Big Data Analytics

S.SAI SRI HARI,

2nd year BTech,

Regd.No:15H41A0552,

DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING,

BVCITS,

AMALAPURAM.

**ABSTRACT:** Big data analytics is the process of examining large data sets to uncover hidden patterns unknown correlations, market trends, customer preferences and other useful business information. Big data consists of data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage and process data with tolerable elapsed time.

By applying big data principles into the concepts of machine intelligence and deep computing, IT Departments can predict potential issues and move to provide solutions before the problems even happens. It also possible to predict winners in a match using big data analytics. Future performance of players could be predicted as well, Thus players’ value and salary is determined by Data collected throughout season. Example: movie: Money Ball.

**Introduction**:

Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crimes and so on,

Data sets grow rapidly in part because they are increasingly gathered by cheap and numerous information like sensing mobile devices, aerial, software logs, cameras, microphones etc.

The world technologies per captia capacity to store capacity to store information has roughly doubled every 40 months since 1980’s as of 2012, everyday 2.5exabytes of data are generated.

Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data.

The work may require “massively parallel software running on tens, hundreds, or even thousands of servers what count as big data varies depending on the capabilities of the user and their tools and expanding capabilities make big data a moving target.

Big data requires a set of techniques and technologies with new forms of integrations to reveal insights from datasets that are diverse, complex and of a massive scale.

1. Increasing volume(amount of data)
2. Velocity(speed of data in and out)
3. Variety (range of data types and sources)

In 2004 Google published a paper on a process called MapReduce.

**MapReduce**:

MapReduce concept provides a parallel processing model, and associated implementation was released to process huge amount of data.

With MapReduce, queries are spit and distributed across parallel nodes and processed in parallel (the Map step). The result are then gathered and delivered (the Reduce step).

This MapReduce framework was adopted by Apache open-source project named Hadoop

  **Working of map and reduce technique**

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment it is part of the apache project sponsored by the apache software foundation.

Hadoop framework mainly dependents on two things

1. Hadoop Distributed File System
2. Hadoop MapReduce

Thus MapReduce undergone for more techniques those are

1. Data placement in Heterogeneous cluster
2. Initial data placement
3. Data redistribution
4. A Parallel genetic algorithm
5. Image based MapReduce using HIPI

The base Apache Hadoop framework is composed of the following modules

1. Hadoop common
2. Hadoop Distributed File System
3. Hadoop Yarn
4. Hadoop MapReduce

**Conclusion:**

Hadoop’s unique storage method is based on distributed file system that basically ‘maps’ data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you were dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes and petabytes in hours.