Module 2 428

Hadoop Cluster Setup and Management" submodule into its components and illustrate their practical applications.

Submodule	**Description**	**Practical Examples**
Hadoop Installation and Configuration	This involves setting up the core Hadoop components (HDFS, YARN, MapReduce) on individual nodes in the cluster and configuring them to work together seamlessly.	* Downloading and installing the Hadoop distribution on each node. * Configuring core-site.xml, hdfs-site.xml, yarn-site.xml, and mapred-site.xml files to define cluster settings like HDFS namenode address, YARN resource manager address, etc.
Hadoop Cluster Architecture	This defines the overall structure of the Hadoop cluster, including the roles of different nodes (namenode, datanodes, resource manager, node managers) and their interconnections.	* Setting up a High Availability (HA) cluster with two namenodes for fault tolerance. * Configuring a cluster with multiple racks to optimize data locality and network bandwidth usage.
Hadoop Cluster Management and Monitoring	This encompasses the tools and techniques used to manage and monitor the health and performance of the Hadoop cluster.	* Using the Hadoop web interface to view cluster metrics, node status, and running jobs. * Setting up alerts and notifications to proactively identify and address issues like node failures or resource bottlenecks. * Employing tools like Ganglia or Nagios for comprehensive cluster monitoring and performance analysis.
Hadoop Security	This deals with securing the Hadoop cluster and its data from unauthorized access and threats.	* Enabling Kerberos authentication to ensure secure communication between cluster components. * Implementing access control lists (ACLs) to restrict access to sensitive data. * Configuring data encryption at rest and in transit to protect data confidentiality.

Key Points to Remember

- The specific tools and techniques used within each submodule can vary depending on the Hadoop distribution (Apache Hadoop, Cloudera, Hortonworks, etc.) and the specific requirements of the cluster.
- Effective cluster setup and management are crucial for ensuring the reliability, performance, and security of Hadoop-based big data applications.

More Examples

Hadoop Installation and Configuration

- Configuring the hadoop-env.sh file to set environment variables like JAVA_HOME.
- Setting up SSH key-based authentication for passwordless communication between nodes.

Hadoop Cluster Architecture

- Configuring a federated cluster with multiple independent namenodes to scale HDFS namespace.
- Setting up a heterogeneous cluster with nodes having different hardware configurations to optimize resource utilization.

Hadoop Cluster Management and Monitoring

- Using Apache Ambari or Cloudera Manager for centralized cluster management, monitoring, and configuration.
- Leveraging log aggregation and analysis tools like Elasticsearch, Logstash, and Kibana (ELK stack) to gain insights into cluster behavior and troubleshoot issues.

Hadoop Security

- Implementing transparent data encryption in HDFS to protect data at rest without requiring application modifications.
- Setting up role-based access control (RBAC) to grant fine-grained permissions to users and groups.