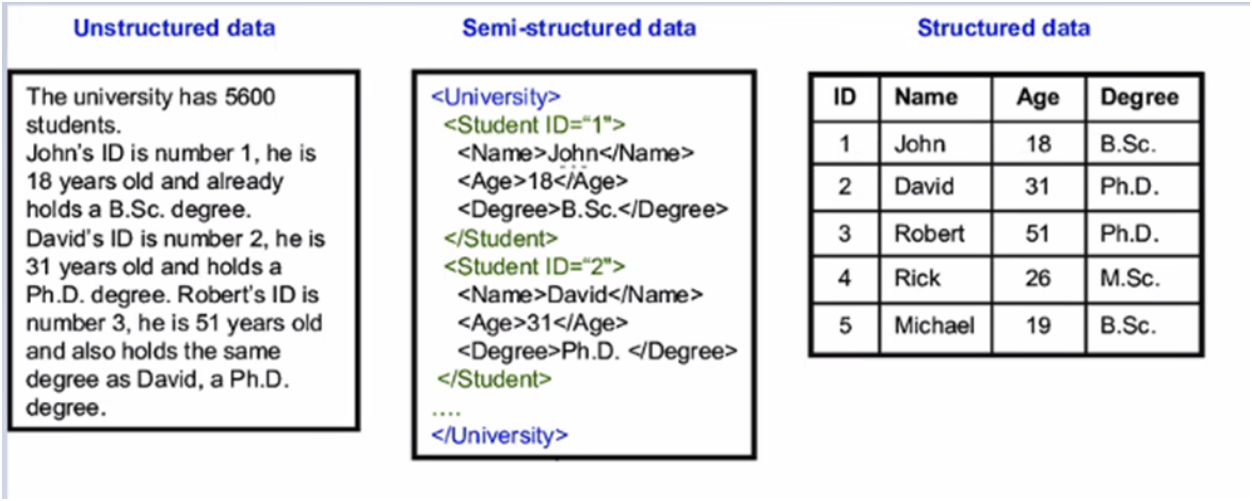
→ **Analysis:** diving deep into the data, to get some insightful info, which would be used to understand the patterns,(in order to take up some actions which would result into the growth of the organization)

→ **Action:** some steps taken up with respect to the insights that we get from analysis, for the growth of the organization

→ Forms of Data

a. Structured Data
    i. RDBMS
    ii. business data in table format
b. Unstructured Data
    i. XML, JSON, excel, Files (Key-Value)
    ii. Usually produced from social media, telematics
c. Semi-strutctured data
    i. Audio, video, text, images, social media

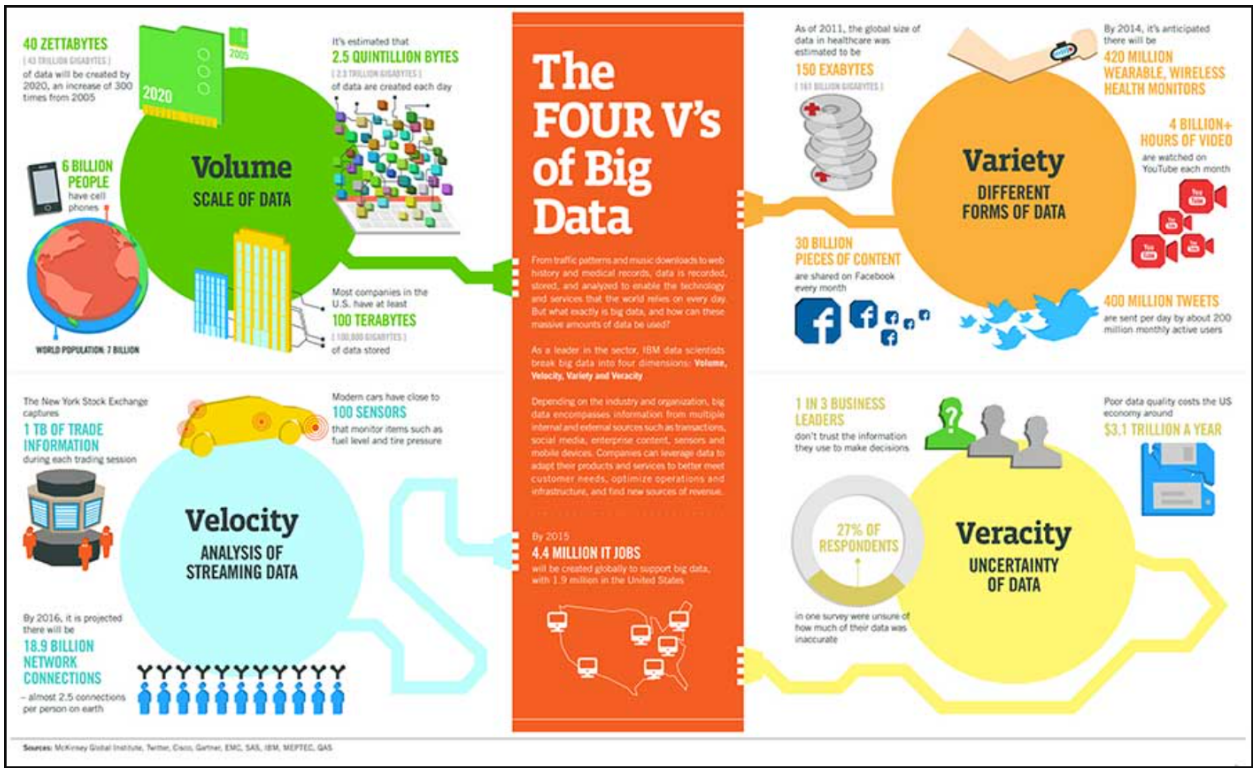


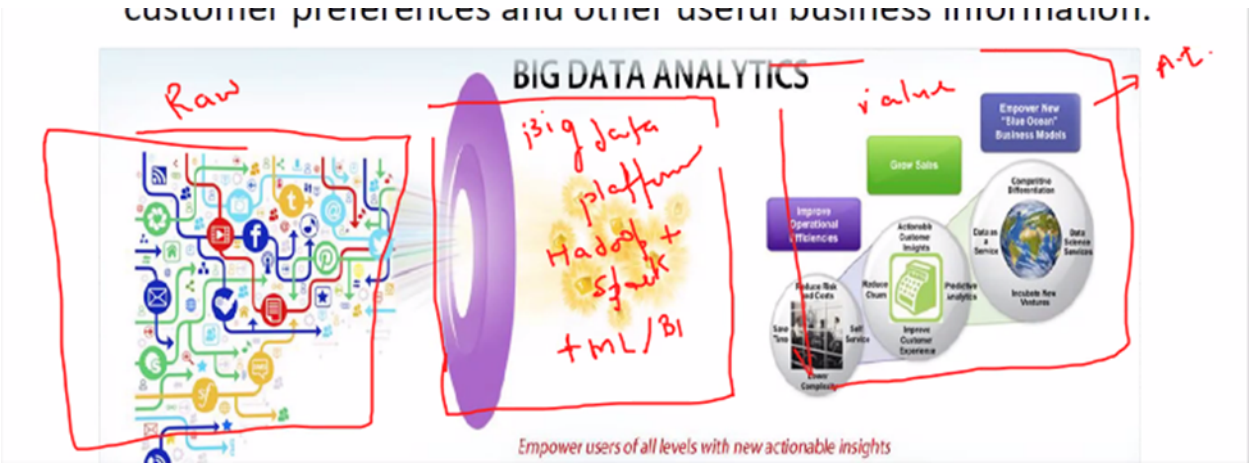
→

→ **Big Data:** any collection of data set that is large and complex

a. to handle huge volume of data, we use hadoop + spark
b. This huge volume of data can be used for analysis to understand patterns & trends
c. more data, more analysis
d. more data, more permutation, better analysis

→ 4Vs of Big Data



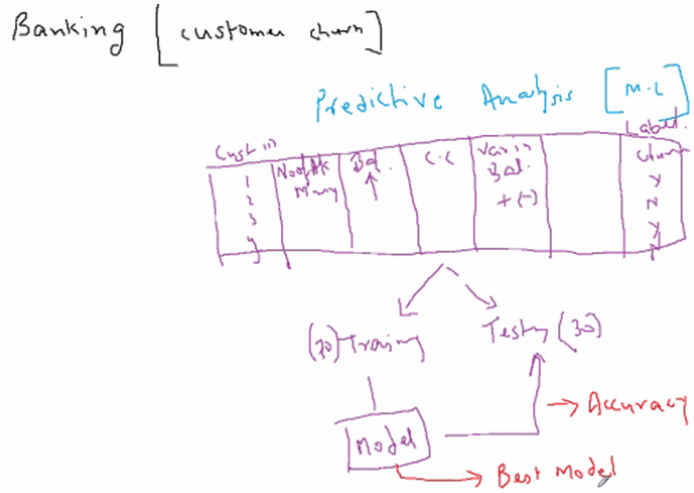


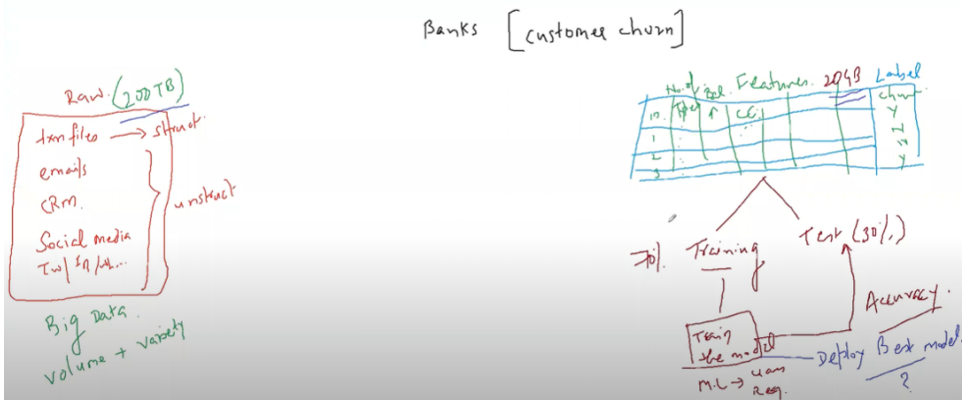
a. Volume
    i. scale of data
    ii. huge amount of data in terms of TBs & PBs
b. Variety
    i. different forms of data
    ii. Nature of data being structured, unstructured or semi-structured
    iii. Structured: has fixed format and length of data
    iv. Semi-structured: does not have a proper format/length like log files
    v. Unstructured: unorganized data, like texts, pictures, videos, etc.

c. Velocity
    i. Massive and continuous flow of data
    ii. analysis of streaming data
    iii. That which is being generated rapidly (real-time data)
    iv. Data from stock markets, location tracking, telematics like car sensors, games like cricket are generated rapidly
d. Veracity
    i. Inconsistency / uncertainty of data
    ii. Uncertainty can be in terms of
        1. Quality:sometimes data is messy and quality/accuracy of data are difficult to control (cases of data that needs to be cleaned first)
        2. Quantity: sometimes the volume of data is very less
    iii. Data which is lacking quality or quantity causes confusion or could convey incomplete information
e. Value
    i. creating value out of data
    ii. Data without any value is of no use to the organization

→ A.I. : actionable insight

→ **Customer Churn:**



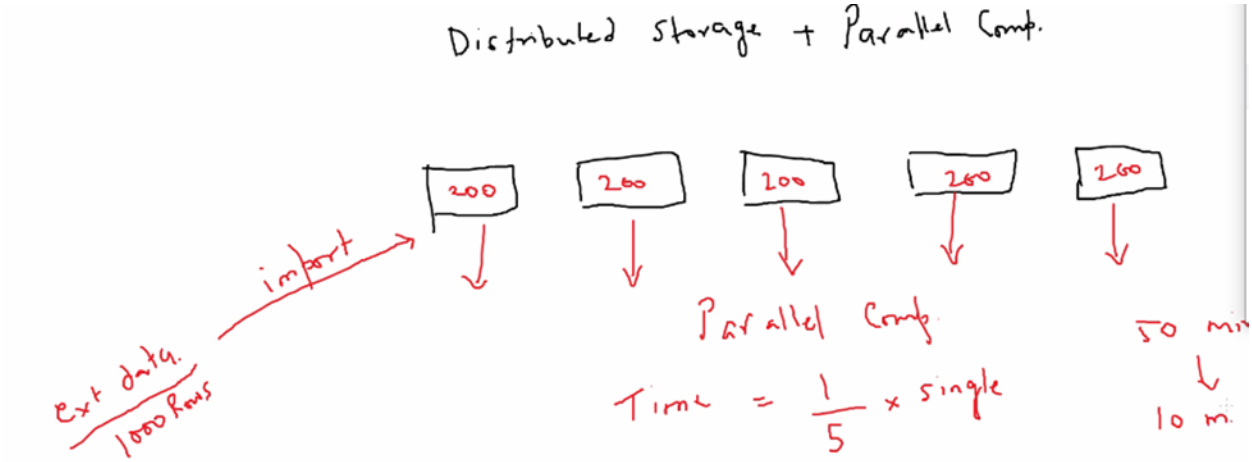


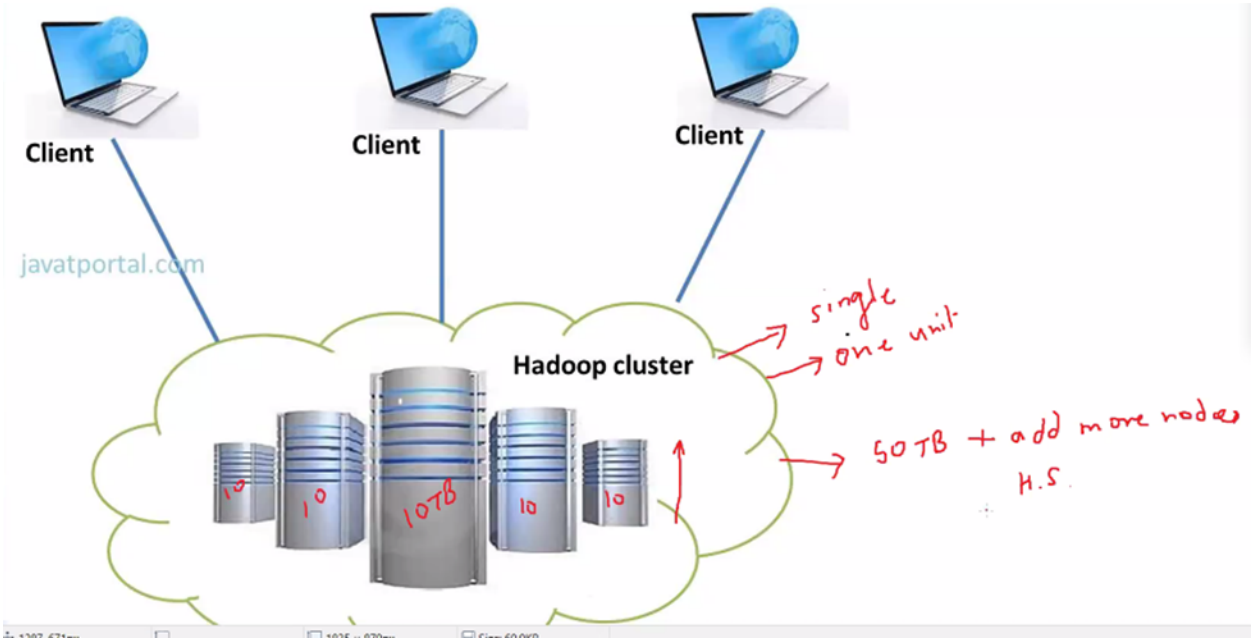
a. loss of customers based on behavior due to any bad customer service
b. We can cater to this problem using big by generating insights regarding the customers who might opt out and taking appropriate actions to save the business
c. Need to use predictive analysis tools like ML
d. Data need to be split into two parts:
    i. Training data 70%
    ii. Testing data 30%

→ spark good for real-time data due to fast processing speed
→ Hadoop recommended for huge data storage and parallel processing

→ Hadoop

a. **Cluster:** lot of machines put together in single network
b. Can store huge amount of data by scaling the hadoop nodes vertically of horizontally
c. Two types of scalability in hadoop
    i. vertical scalability: adding new nodes
    ii. horizontal scalability: increasing size of existing node
d. Can store any type of data in hadoop cluster
e. We use hadoop to store & process(transform) data
f. hadoop is based on ELT framework (Extract→ Load→ Transform)
g. It does high speed, parallel computing, so we need to store data in a distributed manner on hadoop clusters
h. It stores any amount and any type of data





→ one machine vs. 100 clusters

<u>100TB data → one machine [100MB/s]</u>

Time = data / speed

1TB = 1024 GB x 1024 MB x 100 TB /100 MB/s

= 1048576 secs

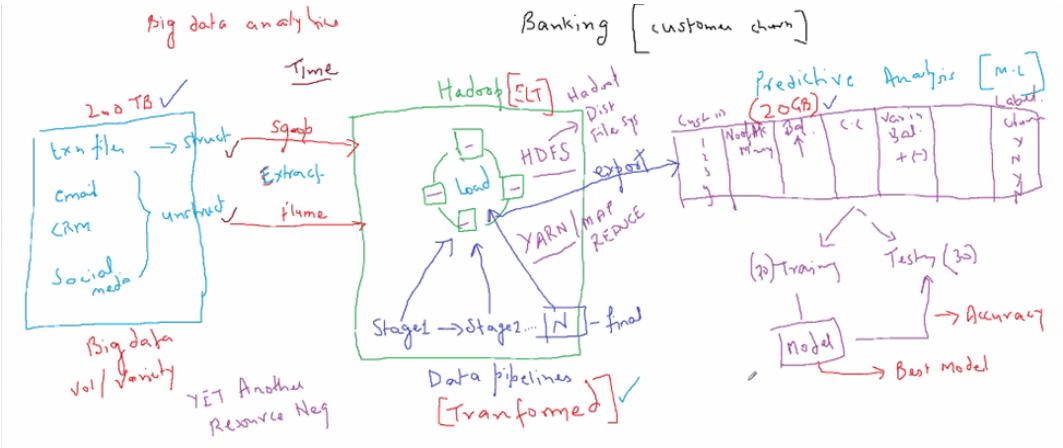
= 291 hrs (approx 12d)

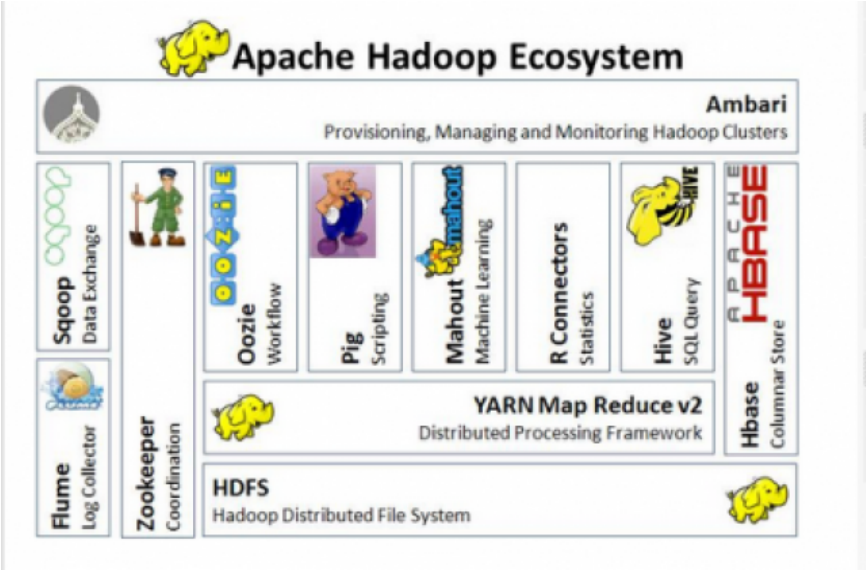
Deploy on 100 node cluster with 1TB on each node
Each node = 1TB
Time for 1TB = 1024 GB x 1024 MB / 100 MB/s
= 10486 secs
= 2.91 secs
~3hrs

500 node cluster with 200GB on each node
Each node = 200GB
Time = 45 min or under 1hr

Note: distribute data over multiple nodes trying to keep lesser/few data on each node, so that data can be accessed and processed quickly

→ Hadoop architecture

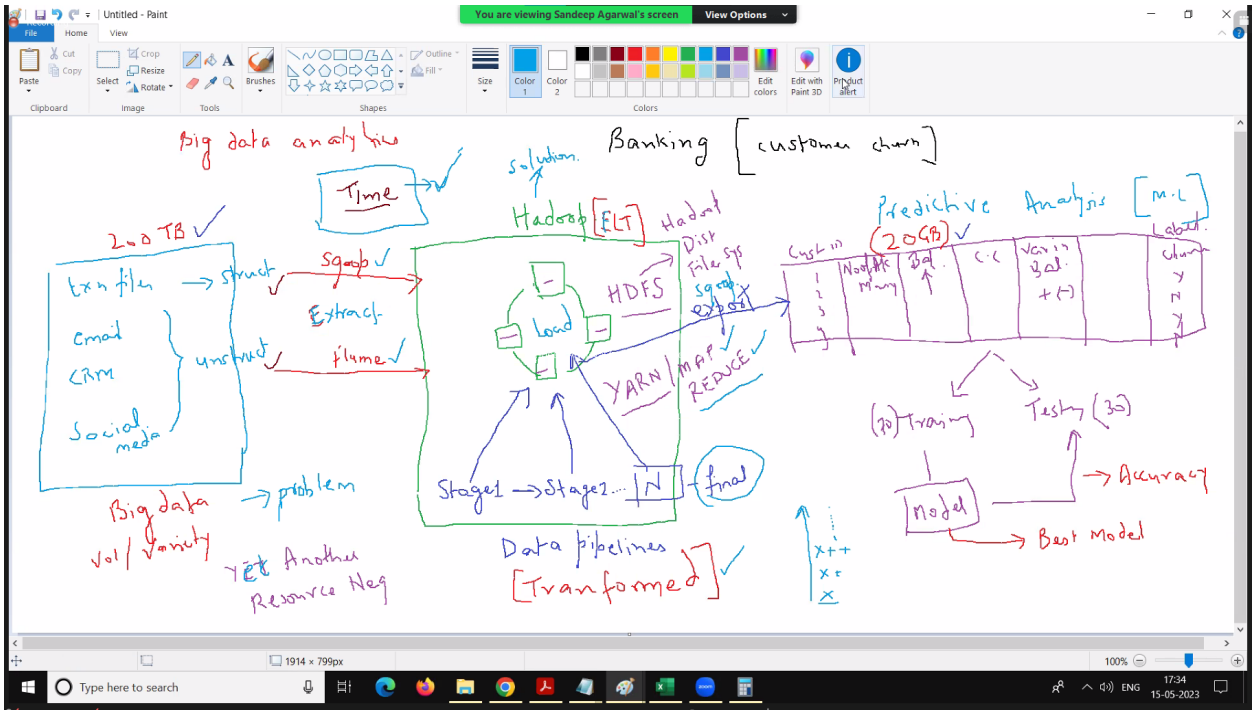




a. to import data to Hadoop
    i. For structured data: Sqoop
    ii. For unstructured data: Flume
b. HDFS:
    i. Hadoop Distributed File System,
    ii. Works on flat files (splits of files being stored)
    iii. File system to store data while processing
    iv. automatically manages data imported to HDFS
c. YARN (Yet-Another Resource Negotiator) framework:
    i. Is a service
    ii. takes Map-Reduce code to execute
    iii. Schedules jobs and tasks / set queries to run on HADOOP system
    iv. responsible for processing of data in distributed/parallel computing manner
d. Map-Reduce framework:
    i. Has two separate Map program & Reduce program
    ii. Map program takes input data and converts it into a dataset which is in key-value pair format
    iii. used to write queries to Hadoop system in parallel computing,
    iv. Written in java language
e. data pipelines
    i. passing multiple programs for parallel computing
f. Hive:
    i. Converts SQL queries to Java programs internally
    ii. Sends converted query to Map-Reduce
    iii. Cannot run SQL queries on flat files in Hadoop directly , so we need to convert to Java program
g. HBase:
    i. No Sql DB

ii. For faster retrieval as it is NoSql

h. Oozie
   i. To automate jobs



→ data is processed in stages, and each stage data is kept on Hadoop
→ once data is processed, it is exported/converted to machine learning/data science stream using Sqoop, thus transforming data
→ all the steps from raw data , fetching using Sqoop/Flume, transforming data on Hadoop & sending data into data science stream using Sqoop

→ Spark:
   a. independent system,
   b. unlike Hadoop , it is processing based system
   c. both do distributed computing, so need to use them separately based on the approach required

→ why is Hadoop important to big data apps?
   a. Big data is just huge data which is problem, and we need hadoop as a solution to transform data quickly so that we can use it in further analytics
   b. Need to process raw big data using hadoop

→ challenges for big data developer

a. Reduce turnaround time
b. Handle the increasing volume of data which increases each day
c. To optimize the processing of data
d. To prepare data & apply analytics to reduce turnaround time & increase monetary gains
e. understand processing in business