# **Microsoft Certified: Azure Data Fundamentals**

Core data concepts

**Identify data formats**

1. **Structured data** is data that adheres to a fixed *schema*, so all the data has the same fields or properties. (database)
2. **Semi-structured data** is information that has some structure, but which allows for some variation between entity instances. (JSON)
3. **Unstructured data** documents, images, audio and video data, and binary files might not have a specific structure.
4. **Data stores ex.** File stores, databases.

**Explore file storage**

1. Delimited text files
2. JavaScript Object Notation (JSON)
3. Extensible Markup Language (XML)
4. Binary Large Object (BLOB)
5. ***Avro*** is a row-based format
6. *ORC* (Optimized Row Columnar format) organizes data into columns rather than rows.
7. *Parquet* is another columnar data format. It was created by Cloudera and X.

**Explore databases**

1. **Relational databases** are used to store and query structured data.
2. **Non-relational databases** are data management systems that don’t apply a relational schema to the data.
   1. **Key-value databases** record consists of a unique key and an associated value, which can be in any format.
   2. **Document databases** are a specific form of key-value database in which the value is a JSON document.
   3. **Column family databases** store tabular data comprising rows and columns.
   4. **Graph databases** – store entities as nodes.

**Transactional data processing???**

**A transaction** is a small, discrete, unit of work.

**Transactional systems** are often high-volume, sometimes handling many millions of transactions in a single day. The work performed by transactional systems is often referred to as **Online Transactional Processing (OLTP).**

* **Atomicity**– each transaction is treated as a single unit, which succeeds completely or fails completely.
* **Consistency**– transactions can only take the data in the database from one valid state to another.
* **Isolation** – concurrent transactions cannot interfere with one another, and must result in a consistent database state.
* **Durability** – when a transaction has been committed, it will remain committed.

**Analytical data processing**

uses read-only (or read-*mostly*) systems that store vast volumes of historical data or business metrics.

***Data lakes*** are common in large-scale data analytical processing scenarios, where a large volume of file-based data must be collected and analyzed.

***Data warehouses*** are an established way to store data in a relational schema that is optimized for read operations – primarily queries to support reporting and data visualization.

Data roles and services

**Job roles in the world of data**

* **Database administrators** manage databases, assigning permissions to users, storing backup copies of data, and restoring data in the event of a failure - design, implementation, maintenance, and operational aspects of on-premises and cloud-based database systems, 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 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 - design and implement data-related workloads
* **Data analysts** explore and analyze data to create visualizations and charts that enable organizations to make informed decisions - enables businesses to maximize the value of their data assets.

**Data services.**

1. **Azure SQL**

* **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.

1. **Open-source databases in Azure**

* **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.

1. **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.
2. **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.

1. **Azure Data Factory** is an Azure service that enables you to define and schedule data pipelines to transfer and transform data.
2. **Microsoft Fabric** is a unified Software-as-a-Service (SaaS) analytics platform based on an open and governed lakehouse that includes functionality to support:
3. **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.
4. **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.
5. **Azure Data Explorer** is a fully managed, standalone, big data analytics platform that offers high-performance querying of log and telemetry data
6. **Microsoft Purview** - provides a solution for enterprise-wide data governance and discoverability.