\* Big Data - data is generated through sensors, online shopping, airline data, hospital data, share market, cc cameras

(depends upon the company to company)

\* 5V's

1.Volume - Large amount of data (PB/HB)

2.Variety - variety of formats (csv, xml, audio, video)

3.Velocity - speed of the data(social media, cctv)

4.Veracity - uncertainity of data due to inconsistency & incompleteness & accuracy of data

5.Value - Organisation working on Bigdata achieving high ROI(Return On Investment)

\* 3 variety of data

1.Structured = in a format(i.e relational db has table format)

2.Unstructured = Videos, audios, posts

3.Semistructured data = .xml like , log files(gmail,yahoo,fb)

\* NOSQL - Not Only SQL

- Works on structured, Unstructured, Semistructured data also.

-------------------------------------Hadoop---------------------------------------------------

\* open source framework by apache s/w foundation

\* processing of large sets of data in a distributed manner across clusters of commodity computers(normal PCs) using simple programming model.

\* hadoop is solution to bigdata. It is not a technology

\* Hadoop has two parts-

1.HDFS(for Storage) - It is specially designed file system for storing huge dataset with clusters of commodity h/w and with WORA without changing the data.

2. mapreduce(for processing) - Java code or PIG(scripting) or HIVE(SQL) tool for mapReduce(they both internally converted into java then)

\* Data is stored in blocks & Default block size in h/w of cluster for 64 MB (hadoop 1.0) and we can customized it.

128 MB (hadoop 2.0)

\* In hadoop clustor setup there is master-slave system. One cluster is master and remaining are slaves.

Master service and slave service are internally talk to each other.

MASTER SLAVE

for storage Name Node Data Node

for processing Job Tracker Task Tracker

SNN(Secondary name node)

\* Name Node has a metadata file - if i want to save the file then first Name node has a information that where that

file is stored in pieces(data node) in which slave machines and same while retrieving.

\* Replication Factor - It stores the replicas of the file in the HDFS (in case of cluster damage , the data is safe)

- we can customized RF also

\* Heart Beat - Heart beat master to slave and vice versa

Heart beat Job Tracker to Task Tracker and vice versa

(in every specific interval of time in sec . we can customized it also)

Name Node & Job Tracker waits for some seconds then considered that Data Node & Task Tracker to be dead.