Department of MCA

Topic: Analyzing The Data with Hadoop Components

COURSE

19CA917

Big Data Analytics

UNIT - II

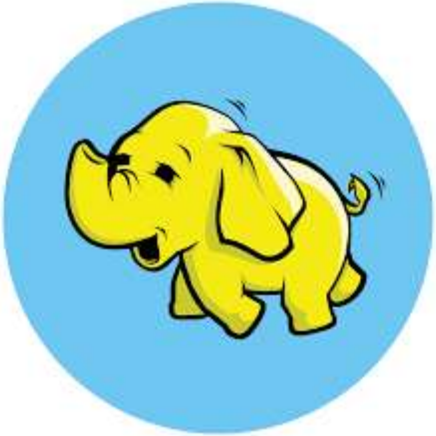
Hadoop

CLASS

III Semester / II MCA

Session Objective

- ❑ Understand the building blocks of HDFS and its functions
- ❑ Demonstrate how read and write operation take place in distributed environment
- ❑ Differentiate HDFS from relational database management system



Problem

- Big data is everywhere, and businesses are always looking for ways to analyze and make sense of it.
- One popular solution is Hadoop, an open-source software framework that allows for the distributed processing of large datasets across clusters of computers.
- Hadoop has several components that work together to provide a powerful and flexible platform for analyzing big data.

Analyzing The Data with Hadoop Components

- The first component of Hadoop is Hadoop Distributed File System (HDFS), which is responsible for storing and managing the data.
- HDFS is a distributed file system that can handle petabytes of data and is designed to be fault-tolerant. It also provides high throughput access to data, making it ideal for big data applications.

Analyzing The Data with Hadoop Components

- The second component is YARN (Yet Another Resource Negotiator), which is responsible for managing the resources in a Hadoop cluster. YARN allows multiple applications to run on the same cluster, ensuring that resources are allocated fairly and efficiently.
- With YARN, businesses can run multiple applications on the same cluster, providing a more streamlined and efficient workflow.

Analyzing The Data with Hadoop Components

- The third component is MapReduce, which is responsible for processing the data. MapReduce is a programming model that allows for the distributed processing of large datasets.
- It breaks the data into smaller chunks and processes them in parallel on different nodes in the cluster. This allows for faster processing times and more efficient use of resources.

Analyzing The Data with Hadoop Components

- In addition to these core components, Hadoop also has several other tools and technologies that enhance its functionality. For example, Hive is a data warehouse system that allows for the querying and analysis of large datasets stored in Hadoop.
- Pig is a high-level scripting language that allows for the processing of large datasets without the need for complex programming.

Analyzing The Data with Hadoop Components

- One of the main benefits of Hadoop is its scalability. Because it is designed to run on clusters of commodity hardware, businesses can easily add more nodes to the cluster as their data needs grow. This makes Hadoop a cost-effective solution for businesses that need to analyze large datasets.
- Another benefit of Hadoop is its flexibility. Because it is an open-source software framework, businesses can customize it to meet their specific needs. They can add their own applications and tools to the Hadoop ecosystem, making it a highly customizable platform.

Analyzing The Data with Hadoop Components

•Hadoop also provides businesses with the ability to analyze data in real-time. With tools like Apache Storm and Spark Streaming, businesses can process data as it is generated, allowing them to make real-time decisions based on the data.

•However, there are some challenges associated with using Hadoop. One of the biggest challenges is the complexity of the system.

•Hadoop has a steep learning curve, and businesses may need to hire specialized personnel to manage and maintain the system. Additionally, because Hadoop is a distributed system, it can be difficult to troubleshoot issues that arise.

Analyzing The Data with Hadoop Components

- Another challenge is the security of the system. Because Hadoop is designed to be an open system, it can be vulnerable to security threats. Businesses need to ensure that they have proper security measures in place to protect their data.
- Despite these challenges, Hadoop remains a popular solution for big data analytics. It provides businesses with a powerful and flexible platform for analyzing large datasets, and its scalability and real-time processing capabilities make it an attractive option for businesses of all sizes.

Analyzing The Data with Hadoop Components

•Hadoop is a powerful open-source software framework that provides businesses with a flexible and scalable platform for analyzing big data.

•Its core components, HDFS, YARN, and MapReduce, work together to provide a powerful platform for storing, managing, and processing large datasets. With additional tools like Hive and Pig, businesses can customize Hadoop to meet their specific needs.

•While there are some challenges associated with using Hadoop, its benefits, including scalability and real-time processing, make it an attractive option for businesses of all sizes.

References

- ❑ Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 2012

https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

https://data-flair.training/blogs/hadoop-hdfs-tutorial/

https://www.simplilearn.com/tutorials/hadoop-tutorial/hdfs (video)

https://www.techbooktutorial.com/hadoop/hadoop-distributed-file-system

https://www.techbooktutorial.com/hadoop/analyzing-the-data-with-hadoop-components