softtech® group

HADOOP

& Its Features

Data Sheet

Hadoop and its Features

What is Hadoop

Apache Hadoop is an open source software platform that is used to build applications for data processing that are run in a distributed computing environment.

Applications developed using HADOOP run on massive data sets spread through commodity computer clusters. Commodity computers are moderately priced and readily available. These are primarily useful for achieving greater low cost computing power.

Hadoop Ecosystem

Apache Hadoop consists of various tools needed by Hadoop to perform various tasks.

MapReduce - MapReduce is a programming model and software system for Hadoop-run applications for writing. These MapReduce programs are capable of processing massive data on large clusters of computing nodes in parallel.

HDFS - The storage aspect of Hadoop applications is taken care of by HDFS. Applications MapReduce consumes data from HDFS. HDFS generates and distributes several data block replicas to compute nodes in a cluster. This distribution allows computations that are accurate and extremely rapid.

Sqoop - sqoop is used to transfer data between Hadoop and external datastores, such as relational databases and warehouses of business data.

Flume - Flume is a distributed service designed to store, compile and transfer vast volumes of log data.

Pig: Pig is used in Hadoop to analyses the data. To perform various operations on the data, it provides a high-level data processing language.

Hive - Hive supports SQL (Hive Query Language) reading, writing, and handling massive data sets residing in the distributed storage.

Spark - Spark is a distributed open-source software engine for the collection and analysis of immense real-time data volumes.

Mahout - Mahout is used to build scalable machine learning algorithms, which are distributed. It has a library for collaborative filtering, classification and clustering that includes in-built algorithms.

Ambari - Ambari is an open source tool which is responsible for monitoring running applications and their status.

Kafka - Kafka is a framework for distributed streaming to store and process record streams. It constructs data streaming pipelines in real time that reliably get data between applications.

Storm - Storm is a processing engine, which processes data streaming at a very high speed in real time. It has the capacity to process over a million jobs on a node in a fraction of seconds.

Oozie - To handle Hadoop jobs, Oozie is a workflow scheduler framework. It has two parts, a working engine and an engine coordinator.

Hadoop Architecture

Using MapReduce and HDFS approaches, Hadoop has a Master-Slave Architecture for data storage and distributed data processing.

NameNode – NameNode described all files and directories that are used in namespace

DataNode – DataNode enables you to monitor the state of an HDFS node and to communicate with the blocks.

MasterNode – The master node lets you use Hadoop MapReduce to perform parallel data processing.

SlaveNodes – The slave nodes are the Hadoop cluster's additional machines that allow you to store data to perform complicated calculations. In addition, the entire slave node comes with a DataNode and a Task Tracker. This helps you to synchronize the processes to the respective NameNode and Work Tracker.

Features of Hadoop

- **Hadoop is Open Source**
- **Hadoop cluster is Highly Scalable**
- **Hadoop provides Fault Tolerance**
- **Hadoop provides High Availability**
- **Hadoop is very Cost-Effective**
- **Hadoop is Faster in Data Processing**
- **Hadoop is based on Data Locality concept**
- **Hadoop provides Feasibility**
- **Hadoop is Easy to use**