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

HADOOP SCALABILITY

HADOOP SCALABILITY

Scalability is a critical aspect of Hadoop, as it is designed to handle large-scale data processing across distributed clusters of commodity hardware. Hadoop achieves scalability through several key features and principles:

1.Distributed Storage (HDFS):

1. Hadoop Distributed File System (HDFS) enables the distributed storage of large datasets across multiple nodes in a cluster. The data is divided into blocks, and each block is replicated across different nodes for fault tolerance. This distributed storage model allows Hadoop to scale horizontally by adding more nodes to the cluster.

2.Distributed Processing (MapReduce and YARN):

1. Hadoop uses a distributed processing model with MapReduce for data processing. MapReduce divides tasks into map and reduce phases, enabling parallel processing across nodes in the cluster. With YARN (Yet Another Resource Negotiator), Hadoop can manage and allocate resources efficiently, allowing multiple applications to share the same cluster.

3.Horizontal Scaling:

1. Hadoop's architecture is built for horizontal scaling, meaning that you can increase the capacity of the system by adding more commodity hardware to the cluster. This allows organizations to scale their Hadoop infrastructure as data volumes and processing requirements grow.

HADOOP SCALABILITY

1.Data Replication for Fault Tolerance:

1. Hadoop ensures fault tolerance by replicating data across multiple nodes. The default replication factor in HDFS is three, meaning that each data block is stored on three different nodes. If a node fails, the system can retrieve the data from one of the replicated copies, ensuring data availability.

2.Decentralized Resource Management (YARN):

1. YARN decouples the resource management and job scheduling functions, allowing for more efficient resource utilization. YARN supports multi-tenancy, enabling multiple applications to share resources on the same Hadoop cluster. This flexibility contributes to scalability.

3.Commodity Hardware:

1. Hadoop is designed to run on commodity hardware, which means using relatively inexpensive, off-the-shelf servers. This approach to hardware lowers the cost of scaling the cluster, making it more cost-effective to add additional nodes.

HADOOP SCALABILITY

1.Dynamic Scalability:

1. Hadoop allows for dynamic scalability, meaning that nodes can be added or removed from the cluster without disrupting ongoing operations. This flexibility is valuable for adapting to changing data processing requirements.

2.Hadoop Ecosystem:

1. The Hadoop ecosystem includes a variety of tools and projects that complement the core Hadoop components. These tools enhance the overall functionality of the platform and provide specialized capabilities. As organizations face different challenges, they can leverage specific tools from the ecosystem to address their unique requirements.

3.Community and Open Source Model:

1. The open-source nature of Hadoop encourages collaboration and innovation within the community. This allows for the continuous improvement of the platform, addressing scalability challenges and incorporating new technologies.

While Hadoop is known for its scalability, it's essential to consider other factors, such as network bandwidth, hardware specifications, and cluster management practices, when planning and optimizing a Hadoop deployment for scalability. Organizations should also be mindful of the specific needs of their use cases and choose appropriate configurations and tools within the Hadoop ecosystem to achieve optimal scalability.