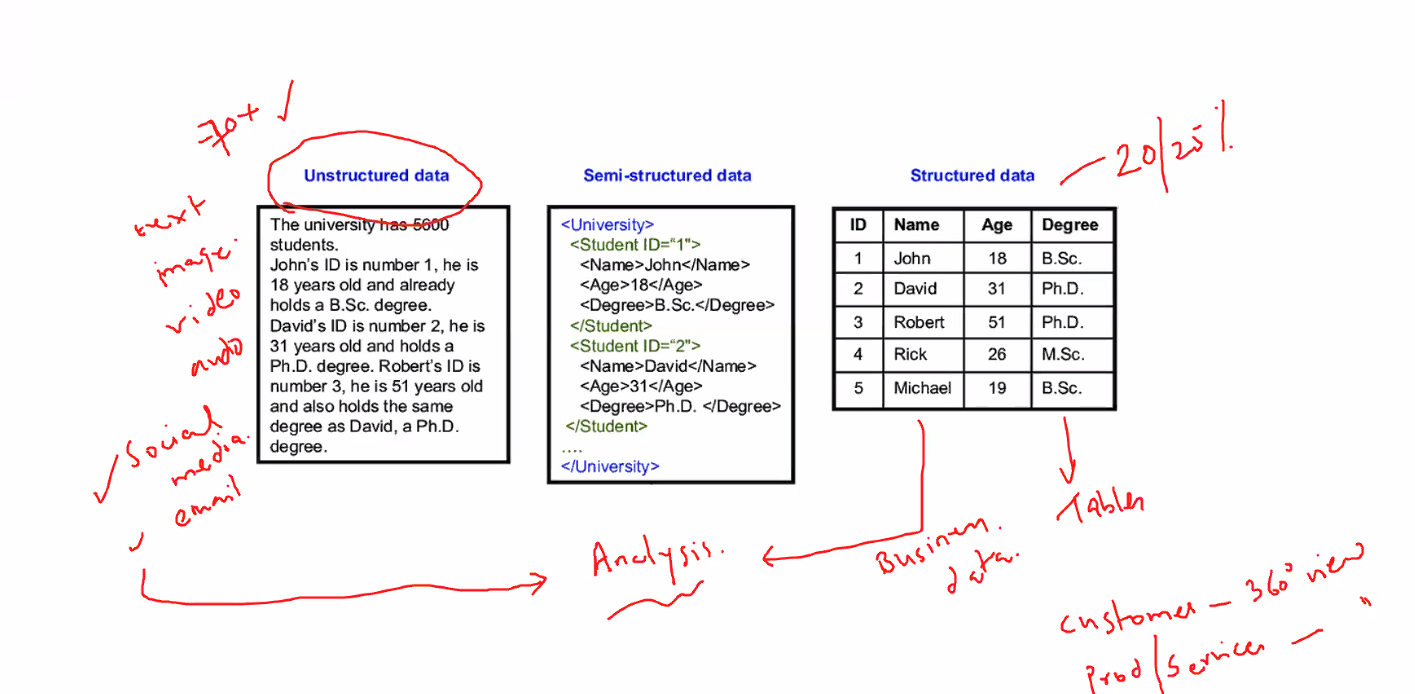
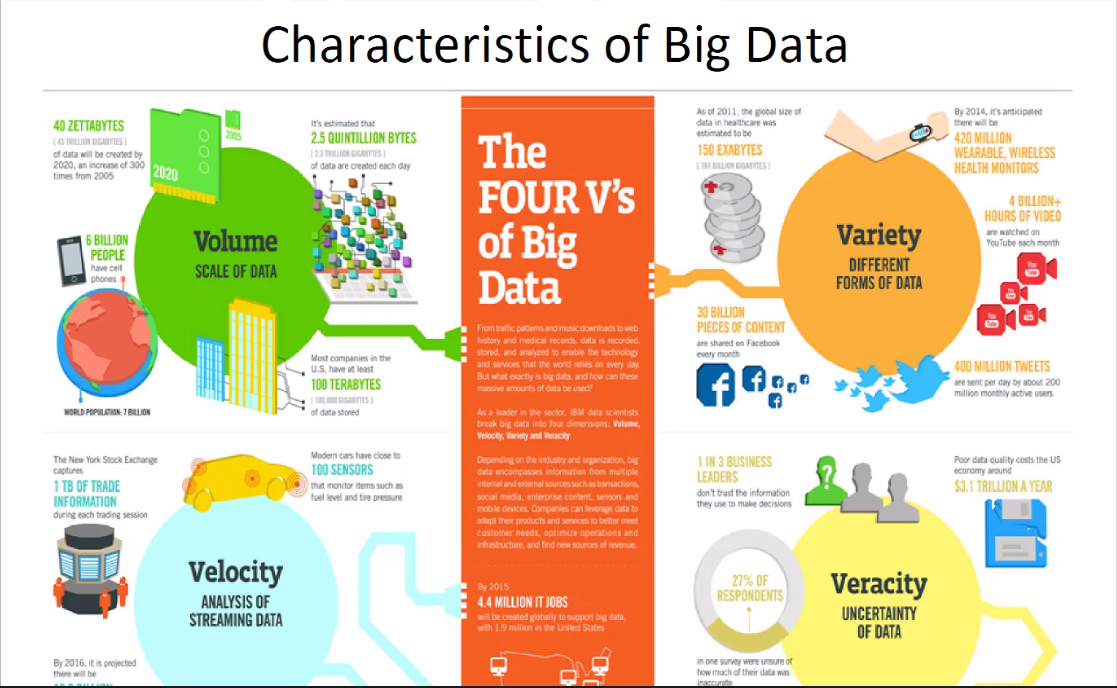
**Big Data**

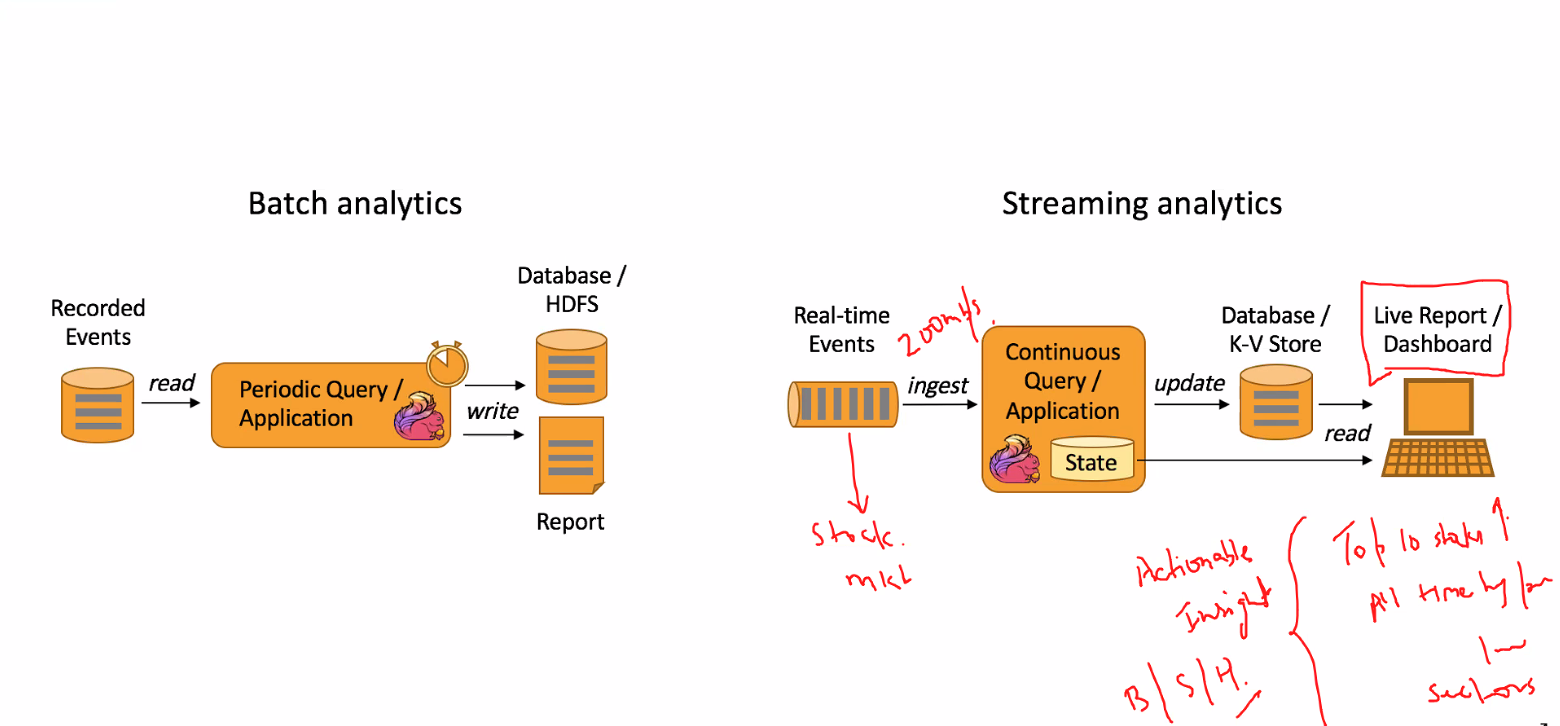
pwd

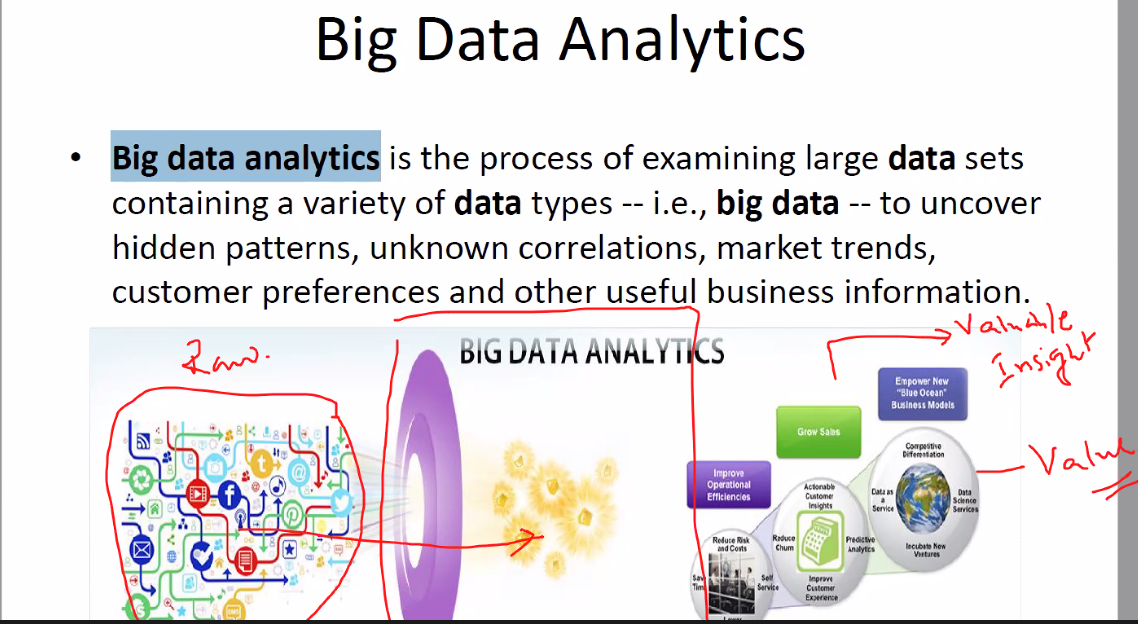
ls

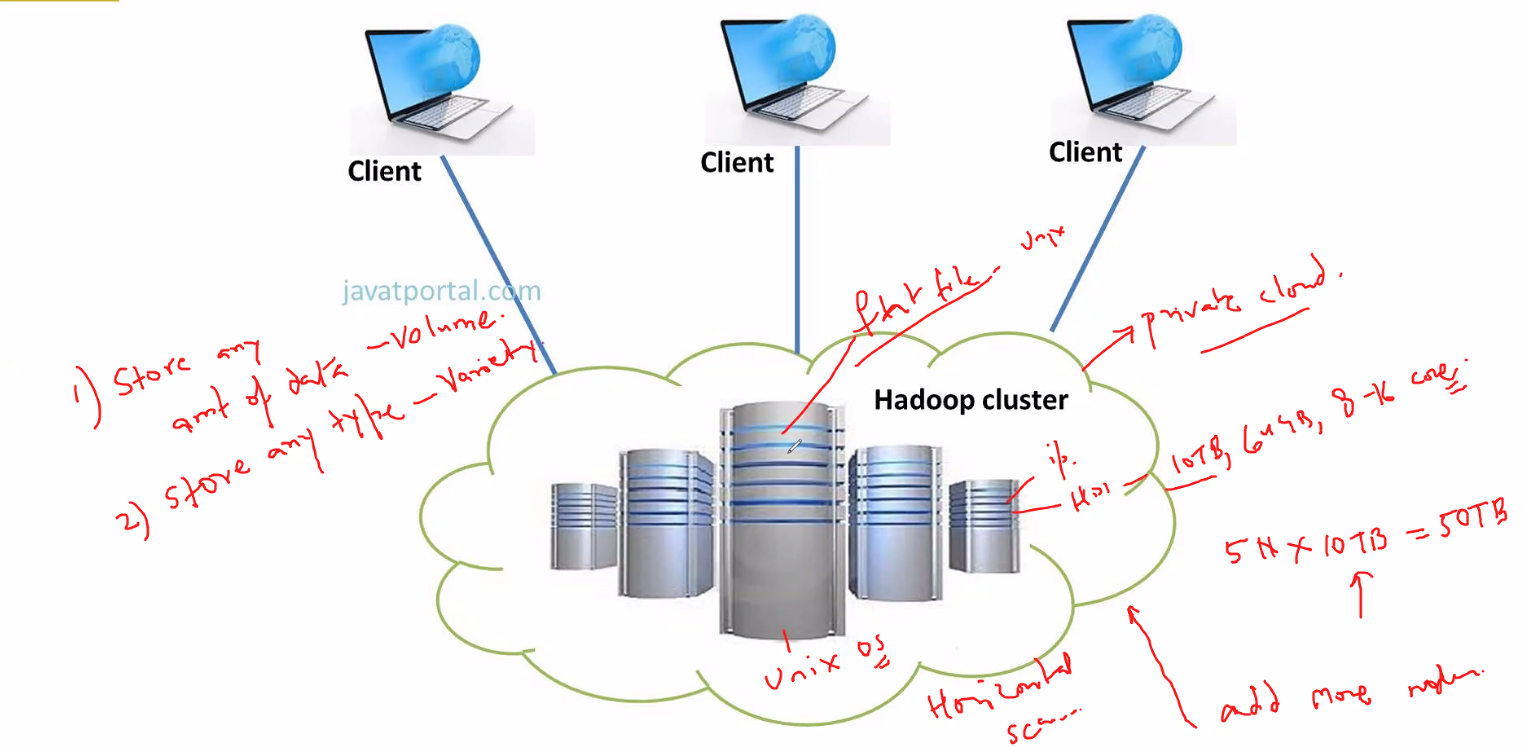




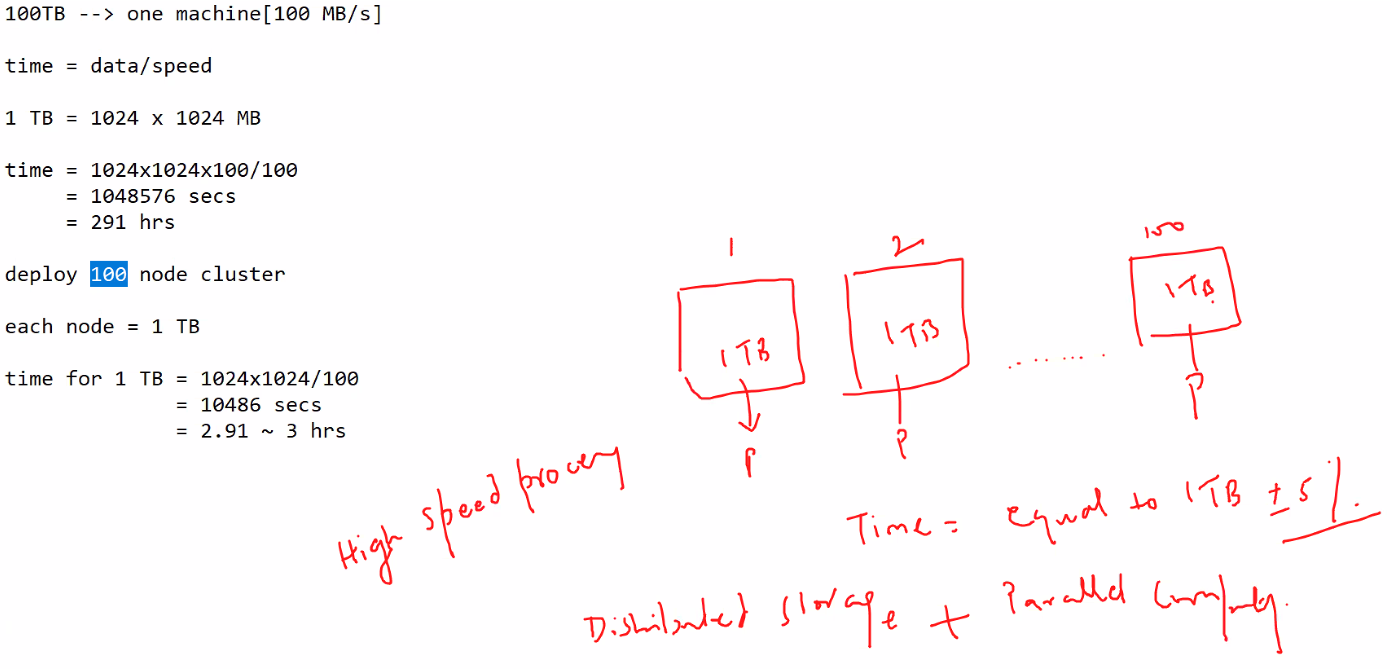
Veracity: Cleansing, preprocessing, validating

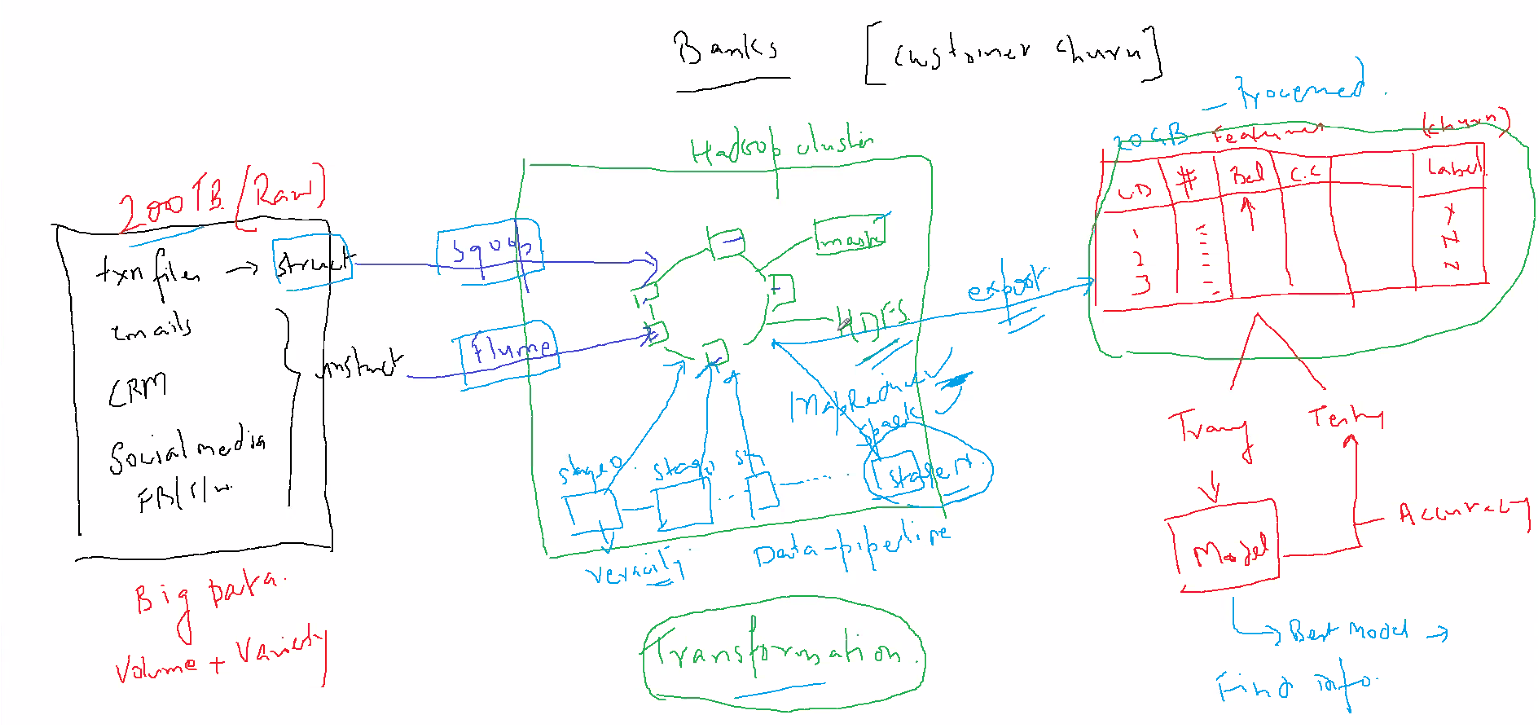






Hadoop -> Transform Unstructured data to structured processed data (processed in distributed storage and parallel computing is used)





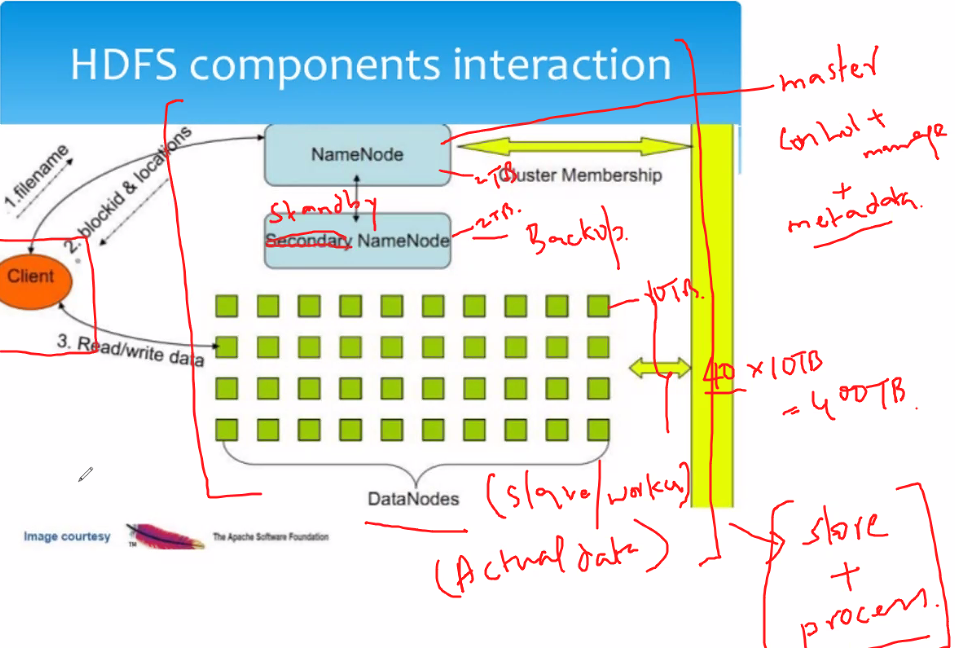
Sqoop => pull structured data faster

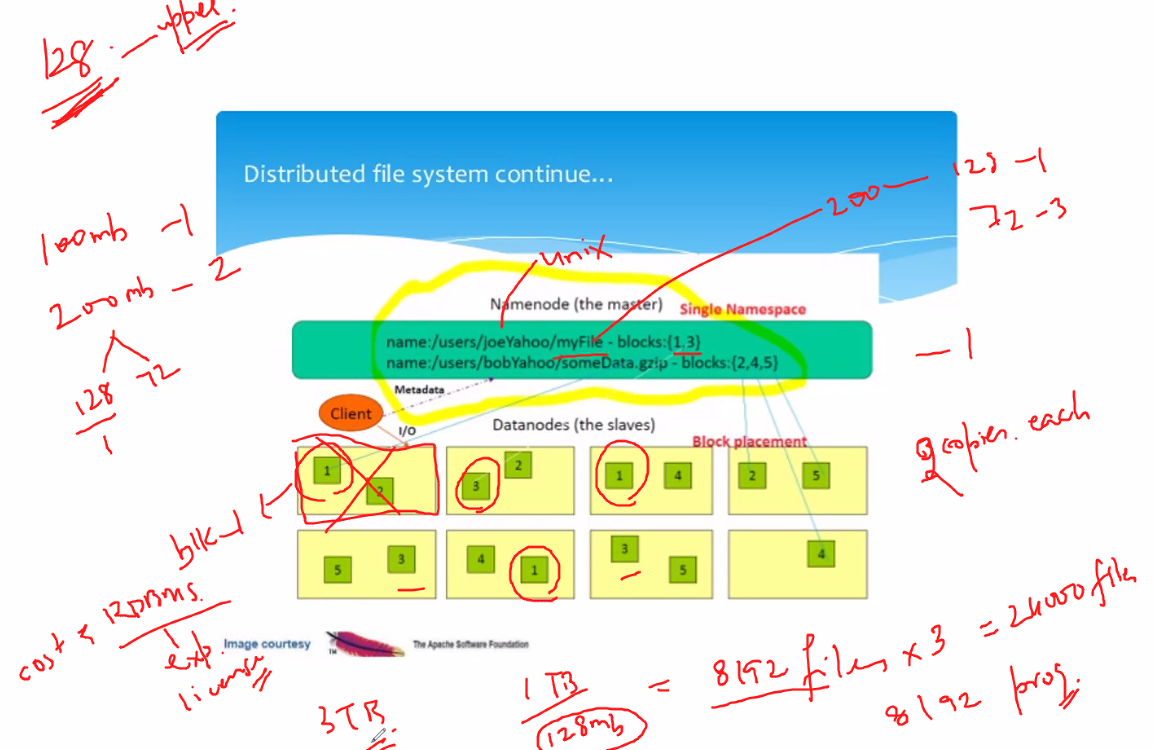
Flume => pull unstructured data faster

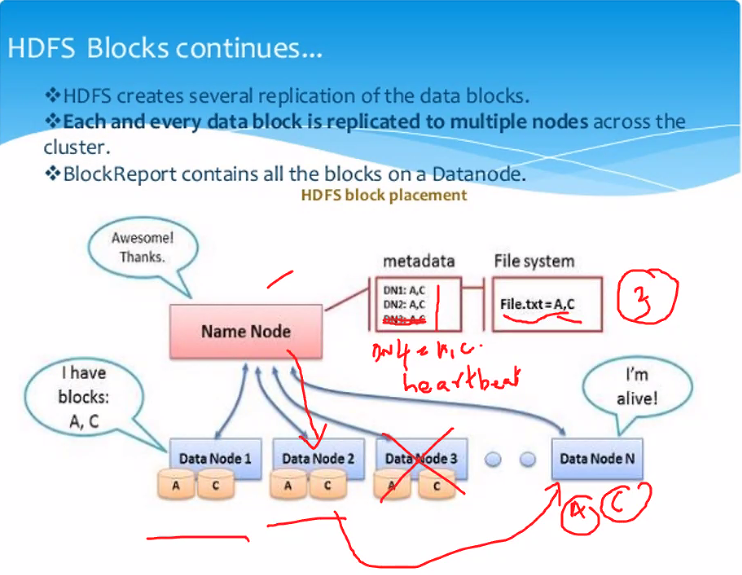
MapReduce/Spark => Algorithms to process distributed data

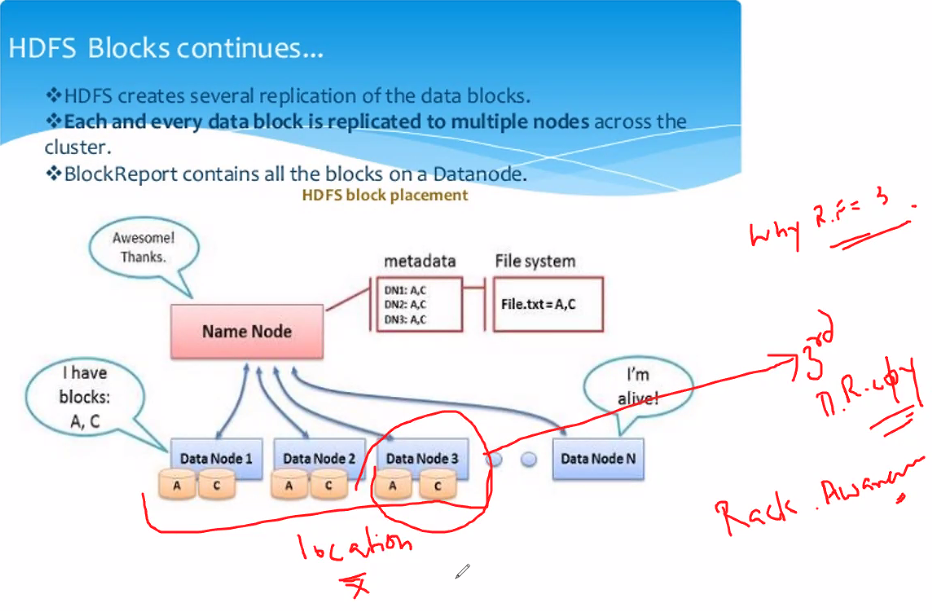
Yarn MapReduce written in Java

Hive => SQL query to Java









3rd copy => Different rack/location

DR => Disaster recovery

RF => Replication factor

