Big data

* It describes large volume of data both structured and unstructured and certain advanced data analytics method to extract value from data.
* 10 v of it are volume(data measured in gb,tb),velocity(speed at which data is generated,refreshed),variety(bigdata handle structured,unstructured and semi structured data),variability(same data but meaning is constantly changing),veracity(refer to reliability of data source),validity(refer to how accurate and correct data is),vulnerability(security concerns),volatility(means how long you need to store data),value(getting value from data),visualization(to visualize data in graph and trees)
* **Gzip file**-it is a file format and a software application used for compression and decompression of files.

**# gzip file format is a 10 byte header containing a magic number(1f8b),a version number and time stamp.**

**#tar.gz file is usually called tarball.**

* zip file and gzip file are not same because zip file compress file individually and cannot take advantage of redundancy between files.
* **MapReduce-**it’s a processing technique and a program model for distributed computing based on java and it consist 2 task i.e. map and reduce.

**#map takes set of data and convert it into another set where individual element are broken into key/value pair**

**#reduce take output from map as input and combile those key/value pair into a smaller set of key/value pair.**

* **Payload-** applications implement the map and reduce functions and form core of job.
* **Mapper-** maps the input key to a set of intermediate keys.
* **Namenode-** it manages hdfs
* **Datanode-** where data is present in advance before any processing takes place.
* **Masternode-** node where jobtracker runs and accept job requests from clients
* **Slavenode-** where map and reduce program runs
* **Jobtracker-** track task and track the assign jobs to tasktracker

**# Namenode maintains a metanode in 2 files:** fsimage and editlogs where fsimage initially loaded when hadoop system is started and it contains directory structure of clusters and data stored.then afterwards for every occuring transaction edit log is updated.

**#** periodically per default every hour fsimage update.

**# editlog will update on disk with new transaction and fsimage will update in main memory.**