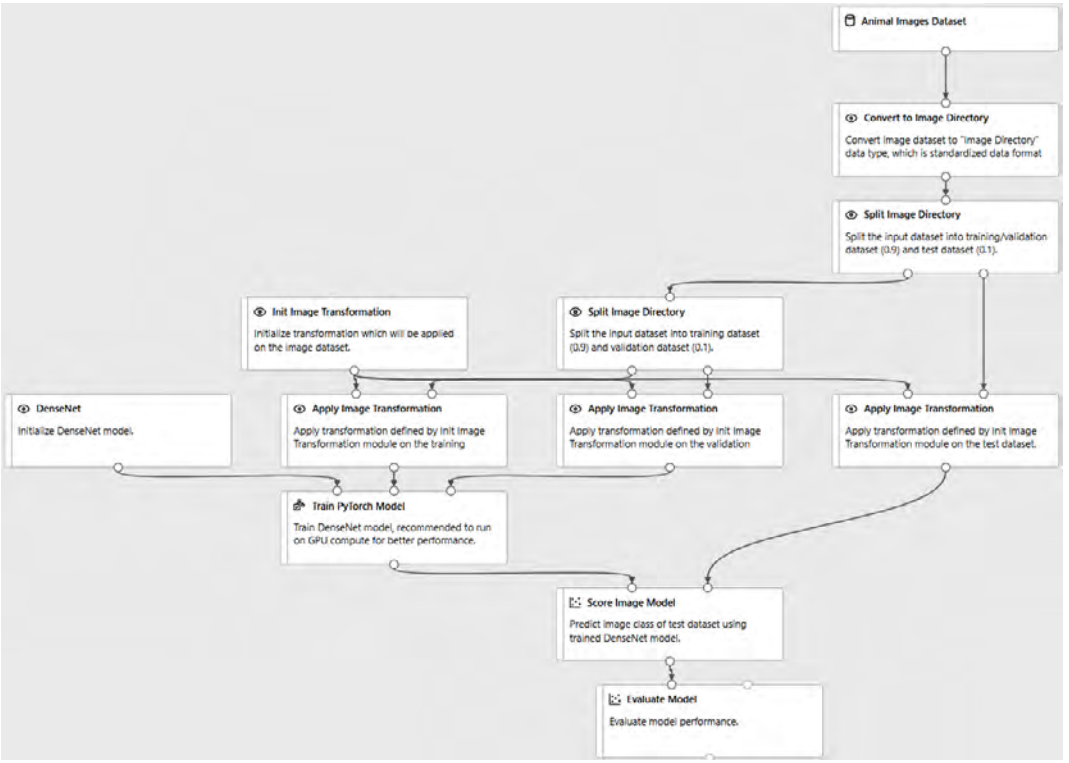


Figure 3.35: Image classification pipeline

In summary, this pipeline takes a set of images from the **Animal images Dataset** repository to use for training the model. It will then apply series of steps and techniques (refer to the flow chart) to train the model until it's ready to be scored.

Scoring the model means creating a measurement on how accurate this model can be if fed with a new image. The last part is the model evaluation, wherein the performance of this model is evaluated. This exercise is what sets your model apart from other models in terms of accuracy and reliability.

2. Click on the **Submit** button in the top-right portion of the page. It will ask you to specify which compute target you are going to use for this experiment, meaning which virtual machine you will use.

   For this exercise, create a new compute target by ticking the **Create new** radio button and assigning it a name. Click on **Save**. It will create that compute target:

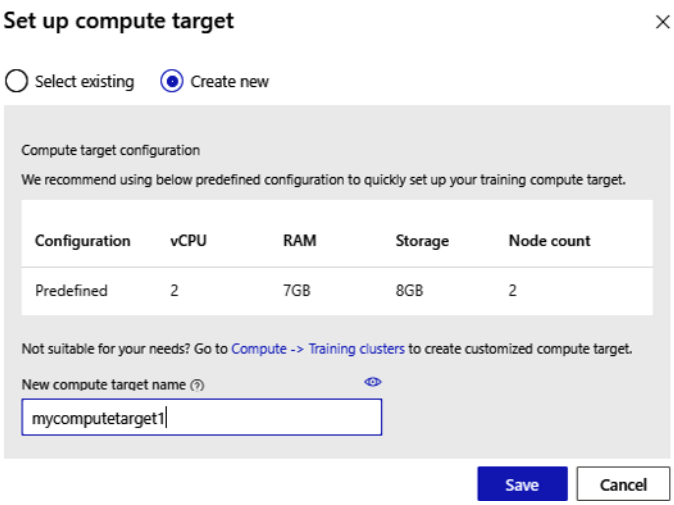


Figure 3.36: Setting a compute target

> **Note**
>
> In Azure Machine Learning, you can also add and configure existing virtual machines as a compute target.

Once the compute target has been created, you are now ready to run the experiment. Try to click the **Submit** button again to trigger the experiment.

3. You will be prompted to select or create an experiment name as shown in *Figure* 3.37. Put an experiment name in the **Create new** option and once you're ready, click on the **Submit** button:

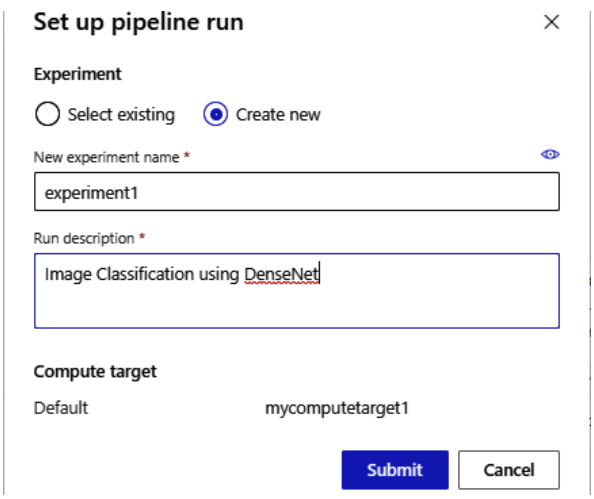


Figure 3.37: Creating a new experiment

It will now run your experiment, and this will take some time. You can see the progress of the experiment visually depending on the status of each step: **queue**, **not started** or **completed**.

Once the experiment is complete, all boxes should be highlighted in green, as shown here:

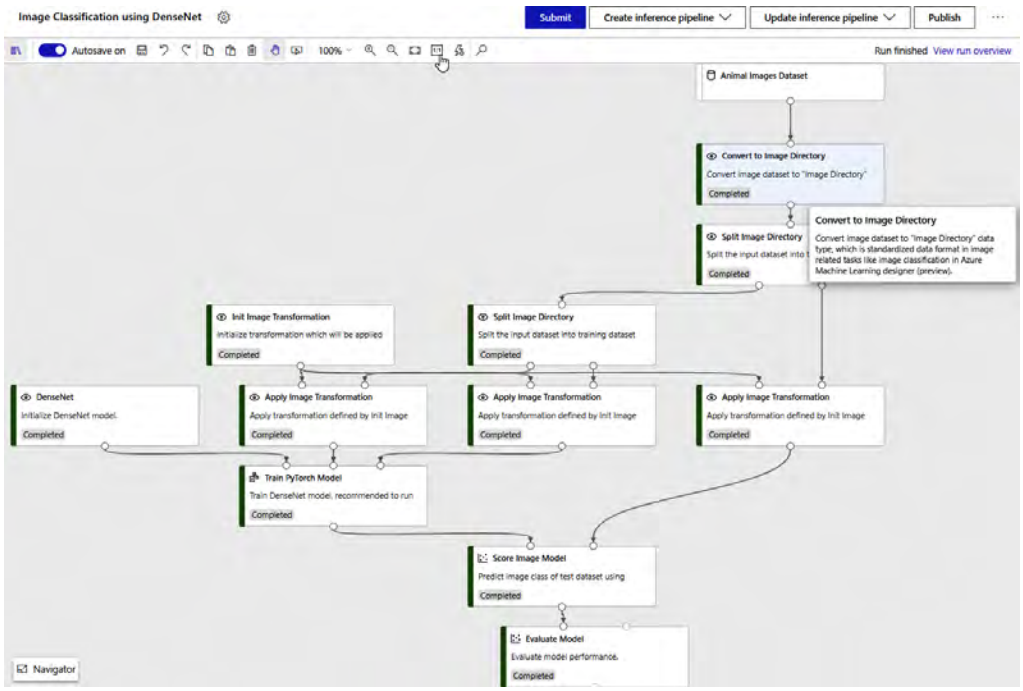


Figure 3.38: Animal imaging experiment pipeline

4. Click on the **Scored Image Model** step. Right-click on it and choose the **Visualise** option. Then click on **Scored dataset**:

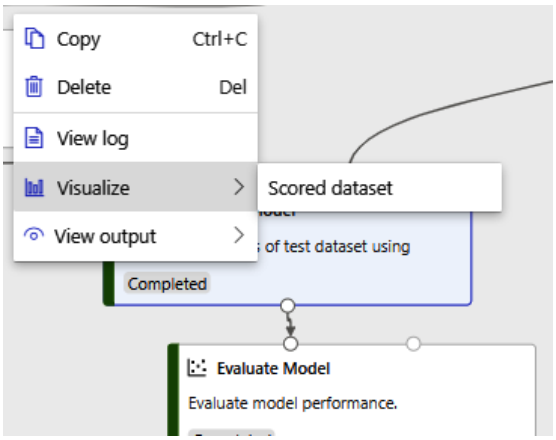
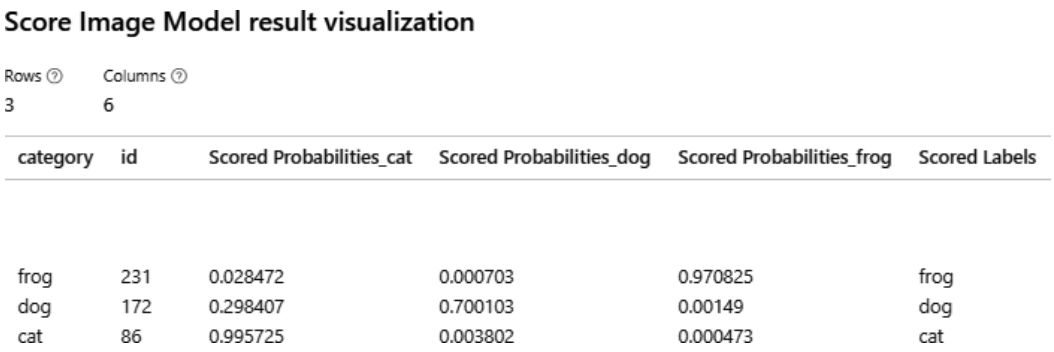


Figure 3.39: Visualising the scored model

The result visualisation shows the probability of an image being a cat, dog or frog. As you can see, this model is better at detecting cats, where it scores 99.57% accuracy, compared with recognising dogs, where it scores only 70.01%:


Score Image Model result visualization

Rows 3 Columns 6

category	id	Scored Probabilities_cat	Scored Probabilities_dog	Scored Probabilities_frog	Scored Labels
frog	231	0.028472	0.000703	0.970825	frog
dog	172	0.298407	0.700103	0.00149	dog
cat	86	0.995725	0.003802	0.000473	cat

Figure 3.40: Scored model visualisation result

This exercise gave you a glimpse of how easy it is to get started on Azure Machine Learning. You may choose to deploy this model to ACI or AKS to get a real-time inference on an image that you want to test, but that's beyond the scope of this book. Do not forget to clean up or delete your resources if you won't be needing them anymore.

Summary

In this chapter, we discussed how Azure Synapse Analytics can serve a unified data model to Power BI. We then explored the use of Power BI Desktop to create reports. Power BI enables you to create rich and meaningful graphs that derive business insights. We then published a report in order to collaborate on it across different media.

We have also learned that there are many available tools and technologies to implement machine learning and artificial intelligence in Azure. We explored Azure Machine Learning, including its features and benefits, and performed machine learning on Azure to create a model that classifies animals. In the next chapter, we will explore different business use cases for Azure Synapse Analytics and related technologies.

4

Business use cases

In the previous chapters, you learnt about cloud-scale analytics and the services Microsoft Azure offers to empower businesses to discover insights. You were also introduced to the new features and functionalities added to the modern data warehouse. In this chapter, you will look at two real-world business use cases to demonstrate high-level solutions using Microsoft Azure. The aim of these use cases is to illustrate how real-time data can be analysed in Azure to derive meaningful insights and make business decisions.

The company names used here are fictitious, and for the implementation demos, we use sample datasets. However, the business use cases, the challenges and the actual problems are real. They illustrate the kinds of data problems you may encounter in your everyday life.

The first business case focuses on helping a company gain actionable insights from its data in near real-time. The second one talks about using data analytics on Azure to address operational issues and offer better services to passengers by improving the utilisation of the infrastructure of Egypt's busiest airport. For each of the use cases, we will first briefly discuss the problem and the challenges, and then look at a potential solution design and the Azure services that enable such solutions.

Use case 1: Real-time customer insights with Azure Synapse Analytics

Imagine a large multinational retail company that has stores in Australia, New Zealand and Japan. Let's call this company Coolies. The company sells consumer goods, electronics and personal care items through its bricks-and-mortar stores and digital online channels (mobile and web applications).

Coolies has appointed a new CEO who is passionate about data, and she has set up a new data analytics team and tasked it with creating and maintaining customer insights in near real-time to drive business decisions.

The problem

Coolies, like many other organisations, is trying to reinvent itself as a data-driven company. There are many indicators that show that this is the right strategy. However, for Coolies to succeed, it must solve many problems. Some of these problems are technical, while others are cultural and organisational.

Coolies' CEO wants the new data analytics team to help the business answer questions that guide operational decisions. There are many questions that the executive team is hoping to answer with data. However, to better articulate the scope of this project, the Coolies data team (in consultation with the CEO) agreed to focus on the question:

How can Coolies increase profits?

More specifically, the Coolies data team is given a 20-day challenge to run a pilot data analytics program to model ways to help Coolies increase its profit margin by 10%.

The team will start by focusing on two key areas:

- Understanding customers' purchasing behaviours to predict product sales. This includes optimising logistics operations, use of shelf space and reducing the waste of expired products.
- Using customer, sales and marketing data to optimise Coolies' spending on promotions and marketing to reach the right customers with the right promotion.

Excited to accomplish this task, the Coolies data team started with a workshop to refine the requirements and technical challenges. Coolies' current data practices, like most companies, are geared toward reporting what happened in the past. Current reports answer questions such as, "How many products were sold?" and "What was the revenue generated by product A?". However, this is very different from what Coolies is trying to achieve, which is to discover patterns to predict what products Coolies should sell and in what quantities they will sell them and doing so in near-real-time. For this to happen, the Coolies data team concluded that they need to tackle the following challenges.

Capturing and processing new data

Coolies interacts with its customers via multiple physical and digital channels. Each of these interactions generates data that can be valuable to Coolies. Think of all the transactions at the store checkout, customer responses to varying advertisements and aisle adjustments on the shop floor, as well as any loyalty card points that they might have earned. Each customer interaction generates a variety of data that Coolies needs to capture.

Furthermore, Coolies' online store has trackers and beacons to record customers' activities and their responses to products that are on promotion. Coolies' mobile application has similar functionality that enables Coolies to create a very good view of what customers like and dislike. Coolies uses a mix of Azure Application Insights and Splunk, as well as internal tools for recording users' click events, navigations, time spent on each page, what products are added to the shopping cart and how many orders are finalised. Combining this data with application logs, network monitoring events and what Coolies already knows about its customers provides a powerful tool for Coolies to predict usage trends and patterns.

Coolies not only needs to capture and store all this data from physical and digital customer interactions, but it also needs to clean, validate, prepare and aggregate all this data. This is a massive task that the team has not had experience with before. The team used to batch process data by loading historical data into the data warehouse to generate daily and weekly reports. This is quite challenging for the team, being an exciting, but also mammoth task.

Bringing all the data together

Usually, data comes in various formats and shapes. Purchase transactions, for instance, are highly structured tabular data that is easy to work with. Application logs, on the other hand, are semi-structured files that list millions of events and trace messages of what happens on the application's servers. Coolies needs to ingest both types of data: structured and unstructured.

To make things even more interesting, social media feeds are unstructured and are in a natural language that customers use to write on the web. These feeds can be very valuable to Coolies, as they inform the company of the actual feedback provided by their customer base. However, for data practitioners, it's hard to capture and organise these loose feeds of natural language posts in the same format and shape as the highly structured transaction data.

The data team at Coolies needs to face the challenges of not only capturing data in all of its varying forms (structured, semi-structured and unstructured), but also find a way to clean and store all this data in one place so that it can be joined and correlated with other forms of data from other sources to discover new insights.

Finding insights and patterns in data

Once all the data is captured, cleaned, validated and stored, the Coolies data team needs to start the challenging task of finding meaningful insights and patterns in the data. However, this can be a complicated problem to solve. When we are talking about gigabytes (or potentially terabytes) of varying datasets, how do you find these patterns? Where do you start?

Traditional reporting and statistical techniques will not scale and can't be used alone to tackle this challenge. Conventional forms of programming are not very useful, as the programmers and the data practitioners themselves do not know what they are looking for or how to find these insights.

Real-time discovery

Coolies needs to discover meaningful insights quickly and action any findings as soon as such insights are discovered. Data usually loses its value over time, and some data loses its value directly after ingestion. For instance, imagine a scenario where Coolies is running a major promotion on a particular product, say, an ABC-brand soft drink. This drink is selling very fast today in stores X, Y and Z. It would have no great value if Coolies discovered this trend tomorrow because, by that time, stores X, Y and Z will have empty shelves and customers will be disappointed that they did not get their desired product, meaning Coolies will have lost good opportunities to sell more.

As a result, Coolies is aiming to discover insights and trends in real-time or near real-time. Coolies defines near-real-time as being 5-10 seconds behind real-time. This gives just enough time for data pipelines to process and analyse live data as it is generated in Coolies.

Coolies' CEO has made it clear to the team that the organisation needs to know in near real-time how its operations are running and how the customers feel about its brand and services. She mentioned a scenario where Coolies had just decided to discontinue selling product A. After this decision, customers started having many discussions on social media platforms. Then, Coolies' CEO posed the following question: What if a large number of Coolies' customers were talking online about potentially switching to a competitor of Coolies just because of that decision? The answer to that question is a piece of information that is critical to Coolies and could be detrimental to its success. The CEO's point is that having the ability to find and analyse, and then act on, insights in real-time can be a massive competitive advantage for Coolies.

To summarise, Coolies is facing the following challenges:

- Coolies wants to capture and store large datasets from varying data sources with potentially high throughput. These data sources include transactional data stores, **Internet of Things** (**IoT**) sensors, Coolies' online stores and application log files.
- The company also wants to combine structured, semi-structured and unstructured data to create a single dataset through joining and correlating data from multiple sources.
- Coolies needs to handle the varying granularity and quality levels of the different data points. The team needs to clean, prepare, transform and join these multiple datasets.
- Coolies wants to draw meaningful insights and patterns from the data in near real-time.
- Finally, the company wants to scale this data discovery process to meet the demands of the business.

Design brainstorming

The following few sections will try to better articulate the requirements and come up with a technical solution that could satisfy these requirements.

Data ingestion

The first task for any data practitioner is to look for data, collect it, clean it, validate it and then start the exciting part of data discovery and exploration. For the current scenario, you need to define the data sources you need to pull data from. You also need to look at how you can load data from different sources to create a single dataset that can be explored and queried easily by data analysts. Some of the source systems that you need for this use case include:

- **Sales transactions:** The sales transactions can not only tell what and how many products were sold at a particular store, but they can also indicate what customers bought what products. This is because Coolies already has a loyalty programme where customers scan their loyalty card as part of the checkout procedure. Coolies has two different data stores for sales transactions: one data store for physical stores, and another one for Coolies' online stores.
- **Customer data:** Coolies has a **Customer Relationship Manager** (**CRM**) system that holds customer data. Customer data includes (among other things) profile details (name, age, address, phone number), last purchase date, favourites and other purchase history.
- **Loyalty programme data:** The loyalty programme data is stored in a different source system, and it helps Coolies link customer data with sales transactions.
- **Digital applications clickstreams and usage data:** This indicates how Coolies' customers are responding to the design and content of Coolies' applications.
- **Sensors and IoT data:** Some of Coolies' stores are equipped with digital sensors to understand customer behaviour on the store floor. Some stores have IoT sensors installed to count how many customers walk past each aisle and at what time. There are also sensors to measure temperature and humidity in Coolies' stores. Coolies uses these sensors to ensure that fresh products such as milk are kept in the right conditions. Moreover, Coolies also has sensors for counting customer headcounts in near real-time. This helps Coolies to deploy more staff during peak hours/rush hours to ensure faster service, so customers do not have to wait in long queues.

- **Other datasets:** To enrich Coolies' data and give it another dimension, the Coolies data team is considering pulling other data, such as weather data, **Geographical Information Systems** (**GIS**)/map data, suburb and city profile data and other similar data from public datasets. These datasets can enrich Coolies' data and add greater context to the trends and patterns in customer behaviour and sales figures. Take the weather, for instance. Coolies might find that the sales of certain products might correlate with certain weather conditions – for example, an increase in the sale of ice cream during the summer season. Similarly, a city profile with certain age group and average income attributes may have a strong correlation with the sales figures of certain products. For instance, suburbs where the average age is 25 might have higher sales of hair-styling products, while stores in suburbs with an average age of 45 might sell far fewer of the same products.

Data storage

As explained previously, Coolies needs to ingest data from a variety of sources. It also estimates the size of its current datasets to be over 400 TB with an average growth of 10-15 GB per day. The formats of these datasets are also quite different. Some of them are highly structured, while others are fully unstructured. One common thing between all these different datasets is that they are all growing rapidly, and they arrive at a high throughput rate. To serve Coolies' needs, we need to have a data store that:

- Is scalable and elastic to grow with the demands of Coolies' data team
- Is a secure and controlled platform to ensure that Coolies' assets and intellectual property are well protected
- Is compatible with other existing systems and tools
- Is reasonably priced
- Is highly available
- Supports high-throughput operations and parallel processing

Data science

Once all the data is collected and stored in a central data store, the Coolies data team will need to have a platform for:

- Cleaning, transforming and exploring the datasets
- Collaborating with other business and technical stakeholders to discover patterns and trends
- Integrating with artificial intelligence frameworks and runtimes to apply machine learning algorithms on the datasets and uncover any patterns
- Training and operationalising the new machine learning models that might come out of the work of the previous integration
- Enabling the team to schedule, run and monitor data pipelines to enable data transformation, cleaning and integration

Dashboards and reports

Developing data analytics solutions can be seen as a continuous conversation between the data practitioners, business stakeholders and the data itself. It requires continuous refinement and hypothesis testing. Therefore, it is imperative for Coolies' data team to develop and maintain interactive reports and dashboards to communicate their work and the results of their data discovery processes to the business.

The first step in creating such reports and dashboards is coming up with a consistent and common data model that facilitates a common understanding across the organisation.

The solution

The Coolies data team decided to use Microsoft Azure to implement Coolies' analytics solution. Among other things, the Coolies team listed Azure's scalability, compliance and regional availability in Coolies' regions (Australia and Japan) as the main factors in making this decision. The team also articulated the reasons why each of the chosen Azure services is fit for purpose, as we will see in the next few sections.

Coolies used the refined requirements and brainstorming ideas (from the previous sections) to come up with a design for the solution architecture. The Coolies data team has come up with the following architecture (as shown in *Figure 4.1*):

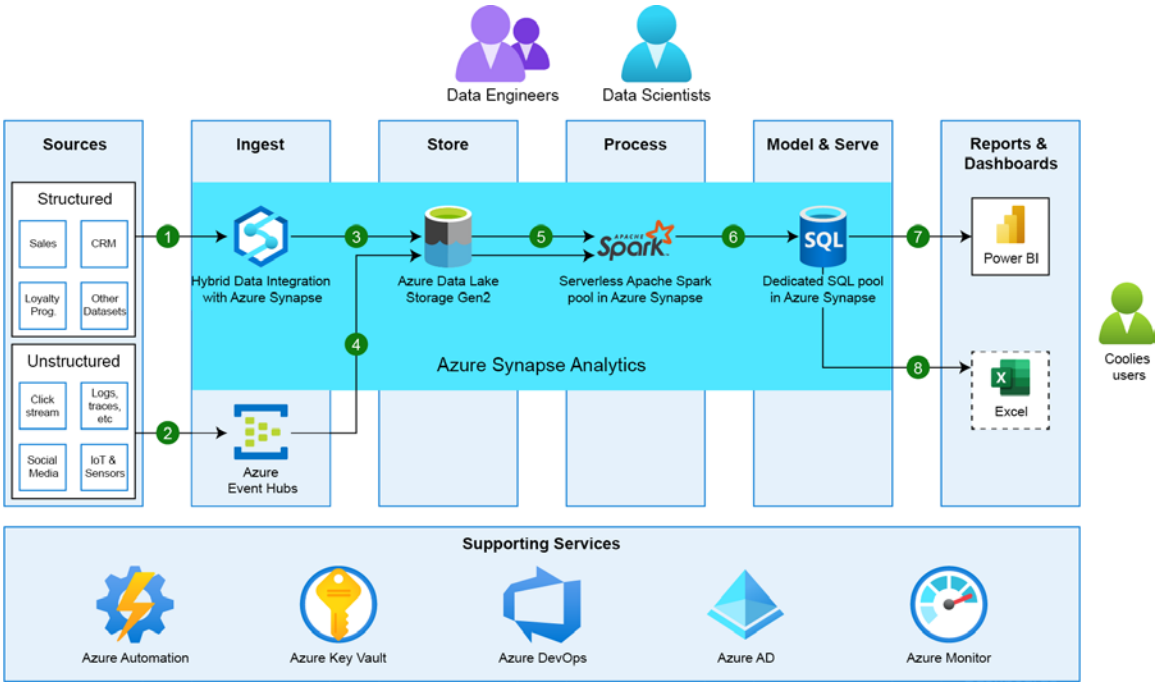


Figure 4.1: Coolies' solution architecture

Data flow

The design in *Figure 4.1* shows the solution architecture and the data flow between the individual components. Here, we explain each of the workflow segments as marked (numbered) in the preceding diagram:

1. Coolies needs to ingest many varying datasets. Some of these contain structured data and others have unstructured data. Coolies will use **Azure Synapse Analytics' capability** for hybrid data integration. This enables Coolies to ingest structured data using periodic (five-minute) batch activities and push this data to **Azure Data Lake Storage Gen2**.

2. For the unstructured data, Coolies is going to use **Azure Event Hub** to capture this data in near-real-time and push it to **Azure Data Lake Storage Gen2**. This makes all new data available for processing by Coolies' analytics solution and enables Coolies to trigger any action on data in near-real-time (every one minute). The unstructured data includes data coming from clickstream analytics (reports on user behaviour on Coolies' digital channels), social media feeds (from Twitter, Facebook, and so on), logs and trace information from Coolies' servers and any data coming from IoT sensors.

3. All ingested data will end up in **Azure Data Lake Storage Gen2**, which will serve a central hub for all data across the organisation.
4. Once the unstructured data is ingested into **Azure Event Hub**, it will be pushed to Azure Data Lake for permanent storage. This is done using the Azure Event Hub data capture feature, which allows us to store streaming data in Azure Data Lake Storage Gen2 quickly and easily.
5. Coolies' data will land in **Azure Data Lake Storage Gen2**. This data is coming from different sources, with various quality levels and different granularities. Thus, Coolies' data team will need to clean, prepare, validate and enrich these datasets. This work will be done using **the Spark pool of Azure Synapse Analytics**. Azure Synapse Analytics provides a managed Spark cluster so that Coolies' data engineers can easily connect to and explore the data in the Data Lake Storage without having to move the data to any other system.
6. While the Coolies team is cleaning and preparing data in **the Spark pool of Azure Synapse**, all the data (fresh/new and historical) needs to be combined in one standard model that is easy to query and serve to business users. For this, the Coolies team is planning to use the data warehouse (SQL pool) of Azure Synapse Analytics. This allows the Coolies data team to unite, model and prepare all their data for consumption by business users. This not only enables Coolies users' systems to run queries and answer questions about the newly arriving data, but also combine this new data with the historical data that is already in the data warehouse to come to a consensus about business performance and customer behaviours.
7. **Power BI** enables Coolies not only to publish reports and dashboards for Coolies users, but it also enables every user to be a data analyst for their domain using a self-service approach and by exploring the published data models. Coolies can use composite data models for large datasets, which is a feature of Power BI Premium.
8. Coolies invested in complex models using **Microsoft Excel**. Some of Coolies' data analysts would like to use **Microsoft Excel** to access data from the Azure Synapse Analytics. This is supported out of the box in both Microsoft Excel and Azure Synapse Analytics.

Azure services

The following sections will elaborate on each of the Azure services that are shown in the solution design of *Figure 4.1*. For each service, it will first explain why this component is needed, then why Azure services are fit for purpose for Coolies, and then finally show a brief practical example of the core part of the implementation.

Azure Data Lake Storage Gen2

Role in the design

Azure Data Lake Storage Gen2 works as Coolies' central data store. This enables Coolies to bring massive amounts of data from varying sources together. Moreover, the type and format of Coolies' datasets vary significantly (structured, semi-structured and unstructured), which requires a more capable data store than mere tabular storage, which is where Azure Data Lake Storage Gen2 is needed. Azure Data Lake Storage Gen2 can store schema-less data as blobs and can handle varying formats (for instance, text files, images, videos, social media feeds and zipped files). The ability to handle schema-less data formats makes it easy for Coolies to ingest data in its raw format, which is essential for advanced analytics as analysis can be done on the original data without any bias from any data aggregation.

Furthermore, the Coolies team needs elastic storage for a sandbox environment to explore and transform the data. Azure Data Lake Storage Gen2 can be used for this, too.

Why Azure Data Lake Storage Gen2?

- Coolies uses Azure Active Directory for access management. The Azure Data Lake Storage Gen2 offers native and out-of-the-box integration with Azure Active Directory to manage access to data using Azure Active Directory as the enterprise control mechanism. This reduces the design complexity and improves security and compliance.
- Studies suggest that Azure Synapse Analytics is much faster and cheaper that other cloud providers.
- Besides being low in cost, Azure Data Lake Storage Gen2 imposes no limits on how much data can be stored. This means that the Coolies team can start small with very minimal costs and scale as needed without worrying about hitting any maximum limits.
- Azure Data Lake Storage Gen2 integrates natively with Azure Synapse Analytics, Power BI and many other Microsoft Azure services. This makes a compelling case for the Coolies team since they are already using Power BI.
- Besides integrating with Azure Active Directory, Azure Data Lake Storage Gen2 offers the security features that Coolies' security team demands. This includes data encryption – at rest and in transit – single sign-on, multi-factor authentication, fine-grained access control for users and groups and full auditing compliance by monitoring every access and configuration change on the data lake.

Sample implementation

When ingesting data into Azure Data Lake Storage Gen2, it is considered a good practice to use namespaces and containers to organise the data in the data lake. This not only makes finding data easier, but it also helps with access control management. *Figure 4.2* shows an example of simple data lake zoning, where a data lake is divided into four zones: **Landing Zone** (ingestion), **Staging Zone**, **Secure Zone** and **Analytics Sandbox**:

- **Landing Zone:** This is where all data (except sensitive data) coming to the data lake will land before being processed, cleaned, aggregated, and so on.
- **Staging Zone:** This is where data will be cleaned/staged before being made ready for consumption.
- **Analytics Sandbox:** This zone is used by data scientists and data engineers as their sandbox for storing data while they are processing and exploring.
- **Secure Zone:** This is where highly sensitive data is stored and processed. Splitting the secure zone from the other zones can enhance access control management. This zone includes sensitive data such as merger and acquisitions data, financial data and other customer data that might be hard to mask, such as customer gender, age and ethnicity data, if known:

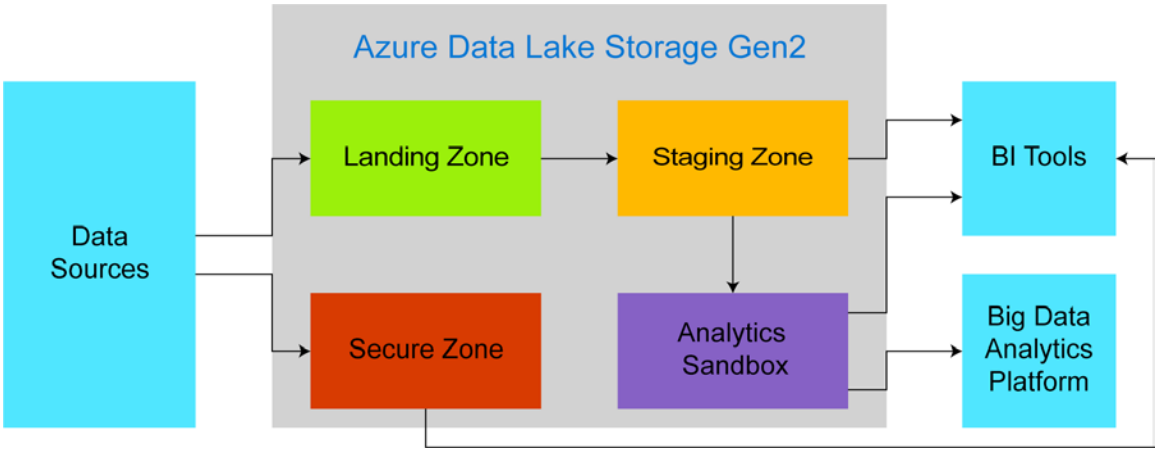


Figure 4.2: Example of data lake storage zoning

Azure Synapse Analytics

Role in the design

The productivity and accelerated time-to-insight that Azure Synapse can bring Coolies is incredible. Coolies' team understands the value of the unified experience that brings together data ingestion, big data analytics and data warehousing – at cloud scale. This unification of these divided worlds can enable the team to massively reduce their project development time. Instead of stitching a multitude of services together or creating a siloed point solution for just one of these three core areas, Azure Synapse unifies data, data tasks and data teams with a limitless analytics service that does it all.

The Coolies team considered adopting other comparable services and platforms. However, upon studying the features of Azure Synapse Analytics, the Coolies team started to appreciate the uniqueness of Azure Synapse Analytics as it is the only platform in the market that unifies big data analytics, data warehouse and data ingestion and visualisation all in one platform.

Why Azure Synapse Analytics?

- Azure Synapse Analytics is a limitless scale analytics platform that unifies big data analytics and data warehousing. It is a fully managed service and dynamically scalable. This makes it very attractive for Coolies because it means less infrastructure to manage, less overhead and a scalable service that can grow with the business.
- Studies suggest that Azure Synapse Analytics is much faster and cheaper than other cloud providers.
- Azure Synapse Analytics offers an Apache Spark runtime out of the box and seamlessly unifies it with its SQL engine. This is very important to Coolies because not only does this mean simpler queries because the runtimes are integrated, but it also means better collaboration as everyone (data analysts, data scientists, data engineers, and so on) can query the data with plain old SQL.
- Coolies' security team has a clear requirement that stipulates the need to protect its data assets. The Coolies security team has demanded that the data warehouse must not be publicly accessible on the web. This is natively supported by Azure Synapse Analytics via Azure Virtual Network integration, where Azure Synapse Analytics is deployed as part of the Coolies network (a virtual private network).
- Coolies also liked the other security features in Azure Synapse Analytics. These include integration with Azure Active Directory, activity auditing, native row- and column-based security, ExpressRoute integration and the out-of-the-box threat detection and data encryption capabilities.

- Some parts of Coolies use other BI tools such as Power BI, Tableau and Qlik, and they wanted to make sure that the new data warehouse supports this integration. Azure Synapse Analytics is compatible with many BI tools, including Coolies' existing BI tools.
- Azure Synapse Analytics can also spin up Apache Spark on demand. This can be remarkably useful for Coolies' data team, as it enables them to use the same open source tooling to work with their data inside Azure Synapse Analytics. This facilitates better productivity, as Spark clusters support multiple languages and frameworks out of the box (Python, R, Scala, and so on). Thus, Coolies' team members can be productive and happy using the tools they are most comfortable with.
- Developing, deploying and managing a Data Warehouse can be a very complex exercise. Azure Synapse Analytics shines in this area, as it draws on Microsoft's wealth of experience as a development company. Part of the reason why Coolies is so interested in Azure Synapse Analytics is its streamlined workload management and excellent developer productivity. Azure Synapse Analytics is the only cloud data warehouse that offers native SSMS and SSDT support, including Visual Studio projects for code and schema management, which are vital to ensure a streamlined development life cycle and reduce the total cost of ownership.
- Moreover, Azure Synapse Analytics provides a rich collaboration environment where multiple Coolies stakeholders can work on the same notebook at the same time. This significantly improves productivity and fosters greater innovation by bringing the knowledge of all team members together.
- Diversity is good for innovation, and so Azure Synapse Analytics supports multiple programming languages and frameworks. Data scientists and data engineers at Coolies can use R, Python, SQL, Scala, Java and C# to write code in Notebooks thanks to Azure Synapse support for Apache Spark. This is especially important for Coolies, where it has been hard to recruit and retain talent in this area.

Azure Synapse Hybrid Integration (Pipelines)

Role in the design: Coolies, like most other enterprises, has many data sources. Some of these data sources reside on-premises, while some others are on the cloud. As discussed previously, Coolies needs to bring all this data together in one place to be able to combine, correlate, model and transform these datasets to discover trends and insights. This requires building and managing many data connectors to move the data from Coolies' source systems to the central data store (the data lake). This is exactly where Azure Synapse shines because it is a managed service that is aimed at simplifying data integration for users of all skill levels.

Why use Azure Synapse Hybrid Integration?

1. Azure Synapse Pipelines offers more than 90 pre-built data connectors. This enables Coolies to connect source systems to this new modern data warehouse quickly and easily at no extra cost. These data connectors are built by Microsoft; they offer efficient and resilient integration, and they take advantage of the Microsoft Azure network, which delivers up to 1.5 GB/s in throughput. This not only offers Coolies a fast time to market, but also provides a platform for orchestrating data movement with minimal overhead.
2. Besides all the pre-built data connectors, Azure Synapse provides a visual interface that empowers everybody to develop comprehensive data movement pipelines with little or no code. Moreover, Azure Synapse's Visual Editor offers the ability to integrate with Git source control repositories to improve flexibility and maintainability. This resonates well with the Coolies team as it improves their productivity and development pace, while at the same time reducing overhead. Using this feature allows the Coolies team to take advantage of Azure Synapse's powerful visual data transformation capabilities and data wrangling in the visual portal, while keeping all the work version controlled.
3. Azure Synapse is a fully managed tool that enables the Coolies team to start small with little or no investment and to scale as needed. This also means that there is no infrastructure to manage, and the Coolies team pays only for what they use.
4. Besides other certifications, Azure Synapse Pipelines is ISO/IEC 27001 and 27018 certified and is available in 25 countries/regions, including Australia and Japan, which is where Coolies operates. This makes Azure Synapse a very compelling service for Coolies as it ticks all the boxes on their security and compliance checklist.
5. Azure Synapse provides the tools to build data pipelines that are resilient in the face of schema drift. This means that when the Coolies team builds pipelines to move data from source A to B, they can be assured that the pipelines will still be functional, even if the scheme of the data from source A has changed. This significantly improves the reliability and resilience of Coolies' data pipelines.
6. Finally, using Azure Synapse provides Coolies with a single control plane to manage all activities of data movement and processing.

Sample implementation

Here is an example of how Coolies configures their Azure Synapse Pipelines to pull data from their sales transactional database (which sits on an Azure SQL server) to Azure Data Lake Storage Gen2:

1. Azure Synapse provides native integration with Azure Data Lake Storage Gen2. Coolies can connect to Azure Data Lake Storage Gen2 by configuring a linked service in Azure Synapse Pipelines as follows:

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

   It's worth mentioning that this example contains placeholders for the main configuration values, such as the Azure Storage account **`name`** and **`accountKey`** fields, and the name of the integration runtime.

2. After creating a linked service in Azure Synapse Pipelines, we need to have an Azure dataset to reference this linked service. This can be done as follows:

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

The preceding code snippet makes use of the Azure Data Lake Storage Gen2 linked service to create a dataset. This dataset will create comma-separated values (**CSV**) files and store them as compressed files (**gzip**).

3. Configure the Azure SQL database as a linked service:

```
{
"name": "CooliesSalesAzureSqlDbLS",
"properties": {
"type": "AzureSqlDatabase",
"typeProperties": {
"connectionString": {
"type": "SecureString",
Azure Services | 163
"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"
}
}
}
```

   Again, the preceding code snippet has placeholders for the following parameters: the Azure SQL Server name, the SQL database name, the SQL server username and password and the name of the integration runtime. Also, the example is for demo purposes only; passwords should always be kept out of the code and should be stored in Azure Key Vault to ensure security.

4. Similar to *step 2*, you need to configure a dataset in Azure Synapse Pipelines for Coolies' sales database. The following code snippet makes use of the Azure SQL Database linked service to create a dataset that references **`sales_table`** in Coolies' SQL database:

```
{
"name": "CooliesSalesDataset",
"properties":
{
"type": "AzureSqlTable",
"linkedServiceName": {
"referenceName": "CooliesSalesAzureSqlDbLS",
"type": "LinkedServiceReference"
},
"schema": [ {optional} ],
```

```
"typeProperties": {
"tableName": "sales_table"
}
}
}
```

5. The following code snippet configures the data movement activity from the sales SQL database to Azure Data Lake. This will create an activity in Azure Synapse Pipelines, and it references the two datasets we created in *step 2* and *step 4*. The activity sets the Azure SQL sales database as the source of the data movement, and Azure Data Lake Storage Gen2 as the destination of the data movement activity:

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

Power BI

Role in the design

The Coolies team needs to visualise and communicate its findings as well as some of the raw data to the business. This is critical to ensure engagement from the business stakeholders and to get feedback quickly and easily. Coolies also needs a platform to enable users to use self-service reports and dashboards and to empower Coolies users to explore data for themselves.

Power BI fills this role by enabling Coolies to visualise data using a variety of visuals and shapes, and by also enabling business and non-technical users to self-service any reporting and/or data needs.

Why Power BI?

Power BI is a business intelligence **Software-as-a-Service** (**SaaS**) offering that allows Coolies to transform its data into interactive visuals and dashboards quickly and easily. Coolies chooses Power BI not just because of its visualisation capabilities, but also to improve collaboration and self-service for data and reporting – core features of the Power BI service. The Coolies data team summarised their rationale for why Power BI is fit for this purpose, as follows:

- Power BI is a fully managed SaaS offering, which means less infrastructure for the Coolies team to manage.
- Power BI has simplified data visualisation and reporting and can empower any Coolies user to be a data analyst. The user experience in Power BI is a major advantage of the platform, as it enables users to explore data and interactive dashboards for themselves. Coolies hopes that this will reduce overhead and data requests for their data team, as well as improve collaboration and business user engagement.
- Power BI provides a desktop application that can be used by Coolies users to explore data, clean data and create visuals. This is very attractive to Coolies since the use of the Power BI desktop app is free and does not require a commercial licence. Moreover, the user experience on the Power BI desktop app and the Power BI cloud-based service is very similar, which means less training and easier knowledge transfer.

- Power BI has native integration with Azure Active Directory, which enables Coolies users to use their existing identities. This simplifies deployments, improves governance and enhances security. Moreover, Power BI has gone through many compliance certifications and is available in many regions around the world, including Australia, which is where Coolies is based.
- Coolies has well-defined branding and promotional guidelines. This means that all visualisations and dashboards must adhere to Coolies' colour styles, best practices, and so on. Coolies believes this improves branding and information comprehension because the reports are consistent and familiar to the user. Power BI supports this requirement by offering several features, such as customisable themes, customisable layouts and other custom visuals.
- Power BI offers out-of-the-box integration with Azure. This enables Coolies to start with any data preparation and transformation locally and scale to Azure when needed. Plus, Power BI has native integration with Azure AI services. This enables Coolies to infuse AI and machine learning capabilities to deliver value more quickly and easily, all from within Power BI.
- Power BI Composite Model enables Coolies to have rich reports that fetch data from multiple sources. Composite Models allows Coolies to seamlessly include data connections from more than one DirectQuery or import data connections in any combination. This simplifies data connections from reports to the data sources and it helps Coolies to build complex data models by combining multiple source systems and to join tables from different datasets. Moreover, using the Storage Mode feature of Composite Models in Power BI Premium can help improve performance and reduce the back-end load. This is because Power BI gives the author of a report the ability to specify which visuals require (or do not require) back-end data sources. Power BI then caches (stores) the visuals that do not require continuous updates from back-end data sources, which, in turn, improves performance and reduces the load on back-end systems.

Sample implementation

Here is an example of Coolies' dashboards. The following reports aim to communicate Coolies' performance in terms of product sales in their respective categories, the current stock of the top-selling products, sales figures per year and the regional distribution of Coolies' customers. The report was built using sample data that is provided by Microsoft:

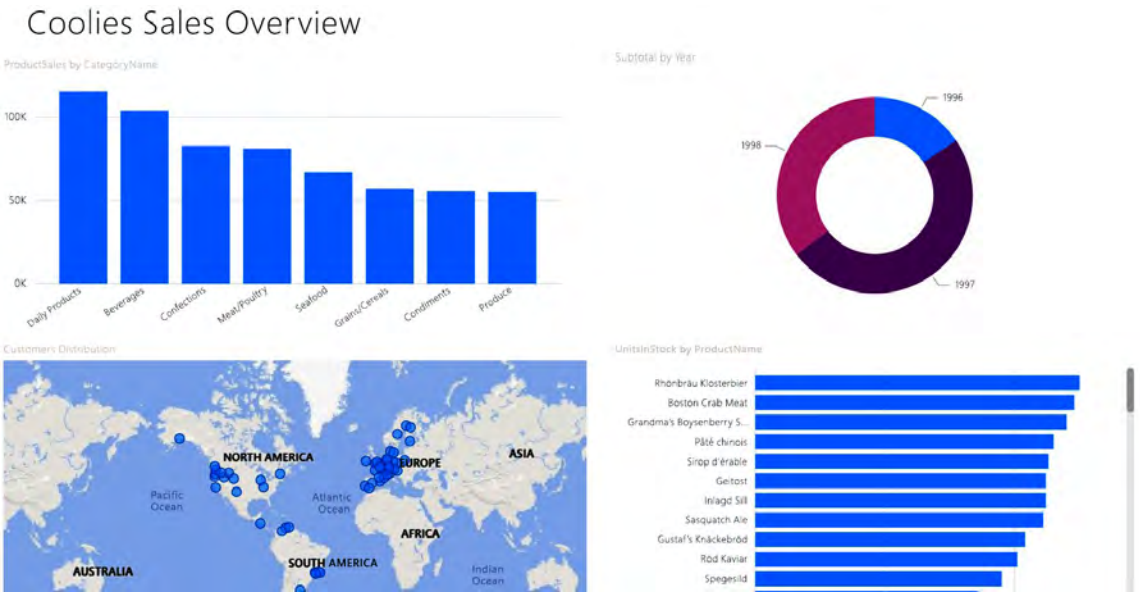


Figure 4.3: An example of Coolies' business performance dashboards

Azure supporting services

Besides all the Azure services that have been covered, Coolies needs a number of other services to support and enable this solution architecture. These services are shown in the **Supporting Services** section of the solution design back in *Figure 4.1*. In this section, we briefly describe each of these Azure services.

Azure Automation

Coolies has several database servers, test machines and other supporting servers. Coolies can use Azure Automation to automate the configuration and installation of updates on these machines. Azure Automation enables Coolies to have a consistent way of managing these servers and ensures security and compliance using serverless runbooks. This simplifies the operations and management overhead and gives the Coolies team time back to focus on the more important issue of discovering insights that add business value.

Azure Key Vault

Every organisation has many encryption keys, passwords, certificates, connection strings and other sensitive data that needs to be well secured. Coolies understands the need to protect this sensitive information and manage it in a secure and well-organised way. Azure Key Vault is designed to solve exactly this problem by safeguarding all such sensitive data in a central place where access can be securely managed, and keys can be organised. Azure Key Vault not only improves security controls, but also simplifies operational tasks such as certificate and key rotation.

Moreover, Coolies wants to use Azure Key Vault to remove the need for individuals and applications to have direct access to keys. Coolies can use Azure's Managed Service Identity so that users and applications can use keys and passwords without having to keep any local copies on their machines. This improves the overall security posture of Coolies, while at the same time streamlining secret management.

Azure DevOps

Azure DevOps provides Coolies with the tools, frameworks and services to run an agile practice to deliver its solution. Azure DevOps is a fully managed service that empowers the Coolies team to:

- Plan, track, discuss and monitor work items using Azure Boards. Coolies is already using agile practices, where currently physical walls are used to track work items, but Coolies realises that physical walls cannot scale for bigger teams and they are limited in their functionality. For instance, Coolies can use Azure Boards to link defects and work items to code changes to monitor and improve code quality.
- Continuously build, test and deploy code changes using Azure Pipelines. Azure Pipelines facilitate agile practices such as continuous integration and continuous delivery, which can significantly improve the quality and pace of delivery. Azure Pipelines also allows Coolies to automate any deployment steps that are needed to push code changes. This reduces overhead and improves confidence in the new deployments.
- Coolies needs a source control system to host its code and scripts. Coolies would like to use Azure Repos for this as it provides enterprise-grade support, an unlimited number of repositories and a collaborative environment for the development team to discuss and review code changes before merging.
- Azure Test Plans can help Coolies in the validation and verification of any code and data changes to give Coolies greater confidence in changes before merging them. Coolies can use Azure Test Plans for manual as well as exploratory testing, and since Azure Test Plans is part of Azure DevOps, Coolies can have great end-to-end traceability for stories, features and defects.

Azure Active Directory

Coolies uses Microsoft Office 365 for office collaboration, which means Coolies is already using Azure Active Directory. Coolies does not want to have multiple identity servers to manage and it understands that managing usernames and passwords is a massive task that is better left to a well-equipped team, such as the Active Directory team. Azure Active Directory has integration with many of the services that Coolies is aiming to use, such as SQL Server, Azure Synapse Analytics, Azure Data Lake Storage and Power BI. This makes it a no-brainer for Coolies to choose Azure Active Directory to enable simple and seamless login to all these services, while at the same time improving security controls over Coolies' data and applications.

Coolies can also benefit from Azure Active Directory's comprehensive identity protection, which includes threat detection and response. Overall, Coolies can reduce overhead remarkably and improve security by using Azure Active Directory.

Azure Monitor

Coolies recognises that the availability and performance of its data platform is of paramount importance to gain the confidence and trust of all stakeholders. To achieve that, Coolies not only needs to collect and store telemetry from all solution areas, but also analyse and action any data. This requires a dedicated service, since it is a major challenge to implement, and that is exactly what Azure Monitor is designed for.

Azure Monitor is a fully managed service that empowers Coolies to easily and quickly collect, analyse and action data from all components of the data platform (including Azure services, virtual machines, network performance and other sources). Azure Monitor offers two fundamental types of data, which are logs and metrics. Coolies can use metrics to learn about the state of its services at any given time, while the logs help the Coolies team analyse and visualise trace messages from the individual solution components. Azure Monitor also offers many charts and visualisations that can be used by Coolies to visualise the state of the system at a glance. Finally, Coolies can use Azure Monitor to trigger actions, such as alerts when certain conditions are met (for instance, when the number of errors goes beyond a certain threshold, the Coolies team can be notified by email or SMS).

Insights and actions

Using Microsoft Azure, the Coolies data team was able to design, build and deploy the solution quickly and easily. Within two weeks, the team found a number of key insights that can help Coolies increase its profit margins. Three of these insights are listed in here.

Reducing waste by 18%

Description: With initial modelling, the Coolies data team was able to reduce waste by 18%. Currently, the organisation loses close to USD 46M per year because of overstocking products with short shelf lives. This includes products such as bread and milk. The team combined historical sales data with other sources, such as weather data and school calendars, which allowed the team to predict the demand for these products with higher accuracy, leading to a significant reduction in waste.

Estimated business value: USD 8.28 million/year

Key data sources: Sales transactions (online and physical store), store data (store locations and stock over time), weather data, suburb profile data, school calendars and public holiday calendar.

Actions: Business stakeholders of Coolies were very impressed and wanted to deploy this quickly. Using Azure Synapse Analytics and Power BI, the Coolies data team was able to deploy the solution for use by store managers quickly. The result is that Coolies store managers now have an interactive dashboard that can predict sales accurately and give recommendations for the amount of stock to have for each product.

Data pipeline: Here is the simplified data pipeline for this initiative:

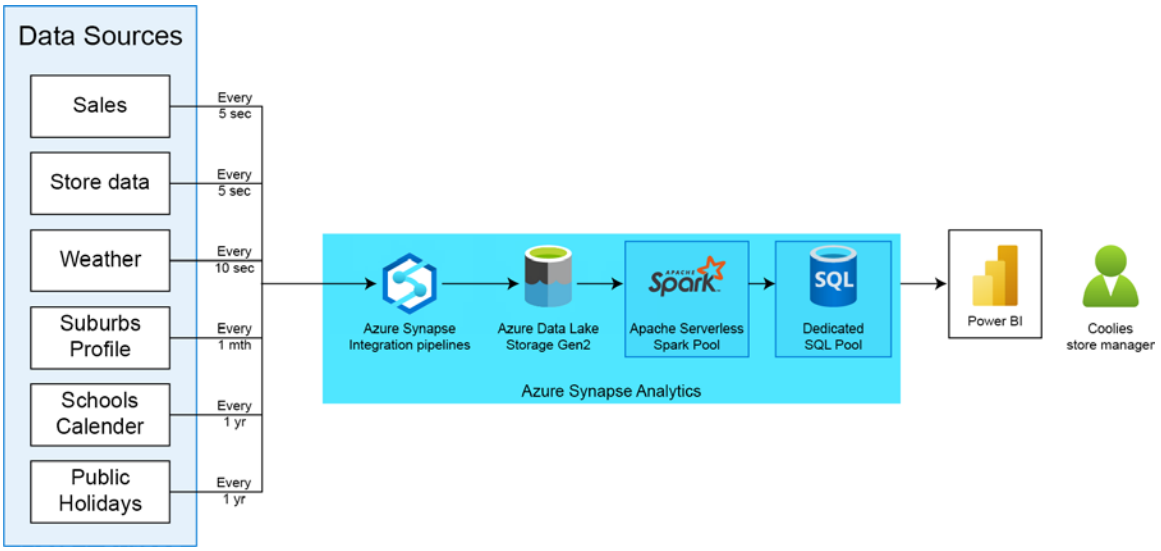


Figure 4.4: Data pipeline for initiative 1 (reducing waste)

Social media trends drive sales up by 14%

Description: The Coolies data team developed a hypothesis that social media trends can increase sales. The team performed initial data discovery to find that such a pattern does indeed exist. Interestingly, the team discovered that timing Coolies' marketing activities with social media trends can help Coolies improve sales by 14%.

One clear example of this was what the team found in the data of Australia's summer season of 2019. In January 2019, there was a huge social media trend related to healthy eating. This was not organised by Coolies. There were more than 4.5 million Australians who tweeted, liked, shared or commented on Twitter and Facebook posts using the **`#BeHealthy`** hashtag. Coincidentally, Coolies had a marketing campaign regarding fruit salad products. The team found that this marketing campaign was exceptionally successful and increased sales by over 25%, which is much higher than the average expected increase of 5-10%.

Estimated business value: USD 15.4 million/year (based on a 14% increase for subject products)

Key data sources: Social media feeds (Twitter, Facebook and Instagram), sales transactions (online and physical store), store data (store locations and stock over time) and marketing campaign data.

Actions: After discussing the results with the Coolies marketing team, it was agreed that Coolies could reproduce the success of their January 2019 promotion by monitoring and aligning their promotions with social media trends. The Coolies data team implemented the data pipeline, as shown in *Figure* 4.5, and deployed it as an interactive and real-time dashboard to inform both the Coolies marketing team and store managers.

Data pipeline: Here is the simplified data pipeline for this initiative:

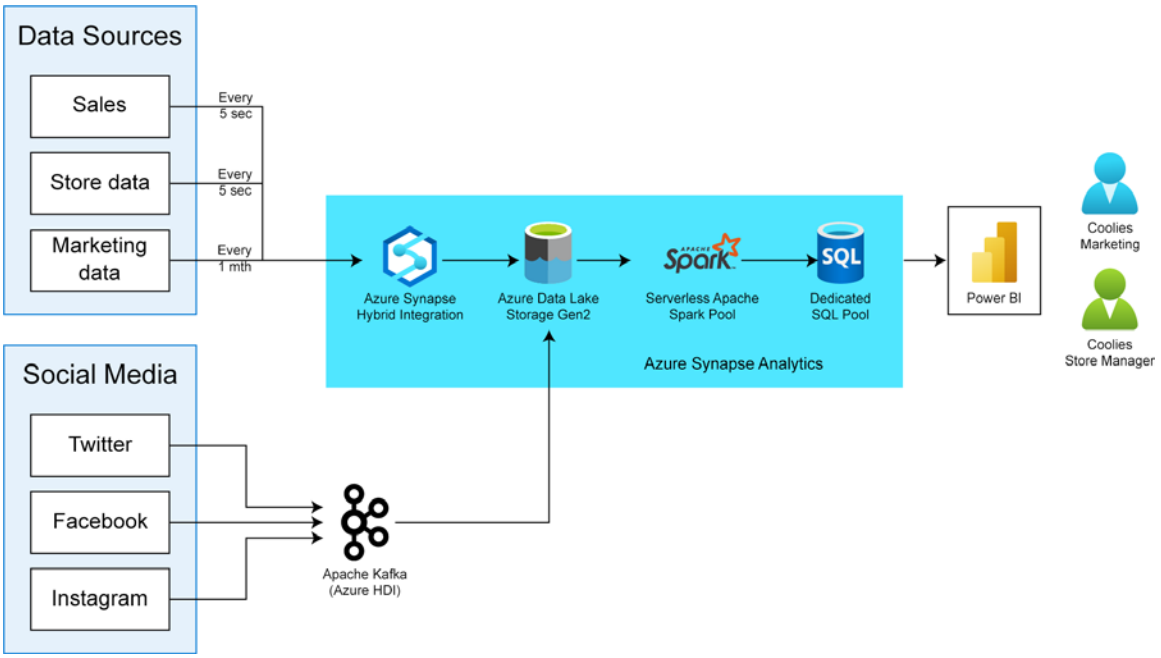


Figure 4.5: Data pipeline for initiative 2 (increasing sales)

Conclusion

You have seen how Coolies (or any other organisation) can take advantage of Microsoft Azure to discover customer insights and add value in near real-time. Microsoft Azure offers a wide range of services for data management and analytics and aims to streamline the development process, while at the same time raising the bar for quality and performance.

Furthermore, Microsoft Azure offers many data and analytics services as fully managed offerings, which means less overhead for Coolies (and any other organisation). The other advantage of using serverless Azure is that organisations and teams can start small with no major investment and scale up as demand grows. This is a great business approach, as it reduces the risks of upfront investment while at the same time mitigating the burden of bureaucratic approval for large expenditure at the start of a data project.

Finally, Microsoft Azure provides a great deal of documentation and learning materials online and aims to break the barrier to entry by offering free credit that can be used by any team or individual to start learning and building with Azure today.

Use case 2: Using advanced analytics on Azure to create a smart airport

Najad is a large city in the northern part of Egypt. The city's main airport, **Najad International Airport** (**NIA**), services 25 million passengers per year, which amounts to 70,000 passengers every day. It is Egypt's busiest airport and sees an average of 200,000 flights every year.

The management of NIA is hoping to adopt data analytics on Azure to improve capacity planning and quality of service. The goal is to use data to address operational issues that are currently hindering NIA's ability to fully utilise its infrastructure and resources. This will, in turn, improve customer satisfaction and enable NIA to scale its operation by serving more passengers and aeroplanes.

The following sections will define the problems that NIA is facing and brainstorm some design ideas. Finally, you will create a possible solution architecture on Microsoft Azure that can solve this problem and see why Azure is the perfect platform for such solutions.

The problem

To properly define the business problem, you first need to consider the challenges from the business perspective. Then you will look at the technical problems hindering the airport's ability to move forward.

Business challenges

As mentioned, NIA serves tens of millions of passengers every year. The volume of these passengers is forecasted to increase by about 20% in the next three to five years. In 2019, the airport suffered a loss of more than USD 370 million due to operational inefficiencies. This included costs of flight delays due to congestion and long queues, lost retail opportunities due to poor passenger experience, poor planning of staffing levels and under-utilisation of airport assets.

NIA's CIO, Zara Hassan, is relatively new (she's only been with NIA for 6 months) and has a background in data and business intelligence. Zara has a vision for turning this massive inefficiency at NIA into a business opportunity. She presented a business case to the NIA board to make small incremental investments in advanced analytics to reduce the airport's overall operational costs, while at the same time improving customer experience.

As a visionary, Zara understands that for the airport to succeed, it must move from observing historical reports to predicting the future. She wants her team to help airport management predict flight delays and mitigate such occurrences. She believes that if airport management has access to the right tools then capacity planning, resource allocation and safety can all be improved.

The proposed approach is to use data and artificial intelligence to model passengers, flights, baggage, assets and other datasets to be able to confidently predict passenger volume and crowd movement, which, in turn, will allow the airport to improve its operations and reduce costs.

The business challenges faced by the NIA data analytics team can be summarised as follows:

- The first major challenge for airport management is to improve capacity planning. Currently, the leadership team at NIA makes these decisions based on assumptions and previous experiences, which does not necessarily reflect reality. So far, NIA has not had a consistent data-driven approach to predict the number of passengers they can expect on a given day. Accuracy in predicting the number of expected passengers is critical for capacity planning, such as managing staffing levels and the purchase of equipment, as well as the planning of infrastructure upgrades. Moreover, NIA does not have a solution in place to predict the airlines that might get delayed or predict the number of security personnel the airport might need on a given day to serve passengers. This leads to overcrowding, long queues and inefficient infrastructure utilisation. Poor capacity planning alone was estimated to have cost NIA close to USD 160 million last year. Add to that new assets such as vehicles and carts that the airport purchased because of the perceived need, while in reality they just needed to improve the utilisation of existing assets.
- Resource allocation is another major concern for NIA's management. Passengers have to wait at the airport in long queues, whether at customs or at the airlines' check-in counters. Most of these long waits are due to the poor allocation of NIA staff to different areas of the airport. NIA's management wants to improve resource allocation, which would then improve customer satisfaction.
- The retail and duty-free shops make up a decent portion of the airport's revenue. NIA has a number of large billboards, and they use customer information to provide some occasional promotions. The NIA management would like to improve customer engagement, and eventually business opportunities, at these airport retail shops.

- A big part of customer service is to provide customers with the information they need when they need it. Travelling through an airport can be a very tiring experience, and it can also be stressful when passengers are running late or have a flight delayed or cancelled. NIA thus needs to update the flight status/delays in near-real-time. This requires NIA's management to think of creative and innovative ways to make the relevant information available to customers when they need it. This will reduce customer confusion and stress and improve the overall customer service.
- NIA needs an infrastructure overhaul in the long term. This would solve the problem of congestion, which has caused minor accidents in the past, and has cost the airport money while negatively impacting customer experience. However, NIA is looking to improve passenger flow and reduce congestion by making proper use of resources as a short-term solution for the near future. Congestion hampers the flow of passengers and creates safety hazards when too many people are forced to go through small halls and/or walkways, especially when there are old people, babies and people with physical disabilities This creates safety incidents, and each of these safety alerts and incidents costs the airport money, puts the lives of passengers at risk and negatively impacts the customer experience. The airport wants to improve passenger flow to reduce congestion and improve safety.

Now that you know the main pain points that the business side of NIA is hoping to address, you'll need to consider the technical challenges so that you can start designing a solution.

Technical challenges

No single source of truth: One main problem that NIA's CIO is trying to solve is the fact that NIA currently has no single source of truth in terms of data sources. Today, the airport relies on reports from a number of old internal systems, as well as reports from partners. These reports usually cover operational aspects of the previous day and week, and have conflicting figures. For instance, flight data is currently held by the individual carrier companies. NIA has more than 35 airline companies, each of which has its own systems and uses different terminology. This makes it extremely difficult for the NIA management to get credible reports in time, let alone have data-driven operations.

Latency in obtaining data and reports: Because NIA does not have control over flight and cargo data, it relies on partners to generate, aggregate and send operational reports. These reports are usually delayed by days or weeks. This significantly reduces the organisation's ability to action any insights from these reports and forces NIA to always be reactive in its operations rather than planning ahead. For instance, if a report is presented to airport management and shows that there were long queues that

caused flight delays yesterday, airport management can't change the situation since it happened in the past. Timely access to this data is critical for NIA and almost all other organisations.

Data availability and access: Innovation requires exploring possibilities and experimenting with options. In terms of data, this requires NIA to continuously explore, enrich and correlate flight and passenger data with external data sources. Unfortunately, NIA cannot do any of these things today because the data is sitting in many silo systems that the airport does not control.

Scalability: NIA currently has a SQL Data Warehouse that is hosted in its virtual data centre. The management team has been reluctant to invest in this data warehouse because it does not hold all the data. This makes the current data warehouse obsolete, because it does not help the business in finding the insights the airport needs. Moreover, this current SQL Data Warehouse does not have the ability to ingest and/or hold all data that NIA can collect.

Security: NIA has clear and strict policies to protect its data and all its customers' data. The airport is required to clear ISO/IEC 27001 and ISO/IEC 27018 certifications to ensure that security measures are properly applied to protect the airport, its suppliers, its customers and all stakeholders. NIA needs to guarantee all these security requirements in any potential solution.

Data serviceability: For any data to be useful, it needs to be provided to the relevant users at the right time. NIA currently serves notices and alarms to passengers using audio announcements, as well as a few large monitors placed in a few locations around the airport. This is highly inefficient because it creates noise, and it does not consider the context of who the user is or what the user wants to know. NIA now acknowledges that it needs to raise its game not only in improving data and report efficacy, but also in how these reports are served to users.

Based on these requirements, NIA's business intelligence team, along with Zara, agree to define the problem statement as follows:

NIA is losing more than USD 350 million a year because of operational inefficiencies, which include long queues, poor staffing levels and under-utilisation of airport assets. The NIA business intelligence team will work to deliver data analytics tools (dashboards, reports and apps) that help the business optimise operations and remove inefficiencies.

Design brainstorming

After defining the problem and articulating the business and technical challenges, the next few sections will help you to brainstorm some design ideas to come up with a solution design for NIA.

Data sources

Data is at the centre of any analytics solution. Hence, you need to start by thinking about the different types of data that NIA would need. Then, you need to think about a design to bring this data together. NIA needs to collect data from the following sources:

- **Customs data:** Customs data holds information about passengers and their declarations as they enter or leave the country. Currently, customs data is held by external systems. However, the airport can pull this data and integrate it into its systems. The current mechanism of integrating with the customs data system is by using a scheduled file dump to a file server. This can be used by NIA to pull customs data to its new platform.
- **Airline/flights data:** Currently, the individual airline systems hold the passenger data, their trips, check-in times and other related details. Although this data is held by the airline systems, the airport can integrate with these systems using integration APIs. The specific implementation of this integration will vary based on the individual airlines, but the airport needs to obtain this data in near-real-time.
- **Parking systems data:** The airport has sensors at all parking facilities that count cars coming in and going out. The parking systems also have an indication at any given point of how many parking spaces are available and where these spaces are. This data will need to be ingested with other sources.
- **IoT and video streams:** NIA has a number of traffic monitoring cameras that are installed throughout the site. These cameras send live video streaming and are used by the control room to direct resources and adapt operational procedures to cope with the traffic. The airport also has IoT sensors installed near the gates to indicate the status of each gate. There are also sensors that are aimed at monitoring crowd distribution in the airport. Data from all these sources (IoT and cameras) can be streamed for real-time analysis to provide NIA's management with actionable insights as traffic issues arise.
- **Baggage systems data:** The airport has an internal system that is used to manage all baggage data. This includes what luggage has arrived, on which flight and where it is now. The airport also serves logistics companies and receives multiple cargo flights every day. It is important to collect and analyse all the relevant data to serve these logistics companies for freight management.

- **Social media feeds:** To provide good customer service, it is essential for NIA to analyse passenger sentiment and feedback as passengers are expected to use social media platforms to share their experiences. This helps NIA to improve its services and address any concerning issues immediately.
- **Other data sources:** As discussed in *Use case one*, it is very common in data analytics to enrich existing datasets with other external sources of data to provide more context to any trends or patterns that are identified. This is especially true for airport operations where things can be heavily impacted by weather data, holiday seasons and other such factors. NIA will need to ingest many of these external data sources to complement its own operational data.

Data storage

The airport estimates its current existing data to be close to 310 TB, which does not include all the partners' data that needs to be collected and stored. To add to that, the airport is aiming to pull camera streaming and social media feeds. This could add an extra 15 GB of data per day, based on historical figures. This requires a highly scalable data storage service that can adapt elastically to the rapidly increasing volumes. To address this requirement, it makes sense to use a cloud-based service such as Azure Data Lake Storage to ensure elastic scalability and the ability to store data in various formats.

Data ingestion

To ensure that data is made available to the airport staff and customers in a timely manner, data needs to be ingested from internal and external sources quickly and efficiently. Based on the data sources that have been discussed, the solution needs to cater to multiple forms of data ingestion. This includes loading file dumps, processing real-time data streams from social media and monitoring cameras and pulling data by calling external APIs. The data team at NIA can either build their own integration and ingestion solution, which would be very expensive and would require plenty of development time, or use a cloud-based data ingestion and orchestration tool such as **Azure Data Factory** (**ADF**). ADF streamlines the data ingestion process by offering more than 80 pre-built data connectors, which can be used to integrate with a variety of source systems, such as SQL databases, blob storage and flat files.

Security and access control

The solution needs to provide the right security controls to NIA's management so that data can be secured and protected. Understandably, the airport has a long list of stakeholders that need to have access to data, including airport staff, security contractors, airline crew, passengers and partners. Therefore, the solution needs to enable NIA to provide row-level security to ensure that users only have access to their data. This requires a fine-grained access control management system that is built into the chosen platform so that the NIA business intelligence team does not need to spend too much time worrying about security. The focus of the NIA business intelligence team should be on finding insights that can help airport management.

Discovering patterns and insights

A key part of Zara's strategy is to empower the business to make decisions intelligently. This intelligence is assumed to be acquired by exploring and discovering trends and patterns in data. The major challenge here is where and how to build these machine learning models. Building such models requires working with many large datasets and an elastic pool of compute resources. The team acknowledges the challenge ahead and is looking to use managed Apache Spark clusters to empower their data engineers. The team was impressed with the scalability, security and wide support of tools and frameworks on Azure Synapse Analytics.

The solution

NIA's CIO, Zara, with the help of the business intelligence team, agreed to use Microsoft Azure as the cloud provider to build the new solution. They summarised their reasoning as follows.

Why Azure for NIA?

- NIA is already using Microsoft technologies such as Windows 10, Office 365 and other tools. Azure has better native integration with all these services than any other cloud provider. Thus, it makes perfect sense to use Azure. Furthermore, NIA is keen to take advantage of the **Open Data Initiative** (**ODI**), which enables organisations to deliver exceptional business insights by combining behavioural, transactional, financial and operational data in a single data store. The initiative simplifies the creation of common data models across the organisation and was developed jointly by Adobe, Microsoft and SAP.
- Using Azure means that NIA can keep using the same centralised identity server, which is managed using Azure Active Directory for Office 365. This means better security for NIA and less overhead when creating and managing new user accounts.

- Azure has more regional data centres than other major cloud providers. This means that Azure can provide NIA with higher resiliency and service availability. In addition to that, Azure is the only cloud provider that has a regional data centre in Africa, which is where NIA is based. This makes Azure the perfect choice, as it ticks all the boxes.
- The NIA business intelligence team found that using Azure is more cost-effective than using other cloud providers. Azure Synapse Analytics is up to 14 times cheaper than AWS or Google services, as explained in the first use case. Moreover, Azure provides the ability to use reserved instances for virtual machines and compute instances, which can give even greater discounts. Furthermore, by utilising its existing enterprise agreement with Microsoft, NIA can get even further discounts on all Azure service retail prices. This makes it very hard to justify choosing any other cloud provider.
- NIA also considered Microsoft's track record in developer technologies and developer experience as a big advantage for Azure. As a software development company, Microsoft provides the best developer experience by employing its wealth of intellectual property in this field. This means that NIA can have a good development and deployment experience when using Azure.
- Azure has achieved more than 30 compliance certifications, including **ISO/IEC 27001** and **ISO/IEC 27018**, which are also required by NIA. Add to that the fact that the Azure business model is not based on using or selling customer data, which is part of the business model of other cloud providers. This gives greater assurances to NIA and its board that their data and their customers' data is well protected.
- NIA is also hoping to use Azure Stack, which provides the airport with the ability to host applications and services on-premises and in the cloud seamlessly using the same base infrastructure, which is powered by Azure and Azure Stack.
- Finally, NIA wants to have the ability to choose a mix of PaaS, SaaS and open source tools. Azure enables NIA to do exactly that by offering great PaaS and SaaS services such as Azure Data Lake, ADF and others, while at the same time supporting native integration with open-source services such as Azure Synapse Analytics.

Solution architecture

Now that the BI team has refined the requirements and a cloud platform has been chosen, it is time to come up with a secure and scalable design. The NIA business intelligence team went with the following solution architecture:

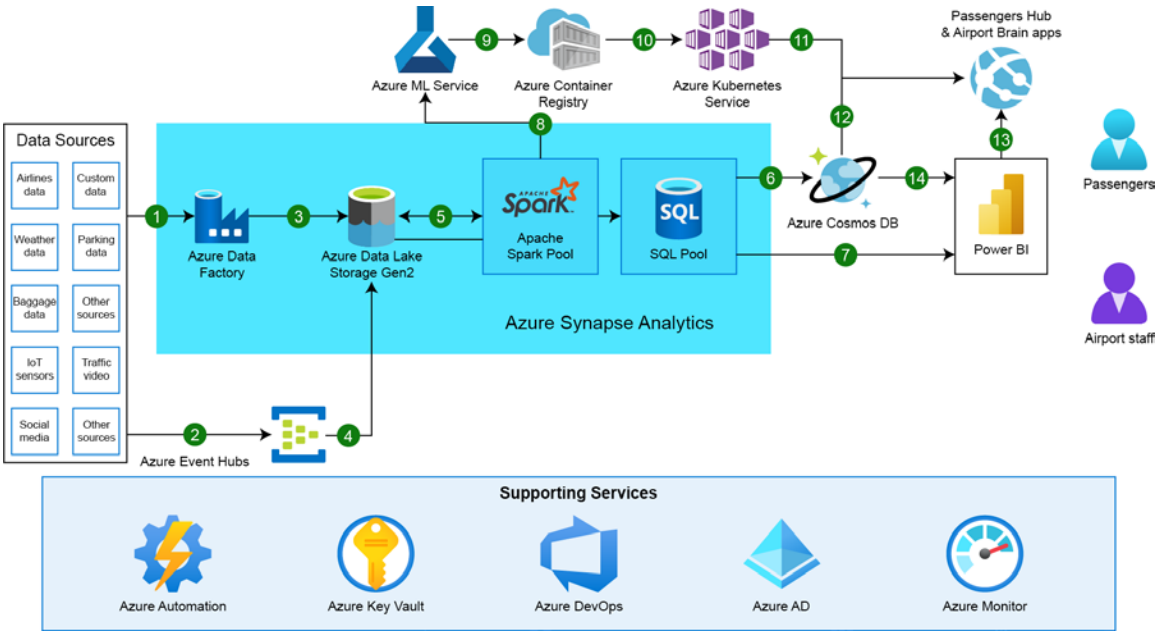


Figure 4.6: Solution architecture for NIA

The design in *Figure 4.6* shows the solution architecture and the data flow between the individual components. Here's an explanation for each of the workflow segments, as marked (numbered) in the diagram:

1. Structured data such as **Airlines Data**, **Custom Data** and **Baggage Data** is ingested using **Azure Data Factory**. This includes other data sources, such as data from the parking systems and weather data. ADF provides the ability for NIA to configure an integration runtime that can be used as a gateway to connect to NIA's on-premises data sources from within Azure.

2. All unstructured data, including **IoT Sensors** data, **Traffic Video** streaming and **Social Media** feeds, will be ingested using **Azure Event Hub**, which provides direct data capture to Azure Data Lake Storage Gen2.

3. **Azure Data Factory** pushes the ingested structured data to **Azure Data Lake Storage Gen2** for storage.
4. The NIA data team is adopting Azure Event Hub to ingest data streaming and easily and quickly capture the ingested data in Azure Data Lake Storage Gen2. The team takes advantage of Azure Event Hub's data capture feature to store the data to the data lake.
5. Using notebooks, NIA data engineers can then read, clean, transform and model the data using Apache Sparks runtime on Azure Synapse Analytics.
6. Cleaned data that is ready for consumption is then pushed from Azure Synapse to **Azure Cosmos DB**. This includes the most recent flight data and baggage data. Airport applications and APIs can then pull this data directly from Cosmos DB to serve passengers and staff.
7. **Power BI** is also configured to read more extensive datasets directly from **Azure Synapse**. Examples of the data that will be pushed using this mechanism include the results of decoding the traffic video stream to create crowd heatmaps.
8. **Azure Machine Learning** is used to manage the machine learning models, datasets, experiments and new model images. Azure Machine Learning has native integration with Azure Synapse Analytics.
9. Azure Machine Learning is used to train and build machine learning models. The resulting models are stored as Docker images in **Azure Container Registry**. Docker images are generally used as a way of packaging machine learning models with all their dependencies (libraries, source code and configuration files) as one single deployable package. This improves the development lifecycle and reduces deployment errors.
10. **Azure Kubernetes Service** is configured with the deployment to take the new machine learning model images from **Azure Container Registry** and run these models as Kubernetes pods. This makes the machine learning models available to generate predictions by making simple HTTP calls. Examples of the machine learning models include a recommendation engine for gate assignments and a parking demand forecasting machine learning model.

11. The NIA business intelligence team can deploy the new machine learning models via web applications, which can be hosted on **Azure Kubernetes Service**. These web applications can then interact with **Azure Cosmos DB** to save machine learning inference data (such as what actions are recommended to the airport staff), as well as to serve curated data such as flight schedule and baggage data. Examples of these web applications are **Passengers Hub** and **Airport Brain**. **Passengers Hub** is designed to be the one central portal to serve all passengers' data, which includes things such as flight details, gate numbers, check-in counters and recommendations to the passenger. Passengers can see all this information on their mobile device by downloading the airport's mobile app. **Airport Brain** is the name given to the new central portal for airport management staff. The goal is to provide NIA's management with the tools to enable efficient operations. The portal uses data to provide recommendations on gate assignments, staffing levels and the distribution of airport assets.
12. Both **Passengers Hub** and **Airport Brain** make use of the curated data that is stored in Azure Cosmos DB. Azure Cosmos DB is also used to store application-specific types of data such as users' sessions and alerts. This is all enabled by Azure Cosmos DB's blazing-fast querying engine and high responsiveness.
13. Both **Passengers Hub** and **Airport Brain** require data visualisation. Power BI is used to build these reports, and then the web page embedding feature is used to present these Power BI reports in the new web applications. The curated data includes passenger-related information such as passenger flight details, any predicted delays and passenger baggage information.
14. The Power BI dashboard serves reports and visuals that combine data from Azure Cosmos DB with curated data from Azure Synapse Analytics.

Azure services

As in the first use case, the following sections will elaborate on each of the Azure services that are shown in *Figure* 4.6. They will first explain why each service is needed, why it is suitable for NIA and then, finally, show a brief practical example of the core part of its implementation. To avoid repetition, the Azure services that are covered in first use case are skipped unless NIA has specific requirements for that service.

Azure Synapse Analytics

Role in the design

Azure Synapse serves as the unified platform for big data analytics as well as data warehousing. Azure Synapse is needed to provide the compute power needed to process data and to foster greater collaboration between the many stakeholders.

Why Azure Synapse?

Besides all the advantages of Azure Synapse that were covered in *Use case 1*, Azure Synapse supports multiple languages, machine learning frameworks (such as TensorFlow and PyTorch) and integrates with many open source tools. NIA's business intelligence team needs a platform that can handle both data engineering and data science workloads. Azure Synapse is designed to serve this purpose by enabling data engineers to clean, merge, transform and curate data, while at the same time empowering data scientists to use any of the popular machine learning frameworks, such as TensorFlow or PyTorch.

Sample implementation

The following code snippet configures a connection to Azure Cosmos DB from a Spark notebook using the Azure Cosmos DB connector. The following Python code has a placeholder for the master key of the Azure Cosmos DB instance and assumes that NIA has an Azure Cosmos DB instance that is called **`NIAAnalytics`**, which has a collection called **`flights_data`**. The code saves a **`flights`** DataFrame (Spark DataFrame) to 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

Role in the design

Azure Cosmos DB serves two main purposes: it stores all the application data for applications such as the Passengers Hub and Airport Brain apps, and it is also used to serve curated data that is ready for consumption by the airport staff and external stakeholders (such as passengers).

Why Azure Cosmos DB?

There are many options for storing NIA's curated data and application data. However, the NIA business intelligence team decided to choose Azure Cosmos DB for the following reasons:

- Azure Cosmos DB provides out-of-the-box turnkey global distribution, which is great for ensuring the availability and resiliency of the NIA platform. Understandably, NIA cannot afford downtime because it serves millions of passengers all year round. Thus, its new platform needs to have high availability that can be powered by Azure Cosmos DB.
- The NIA platform needs to provide data in near-real-time. Therefore, it is important to reduce latency. Azure Cosmos DB enables NIA to have a single-digit millisecond latency. This is also complemented by Azure Cosmos DB's impressive SLA of 99.999%.
- As mentioned before, NIA estimates its current data to be over 310 TB, with a growth rate of 15 GB per day. This does not yet include data coming from airline partners and external data sources such as weather and traffic. For this reason, the team chose Azure Cosmos DB for its elastic and unlimited scalability. Azure Cosmos DB provides NIA with the scalability it needs, with the option to only pay for what is used in terms of storage and throughput.
- The airport currently has multiple internal systems to hold its data, including SQL servers and MongoDB servers. The team wants to have greater compatibility with all these existing source systems and to enable existing applications to work with the new database without having to make any changes. Azure Cosmos DB is the perfect choice for this requirement because it provides a multi-model engine with a wire protocol-compatible API endpoint. This means that NIA applications can connect to the same Azure Cosmos DB instance using multiple drivers, such as MongoDB, SQL and Gremlin. This simplifies the development and deployment effort because it uses the same drivers' APIs, and it also reduces the total cost of ownership because of the room for knowledge transfer and the reduction of the need to rework.

- Another feature of Azure Cosmos DB that appealed to the NIA business intelligence team was the ability to do real-time operational analytics and AI on top of Cosmos DB. Azure Cosmos DB has out-of-the-box integration with Apache Spark and enables running Jupyter Notebooks to work with data in Cosmos DB directly without further integration or custom development work.
- Commercially, Azure Cosmos DB is a cost-effective option because it offers the business intelligence team the flexibility and control that is needed. The beauty of using Azure Cosmos DB is its ability to offer planet-scale functionality with the ability to control the costing model based on the storage and throughput that is needed. This means that when an update is executed on a record in Azure Cosmos DB, every user in the world can see this update within milliseconds.
- Finally, Azure Cosmos DB is a fully managed service, which means the NIA team will only need to worry about the data it stores in Cosmos DB, and not the infrastructure. Moreover, this allows the team to start quickly and cheaply, and to scale as they start on-boarding more datasets and demonstrating more business value.

Sample implementation

One of the nice things about Azure Cosmos DB is its compatibility with many querying models and drivers. The following code snippets show how Cosmos DB can be queried using SQL or MongoDB. Both samples are written in C#:

1. The first code snippet queries records from the **`passengers`** table, looking up passengers with the name **`Bob`**. Then, it iterates through all the returned results and prints the name of the passenger to the console:

```
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. The second code snippet performs a similar query, but it uses the MongoDB API. It creates **`MongoClientSettings`** first, and then **`MongoClient`**, which is then used to get a reference to Azure Cosmos DB. The code assumes that the configuration settings have already been configured at this point. The code creates a reference to NIA's Azure Cosmos DB (**`NIAAnalytics`**) and queries **`passengerCollection`**:

```
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

Role in the design

Azure Machine Learning is used by the NIA business intelligence team to operationalise their machine learning models. To optimise resource allocation, the team needs to build several machine learning models to predict the number of passengers and to create a recommendation for gate allocation. Azure Machine Learning gives the business intelligence team a consistent and reproducible way of generating machine learning models, while keeping track of all machine learning experiments, datasets and machine learning training environments at the same time. This is critical for any machine learning model implementation, where explainability is a basic requirement for customers and stakeholders.

Why Azure Machine Learning?

- Azure Machine Learning enables NIA to streamline and accelerate the whole machine learning life cycle, from data clean-up and feature engineering to model creation and validation. Azure Machine Learning makes it easy to automate many parts of the pipeline. This in turn reduces overhead, improves quality and allows the NIA team to innovate more quickly.
- Versioning and maintaining multiple snapshots of datasets is a common practice when creating and experimenting with machine learning models. It can be a very tedious and confusing process to maintain multiple versions of the same datasets. Azure Machine Learning provides a full set of features that aim to help customers such as NIA tackle this challenge. Azure Machine Learning datasets enable NIA to track, version and validate datasets with ease, as can be seen in the *Sample implementation* section.

- Part of the challenge for any advanced analytics team is finding the right algorithm to use to create a machine learning model. Not only does the NIA business intelligence team need to pick the right algorithm, but they also need to fine-tune any hyperparameters. Azure Machine Learning automates this whole process so that any data analyst can be a data scientist. Azure AutoML enables NIA's business intelligence team to automate the process of creating machine learning models quickly, easily and cheaply.
- Compatibility is also another big plus of Azure Machine Learning. Azure Machine Learning integrates nicely with open source tools such as Apache Spark on Azure Synapse. It also enables NIA to use any machine learning frameworks (such as TensorFlow and PyTorch), while at the same time taking full advantage of Azure Machine Learning.
- Azure provides organisations such as NIA with all the latest breakthrough innovations in data and AI. One of these breakthroughs is the concept of abstracting computing from the actual data and its pipeline. This enables the NIA business intelligence team to write their code once and run it on any compute. This includes data transformation code and machine learning model code. The NIA team can build their machine learning model, run it locally on their development machines and when ready, move that code to run on the cloud. This provides developers and organisations with great flexibility in terms of development and operational costs. NIA can pay only for the compute they use and only run the training of the machine learning models on the cloud when large computing resources are needed.
- NIA's CIO is a strong believer in DevOps and the benefits it brings to an organisation. Support for DevOps processes was a major factor in deciding to choose Azure Machine Learning. Azure Machine Learning has native integration with Azure DevOps, which allows NIA to create and deploy machine learning models with ease.
- Security, reproducibility and governance are all significant concerns for any advanced analytics team. Microsoft Azure addresses all these nicely and elegantly through native integration with other Azure services that are all battle-tested for enterprises. Azure Machine Learning offers out-of-the-box integration with Azure AD and Azure Monitor. Moreover, by using Azure Resource Manager templates and Azure Blueprints, organisations such as NIA can enforce proper governance and standards.

Sample implementation

Azure Machine Learning makes it easy to version, track and work with multiple versions of a dataset for machine learning purposes. The following code snippet first creates a data store to tell Azure Machine Learning where it should store the data:

```
# 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')
```

The preceding code snippet written in Python first creates an Azure Machine Learning workspace from an existing configuration file. The code then creates a datastore by registering an Azure Blob container as the data store. The sample names the datastore **`NIA_airport_datastore`** and has placeholders for the Azure Storage account name and key. Finally, the sample creates a reference to a datastore that already exists by using its name.

The following code snippet registers a new dataset and provides a name, a description and a tag to make it easier to find this dataset in the future:

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

The following code snippet retrieves an existing dataset by name and/or version ID. This is very useful when we have multiple versions of the same dataset:

```
#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

Role in the design

Azure Machine Learning Services enables the NIA business intelligence team to create their machine learning models as standard containers that can be run on any container engine, such as Docker and Kubernetes. The team uses Azure Container Registry to securely host and share these Docker containers, which hold their machine learning models.

Why Azure Container Registry?

- Azure Container Registry enables NIA to store images for all types of containers. Azure Container Registry abstracts the hosting of the images from the deployment of these images to the different deployment targets, such as Docker Swarm and Kubernetes. This enables NIA to use one container registry (Azure Container Registry) to host images for all types of containers.
- Azure Container Registry builds on the functionalities of the standard container registries. For instance, Azure Container Registry integrates with Azure AD to improve security. Moreover, Azure Container Registry provides a simple way to integrate with container actions using triggers. For instance, NIA can configure a webhook to trigger Azure DevOps Services when a new image is added to Azure Container Registry.
- Azure Container Registry is fully compatible with the standard Docker Registry v2. This means that the NIA team can use the same open-source Docker **command-line interface** (**CLI**) tools to interact with both registries (Azure Container Registry and Docker Registry v2).
- Azure Container Registry supports multi-region replication. This appeals to NIA because it helps with two things. Firstly, it reduces network latency and cost by keeping the container registry close to the deployment targets. Second, it improves business continuity and disaster recovery since the same container registry is replicated across multiple regions.

Sample implementation

The following code is part of the **Azure Resource Manager** (**ARM**) template that NIA uses to create the Azure Container Registry instance. It creates an Azure Container Registry (Standard tier) instance in Azure's South Africa North data centre.
The template also enables the Admin User account to manage the registry. The ARM template has two parameters, one parameter for the name of the registry and another parameter for the 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

Role in the design

Azure Kubernetes Service (**AKS**) is used to serve machine learning models as consumable APIs. An example of these machine learning models is a model that predicts crowd movement through the airport. Such models are trained by the team using historical data in Azure Synapse. Then, using Azure Machine Learning, the model is pushed as a Docker image. AKS runs these models and other apps, such as Passenger Hub. Moreover, AKS helps manage the service discovery of these apps, provides autoscaling mechanisms and facilitates self-healing policies for handling errors and failures.

Why AKS?

- Managing a cluster of computers is a hard task, and it is even harder to manage and configure a Kubernetes cluster. That is because Kubernetes has many moving parts and requires lots of configuration. AKS simplifies this by offering a managed cluster. This means Microsoft Azure manages the master nodes and the NIA team only needs to configure and use the slave nodes for deploying their workloads. This reduces the overhead for NIA significantly.
- Using concepts such as **virtual node** and **virtual kubelet**, AKS allows NIA to provision additional capacity elastically at any time. This is important for NIA because it is very hard to predict the load and the capacity needed, and therefore it is important to have this elastic provisioning when needed.
- The native integration and support for AKS in Azure DevOps is another advantage of AKS. This simplifies configuring and automating deployments of NIA workloads into AKS. AKS also has native integration with services such as Azure Monitor and Azure Key Vault.
- The NIA team can improve and speed up the end-to-end development experience using the Visual Studio Code's support for AKS.
- Besides the native integration with other Azure services, AKS integrates nicely with Azure Active Directory. This means that NIA can improve security by taking advantage of this integration. Furthermore, NIA can use Azure Policy to enforce governance across the whole organisation.
- Azure provides great support for open source tools such as Kubernetes, not only in the cloud, but also on the edge. The NIA team understands that there are scenarios where pushing computing to the edge might be the best option. An example of this is their plan to push machine learning models close to traffic monitoring cameras to trigger alerts when a safety event occurs. Microsoft Azure has good support for running Kubernetes on Azure IoT Edge for such scenarios. Therefore, using AKS will be a good option for future plans to push machine learning models to the edge using Kubernetes with Azure IoT Edge.

Sample implementation

The following code snippet is part of NIA's Azure DevOps Services pipeline that deploys the new Airport Brain web application. The code takes advantage of Azure DevOps' support for Kubernetes by using the **`KubernetesManifest`** task type. The task deploys the Docker image at **`nia/aiportbrain:lastest`** to the pre-configured AKS by using **`NIAairport_AksServiceConnection`**. The following code has a placeholder for **`imagePullSecret`**, which is used as the authentication mechanism to pull images from Azure Container Registry to the deployment target (AKS):

```
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

Role in the design

Part of Zara's strategy for NIA's reporting is to have Power BI as the visualisation tool within NIA. Power BI can be used to generate reports and dashboards, as well as for self-service purposes. The BI team is hoping to also take advantage of Power BI's ability to be embedded into other web apps to reuse Power BI visualisations inside other new apps, such as Passenger Hub.

Why Power BI?

Besides all the benefits of Power BI that were mentioned in Use case one, Power BI reports and dashboards can be embedded into other web applications. The NIA business intelligence team wants to take advantage of the simplicity and power of Power BI's visuals to build the dashboards for the Passenger Hub and Airport Brain apps. Using embedded Power BI reports enables NIA's business intelligence team to build and ship reports quickly and easily, while at the same time serving them securely through NIA's new web applications.

Sample implementation

In Power BI, you can embed any Power BI report inside any web page. From the Power BI service, while displaying the report, click on the **Share** option and then select **Embed report**, followed by **Website** or **Portal**. This creates a dialogue box that includes your embed code to be used on your destination website or portal:

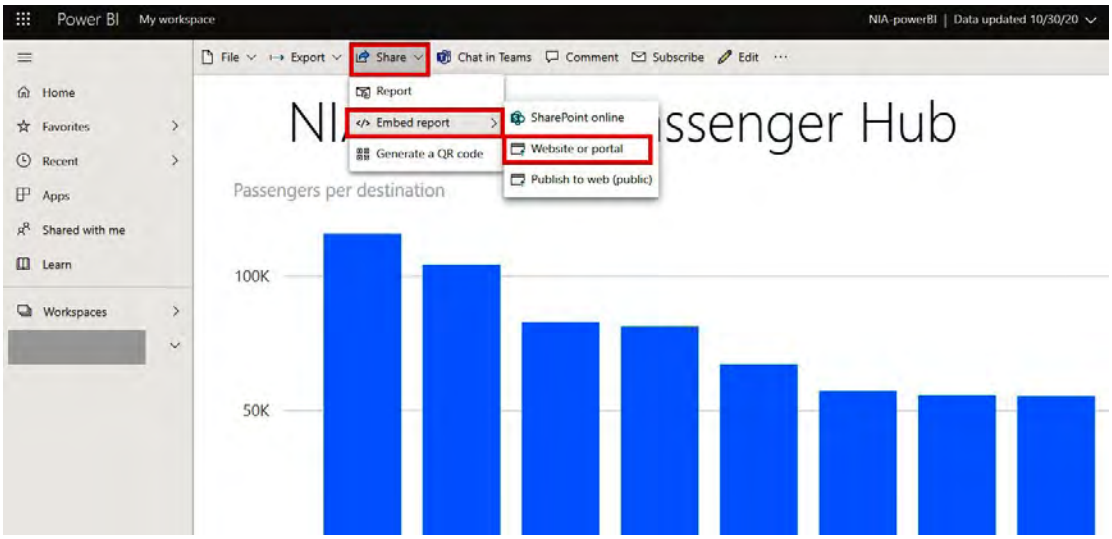


Figure 4.7: Creating an embed code

The dialogue box will show an HTML **`iFrame`** code that can be used on any HTML web page. The next dialogue box allows the NIA team members to configure the properties of the **`iFrame`** code, such as the **width** and **height**. Then, using the **`iFrame`** code, the report can be embedded into any web application:

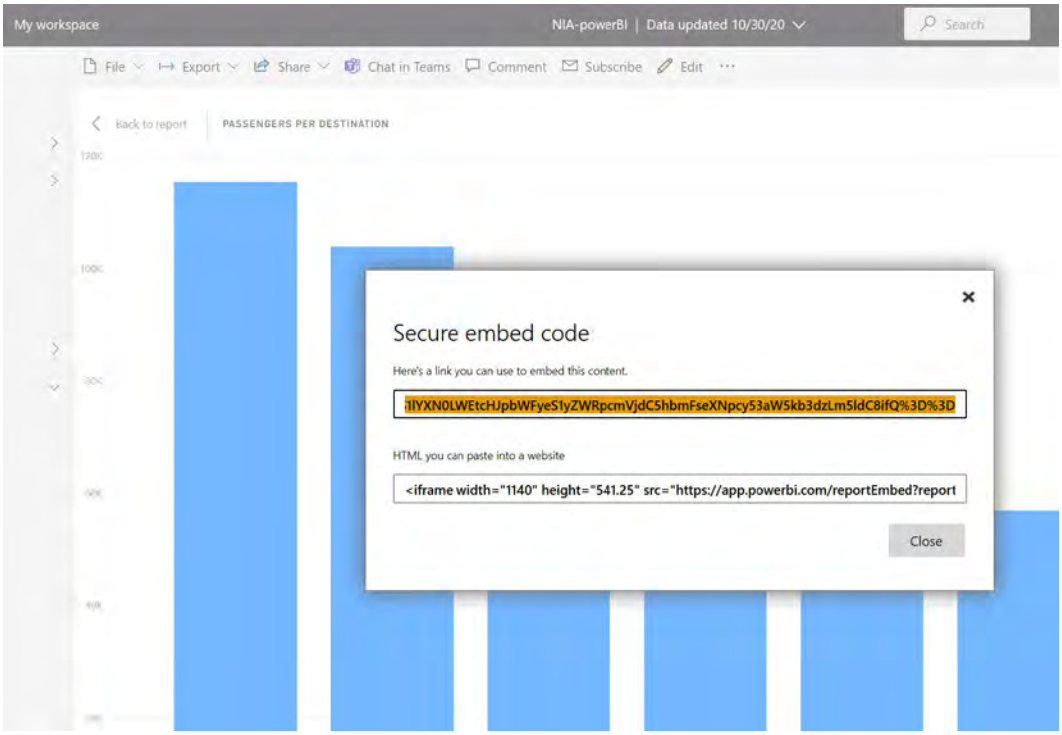


Figure 4.8: Configuring the iFrame properties

Supporting services

NIA wants to ensure that the new solution is secure and scalable and has a good level of monitoring and support. Azure has many services that enable organisations such as NIA to secure, scale and monitor their solutions. This includes all the services listed in the first use case, such as Azure DevOps, Azure Key Vault and Azure Monitor.

Insights and actions

Azure helped NIA draw meaningful insights after analysis and deploy necessary measures, as discussed in the following sections.

Reducing flight delays by 17% using predictive analytics

Description: While performing initial data discovery and exploration, the NIA business intelligence team found that inefficient gate assignment was a major contributor to flight delays. Flight delays have a snowball effect because a delay in one flight can impact the next flight and the one after that. There is also the negative passenger experience that it produces. Currently, the assignment of gates at NIA is based on the capacity of their waiting area and the maximum capacity of the aeroplanes. This assumes that all flights are full, which is not necessarily true.

Combining weather data, city traffic data, historical flight delay data and other sources allowed the business intelligence team to produce a better recommendation engine for gate assignment. The new recommendation engine, which was built using machine learning, looks at contextual (weather and traffic) data and historical data to estimate the number of passengers on a given flight and assign a gate accordingly. During initial modelling and validation, the team found that deploying such a recommendation engine in the Airport Brain app can reduce flight delays by 17%.

Estimated business value: USD 14.7 million/year

Key data sources: Airlines flight data, airport data (layout and gates), weather data, city traffic data, school calendars and public holiday calendar.

Actions: The NIA business intelligence team deployed the solution using the architecture shown in *Figure* 4.6. As a result of this solution, airport management now has a new tool as part of the new portal (**Airport Brain**) to provide real-time recommendations for assigning gates. This improves efficiency and reduces operational overhead by excluding assumptions in planning and introducing operational decisions that are made based on facts and science.

Data pipeline: The simplified data pipeline for this initiative is shown in *Figure* 4.9:

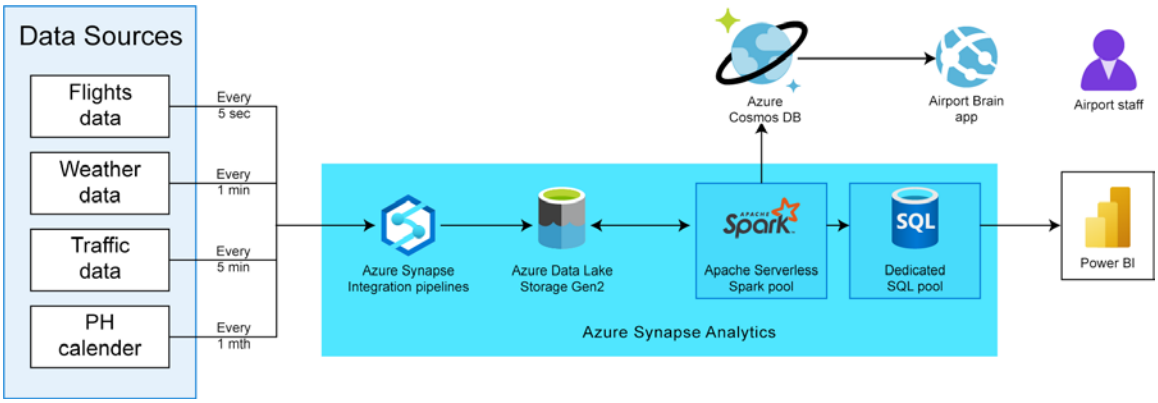


Figure 4.9: Data pipeline for initiative 1

Reducing congestion and improving retail using smart visualisation

Description: Another encouraging discovery that the NIA business intelligence team made was the correlation between passengers' arrival times by car and long queues. The team found that when a large number of passengers arrived at the airport more than four hours before their flight departure time, long queues and overcrowding became an issue. This can be attributed to the fact that the airport management team did not expect/plan for these passengers to be there at this time, which caused the long queues and congestion. Another explanation that one of the senior managers at the airport had was the fact that these early-arriving passengers were going directly to the gate and not to the airport's other facilities.

Therefore, the team decided to address this issue by directing these early-arriving passengers to the airport's other facilities, such as the duty-free area, the cinema and the rest areas. Based on initial testing, the team estimates that this can increase retail opportunities by 11% while at the same time reduce overcrowding at the airport gates by approximately 15%.

Estimated business value: USD 9.3 million/year

Key data sources: Airline flight data, airport data (layout and gates), weather data, passenger info, airport retail data and public holiday calendar.

Actions: Based on these findings, the team created new dashboards in the Passenger Hub app. When passengers arrive early and scan their ID, the Passenger Hub app shows them their flight details and guides them to rest areas, duty-free shops and the airport cinema. The team also used real-time traffic monitoring data to create signs to be used on large screens across the airport so that users can see them without even scanning.

Data pipeline: The simplified data pipeline for this initiative is as shown in *Figure 4.10*:

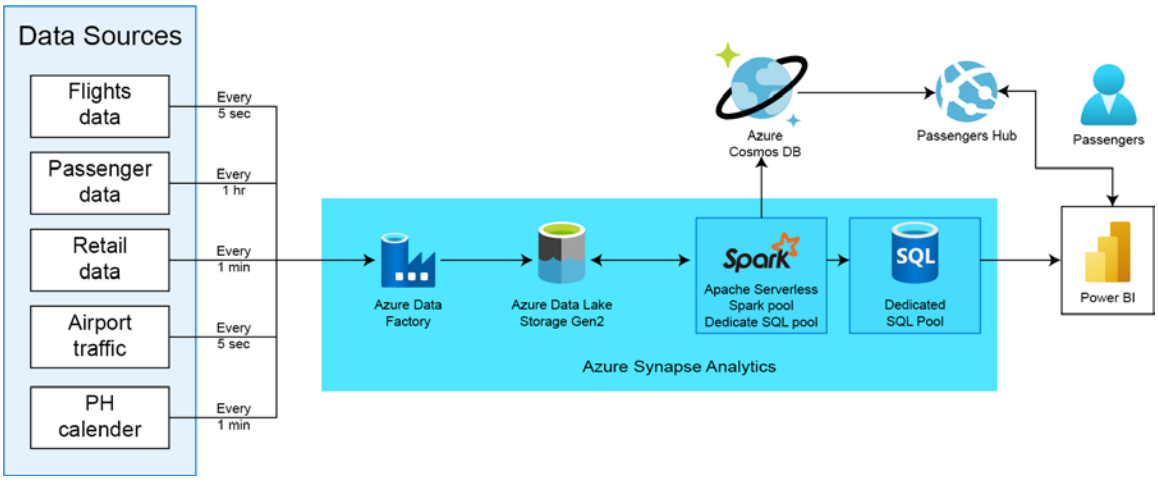


Figure 4.10: Data pipeline for initiative 2

Conclusion

Airports have complex operations and procedures, and they run around the clock. Thus, even making small improvements can provide great savings and can improve safety and customer satisfaction.

In the previous few pages, you have considered a practical example of a large airport. Although the names are fictitious, many of the numbers discussed here are based on an actual use case that the author was involved in. You have seen how advanced analytics can be used to improve efficiency and save an organisation millions of dollars. Data can be used not only to help airports save on their operational costs, but also to create a competitive advantage.

You have also looked at how a data-driven solution can be implemented using Azure and seen why Azure is the perfect platform for running such workloads. Azure is affordable, secure and provides organisations with the ability to be agile and scalable.

5

Conclusion

Nowadays, data is the driving force behind corporate success. Every organisation is utilising data to formulate business decisions. Using the enormous increase in data being generated and collected by organisations from a multitude of data sources (whether it's structured, semi-structured or unstructured data), Azure Synapse Analytics delivers a limitless analytics service that brings together data ingestion, enterprise data warehousing and big data analytics.

The unified experience in Azure Synapse Analytics allows customers to build end-to-end analytics solutions and perform data ingestion, data exploration, data warehousing, big data analytics and machine learning tasks from a single, streamlined environment. Azure Synapse offers a promising means of analysing data to get real-time insights. This is essential for making business decisions and deriving business strategies.

This book took you through a journey on cloud analytics with Microsoft Azure by showing you how you can build your data warehouse with Azure Synapse, process and visualise data, and build end-to-end analytics solutions using the unified environment in the Synapse Studio.

In the following section, we will recap the book, chapter by chapter, to remind you of the materials and technologies that we have covered.

Chapter 1, *Introducing analytics on Azure*, discussed the importance of data analytics and highlighted several reasons why Microsoft Azure is an ideal platform for achieving business intelligence capabilities on the cloud. It touched on some fundamental concepts around big data, machine learning and DataOps. You also learned about some of the business drivers for adopting data analytics on the cloud. Lastly, you gained a high-level view of what it takes to have a modern data warehouse.

Chapter 2, *Introducing Azure Synapse Analytics workspaces and the Synapse Studio*, introduced you to the new unified experience in the Azure Synapse workspace and the Synapse Studio.

The Synapse Studio provides an all-in-one streamlined environment for data prep, data management, data warehousing, big data analytics and AI tasks. It offers the following features:

- Code-free visual environments for managing data pipelines
- Automated query optimisation
- The functionality to build proofs of concept in minutes
- Serverless on-demand queries
- The option to securely access datasets and use Power BI to build dashboards in minutes – all while using the same analytics service

You learned how to get started with Azure Synapse by creating a workspace through our step-by-step guide. We also demonstrated how you can start building your end-to-end analytic solution with the tools in the Synapse Studio. We explored some of the key capabilities in Azure Synapse, including serverless SQL on-demand. With serverless SQL on-demand, we can instantly perform data exploration and data analysis using familiar T-SQL syntax without having to provision resources.

To complete the journey, you also learned how to:

- Provision an SQL pool, ingesting data and analysing the data in the SQL pool
- Create an Apache Spark pool, ingesting data and exploring data using Spark
- Copy data to/from an SQL pool and an Apache Spark pool
- Pull external data from linked data sources
- Analyse data using serverless SQL on-demand
- Integrate with Pipelines

Chapter 3, *Processing and visualising data*, focused on the analytical side of the modern data warehouse, where we demonstrated how you can process and visualise data using Power BI and implement machine learning. With Azure Synapse and Power BI, you can perform powerful, customisable, self-service data analytics to find and share data insights. While Azure Synapse is the engine that powers these insights, Power BI is a visualisation tool that empowers users to analyse data for themselves.

For advanced analytics, Azure Machine Learning gives you the infrastructure and tools to analyse data, create high-quality data models and train and orchestrate machine learning as you build intelligent apps and services. The benefit of using Azure Machine Learning is that it delivers the predictive intelligence that businesses need to stay competitive.

Chapter 4, *Business use cases*, contained real-world use cases on how all of these technologies integrate with one another to provide complete end-to-end data warehouse solutions. The two real-world business use cases in this chapter demonstrated high-level solutions using Microsoft Azure. They also illustrated how real-time data can be analysed in Azure to derive meaningful insights and make business decisions. The sample implementations and use cases demonstrate how real organisations have used Azure technologies in different sectors to make the most out of data, giving you an idea of how you can leverage this powerful technology to help your own business.

Final words

Azure Synapse brings the worlds of enterprise data warehousing and big data analytics together into a unified experience that helps you accelerate your time to get insights. The cloud model for modern data warehouses is not only flexible and scalable, but it is also cost-effective due to its unique elastic properties. Analytics workloads are one scenario where the elasticity truly shines.

With Azure Synapse, data professionals of varying skill sets can collaborate, manage and analyse their most important data with ease, all within the same service. From Apache Spark integration, with the powerful and trusted SQL engine, to code-free data integration and management, Azure Synapse is built for every data professional.

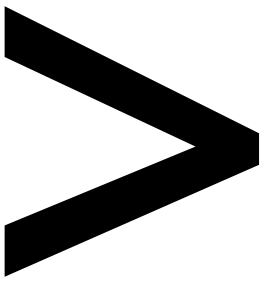
Now that you have reached the end of this book, you are armed with the knowledge of the services and tools you need to build your own complete data analytics solution on Azure. The best way to begin is to start small, by integrating some of the technologies this book has introduced into your existing workflow. Then, gradually add more capabilities in the future as your needs evolve.

Best of luck with your modern data warehouse and cloud analytics journey!

In the following section, we provide you with additional resources for further learning.

For further learning

- **Get started with 12 months of free Azure services:** https://azure.microsoft.com/free/synapse-analytics/
- **Azure Synapse Analytics:** https://azure.microsoft.com/services/synapse-analytics/
- **Azure Synapse Analytics Toolkit:** https://azure.microsoft.com/resources/azure-synapse-analytics-toolkit/
- **Compare the price and performance of Azure Synapse Analytics versus Google BigQuery and Amazon RedShift:** https://azure.microsoft.com/services/synapse-analytics/compare/
- **Four Steps to Supercharge your Analytics** (PDF eBook): https://clouddamcdnprodep.azureedge.net/gdc/gdcEbYaLj/original



Index

About

All major keywords used in this book are captured alphabetically in this section. Each one is accompanied by the page number of where they appear.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z