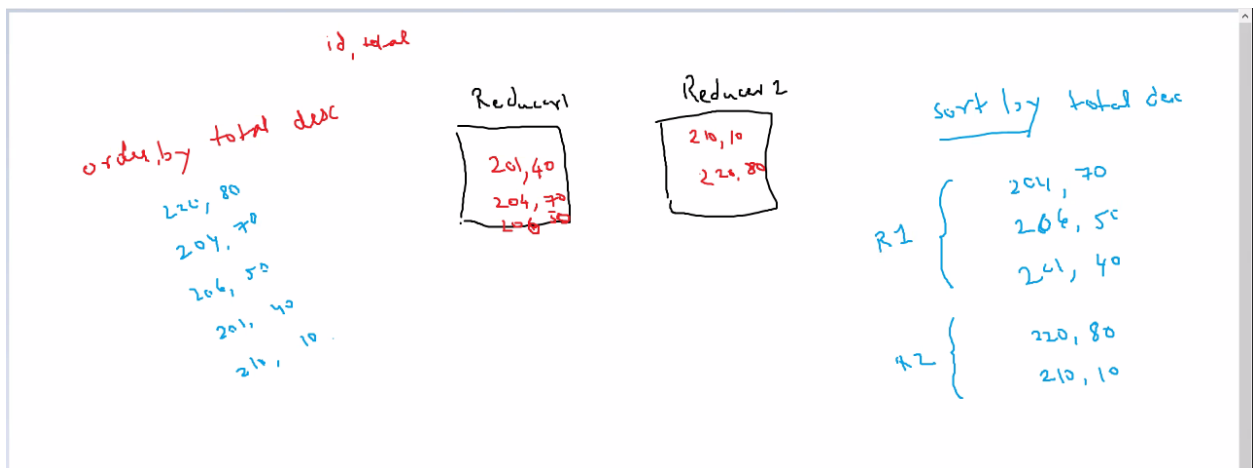


→ ORDER BY vs. SORT BY

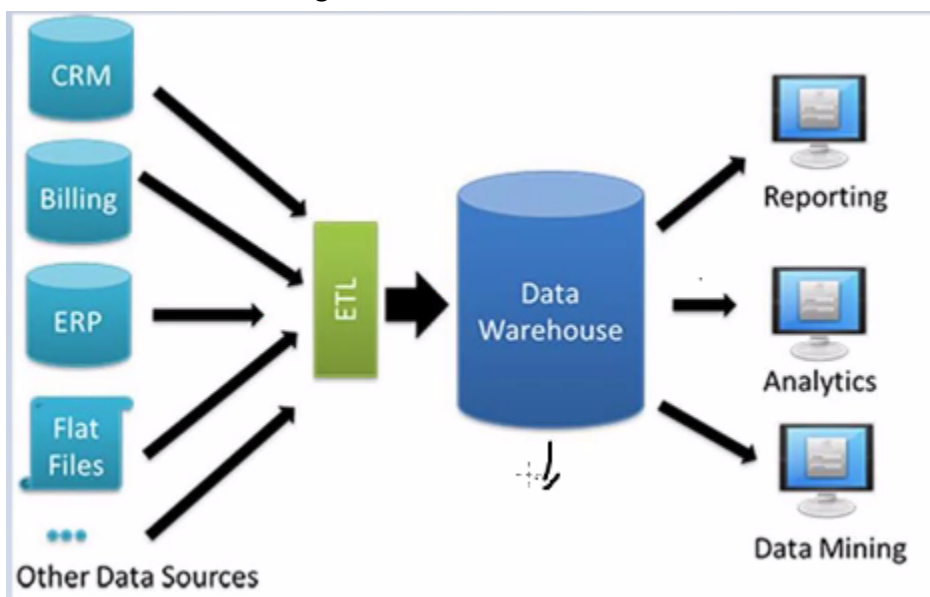
- ORDER BY : sorts all of data on all reducers
- SORT BY : sorts only the data within one reducer

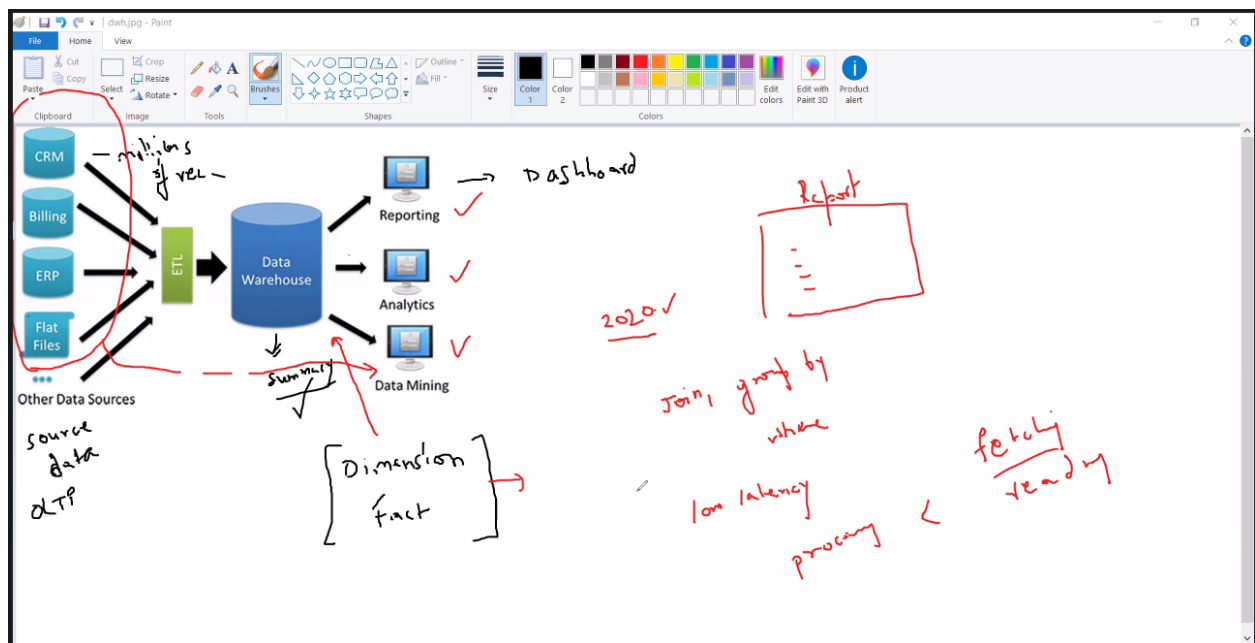


→ OLTP vs. OLAP

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

→ Data Warehousing



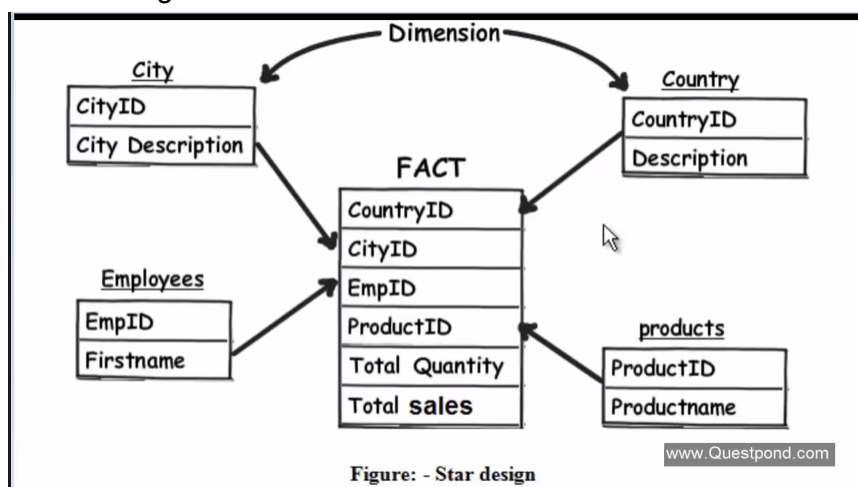


Data warehouse has two set of tables:

- 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.

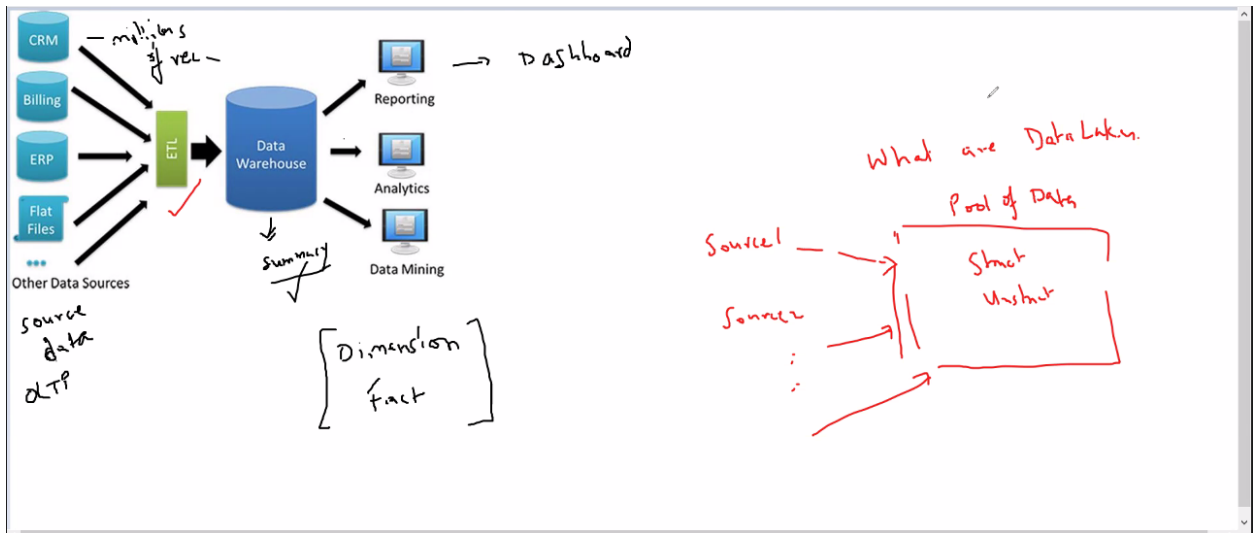
→ Star design of data warehouse



Sales Data						In Datawarehouse															
sale id	date	prod	cust	qty	amt	Fact table															
1	01.01.23	P1	C1	100	2000	Month	Prod	Cust	Total_qty	Total_amt	<div>Query (OLAP)</div> <div>Analyst</div> <div>Dashboard</div> <div>4800 96000 94000</div> <div>cust</div>										
2	10.01.23	P1	C1	100	2000	Jan	P1	C1	300	6000											
3	20.01.23	P1	C1	100	2000	Jan	P2	C1	200	4000											
4	20.01.23	P2	C1	200	4000	Jan	P1	C2	100	2000											
5	20.01.23	P1	C2	100	2000	Jan	P2	C2	1000	20000											
6	20.01.23	P2	C2	300	6000	Feb	P1	C1	300	6000											
7	25.01.23	P2	C2	300	6000	Feb	P2	C1	200	4000											
8	28.01.23	P2	C2	400	8000	Feb	P1	C2	100	2000											
9	02.02.23	P1	C1	100	2000	Feb	P2	C2	300	6000											
10	02.02.23	P1	C1	100	2000	Mar	P2	C2	1700	34000											
11	02.02.23	P1	C1	100	2000	Mar	P1	C1	300	6000											
12	02.02.23	P2	C1	200	4000	Mar	P2	C1	200	4000											
13	02.02.23	P1	C2	100	2000	Mar	P2	C1	200	4000											
14	02.02.23	P2	C2	300	6000	Mar	P1	C2	100	2000											
17	10.03.23	P1	C1	100	2000	Product Dimension															
18	10.03.23	P1	C1	100	2000	Prod Id	Name														
19	10.03.23	P1	C1	100	2000	P1	Rice														
20	10.03.23	P2	C1	200	4000	P2	Grain														
21	10.03.23	P1	C2	100	2000																
22	10.03.23	P2	C2	300	6000																
23	15.03.23	P2	C2	300	6000																
24	16.03.23	P2	C2	400	8000																
25	15.03.23	P2	C2	300	6000																
26	15.03.23	P2	C2	300	6000																
27	16.03.23	P2	C2	400	8000																
28	15.03.23	P2	C2	300	6000																
29	16.03.23	P2	C2	400	8000																
30																					
31																					
32																					
33																					
34																					
35																					
36																					
37																					

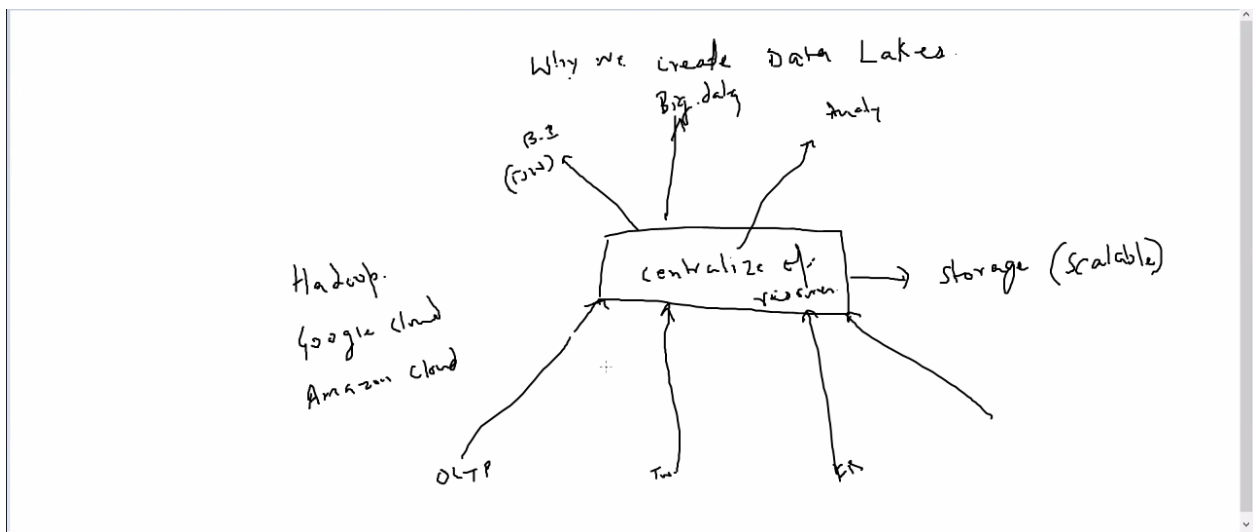
→ Data Lakes

- Centralized repository/Collection of data from different sources
- No data size limits
- Can store any type of data
- Hadoop may be called as data lake

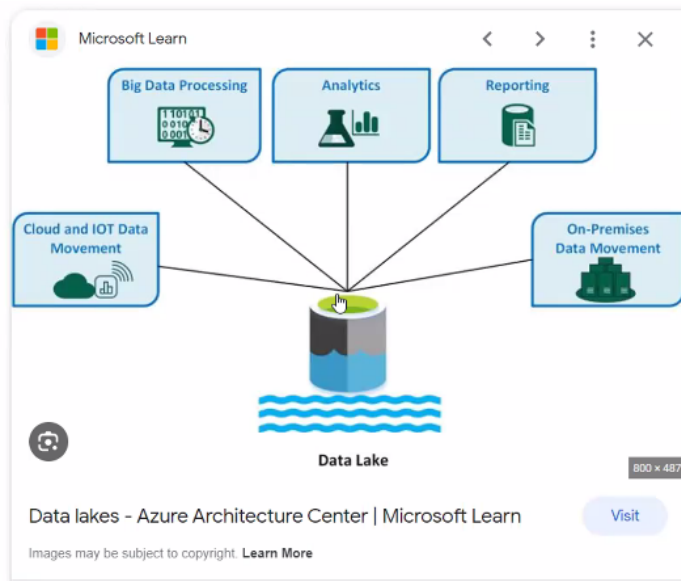


→ need of data lakes

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

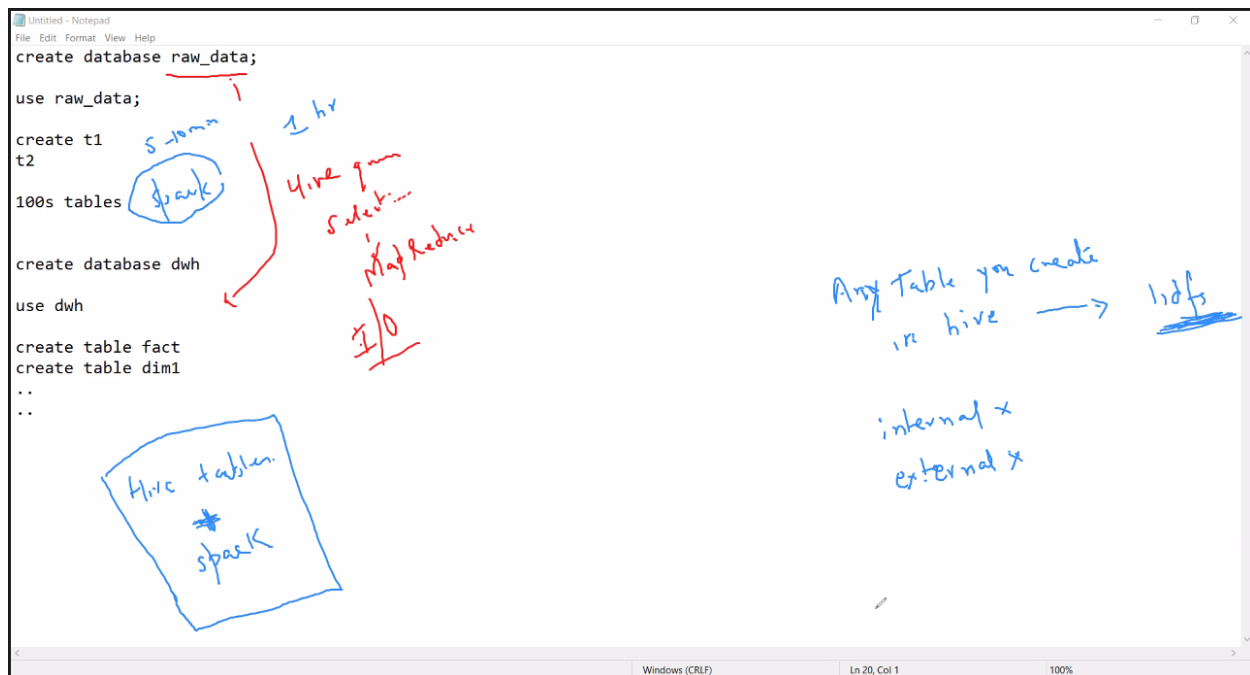
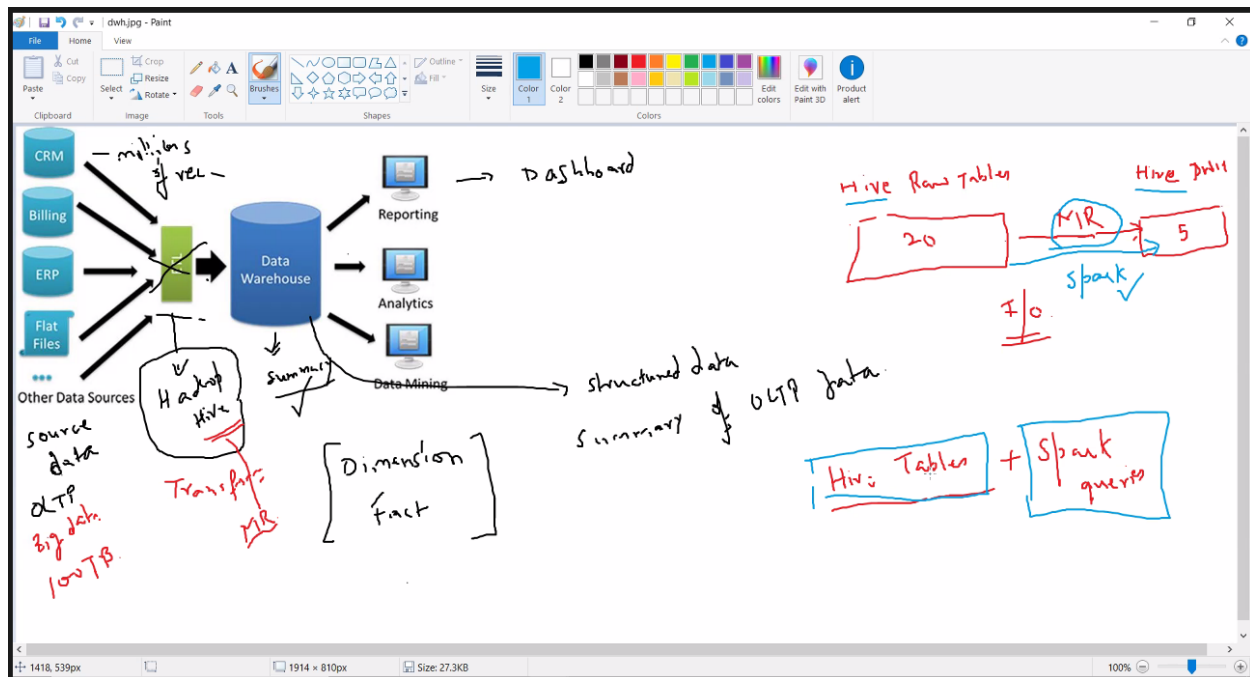


→ Dashboard



- Always create data warehouse on top of data lake while creating dashboard
- Not always realtime
- Usually updated daily or weekly based on frequency of dashboard being checked
-

→ MapReduce vs. Spark



- MR mapReduce is very slow while processing
- 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