

→ **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-structured data
 - i. Audio, video, text, images, social media

Unstructured data	Semi-structured data	Structured data																								
<p>The university has 5600 students.</p> <p>John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.</p> <p>David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.</p>	<pre> <University> <Student ID="1"> <Name>John</Name> <Age>18</Age> <Degree>B.Sc.</Degree> </Student> <Student ID="2"> <Name>David</Name> <Age>31</Age> <Degree>Ph.D. </Degree> </Student> </University> </pre>	<table> <tr> <th>ID</th><th>Name</th><th>Age</th><th>Degree</th></tr> <tr> <td>1</td><td>John</td><td>18</td><td>B.Sc.</td></tr> <tr> <td>2</td><td>David</td><td>31</td><td>Ph.D.</td></tr> <tr> <td>3</td><td>Robert</td><td>51</td><td>Ph.D.</td></tr> <tr> <td>4</td><td>Rick</td><td>26</td><td>M.Sc.</td></tr> <tr> <td>5</td><td>Michael</td><td>19</td><td>B.Sc.</td></tr> </table>	ID	Name	Age	Degree	1	John	18	B.Sc.	2	David	31	Ph.D.	3	Robert	51	Ph.D.	4	Rick	26	M.Sc.	5	Michael	19	B.Sc.
ID	Name	Age	Degree																							
1	John	18	B.Sc.																							
2	David	31	Ph.D.																							
3	Robert	51	Ph.D.																							
4	Rick	26	M.Sc.																							
5	Michael	19	B.Sc.																							

→

→ **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

The FOUR V's of Big Data

Volume: SCALE OF DATA

- 40 ZETTABYTES (40 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005
- 6 BILLION PEOPLE have cell phones
- WORLD POPULATION: 7 BILLION
- It's estimated that 2.5 QUINTILLION BYTES (2.3 TRILLION GIGABYTES) of data are created each day
- Most companies in the U.S. have at least 100 TERABYTES (100,000 GIGABYTES) of data stored

Velocity: ANALYSIS OF STREAMING DATA

- The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure
- By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

Veracity: UNCERTAINTY OF DATA

- As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES (151 BILLION GIGABYTES)
- 30 BILLION PIECES OF CONTENT are shared on Facebook every month
- 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users
- 1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions
- 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate
- Poor data quality costs the US economy around \$3.1 TRILLION A YEAR

Veracity: UNCERTAINTY OF DATA

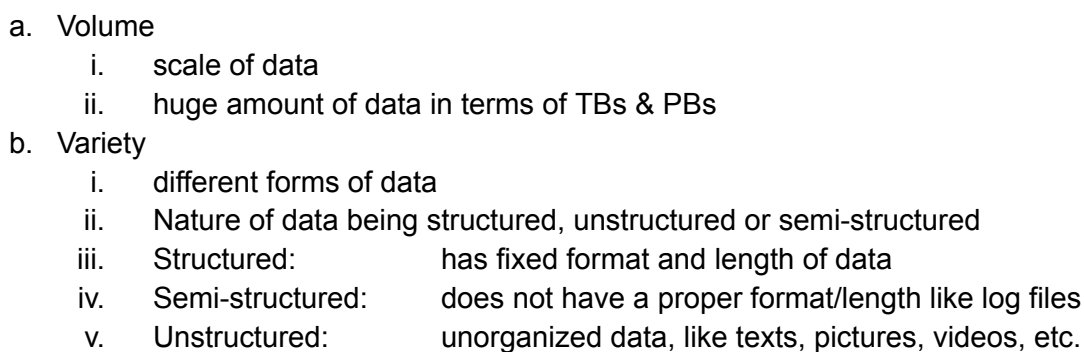
- By 2015, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

Veracity: UNCERTAINTY OF DATA

- Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Veracity: UNCERTAINTY OF DATA

- As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**



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

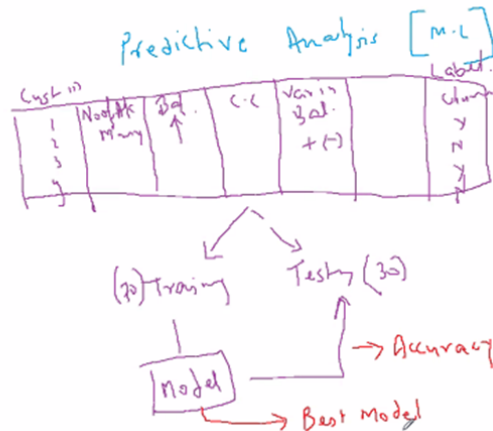
e. Value

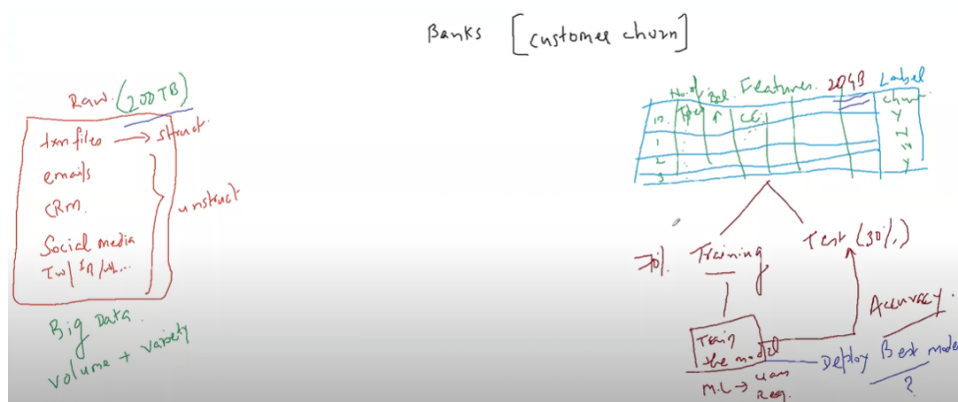
- i. creating value out of data
- ii. Data without any value is of no use to the organization

→ A.I. : actionable insight

→ **Customer Churn:**

Banking [customer churn]





- loss of customers based on behavior due to any bad customer service
- 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
- Need to use predictive analysis tools like ML
- Data need to be split into two parts:
 - Training data 70%
 - Testing data 30%

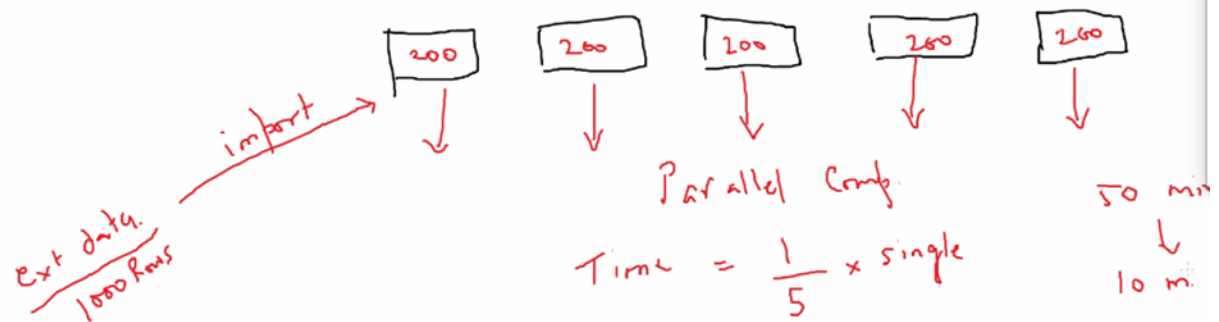
→ spark good for real-time data due to fast processing speed

→ Hadoop recommended for huge data storage and parallel processing

→ Hadoop

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

Distributed storage + Parallel Comp.



→ one machine vs. 100 clusters

100TB data → one machine [100MB/s]

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 = $1024 \text{ GB} \times 1024 \text{ MB} / 100 \text{ MB/s}$

= 10486 secs

= 2.91 secs

~3hrs

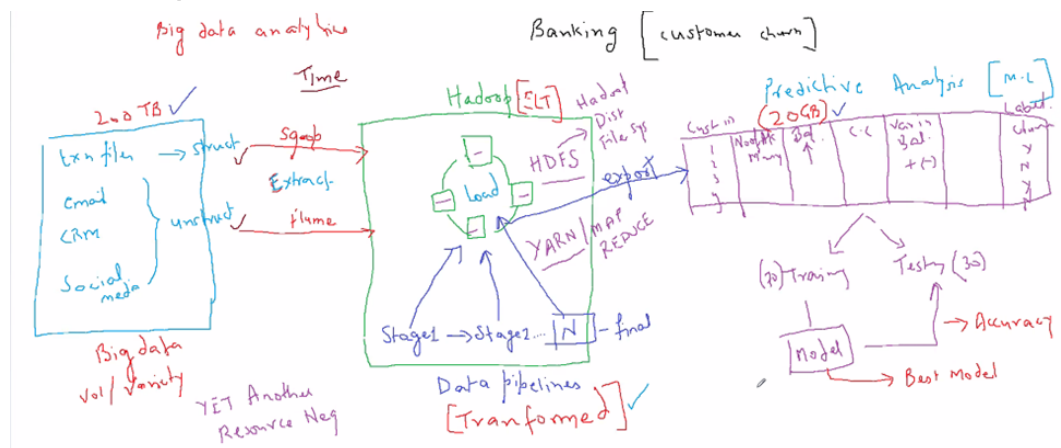
500 node cluster with 200GB on each node

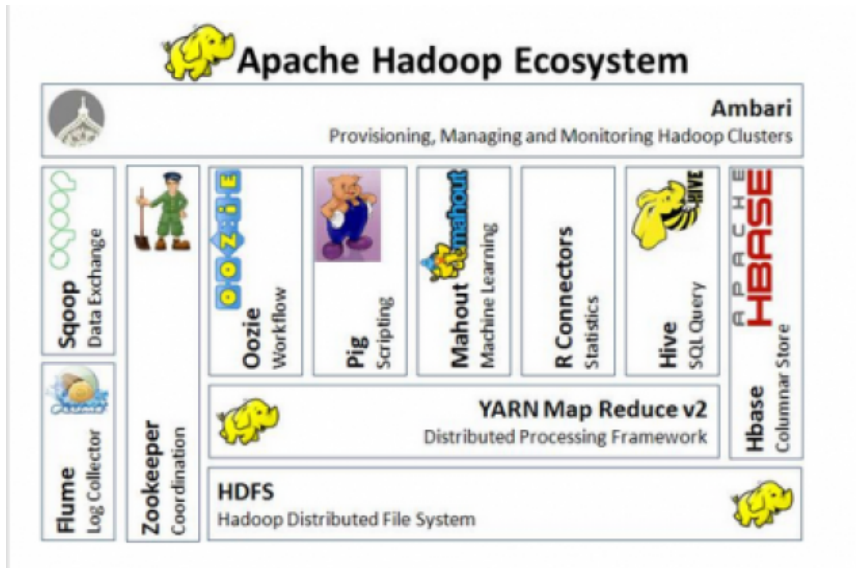
Each node = 200GB

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



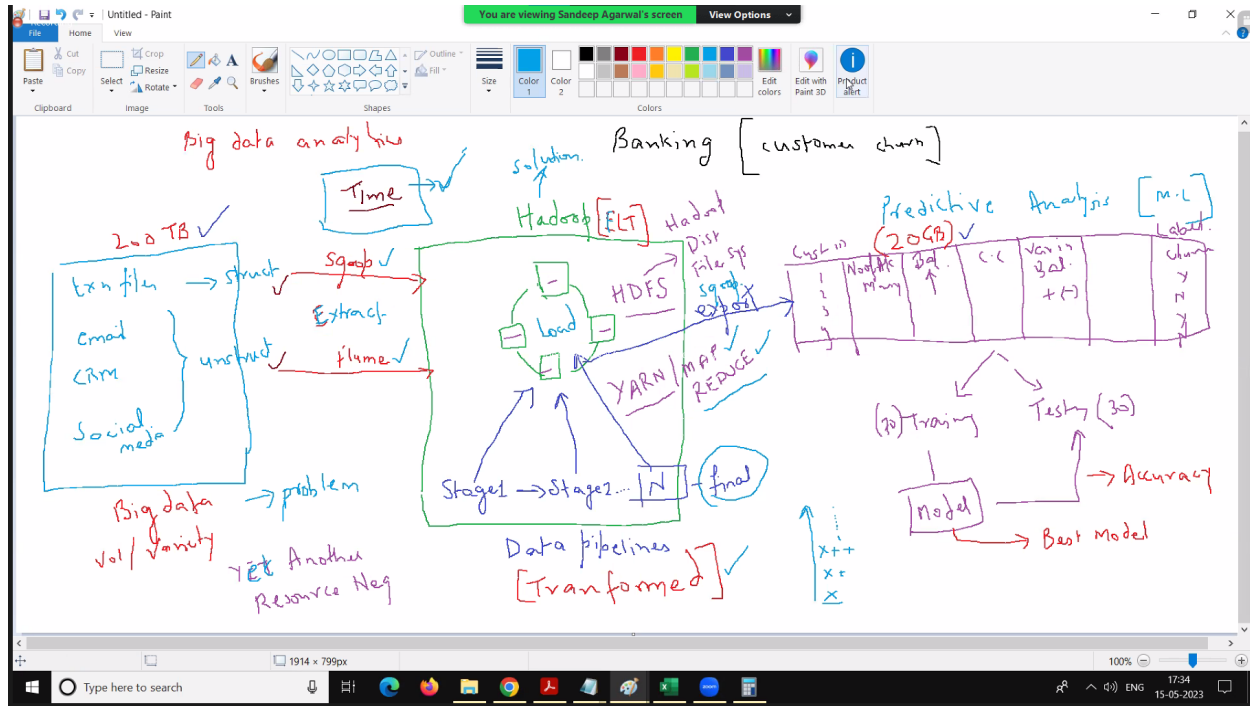


- a. to import data to Hadoop
 - i. For structured data: Sqoop
 - ii. For unstructured data: Flume
- b. HDFS:
 - i. Hadoop Distributed File System,
 - ii. Works on flat files (splits of files being stored)
 - iii. File system to store data while processing
 - iv. automatically manages data imported to 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:

- independent system,
- unlike Hadoop , it is processing based system
- both do distributed computing, so need to use them separately based on the approach required

→ why is Hadoop important to big data apps?

- 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
- 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