***Big Data and The role of Hadoop Framework.***

**ABSTRACT**

TEAM DETAILS:

Name: K.Venkatesh

College: P.V.P.S.I.T

E-mail id: kolli.sunny5@gmail.com

Mobile number:8096330923

Name: M.Venkatesh

College: P.V.P.S.I.T

E-mail id: venkateshchowdarysehwag@gmail.com

Mobile number:9014492869

Transaction id: MOJO7215000N42868492

Right now if we observe, we are living in the data world. So the important thing is how to store and process the data.

**Understanding of Big-data**: Big-Data is a high volume of data which can’t be stored and processed using the traditional approach with in the given time frame.

**When do we call a data as big data**: even a small data can also be referred as big-data depending on the context we are speaking.

**Example**: when a document with 100MB is tried to attach to the email, It doesn’t accept it. It means that 100MB is a big data with reference to Email.

**Characteristics of Big-data (given by IBM)**: 3V’s.

**1.Volume**: Rapid increasing of data in our day to day life.

**2.Velocity**: Higher volumes of data entering your organizations per sec.

**3.Variety**: Basically we use RDBMS for storing structured Data, but now a days there is also a generation of unstructured(photos,MP3) and semi structured data(log files).

**Traditional approach followed to access more data:**

->The entire data is given as input to the ETL system, then the ETL system extracts data and convert this data into specific format and load this data into database.

Then the end user process the data from database through the code. But as the data is increased rapidly, it became challenge to manage and process.

**Disadvantages of traditional approach**:

1.expensive.

2.Time consuming.

3.Scalability.

-> so the basic thing is COMPUTATION IS PROCESSOR BOUND.

->It means The data is increasing but processing power is not increasing.

->Storing data some where and accessing is not a reliable thing, so it is better to Access data in your local file System.

\* So in that sense Hadoop had came into existence.

**Understanding of hadoop:**

->open source framework developed by Doug cutting, which is managed by Apache software Foundation.

->Large scale processing of data sets on clusters of commodity hardware.

**History of Hadoop**: Google had presented a paper with two techniques in 2004.they are:

->Google File System(GFS) for storing data.

->Map Reduce for processing that data.

Later YAHOO implemented that two techniques as:

->Hadoop Distributed File System for storing data.

->Map Reduce to process that data.

The name Hadoop is nothing but Doug Cutting son’s Toy(Elephant) name.

**How data is managed and processed?**

->It divides the existed data into smaller chunks, each data part is stored in each node with in a cluster. Replication facility is there in Hadoop.

**Hadoop Features:**

1.Cost effective System.(Commodity-hardware).

1. It supports Large cluster of nodes.
2. It supports parallel processing.
3. Distributed Data.
4. Automatic Failover Management.
5. Data Locality Optimization.
6. Heterogeneous cluster.
7. Scalability.

**Hadoop Ecosystem:** An Array of Related Software.(to increase the efficiency)

They are:

1. Apache PIG -> Scripting Language.

2. Apache HBASE -> Column oriented Database.

3. Apache HIVE -> SQLite Language.

4. Apache SQOOP ->For Transferring Data.

5.Apache FLUME ->For Streaming Data.

6.Apache ZOOKEEPER->take care of Functionality of above 5 softwares.

**Conclusion**:

->This technology is still in the progress of further advance development.

->So at present situation, Based on three characteristics All organizations, social networks etc… are running on the base of hadoop technology.

**References**:

* Hadoop for Dummies(by Dirk deRoos).
* https://www.youtube.com/watch?v=Pq3OyQO-l3E.