Apache Hadoop

Malhari Kambale

Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur

Third Year Engineering Student

Introduction:-

Now a day in enterprises tremendous amount of data is present, to store that data and maintain it becomes very difficult .How to process big data is the biggest problem occurred in front of enterprises?

For example in bank there are no of customers opens their account, so to store and maintain the information of that particular customer (Big data) is occupied large amount of space as well as time and it takes more cost to maintain it by other software's. So to avoid it the hadoop is come into existence.

"Apache Hadoop "is open source software which processes the big data within less time as well as less cost. Due to this it becomes most popular than the others software. The hundreds of enterprises are using it because of its beauty of effective processing, storing, and maintaining big data.

Architecture of Hadoop:-

Basically hadoop is built on the two main components that is

1. HDFS (Hadoop Distributed File System)
2. MapReduce

1. HDFS:-

The file is stored in the large blocks on the node on cluster of hadoop. The blocks are placed on the separate machines that the data can be replicated on three separate machines that is three separate nodes to protect against future failure.

Because in future if any individual node get fails the data can be read from another node so that there is no need to worry about reading or accessing data ,so it gives better efficiency as well as reliability.

2. Map Reduce:-

- MapReduce is programming model that allows to process large amount of data.
- Map and Reduce are two methods which run parallel at the same time.
- MapReduce designed to work with HDFS.
- Each Map task operates on a discrete portion of the overall
- Dataset
- Typically one HDFS data block
- After all Maps are complete, the MapReduce system distributes
- The intermediate data to nodes which perform the Reduce phase– Much more on this later!

Architecturally, the reason you're able to deal with lots of data is because Hadoop spreads it out. And the reason you're able to ask complicated computational questions is because you've got all of these processors, working in parallel, harnessed together.

Hadoop Data Services:-

Hadoop having following data services:-

- **Apache Hive**
- **Apache Pig**
- **Apache HCatalog**

Future of Hadoop:-

These people, who want to improve the performance of their companies and unlock new business opportunities, realize that including Apache Hadoop as a deeply integrated supplement to their current data architecture offers the fastest path to reaching their goals while maximizing their existing investments.