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

INTRODUCTION TO HADOOP

INTRODUCTION TO HADOOP

Hadoop is an open-source framework designed for the distributed storage and processing of large data sets using a cluster of commodity hardware. It is part of the Apache Software Foundation and is widely used in big data analytics and processing. Hadoop provides a scalable, fault-tolerant, and cost-effective solution for handling massive amounts of data.

Core Concepts:

1.Data Nodes:

1. Machines in the Hadoop cluster that store the actual data.

2.Name Node:

1. The central server that manages metadata and controls access to files stored in HDFS.

3.Job Tracker:

1. In earlier versions of Hadoop (pre-YARN), Job Tracker was responsible for managing MapReduce jobs. In YARN, this functionality is distributed across Resource Manager and Application Master.

4.Task Tracker:

1. In earlier versions of Hadoop (pre-YARN), Task Tracker managed individual tasks. In YARN, this functionality is part of Node Manager.

5.Resource Manager:

1. YARN component that allocates resources and schedules applications.

6.Node Manager:

1. YARN component responsible for managing resources on individual nodes.

INTRODUCTION TO HADOOP

How Hadoop Works:

1.Storage:

1. Data is stored in HDFS, which distributes the data across nodes in the cluster.

2.Processing:

1. MapReduce programming paradigm is used for distributed processing. Map tasks process data in parallel across the nodes, and reduce tasks aggregate the results.

3.Fault Tolerance:

1. Hadoop provides fault tolerance by replicating data across multiple nodes. If a node fails, tasks are redirected to other nodes with copies of the data.

Use Cases:

1.Batch Processing:

1. Hadoop is well-suited for batch processing of large volumes of data, making it a key tool for tasks like log analysis, data warehousing, and ETL (Extract, Transform, Load) processes.

2.Data Storage and Retrieval:

1. HDFS is used for storing vast amounts of data, and Hadoop provides mechanisms for efficient retrieval and processing.

3.Data Transformation:

1. Hadoop facilitates the transformation of raw data into a structured and usable format through its MapReduce paradigm.

4.Data Analysis and Exploration:

1. Hadoop is commonly employed for data analysis tasks, enabling organizations to derive insights from large datasets.

INTRODUCTION TO HADOOP

Ecosystem:

Hadoop has a rich ecosystem of related projects and tools that extend its capabilities. Some notable components include:

- **Apache Hive:** Data warehousing and SQL-like query language for Hadoop.
- **Apache Pig:** Platform for analyzing large datasets using a high-level scripting language.
- **Apache HBase:** Distributed, scalable, and NoSQL database for Hadoop.
- **Apache Spark:** In-memory data processing engine for faster and more flexible data processing.

Challenges:

- **Complexity:** Setting up and configuring a Hadoop cluster can be complex.
- **Skill Requirements:** Users need to be familiar with Java and the Hadoop ecosystem tools.
- **Data Movement Overhead:** Data needs to be moved to the Hadoop cluster, which can incur additional overhead.

Evolution:

Hadoop has evolved over the years, with advancements such as the introduction of YARN, improvements in performance, and the integration of new tools like Apache Spark. As the big data landscape continues to evolve, Hadoop remains a foundational technology for many organizations dealing with large-scale data processing and analytics.