Name: Tejas Redkar
Roll No: PC-44
PRN: 1032210937
Panel-C
BDT-2 Batch-1 (22)

BDT Lab Assignment -5

* Problem Statement:

Perform Data Analysis using Map-Reduce in Hadoop/ Pyspark.

* Objectives:

1) To learn concepts of Map-Reduce
2) To learn how to analysis in Hadoop

* Theory:

1) Map-Reduce:

It is a programming model & data processing technique used for processing large volumes of data in a distributed & parallel manner. It was introduced by Google & is widely used in the Hadoop ecosystem. The fundamental idea behind Map-Reduce is to split a task into smaller sub-tasks, process them in parallel, & then combine the results.

In MapReduce, data processing is divided into two phases:

1) Map Phase: The input data is divided into chunks & processed in parallel by a set of map tasks.
2) Reduce Phase: Intermediate key-value pairs are sorted & grouped by key, & then reduce tasks apply a user defined function to aggregate & generates output.

2) Working of MapReduce:

Suppose we have a large text document that you want to analyze to count the frequency of each word. We can use Map-Reduce as follows:

- Map Phase:
  - Each mapper takes a portion of the document & tokenizes it into words.
  - For each word, the mapper emits a key-value pair where the word is the key, & the value is 1.
  - For eg. "apple" → (apple, 1), "banana" → (banana, 1), etc.
- Shuffle & Sort:
  - The framework sorts & groups the intermediate key value pairs key by key.
- Reduce Phase:

- Each Reducer receives a group of key value pairs with the same word as the key.
- The reducer sums up the values for each key, which gives the word count
- The final output will be a list of words & their respective counts.

* Platform : 64-bit Open Source Linux / Windows.

* Conclusion : Hence, I learned ~~to~~ the ~~masterpiece~~ mapreduce concept applying ~~&~~ on dataset in Hadoop environment.

* FAQ's

Q1) Explain the DFS, YARN services in Hadoop.

Ans DFS : Distributed File System is the primary storage system in Hadoop. It is designed to store large files & distributed them across multiple nodes in a Hadoop cluster. DFS is fault-tolerant, meaning it can recover from node failures. It is the foundation for storing & ~~manging~~ managing data in Hadoop.

YARN (Yet Another Resource Negotiator) : It is a resource management layer in Hadoop that is responsible for managing & allocating resources to applications running in a Hadoop cluster.

YARN seperates the resource management & job scheduling functions, making Hadoop more flexible & efficient in resource allocation.

Q2) What are the advantages of using MapReduce with Hadoop?

Ans Scalability : Easily scale to process large datasets by adding more hardware.

Fault tolerance : Automatic data replication & task recovery ensure reliability.

Parallel processing : Enables efficient parallel data processing for big data analytics.

Cost effective

Flexibility : Handles diverse data types & is versatile for various applications.

Q3) What is shuffling & sorting in MapReduce?

Ans Shuffling & sorting in MapReduce refer to the process of organizing & rearranging intermediate key-value pairs before they are processed by reduce tasks. During the shuffle & sort phase:

- Intermediate key-value pairs generated by map tasks
- The grouped data is sorted by key.
- Data with the same key is brought together, ensuring that each ~~other~~ reducer receives all data associated with a specific key.

21/11/23