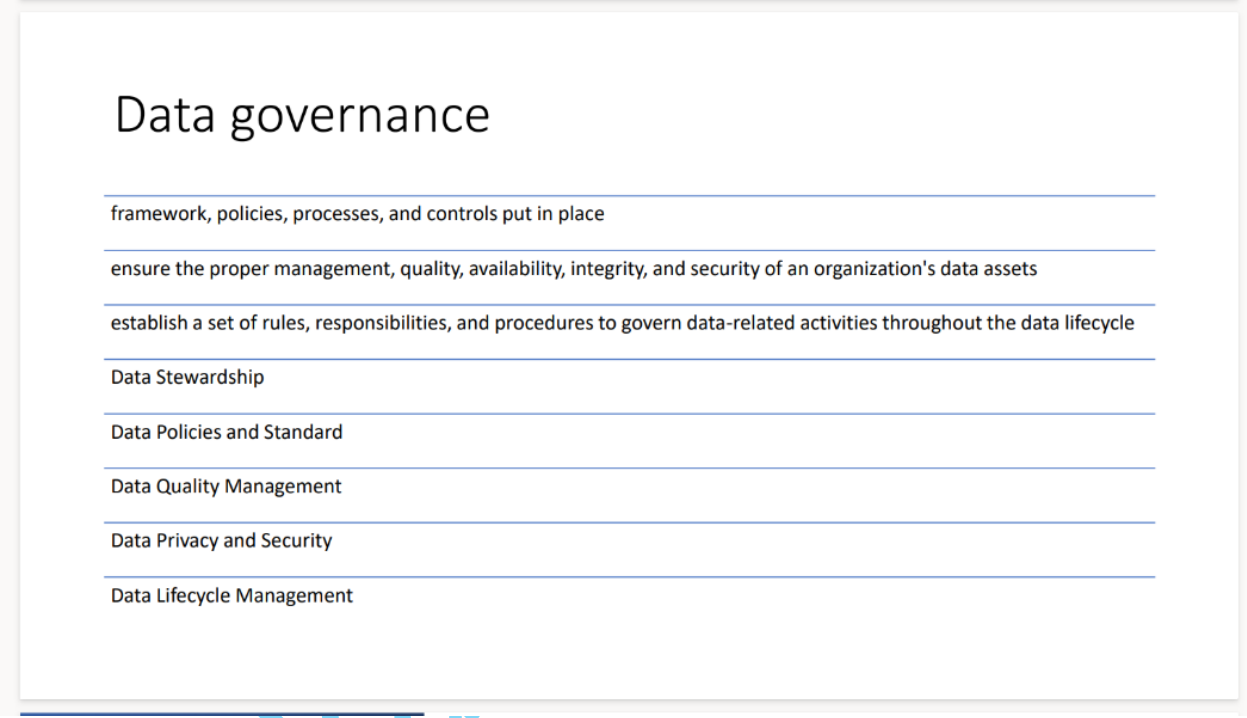
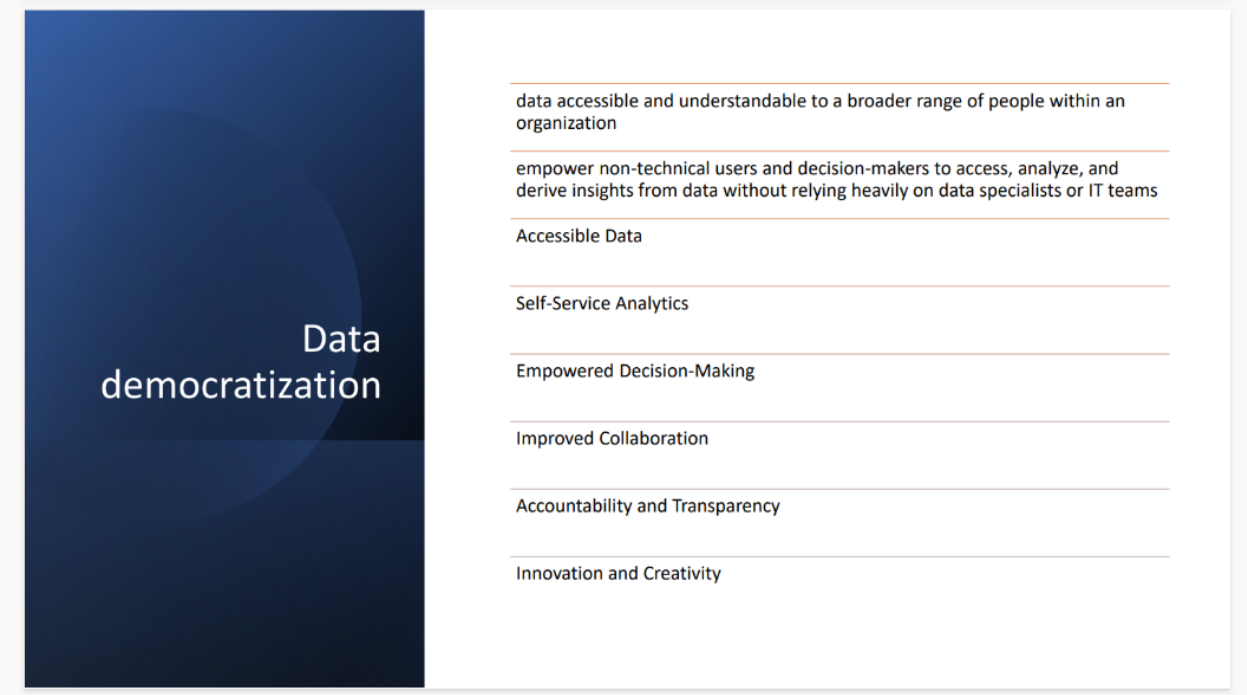
Data governance –

Data governance refers to the overall management of data within an organization. It involves establishing policies, processes, and standards to ensure the effective and efficient use of data. Data governance aims to ensure that data is accurate, consistent, secure, and accessible to those who need it. It also addresses issues related to data ownership, data quality, data security, and data compliance.



Data democratization –

Data democratization refers to the process of making data accessible and usable to a wider range of people within an organization, regardless of their technical expertise or role. It involves breaking down data silos, simplifying data access, and providing tools and resources that enable users to analyze and interpret data for informed decision-making.



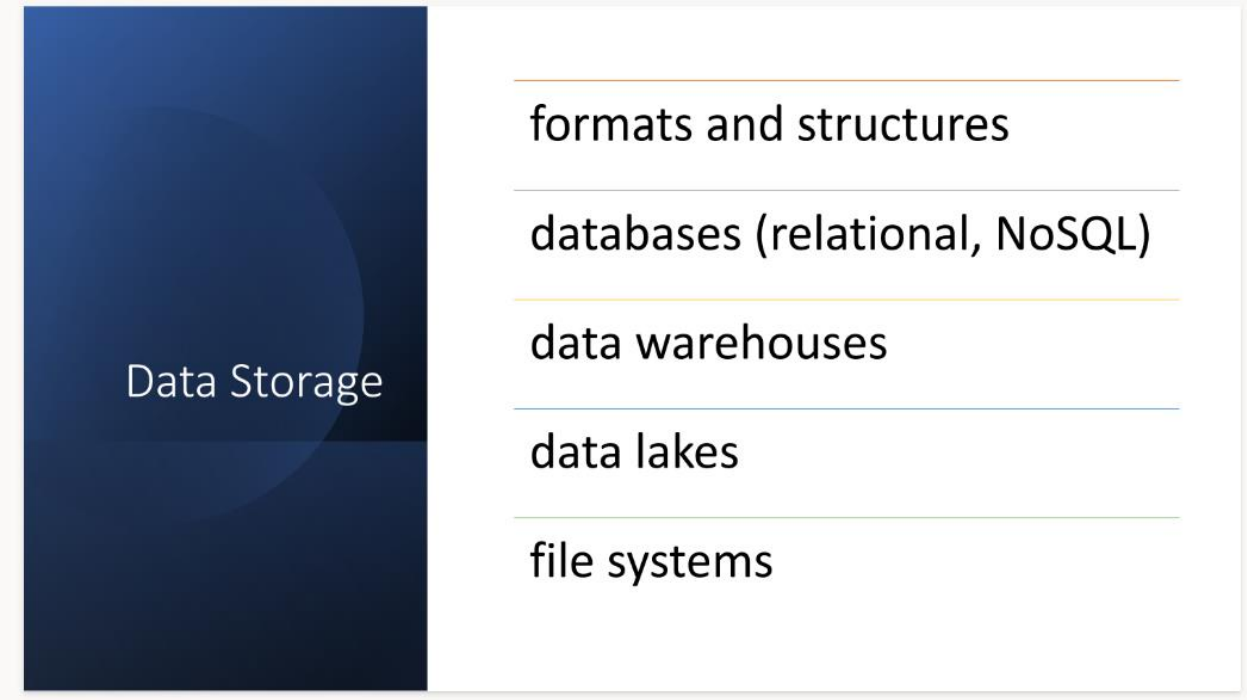
➔

Data Storage –

Data storage refers to the methods and technologies used to store and manage data. It involves the physical storage of data on devices such as hard disk drives (HDDs), solid-state drives (SSDs), and

magnetic tapes, as well as the logical organization and management of data in databases and file systems.

There are various types of data storage technologies, each with its own advantages and disadvantages. HDDs are traditional mechanical storage devices that use spinning disks to store data. They offer large storage capacities at a relatively low cost but have slower read/write speeds compared to SSDs. SSDs, on the other hand, use flash memory to store data and offer much faster read/write speeds, making them ideal for applications that require quick access to data.



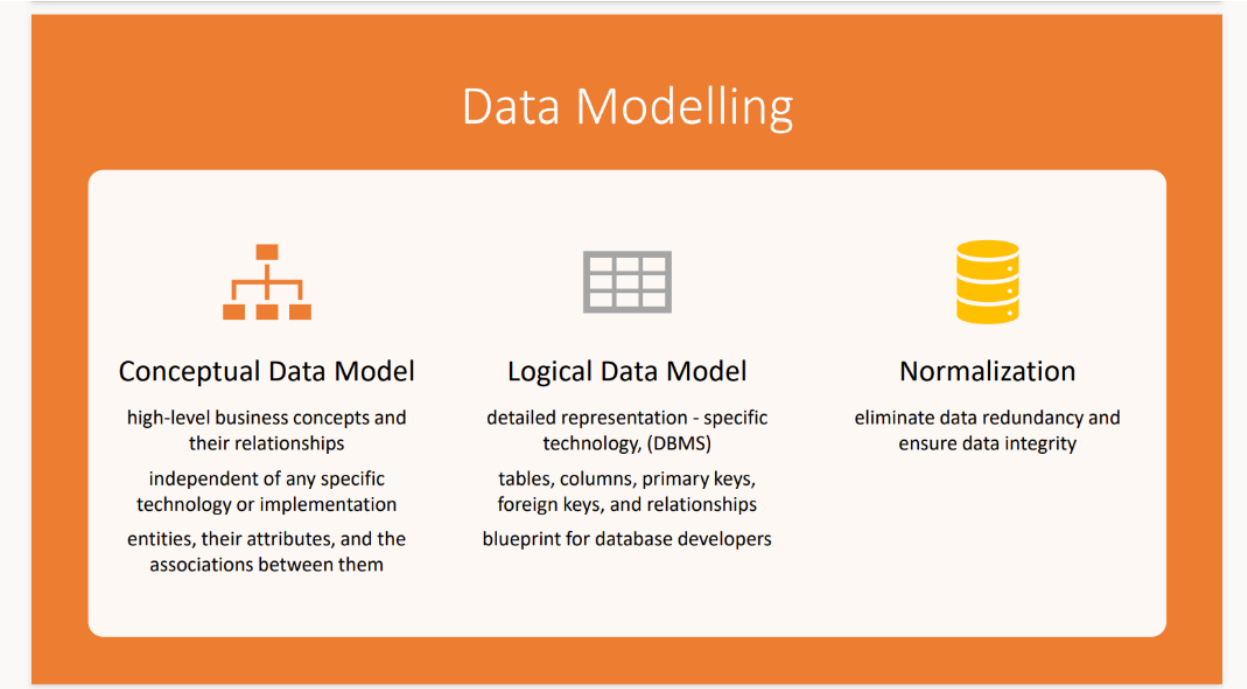
Data Modeling-

Data modeling is the process of creating a data model, which is a logical representation of data and its relationships. It is a fundamental step in data management and is used to organize and structure data in a way that makes it easy to understand and use.

There are many different types of data models, each with its own strengths and weaknesses. Some of the most common types of data models include:

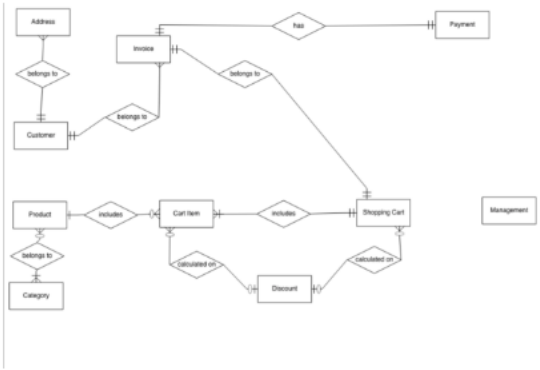
- **Entity-relationship diagrams (ERDs):** ERDs are graphical representations of data entities and their relationships. They are often used to design relational databases.

- **Object-oriented data models (OODMs):** OODMs are data models that are based on the object-oriented programming paradigm. They are often used to design object-oriented databases.
- **NoSQL data models:** NoSQL data models are data models that are designed for non-relational databases. They are often used to design databases that need to store large amounts of unstructured data.



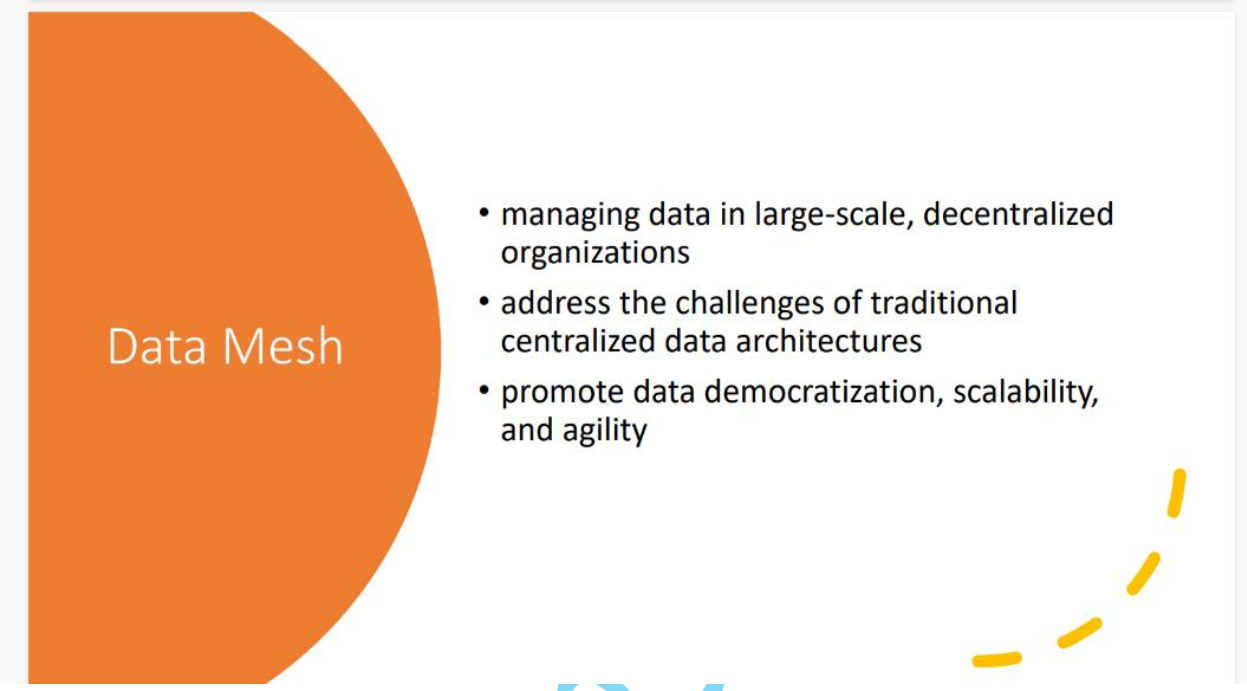
Data Modelling

- Entity-Relationship Diagram (ERD)
  - visual representation of the data model
  - illustrats the entities, attributes, and relationships using symbols and connectors
  - visualize and understand the structure and flow of data



Data Mesh –

A data mesh is a decentralized data architecture that organizes data by a specific business domain—for example, marketing, sales, customer service, and more—providing more ownership to the producers of a given dataset.



Domain-Oriented Ownership

- decentralizing data ownership
- management to domain-oriented teams
- Each team takes ownership of the data within their specific domain
- data collection, storage, processing, and governance
- sense of responsibility and accountability for data quality and usability

Data as a Product

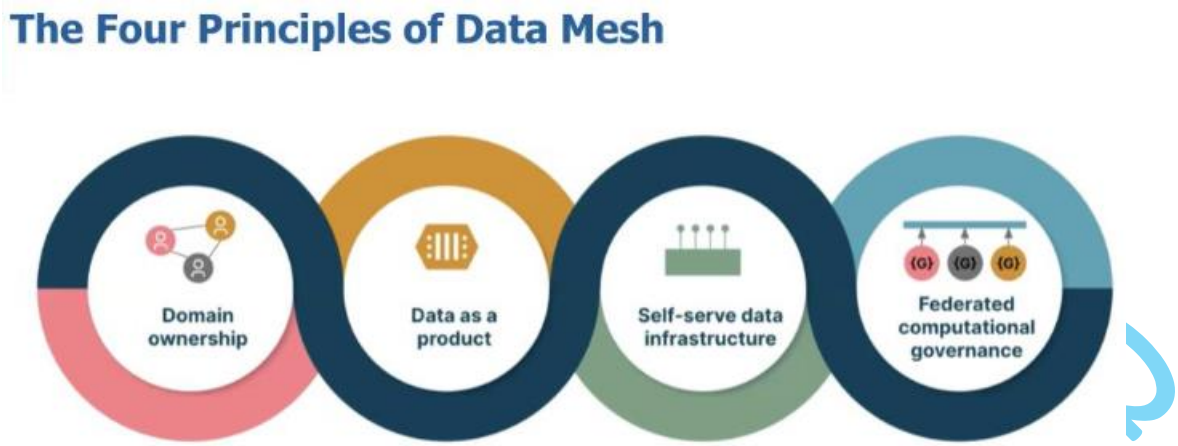
- designed, developed, and delivered to data consumers within the organization
- understanding the needs of data consumers
- creating well-defined data products with clear contracts and APIs
- continuously improving the data products based on feedback

Self-Serve Data Infrastructure

- development of self-serve data infrastructure
- platforms that empower domain teams to manage their data independently
- tooling, frameworks, and platforms for data collection, storage, processing, and analysis
- minimal dependencies on centralized data engineering teams

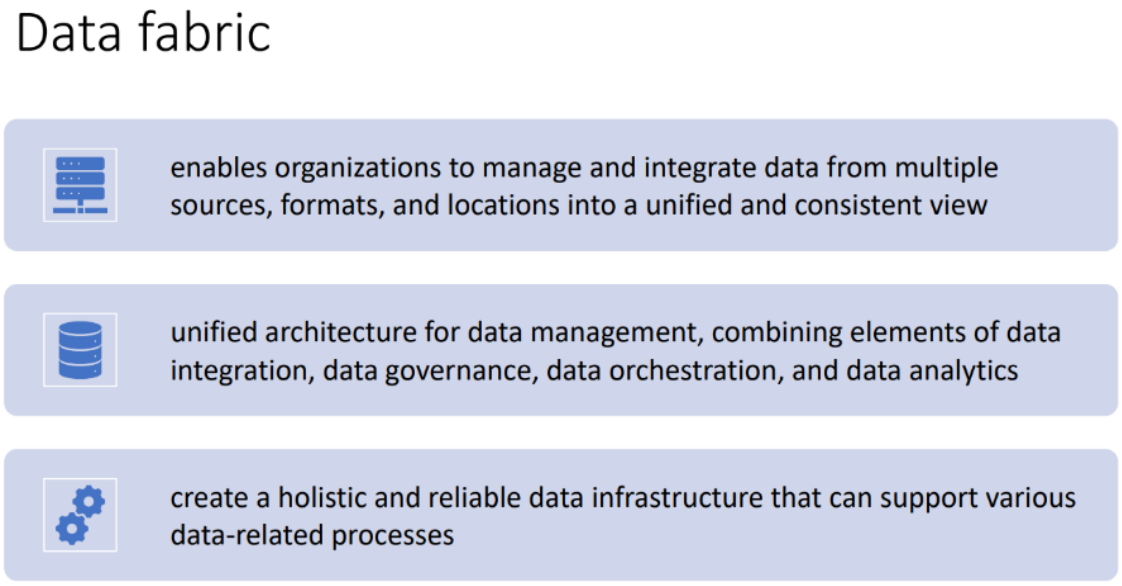
Federated Data Governance

- Rather than relying on a centralized data governance team, governance responsibilities are distributed across domain teams
- defining and enforcing data quality, security, compliance, and privacy within their domain
- Standards and guidelines are established collaboratively



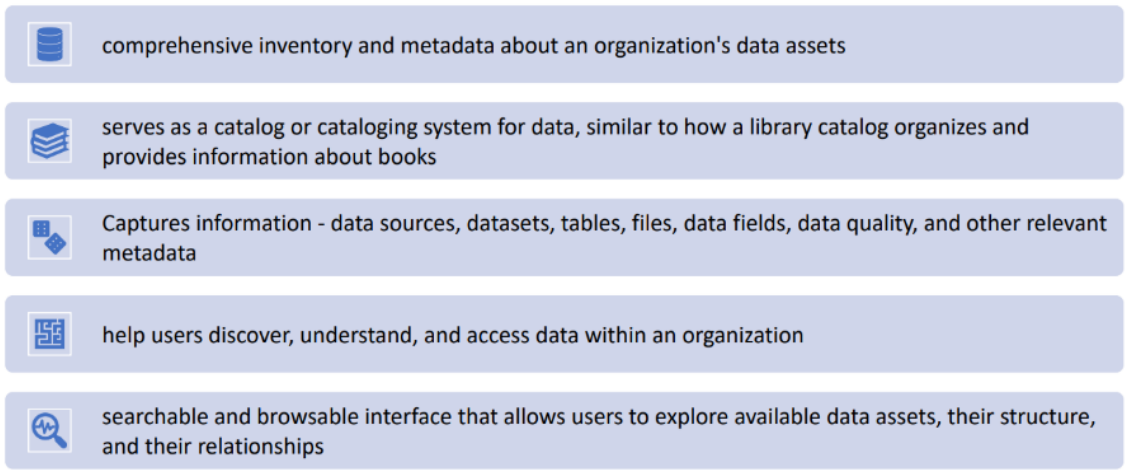
Dat Fabric –

A **data fabric** abstracts away the technological complexities engaged for **data** movement, transformation and integration, making **all data** available across the enterp**rise. Data fabric** architectures operate around the idea of loosely c**oupling data** in platforms with applications that need it.



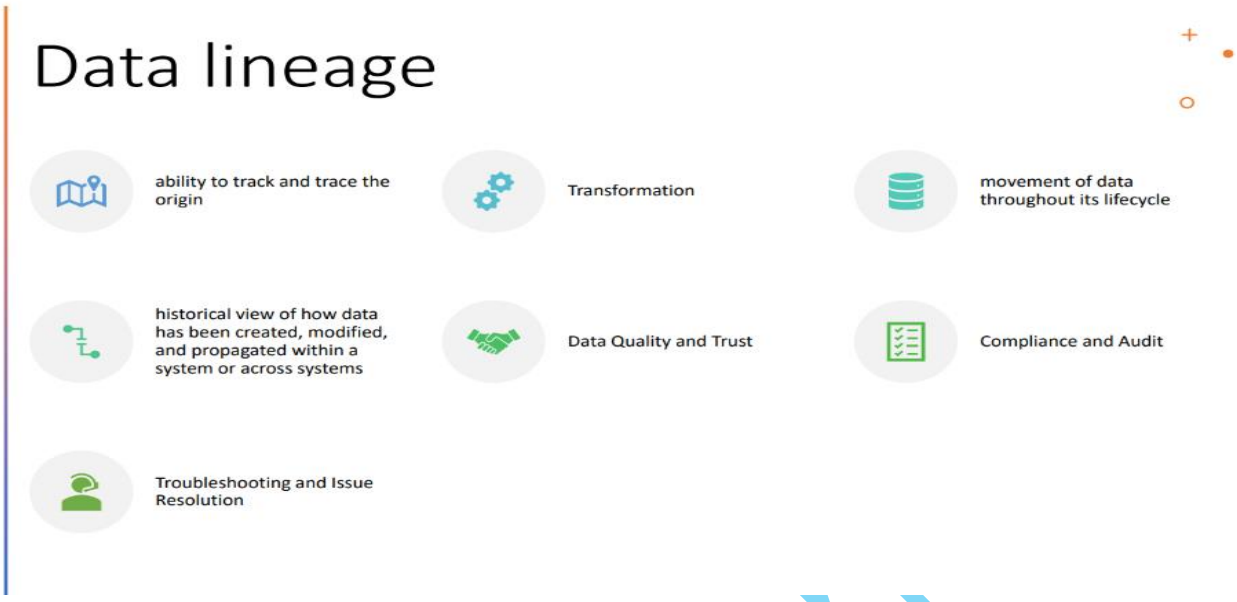
Data Catalog –

A **data catalog** is a detailed inventory of all **data** assets in an organization, designed to help **data** professionals quickly find the most appropriate **data** for any analytical or business purpose. Scale AI workloads, for all your **data**, anywhere. Try watsonx.**data**. **IBM** Knowledge **Catalog**

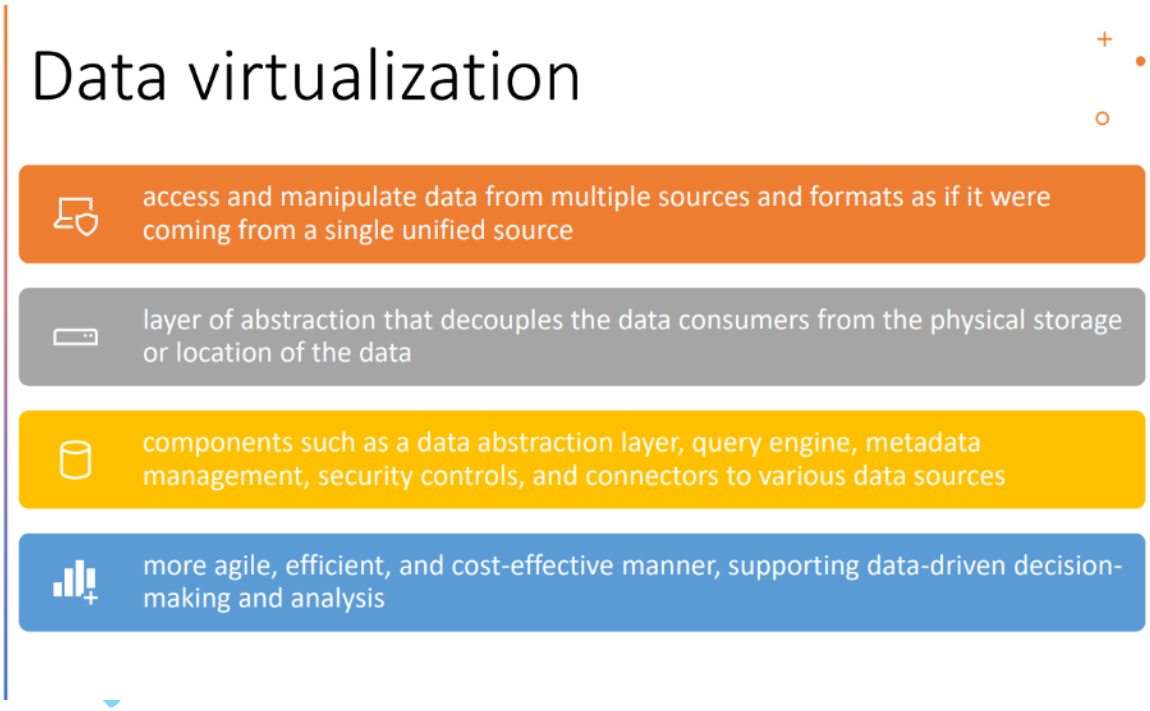


Data Lineage-

Data lineage is **the process of tracking the flow of data over time, providing a clear understanding of where the data originated, how it has changed, and its ultimate destination within the data pipeline.**



Data Virtualization -



Data caching

- temporarily store a copy of frequently accessed or expensive-to-retrieve data in a cache
- high-speed storage layer
- Data Access - requested data is already present in the cache
- Cache Hit and Cache Miss
- Cache Replacement Policies
  - Least Recently Used (LRU), First-In-First-Out (FIFO), and Least Frequently Used (LFU)

➔ Cache replacement policies are algorithms used to determine which data to remove from a cache when it is full and new data needs to be added. The goal of a cache replacement policy is to maximize the efficiency of the cache by keeping the most frequently used data in the cache and removing the least frequently used data :

- Least Recently Used (LRU): The LRU policy replaces the data that has not been used for the longest period of time. This policy is based on the assumption that the data that has not been used recently is less likely to be used in the near future.
- First-In-First-Out (FIFO): The FIFO policy replaces the data that was added to the cache first. This policy is based on the assumption that the data that was added to the cache first is less likely to be used in the near future.
- Least Frequently Used (LFU): The LFU policy replaces the data that has been used the least number of times. This policy is based on the assumption that the data that has been used the least number of times is less likely to be used in the near future.

The choice of which cache replacement policy to use depends on the specific application. For example, the LRU policy is often used for web caching, while the FIFO policy is often used for file caching.