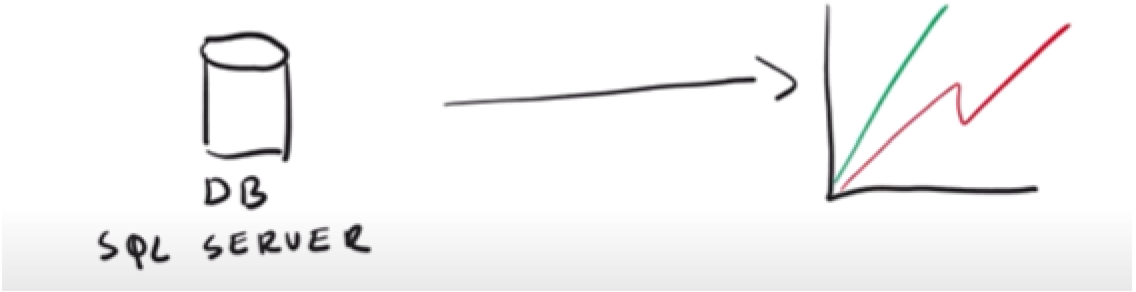
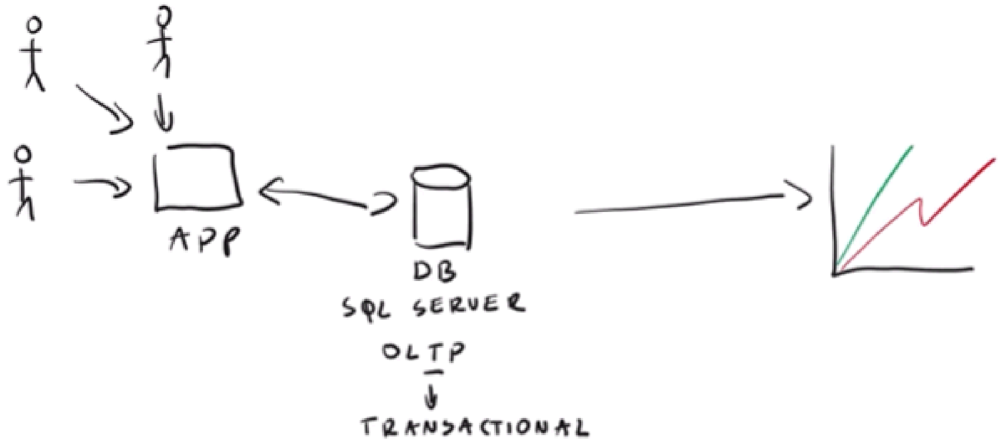
Scenario 1 : Direct Connection to Source DB

1. Assume our only data source is this relational DB(SQL server) and we dont have any other sources for data ..like API and all

2. And we want to create various reports from this data

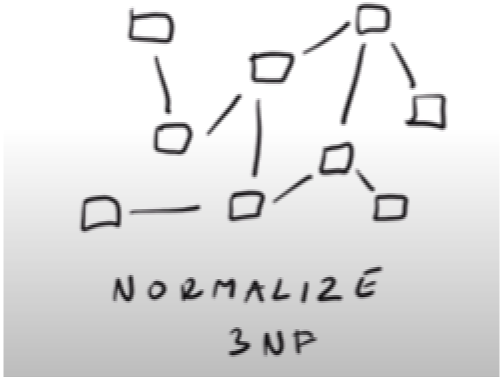


3. SO here why cant we directly connect reporting tool with the DB and make reports? Its a valid question
4. Here assume there is an app…which stores and read the data from the DB

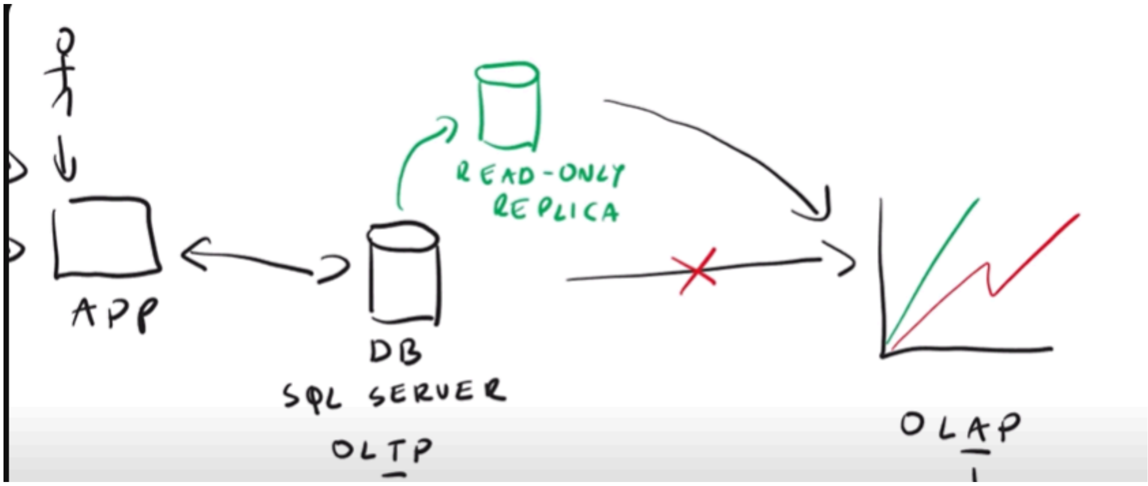


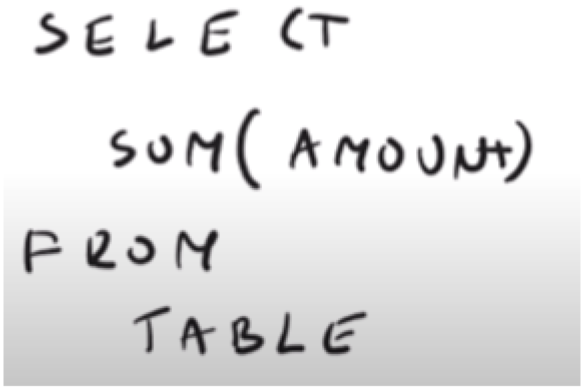
5. Also Assume there are many user’s who are placing order’s…and these order’s data must be stored in DB

6. And here we have tables of customer details spread across our database which follows 3NF normalization



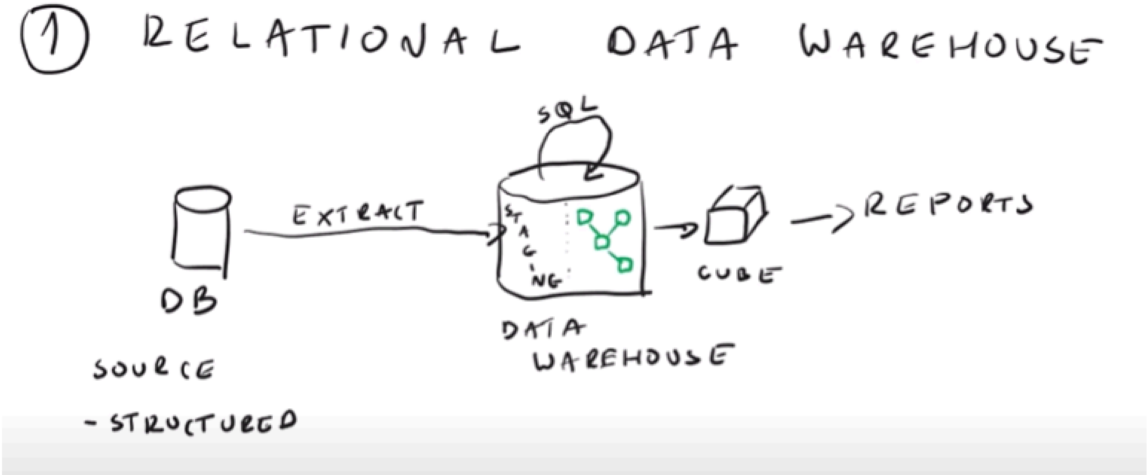
7. So here for the end users to directly connect and retrieve data from DB…they have to implement complex joins which will be hard to implement
8. And also there might be security issues as well
9. So one solution is to having a read-only replica of our DB



10. And also queries like  `SELECT SUM(AMOUNT) FROM TABLE` will read the entire rows and it takes more time to execute
11. So to avoid this we dont recommend directly connecting with the DB

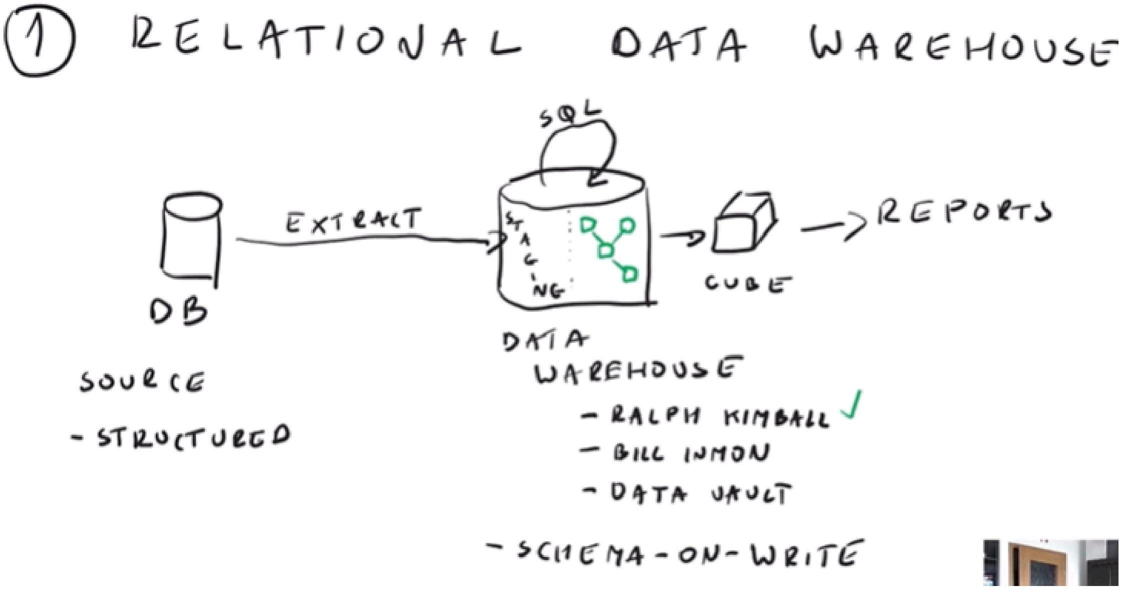
Relational Data Warehouse(Work Around)

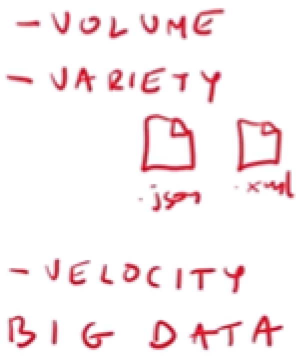
1. Here we'll be using Data warehouse
2. Here the data will be extracted from the DB into the staging layer of DW and from there using SQL we transform the data implement data modeling techniques and store it in the other half of DW



   The reports team can get the data from the DW faster…and to make it even faster..we can use CUBE

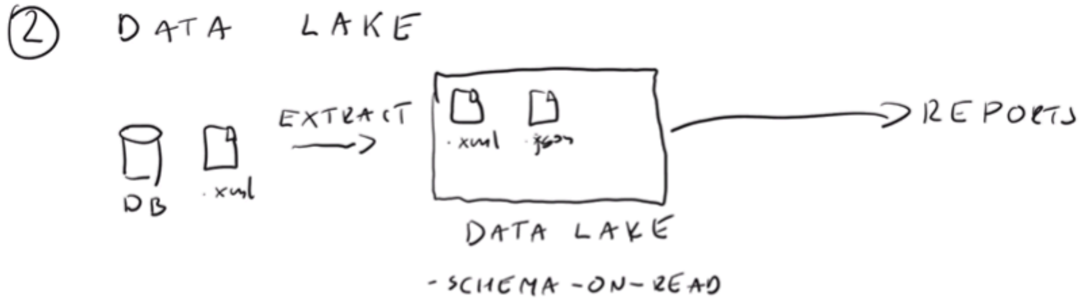
3. Here for modeling the data we'll use KIMBALL model
4. And in data warehouse…we have to implement schema before we load the data



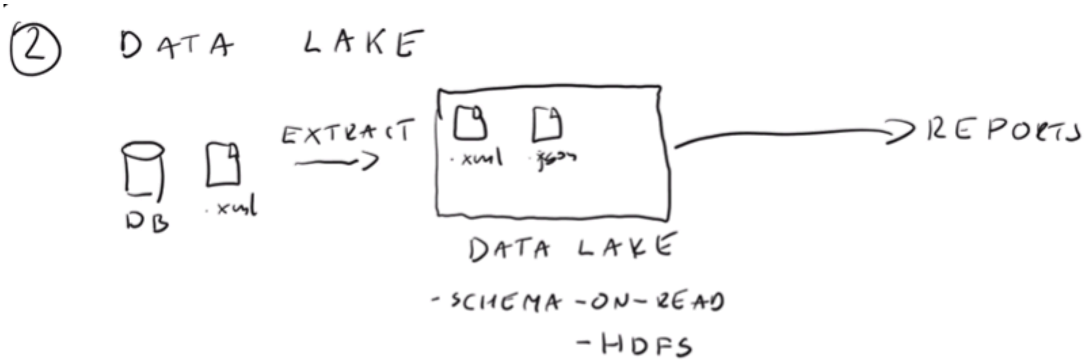
5. But to  the cost of data warehouse were increasing

Data Lake (Another Work Around)

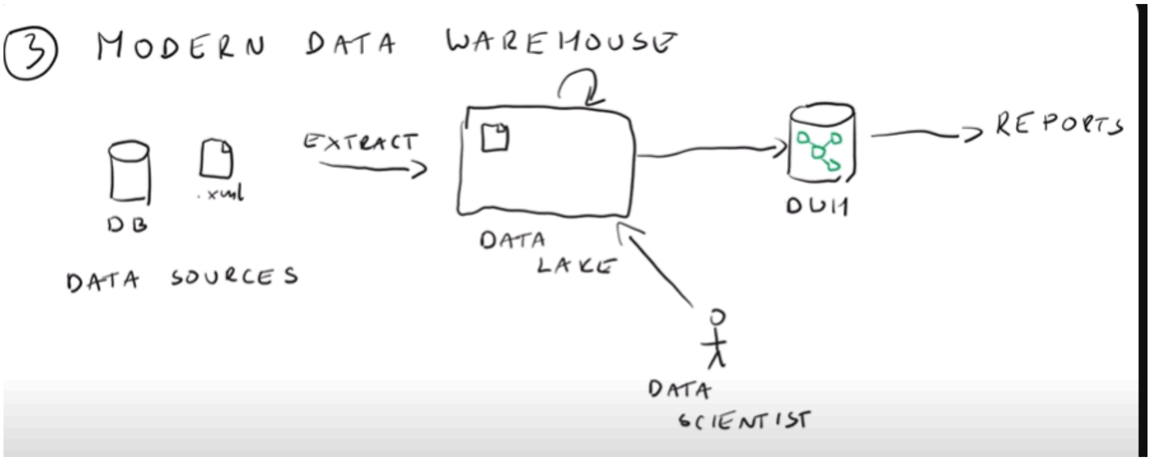
1. Data Lake can store data of both structured and non-structured data
2. We can think of it as Windows LocalDisk



3. To process the data in datalake ..hadoop was used initially

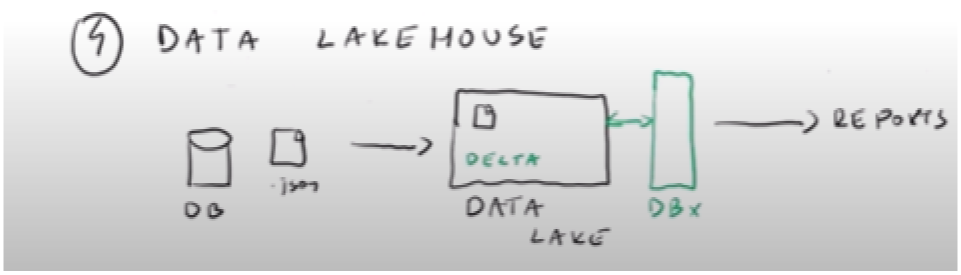


Modern Data Warehouse

1. 

   here data comes in multiple formats and from different sources into the datalake
2. And we'll also be having a Datawarehouse that models the data based on KIMBALL approach..and reporters can get data easily from there
3. This architecture is most common now a days
4. But the drawback is the data is getting duplicated in Datalake and Datawarehouse

DataLake House

1. 

DataMesh

1.