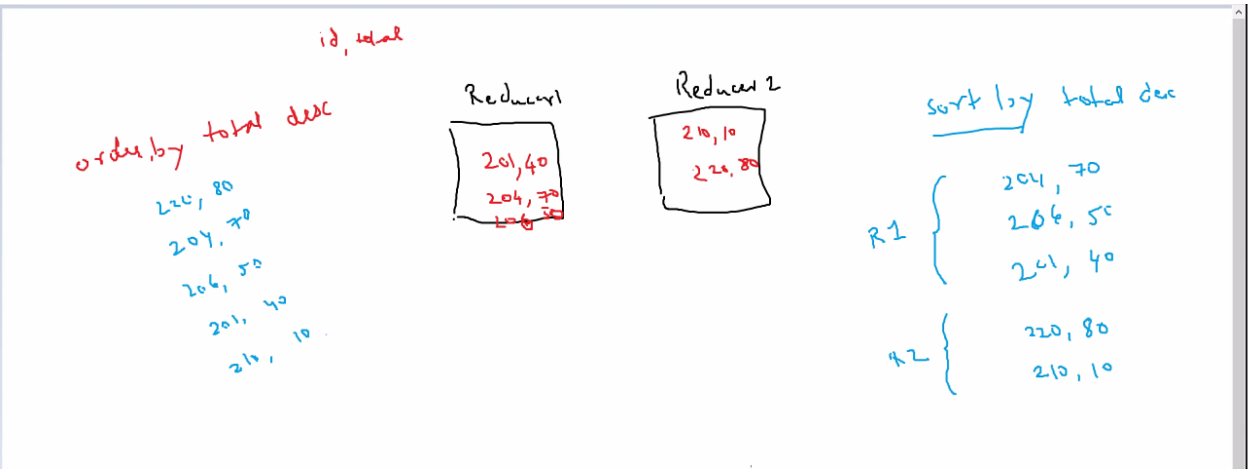
→ ORDER BY vs. SORT BY

a. ORDER BY
   i. sorts all of data on all reducers
   ii. Is slow as it takes up all the reducers

b. SORT BY :
   i. sorts only the data within one reducer
   ii. Is faster as it works with only one reducer at the end

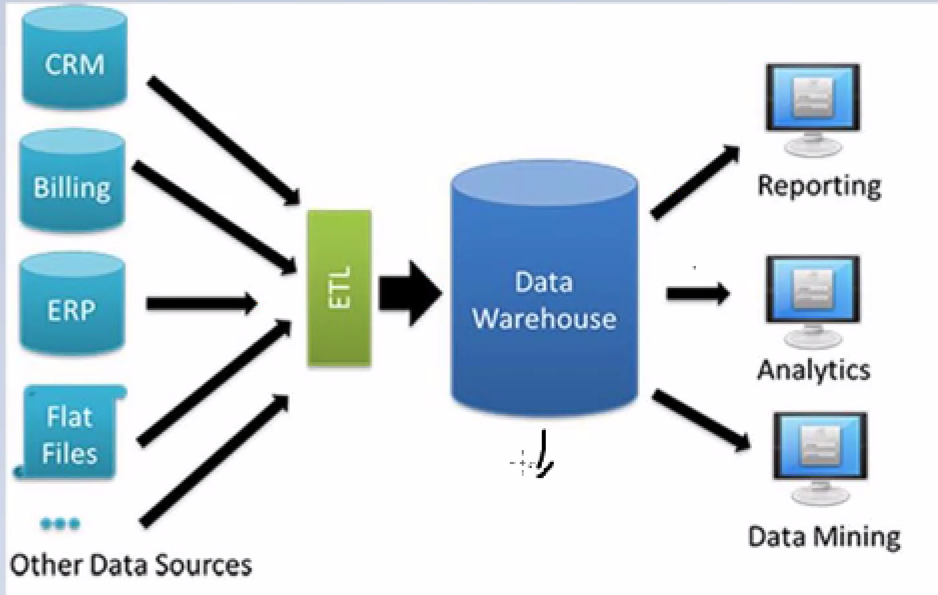


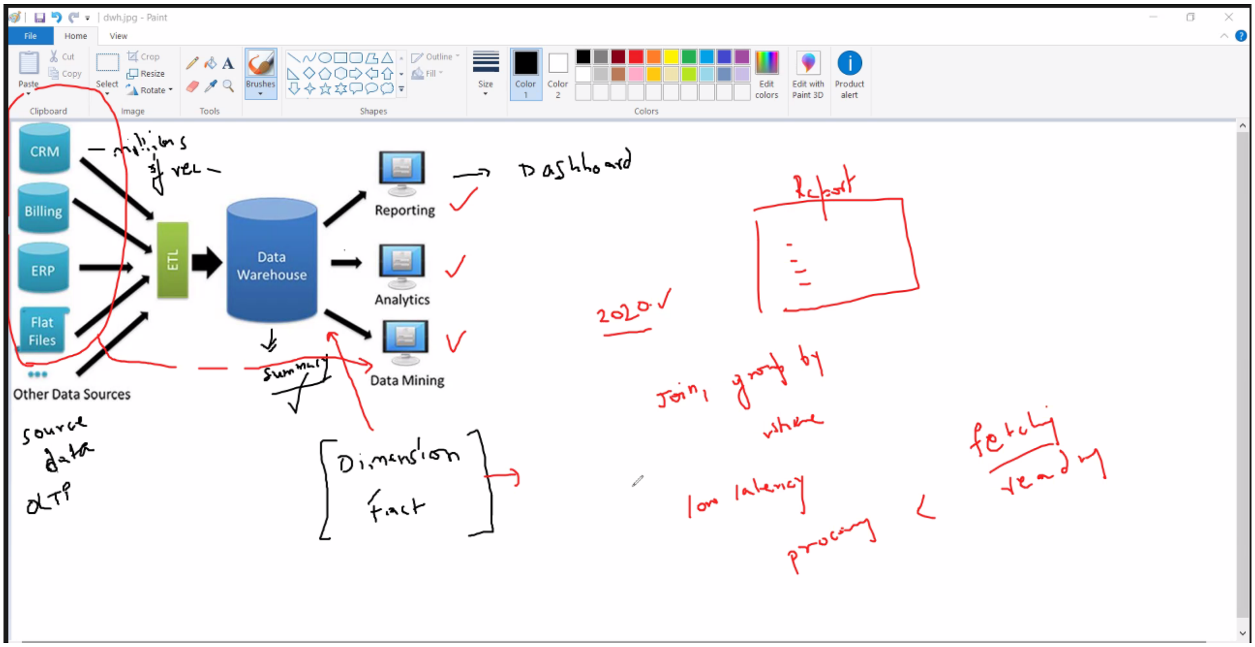
→ OLTP vs. OLAP

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

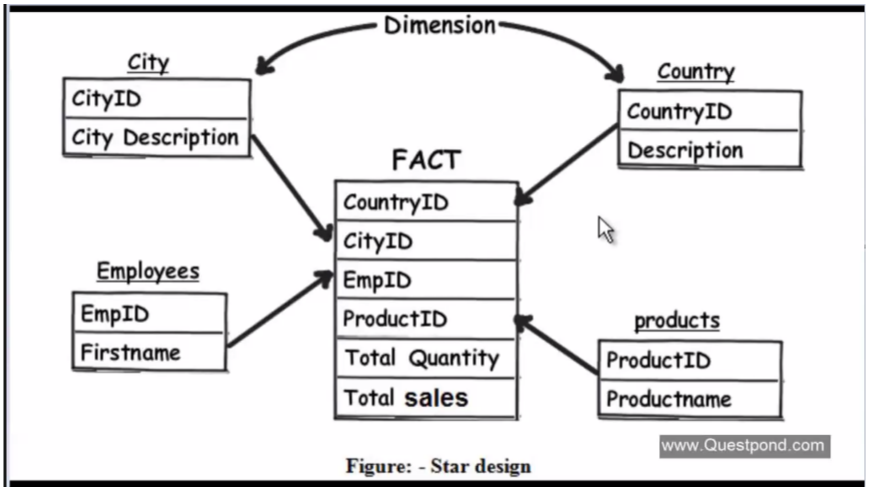
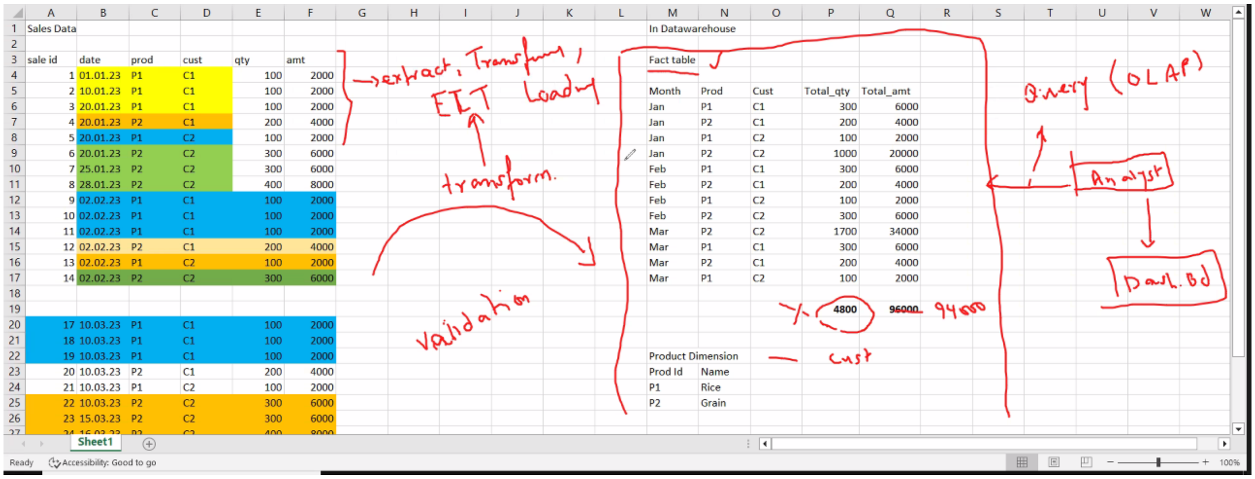


Figure: - Star design



Sales Data

sale id	date	prod	cust	qty	amt
1	01.01.23	P1	C1	100	2000
2	10.01.23	P1	C1	100	2000
3	20.01.23	P1	C1	100	2000
4	20.01.23	P2	C1	200	4000
5	20.01.23	P1	C2	100	2000
6	20.01.23	P2	C2	300	6000
7	25.01.23	P2	C2	300	6000
8	28.01.23	P2	C2	400	8000
9	02.02.23	P1	C1	100	2000
10	02.02.23	P1	C1	100	2000
11	02.02.23	P1	C1	100	2000
12	02.02.23	P2	C1	200	4000
13	02.02.23	P1	C2	100	2000
14	02.02.23	P2	C2	300	6000
17	10.03.23	P1	C1	100	2000
18	10.03.23	P1	C1	100	2000
19	10.03.23	P1	C1	100	2000
20	10.03.23	P2	C1	200	4000
21	10.03.23	P1	C2	100	2000
22	10.03.23	P2	C2	300	6000
23	15.03.23	P2	C2	300	6000

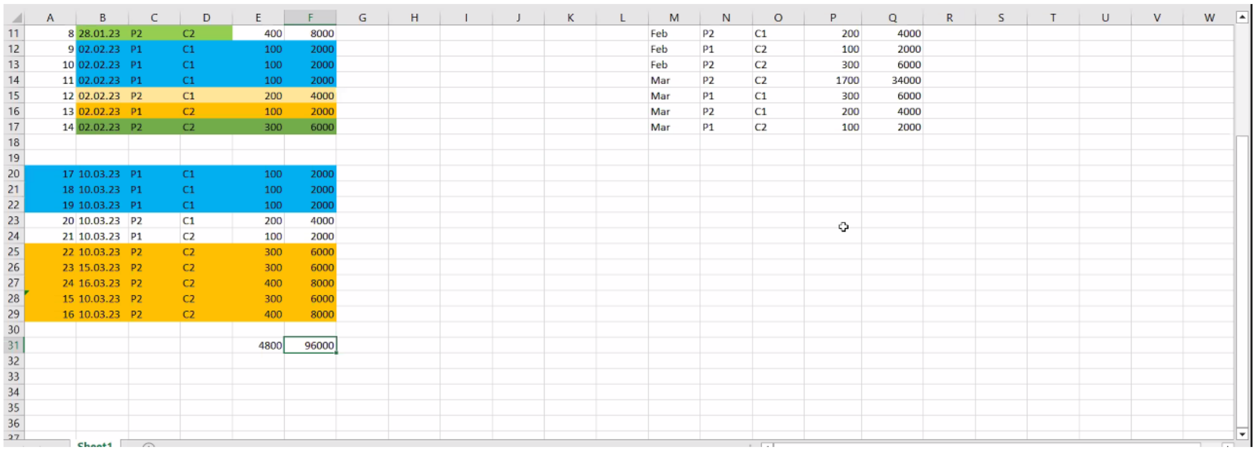
In Datawarehouse

Fact table ✓

Month	Prod	Cust	Total_qty	Total_amt
Jan	P1	C1	300	6000
Jan	P2	C1	200	4000
Jan	P1	C2	100	2000
Jan	P2	C2	1000	20000
Feb	P1	C1	300	6000
Feb	P2	C1	200	4000
Feb	P1	C2	100	2000
Feb	P2	C2	300	6000
Mar	P2	C2	1700	34000
Mar	P1	C1	300	6000
Mar	P2	C1	200	4000
Mar	P1	C2	100	2000
			4800	~~96000~~

Product Dimension

Prod Id	Name
P1	Rice
P2	Grain

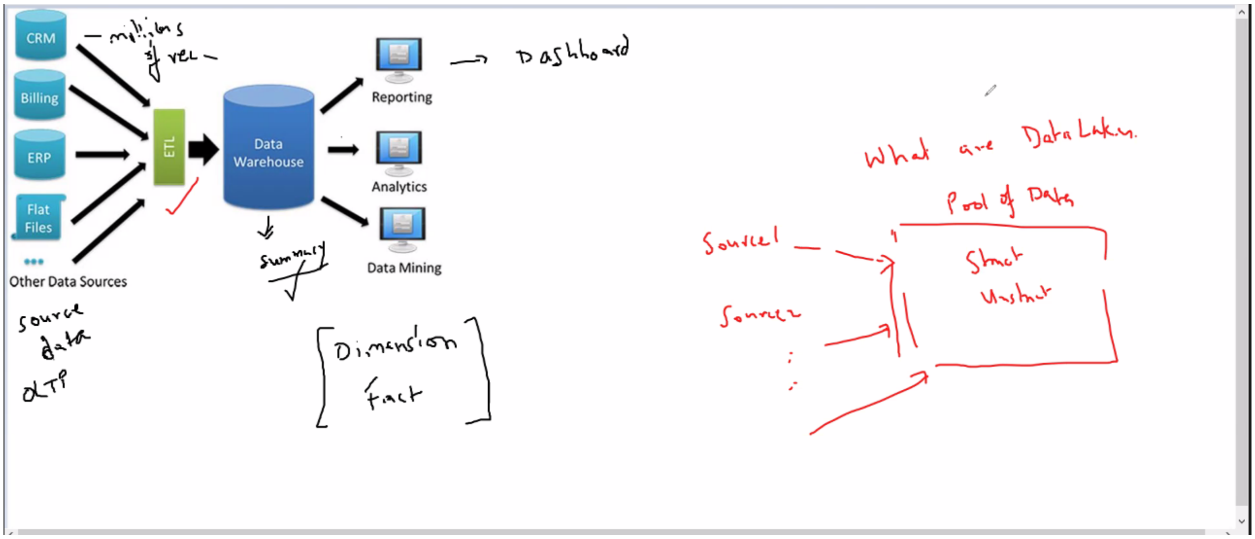


8	28.01.23	P2	C2	400	8000	Feb	P2	C1	200	4000
9	02.02.23	P1	C1	100	2000	Feb	P1	C2	100	2000
10	02.02.23	P1	C1	100	2000	Feb	P2	C2	300	6000
11	02.02.23	P1	C1	100	2000	Mar	P2	C2	1700	34000
12	02.02.23	P2	C1	200	4000	Mar	P1	C1	300	6000
13	02.02.23	P1	C2	100	2000	Mar	P2	C1	200	4000
14	02.02.23	P2	C2	300	6000	Mar	P1	C2	100	2000
17	10.03.23	P1	C1	100	2000					
18	10.03.23	P1	C1	100	2000					
19	10.03.23	P1	C1	100	2000					
20	10.03.23	P2	C1	200	4000					
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					
15	10.03.23	P2	C2	300	6000					
16	10.03.23	P2	C2	400	8000					
				4800	96000					

→ Data Lakes

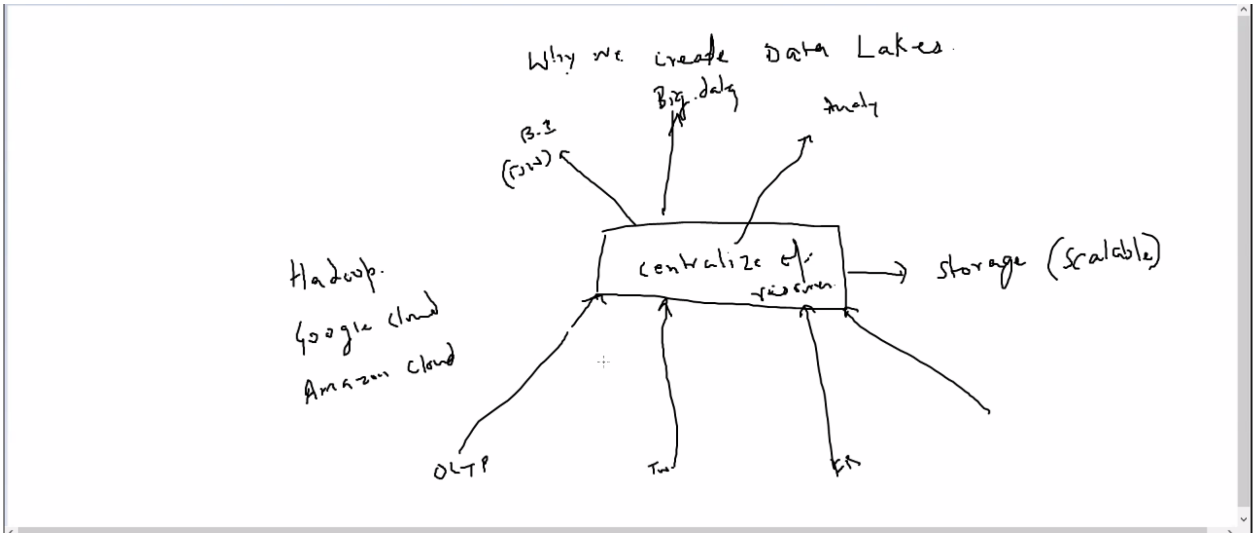
a. Centralized repository/Collection of data from different sources
b. No data size limits
c. Can store any type of data

d. Hadoop may be called as data lake

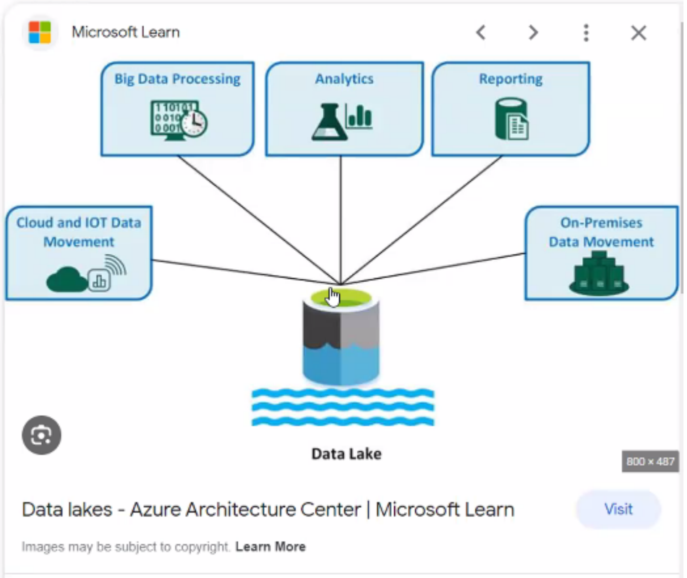


→ Need of data lakes

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



→ Dashboard



Data lakes - Azure Architecture Center | Microsoft Learn

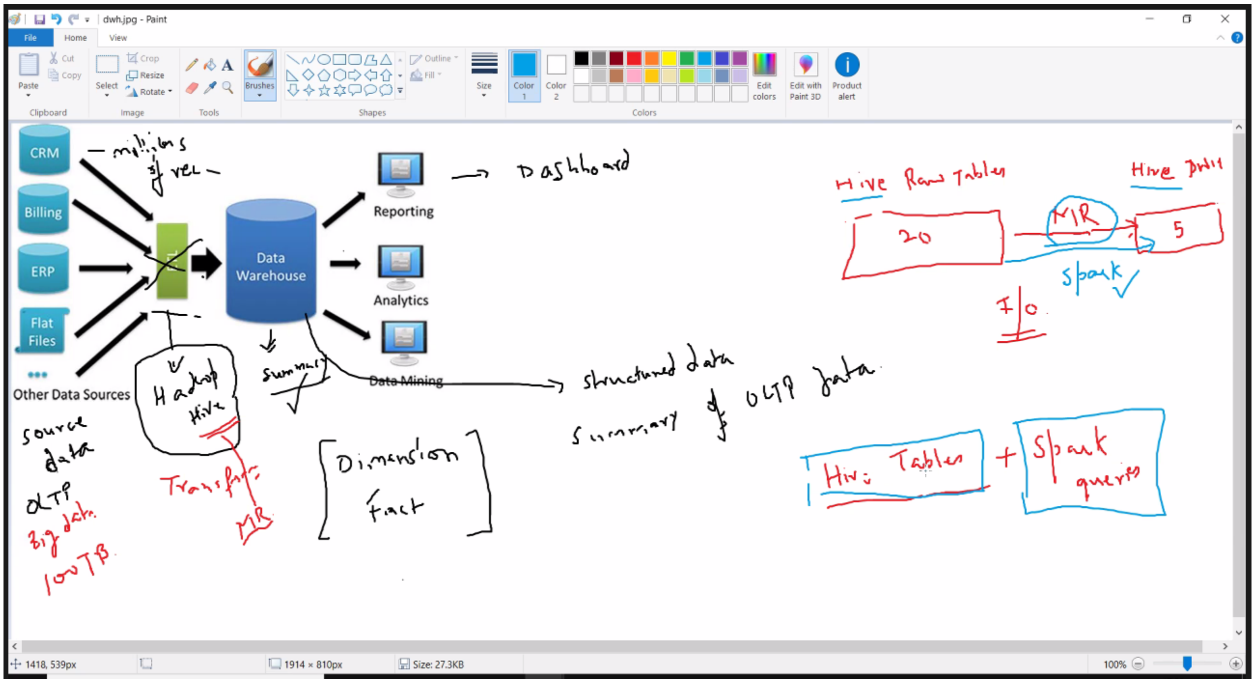
Images may be subject to copyright. **Learn More**

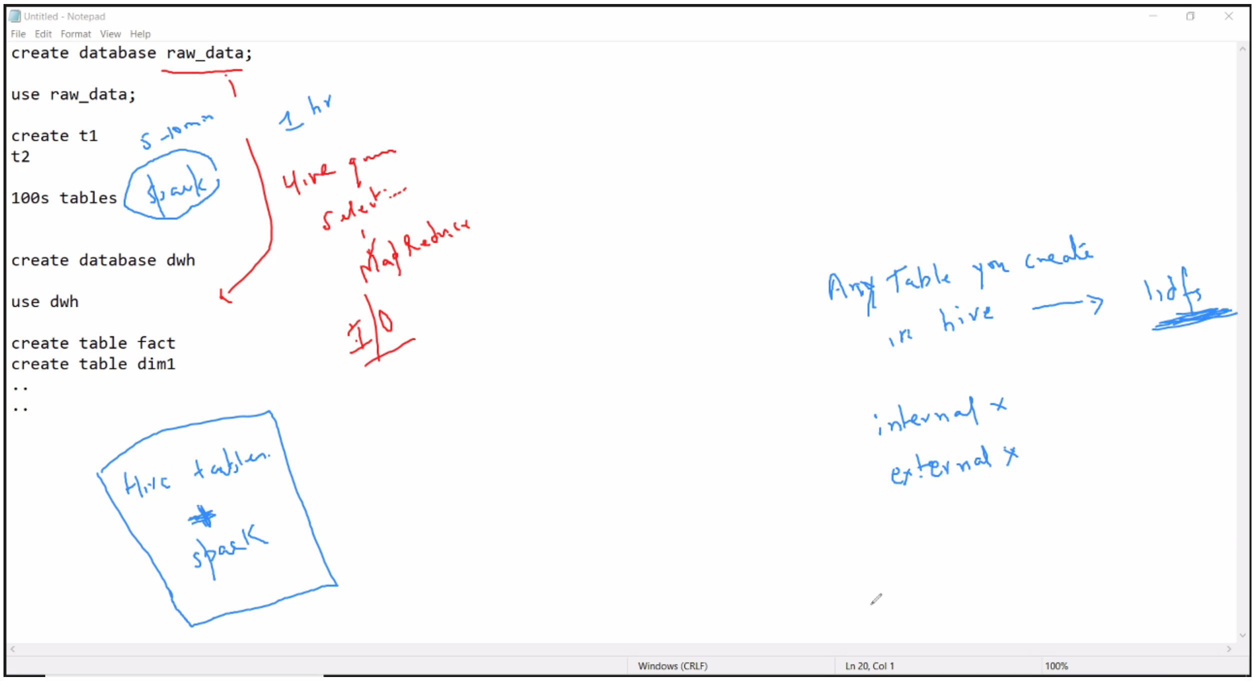
a. Always create data warehouse on top of data lake while creating dashboard
b. Not always realtime
c. Usually updated daily or weekly based on frequency of dashboard being checked

→ 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





```
create database raw_data;

use raw_data;

create t1
t2

100s tables


create database dwh

use dwh

create table fact
create table dim1
..
..
```