**Data Analytics Laboratory**

**Task 7**

**Hadoop File Management**

## Aim

Performing Linux and Hadoop file system commands to store, edit, retrieve and delete a file in Hadoop distributed file system.

## Introduction to Hadoop Distributed Filesystems

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is now an Apache Hadoop subproject.

## Prerequisites

**1. What are distributed Systems?**

A distributed system is a network that consists of autonomous computers that are connected using a distribution middleware. They help in sharing different resources and capabilities to provide users with a single and integrated coherent network.

**2. What is distributed Computing?**

Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance.

**3. Applications of distributed computing.**

The very nature of an application may require the use of a communication network that connects several computers: for example, data produced in one physical location and required in another location. B) A distributed system may be easier to expand and manage than a monolithic uniprocessor system

**4. What is Hadoop?**

Ans. Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

**5. Advantages of Distributed Computing?**

Ans. A distributed ledger is a database that is consensually shared and synchronized across multiple sites, institutions or geographies. Underlying the distributed ledger technology is the blockchain, which is the technology that underlies bitcoin.

## In-Lab Tasks

To complete this lab, you will need the following:

* Hadoop Environment installed preferably in Linux Subsystem.

|  |  |
| --- | --- |
| Step | Action |
| 1 | **Boot up Linux Subsystem**Make sure you choose Linux Subsystem while booting, instead of Windows. |
| 2 | **User Login**Login to with provided credentials |
| 3 | **Terminal**Open Terminal Window by searching in start-menu or can use the shortcut ‘ctrl+alt+t’ |
| 4 | **Login to hadoop**hadoop@vrsec:$ su hadoopEnter the password for logging in as hadoop user (password: 12345) |
| 5 | **Start the Hadoop components (daemons) on the Linux Subsystem.** You can practice starting all components with these commands. Please note they will take a few minutes to run:hadoop@vrsec:$ startCDH.sh |
| 6 | **Check the running hadoop services**hadoop@vrsec:$ jps |
| 7 | **Exploring Hadoop Distributed File System (HDFS)**Command-line approachhdfs dfs <args> |
| 8 | **To recursively list all the files in each directory –lsr command is used instead of –ls**hdfs dfs -ls -R /**To view the contents of a file**hdfs dfs fs -cat <file path> |
| 9 | **Command to create the directory**hdfs dfs -mkdir <Foldername>**To copy a file in HDFS**hdfs dfs –put <path of src file/direc> <path of dest file>**To copy a directory in HDFS**hdfs dfs –put <path of src directory> <path of dest directory> |
| 10 | **To remove a file in HDFS**hdfs dfs –rm <path of file> |
| 11 | **To move files between your regular Linux filesystem and HDFS you can use the put and get commands.**hdfs dfs –moveFromLocal <local system file path of file> <hdfs directory path>hadoop fs –moveToLocal <hdfs file path> <local system directory path>**To change the replication factor of the file in HDFS**hdfs fs –setrep <number> <path to file> |
| 12 | **To Stop the Hadoop Services**hadoop@vrsec: $ stopCDH.sh |

## Post-Lab Task

**1. What is a cluster?**

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. Typically, one machine in the cluster is designated as the NameNode and another machine the as JobTracker these are the masters.

**2. Types of cluster?**

There are 3 types of clusters

1. Single Node Cluster
2. Pseudo Distributed Cluster
3. Multi-node Cluster

**3. Difference between parallel and distributed computing?**

One of the major differences between Parallel and Distributed Computing is the underlying architecture of memory sharing. In Parallel Computing, all the different "processor" has the access to a shared memory. In Distributed Computing, all the different processors have their own private non-shareable memory

**4. Use cases of cluster.**

Some classes of use cases of clustering: As a pre-processing step for supervised learning - for example, clustering the data and using the cluster id as a feature to predict the label

As a way to detect patterns - for example, partitioning customers into categories, so we can recommend different products to each group.

As a way to reduce data size - for example, if you are trying to detect and classify different objects in a 3d point cloud, it may be useful to focus on geometrically close clusters of points, rather than dealing with the entire dataset.

**5. Advantages and disadvantages of a cluster.**

Advantages of Cluster:

1. Increased resource availability: If one Intelligence Server in a cluster fails, the other Intelligence Servers in the cluster can pick up the workload. This prevents the loss of valuable time and information if a server fails.
2. Strategic resource usage: You can distribute projects across nodes in whatever configuration you prefer. This reduces overhead because not all machines need to be running all projects and allows you to use your resources flexibly.
3. Increased performance: Multiple machines provide greater processing power.
4. Greater scalability: As your user base grows and report complexity increases, your resources can grow.
5. Simplified management: Clustering simplifies the management of large or rapidly growing systems.

Disadvantages of Cluster:

1. Cost is high. Since the cluster needs good hardware and a design, it will be costly comparing to a non-clustered server management design. Being not cost effective is a main disadvantage of this design
2. Since clustering needs more servers and hardware to establish one, monitoring and maintenance is hard.

**Results**

The program is implemented in python and the output is observed.

**Faculty Signature**