15/2/24 :

Database :-

* Collection of data or information
* These are typically accessed electronically and are used to support OLTP.
* DBMS store data in database & enable users & applications to interact with data.

Characteristics

* All databases store information but each database will have its own characteristics.

Ex:- Relational databases - store data in tables (fixed rows & columns)

Non-Relational dbs (NoSQL dbs) - store data in variety of models JSON, BSON,
↓
Binary JSON
key-value pairs, nodes & edges

* Databases store structured & semi-structured data

(1) Security features to ensure data can only be accessed by authorized users.

(2) ACID transactions to ensure data integrity

(3) Query languages & APIs to easily interact with data in database

(4) Indexes to optimize query performance

(5) Full-text search

(6) optimizations for mobile devices

⑦ Flexible deployment topologies to isolate workloads (e.g., analytics workloads) to specific set of resources)
Azure SQL is used here to store data in the cluster

Uses of database

* Applications across industries and use cases are built on databases.

1) Patient medical records
2) Items in an online store
3) Financial records
4) Articles & blog entries
5) sports scores & statistics
6) online gaming info
7) Student grades & scores
8) IOT device readings
9) Mobile appln information

OLAP

* Both data warehouses & datalakes are meant to support OLAP.

* OLAP will
↓
Collect data from variety of sources
↓
data is used to power a range of analytical use cases ranging from BI and reporting to forecasting

Data warehouse

* It is a system that stores highly structured info from various sources.

* Datawarehouses store current & historical data from one or more systems

** Data warehouse is a giant database used for analytics

* Goal - combine disparate data sources

↓

analyze data

↓

look for insights

↓

create BI in form of reports & dashboards

Business Intelligence

tracks records of org

Database examples

1) Relational databases - Oracle, MySQL, Microsoft SQL server, and PostgreSQL

2) Document databases - MongoDB, CouchDB

3) Key-value databases - Redis, DynamoDB

4) Wide-column stores - Cassandra, HBase

5) Graph databases - Neo4j and Amazon Neptune

Characteristics of Dataware House :-

1) Store large amounts of current & historical data from various sources.

2) Range of data ⇒ from raw ingested data to highly curated, cleansed, filtered & aggregated data.

keeping

3) ETL processes move data from its original source to data warehouse.

Extract ⇒ Transform ⇒ Load

* Extract data (raw data from sources)
* Apply transformation ⇒ O/P of transformation is highly curated, cleaned, filtered & aggregated data
* Load data into DW, Data brick cluster, datalake, database
* All these processes are done in one go ⇒ all at a time ⇒ using datapipeline.
* ETL processes move data on regular schedule (Eg:- hourly, daily).

④ DWs have pre-defined & fixed relational schema i.e., structured data

⑤ Some dws support semi-structured data also

⑥ Business Analysts can ~~can~~ Connect data warehouses with BI tools ⇒ to explore data, look for insights & generate reports.

Why use data warehouse?

1) to store large amount of data
2) perform in-depth analysis
3) Due to highly structured nature, analyzing data in data warehouses is relatively straight forward.

Datawaree house Examples

1) Amazon Redshift

2) Google big query

3) Microsoft Azure Synapse

4) IBM Db2 warehouse

5) oracle Autonomous DW

6) Snowflake

7) Teradata Vantage.

Data Lake (ELT process is used)

* It is a repository of data from disparate sources that is stored in its original, raw format

* stores large amount of data (current & historical) of variety of formats - JSON, BSON, CSV,

* main purpose ⇒ Analyze data to gain insights

* Data insights ⇒ knowledge gained by analyzing dat about a specific topic or situation.

Is datalake a database?

* A data lake is a repository for data stored in a variety of ways including databases

* It forms a storage layer of database

* Ex:- Tools like Starburst, Presto, Dermio and Atlas Data lake can give database - like view.

-x-

Datalake characteristics

① store large amounts of structured, semi-structured and unstructured data.

Ex:- Relational data to JSON docs to PDFs to audio files;

② No need of transforming data before adding to data lake (ELT process).

③ Primary users of datalake => based on structure of data

Business analysts => gain insights from structured data.
⇓
look for unexpected patterns & insights.

④ Data in datalake can be processed with variety of OLAP systems and BI tools

⑤ Cost -effective } use

⑥ support machine learning & gain insights from data

Data lake examples

i) AWS S3
ii) Azure Data Lake Storage Gen2
iii) Google cloud storage

} Provide flexible and scalable storage for building data lakes.

~~Key differences b/w~~

Another examples => organizing & querying data in data lakes

1) MongoDB Atlas Datalake
2) AWS Athena
3) Presto
4) Starburst
5) Databricks SQL Analytics

	Database	Data lake	Data warehouse
Workloads	operational and transactional	Analytical	Analytical
Datatype	Structured or semi-structured	structured, semi- " , unstructured	structured / semi structured
Schema Flexibility	Rigid or flexible schema depending on database type	No schema defn required to ingest	Pre-defined & fixed schema definition for ingest
Data freshness	Real time	May not be up to date based on freq of ETL processes	⟹ same
Users	Appln Developers	BA, Appln Dev, Data scientists	Business analysts & data scientists
Pros	Fast queries for storing & updating data	1) Easy storage simplifies ingesting raw data. 2) Schema is applied after 3) Separate storage & compute	Fixed schema makes working with data easy for BA.
Cons	limited analytical capabilities	Requires effort to organize and prepare data for use	difficult to desig Scaling issues