|  |  |
| --- | --- |
| Ex No: 10 | **Study on Hadoop and MapReduce** |
| Date: 01-06-2022 |

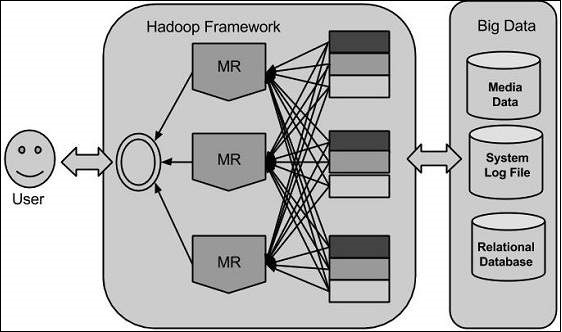
**Aim:**

To study on Hadoop and MapReduce.

**Procedure:**

**Hadoop:**

Hadoop is an open-source dsystem for storing and processing large amounts of data in a distributed setting using basic programming concepts across clusters of machines. It's built to expand from a single server to thousands of devices, each with its own computation and storage capabilities. Hadoop uses the MapReduce method to execute applications, which allows data to be processed in parallel with other data. In a nutshell, Hadoop is used to create systems that can run extensive statistical analyses on massive amounts of data.



**Hadoop Architecture:**

Hadoop is made up of two layers at its core:

* Processing/Computation layer (MapReduce), and
* Storage layer (Hadoop Distributed File System)



**MapReduce:**

MapReduce is a parallel programming methodology for creating distributed applications developed at Google for processing enormous volumes of data (multi-terabyte data sets) on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. Hadoop, an Apache open-source framework, is used to run the MapReduce algorithm.

**Hadoop Distributed File System:**

HDFS (Hadoop Distributed File System) is a distributed file system based on the Google File System (GFS) that is meant to run on commodity hardware. It has a lot in common with other distributed file systems. However, there are considerable distinctions between it and other distributed file systems. It's meant to run on low-cost hardware and is highly fault-tolerant. It allows high-throughput application data access and is well-suited to applications with massive datasets.

Aside from the two fundamental components described above, the Hadoop framework additionally includes the following two modules:

* Hadoop Common are Java libraries and utilities that other Hadoop modules rely on.
* Hadoop YARN is a job scheduling and cluster resource management system.

**How Does Hadoop Work:**

Larger servers with heavy configurations for large-scale processing are quite expensive, but you can tie together many commodity computers with single-CPU as a single functional distributed system, and the clustered machines can read the dataset in parallel and provide significantly higher throughput. Furthermore, it is less expensive than purchasing a single high-end server. The fact that Hadoop runs on clustered and low-cost servers is the initial motivating point for employing it.

Hadoop is a distributed computing system that runs code across a group of computers. Hadoop conducts the following fundamental actions as part of this process:

* Initially, data is organized into folders and files. The files are separated into 128M and 64M blocks of uniform size (preferably 128M).
* These files are subsequently transferred to other cluster nodes for processing.
* The processing is overseen by HDFS, which sits on top of the local file system.
* To handle hardware failure, blocks are copied.
* Verifying that the code was successfully executed.
* Executing the sort that occurs between the map and reduce steps.
* Sending the data that has been sorted to a certain computer.
* For each job, create debugging logs.

**Advantages:**

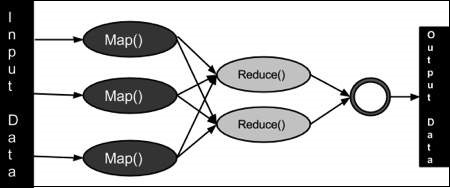
* The Hadoop framework enables users to write and test distributed systems quickly. It's efficient because it distributes data and work across computers automatically, taking advantage of the CPU cores' inherent parallelism.
* Hadoop's fault-tolerance and high availability (FTHA) is not based on hardware; rather, the Hadoop library has been intended to detect and handle faults at the application layer.
* Hadoop can be dynamically added or removed from the cluster, and it will continue to run without interruption.
* Another significant benefit of Hadoop is that, in addition to being open source, it is platform agnostic due to its Java foundation.

**MapReduce:**

MapReduce is a distributed computing processing technology and program architecture based on Java. Map and reduce are two fundamental tasks in the MapReduce algorithm. Map turns a set of data into another set of data by breaking down individual pieces into tuples (key/value pairs). Second, there's the reduction job, which takes the result of a map as an input and condenses the data tuples into a smaller set. The reduction work is always executed after the map job, as the name MapReduce suggests. MapReduce's main advantage is that it's simple to expand data processing over several computing nodes. The data processing primitives of the MapReduce model are known as mappers and reducers. It's not always easy to break down a data processing application into mappers and reducers. Scaling an application to run over hundreds, thousands, or even tens of thousands of servers in a cluster is only a configuration modification once we build it in MapReduce form. Many programmers have been drawn to the MapReduce approach because of its simple scalability.

**MapReduce Algorithm:**

* The MapReduce paradigm is built on sending the computer to the location of the data!
* The MapReduce program is broken down into three stages: map, shuffle, and reduce.
  + The map or mapper's job is to process the input data in the map stage. In most cases, the input data is stored in the Hadoop file system as a file or directory (HDFS). Line by line, the input file is supplied to the mapper function. The mapper divides the data into little bits and analyses it.
  + This stage is a mixture of the Shuffle and Reduce stages. The Reducer's job is to take the data from the mapper and process it. It generates a new set of outputs after processing, which will be stored in the HDFS.
* Hadoop assigns the Map and Reduce tasks to the relevant computers in the cluster during a MapReduce job.
* The framework handles all of the data-passing aspects, such as issuing tasks, checking task completion, and moving data between nodes in the cluster.
* The majority of computing is done on nodes, with data stored on local storage, which decreases network traffic.
* The cluster collects and reduces the data to generate an acceptable result, then delivers it back to the Hadoop server after the tasks are completed.



**Hadoops Importance to users:**

Despite the rise of other solutions, particularly in the cloud, Hadoop remains a critical and beneficial technology for big data consumers for the following reasons:

* It has the ability to quickly store and handle large amounts of structured, semi structured, and unstructured data.
* It guards against hardware faults in application and data processing. Processing jobs are automatically transferred to other nodes whenever one node in a cluster fails, ensuring that applications continue to execute.
* It does not necessitate the preprocessing of data prior to storage. Organizations can keep unprocessed data in HDFS and determine how to process and filter it later for specific analytics purposes.
* It's scalable, so businesses can quickly add more nodes to increase the amount of data their systems can process.
* It can handle batch workloads for historical analysis as well as real-time analytics to aid in better operational decision-making.

**Hadoop application and use cases:**

Interactive querying, stream processing, and real-time analytics are just a few of the new applications that YARN has added to Hadoop clusters. Manufacturers, utilities, oil and gas industries, and other enterprises, for example, use real-time data from IoT devices streaming into Hadoop systems in predictive maintenance applications to try to detect equipment issues before they happen. Other real-time use cases include fraud detection, website customization, and customer experience scoring.

The following are some other common Hadoop use cases:

* Customer data is analyzed. Predicting customer turnover, analyzing clickstream data to better target online ads to web users, and tracking consumer sentiment based on social media comments are just a few examples.
* Risk management is the management of risks. Hadoop clusters are used by financial services businesses to generate more accurate risk analysis models for internal and external use. In Hadoop-based big data systems, they also design financial models and trading algorithms.
* Intelligence in the field of operations. For capacity planning and management, Hadoop can assist telecommunications firms better analyze switching performance, network and frequency utilisation, and so on. Telcos may also find the ideal spots to install new cell towers and respond more rapidly to network problems by studying how mobile services are utilized and the available bandwidth in geographic regions.
* Management of the supply chain. Hadoop systems are used by manufacturers, retailers, and trucking companies to track the movement of goods and trucks in order to calculate the costs of various transportation choices. They can also use past, time-stamped location data to map out potential delays and optimize delivery routes.

The technology has also been used for a variety of other purposes. Insurers, for example, employ Hadoop for applications like policy price analysis and maintaining safe driver discount programs. With Hadoop's help, healthcare institutions are also looking for methods to better therapies and patient outcomes.

**Result:**

Thus, study on Hadoop and MapReduce is completed successfully.