Azure Data Fundamental DP-900

HASSAN RAFIQUE *MSc Data Science*

- Describe core data concepts
- Identify considerations for relational data on Azure
- Describe considerations for working with non-relational data on Azure
- Describe an analytics workload on Azure
- Azure Synapse Analytics
- Azure cosmos DB
- Azure Data Factory
- Stream and Batch processing
- Data warehouse

Azure DP-900

1 - Explore Core Data Concept:

Key-value, Document DB, Graph DB

Transaction DB

Ex: Banking data.
It must deal to ACID.

✓ **Automicity**:
- Execute in single unit
Fail or Pass.

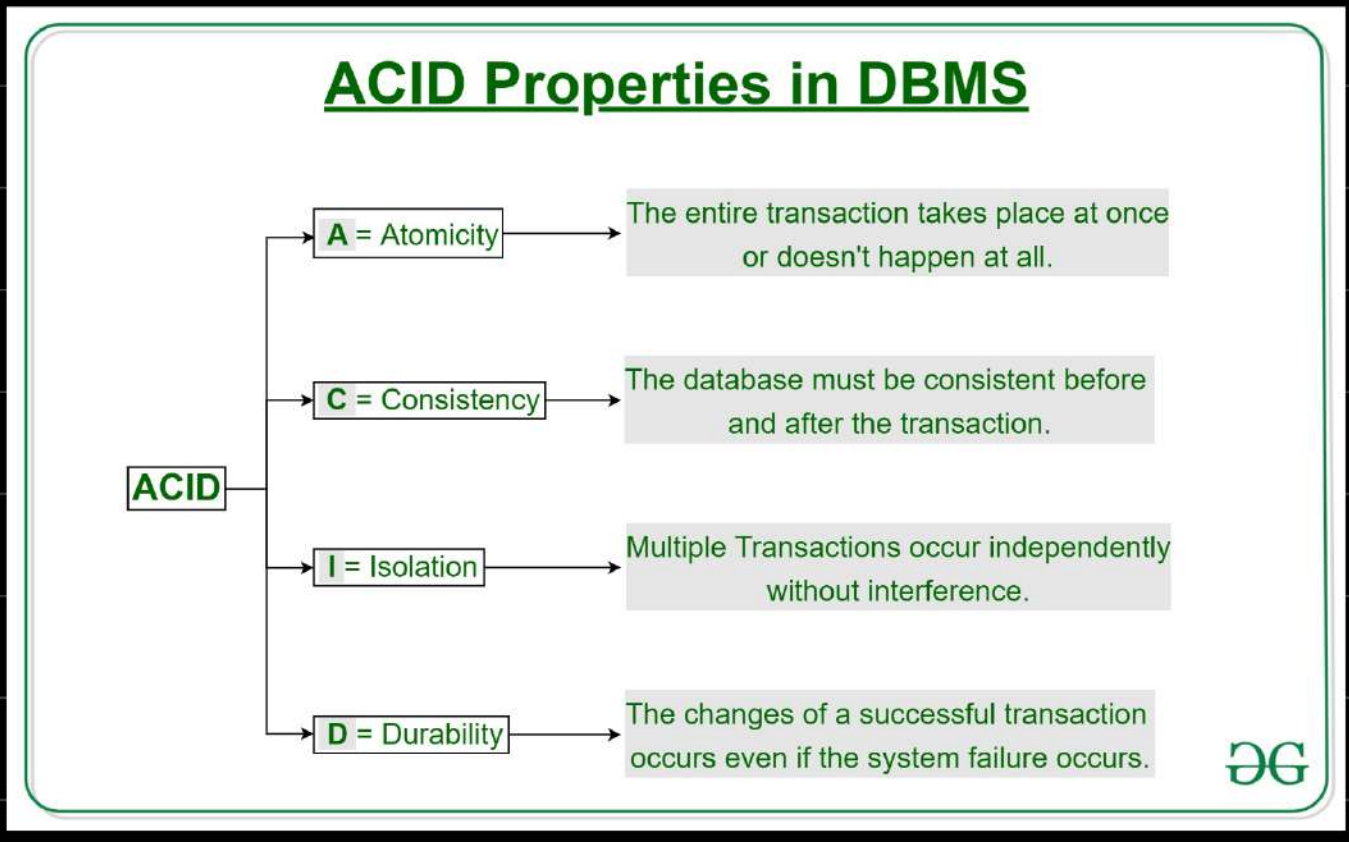
✓ **Consistency**:
- Take data from valid state to another state.

state-1 ⟶ state-2

✓ **Isolation**: multiple trasaction occur independ

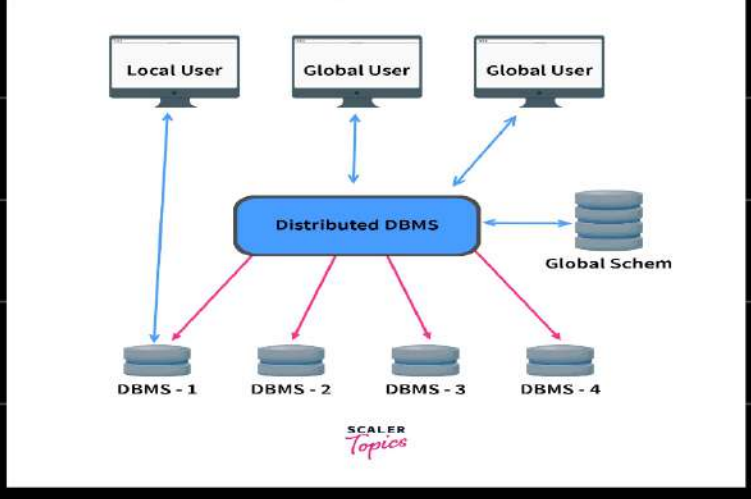
Durabifity:-

Commited transadions renein Commited even system failure occur.



Distributed DB.

Data Store in different location on same or different network.



Batch and Stream Process:-

★ data in the form of group is Batch.
- apply OLAP.
- Data warehouse solution, Monthly analysis

★ Stream = OLTP. trasactional data. New incoming data.

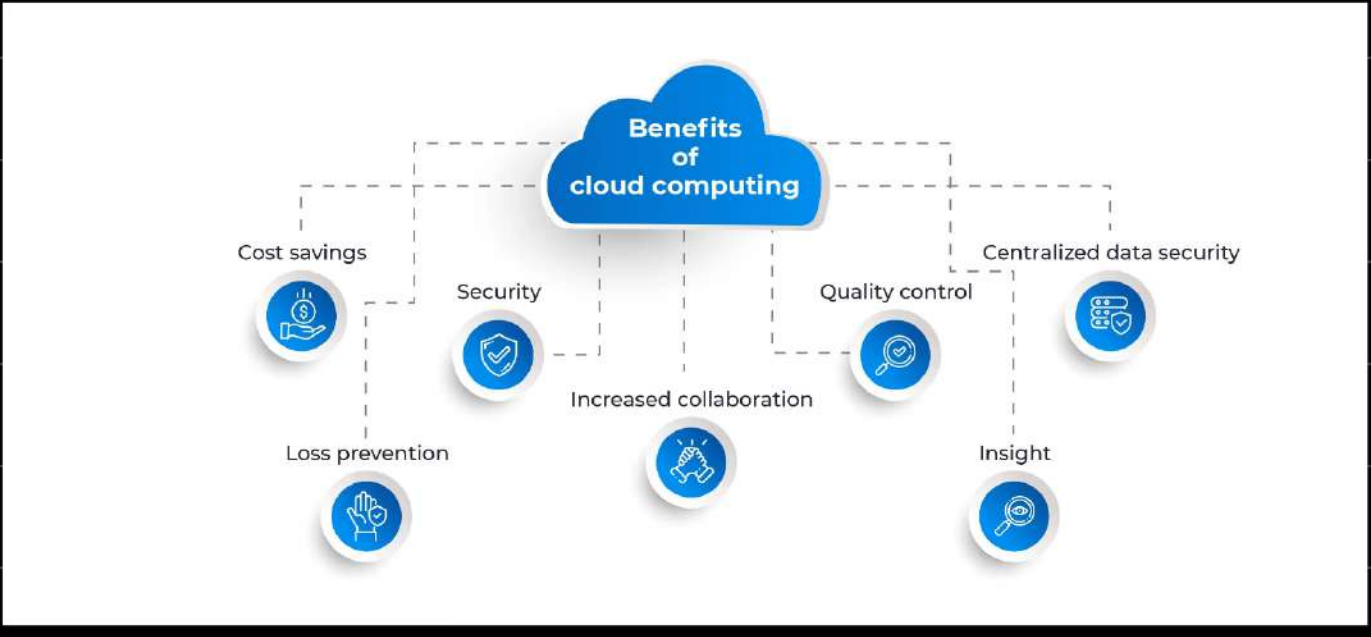
Ex: Stock price
online gaming company
Real-estate website.

1 Role and Responsibilities in Data:

- Data Administrator.
- Handle data
- Modify DB structure

- Control and monitoring
- Backup database
- Data replication.
- Install and upgrad DB Server.

- EX:- Azure Data Studio
  SQL Server Mangmet Studio.



Relational and Non relational data.

Entity, field, Table. Name
Relation DB structure is fixed Format.

★ Primary Key : Identity

✯ Foreign Key :-

Primary key in an other table called foreign key.

✯ Cluster Index / Non Cluster Index.

- Insert data in table with order Formate
- Fast search
- Primary key creat cluster automatically.

ID	Name
1	A
2	B
3	C

✯ View :-

Immediate accel to relative data.
A virtual Table.

"Create VIEW as "myview"
Select "Name" From "mytable";

Non-Relational

Azure Cosmos DB.

- semi-structure
- unstructured data } data structure.

JSON - semi-structure, XML,

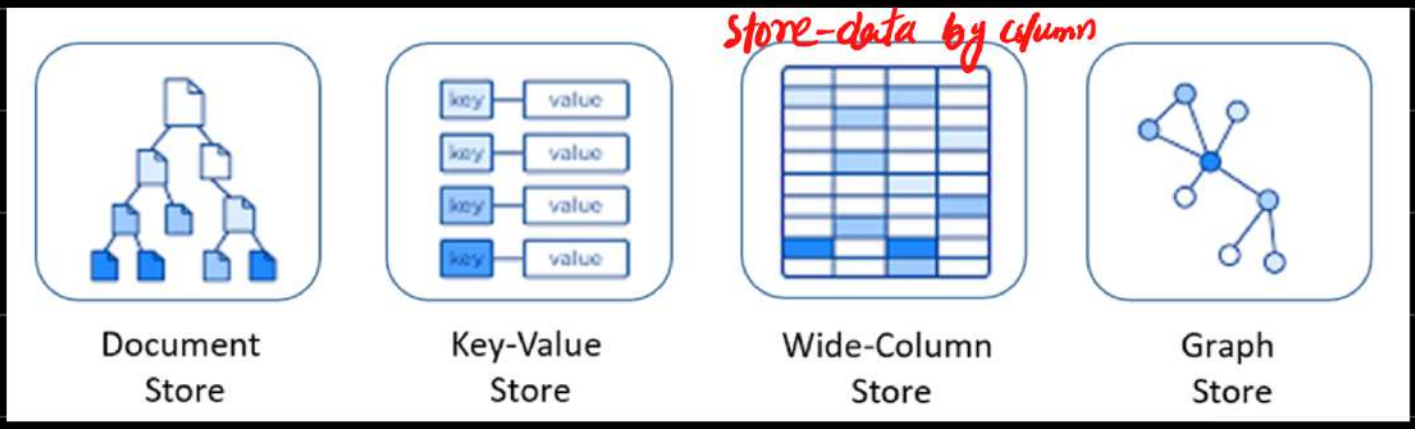
- **AVRO - Apache** –
  Row based format. Header store JSON information store in Binary.
- **Apach ORC**

- **Parquet** :-
  Column data base.

TYPES

Azure Cosmos DB.

- key-value — Azure Table Store.
- Column — Column family – Apach Cassandra
- graph — neo4J,
- Document – Cosmos DB.



Ingestion → Process → Explore

ETL, ELT.

ELT use for complex data model.

Azure data factory.

Cloud base _ data transform tools.

2- SQL:

DBMS:-

on premiss _ Cloud.

SQL Server.

Stored Procedure

linked server (SQL server we join to one server to another server)

PaaS. Azure data Services are available to create.

< SQL, Maria dB, Postgress > Relational-DB

My SQL server are populer tools.

On-Premonis Data —migrate→ Azure cloud. Data

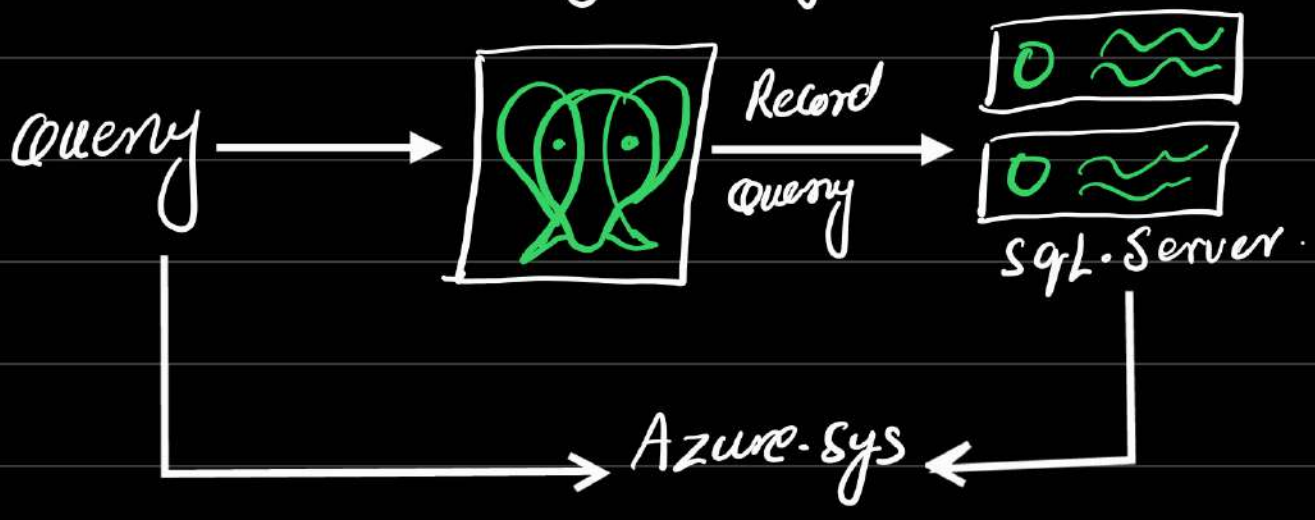
- Azure resources manged by server.
  - Availability
  - Scalability
  - Security. [AD, key vault]
  - threat detection
  - Automatic update

SQ server login ⟶ Azure AD

SQL server managed is good for connecting more server.

Postgres (Pg_admin tool)

- Hybrid relational-object database
- Relational and non-relational store
- Store geometric data [2d-data]
- Pgsql own Query Language.



- Support ultra heigh performence.
- Create read Replices in Azure to 5 number.

Virtual Network.

Azure VM ↔ < ... > ↔ Azure Service
AVN

RBAC:- Role base access control.

allow resources who are use Resource.

owner, Reader, Contributor.

Azure DB Communicate over Port 1433.

3 Azure Cosmos DB

Azure database services for non relational.

(1) Az Cosmos DB.
(2) Table Storage
(3) Blob Storage
(4) File Storage [Continuous Availability]
[tranfer legacy system]

✓ Table Storage.
- un-structure data Store
- Easy and Scalable.
- Partion key, Rowkey.

✓ Blob Storage.
image, video, Storage
① - Block blob
② - Page - 8Tb - VM

③ – appendel –

Blob ⟶ Container ⟶ Blobs

✓ File Shore :-
- Data replicated
- Scalability
- move data for legacy system to Cloud.
- automatically handle delta.
- Back-up ✓ File

✓ Cosmos DB -
- Multimodel NOSQL system. I store document data.
- API Cosmos DB support.
  - SQL API
  - Mongo db API
  - Table API
  - Cassandra API
  - Gremlin API (graph API)

Document in Cosmos db

Partion-key — Container.

- automatically allocating space.
- Height availability
- Replicas

NON-Relational.

Provision:-
action of providing and supply something
like : Azure-portal

Cosmos DB → Container → data

Data Lake Storage:

- use to store large amount of data.

Security Component.

Firewall

Active directory (AD)

- user authentication.
- multi-user authentication.

etc = mobile

- Add user.
- Access Control.

COSMOS-DB

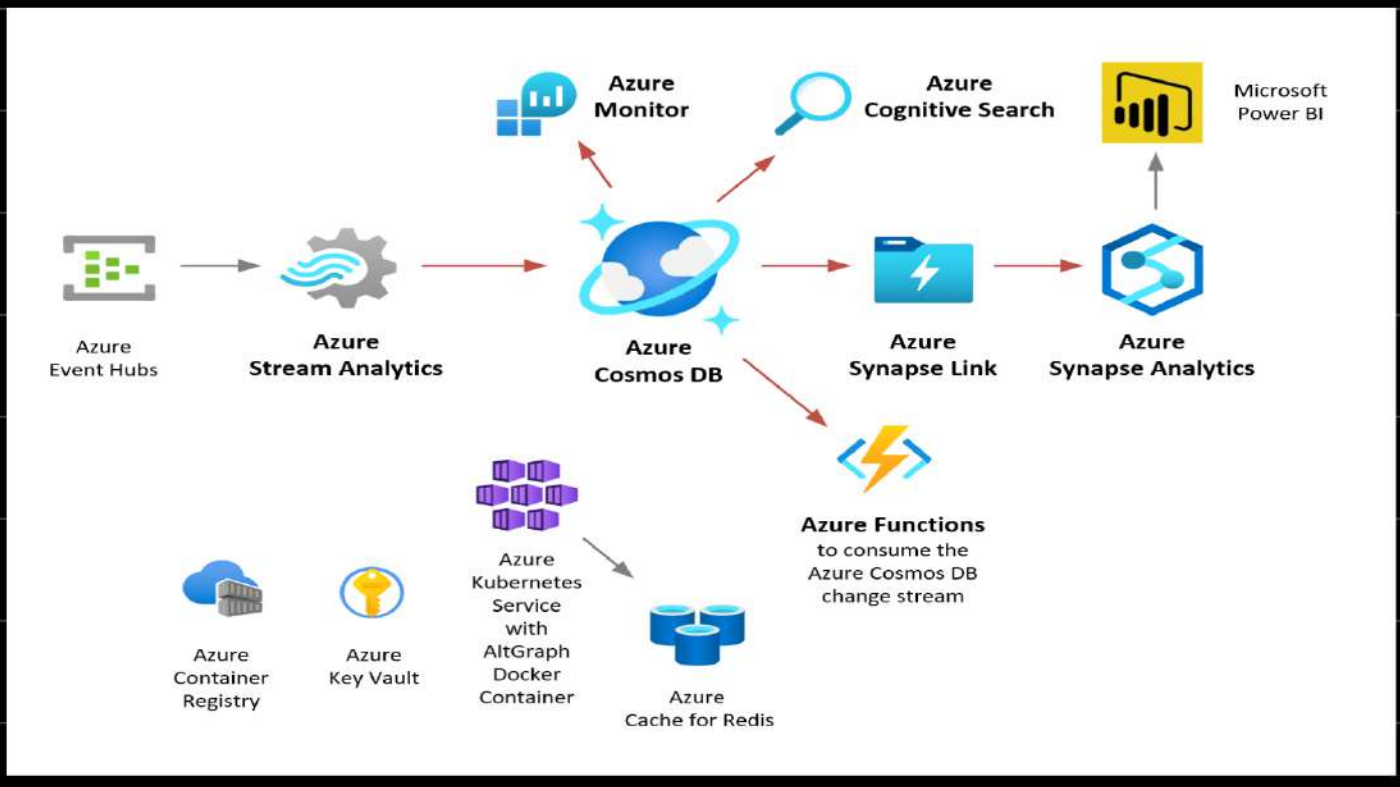
- Replication
  - in region - 4 time
  - outside region

(☆)

- It manage data as set of document.
- select partition to get all related container.

Partition key	container
1	" ____ "
1	" ____ "

(☆) Native API of Cosmos DB is SQL API.



Data Warehouse:-

- Data ingestion
- Store
- PowerbI
- Data Factory. (handle ETL service)

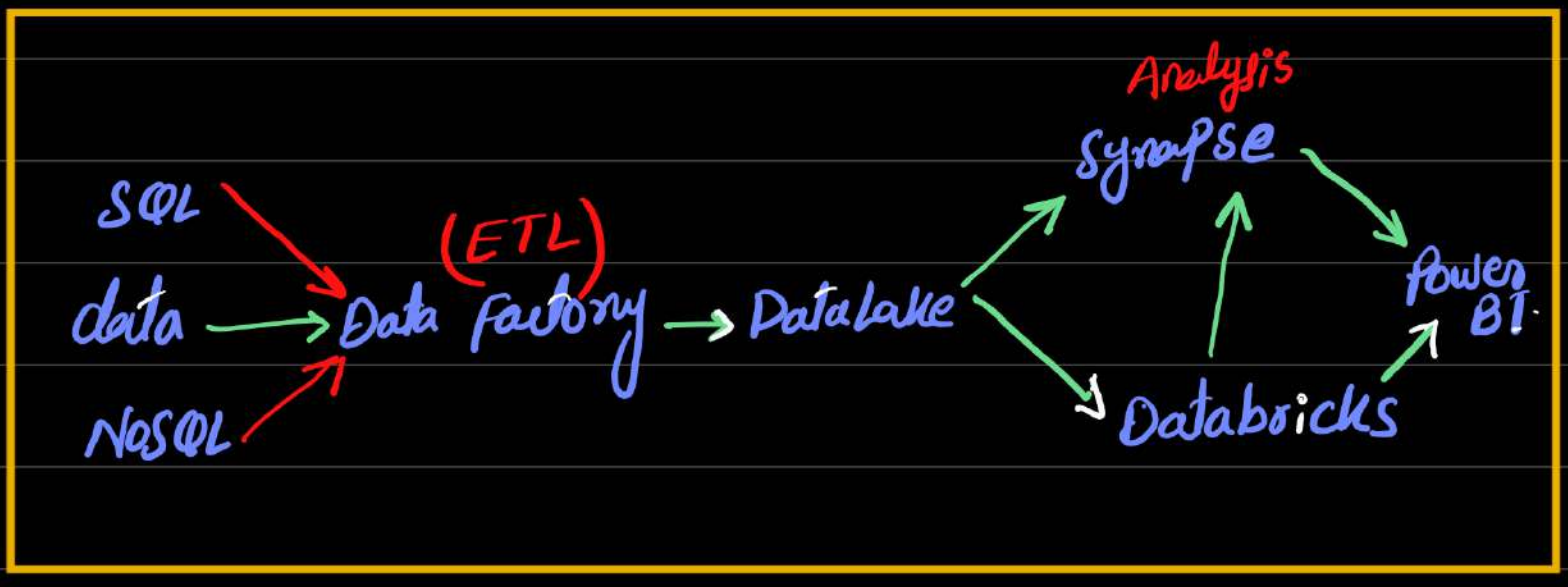
Synapse:-

Datawarehouse + Bigdata Analytics.

4 Azure service to build datawarhous Solution.

Data Factory , Data bricks , Synapse
Powerbi.

Stream and batch data
Data warehouse. Store Structure data.

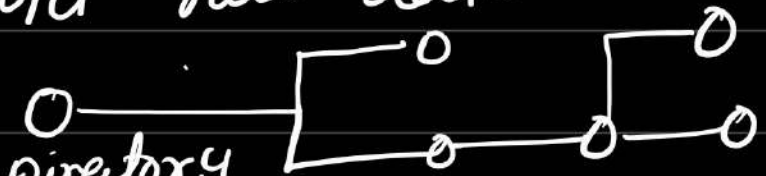


① Data Factory:-

- ETL / ELT Process.
- Ingest data
- Extract Intresting Data.
- Data Integration Service.

② Data Lake

Azure data lake Gen(2)

- hold raw data.
- 

Store data in Folder in Folder.

- hierarchical Name Space.

Note - Blob only mimic data

- HDF

Data lake → Factory → Synapse → Report

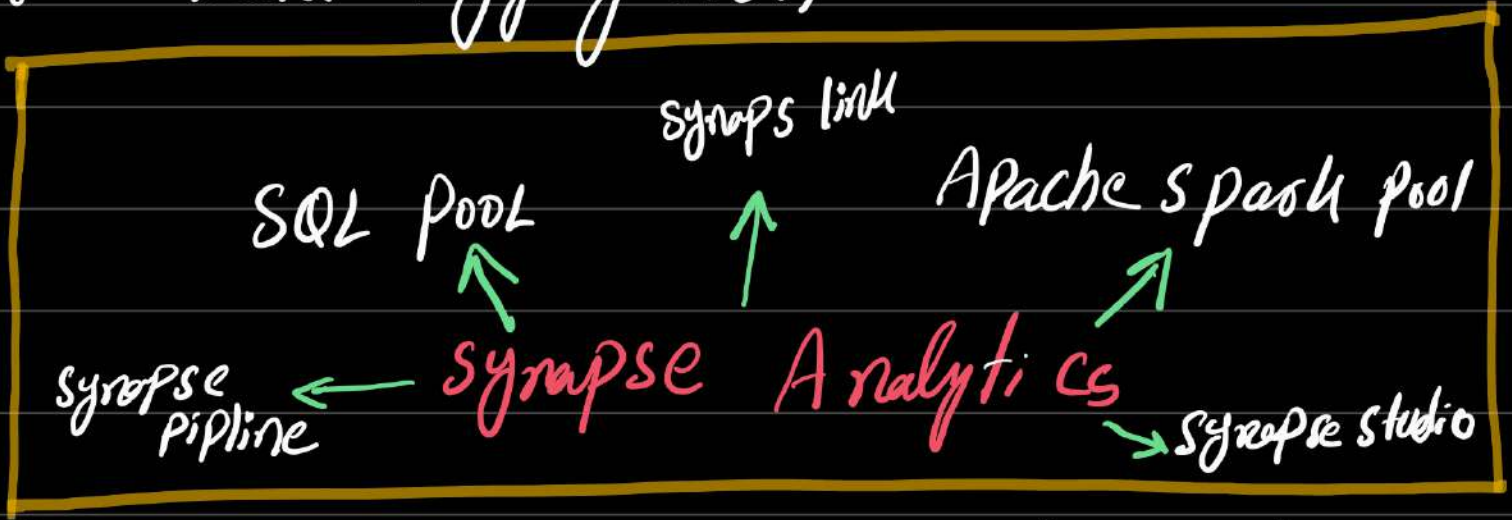
③ Data bricks.

is an apache spark environment running on Azure, use for machine learning.

Data Processing.

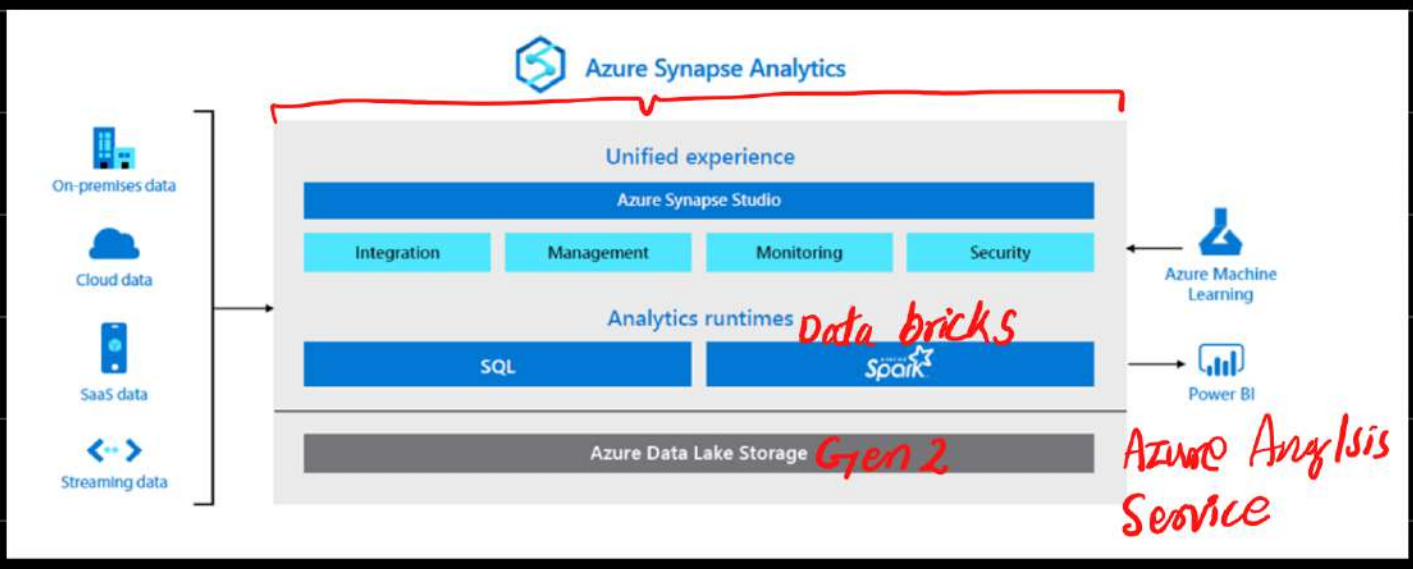
④ Synapse:-

- ingust data from different source
- Data Aggregation



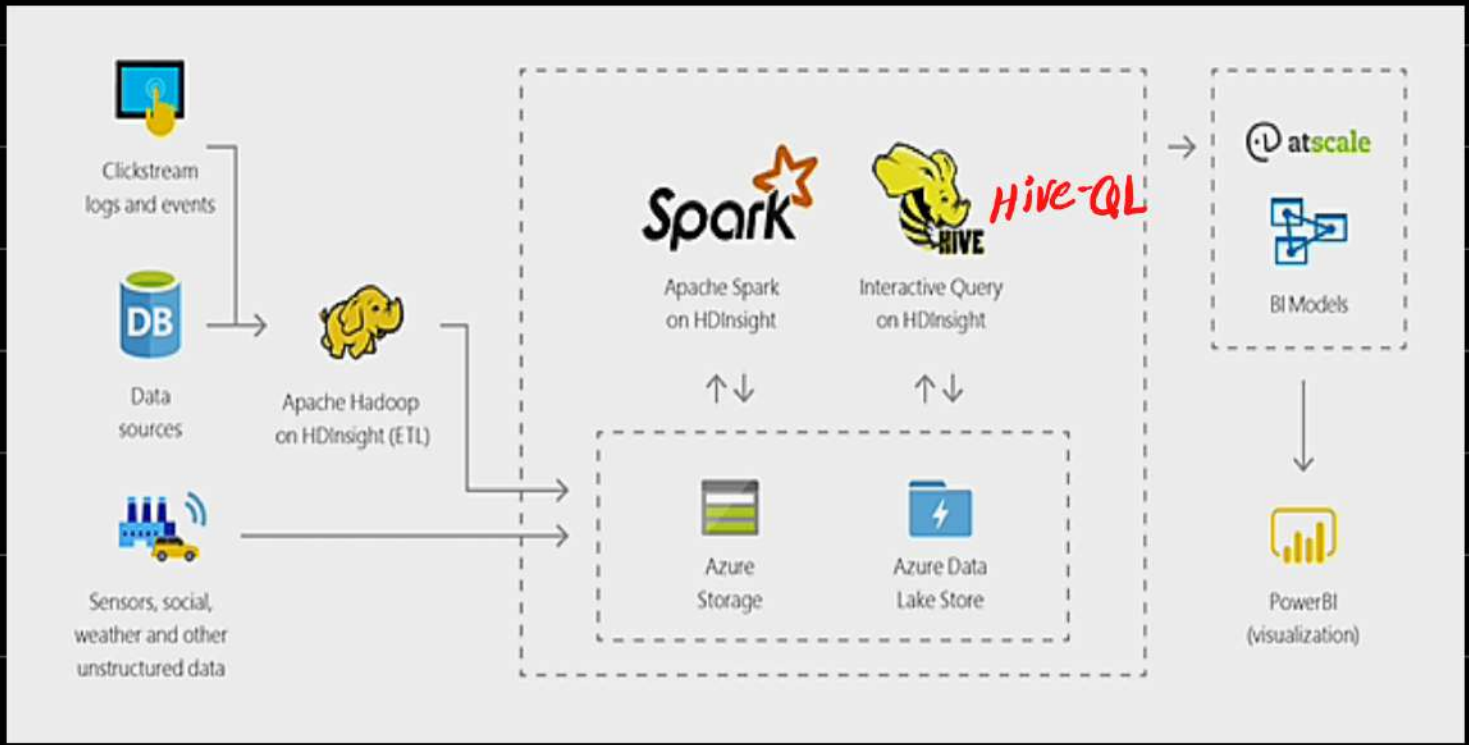
(☆) Spark is in memory computation tool.

- ETL operation



Azure HD insight

- Big data Processing
- implement Cluster model.
- Support Streaming technology such as (Kafka).



Azure Data bricks :-

- parallel process.
- Streaming data
- use Apache Spark to distribute data.

Data Storage and Processing :

- ☆ Azure Synapse Analytics.
- ☆ Data Factory
- ☆ Data bricks
- ☆ Data Lake

} Data Process

☆ PolyBase :-

- Polybase is technique to make external data → look like Sql table.
- Run Queries against these table directly.

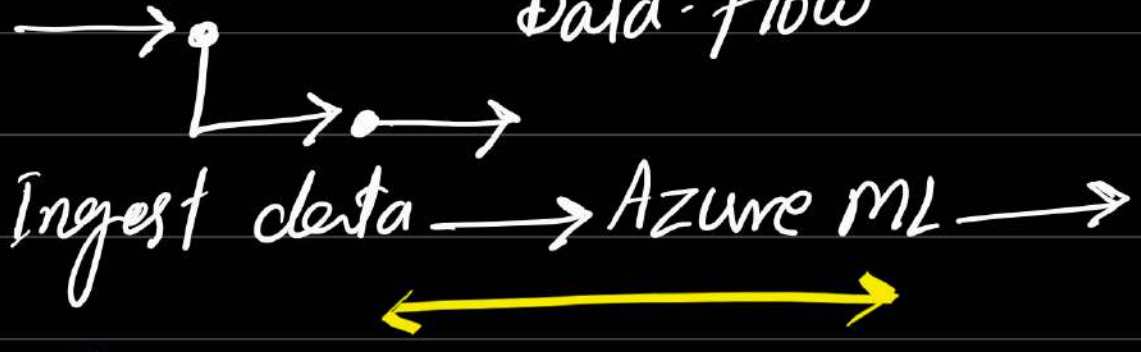
Pool used

- Complex data
- data injust from many source

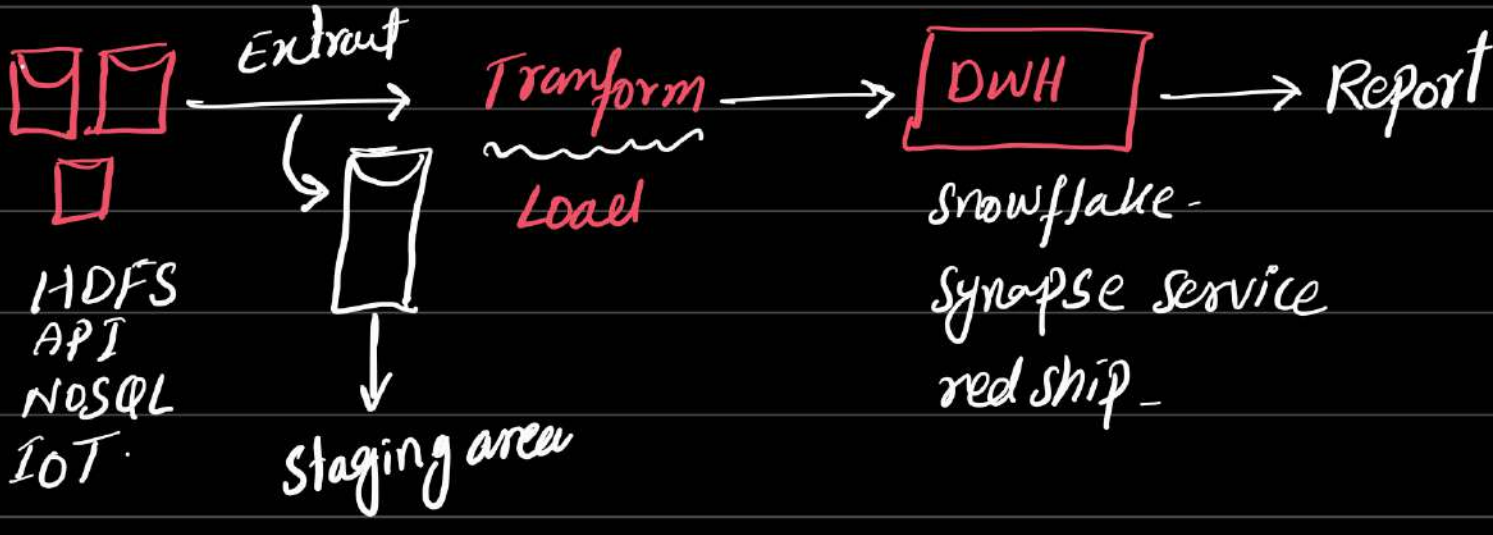
- Synapse Spark to write code baye on Apache Spark.
  To train ML model.
- In memory computing more Faster.

Component of Synapse

Pipeline



Ineuron :



[ELT]

Data Lake
Gen(2)

{ HDFS , Amazon S3
Azure Blob }

Data Lake :-

Categorical data in DWH for Specific Purpose called data-mart.

OLAP / OLTP

OLTP :-

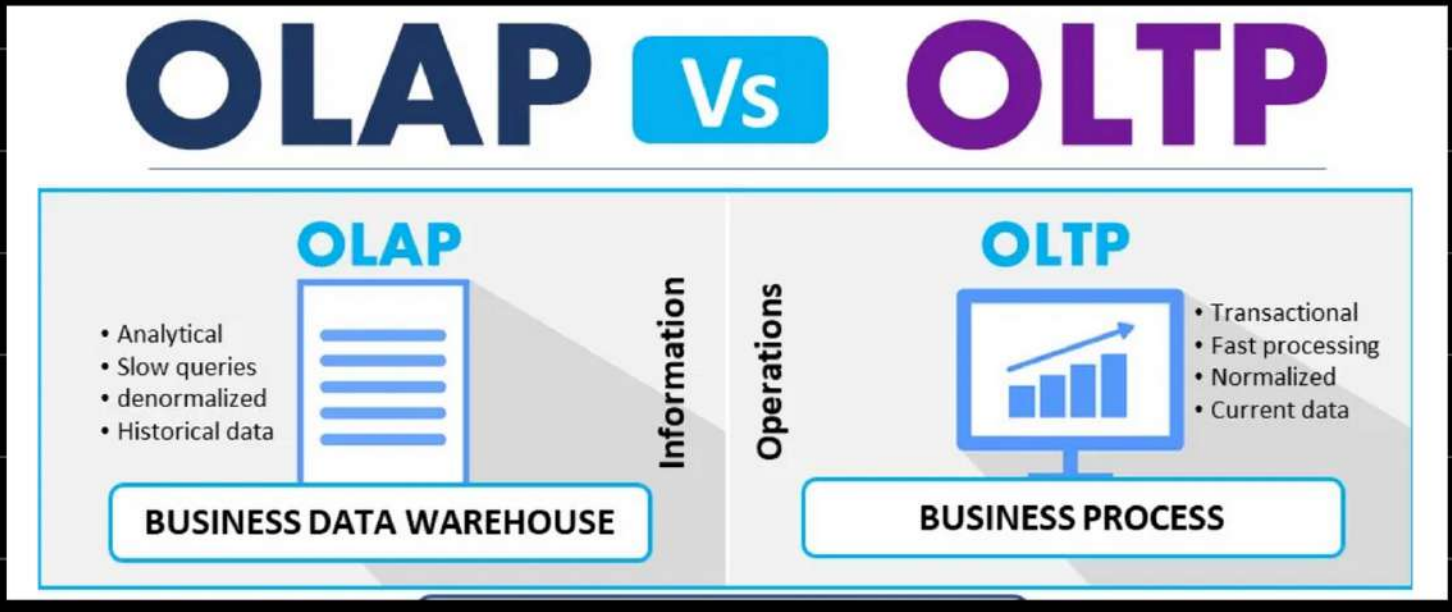
It support ACID Property and

main goal to handle realtime data.

- RDMS
- MSQL, oracle.

OLAP::

- to handle large data.
- Analytics on History data
- Data ware-house
  Hive, Azure synapse.

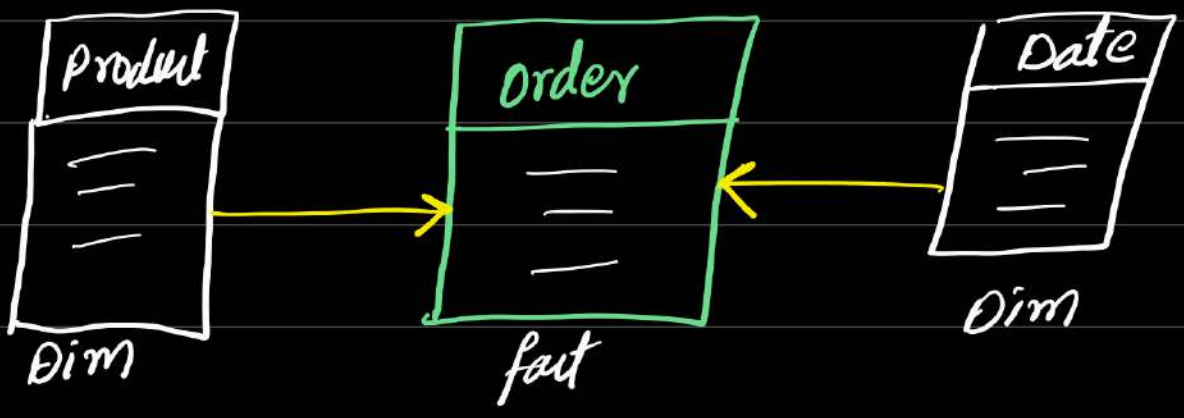


Important Termonology in warehouse.

- Dimension table
- Fact table

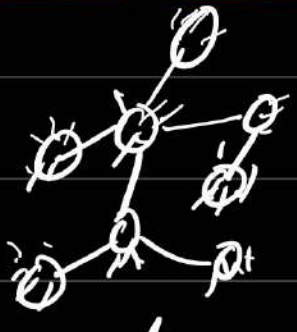
- Star & Snowflake.

Dimension table connected with fact table using foreign key.



Dimension is de-normalize table.

Star Schema Vs Snowflake.



- Store schema denormalize
- Snowflake normalized & less space because remove retudancy.
- Query Fast in Star schema because less number of joins.