DP-900

OLTP VS OLAP (Transaction processing vs Analytic Processing)
Batch Processing vs Stream Processing

Relation DB best suites for OLTP (Online Transaction Processing)

Data Normalization: Normalization is the process that is used to split an entity into multiple tables.
Reducing duplicacy of data

A clustered index is a data structure associated with a table that defines the order in which
rows are stored on a disk.

Data Engineer : One who collects the data and ingest in to database

Data Scientist : Make use of data and provide future insights

Data Analyst : Make use of data and visulation past records

Data Ingestion --> Data Processing --> Data Exploration

MYSQL : opensource DB, (In Azure - HA,scalable,point in time restoration (35 days)

Maria DB: New DBMS created by original developers of Mysql, compatible with Oracle DB, optimized to
improve performance, built-in support for temporal data (versioned data)

PostgreSQL : Hybrid relation object DB, enables to store custom data type, code modules can be added
manipulate geometric data like lines, circles & polygons

1) Azure portal shell
2) SQL Studio
3) Azure data Studio
4) Azure portal bash
5) sqlcmd

NO SQl :

SQL Disadvantages : Not scalable Not Flexible
can only scale vertical

NO SQL can scale horizantally
No structure is required

Azure Storage Service :
Authentication:
Storage Account Keys
Shared Access Signature
Azure Active Directory
Access Control
RBAC
ACL
Network Access
Firewall and Virtual network

Azure Cosmos DB: Globally Distributed, sclable DB & Multi Model DB

Key-Value : Table API
wide-column : Cassandra
Graph : Gremlin API
Document : MongoDB,
SQL : Core SQL

Cosmos DB Consistency Levels:

1. Strong --> No dirty reads, High latency, Cost high, close to RDBMS
2. Bounded Staleness --> Dirty reads possible, bounded by time and updates
3. Session --> No dirty reads for writers in the same session and possible for other users
4. Consitent Prefix --> Dirty reads possible but sequence maintained
5. Eventual --> Dirty reads, No guaranteed order, but eventually everything gets in order

Cosmos DB Security
1. RBAC
2. Network Security
3. Access Security Keys
4. CORS
5. Azure Private Endpoint

6. Advanced Security Option

Modern Data Ware House

1. Azure Data Factory - integration product -> bring data from any source(lot of plugins) to data lake
2. Azure Data Lake Storage Gen 2 -> used to store massive amount of data at cheapest cost
3. Data Bricks : Explore data present in ADLS using any language (preparation, cleaning etc)
4. Azure SQL Data ware house : query data from ADLS or move data from ADLS to ware house using polybase

Azure synapse : advance to sql data ware house
Combination of Big data, ETL, Data ware house and all

Loading Methods:
single client loading Methods --> can add some parallel processing capabilities but adds bottleneck
at control node

    1. SSIS
    2. Azure Data Factory
    3. BCP (Bulk control processing

Parallel reading loading Methods -->

    1. Polybase (reads data from blob storage and loads into Azure Data ware house, bypasses control node
    and loads directly into compute nodes)

HD Insight(hadoop on cloud) : Cloud distribution of Hadoop components
Process massive amount of data

Batch processing
1.U-SQL
2.Hive (Hadoop)
3.Pig
4.Sparks (Data bricks)

Azure Data Factory : SSIS in on premise
1. Copy Data
2. Transform Data

BUilding Blocks of Power Bi

1. Visualizations
2. Datasets
3. Reports (paginated Reports, Interactive Reports (requires Power BI server))
4. Dashboards
5. Tiles

Report Builder : To Build and preview reports , Power BI - to publish the reports