# 1. Explore core data concepts

A . DATA FORMATS

1. Structured data
2. Semi-structured data
3. Unstructured data

B. DATA STORES

1. File Stores

2. Databases

1. File storage

Delimited text files , JSON , XML , BLOB ,AVRO, ORC , Parquet

1. Databases

Relational databases,

Non-relational databases

There are 4 types of non-relational databases:

1. Key-value databases
2. Document databases
3. Column family database
4. Graph database

TRANSACTIONAL DATA PROCESSING

OLAP (Online analytical processing)

OLTP (Online transactional processing)

Properties of Relation databases:

1. Atomicity – if a transaction is processed it will be processed completely if not completed it will fail completely.
2. Consistency – If a transaction is sent from A to B and it is processed from A completely but before reaching B it fails. This makes the result inconsistent. The transaction must be executed in its entirety in order to ensure the correctness of database.
3. Isolation – This makes sure that multiple processes can run concurrently without any interference.
4. Durability – When the process is completed is written in memory, then incase the process fails somehow still the database will be able to complete the transaction, since process is written in memory, so it d’snt have any effect of failure

## Data Services

1. Azure SQL – It is collective name for a family of relational database solutions based on SQL server engine.
2. Azure SQL database - fully managed PAAS database hosted in azure.
3. Azure SQL Managed Instance – hosted instance of SQL server with automated maintenance which allows flexible configuration than Azure SQL DB.
4. Azure SQL VM – Virtual machine with installation of SQL server, allowing max configurability with full management responsibility.
5. Azure database for open-source relational databases
6. Azure database for MySQL
7. Azure database for MariaDB
8. Azure database for PostgreSQL
9. Azure COSMOS DB

It is a global-scale non-relational database system that support multiple API’s enabling to store and managed data as JSON documents, key-value pairs, column families and graphs

1. Azure Storage

1. Blob container
2. File-Shares
3. Tables
4. Azure Data Factory

It enables to define and schedule data pipelines to transfer and transform data. Used for ETL.

1. Azure synapse analytics

It is a unified data analytics solution that provide single service interface for multiple analytical services eg: Pipelines, SQL, Apache Spark, Synapse data explorer .

1. Databricks

It combines Apache spark data processing platform with SQL database and an integrated management interface to enable large-scale data analytics.

1. HDInsight

It provides azure-hosted cloud clusters for Apache open-source big data processing technologies.

1. Stream analytics

It is a real-time stream processing engine that captures a stream of data from input, applies query to extract and manipulate data from input and write result to output for analysis.

1. Data Explorer

For analyzing log and telemetry data files

1. Purview

Solution for enterprise-wide data governance and discoverability. Used to map data and track data lineage across multiple data sources and systems.

1. Power BI

Visualization, Reporting

# 2. Explore relational data in Azure

# Normalization – It is used for professional schema design process that minimizes data duplication and enforces data integrity.

# Separate each entity into its own table

# Separate each discrete attribute into its own column

# Uniquely identify each entity instance using primary row

# Use foreign key columns to links related entities

# 3.Explore non-relational data in Azure

# 1. Azure Blob Storage

# Enables to store massive amount of unstructured data as binary large objects/blobs in cloud.

# Block Blobs, Page Blobs, Append Blobs

# Datalake storage – separate service for hierarchical data storage for analytical data lakes, often used by big-data analytical solutions that work with structured, semi and unstructured data stores in files.

# Azure file shares - way to create cloud-based network shares, through which data of multiple locations around the globe, can be accessed just like data from same site is accessed.

# Azure Tables - It is a No-SQL storage solution that make use of table containing key/value data items. Azure table is not like relational database, it stores semi-structured data

# Azure Cosmos DB – It is a globally supported NO-SQL database, supports multiple API’s that enable developers to use programming semantics of many common kinds to data store in cosmos db database.

# For eg. Ben is the manager of Sue, in this entities (Ben and Sue) are called Vertices and the relation between (Manager -Employee) them is called is called Edges

# 4. Explore data analytics in Azure

# Data visualization methods:

# Tables and Text – used to display numeric data and text data to be displayed

# Bar and column chart – to compare numeric values for discrete categories

# Line Chart – used to compare categorized values and to examine trends

# Pie charts – used to compare categorized values as proportions of total

# Scatter-Plot – used to compare two numeric measures and identify correlation between them

# Maps – display geographic locations

# Batch Processing and Stream Processing

# Data warehouse architecture

# Data ingestion and processing - ETL/ELT (Extract transform load/Extract Load Transform)

# Analytical data store – data-warehouse, file-system based data lakes.

# Analytical data model –

# Data visualization –making reports and dashboards

# Data warehouse – It is a relational database in which data is stored in a schema that is optimized for data analytics rather than transactional workloads.

# Data Lakes – It is a file store, usually on a distributed file system for high performance data access.

# ### ETL vs ELT

# ETL

# Appropriate for smaller datasets which require complex transformation and have been predetermined as being relevant to analysis goals

# Used for relational or structured data

# Not appropriate for data-lakes

# Transforming larger datasets can take-up long time but analysis can take place immediately

# Optimize data privacy

# ELT

# Used for relational and non-relational data both

# Suitable for pipelines and data-lakes

# Suitable for larger datasets also unlike ETL

# Transformation step takes little time but querying takes long time.

# Data privacy issues.

# Azure SQL

# Azure sql database -- PAAS solution

# Azure SQL managed instance – hosted instance of SQL server with automated maintenance and more authority

# Azure SQL VM – VM installation of SQL server, allowing maximum configurability with full management responsibility.

# Azure COSMOS database

# Global scale non-relational database that support multiple API’s.

# Azure data factory

# It enables to define and schedule data pipelines to transfer and transform data. Pipelines can be integrated with other azure services. used by data engineers for ETL/ELT.

# Azure synapse analytics

# Unified data analytics solution that provides a single service interface for multiple analytical services. Pipelines, SQL, Apache, azure, synapse data explorer.

# Azure Databricks

# It combines Apache spark data processing with SQL database.

# Azure HDInsight

# It provides azure hosted clusters for big data processing

# Stream analytics

# Real time stream processing engine that captures data applies ETL and write result to an output for analysis.

# Data explorer

# Used for querying log and telemetry data using KQL

# Purview

# Solution for data governance and discoverability. Used to create map and track data lineage across multiple data sources.

# Power BI

# Used for data visualization and reporting, dashboarding