SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution
Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF MECHATRONICS ENGINEERING

UNIT 3 – HADOOP

ANATOMY OF HADOOP

ANATOMY OF HADOOP

The anatomy of Hadoop refers to the core components and architecture of the Hadoop ecosystem, which is designed to handle the distributed storage and processing of large-scale data sets. Here's an overview of the key components that make up the anatomy of Hadoop:

Hadoop Distributed File System (HDFS):

1.NameNode:

1. The NameNode is the master server that manages the metadata of the files and directories stored in HDFS. It keeps track of the location and structure of the data.

2.DataNodes:

1. DataNodes are the worker nodes responsible for storing and managing the actual data blocks. They communicate with the NameNode to report the status of the data blocks they store.

3.Block Size:

1. HDFS divides data into fixed-size blocks (typically 128 MB or 256 MB). Each block is replicated across multiple DataNodes for fault tolerance.

4.Replication:

1. Data replication is a key feature of HDFS. By default, each block is replicated three times across different DataNodes to ensure data availability and fault tolerance.

ANATOMY OF HADOOP

MapReduce:

1.Map Task:

1. The Map phase of MapReduce processes input data and produces a set of intermediate key-value pairs.

2.Shuffle and Sort:

1. After the Map phase, the system performs a shuffle and sort operation to group and order the intermediate key-value pairs before passing them to the Reduce tasks.

3.Reduce Task:

1. The Reduce phase aggregates and processes the intermediate key-value pairs, producing the final output.

4.JobTracker (Deprecated):

1. In older versions of Hadoop, the JobTracker was responsible for managing and scheduling MapReduce jobs. However, with the introduction of YARN, this functionality has been distributed across ResourceManager and ApplicationMaster.

5.TaskTracker (Deprecated):

1. In older versions, the TaskTracker managed individual tasks (Map and Reduce). With YARN, this functionality is part of the NodeManager.

ANATOMY OF HADOOP

Yet Another Resource Negotiator (YARN):

1.ResourceManager:

1. The ResourceManager is the master node that manages and allocates resources across the Hadoop cluster. It receives resource requests from clients and schedules applications.

2.NodeManager:

1. NodeManagers run on individual nodes and are responsible for managing resources on that node. They report the available resources back to the ResourceManager and execute tasks.

3.ApplicationMaster:

1. The ApplicationMaster is responsible for negotiating resources with the ResourceManager and coordinating the execution of tasks within a specific application.

ANATOMY OF HADOOP

Hadoop Ecosystem:

1.Apache Hive:

1. A data warehousing and SQL-like query language for Hadoop that facilitates querying and managing large datasets.

2.Apache Pig:

1. A high-level scripting language and platform for analyzing large datasets. Pig scripts are executed in a multi-step process, making it easier to process complex data flows.

3.Apache HBase:

1. A NoSQL database that provides real-time, random read and write access to large datasets. HBase is designed to scale horizontally.

4.Apache Spark:

1. An in-memory data processing engine that provides a faster and more flexible alternative to MapReduce. Spark can be used for various data processing tasks, including batch processing, machine learning, and streaming.

ANATOMY OF HADOOP

1.Apache Sqoop:

1. A tool for efficiently transferring data between Hadoop and relational databases.

2.Apache Flume:

1. A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

3.Apache Oozie:

1. A workflow scheduler for managing Hadoop jobs. Oozie allows users to define, schedule, and manage workflows with multiple Hadoop jobs.

4.Apache ZooKeeper:

1. A distributed coordination service that provides distributed synchronization and maintenance of configuration information, critical for managing distributed systems.

5.Cloudera Impala, Apache Drill, PrestoDB:

1. Interactive query engines that provide faster SQL queries for data stored in Hadoop.