# **Microsoft Certified: Azure Data Fundamentals (DP-900)**

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 (CSV, TSV comma-separated and tab-separated values)
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.

Different types of **user** might perform data analytical work at different stages of the overall architecture.

* Data scientists might work directly with data files in a data lake to explore and model data.
* Data Analysts might query tables directly in the data warehouse to produce complex reports and visualizations.
* Business users might consume pre-aggregated data in an analytical model in the form of reports or dashboards.

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.

Fundamental relational data concepts

**Relational data**

In a relational database, you model collections of entities from the real world as **tables.**

**Normalization**

Normalization is a term used by database professionals for a schema design process that minimizes data duplication and enforces data integrity.

1. Separate each *entity* into its own table.
2. Separate each discrete *attribute* into its own column.
3. Uniquely identify each entity instance (row) using a *primary key*.
4. Use *foreign key* columns to link related entities.

**SQL**

Some popular dialects of SQL include:

* *Transact-SQL (T-SQL)*. This version of SQL is used by Microsoft SQL Server and Azure SQL services.
* *pgSQL*. This is the dialect, with extensions implemented in PostgreSQL.
* *PL/SQL*. This is the dialect used by Oracle. PL/SQL stands for Procedural Language/SQL.

**Database objects**

* Views - a virtual table based on the results of a **SELECT** query.
* Stored procedures - defines SQL statements that can be run on command.
* Indexes helps you search for data in a table.

**Relational database services in Azure.**

Azure SQL services and capabilities.

A ***hybrid deployment*** is a system where part of the operation runs on-premises, and part in the cloud. Your database might be part of a larger system that runs on-premises, although the database elements might be hosted in the cloud.

**Describe core data concepts (15-20%)**

**Describe types of core data workloads**

* Describe streaming data
* Describe the difference between batch and streaming data
* Describe the characteristics of relational data

**Describe data analytics core concepts**

* Describe data visualization (e.g., visualization, reporting, business intelligence (BI))
* Describe basic chart types such as bar charts and pie charts
* Describe analytics techniques (e.g., descriptive, diagnostic, predictive, prescriptive, cognitive)
* Describe ELT and ETL processing
* Describe the concepts of data processing

**Describe how to work with relational data on Azure (25-30%)**

**Describe relational data workloads**

* Identify the right data offering for a relational workload
* Describe relational data structures (e.g., tables, index, views)

**Describe relational Azure data services**

* describe and compare PaaS, IaaS, and SaaS solutions
* Describe Azure SQL database services including Azure SQL Database, Azure SQL Managed Instance, and SQL Server on Azure Virtual Machine
* describe Azure Synapse Analytics
* Describe Azure Database for PostgreSQL, Azure Database for MariaDB, and Azure Database for MySQL

**Identify basic management tasks for relational data**

* Describe provisioning and deployment of relational data services
* Describe method for deployment including the Azure portal, Azure Resource Manager templates, Azure PowerShell, and the Azure command-line interface (CLI)
* Identify data security components (e.g., firewall, authentication)
* Identify basic connectivity issues (e.g., accessing from on-premises, access with Azure VNets, access from Internet, authentication, firewalls)
* Identify query tools (e.g., Azure Data Studio, SQL Server Management Studio, sqlcmd utility, etc.)

**Describe query techniques for data using SQL language**

* Compare Data Definition Language (DDL) versus Data Manipulation Language (DML)
* Query relational data in Azure SQL Database, Azure Database for PostgreSQL, and Azure Database for MySQL

**Describe how to work with non-relational data on Azure (25-30%)**

**Describe non-relational data workloads**

* Describe the characteristics of non-relational data
* Describe the types of non-relational and NoSQL data
* Recommend the correct data store
* Determine when to use non-relational data

**Describe non-relational data offerings on Azure**

* Identify Azure data services for non-relational workloads
* Describe Azure Cosmos DB APIs
* Describe Azure Table storage
* Describe Azure Blob storage
* Describe Azure File storage

**Identify basic management tasks for non-relational data**

* Describe provisioning and deployment of non-relational data services
* Describe method for deployment including the Azure portal, Azure Resource Manager templates, Azure PowerShell, and the Azure command-line interface (CLI)
* Identify data security components (e.g., firewall, authentication, encryption)
* Identify basic connectivity issues (e.g., accessing from on-premises, access with Azure VNets, access from Internet, authentication, firewalls)
* Identify management tools for non-relational data

**Describe an analytics workload on Azure (25-30%)**

**Describe analytics workloads**

* Describe transactional workloads
* Describe the difference between a transactional and an analytics workload
* Describe the difference between batch and real time
* Describe data warehousing workloads
* determine when a data warehouse solution is needed

**Describe the components of a modern data warehouse**

* describe Azure data services for modern data warehousing such as Azure Data Lake, Azure Synapse Analytics, Azure Databricks, and Azure HDInsight
* Describe modern data warehousing architecture and workload

**Describe data ingestion and processing on Azure**

* Describe common practices for data loading
* Describe the components of Azure Data Factory (e.g., pipeline, activities, etc.)
* Describe data processing options (e.g., Azure HDInsight , Azure Databricks, Azure Synapse Analytics, Azure Data Factory)

**Describe data visualization in Microsoft Power BI**

* Describe the role of paginated reporting
* Describe the role of interactive reports
* Describe the role of dashboards

Describe the workflow in Power BI