1. Explore Core Data Concepts:

Data is a collection of facts such as numbers, descriptions, and observations used in decision making.

\* OLAP vs OLTP:-

To support fast processing, the data in a transactional system is often divided into small pieces.

Splitting tables out into separate groups of columns like this is called normalized.

Normalization can enable a transactional system to use cache & process InMemory. But it can also make queries complex (joins).

Analytical systems are concerned with capturing raw data, and using it to generate insights.

\* Four Stages of Analytical system:-

Data Ingestion ==> Data Transformation/Cleaning => Querying/Analytical Processing => Visualization

\* Transactional workload VS Analytical workload:

A transaction is a sequence of operations that are atomic. Means all operationsin the sequence should be complete successfully, if any operation fails, all operations should be revert back to previous stage. (Ex. bank transactions).

A transactional database must adhere to the ACID (Atomicity, Consistency, Isolation, Durability) properties.

1. Atomicity: Each transaction is treated as a single unit, which either succeeds completely, or fails completely.
2. Consistency: A transaction can only take the data in the database from one valid state to another. (If fund is credited, it should be debited from another account or a record which describe frome where this fund is coming externally ).
3. Isolation: The intermediate state of a transaction is invisible to other transactions. As a result, transactions that run concurrently appear to be serialized. ie. transferred funds in one account or the other, but not in both, nor in neither.
4. Durability: Once a transaction has been committed, it will remain committed even if there's a system failure or power crash.

Lock vs eventual consistency:-

Many systems implement relational consistency and isolation by applying locks to data when it is updated. The lock prevents another process from reading the data until the lock is released. The lock is only released when the transaction commits or rolls back. Extensive locking can lead to poor performance, while applications wait for locks to be released.

In Distributed dB, locks may be retained for a very long time, especially if there's a network failure between databases at a critical time. To counter this problem, distributive DB implements "eventual consistency." In this form of consistency, as an application writes data, each change is recorded by one server and then propagated to the other servers in the distributed database system asynchronously.

\* Batch vs Stream processing:

Newly arriving data elements are collected into a group. The whole group is then processed at a future time as a batch. Any Problems with data, errors, and program crashes that occur during batch jobs bring the whole process to a halt. The input data must be carefully checked before the job can be run again. Even minor data errors, such as typographical errors in dates, can prevent a batch job from running.

In stream processing, each new piece of data is processed when it arrives. For example, data ingestion is inherently a streaming process.

*Explore Roles & Responsibility:*

\* Common Roles: Database Administrators, Data Engineers, Data Analysts.

DBA Tools : Azure data stdio, SQL Server Management Stdio, Azure portal.

Data Engineering Tools:

*Describe RDBMS Concepts:*

**Index** : Create index on a column to search it fast. However indexes aren't free. An index might consume additional storage space, and each time you insert, update, or delete data in a table, the indexes for that table must be maintained. This additional work can slow down insert, update, and delete operations, and incur additional processing charges. Some relational database management systems also support 'clustered indexes'. A clustered index physically reorganizes a table by the index key (Grouping through Index Key.)

**View :** A view is a virtual table based on the result set of a query.

*Describe Non-RDBMS Concepts:*

\* Semi Structured Data: JSON, Avro, OCR, Parquet

\* Transparent VS Opaque:

DBMS just sees the value as an unstructured block. Only the application understands how the data in the value is structured and what fields it contains.

In Transparent, DBMS understands how the fields in the data are organized. (Ex. RDBMS )

\* NoSQL: Key-Value, Document DB, Column family dB, Graph DB

**Key-Value** : data is oragined with Key and sorted with Key ASC. This is very fast for ingestion.

**Document DB** : Document store entire data of an entity (ie. MongoDB) The focus of a document database is its query capabilities.

**Column DB** : Whole document is devided into column family (ie. one column for name, one is for address info). ie Cassandra.

**Graph DB:**

*Explore Concept of Data Analytics*

\* Data Ingestion :

\* Data Processing :

**ETL :** Suitable for systems that only require simple models, with little dependency between items.

**ELT :** suitable for constructing complex models that depend on multiple items in the database, often using periodic batch processing. ELT is a scalable approach that is suitable for the cloud.

\* Data Analytics :

**Descriptive Analytics** : What has happened.

**Diagnostic Analytics** : Why things happened. This generally occurs in three steps:

Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.

Collect data that's related to these anomalies.

Use statistical techniques to discover relationships and trends that explain these anomalies.

**Predictive Analytics :** What will happen in the future.

**Prescriptive Analytics :** What action should be taken to acheive target or goal.

**Cognitive Analytics :** draw inference from existing data and patterns. (a self-learning feedback loop)

**2.Explore Relational Data in Azure:**

*Explore Relational Data Services in Azure*

\* Pass : You specify resources and Azure will automatically VM, network & other devices for you.

\* Azure data services falls is PaSS CATEGORY.

**\* SQL server on Azure VM:** 'lift-&-Shift', easily migrate on-premise SQL server DB to cloud.

**\* Azure SQL DB**: Scaling up or out will take effect without restarting the SQL database. 3 options are available.

- **Single DB :** Quickly setup & run a single DB. You can also specify a serverless configuration. In this configuration, Microsoft creates its own server, which might be shared by a number of databases belonging to other Azure subscribers. Microsoft ensures the privacy of your database. Your database automatically scales and resources are allocated or deallocated as required.

- **Elastic Pool :** Pool of servers which shares same resources. This model is useful if you have databases with resource requirements that vary over time, and can help you to reduce costs. Elastic Pool enables you to use the resources available in the pool, and then release the resources once processing has completed.

- **Managed Instance:** A fully controllable instance of SQL Server in the cloud. You can install multiple databases on the same instance. You have complete control over this instance, much as you would for an on-premises server. The Managed instance service automates backups, software patching, database monitoring, and other general tasks.

*Azure RDBMS Offerings:*

MySQL :

MariaDB :

Azure SQL:

PostGreSQL : Single server, HyperScale(distributed via partition/shrading key).

*Provising & Deploying RDBMS*

\* Provision RDBMS on Azure

\* Configuring Services related to RDBMS

- Configure connectivity to virtual network & on-premise servers.

- Configure private endpoints.( )

- Configure Authentication (ADDS)

Querying RDBMS in Azure

\* DML Commands: SIDU ( SELECT, INSERT, DELETE, UPDATE )

\* DDL Commands: CARD ( CREATE, ALTER, RENAME, DROP )

Running Query - sqlcmd : a cmd utility to run sql queries from cloud shell or from local cmd.

- Azure data Stdio : Cross platform GUI tool

- SQL Server Management Stdio :

- SQl Server data tool in Visual Stdio

**3.Explore Non-Relational Data in Azure:**

*Non Relational Offerring in Azure*

**\* Azure Table storage:**

- Key Value based => unique keys with 'array of Values'.

- Azure Table Storage have no concept of relationships, stored procedures, secondary indexes, or foreign keys.

- You can group rows based on common properties index via a partiton key. (partition + row = clustered Index)

- Partition key => Row key => ARRAY OF VALUES.

- Items in the same partition are AUTO INDEXED & stored in row key order.

**\* Point Query VS Range Query**

- Point Query retrieve a single row; Range query retrieve a contiguous block of rows.

- A table can have 252 columns(arrays of value), a single column can hold data to 64 KB.

**\* Blob storage:** binary object. For unsuctured data.

**Block Blob :** To store large, blob that donot change frequently. (50,000 blocks of upto 100 MB = 4.7 TB MAX)

**Page Blob** : Random read/write. (512 byte pages, max 8 Tb) => AZure use this for VM storage.

**Append Blobs :** Max 195 Gb, average size is 4 MB. Only Append, no update or delete.

- In Azure , you create blobs inside containers, and you can organize blobs in a hierarchy of folders.

- Blob storage provides three access tiers, which help to balance access latency and storage cost: Hot, Cold & Archieve

You can create lifecycle management policies for blobs in a storage account. A lifecycle management policy can automatically move a blob from Hot to Cool, to the Archive tier, as it ages and is used less frequently (policy is based on the number of days since modification). A lifecycle management policy can also arrange to delete outdated blobs.

Other features available with Azure Blob storage include:

* Versioning: You can maintain and restore earlier versions of a blob.
* Soft delete: Recover a blob that has been removed or overwritten, by accident or otherwise.
* Snapshots: A snapshot is a read-only version of a blob at a particular point in time.
* Change Feed: The change feed for a blob provides an ordered, read-only, record of the updates made to a blob. You can use the change feed to monitor these changes, and perform operations:

**\* File Storage:** File Storage enables you to create files shares in the cloud, and access these file shares from anywhere with internet connection. It works on SMB 3.0 MAX size of a single file is 1 TB, it supports 2000 concurent connections to a single file.

- Tools for file storage: AzCopy, Azure File Synce Copy, Azure Portal.

**NOTE:** Don't use Azure File Storage for files that can be written by multiple concurrent processes simultaneously. Multiple writers require careful synchronization, otherwise the changes made by one process can be overwritten by another.

**\* Cosmos DB:** A Multi-model NoSQL database management system. Cosmos DB manages data as a partitioned set of documents. A document can hold up to 2 MB of data.

- Documents in a Cosmos DB are organized into containers. The documents in a container are grouped together into partitions. A partition holds a set of documents that share a common partition key.

- Cosmos DB & Table Storage both uses partiton ID. Unlike Azure Table storage, documents in a Cosmos DB partition aren't sorted by ID. Instead, Cosmos DB maintains a separate index. This index contains not only the document IDs, but also tracks the value of every other field in each document. This index is created and maintained automatically.

- Cosmos DB automatically allocates space in a container for your partitions, and each partition can grow up to 10 GB in size.

*Describe common practices for data loading*

1. Ingest data using Azure Data Factory.
   * linked services, datasets, and pipelines.
2. Ingest data using PolyBase (Synapse Analytics, SQL Server)
3. Ingest data using SQL Server Integration Services.
4. Ingest data using Azure Databricks.

**\*Azure Data Factory:** Data Factory can clean, transform, and restructure the data, before loading it into a repository such as a data warehouse (ETL). Once the data is in the data warehouse, you can analyze it.

Data Factory provides an orchestration engine. Data Factory uses orchestration to combine and automate sequences of tasks that use different services to perform complex operations.

**A linked service** provides the information needed for Data Factory to connect to a source or destination.

**A dataset** in Azure Data Factory represents the data that you want to ingest (input) or store.

**A pipeline** is a logical grouping of activities that together perform a task. The activities in a pipeline define actions to perform on your data. For example, you might use an Azure Function activity to run an Azure Function to modify and filter data, or an Azure Databricks Notebook activity to run a notebook that performs more advanced processing.

Pipelines don't have to be linear. You can include logic activities that repeatedly perform a series of tasks while some condition is true using a *ForEach* activity, or follow different processing paths depending on the outcome of previous processing using an *If Condition* activity.

**\* Ingest data using PolyBase:**

PolyBase is a feature of SQL Server and Azure Synapse Analytics that enables you to **run T-SQL** queries that read data from external data sources. PolyBase makes these external data sources appear like tables in a SQL database.

Using PolyBase, you can read data managed by **Hadoop, Spark,** and **Azure Blob Storage**, as well as other database management systems such as **Cosmos DB, Oracle, Teradata,** and **MongoDB.**

\*\* Azure SQL Database does not support PolyBase.

**\* Ingest data using SQL Server Integration Services:**

You can use SSIS to solve complex business problems by copying or downloading files, loading data warehouses, cleaning and mining data, and managing SQL database objects and data. *SSIS is part of Microsoft SQL Server.*

SSIS can extract and transform data from a wide variety of sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations.

SSIS is an on-premises utility. However, Azure Data factory allows you to run your existing SSIS packages as part of a pipeline in the cloud. This allows you to get started quickly without having to rewrite your existing transformation logic.

**\* Ingest data using Azure Databricks:**

Databricks can process data held in many different types of storage, including Azure Blob storage, Azure Data Lake Store, Hadoop storage, flat files, SQL databases, and data warehouses, and Azure services such as Cosmos DB. Databricks can also process streaming data. For example, you could capture data being streamed from sensors and other devices.

Azure Data Factory can incorporate Azure Databricks notebooks into a pipeline. A pipeline can pass parameters to a notebook.

*Describe**data storage and processing with Azure:*

1. Azure Synapse Analytics.
   * T-SQL
   * Spark.
2. Azure DataBricks
3. Azure HDInsight
4. Azure DataFactory
5. Azure Data Lake
   1. Data Lake Store
   2. Data Lake Analytics
   3. HDInsight

**\*Azure Synapse Analytics:** Azure Synapse Analytics is a generalized analytics service. You can use it to read data from many sources, process this data, generate various analyses and models, and save the results. You can select between two technologies T-SQL, Spark.

*It uses clustered architecture*:- Control node, Compute Node.

**\*Azure Databricks**: Databricks can process data held in many different types of storage, including Azure Blob storage, Azure Data Lake Store, Hadoop storage, flat files, databases, and data warehouses. Databricks can also process streaming data*. Databricks uses an extensible architecture based on drivers*.

*A driver is a piece of code that connects to a specific data source and enables you to read and write that source. A driver is typically provided as part of a library that you can load into the Databricks environment.*

**\*Azure HDInsight:** Azure HDInsight is a managed Analytics service in the cloud. *HDInsight uses a clustered model*, similar to that of Synapse Analytics. *HDInsight stores data using Azure Data Lake storage.* You can use HDInsight to analyze data using frameworks such as Hadoop Map/Reduce, Apache Spark, Apache Hive, Apache Kafka, Apache Storm, R, and more.

Hadoop Map/Reduce uses a simple framework to split a task over a large dataset into a series of smaller tasks over subsets of the data that can be run in parallel, and the results then combined.

**\*Azure DataFactory:** Data Factory provides a scalable and programmable ingestion engine that you can use to implement complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.

**\*Azure Data Lake Storage:** Azure Data Lake is a collection of analytics and storage services that you can combine to implement a big data solution. It comprises three main elements:

* Data Lake Store
* Data Lake Analytics
* HDInsight
* Data Lake Store provides a file system that can store near limitless quantities of data. It uses a hierarchical organization (like the Windows and Linux file systems), but you can hold massive amounts of raw data (blobs) and structured data. It is optimized for analytics workloads.
* Azure Data Lake Store provides granular security over data, using Access Control Lists. An Access Control List specifies which accounts can access which files and folders in the store. If you are more familiar with Linux, you can use POSIX-style permissions to grant read, write, and search access based on file ownership and group membership of users.
* Services such as Azure Data Factory, Azure Databricks, Azure HDInsight, Azure Data Lake Analytics, and Azure Stream Analytics can read and write Data Lake Store directly.

**\*Azure Data Lake Analytics:** Azure Data Lake Analytics is an on-demand analytics job service that you can use to process big data. It provides a framework and set of tools that you use to analyze data held in Microsoft Azure Data Lake Store, and other repositories. You write jobs that contain queries to transform data and extract insights.

You define a job using a language called **U-SQL**. This is a hybrid language that takes features from both SQL and C#, and provides declarative and procedural capabilities that you can use to process data.

*Describe**Azure Synapse Analytics:*

Azure Synapse is composed of the following elements:

* **Synapse SQL pool:** This is a collection of servers running Transact-SQL. Transact-SQL is the dialect of SQL used by Azure SQL Database, and Microsoft SQL Server. You write your data processing logic using Transact-SQL.

Use SQL pools in Synapse Analytics for the following scenarios:

* **Complex reporting**: You can use the full power of Transact-SQL to run complex SQL statements that summarize and aggregate data.
* **Data ingestion:** PolyBase enables you to retrieve data from many external sources and convert it into a tabular format. You can reformat this data and save it as tables and materialized views in Azure Synapse.
* **Synapse Spark pool:** This is a cluster of servers running Apache Spark to process data. You write your data processing logic using one of the four supported languages: Python, Scala, SQL, and C# (via .NET for Apache Spark). Spark pools support Azure Machine Learning through integration with the SparkML and AzureML packages.
* **Synapse Pipelines:** A Synapse pipeline is a logical grouping of activities that together perform a task.
* **Synapse Link:** This component allows you to connect to Cosmos DB. You can use it to perform near real-time analytics over the operational data stored in a Cosmos DB database.
* **Synapse Studio**: This is a web user interface that enables data engineers to access all the Synapse Analytics tools. You can use Synapse Studio to create SQL and Spark pools, define and run pipelines, and configure links to external data sources.
* **Note**: Any data stored in Azure Synapse Analytics can be used to build and train models with Azure Machine Learning.
* On-demand pools only allow you to query data held in external files. If you want to ingest and load the data into Synapse Analytics, you must create your own SQL pool.
* Spark pools and SQL pools can coexist in the same Azure Synapse Analytics instance.