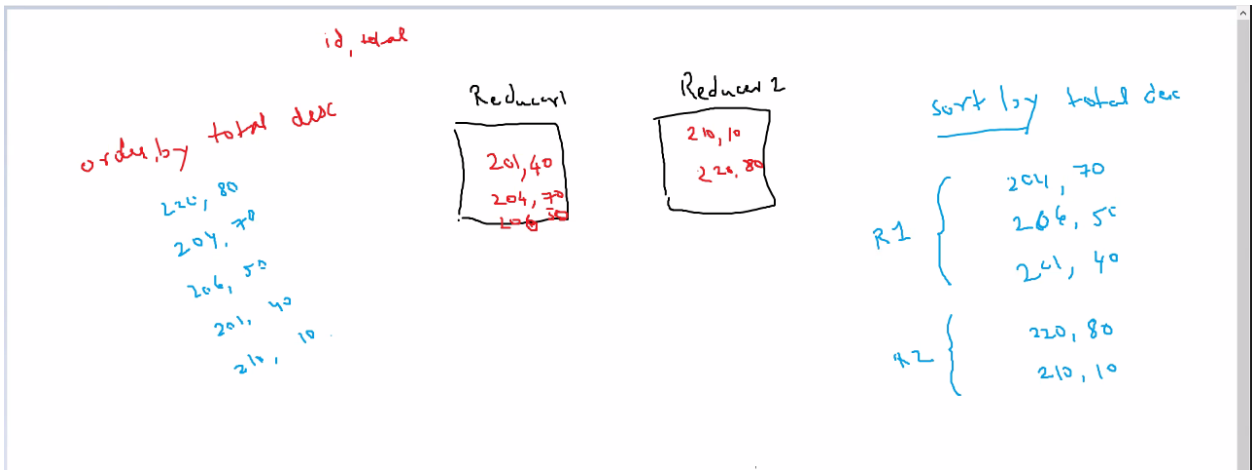
→ ORDER BY vs. SORT BY

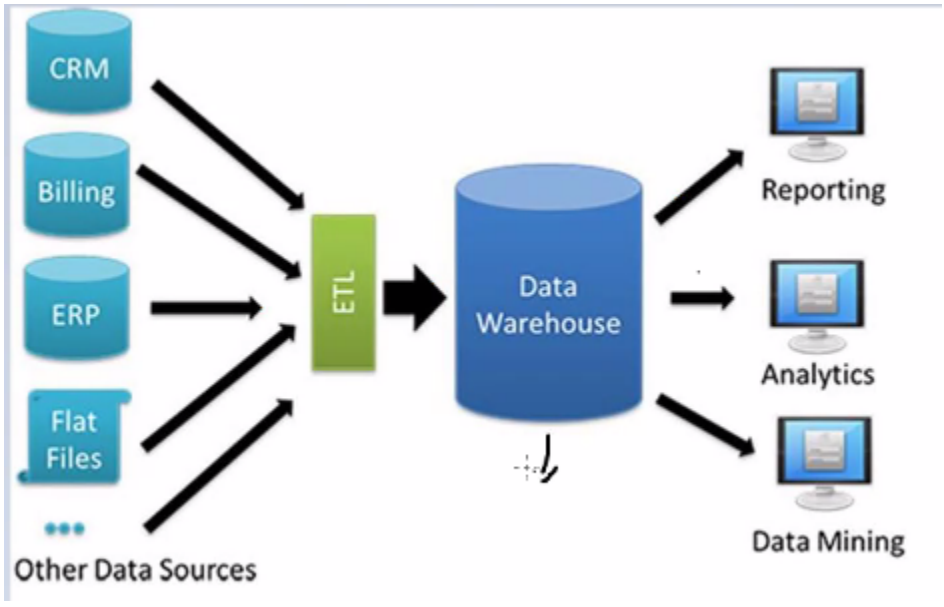
a. ORDER BY : sorts all of data on all reducers
b. SORT BY : sorts only the data within one reducer

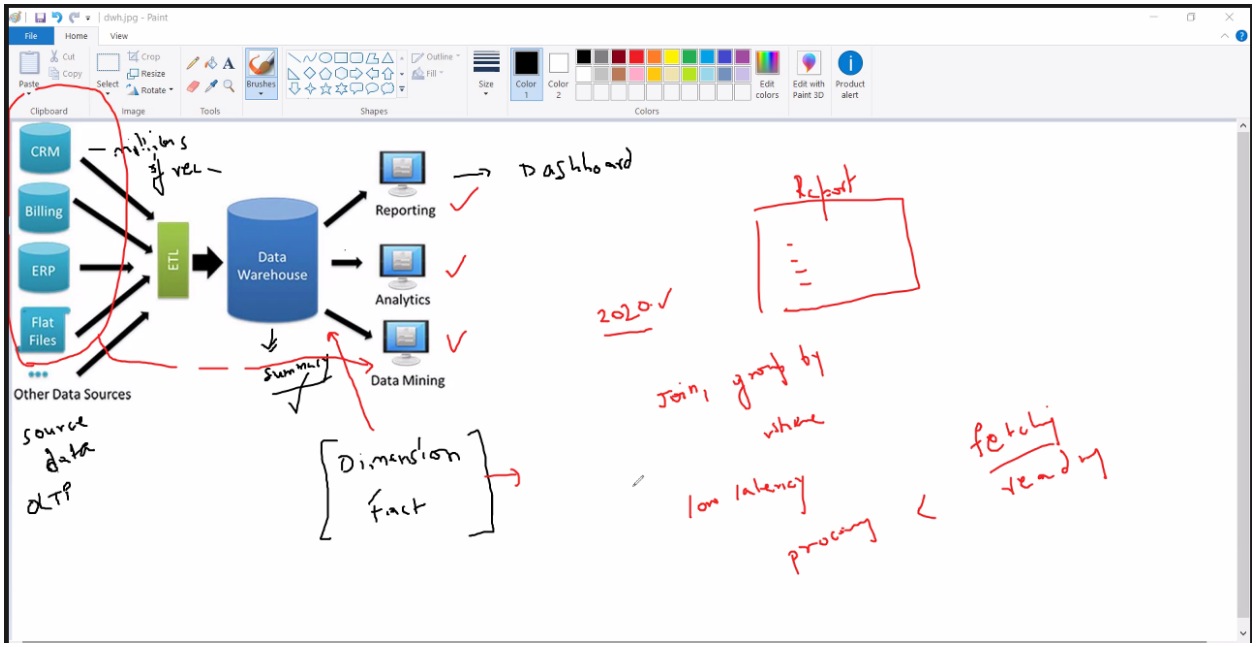


→ OLTP vs. OLAP

a. OLTP :
    i. OnLine Transaction Processing
    ii. Manual punching/Entry of source data into database
    iii. e.g. - depositing money into bank account, so cashier makes a manual entry
b. OLAP :
    i. OnLine Analytical Processing
    ii. Automated entry into database using OLAP services to update fact table and showing data at dashboard

→ Data Warehousing





Data warehouse has two set of tables:

a. Dimension tables
b. Fact tables

a. Data warehouse design is as per report/dashboard
b. Stores structured data in summary
c. Idea is to read data, not processing
d. ETL is done by hadoop dev to get data for analytical tools like informatica, etc.

→ Star design of data warehouse

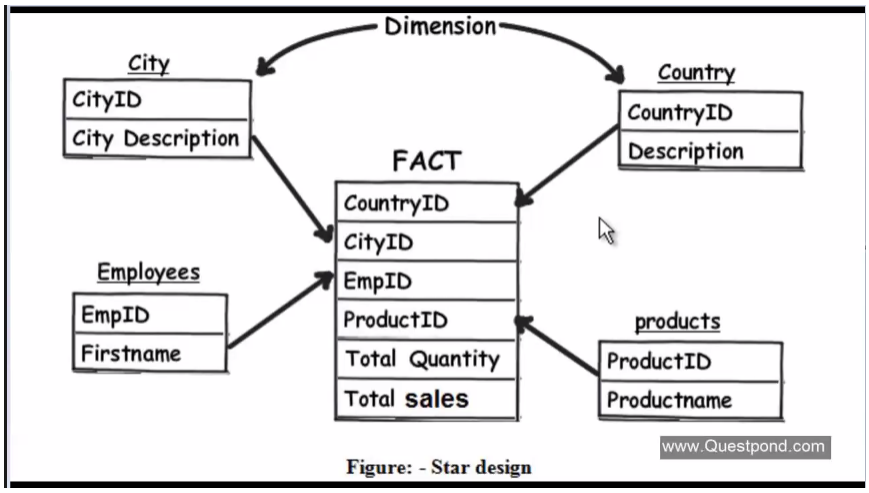
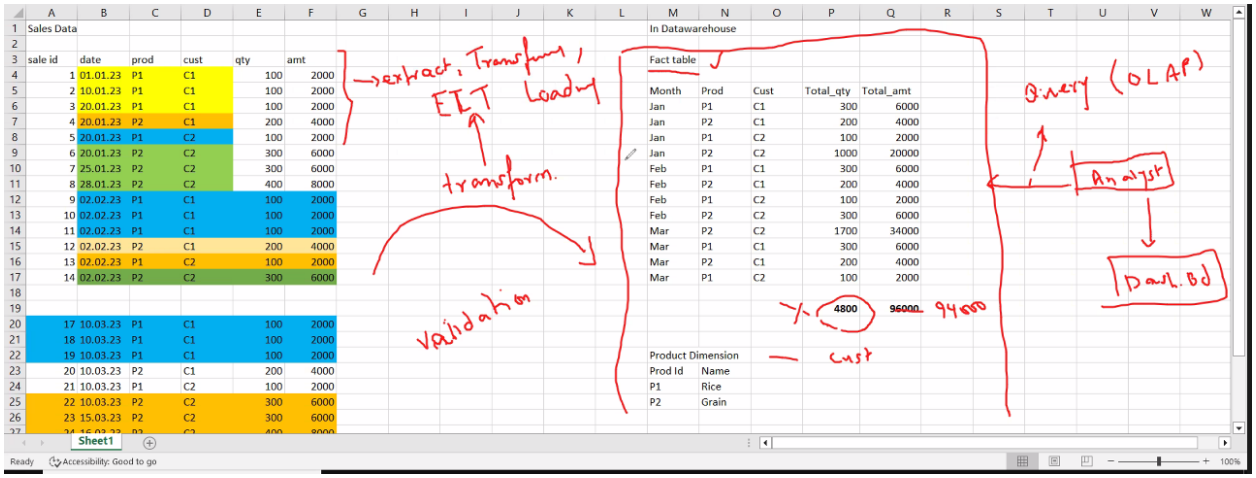
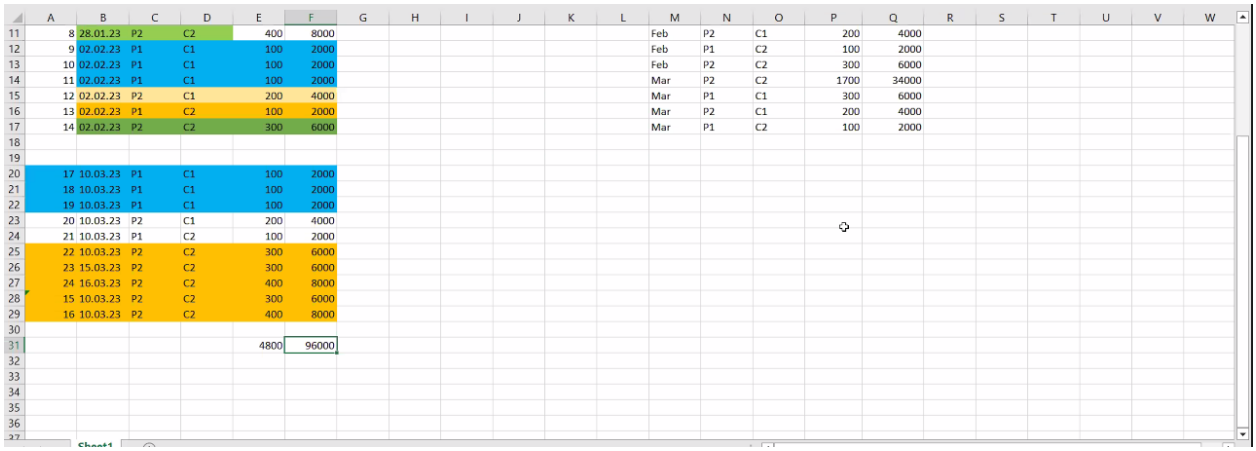


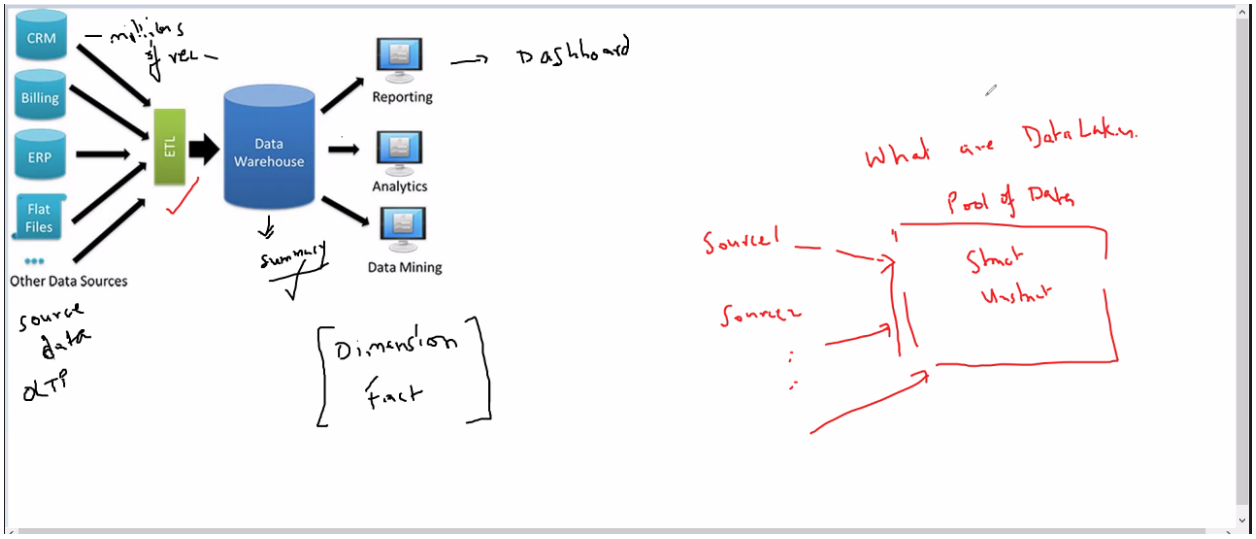
Figure: - Star design





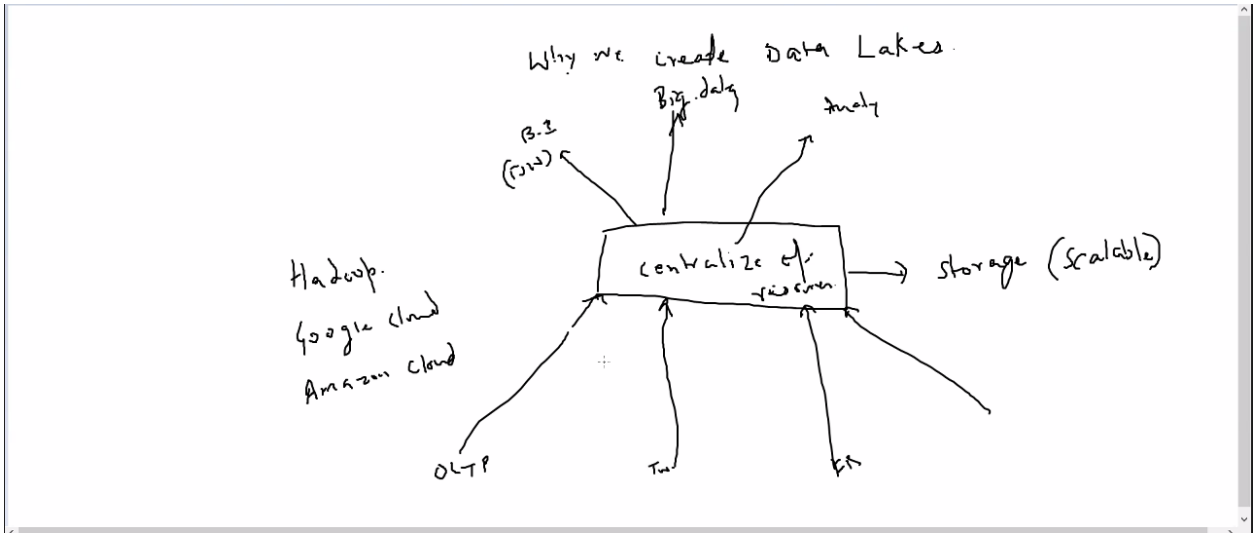
→ Data Lakes

a. Centralized repository/Collection of data from different sources
b. No data size limits
c. Can store any type of data
d. Hadoop may be called as data lake

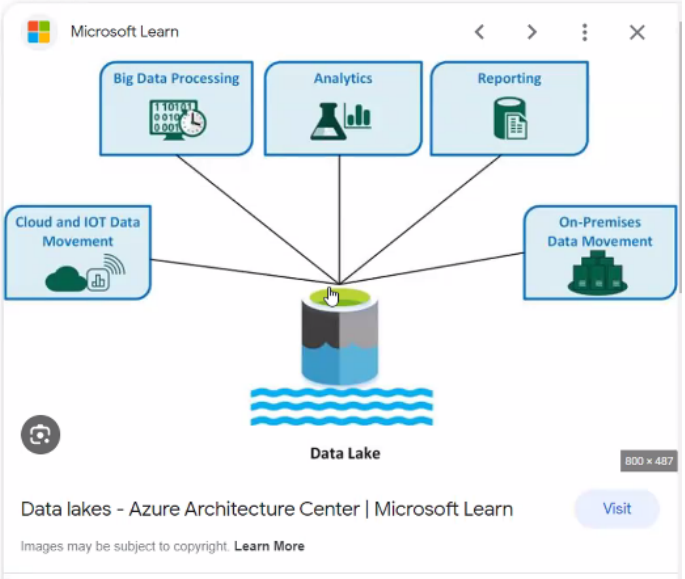


→ need of data lakes

a. Centralized location
b. Scalable storage to store any amount of data
c. E.g. hadoop is good example of data lake

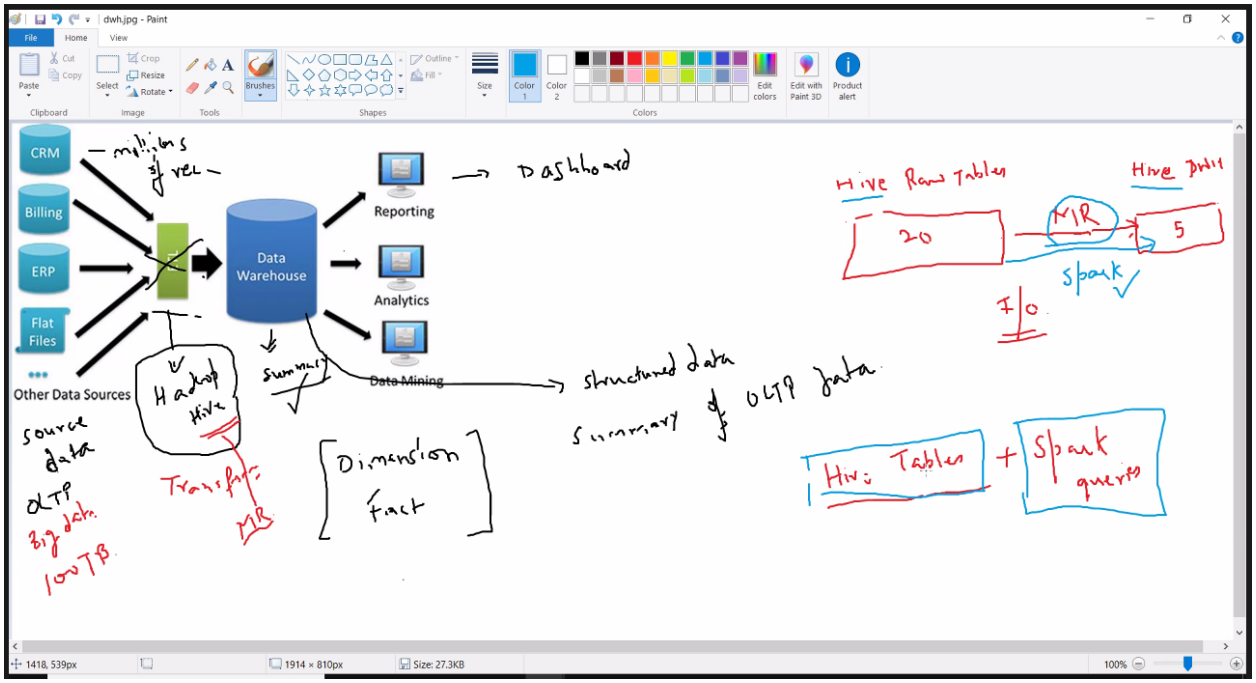


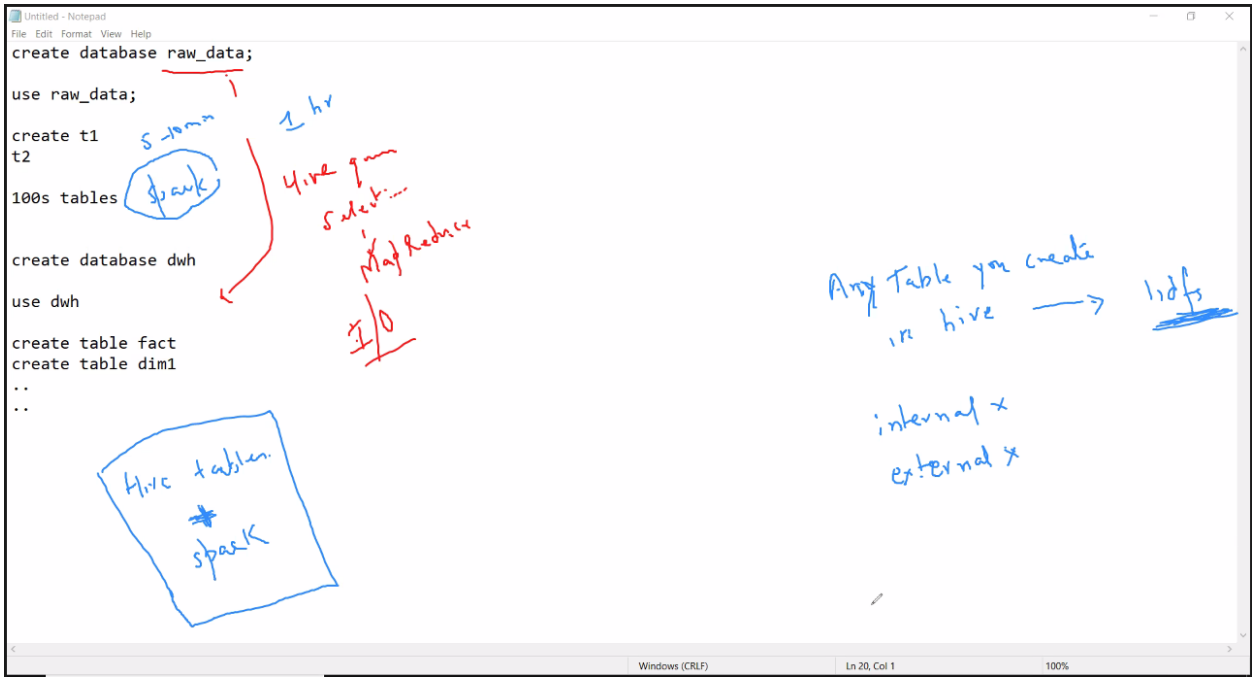
→ Dashboard



a. Always create data warehouse on top of data lake while creating dashboard
b. Not always realtime
c. Usually updated daily or weekly based on frequency of dashboard being checked
d.

→ MapReduce vs. Spark





a. MR mapReduce is very slow while processing
b. Spark is very fast while processing data, so industries are using hive tables to store data, but instead of using MapReduce queries to process data, they're using spark queries to process data