Data Engineering Day 21

The credit for this course goes to Coursera. Click More

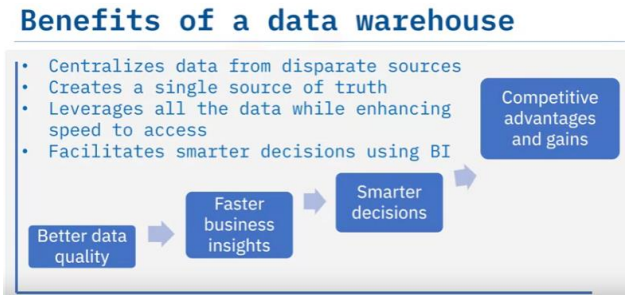
Another link : Azure data Engineer

Getting started with Data Warehousing and Business Analytics

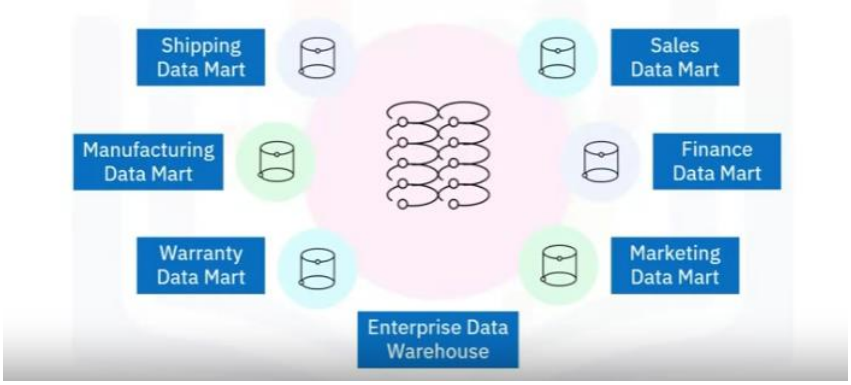
#Introduction to Data Warehousing, Data Mart and Data lake:

- **Data Watehouse:**
  - A data warehouse is **a centralized system designed to enable and support business intelligence (BI) activities, particularly analytics**.
  - It stores large amounts of historical data obtained from various sources, such as transactional systems, relational databases, and other sources, and provides access to business analysts, data engineers, data scientists, and decision-makers through business intelligence tools, SQL clients, and other analytics applications.
  - Data warehouses collect, cleanse, and transform data from multiple sources through a process known as Extract, Transform, and Load (ETL) or Extract, Load, and Transform (ELT).
  - They are used for reporting, data analysis, and informing decision-making in organizations.
  - The figure below represents typical examples of Data warehouse.

  - Data ware house are used for almost all the platforms such as e-commerce, transportations, medical, bangkings, social media, governments etc.
  - Benefits of dataware house includes, centerlizing data from various sources, facilates smater decisions using BI



- **Data Mart:**
  - Datamart is a subset of dataware house built for serving particular roles that include business intiligent or any other fields.
  - Examples includes, sale data mart, finance data mart, enterprice data ware house etc.



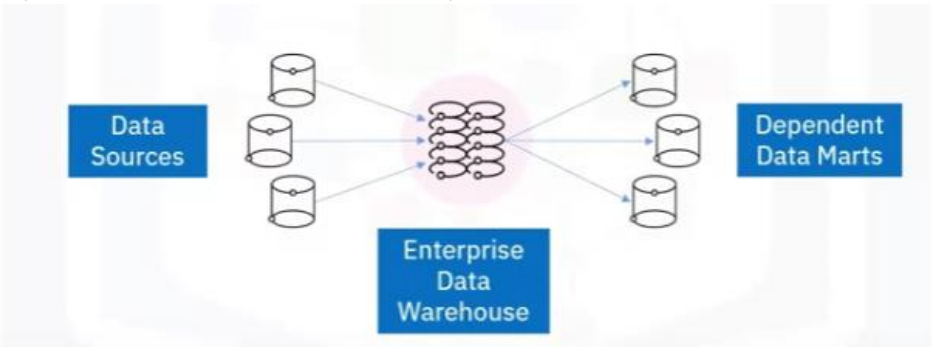
  - Data mart are used for technical decision making, helps end users to focus only one goal, etc.
  - A typical difference between datamart vs database and data mart vs data warehouse is being shown below.

Data repository comparisons

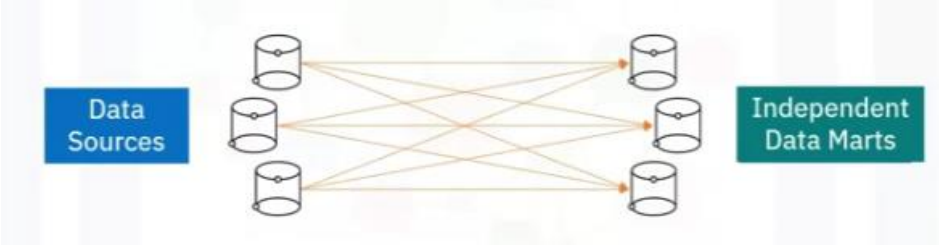
Data Marts	Databases
OLAP systems – read intensive	OLTP systems – write intensive
Use Txn DBs or warehouses as data sources	Use operational applications as sources of data
Contain clean, validated analytical data	Contain raw, unprocessed transactional data
Accumulate history for trend analysis	May not always store history

Data Marts	Data Warehouses
Small data warehouses with tactical scope	Large repositories with broad, strategic scope
Lean and fast	Large and slow

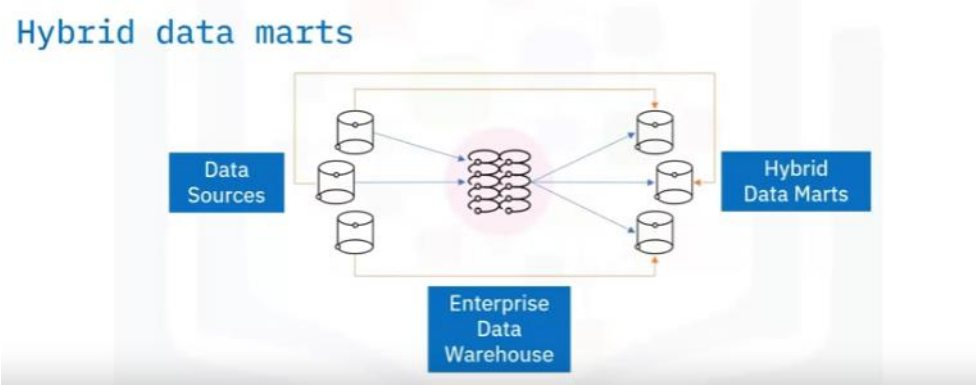
  - There are basically three types of Datamart which are dependent, independent,and the hybrid data mart.
    Dependent data mart draws data from enterprice warehouse.



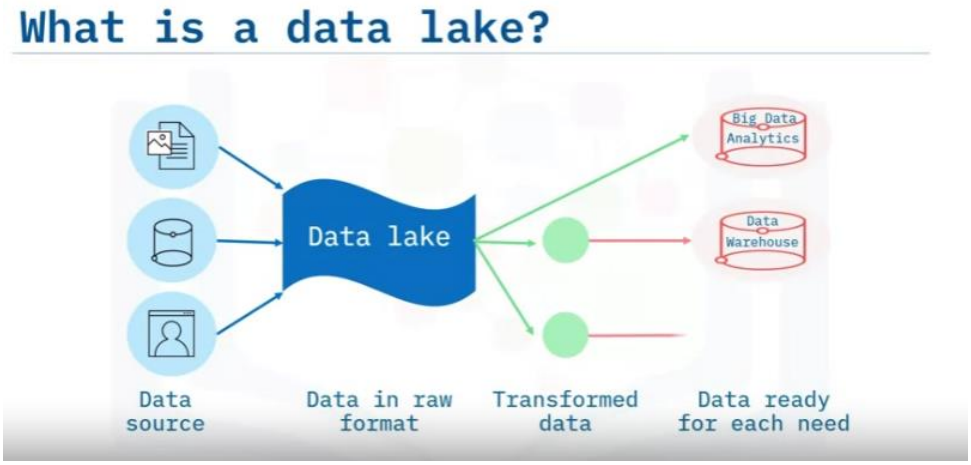
  - Independent datamart are directly created from the sources of the data and it bypass data warehouse.



- Hybride data marts partially depends enterprice dataware house as it combines data from operational systems and other exteral systems from outside.



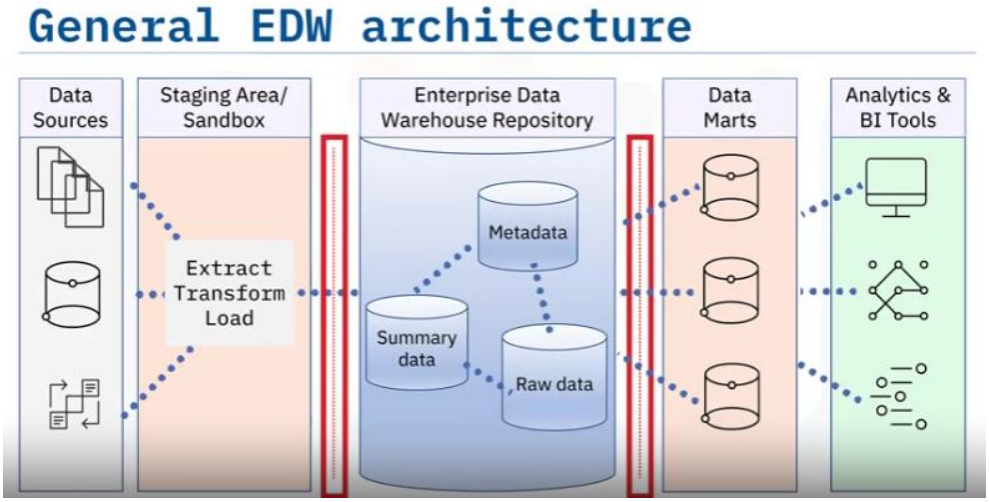
- **Date lake:**
  - A data lake is a storage repository that can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata.
  - Data can be loaded without defining the structure or the scheme of data.
  - It exists as a repository for the raw data.



Data lake	**Data warehouse**
Loaded in its raw and unstructured formats.	has been processed prior to loading
Donot need to define schemas	Need to design a schema before loading
It has a raw data and the quality of the data might be low as data is not organised according to the guidelines.	It follows particluar formates or it basically organised data to be stored there.
Users are data scientist, data developers,and business analytics	Uses are typically data analytics and business analytics.

#Designing modelling and implementing Dataware house:

- **Data ware house architecure:**
  - Generally the architecture of dataware house are built based on their use cases such as report generations and dashboarding, explotery data analysis, automations and machine learning and finally self analytics.



- **Cubes, Rollups, and Materialized Views and Tables:**
  - A data cube, also known as a data cube or an OLAP cube, is **a multi-dimensional data structure that represents data along some measure of interest**.
  - It enables users to analyze data from different perspectives by consolidating or aggregating relevant data into the cube and then drilling down, slicing, dicing, or pivoting data to view it from various angles.
  - Data cubes are commonly used in decision support systems, data warehousing, and business intelligence.

Here are some key characteristics of data cubes: Know more about Data Cube? click me

- **Multi-dimensional**: Data cubes support multi-dimensional data views, making it easier for users to analyze data from different perspectives.
- **Cells, dimensions, and hierarchies**: A typical data cube is composed of cells that represent facts or measures of interest, dimensions that are the categories by which data is classified, and hierarchies that help in drilling data up or down.
- **Association data cubes**: These are data cubes that are formatted as a member of a dimension and have one or more attached dimensions. They associate two dimensions, enabling the end user to group members of one dimension into categories that are defined by the members of a different dimension.
- **Data cube types**: Data cubes can be classified into input, calculation, association, and virtual types, each with specific characteristics and allowed formulas or no formulas.
- **Virtual data cubes**: These data cubes do not store value data in the database, which can benefit the size of the database and the time to load data from it.

Overall, data cubes provide a powerful tool for users to explore and analyze complex datasets, enabling them to make more informed decisions and gain valuable insights.

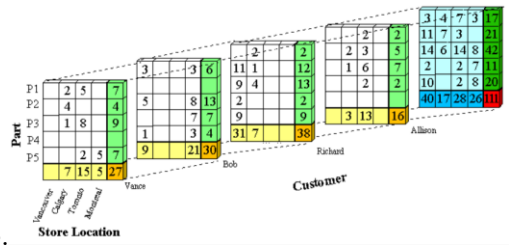
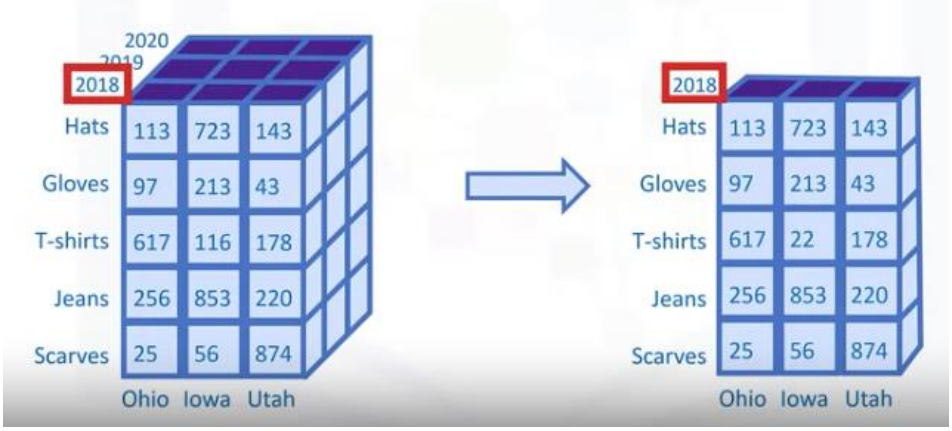


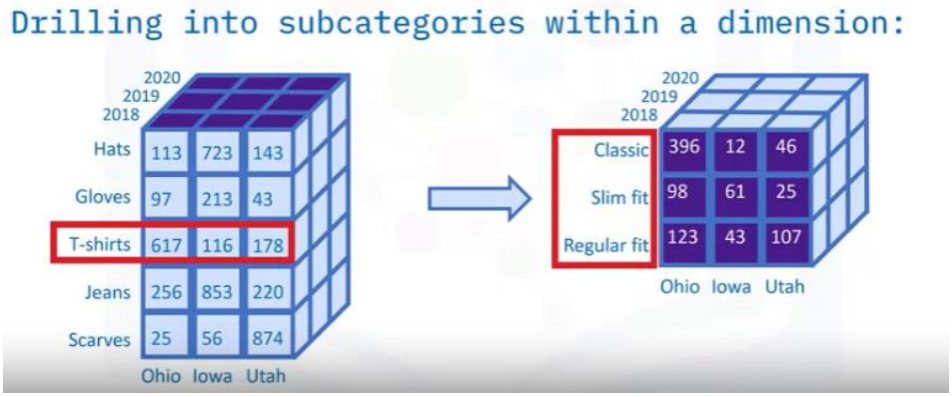
Figure on right handside represents the example of Data Cube.

- Some of the properties that we can apply on a data cube are Slicing,Dicing, Drilling up and down, pivoting, Rolling up. Following the pictures that I have taken screen shot from the IBM Data Engineering course.

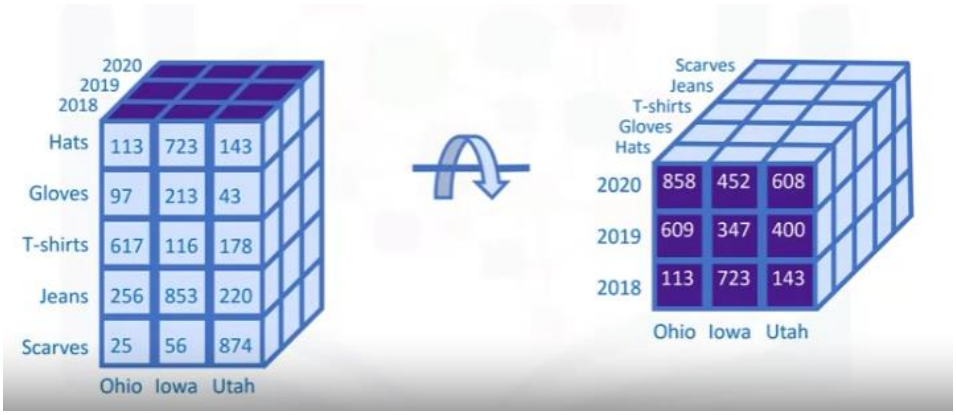
- Slicing a data cube:
  - Slicing redices the cell's dimensions by 1. click me to find more about cubes



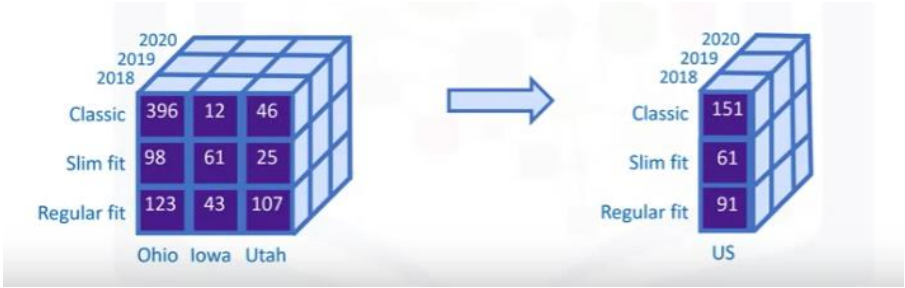
- Drilling a data cube:



- Pvoting data cube :
  - involves the rotations of the data cube by changing the pont of view.



- rolling up :
  - it summarizes the averages of the data cube
  - it aggregate using COUNT, MIN, MAX, SUM,AVG etc.



  - A general example of data cube is shown below.