Data roles and services

Introduction

Over the last decade, the amount of data that systems and devices generate has increased significantly. Because of this increase, new technologies, roles, and approaches to working with data are affecting data professionals. Data professionals typically fulfill different roles when managing, using, and controlling data. In this module, you'll learn about the various roles that organizations often apply to data professionals, the tasks and responsibilities associated with these roles, and the Microsoft Azure services used to perform them.

Learning objectives

In this module you will learn how to:

- Identify common data professional roles
- Identify common cloud services used by data professionals

Data professional roles

There's a wide variety of roles involved in managing, controlling, and using data. Some roles are business-oriented, some involve more engineering, some focus on research, and some are hybrid roles that combine different aspects of data management. Your organization may define roles differently, or give them different names, but the roles described in this unit encapsulate the most common division of tasks and responsibilities.

The three key job roles that deal with data in most organizations are:

- **Database administrators** manage databases, assigning permissions to users, storing backup copies of data and restore data in the event of a failure.
- **Data engineers** manage infrastructure and processes for data integration across the organization, applying data cleaning routines, identifying data governance rules, and implementing pipelines to transfer and transform data between systems.
- **Data analysts** explore and analyze data to create visualizations and charts that enable organizations to make informed decisions.

NOTE: The job *roles* define differentiated tasks and responsibilities. In some organizations, the same *person* might perform multiple roles; so in their role as database administrator they might provision a transactional database, and then in their role as a data engineer they might create a pipeline to transfer data from the database to a data warehouse for analysis.

Database Administrator

A database administrator is responsible for the design, implementation, maintenance, and operational aspects of on-premises and cloud-based database systems. They're responsible for the overall availability and consistent performance and optimizations of databases. They work with stakeholders to implement policies, tools, and processes for backup and recovery plans to recover following a natural disaster or human-made error.

The database administrator is also responsible for managing the security of the data in the database, granting privileges over the data, granting or denying access to users as appropriate.

Data Engineer

A data engineer collaborates with stakeholders to design and implement data-related workloads, including data ingestion pipelines, cleansing and transformation activities, and data stores for analytical workloads. They use a wide range of data platform technologies, including relational and non-relational databases, file stores, and data streams.

They're also responsible for ensuring that the privacy of data is maintained within the cloud and spanning from on-premises to the cloud data stores. They own the management and monitoring of data pipelines to ensure that data loads perform as expected.

Data Analyst

A data analyst enables businesses to maximize the value of their data assets. They're responsible for exploring data to identify trends and relationships, designing and building analytical models, and enabling advanced analytics capabilities through reports and visualizations.

A data analyst processes raw data into relevant insights based on identified business requirements to deliver relevant insights.

NOTE: The roles described here represent the key data-related roles found in most medium to large organizations. There are additional data-related roles not mentioned here, such as *data scientist* and *data architect*; and there are other technical professionals that work with data, including *application developers* and *software engineers*.

Microsoft cloud services for data

Microsoft Azure is a cloud platform that powers the applications and IT infrastructure for some of the world's largest organizations. It includes many services to support cloud solutions, including transactional and analytical data workloads.

Some of the most commonly used cloud services for data are described below.

NOTE: This topic covers only some of the most commonly used data services for modern transactional and analytical solutions. Additional services are also available.

Azure SQL

Azure SQL is the collective name for a family of relational database solutions based on the Microsoft SQL Server database engine. Specific Azure SQL services include:

- **Azure SQL Database** – a fully managed platform-as-a-service (PaaS) database hosted in Azure
- **Azure SQL Managed Instance** – a hosted instance of SQL Server with automated maintenance, which allows more flexible configuration than Azure SQL DB but with more administrative responsibility for the owner.
- **Azure SQL VM** – a virtual machine with an installation of SQL Server, allowing maximum configurability with full management responsibility.

Database administrators typically provision and manage Azure SQL database systems to support line of business (LOB) applications that need to store transactional data.

Data engineers may use Azure SQL database systems as sources for data pipelines that perform *extract*, *transform*, and *load* (ETL) operations to ingest the transactional data into an analytical system.

Data analysts may query Azure SQL databases directly to create reports, though in large organizations the data is generally combined with data from other sources in an analytical data store to support enterprise analytics.

Azure Database for open-source relational databases

Azure includes managed services for popular open-source relational database systems, including:

- **Azure Database for MySQL** - a simple-to-use open-source database management system that is commonly used in *Linux*, *Apache*, *MySQL*, and *PHP* (LAMP) stack apps.
- **Azure Database for MariaDB** - a newer database management system, created by the original developers of MySQL. The database engine has since been rewritten and optimized to improve performance. MariaDB offers compatibility with Oracle Database (another popular commercial database management system).
- **Azure Database for PostgreSQL** - a hybrid relational-object database. You can store data in relational tables, but a PostgreSQL database also enables you to store custom data types, with their own non-relational properties.

As with Azure SQL database systems, open-source relational databases are managed by database administrators to support transactional applications, and provide a data source for data engineers building pipelines for analytical solutions and data analysts creating reports.

Azure Cosmos DB

Azure Cosmos DB is a global-scale non-relational (*NoSQL*) database system that supports multiple application programming interfaces (APIs), enabling you to store and manage data as JSON documents, key-value pairs, column-families, and graphs.

In some organizations, Cosmos DB instances may be provisioned and managed by a database administrator; though often software developers manage NoSQL data storage as part of the overall application architecture. Data engineers often need to integrate Cosmos DB data sources into enterprise analytical solutions that support modeling and reporting by data analysts.

Azure Storage

Azure Storage is a core Azure service that enables you to store data in:

- **Blob containers** - scalable, cost-effective storage for binary files.
- **File shares** - network file shares such as you typically find in corporate networks.
- **Tables** - key-value storage for applications that need to read and write data values quickly.

Data engineers use Azure Storage to host *data lakes* - blob storage with a hierarchical namespace that enables files to be organized in folders in a distributed file system.

Azure Data Factory

Azure Data Factory is an Azure service that enables you to define and schedule data pipelines to transfer and transform data. You can integrate your pipelines with other Azure services, enabling you to ingest data from cloud data stores, process the data using cloud-based compute, and persist the results in another data store.

Azure Data Factory is used by data engineers to build *extract*, *transform*, and *load* (ETL) solutions that populate analytical data stores with data from transactional systems across the organization.

Azure Synapse Analytics

Azure Synapse Analytics is a comprehensive, unified data analytics solution that provides a single service interface for multiple analytical capabilities, including:

- **Pipelines** - based on the same technology as Azure Data Factory.
- **SQL** - a highly scalable SQL database engine, optimized for data warehouse workloads.
- **Apache Spark** - an open-source distributed data processing system that supports multiple programming languages and APIs, including Java, Scala, Python, and SQL.
- **Azure Synapse Data Explorer** - a high-performance data analytics solution that is optimized for real-time querying of log and telemetry data using Kusto Query Language (KQL).

Data engineers can use Azure Synapse Analytics to create a unified data analytics solution that combines data ingestion pipelines, data warehouse storage, and data lake storage through a single service.

Data analysts can use SQL and Spark pools through interactive notebooks to explore and analyze data, and take advantage of integration with services such as Azure Machine Learning and Microsoft Power BI to create data models and extract insights from the data.

Azure Databricks

Azure Databricks is an Azure-integrated version of the popular Databricks platform, which combines the Apache Spark data processing platform with SQL database semantics and an integrated management interface to enable large-scale data analytics.

Data engineers can use existing Databricks and Spark skills to create analytical data stores in Azure Databricks.

Data Analysts can use the native notebook support in Azure Databricks to query and visualize data in an easy to use web-based interface.

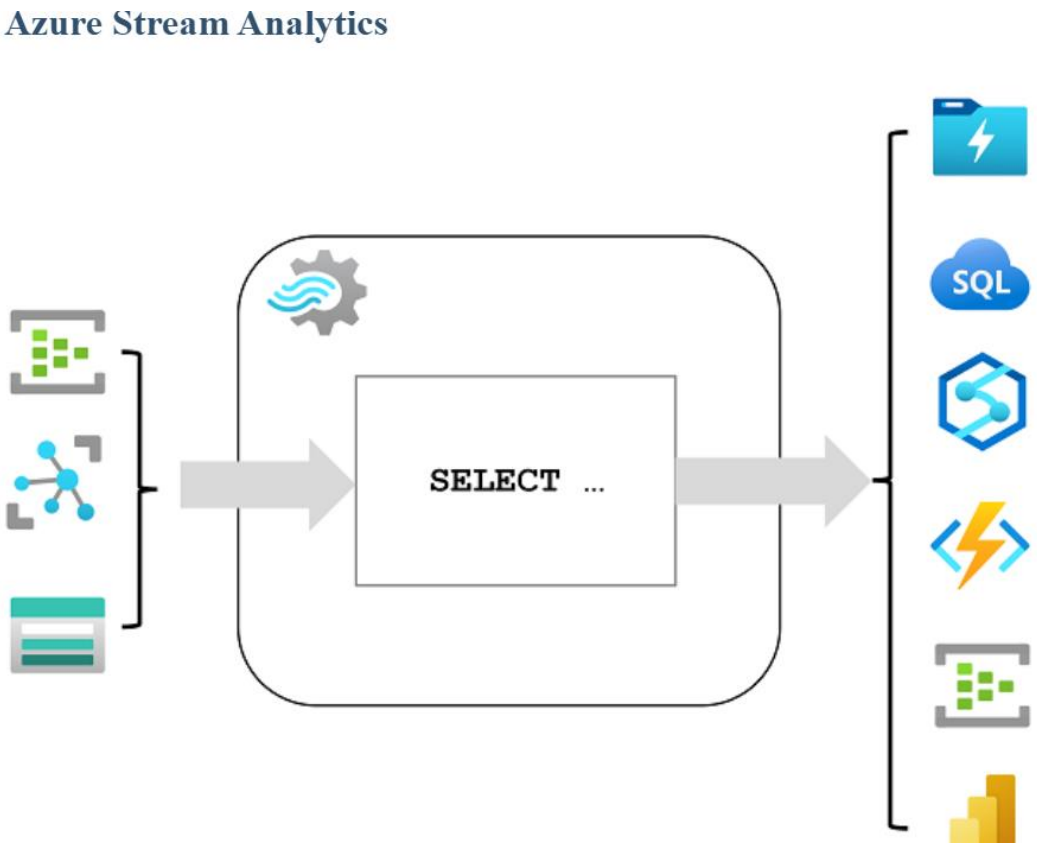
Azure HDInsight

Azure HDInsight is an Azure service that provides Azure-hosted clusters for popular Apache open-source big data processing technologies, including:

- **Apache Spark** - a distributed data processing system that supports multiple programming languages and APIs, including Java, Scala, Python, and SQL.
- **Apache Hadoop** - a distributed system that uses *MapReduce* jobs to process large volumes of data efficiently across multiple cluster nodes. MapReduce jobs can be written in Java or abstracted by interfaces such as Apache Hive - a SQL-based API that runs on Hadoop.
- **Apache HBase** - an open-source system for large-scale NoSQL data storage and querying.
- **Apache Kafka** - a message broker for data stream processing.
- **Apache Storm** - an open-source system for real-time data processing through a topology of *spouts* and *bolts*.

Data engineers can use Azure HDInsight to support big data analytics workloads that depend on multiple open-source technologies.

Azure Stream Analytics



Azure Stream Analytics is a real-time stream processing engine that captures a stream of data from an input, applies a query to extract and manipulate data from the input stream, and writes the results to an output for analysis or further processing.

Data engineers can incorporate Azure Stream Analytics into data analytics architectures that capture streaming data for ingestion into an analytical data store or for real-time visualization.

Azure Data Explorer

Azure Data Explorer is a standalone service that offers the same high-performance querying of log and telemetry data as the Azure Synapse Data Explorer runtime in Azure Synapse Analytics.

Data analysts can use Azure Data Explorer to query and analyze data that includes a timestamp attribute, such as is typically found in log files and *Internet-of-things* (IoT) telemetry data.

Microsoft Purview

Microsoft Purview provides a solution for enterprise-wide data governance and discoverability. You can use Microsoft Purview to create a map of your data and track data lineage across multiple data sources and systems, enabling you to find trustworthy data for analysis and reporting.

Data engineers can use Microsoft Purview to enforce data governance across the enterprise and ensure the integrity of data used to support analytical workloads.

Microsoft Power BI

Microsoft Power BI is a platform for analytical data modeling and reporting that data analysts can use to create and share interactive data visualizations. Power BI reports can be created by using the Power BI Desktop application, and the published and delivered through web-based reports and apps in the Power BI service, as well as in the Power BI mobile app.

Summary

Managing and working with data is a specialist skill that requires knowledge of multiple technologies. Most organizations define job roles for the various tasks responsible for managing data.

In this lesson you've learned how to:

- Identify common data professional roles
- Identify common cloud services used by data professionals