**Get started with data engineering on Azure**

In most organizations, a data engineer is the primary role responsible for integrating, transforming, and consolidating data from various structured and unstructured data systems into structures that are suitable for building analytics solutions. An Azure data engineer also helps ensure that data pipelines and data stores are high-performing, efficient, organized, and reliable, given a specific set of business requirements and constraints.

# What is data engineering

The data engineer will often work with multiple types of data to perform many operations using many scripting or coding languages that are appropriate to their individual organization.

## Types of data

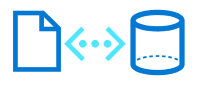
There are three primary types of data that a data engineer will work with.

| **Structured** | **Semi-structured** | **Unstructured** |
| --- | --- | --- |
| Diagram of Structured data type. | Diagram of Semi-structured data type. | Diagram of Unstructured data type. |
| Structured data primarily comes from table-based source systems such as a relational database or from a flat file such as a comma separated (CSV) file. The primary element of a structured file is that the rows and columns are aligned consistently throughout the file. | Semi-structured data is data such as JavaScript object notation (JSON) files, which may require flattening prior to loading into your source system. When flattened, this data doesn't have to fit neatly into a table structure. | Unstructured data includes data stored as key-value pairs that don't adhere to standard relational models and Other types of unstructured data that are commonly used include portable data format (PDF), word processor documents, and images. |

## Data operations

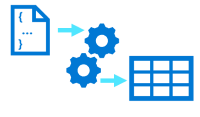
As a data engineer some of the main tasks that you'll perform in Azure include data integration, data transformation, and data consolidation.

### Data integration



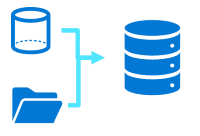
Data Integration involves establishing links between operational and analytical services and data sources to enable secure, reliable access to data across multiple systems. For example, a business process might rely on data that is spread across multiple systems, and a data engineer is required to establish links so that the required data can be extracted from all of these systems.

### Data transformation



Operational data usually needs to be transformed into suitable structure and format for analysis, often as part of an extract, transform, and load (ETL) process; though increasingly a variation in which you extract, load, and transform (ELT) the data is used to quickly ingest the data into a data lake and then apply "big data" processing techniques to transform it. Regardless of the approach used, the data is prepared to support downstream analytical needs.

### Data consolidation



Data consolidation is the process of combining data that has been extracted from multiple data sources into a consistent structure - usually to support analytics and reporting. Commonly, data from operational systems is extracted, transformed, and loaded into analytical stores such as a data lake or data warehouse.

## Common languages

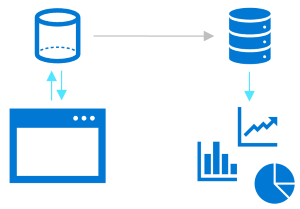
Data Engineers must be proficient with a range of tools and scripting languages - in particular SQL and Python, and potentially others.

* **SQL** - One of the most common languages data engineers use is SQL, or Structured Query Language, which is a relatively easy language to learn. SQL uses queries that include SELECT, INSERT, UPDATE, and DELETE statements to directly work with the data stored in tables.
* **Python** - Python is one of the most popular and fastest growing programming languages in the world. It's used for all sorts of tasks, including web programming and data analysis. It has emerged as the language to learn for machine learning, and is increasing in popularity in data engineering with the use of notebooks.
* **Others** - Depending upon the needs of the organization and your individual skill set, you may also use other popular languages within or outside of notebooks including R, Java, Scala, .NET, and more. The use of notebooks is growing in popularity, and allows collaboration using different languages within the same notebook.

# Important data engineering concepts

There are some core concepts with which data engineers should be familiar. These concepts underpin many of the workloads that data engineers must implement and support.

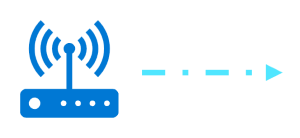
## Operational and analytical data



Operational data is usually transactional data that is generated and stored by applications, often in a relational or non-relational database. Analytical data is data that has been optimized for analysis and reporting, often in a data warehouse.

One of the core responsibilities of a data engineer is to design, implement, and manage solutions that integrate operational and analytical data sources or extract operational data from multiple systems, transform it into appropriate structures for analytics, and load it into an analytical data store (usually referred to as ETL solutions).

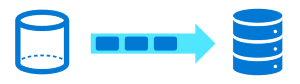
## Streaming data



Streaming data refers to perpetual sources of data that generate data values in real-time, often relating to specific events. Common sources of streaming data include internet-of-things (IoT) devices and social media feeds.

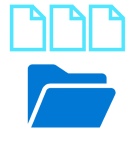
Data engineers often need to implement solutions that capture real-time stream of data and ingest them into analytical data systems, often combining the real-time data with other application data that is processed in batches.

## Data pipelines



Data pipelines are used to orchestrate activities that transfer and transform data. Pipelines are the primary way in which data engineers implement repeatable extract, transform, and load (ETL) solutions that can be triggered based on a schedule or in response to events.

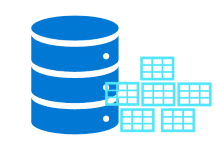
## Data lakes



A data lake is a storage repository that holds large amounts of data in native, raw formats. Data lake stores are optimized for scaling to massive volumes (terabytes or petabytes) of data. The data typically comes from multiple heterogeneous sources, and may be structured, semi-structured, or unstructured.

The idea with a data lake is to store everything in its original, untransformed state. This approach differs from a traditional data warehouse, which transforms and processes the data at the time of ingestion.

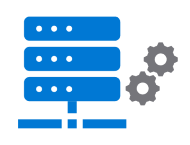
## Data warehouses



A data warehouse is a centralized repository of integrated data from one or more disparate sources. Data warehouses store current and historical data in relational tables that are organized into a schema that optimizes performance for analytical queries.

Data engineers are responsible for designing and implementing relational data warehouses, and managing regular data loads into tables.

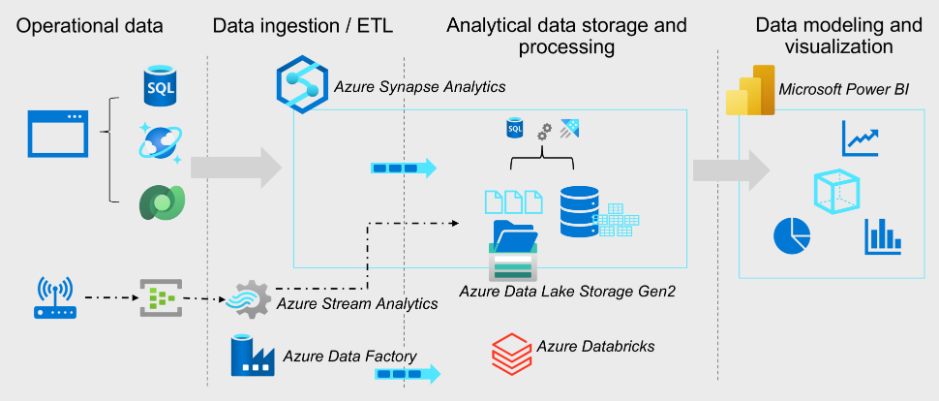
## Apache Spark



Apache Spark is a parallel processing framework that takes advantage of in-memory processing and a distributed file storage. It's a common open-source software (OSS) tool for big data scenarios.

Data engineers need to be proficient with Spark, using notebooks and other code artifacts to process data in a data lake and prepare it for modeling and analysis.

# Data engineering in Microsoft Azure



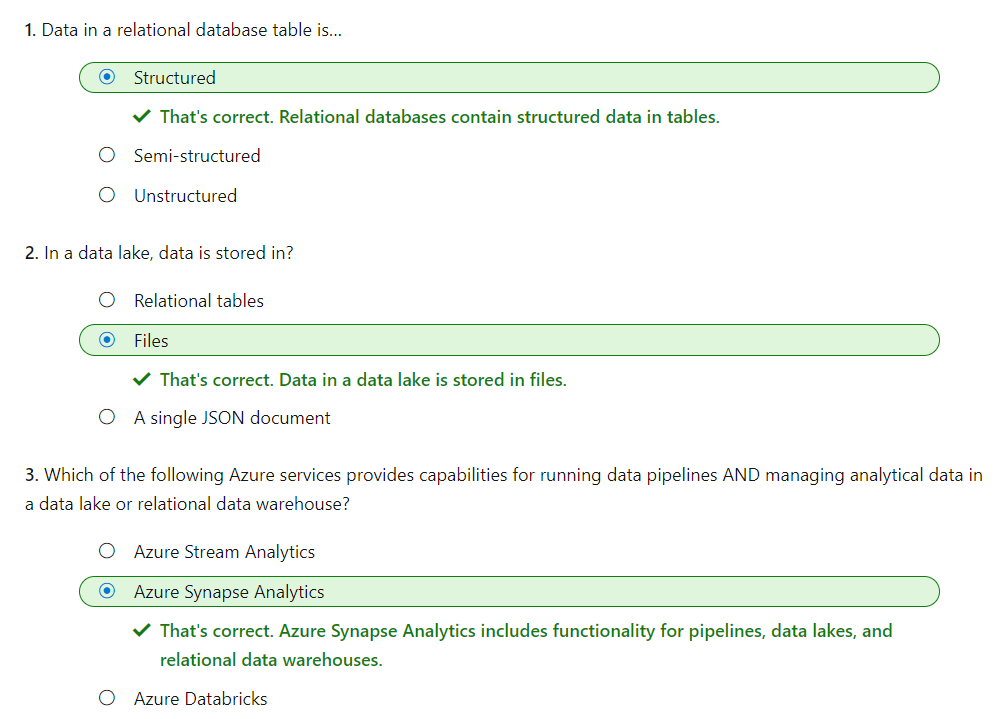
Microsoft Azure includes many services that can be used to implement and manage data engineering workloads.

The diagram displays the flow from left to right of a typical enterprise data analytics solution, including some of the key Azure services that may be used. Operational data is generated by applications and devices and stored in Azure data storage services such as Azure SQL Database, Azure Cosmos DB, and Microsoft Dataverse. Streaming data is captured in event broker services such as Azure Event Hubs.

This operational data must be captured, ingested, and consolidated into analytical stores; from where it can be modeled and visualized in reports and dashboards. These tasks represent the core area of responsibility for the data engineer. The core Azure technologies used to implement data engineering workloads include:

* Azure Synapse Analytics
* Azure Data Lake Storage Gen2
* Azure Stream Analytics
* Azure Data Factory
* Azure Databricks

The analytical data stores that are populated with data produced by data engineering workloads support data modeling and visualization for reporting and analysis, often using sophisticated visualization tools such as Microsoft Power BI.



# Introduction to Azure Data Lake Storage Gen2

Data lakes are a core element of data analytics architectures. Azure Data Lake Storage Gen2 provides a scalable, secure, cloud-based solution for data lake storage.

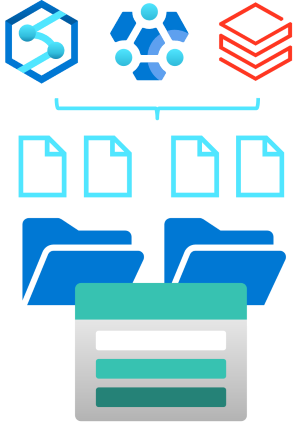
## Learning objectives

In this module you will learn how to:

* Describe the key features and benefits of Azure Data Lake Storage Gen2
* Enable Azure Data Lake Storage Gen2 in an Azure Storage account
* Compare Azure Data Lake Storage Gen2 and Azure Blob storage
* Describe where Azure Data Lake Storage Gen2 fits in the stages of analytical processing
* Describe how Azure data Lake Storage Gen2 is used in common analytical workloads
* Many organizations have spent the last two decades building data warehouses and business intelligence (BI) solutions based on relational database systems. Many BI solutions have lost out on opportunities to store unstructured data due to cost and complexity in these types of data in databases.
* Data lakes have become a common solution to this problem. A data lake provides file-based storage, usually in a distributed file system that supports high scalability for massive volumes of data. Organizations can store structured, semi-structured, and unstructured files in the data lake and then consume them from there in big data processing technologies, such as Apache Spark.
* Azure Data Lake Storage Gen2 provides a cloud-based solution for data lake storage in Microsoft Azure, and underpins many large-scale analytics solutions built on Azure.

# Understand Azure Data Lake Storage Gen2

A data lake is a repository of data that is stored in its natural format, usually as blobs or files. Azure Data Lake Storage is a comprehensive, massively scalable, secure, and cost-effective data lake solution for high performance analytics built into Azure.



Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. Data Lake Storage builds on Azure Blob storage capabilities to optimize it specifically for analytics workloads. This integration enables analytics performance, the tiering and data lifecycle management capabilities of Blob storage, and the high-availability, security, and durability capabilities of Azure Storage.

## Benefits

Data Lake Storage is designed to deal with this variety and volume of data at exabyte scale while securely handling hundreds of gigabytes of throughput. With this, you can use Data Lake Storage Gen2 as the basis for both real-time and batch solutions.

### Hadoop compatible access

A benefit of Data Lake Storage is that you can treat the data as if it's stored in a Hadoop Distributed File System. With this feature, you can store the data in one place and access it through compute technologies including Azure Databricks, Azure HDInsight, and Azure Synapse Analytics without moving the data between environments. The data engineer also has the ability to use storage mechanisms such as the parquet format, which is highly compressed and performs well across multiple platforms using an internal columnar storage.

### Security

Data Lake Storage supports access control lists (ACLs) and Portable Operating System Interface (POSIX) permissions that don't inherit the permissions of the parent directory. In fact, you can set permissions at a directory level or file level for the data stored within the data lake, providing a much more secure storage system. This security is configurable through technologies such as Hive and Spark or utilities such as Azure Storage Explorer, which runs on Windows, macOS, and Linux. All data that is stored is encrypted at rest by using either Microsoft or customer-managed keys.

### Performance

Azure Data Lake Storage organizes the stored data into a hierarchy of directories and subdirectories, much like a file system, for easier navigation. As a result, data processing requires less computational resources, reducing both the time and cost.

### Data redundancy

Data Lake Storage takes advantage of the Azure Blob replication models that provide data redundancy in a single data center with locally redundant storage (LRS), or to a secondary region by using the Geo-redundant storage (GRS) option. This feature ensures that your data is always available and protected if catastrophe strikes.

**Tip**

Whenever planning for a data lake, a data engineer should give thoughtful consideration to structure, data governance, and security. This should include consideration of factors that can influence lake structure and organization, such as:

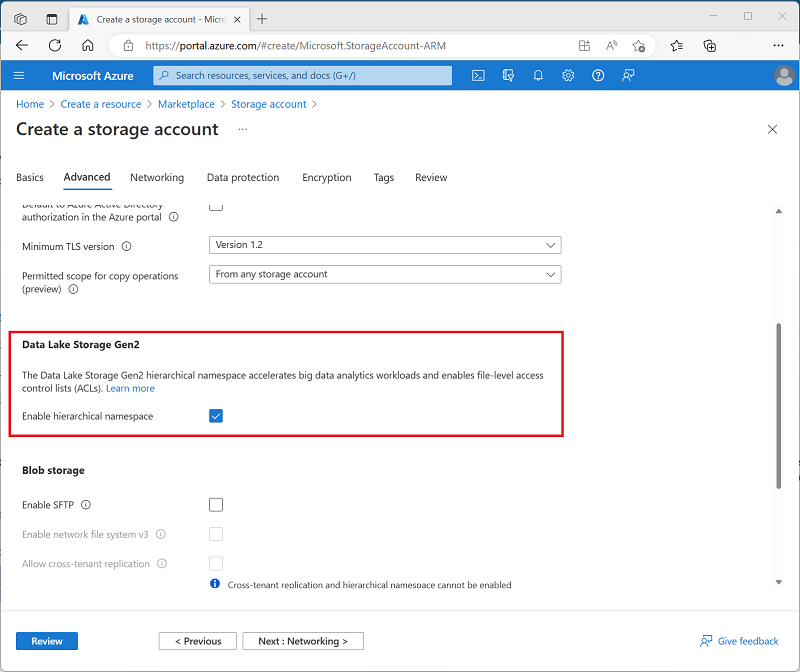
* Types of data to be stored
* How the data will be transformed
* Who should access the data
* What are the typical access patterns

This approach will help determine how to plan for access control governance across your lake. Data engineers should be proactive in ensuring that the lake doesn't become the proverbial data swamp which becomes inaccessible and non-useful to users due to the lack of data governance and data quality measures. Establishing a baseline and following best practices for Azure Data Lake will help ensure a proper and robust implementation that will allow the organization to grow and gain insight to achieve more.

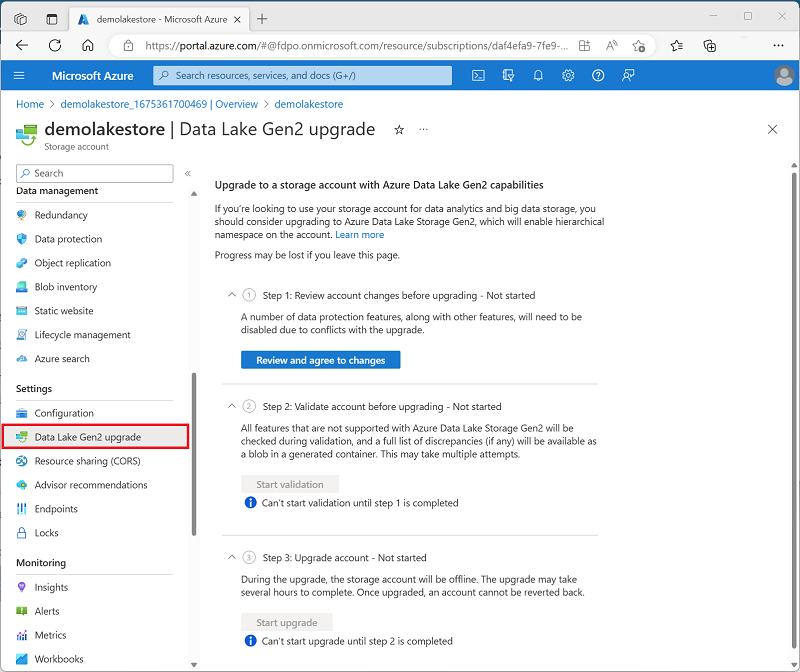
# Enable Azure Data Lake Storage Gen2 in Azure Storage

Azure Data Lake Storage Gen2 isn't a standalone Azure service, but rather a configurable capability of a **StorageV2 (General Purpose V2)** Azure Storage.

To enable Azure Data Lake Storage Gen2 in an Azure Storage account, you can select the option to **Enable hierarchical namespace** in the **Advanced** page when creating the storage account in the Azure portal:

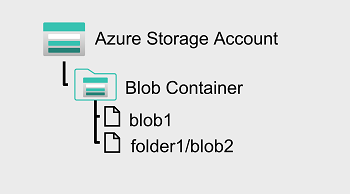


Alternatively, if you already have an Azure Storage account and want to enable the Azure data Lake Storage Gen2 capability, you can use the **Data Lake Gen2 upgrade** wizard in the Azure portal page for your storage account resource.



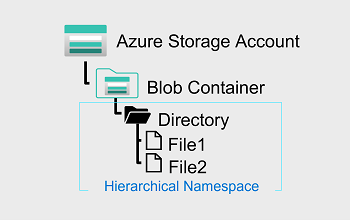
# Compare Azure Data Lake Store to Azure Blob storage

In Azure Blob storage, you can store large amounts of unstructured ("object") data in a flat namespace within a blob container. Blob names can include "/" characters to organize blobs into virtual "folders", but in terms of blob manageability the blobs are stored as a single-level hierarchy in a flat namespace.



You can access this data by using HTTP or HTTPs

Azure Data Lake Storage Gen2 builds on blob storage and optimizes I/O of high-volume data by using a hierarchical namespace that organizes blob data into *directories*, and stores metadata about each directory and the files within it. This structure allows operations, such as directory renames and deletes, to be performed in a single atomic operation. Flat namespaces, by contrast, require several operations proportionate to the number of objects in the structure. Hierarchical namespaces keep the data organized, which yields better storage and retrieval performance for an analytical use case and lowers the cost of analysis.



**Tip**

If you want to store data *without performing analysis on the data*, set the **Hierarchical Namespace** option to **Disabled** to set up the storage account as an Azure Blob storage account. You can also use blob storage to archive rarely used data or to store website assets such as images and media.

If you are performing analytics on the data, set up the storage account as an Azure Data Lake Storage Gen2 account by setting the **Hierarchical Namespace** option to **Enabled**. Because Azure Data Lake Storage Gen2 is integrated into the Azure Storage platform, applications can use either the Blob APIs or the Azure Data Lake Storage Gen2 file system APIs to access data.

# Understand the stages for processing big data

Data lakes have a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:

* An enterprise data warehouse.
* Advanced analytics against big data.
* A real-time analytical solution.

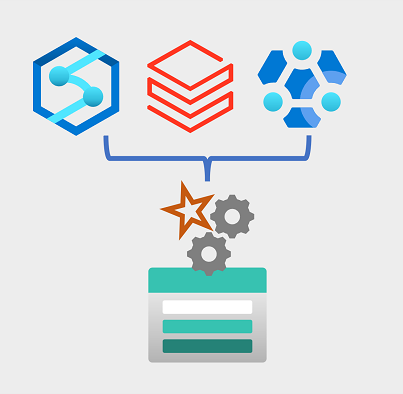
There are four stages for processing big data solutions that are common to all architectures:

* **Ingest** - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the data lake. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, pipelines in Azure Synapse Analytics or Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate choice.
* **Store** - The store phase identifies where the ingested data should be placed. Azure Data Lake Storage Gen2 provides a secure and scalable storage solution that is compatible with commonly used big data processing technologies.
* **Prep and train** - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for machine learning solutions. Common technologies that are used in this phase are Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Machine Learning.
* **Model and serve** - Finally, the model and serve phase involves the technologies that will present the data to users. These technologies can include visualization tools such as Microsoft Power BI, or analytical data stores such as Azure Synapse Analytics. Often, a combination of multiple technologies will be used depending on the business requirements.

# Use Azure Data Lake Storage Gen2 in data analytics workloads

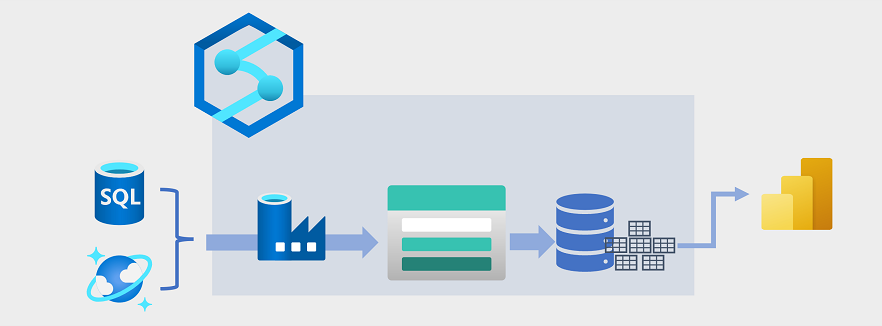
Azure Data Lake Store Gen2 is an enabling technology for multiple data analytics use cases. Let's explore a few common types of analytical workload, and identify how Azure Data Lake Storage Gen2 works with other Azure services to support them.

## Big data processing and analytics



Big data scenarios usually refer to analytical workloads that involve massive volumes of data in a variety of formats that needs to be processed at a fast velocity - the so-called "three v's". Azure Data Lake Storage Gen 2 provides a scalable and secure distributed data store on which big data services such as Azure Synapse Analytics, Azure Databricks, and Azure HDInsight can apply data processing frameworks such as Apache Spark, Hive, and Hadoop. The distributed nature of the storage and the processing compute enables tasks to be performed in parallel, resulting in high-performance and scalability even when processing huge amounts of data.

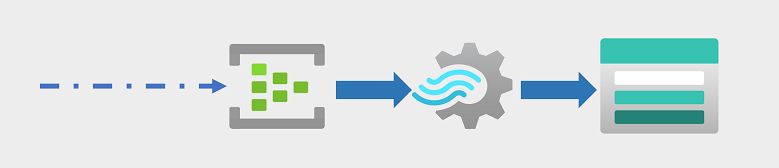
## Data warehousing



Data warehousing has evolved in recent years to integrate large volumes of data stored as files in a data lake with relational tables in a data warehouse. In a typical example of a data warehousing solution, data is extracted from operational data stores, such as Azure SQL database or Azure Cosmos DB, and transformed into structures more suitable for analytical workloads. Often, the data is staged in a data lake in order to facilitate distributed processing before being loaded into a relational data warehouse. In some cases, the data warehouse uses external tables to define a relational metadata layer over files in the data lake and create a hybrid "data lakehouse" or "lake database" architecture. The data warehouse can then support analytical queries for reporting and visualization.

There are multiple ways to implement this kind of data warehousing architecture. The diagram shows a solution in which Azure Synapse Analytics hosts pipelines to perform extract, transform, and load (ETL) processes using Azure Data Factory technology. These processes extract data from operational data sources and load it into a data lake hosted in an Azure Data Lake Storage Gen2 container. The data is then processed and loaded into a relational data warehouse in an Azure Synapse Analytics dedicated SQL pool, from where it can support data visualization and reporting using Microsoft Power BI.

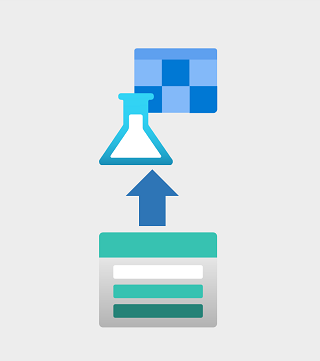
## Real-time data analytics



Increasingly, businesses and other organizations need to capture and analyze perpetual streams of data, and analyze it in real-time (or as near to real-time as possible). These streams of data can be generated from connected devices (often referred to as internet-of-things or IoT devices) or from data generated by users in social media platforms or other applications. Unlike traditional batch processing workloads, streaming data requires a solution that can capture and process a boundless stream of data events as they occur.

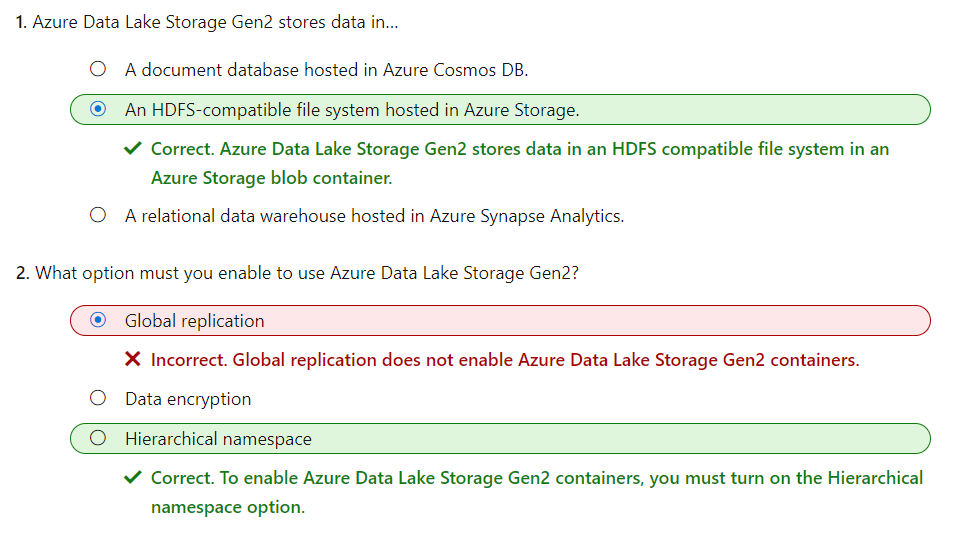
Streaming events are often captured in a queue for processing. There are multiple technologies you can use to perform this task, including Azure Event Hubs as shown in the image. From here, the data is processed, often to aggregate data over temporal windows (for example to count the number of social media messages with a given tag every five minutes, or to calculate the average reading of an Internet connected sensor per minute). Azure Stream Analytics enables you to create jobs that query and aggregate event data as it arrives, and write the results in an output sink. One such sink is Azure Data Lake Storage Gen2; from where the captured real-time data can be analyzed and visualized.

## Data science and machine learning



Data science involves the statistical analysis of large volumes of data, often using tools such as Apache Spark and scripting languages such as Python. Azure Data Lake Storage Gen 2 provides a highly scalable cloud-based data store for the volumes of data required in data science workloads.

Machine learning is a subarea of data science that deals with training predictive models. Model training requires huge amounts of data, and the ability to process that data efficiently. Azure Machine Learning is a cloud service in which data scientists can run Python code in notebooks using dynamically allocated distributed compute resources. The compute processes data in Azure Data Lake Storage Gen2 containers to train models, which can then be deployed as production web services to support predictive analytical workloads.



# Introduction to Azure Synapse Analytics

The volume of data generated by individuals and organizations is growing at a phenomenal rate. This data powers businesses and other organizations by providing a basis for descriptive, diagnostic, predictive, and prescriptive analytical solutions that support decision making and autonomous systems by providing real-time insights into established and emerging patterns.

Organizations have a choice of many tools and techniques for data analytics, often requiring expertise across multiple systems and complex integration of infrastructure and administrative operations. Azure Synapse Analytics provides a single, cloud-scale platform that supports multiple analytical technologies; enabling a consolidated and integrated experience for data engineers, data analysts, data scientists, and other professionals who need to work with data.

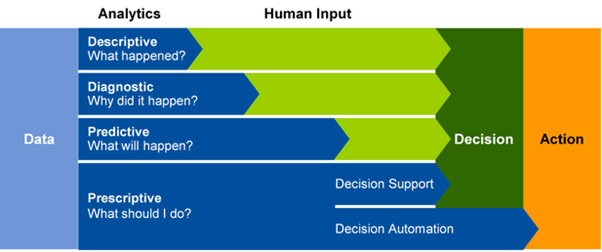
In this module, you'll learn how to:

* Identify the business problems that Azure Synapse Analytics addresses.
* Describe core capabilities of Azure Synapse Analytics.
* Determine when to use Azure Synapse Analytics.

# What is Azure Synapse Analytics

The technological research and consulting firm Gartner defines four common types of analytical technique that organizations commonly use:

* **Descriptive analytics**, which answers the question “What is happening in my business?”. The data to answer this question is typically answered through the creation of a data warehouse in which historical data is persisted in relational tables for multidimensional modeling and reporting.
* **Diagnostic analytics**, which deals with answering the question “Why is it happening?”. This may involve exploring information that already exists in a data warehouse, but typically involves a wider search of your data estate to find more data to support this type of analysis.
* **Predictive analytics**, which enables you to answer the question “What is likely to happen in the future based on previous trends and patterns?”
* **Prescriptive analytics**, which enables autonomous decision making based on real-time or near real-time analysis of data, using predictive analytics.



Azure Synapse Analytics provides a cloud platform for all of these analytical workloads through support for multiple data storage, processing, and analysis technologies in a single, integrated solution. The integrated design of Azure Synapse Analytics enables organizations to leverage investments and skills in multiple commonly used data technologies, including SQL, Apache Spark, and others; while providing a centrally managed service and a single, consistent user interface.

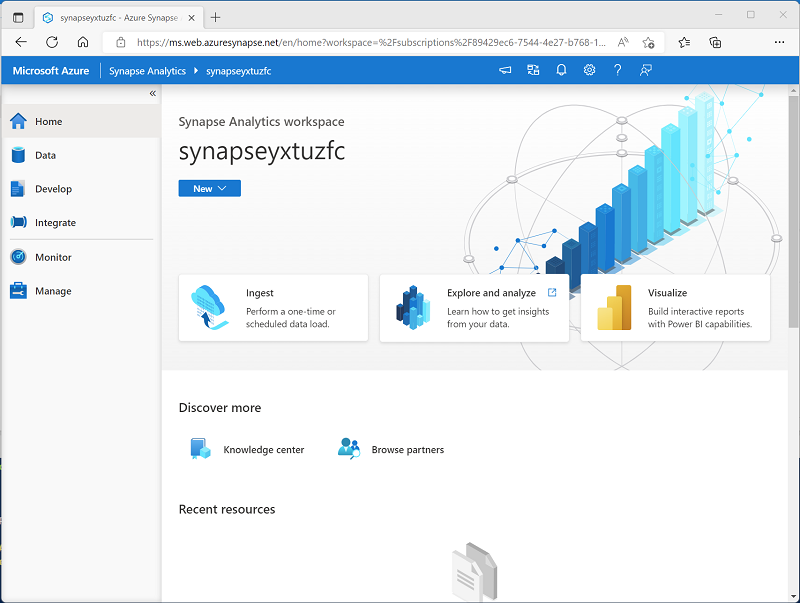
# How Azure Synapse Analytics works

To support the analytics needs of today's organizations, Azure Synapse Analytics combines a centralized service for data storage and processing with an extensible architecture through which linked services enable you to integrate commonly used data stores, processing platforms, and visualization tools.

## Creating and using an Azure Synapse Analytics workspace

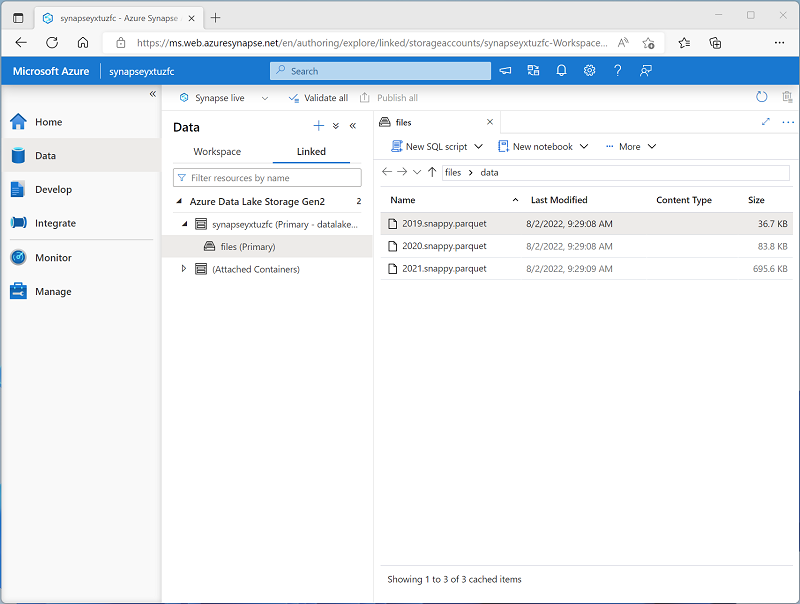
A Synapse Analytics workspace defines an instance of the Synapse Analytics service in which you can manage the services and data resources needed for your analytics solution. You can create a Synapse Analytics workspace in an Azure subscription interactively by using the Azure portal, or you can automate deployment by using Azure PowerShell, the Azure command-line interface (CLI), or with an Azure Resource Manager or Bicep template.

After creating a Synapse Analytics workspace, you can manage the services in it and perform data analytics tasks with them by using Synapse Studio; a web-based portal for Azure Synapse Analytics.

[](https://learn.microsoft.com/en-us/training/wwl-data-ai/introduction-azure-synapse-analytics/media/synapse-studio.png#lightbox)

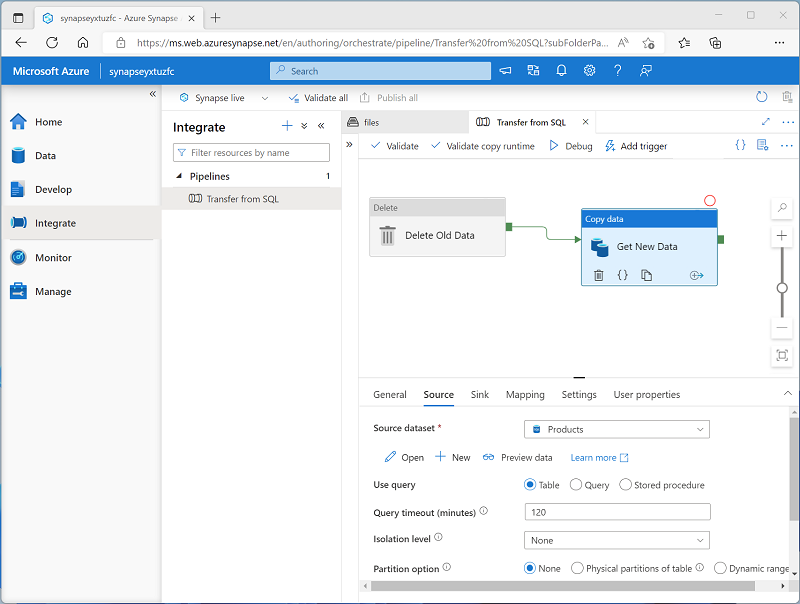
## Working with files in a data lake

One of the core resources in a Synapse Analytics workspace is a data lake, in which data files can be stored and processed at scale. A workspace typically has a default data lake, which is implemented as a linked service to an Azure Data Lake Storage Gen2 container. You can add linked services for multiple data lakes that are based on different storage platforms as required.

[](https://learn.microsoft.com/en-us/training/wwl-data-ai/introduction-azure-synapse-analytics/media/data-lake-store.png#lightbox)

## Ingesting and transforming data with pipelines

In most enterprise data analytics solutions, data is extracted from multiple operational sources and transferred to a central data lake or data warehouse for analysis. Azure Synapse Analytics includes built-in support for creating, running, and managing pipelines that orchestrate the activities necessary to retrieve data from a range of sources, transform the data as required, and load the resulting transformed data into an analytical store.

[](https://learn.microsoft.com/en-us/training/wwl-data-ai/introduction-azure-synapse-analytics/media/synapse-pipeline.png#lightbox)

**Note**

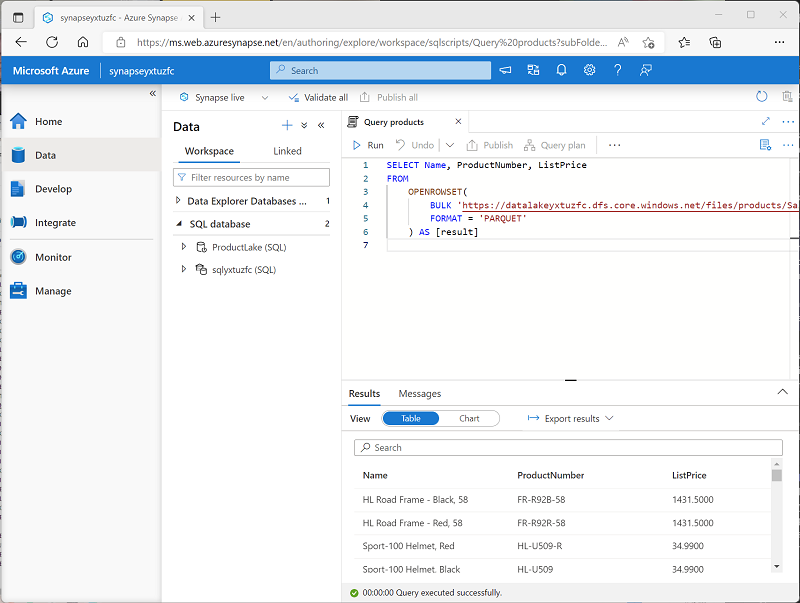
Pipelines in Azure Synapse Analytics are based on the same underlying technology as Azure Data Factory. If you are already familiar with Azure Data Factory, you can leverage your existing skills to build data ingestion and transformation solutions in Azure Synapse Analytics.

## Querying and manipulating data with SQL

Structured Query Language (SQL) is a ubiquitous language for querying and manipulating data, and is the foundation for relational databases, including the popular Microsoft SQL Server database platform. Azure Synapse Analytics supports SQL-based data querying and manipulation through two kinds of SQL pool that are based on the SQL Server relational database engine:

* A built-in serverless pool that is optimized for using relational SQL semantics to query file-based data in a data lake.
* Custom dedicated SQL pools that host relational data warehouses.

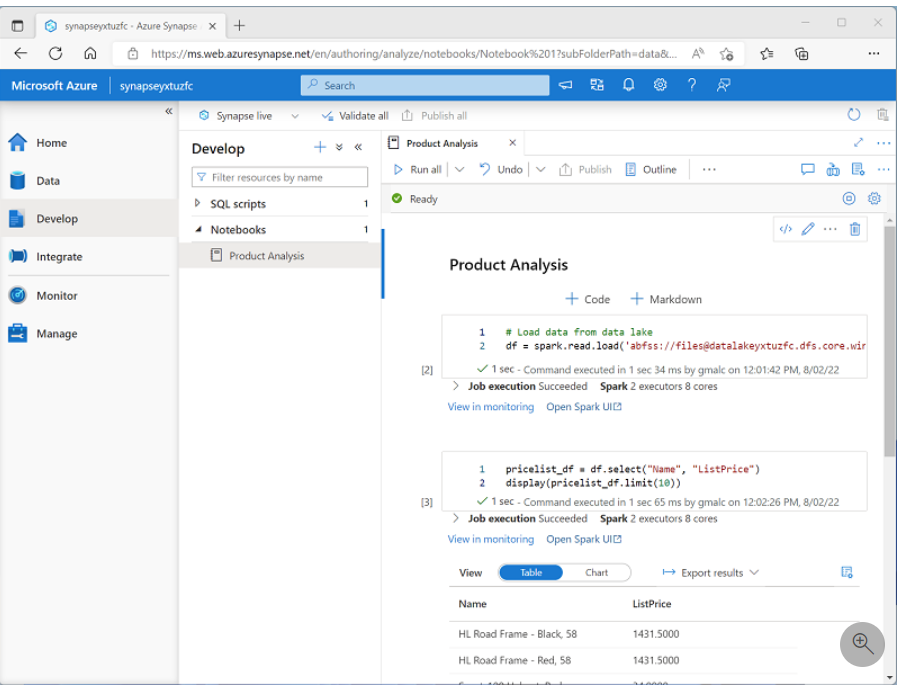
The Azure Synapse SQL system uses a distributed query processing model to parallelize SQL operations, resulting in a highly scalable solution for relational data processing. You can use the built-in serverless pool for cost-effective analysis and processing of file data in the data lake, and use dedicated SQL pools to create relational data warehouses for enterprise data modeling and reporting.

[](https://learn.microsoft.com/en-us/training/wwl-data-ai/introduction-azure-synapse-analytics/media/synapse-sql.png#lightbox)

## Processing and analyzing data with Apache Spark

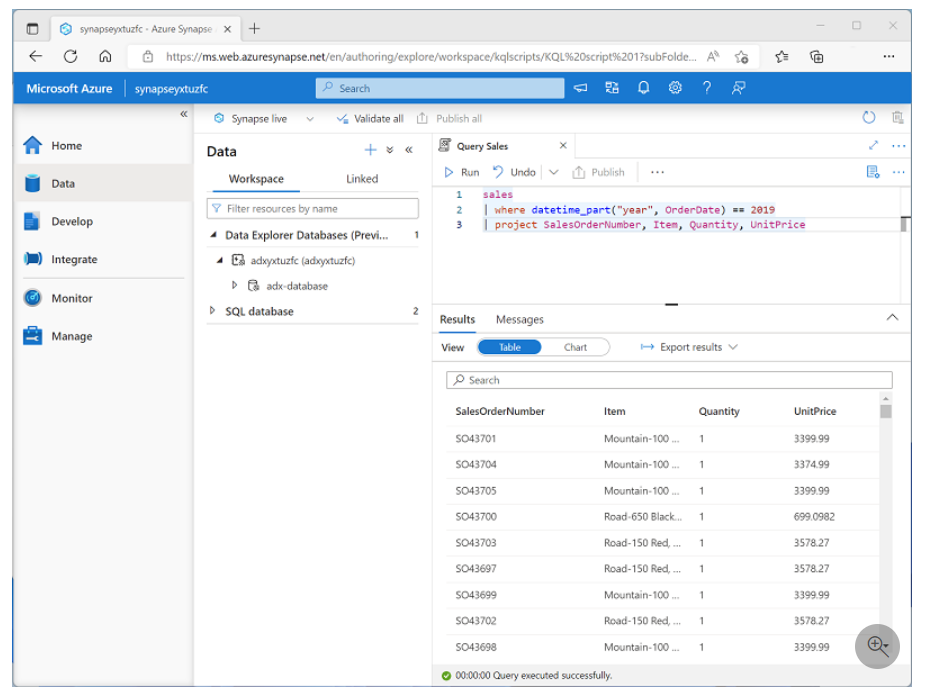
Apache Spark is an open source platform for big data analytics. Spark performs distributed processing of files in a data lake by running jobs that can be implemented using any of a range of supported programming languages. Languages supported in Spark include Python, Scala, Java, SQL, and C#.

In Azure Synapse Analytics, you can create one or more Spark pools and use interactive notebooks to combine code and notes as you build solutions for data analytics, machine learning, and data visualization.



## Exploring data with Data Explorer

Azure Synapse Data Explorer is a data processing engine in Azure Synapse Analytics that is based on the Azure Data Explorer service. Data Explorer uses an intuitive query syntax named Kusto Query Language (KQL) to enable high performance, low-latency analysis of batch and streaming data.



## Integrating with other Azure data services

Azure Synapse Analytics can be integrated with other Azure data services for end-to-end analytics solutions. Integrated solutions include:

* **Azure Synapse Link** enables near-realtime synchronization between operational data in Azure Cosmos DB, Azure SQL Database, SQL Server, and Microsoft Power Platform Dataverse and analytical data storage that can be queried in Azure Synapse Analytics.
* **Microsoft Power BI** integration enables data analysts to integrate a Power BI workspace into a Synapse workspace, and perform interactive data visualization in Azure Synapse Studio.
* **Microsoft Purview** integration enables organizations to catalog data assets in Azure Synapse Analytics, and makes it easier for data engineers to find data assets and track data lineage when implementing data pipelines that ingest data into Azure Synapse Analytics.
* **Azure Machine Learning** integration enables data analysts and data scientists to integrate predictive model training and consumption into analytical solutions.

# When to use Azure Synapse Analytics

Across all organizations and industries, the common use cases for Azure Synapse Analytics are identified by the need for:

## Large-scale data warehousing

Data warehousing includes the need to integrate all data, including big data, to reason over data for analytics and reporting purposes from a descriptive analytics perspective, independent of its location or structure.

## Advanced analytics

Enables organizations to perform predictive analytics using both the native features of Azure Synapse Analytics, and integrating with other technologies such as Azure Machine Learning.

## Data exploration and discovery

The serverless SQL pool functionality provided by Azure Synapse Analytics enables Data Analysts, Data Engineers and Data Scientist alike to explore the data within your data estate. This capability supports data discovery, diagnostic analytics, and exploratory data analysis.

## Real time analytics

Azure Synapse Analytics can capture, store and analyze data in real-time or near-real time with features such as Azure Synapse Link, or through the integration of services such as Azure Stream Analytics and Azure Data Explorer.

## Data integration

Azure Synapse Pipelines enables you to ingest, prepare, model and serve the data to be used by downstream systems. This can be used by components of Azure Synapse Analytics exclusively.

## Integrated analytics

With the variety of analytics that can be performed on the data at your disposal, putting together the services in a cohesive solution can be a complex operation. Azure Synapse Analytics removes this complexity by integrating the analytics landscape into one service. That way you can spend more time working with the data to bring business benefit, than spending much of your time provisioning and maintaining multiple systems to achieve the same outcomes.

# Exercise - Explore Azure Synapse Analytics

Now it's your chance to explore the capabilities of Azure Synapse Analytics for yourself. In this exercise, you'll use a provided script to provision an Azure Synapse Analytics workspace in your Azure subscription; and then use Azure Synapse Studio to perform core data analytics tasks.

**Note**

To complete this lab, you will need an [**Azure subscription**](https://azure.microsoft.com/free) in which you have administrative access.

# Exercise - Explore Azure Synapse Analytics

<https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/01-Explore-Azure-Synapse.html>

