packt

FIRST EDITION

The Modern Data Estate: Foundation to Accelerate Digital Transformation

Unify Complex Data to Gain Business Insights

JACK LEE

The Modern Data Estate: Foundation to Accelerate Digital Transformation

Introduction

Data is a dependable strategic asset. As we navigate these unprecedented times, businesses that can harness the power of data are better equipped, more resilient and can take action to emerge stronger.

Businesses can transform their fragmented data infrastructure to deliver exceptional customer experiences by unlocking their data's value and turning it into insights. In order to deliver the best products and services, organisations must take on a strategy that engages customers, empowers employees, optimises operations and transforms products. This is known as the four pillars of digital transformation, as seen in *Figure 1*:

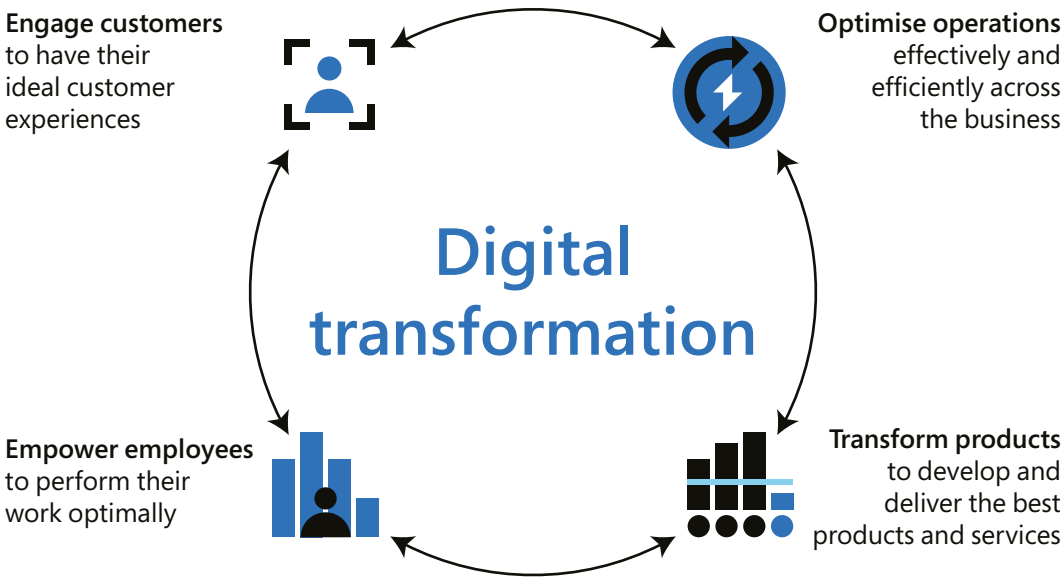

Figure 1: The four pillars of digital transformation

To drive key business outcomes, such as revenue and customer satisfaction, businesses need to ensure that the underlying digital infrastructure enables actionable insights. This can be achieved by bringing disparate data sources together, and enriching data estates with analytics. Alongside this work, it is essential that effective data security, monitoring and compliance are in place. The insights from a modern data estate can then be activated across the organisation, enabling business stakeholders to accelerate decision making and optimise resources.

What is a data estate?

A data estate is the collection of data across operational databases, analytics, AI and data governance on your data platform.

Organisations produce, consume and analyse data every single day. Data can come from a myriad of sources and environments in various shapes and formats. Many organisations have their data stored in legacy database systems, data warehouses or in data lakes. As the volume, velocity and variety of data grow, it becomes a daunting task to attempt to draw valuable insights and make informed decisions from data. To keep up with these demands, organisations need a faster and more scalable data solution, which can only be achieved by migrating and modernising their legacy data estate. Modern database technologies provide the ability to capture and transform any type of data, as well as ensuring availability, security and performance.

As we will explain later, by migrating and modernising your legacy data estate to Azure, you will gain access to the latest cloud-based data services, enabling your organisation to quickly derive important insights to fuel business innovations. These innovations include optimising business operations, becoming more engaged with customers and improving the overall products and services.

In the next section, we will look at what is involved in migrating and modernising your data estate.

Benefits of migrating and modernising your data estate

Modernising a data estate is a continual process of making progressive changes to the way you collect, store and manage your data. The benefit of data estate modernisation is that it allows your organisation to take full advantage of new technologies and unlock new insights to drive your digital transformation initiatives.

Some of the common business drivers and justifications for migrating and modernising your data estate include:

- **End of support for legacy database applications**: When future software and security updates are in doubt or have been discontinued, you need to upgrade to modern database software where software and security updates continue to be available.

- **Cost savings and data centre consolidation**: To lower costs and spending, your organisation has decided to consolidate the on-premises data centres. By migrating to Azure, you can drastically lower your total cost of ownership (TCO).
- **Business agility**: To maintain competitive advantage, organisations need to be able to continuously evolve, innovate and adapt rapidly. By migrating to Azure, you will be able to leverage the latest innovations in cloud technologies.

Azure provides organisations with many options for migrating and modernising data. We will look at some of these options in the next section.

Options for migrating and modernising your data estate on Azure

You can migrate your servers, databases and applications to Azure and modernise your data with various Azure data offerings.

Here are a few approaches organisations can take to migrate their data from legacy data platforms to Azure.

Azure SQL Database	• Azure SQL Database is a fully managed cloud-based relational database service. • It offers the broadest SQL Server engine compatibility and industry-leading TCO. • With built-in AI and high availability, it can maintain peak performance and durability with a Service Level Agreement (SLA) of up to 99.995%. Further reading: • Azure SQL Database https://azure.microsoft.com/products/azure-sql/database/

Azure SQL Managed Instance	• Azure SQL Database Managed Instance is an expansion of the existing SQL Database service. • With Azure SQL Database Managed Instance, databases can be easily migrated with little to no code changes and near-zero downtime. • It is designed to enable friction-free database migration to a fully managed database service, with capabilities such as automatic patching and version updates, automated backups, high availability and reduced management overhead. • Customers can save up to 93% compared to Amazon Web Services (AWS) when migrating to Azure SQL Managed Instance (read the report). Further reading: • Azure SQL Managed Instance https://azure.microsoft.com/products/azure-sql/managed-instance/#overview
SQL Server on Azure Virtual Machines	• Organisations with unique requirements may choose to run SQL Server on Azure Virtual Machines. • This includes organisations that require a full operating system and database control or have special BI requirements such as SQL Server Reporting Services or ISV certification. • Customers can choose between Windows VM and Linux VM to suit their environments. • Customers can save up to 43% compared to AWS when migrating to SQL Server on Azure Virtual Machines. Further reading: • SQL Server on Azure Windows virtual machines https://azure.microsoft.com/products/virtual-machines/#overview • SQL Server on Azure Linux virtual machines https://azure.microsoft.com/products/virtual-machines/sql-server/

Azure Cosmos DB	• Azure Cosmos DB is a globally distributed NoSQL database. It is for building fast, scalable applications with native support for NoSQL. • With globally distributed data, applications can transparently replicate data anywhere in the world, enabling a fast response time and high availability. • It offers turnkey global distribution across any number of Azure regions by transparently scaling and replicating data wherever the users are. • Organisations pay for what they use and are supported by comprehensive service level agreements. Further reading: • Azure Cosmos DB https://azure.microsoft.com/products/cosmos-db/
Open-Source Software (OSS) • MySQL • PostgreSQL • MariaDB	• Organisations using open-source databases can move to the Microsoft fully managed Azure offerings for MySQL, PostgreSQL and MariaDB. • Azure Database Migration Service eases migrations from on-premises to these cloud offerings, helping organisations move to Azure. • With these OSS offerings, customers can save up to 50% on high availability compared to AWS. Further reading: • Azure Database for MySQL https://azure.microsoft.com/products/mysql • Azure Database for PostgreSQL https://azure.microsoft.com/products/postgresql • Azure Database for MariaDB https://azure.microsoft.com/products/mariadb

Hybrid deployment option	• The hybrid deployment option is ideal for organisations that prefer to manage their data estate in a hybrid fashion, with some server applications running on-premises and some in the cloud. • Microsoft offers the most consistent set of products and services, enabling organisations to seamlessly manage their data in both environments. • Microsoft SQL Server 2022 can meet the needs of modernising your on-premises data estate and is a step toward modernising to Azure. Further reading: • Microsoft SQL Server 2022 https://www.microsoft.com/sql-server/sql-server-2022
Azure Arc-enabled data services	• Azure Arc is a bridge that extends the Azure platform to help you build applications and services with the flexibility to run across data centres, at the edge and in multicloud environments. • Azure Arc-enabled data services extend Azure capabilities to any infrastructure across multicloud, on-premises and edge environments. • Develop cloud-native applications with a consistent development, operations and security model. • Azure Arc runs on both new and existing hardware, virtualisation and Kubernetes platforms, IoT devices and integrated systems. • Do more with less by leveraging your existing investments to modernise with cloud-native solutions. Further reading: • Azure Arc-enabled data services • https://azure.microsoft.com/products/azure-arc/hybrid-data-services/

Azure Synapse Link for SQL	• Azure Synapse Link for SQL enables near real-time analytics over operational data in Azure SQL Database or SQL Server 2022. • Seamlessly integrates between operational stores including Azure SQL Database and SQL Server 2022 and Azure Synapse Analytics. • Enables you to run analytics, business intelligence and machine learning scenarios on your operational data with minimum impact on source databases with a new change feed technology. Further reading: • Azure Synapse Link for SQL https://learn.microsoft.com/azure/synapse-analytics/synapse-link/sql-synapse-link-overview
Azure Synapse Analytics	• Azure Synapse Analytics is a limitless analytics service that brings together data integration, enterprise data warehousing and big data analytics. • It gives you the freedom to query data on your terms, using either serverless or dedicated options – at scale. Azure Synapse brings these worlds together with a unified experience to ingest, explore, prepare, transform, manage and serve data for immediate BI and machine learning needs. Further reading: • Azure Synapse Analytics https://azure.microsoft.com/products/synapse-analytics/

In the next section, we will talk about tools that help you streamline your database migration process.

Orchestrating your data migration to Azure

Azure Migrate is a service that provides a central hub of tooling for data centre migration to Azure that includes tools for apps, data and infrastructure migrations. For data migration, the Azure Database Migration Service is a built-in tool within Azure Migrate that can help streamline your database migration to Azure.

Azure Migrate

To plan and accelerate your cloud migration decision-making process, the business case capability in Azure Migrate helps estimate cost savings and ROI of migrating your Windows Server and SQL Server estate. Business assessments are built into Azure Migrate. You can model your own plan and budget with just a few clicks to quickly understand how Azure can bring the most value to your business.

Few key highlights of this capability include:

- On-premises versus Azure Total Cost of Ownership (TCO) analysis for IaaS and/or PaaS migrations, including year on year cashflow/ROI views.
- Long-term cost savings by moving from a capital expenditure model to an operating expenditure model (i.e., pay-for-use) including cost savings offers like Azure Hybrid Benefit.
- Resource utilisation-based insights to identify servers and workloads that are ideal for cloud, including right-sized recommendations on Azure.
- Quick wins for migration and modernisation including Windows Server and SQL Server 2012 end of support workloads.

Azure Database Migration Service

When you are ready to start your migration process, Azure Database Migration Service will execute all the necessary steps for you. This tool gives you peace of mind, knowing that the process will fully take advantage of Microsoft's best practices.

Using Azure Database Migration Service, you can modernise your existing on-premises data estate to the following database service equivalents on Azure:

- Azure SQL
- SQL Server on Azure Virtual Machines
- Azure Cosmos DB
- Azure Database for MySQL
- Azure Database for PostgreSQL
- Azure Database for MariaDB

The following diagram illustrates how you can modernise your legacy data estate to Azure:

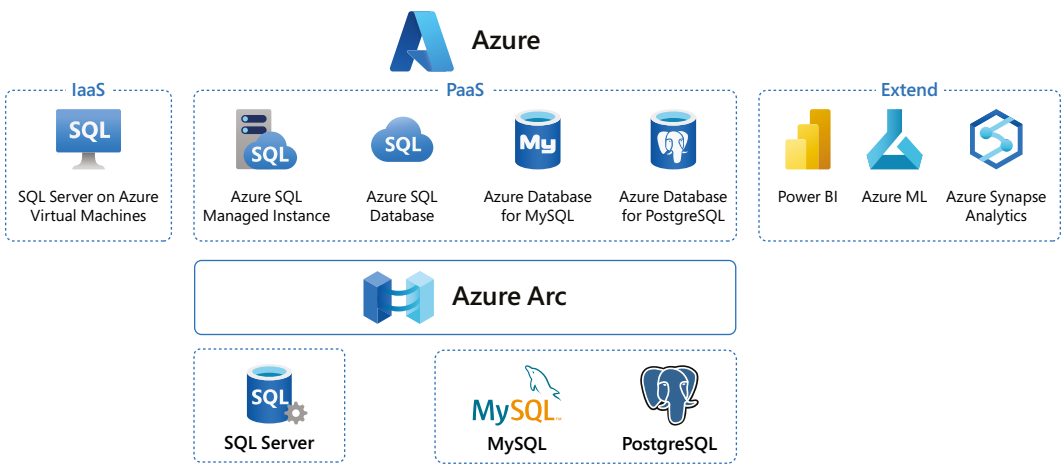

Figure 2: Modernising your legacy data estate to Azure

Azure SQL Migration extension for Azure Data Studio

Azure Data Studio is another tool that can be considered for your data migration needs. Azure Data Studio is built for data professionals who use SQL Server and Azure database on-premises or in multicloud environments. The Azure SQL Migration extension for Azure Data Studio allows you to assess, attain right-sized SKU recommendations based on your Azure SQL target and migrate your SQL Server databases to Azure.

Azure Migration and Modernisation Programme

If you prefer a more guided approach to migration, you can join the Azure Migration and Modernisation Programme. You will get the help you need, save money and bring your workloads to Azure with confidence. The Azure Migration and Modernisation Programme provides the following benefits:

- Step-by-step expert guidance from FastTrack for Azure Engineers and expert partners.
- Gain new skills and learn how to use Azure with technical training for platform fundamentals and advanced migration or modernisation topics.
- Free Azure migration tools to plan, move your workloads and optimise cost and performance.
- Offers to lower your costs.

With the Azure Migration and Modernisation Programme, you can bring your project to success with the right combination of resources customised to your business needs.

Further reading :

- Azure Database Migration Guides – step-by-step guidance for modernising your data assets https://learn.microsoft.com/data-migration/
- Azure Database Migration Service https://azure.microsoft.com/products/database-migration/
- Azure Migrate https://azure.microsoft.com/products/azure-migrate/
- Azure Migration and Modernisation Programme https://azure.microsoft.com/solutions/migration/migration-modernization-program/
- Azure Data Studio https://azure.microsoft.com/products/data-studio/
- Azure SQL Migration extension for Azure Data Studio https://learn.microsoft.com/sql/azure-data-studio/extensions/azure-sql-migration-extension

With your legacy databases in your data estate migrated and modernised, you have now enabled your organisations to take advantage of modern data analytics tools. In the next section, we will deep dive into producing meaningful insights from your data estate.

Producing meaningful insights from your data estate

A data warehouse is a centralised repository that aggregates different (often disparate) data sources – it is the glue that brings your data estate together. 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 primarily generate analytics, business intelligence and 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 (called cold data).

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

- Bring your data together from multitudinous data sources
- Highly scalable and available
- Provides insights from analytical dashboards in real time
- Supports machine learning environments

The one product that makes this possible is Azure Synapse Analytics:

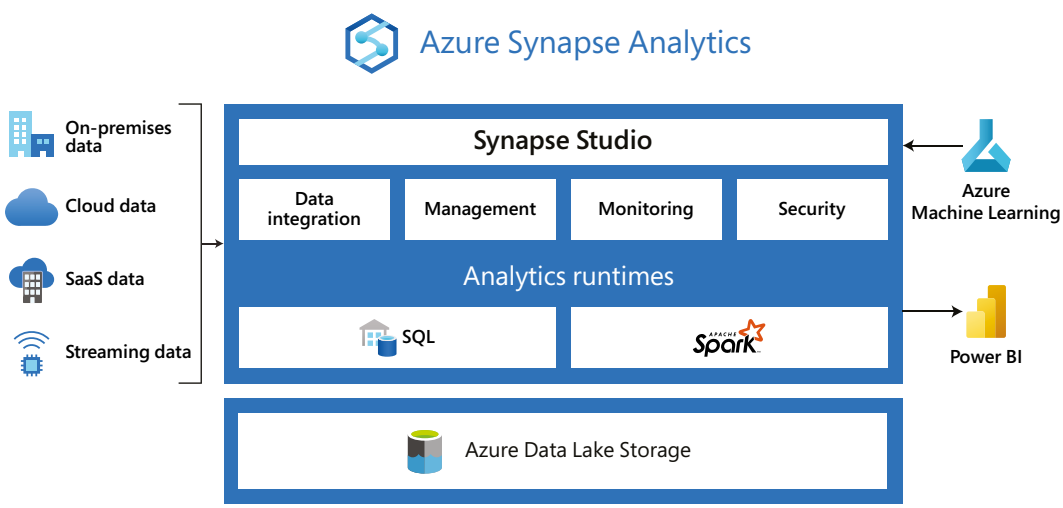

Figure 3: Modern data warehouse using Azure Synapse Analytics

As shown in the preceding diagram, Azure Synapse Analytics enables you to produce meaningful insights from your data estate by bringing together data integration, enterprise data warehousing and big data analytics in a unified experience. You can query data at scale using serverless or dedicated options. Azure Synapse Analytics allows you to ingest, explore, prepare, transform, manage and serve data to meet your BI and machine learning requirements.

In the next section, we will further expand on the idea of bringing your data together in your modern data estate.

Bringing your data together

Traditionally, databases were often the only source of data for your applications. Today, you can have hundreds or even thousands of different data sources in your data estate. The data coming from these different sources are of different data types – some structured, some unstructured and some semi-structured. We will review each of these data types in the following section.

Structured data

The word 'structured' suggests that there is a pattern that can be easily interpreted. Structured data usually comes with a predefined set of models and a schema. A relational database management system (RDBMS) such as Azure SQL Database is a prime example of a database solution that is structured. Your database schema and table columns will determine the structure in which your data is stored.

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

Exploration for your structured data tends to be restrictive because you can only base it on a dataset that is structured and with a semantic form. What if you want to consider other data sources that are unstructured, such as free-form text? One example would be 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

Unstructured data is more or less 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-delimited.

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.

Semi-structured data

IoT devices and sensors, similar to 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 only make a lot of sense once you do some analysis of them. The primary focus is to discern patterns and anomalies. For example, you may want to identify trends in news and topics and remove irrelevant data, such as tweets coming from fake accounts.

Now that we have established a firm understanding of our data sources, we will look at creating a data pipeline to leverage the data in our data estate next.

Creating a data pipeline

Once you have identified your data sources, the next step is creating a data pipeline (sometimes 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:

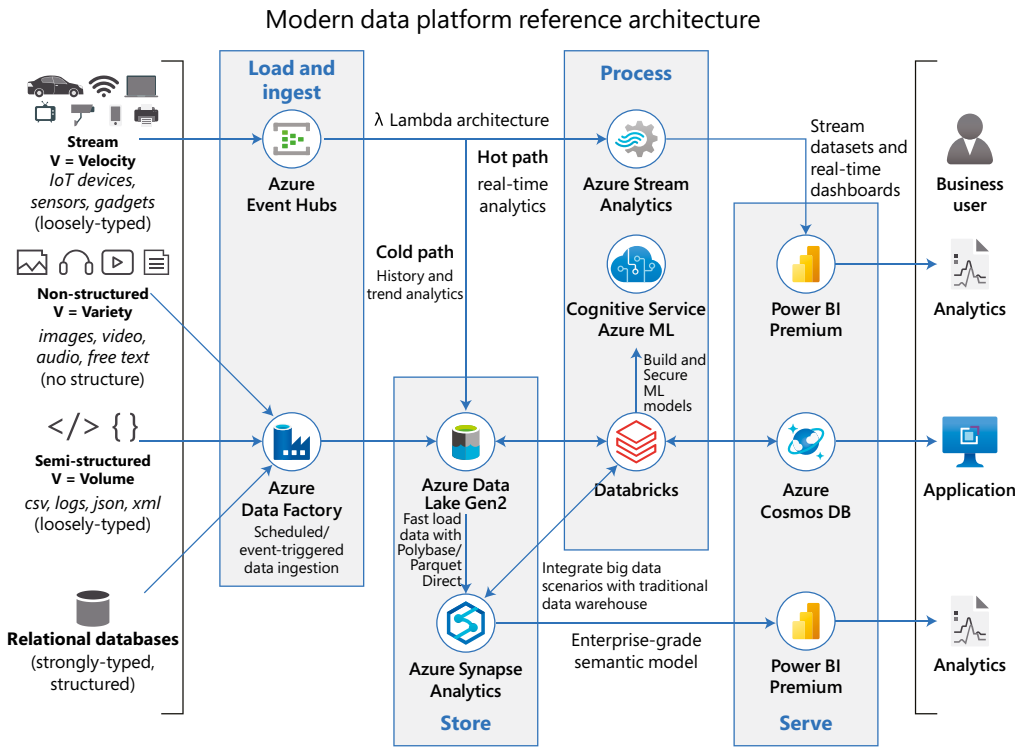

Figure 4: Modern data pipeline

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, to analysts who interpret the data for the business and even the CEO if they want to see what's happening with the business in real time.

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 Azure Synapse Analytics using data integration to transfer data from various sources, such as on-premises databases or 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 data lake will still be in a raw format and include both structured and unstructured data formats.

Data pipeline orchestration and monitoring

In a modern data warehouse scenario, it is 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. Azure Synapse Analytics does not perform the transfer operation, 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 when a data transfer from the source to the destination is 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 with 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 it 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.

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.

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 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.

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

Business intelligence uses data visualisation to analyse performance through graphs and charts. Tools like Power BI help analysts to get the most out of data. Working with Azure, you are not just limited to Power BI, but can also use third party 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 make data-driven decisions. Analysts can use desktop or web application tools to create meaningful representations of their data.

Once the reports are generated, they can be exported to a workspace where people can work together to improve the reports. Users can add comments and annotations to the report, allowing a faster feedback loop for analysts. 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.

In the next section, we will review two real-world business use cases to illustrate high-level solutions using the Azure data platform.

Rationalising and unifying a complex data estate

In this section, we will demonstrate how a complex data estate can be rationalised into a unified solution using Azure through the lens of two business use cases. The aim is to explain how real-time data can be analysed in Azure to derive meaningful insights to drive business decisions.

The company names used here are fictional 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 use case focuses on helping a company gain actionable insights from its data in near real time. The second use case 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 a busy airport. For each use case, 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.

Business use case 1: Migrating apps and data for cost savings and business agility with Azure databases

Contoso Corporation is a fictional global retail company that has stores in North America, Europe and Asia. The company sells consumer goods, electronics and personal care items through its brick-and-mortar stores and digital online channels (mobile and web applications).

Contoso has appointed a new CEO who is passionate about data and they have 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 challenges

- Contoso is currently using a legacy on-premises database system. The vendor of this legacy database system has discontinued this product and it has already announced that support will be ending in less than 90 days.
- Contoso 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, Contoso's 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.
- Contoso 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.
- Contoso 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.

The solution

The Contoso data team decided to use Azure for its scalability, compliance and regional availability.

One of the most pressing concerns that Contoso is facing is its discontinued legacy on-premises database system. Contoso can take advantage of this situation by modernising its legacy database system. The most cost-effective solution is to migrate this legacy system to Azure. Fortunately, with Azure Migrate, the Contoso team can easily migrate its legacy system to Azure SQL Database. Azure Migrate is a central hub for migrating existing databases, apps and infrastructure to Azure. Using the Data Migration Assistant, which is part of Azure Migrate, the Contoso team can first analyse the existing legacy database system for compatibility and identify any potential roadblocks before proceeding with the migration process. The Data Migration Assistant generates an assessment report with recommendations that the team can use to remediate any blockers before the migration takes place. When the team is ready to migrate, simply use Azure Database Migration Service to start the migration process.

With the database migration solution in place, the team then worked on the following solution architecture to address its remaining challenges:

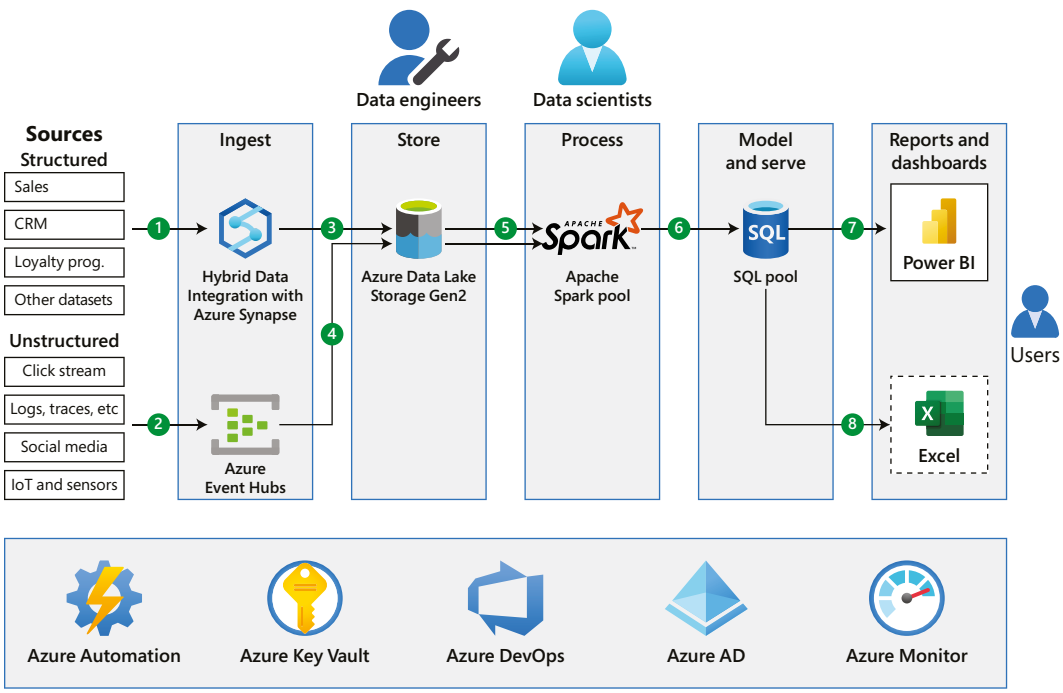

Figure 5: Contoso's solution architecture

We will explain each of the workflow segments numbered in *Figure 4*:

1. Contoso needs to ingest many varying datasets. Some of these contain structured data and others have unstructured data. Worth mentioning is that Contoso successfully migrated their existing RDBS data from its discontinued legacy on-premises database system to Azure SQL Database using Azure Migrate. Contoso will use Azure Synapse Analytics' capability for hybrid data integration. This enables Contoso 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, Contoso is going to use Azure Event Hubs 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 Contoso's analytics solution and enables Contoso 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 Contoso's digital channels), social media feeds (from Twitter, Facebook and so on), logs and trace information from Contoso's servers and any data coming from IoT sensors.
3. All ingested data will end up in Azure Data Lake Storage Gen2, which will serve as a central hub for all data across the organisation.
4. Once the unstructured data is ingested into Azure Event Hubs, it will be pushed to Azure Data Lake Storage Gen2 for permanent storage. This is done using the Azure Event Hubs data capture feature, which allows us to store streaming data in the data lake quickly and easily.
5. Contoso's data will land the data lake. This data is coming from different sources, with various quality levels and different granularities. Thus, Contoso's 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 Contoso's data engineers can easily connect to and explore the data in Azure Data Lake Storage Gen2 without having to move the data to any other system.

6. While the Contoso team is cleaning and preparing data in the Spark pool of Azure Synapse Analytics, 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 Contoso team is planning to use the data warehouse (SQL pool) of Azure Synapse Analytics. This allows the Contoso data team to unite, model and prepare all their data for consumption by business users. This not only enables Contoso 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 Contoso not only to publish reports and dashboards for Contoso users; 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. Contoso can use composite data models for large datasets, which is a feature of Power BI Premium.
8. Contoso invested in complex models using Microsoft Excel. Some of Contoso's data analysts would like to use Microsoft Excel to access data from Azure Synapse Analytics. This is supported out of the box in both Microsoft Excel and Azure Synapse Analytics.

Summary

You have seen how companies like Contoso can modernise their legacy database systems and migrate to Azure using migration tools offered by Azure Migrate. Azure Database Migration Service enables organisations to migrate their existing databases to Azure SQL. With the data estate modernised, companies can take advantage of modern technologies such as Azure Synapse Analytics to discover customer insights and add value in near real time. Azure Synapse Analytics delivers 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, Azure offers many data and analytics services as fully managed offerings, which means less overhead. The other advantage of using fully managed services 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, 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.

Business use case 2: Extending your modern data estate for analytics and intelligence

Fictional company **Fabrikam International Airport** (**FIA**) services 25 million passengers per year, which amounts to 70,000 passengers every day. It sees an average of 200,000 flights every year.

The management team of FIA 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 FIA's ability to fully utilise its infrastructure and resources. In turn, this will improve customer satisfaction and enable FIA to scale its operation by serving more passengers and aeroplanes.

The following sections define the challenges that FIA is facing and brainstorm some design ideas. We will create a possible solution architecture on Azure that can solve these challenges and illustrate why Azure is the perfect platform for the solution.

Business challenges

In the past year, FIA suffered a loss in revenue due to operational inefficiencies. This included costs of flight delays, long passenger queues, poor passenger experience, inefficient planning of staffing levels, under-utilisation of airport assets and lost retail opportunities.

FIA's CIO has a vision for turning this massive inefficiency at FIA into a business opportunity. She presented a business case to the FIA 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, the CIO 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 FIA 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 FIA makes these decisions based on assumptions and previous experiences that do not necessarily reflect reality. So far, FIA 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, FIA 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 FIA close to USD 160 million last year. Add to those 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 FIA'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 FIA staff to different areas of the airport. FIA's management wants to improve resource allocation, which would then improve customer satisfaction.
- 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. FIA thus needs to update the flight status/delays in near real time. This requires FIA'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 overall customer service.
- FIA 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, FIA 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 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 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 FIA 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 FIA's CIO is trying to solve is the fact that FIA 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. FIA has more than 35 airline companies, each of which has its own systems and uses different terminology. This makes it extremely difficult for the FIA management to get credible reports in time, let alone have data-driven operations.
- **Latency in obtaining data and reports**: Because FIA 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 FIA 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 FIA and almost all other organisations.
- **Data availability and access**: Innovation requires exploring possibilities and experimenting with options. In terms of data, this requires FIA to continuously explore, enrich and correlate flight and passenger data with external data sources. Unfortunately, FIA cannot do any of these things today because the data is sitting in many silo systems that the airport does not control.
- **Scalability**: FIA currently has a 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 data warehouse does not have the ability to ingest and/or hold all data that FIA can collect.
- **Security**: FIA 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. FIA 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. FIA 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. FIA 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, FIA's business intelligence team, along with the CIO, agree to define the problem statement as follows:

FIA 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 FIA business intelligence team will work to deliver data analytics tools (dashboards, reports and apps) that help the business optimise operations and remove inefficiencies.

The solution

FIA's CIO with the help of the business intelligence team, picked Azure as the cloud provider to build the new solution. The following is the solution architecture that the team has come up with:

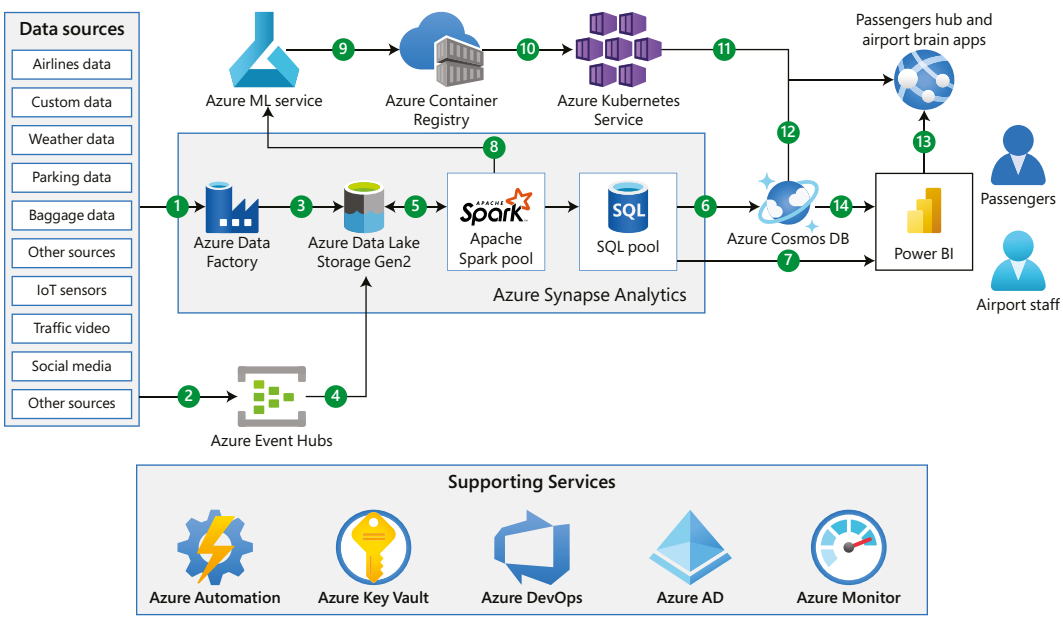

Figure 6: Solution architecture for FIA

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

1. FIA has recently started migrating some of its databases from its on-prem SQL Server to **Azure SQL Managed Instance** which is fully managed and has near 100% compatibility with the latest SQL Server database engine. Structured data, such as Airlines Data, Custom Data and Baggage Data, will be 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 FIA to configure an integration runtime that can be used as a gateway to connect to FIA'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 Hubs**.
3. **Azure Data Factory** pushes the ingested structured data to **Azure Data Lake Storage Gen2** for storage.
4. The FIA data team is adopting Azure Event Hubs 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 Hubs'** data capture feature to store the data to the data lake.
5. Using notebooks, FIA data engineers can then read, clean, transform and model the data using Apache Spark's runtime on **Azure Synapse Analytics**.
6. Cleaned data that is ready for consumption is then pushed from **Azure Synapse Analytics** 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 Analytics. 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 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 machine learning models include a recommendation engine for gate assignments and a parking demand forecasting machine learning model.
11. The FIA 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 devices 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 FIA'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.

Summary

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

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 why Azure is the perfect platform for running such workloads. Azure is affordable and secure and provides organisations with the ability to be agile and scalable.

Conclusion

Data has become an important strategic asset, especially in the aftermath of the recent global pandemic. As we navigate these unprecedented times, businesses that can harness the power of data are better equipped to take appropriate actions and emerge stronger.

Many organisations have their data estates stored in legacy database systems or in data lakes. Deriving insights from legacy data estates is a daunting task. As we learned in this book, to keep up with these demands, organisations need a faster and more scalable data solution, which can only be achieved by migrating and modernising their legacy data estate.

Database management has become top of mind for almost every organisation where data is fluidly moving across hybrid and multicloud environments. With continued innovations across performance, security and availability, Azure unifies operational databases, analytics and data governance. This allows organisations to gain near real-time insights and achieve greater impacts.

We've explained how migrating and modernising your legacy data estate to Azure will gain you access to the latest cloud-based data services and help you grow. This enables organisations to quickly derive important insights and fuel business innovations. These innovations include optimising business operations, becoming more engaged with customers and improving overall products and services.

Resources

Here are some resources to help you get started on your data estate migration and moderation journey:

- Azure SQL Database
  https://azure.microsoft.com/products/azure-sql/database/
- Azure SQL Managed Instance
  https://learn.microsoft.com/azure/azure-sql/managed-instance/sql-managed-instance-paas-overview
- SQL Server on Azure Windows virtual machines
  https://learn.microsoft.com/azure/azure-sql/virtual-machines/windows/sql-server-on-azure-vm-iaas-what-is-overview
- SQL Server on Azure Linux virtual machines
  https://learn.microsoft.com/azure/azure-sql/virtual-machines/linux/sql-server-on-linux-vm-what-is-iaas-overview
- Azure Cosmos DB
  https://azure.microsoft.com/products/cosmos-db/
- Azure Database for MySQL
  https://azure.microsoft.com/products/mysql
- Azure Database for PostgreSQL
  https://azure.microsoft.com/products/postgresql/
- Azure Database for MariaDB
  https://azure.microsoft.com/products/mariadb
- Azure Arc
  https://azure.microsoft.com/products/azure-arc/
- Azure Arc-enabled data services
  https://azure.microsoft.com/products/azure-arc/hybrid-data-services/
- Azure Database Migration Guides – Step-by-step guidance for modernising your data assets https://learn.microsoft.com/data-migration/

- Azure Database Migration Service
  https://azure.microsoft.com/products/database-migration/
- Azure Migrate
  https://azure.microsoft.com/products/azure-migrate
- Azure Migration and Modernisation Programme
  https://azure.microsoft.com/solutions/migration/migration-modernization-program/
- How to migrate your relational databases to Azure
  https://learn.microsoft.com/dotnet/architecture/modernize-with-azure-containers/migrate-your-relational-databases-to-azure
- Azure Data Studio
  https://azure.microsoft.com/products/data-studio/
- Azure SQL Migration extension for Azure Data Studio
  https://learn.microsoft.com/sql/azure-data-studio/extensions/azure-sql-migration-extension
- Azure Synapse Link for SQL
  https://learn.microsoft.com/azure/synapse-analytics/synapse-link/sql-synapse-link-overview
- Azure Synapse Analytics
  https://azure.microsoft.com/products/synapse-analytics/
- Microsoft SQL Server 2022
  https://www.microsoft.com/sql-server/sql-server-2022

Microsoft

Turn data into insights

Get a unified experience across your entire database portfolio and a full range of deployment options. Intelligent and secure Azure database services make it easy to:

Migrate your data workloads with Azure Migrate.

Support modern cloud applications with Azure SQL database.

Modernise your existing applications or enjoy hybrid flexibility with Azure SQL Managed Instance.

Extend your applications to IoT edge gateways and devices with Azure SQL Edge.

Explore Azure Database Migration Guides On Microsoft Learn.

Learn more about saving time and money with the Azure Migration and Modernisation Centre.

Get started with an Azure free account

Find an Azure partner to get expert help

Connect with an Azure sales specialist to learn more