

## 1.1. Explore core data concepts

**Structured data** (Tabular form)

Rows represent instance of a data entity

Columns represent attributes of the entity

**Semi structured data** (JSON etc)

**Unstructured data** (doc, videos etc)

**Broad categories of data store:**

1. File stores
2. Databases

- **File storage**

A dedicated file for managing data records.

**Delimited text files**

**Javascript object notation (JSON)**

Objects in { }, Collection are in [ ]

**Extensible Markup Language( XML)**

in <./>

**Binary language object(BLOB)**

**Optimized file formats**

Avro: row based format

ORC(optimized row columnar format): column based

Parque: columnar format

- **Databases**

A dedicated system for managing data records.

1. Relational (query structured data having primary key, queried using SQL)
2. Non-relational (referred to as NoSQL database, or few support SQL)  
Four types are
  1. Key value database (each record has unique key having its own value)
  2. Document database (form of key-value database, value is a JSON document)
  3. Column family database: (stores tabular data , columns can be divided in groups)
  4. Graph databases: (define relationships by creating entities as nodes)

**Transactional data processing**

**Online transactional processing (oltp)**

**CRUD** operations (create, retrieve, update, delete)

**ACID semantics**

Atomicity: transaction is treated as a single unit. *all-or-nothing* property.

Consistency: ensures that a transaction brings the database from one valid state to another. *Sets rules* before and after transaction

Isolation: ensures the transactions are executed in isolation from each other. *Not visible* to other transactions

Durability: guarantees that once a transaction is committed, the changes are permanent and will survive any subsequent failure. Good in *backup* sense

### **Analytical data processing**

1. Data lake
2. Data warehouse (ETL)
3. OLAP (online analytical processing) model
4. Visualization

## **1.2. Explore Data Roles and Services**

### **Key roles in data field**

1. Database administrators
2. Data engineers
3. Data analysts

**Data Services** (supports transactional and analytical solutions) RDBMS

1. **Azure SQL**
  - Azure sql VM (IaaS)
  - Azure sql managed instance (PaaS)
  - Azure sql database (PaaS managed)
2. **Azure Database for open-source relational database**
  - Azure database for mysql (PaaS)
  - Azure database for MariaDB
  - Azure database for postgresSQL
3. **Azure Cosmos DB**  
A NoSQL database
4. **Azure Storage**  
Data engineers use azure storage to host data lakes. it contains data in
  - Blob containers
  - File shares
  - Tables
5. **Azure Data Factory**  
Data engineers use azure storage for ETL solutions
6. **Azure Synapse Analytics**  
it is a unified data solution that provides a single service interface for multiple analytical capabilities including:
  - Pipelines

- SQL
- Apache Spark
- Azure Synapse data explorer (use Kusto Query language KQL)

## 7. Azure DataBricks

Combines the Apache Spark data processing platform with SQL database semantics

## 8. Azure HDInsight

Provides Azure hosted clusters for technologies like

- Apache Spark (includes Java, python, scala, sql)
- Apache Hadoop (uses MapReduce jobs)
- Apache HBase
- Apache Kafka

## 9. Azure Stream Analytics

## 10. Azure Data Explorer

## 11. Microsoft Purview

## 12. Microsoft PowerBI

# 2.1. Fundamental Relational Data Concepts

## Relational Data

That is structured in tabular form. In addition to tables, it also contains other structures as:

- Views
- Stored procedures
- Indexes

## Normalization:

Schema design process that minimizes data duplication and enforces data integrity.

Convergency it is about to

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.

## Structured Query Language (SQL)

Used to communicate with a relational database. RDBMS that use SQL are Microsoft SQL Server, MySQL, PostgreSQL, MariaDB, and Oracle.

Some dialects of SQL are:

- Transact-SQL (T-SQL): used by Microsoft SQL Server and Azure SQL services
- PgSQL: extension is implemented in PostgreSQL
- PL/SQL: used by Oracle. Stands for Procedural Language.

## SQL statement Types

1. Data Definition Language (DDL) **CARD**
  - Create
  - Alter
  - Rename
  - Drop
2. Data Control Language (DCL)
  - Grant
  - Deny
  - Revoke
3. Data Manipulation Language (DML)
  - Select
  - Insert
  - Update
  - Delete

### View?

View is a virtual table based on the results of a SELECT query

### Stored Procedure?

Defines sql statements that can be run on command.

### Index?

Helps you search for data in a table.

## 2.2. Explore Relational Database Services in Azure

### Prewritten things:

**Data Services** (supports transactional and analytical solutions) RDBMS

#### 13. Azure SQL

- Azure sql VM (IaaS)
- Azure sql managed instance (PaaS)
- Azure sql database (PaaS managed)

#### 14. Azure Database for open-source relational database

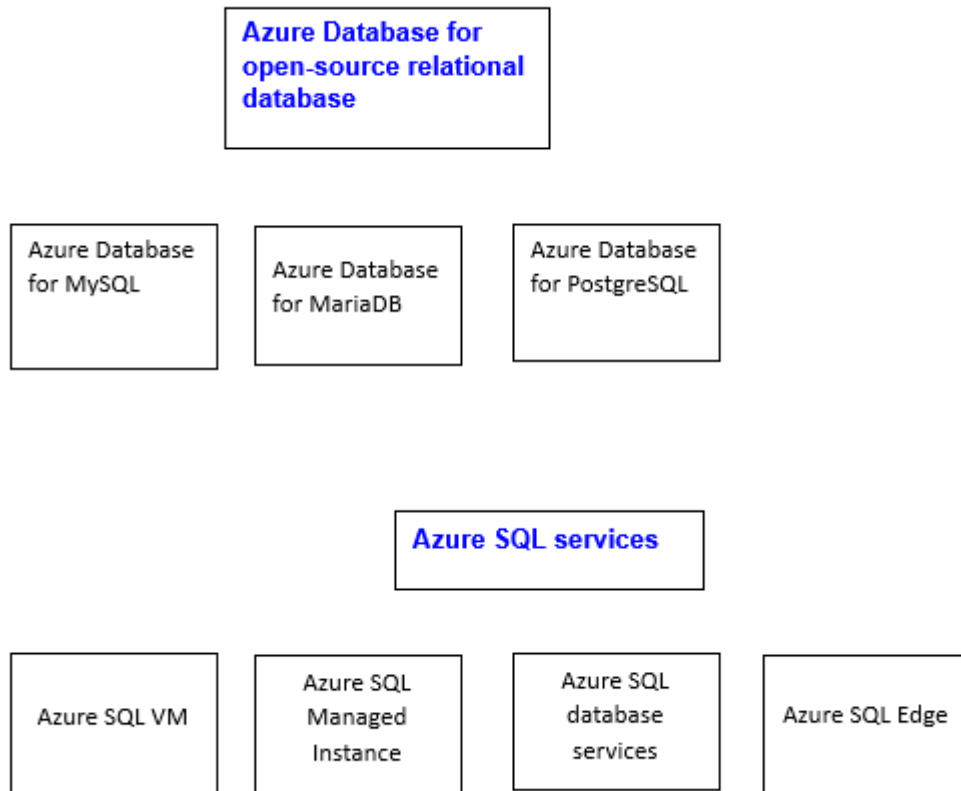
- Azure database for mysql (PaaS)
- Azure database for MariaDB
- Azure database for postgresQL




### Azure SQL services

Azure SQL is a collective term for a family of Microsoft SQL Server based database services on Azure. Specific Azure SQL services include:

- SQL server on Azure Virtual Machines (VMs) (makes it IaaS)
- Azure SQL managed instance (PaaS)
- Azure SQL database (PaaS, provides 100% compatibility with on-premises SQL server instances)

- Azure SQL Edge (optimized for IOT)



	SQL Server on Azure VMs	Azure SQL Managed Instance	Azure SQL Database
			
Type of cloud service	IaaS	PaaS	PaaS
SQL Server compatibility	Fully compatible with on-premises physical and virtualized installations. Applications and databases can easily be "lift and shift" migrated without change.	Near-100% compatibility with SQL Server. Most on-premises databases can be migrated with minimal code changes by using the <a href="#">Azure Database Migration service</a>	Supports most core database-level capabilities of SQL Server. Some features depended on by an on-premises application may not be available.
Architecture	SQL Server instances are installed in a virtual machine. Each instance can support multiple databases.	Each managed instance can support multiple databases. Additionally, <i>instance pools</i> can be used to share resources efficiently across smaller instances.	You can provision a <i>single database</i> in a dedicated, managed (logical) server; or you can use an <i>elastic pool</i> to share resources across multiple databases and take advantage of on-demand scalability.

Availability	99.99%	99.99%	99.995%
Management	You must manage all aspects of the server, including operating system and SQL Server updates, configuration, backups, and other maintenance tasks.	Fully automated updates, backups, and recovery.	Fully automated updates, backups, and recovery.
Use cases	Use this option when you need to migrate or extend an on-premises SQL Server solution and retain full control over all aspects of server and database configuration.	Use this option for most cloud migration scenarios, particularly when you need minimal changes to existing applications.	Use this option for new cloud solutions, or to migrate applications that have minimal instance-level dependencies.

1. Which deployment option offers the best compatibility when migrating an existing SQL Server on-premises solution?

Azure SQL Database (single database)

Azure SQL Database (elastic pool)

Azure SQL Managed Instance

Correct. Azure SQL Managed Instance offers near 100% compatibility with SQL Server.

2. Which of the following statements is true about Azure SQL Database?

Most database maintenance tasks are automated

Correct. Azure SQL Database automates most maintenance tasks.

You must purchase a SQL Server license

It can only support one database

3. Which database service is the simplest option for migrating a LAMP application to Azure?

Azure SQL Managed Instance

Azure Database for MySQL

Correct. LAMP standard for Linux, Apache, MySQL, and PHP.

Azure Database for PostgreSQL

Incorrect. That's incorrect. LAMP standard for Linux, Apache, MySQL, and PHP.

InshaAllah

# Exercise: Explore Azure relational database services

<https://microsoftlearning.github.io/DP-900T00A-Azure-Data-Fundamentals/Instructions/Labs/dp900-01-sql-lab.html>

<https://learn.microsoft.com/en-us/training/modules/explore-provision-deploy-relational-database-offerings-azure/4-exercise-provision-relational-azure-data-services?pivots=azuresql>

## 3.1. Non-relational data in Azure

### Azure Blob storage

Enables you to store massive amount of unstructured data as blob (binary large objects) in the cloud.

In an azure storage account, blobs are stored in containers. It groups related blobs together.

In containers, blobs are organised in a hierarchy of virtual folders, so we cannot perform folder level operation on it to control access.

### Types of blob:

Block: set of blocks upto 50,000 blocks in one blob of max size 4.7 TB, each having diff size upto 100 MB.

Page: collection of fixed size pages of 512 bytes. Cmax size 8TB.

Append: can append blob blocks. Each size varies but upto 4MB. Max size is 195GB. It only appends blocks and does not change, or delete existing ones.

### Access tiers

Hot: High performance, used when data is accessed frequently. Latency can be of a few milliseconds.

Cool: lower performance, less storage charges, used when data is accessed infrequently.

Latency can be of a few milliseconds. migration can be done between hotly tier to cool and vice versa.

Archive: lowest storage cost, Latency can take hours. used for historical data storage or can say backup data.

## Azure DataLake Storage Gen2

Advantage of scalability of blob storage, cost control of storage tier, hierarchical file system capabilities and compatibility with major analytical systems of Azure Data Lake Store.

*Upgrade the account to enable hierarchical namespace and create a blob container in order to support a data lake for Azure Synapse Analytics.*

Structure:

Azure storage account

- Blob container
- —Directory
  - File1
  - File2

Hadoop example. Hierarchical namespace is a must, one-way process.

## Azure files

File share: enables to store a file on one computer and grant access to that file on other computers, it works well for LAN.

As like file share, azure file is a way to create cloud based network shares to share files.

Can share upto 100TB data in a single storage account. The max size of single file is 1 TB.

Currently, azure file storage supports upto 2000 concurrent connections per shared file. To upload files we can use azure portal, AzCopy utility, and Azure File Sync service. It enables users at different sites to share files.

### Performance tiers of Azure File Storage:

- Standard tier: Uses hard-disk based hardware in a datacenter.
- Premium tier: Uses solid-disk based hardware in a datacenter. Higher rate, greater throughout

### Network file share protocols for Azure files:

- Server Message Block (SMB): Used by multiple OS (Windows, Linux, macOS)
- Network File System (NFS): Used by some Linux and macOS versions otherwise use premium tier

## Azure Tables

Azure table storage is a NoSQL storage solution that uses key/value items. Store semi-structured data. Elements of Azure table storage key are partition key and row key.

### Fundamentals of Azure Cosmos DB

- Microsoft's fully managed serverless distributed database for applications of any size or scale.
- Azure cosmos db supports multiple APIs to work in a cosmos db database.
- It is highly scalable database management system.
- Cosmos DB uses indexes and partitioning.
- It automatically allocates space in a container for your partitions and each partition can grow upto 10 GB in size.

### Azure CosmosDB for NoSQL

- Microsoft's native non-relational service
- Manages data in JSON document format and uses SQL syntax to work on data.

### Azure CosmosDB for MongoDB

- Open source database
- Data is stored in BSON (Binary JSON)



- MQL (MongoDB Query Language) uses object oriented syntax to use objects to call methods.

#### **Azure CosmosDB for PostgreSQL**

- Relational database management system (RDBMS)

#### **Azure CosmosDB for Table**

- Work with data in key value tables
- Similar to azure table storage, but have greater scalability and performance.

#### **Azure CosmosDB for Apache Cassandra**

- Uses column-family storage structure.

#### **Azure CosmosDB for Apache Gremlin**

- Used with data in graph structure
- Entities as vertices, and nodes as relationships.

#### *Question:*

How can you enable globally distributed users to work with their own replica of a cosmos DB database?

Enable multi-region writes and add the regions where you have users.

### 4.0. Data Analytics in Azure

#### **4.1. Fundamentals of large scale data warehousing**

- Combines conventional data warehousing for BI with 'big data' like techniques.
- Big data processing solutions are stored in a data lake from which Apache Spark like distributed processing engines process.

#### Data Warehousing Architecture

1. Data ingestion and processing  
Includes both batch processing of static data and real time processing of streaming data
2. Analytical data store  
includes relational data warehouse, data lakes or lake database.
3. Analytical data model  
For pre-aggregation of data and making models to produce reports. Often these data models are described as cubes.
4. Data visualization

#### Data Ingestion pipeline

We can create and run pipelines using azure data factory and in azure synapse analytics. Pipelines are linked services to load and process data.

#### **Analytical data stores**

Types of analytical data stores:

1. Data warehouse

## 2. Data lakes

### Data warehouse

- Relational database in which data is stored in schema.
- Star schema/ snowflake schema

### Data Lake

- A file store, usually on a distributed file system.
- Spark and Hadoop uses schema-on-read approach.

### Hybrid approach

- Combines both features in a lake database or data lakehouse.

### Azure Services for Analytical Stores

- Azure Synapse Analytics: high performance sql server based relational data warehouse along with data lake and apache spark.
- Azure Databricks: build on apache spark.
- Azure HDInsight: suitable if analytics solution relies on multiple open source frameworks or if to migrate an on-premises Hadoop-based solution to the cloud.

### Look into the exercise MCQs

## 4.2. Fundamentals of real-time analytics

Batch processing: multiple data records are collected and stored before performing a single operation.

Stream processing: source of data is constantly monitored in real time as new data event occurs.

### Real time analytics in Azure

- Azure stream Analytics: PaaS solution
- Spark Structured Streaming: enables to develop complex streaming solutions on Apache Spark based services, including Azure Synapse Analytics, databricks and HDInsight.
- Azure Data Explorer: time series element and can be used as Azure Synapse Data Explorer Runtime in an Azure Synapse Analytics workspace.

### Sources for stream processing

- Azure event hubs: in order, at once.
- Azure IoT hub: similar to AEH but data from IoT devices.
- Azure Data Lake Store Gen2: can be used for both batch and
- Apache Kafka: open source data ingestion solution used with Apache Spark. Azure HDInsight can be used for creating kafka cluster.

### Sinks for stream processing (output sent to these services):

- Azure Event Hubs
- Azure Data Lake Store Gen2
- Azure SQL database / Azure Synapse Analytics / Azure DataBricks

- Microsoft PowerBI

Azure Stream Analytics jobs and clusters

To use Azure Stream Analytics we create Stream Analytics job in Azure subscription.

Apache Spark on Azure

Can be used for both batch and stream processing. We can use Spark in services like:

- Azure Synapse Analytics
- Azure DataBricks
- Azure HDInsight

Spark Structured Streaming

- To process streaming data on Spark we can use Spark Structured Streaming library that provides API for ingesting, processing, and outputting results.
- It is built on a structure called a dataframe.

Delta Lake

Can be used in Spark to define relational tables for both processings.

Question

Language that use to query real-time log data in Azure Synapse Data Explorer ----- KQL

#### 4.3. Fundamentals of Data Visualization