S. B. JAIN INSTITUTE OF TECHNOLOGY, MANAGEMENT & RESEARCH, NAGPUR.

Practical No. 9

Aim: To implement the Cloudera Data Platform (CDP) and analyze the working of Cloudera Manager in a distributed environment & also use basic hadoop command using cloudera terminal.

Name of Student: Shrutika Pradeep Bagdi

Roll No.: CS22130

Semester/Year: 7th / 4th

Academic Session: 2024-2025

Date of Performance: _____

Date of Submission: _____

AIM: To implement the Cloudera Data Platform (CDP) and analyze the working of Cloudera Manager in a distributed environment & also use basic hadoop command using cloudera terminal.

OBJECTIVE/EXPECTED LEARNING OUTCOME:

The objectives and expected learning outcome of this practical are:
Understand Cloudera Data Platform (CDP): Gain insight into the architecture and components of the Cloudera Data Platform.
Deploy and Configure CDP Services: Install and configure Cloudera Manager and other Hadoop components (HDFS, YARN, MapReduce, Hive) in a multi-node setup.

HARDWARE AND SOFTWARE REQUIRMENTS:

Hardware Requirement:
64-bit operating system (Linux, CentOS, or Ubuntu)
At least 8 GB of RAM for each node and sufficient storage

Software Requirement: Cloudera Data Platform (CDP)
VirtualBox-7.0.20-163906-Win
Cloudera Manager (for deployment and management)
Hadoop components (HDFS, YARN, MapReduce, HBase, Hive, etc.)

THEORY:

Introduction: Cloudera is a leading provider of a cloud-native data management and analytics platform that delivers machine learning, analytics, and operational services. It specializes in enterprise data management solutions based on Apache Hadoop and other open-source technologies. Cloudera empowers organizations to manage, process, and analyze large-scale data (Big Data) in distributed environments.

Key Components of Cloudera

1. **Cloudera Data Platform (CDP):** Cloudera Data Platform (CDP) is a comprehensive platform that allows organizations to manage data from on-premises to multi-cloud environments. It provides a unified and secure platform for data analytics, enabling modern data-driven operations.
2. **Cloudera Distribution for Hadoop (CDH):** CDH is Cloudera's open-source distribution of Apache Hadoop, bundled with enterprise features like security, management, and high availability. CDH includes core Hadoop components (HDFS, MapReduce, YARN) and additional tools such as Apache Hive, HBase, Impala, and Spark.
3. **Cloudera Manager:** Cloudera Manager is the central management tool provided by Cloudera for deploying, configuring, managing, and monitoring clusters. It allows for streamlined Hadoop ecosystem deployment and cluster lifecycle management through a web-based interface.
4. **Cloudera Navigator:** Cloudera Navigator is a part of the platform that enables data governance, compliance, and security. It provides features like auditing, lineage tracking, and encryption to ensure that enterprise data is well-governed.
5. **Cloudera Data Engineering (CDE):** CDE is a modern platform for building, running, and managing data pipelines with the use of Apache Spark in the cloud or on-premises. It integrates with existing CDP environments and simplifies the process of creating scalable data workflows.

6. **Cloudera Data Warehouse (CDW):** CDW is a service that enables data warehousing in a cloud-native architecture. It leverages the power of SQL-based engines like Apache Hive and Impala to process structured and semi-structured data at a large scale.
7. **Cloudera Machine Learning (CML):** CML is Cloudera's machine learning platform that enables building, training, and deploying machine learning models at scale. It supports the entire ML lifecycle, from data preparation to model training and inference, in a secure and governed environment.
8. **Cloudera DataFlow (CDF):** CDF is a tool used for collecting, curating, and analyzing real-time data streams. It is powered by Apache NiFi and provides a way to process large volumes of streaming data from different sources in real time.

Features of Cloudera

1. **Scalability:** Cloudera's platform is designed to handle petabytes of data across distributed clusters. It provides seamless scalability to meet growing data demands.
2. **Unified Data Management:** Cloudera enables management of structured, semi-structured, and unstructured data across the entire data lifecycle. This makes it suitable for diverse data types and sources.
3. **Data Security and Governance:** Cloudera's platform has built-in security and governance tools such as Cloudera Navigator for ensuring compliance, managing data privacy, and tracking data lineage.
4. **Enterprise-Grade Management:** Cloudera Manager simplifies the deployment and management of Hadoop clusters with features like automated configuration, performance monitoring, and service management.
5. **Support for Hybrid and Multi-Cloud Deployments:** Cloudera offers flexibility for hybrid and multi-cloud environments, enabling organizations to deploy and manage workloads across on-premises, public, and private clouds.

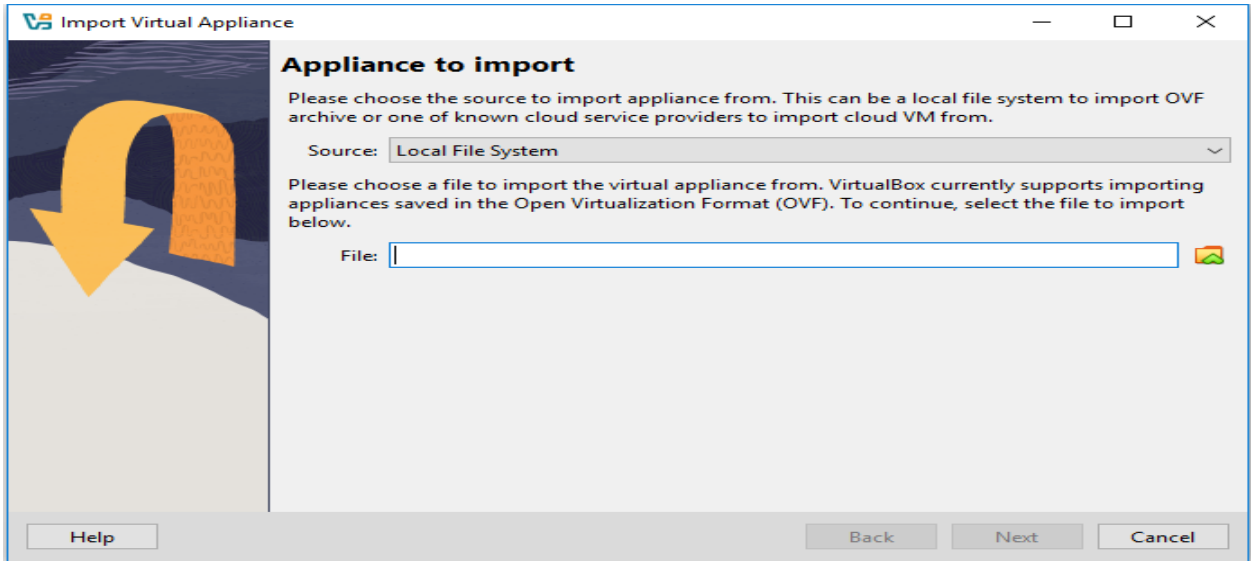
INPUT / OUTPUT (SCREENSHOTS):

Steps for Installing VirtualBox on Windows

Step 1: Download VirtualBox Installer

Step 2: Run the VirtualBox Installer

Step 3: Begin the Installation

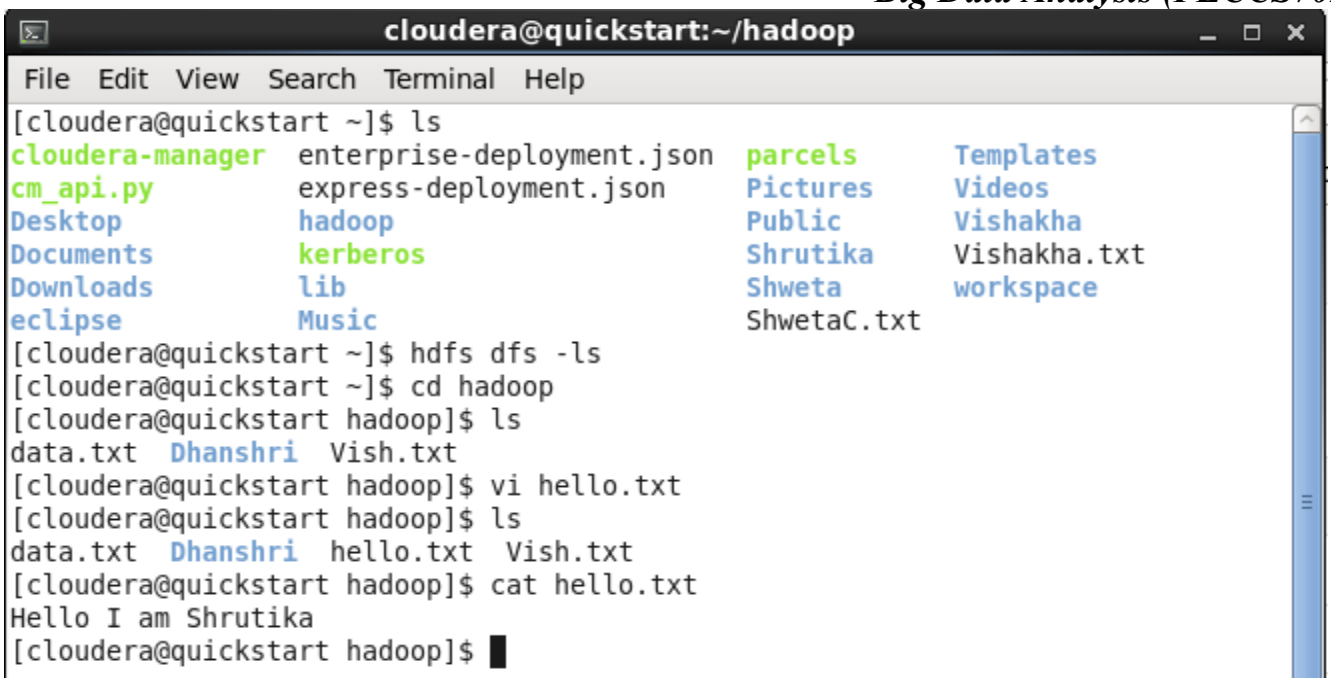


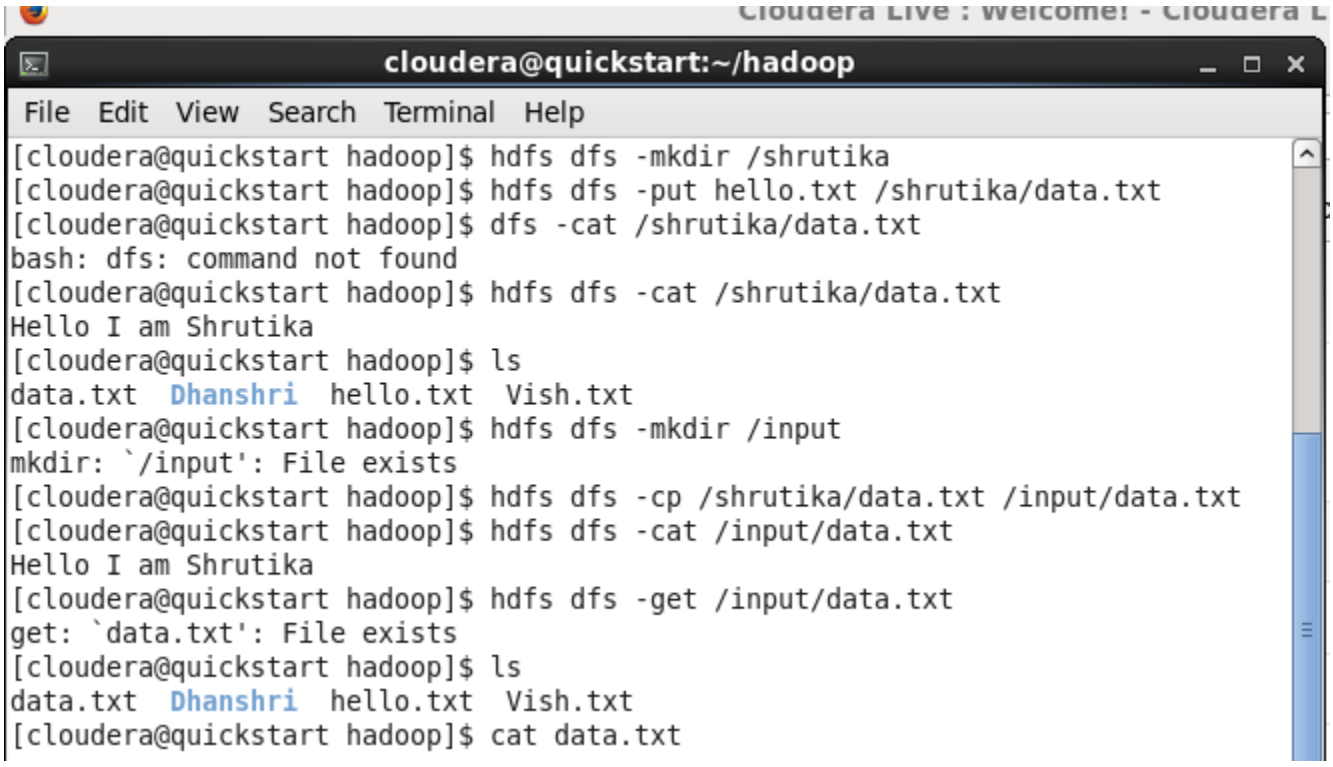
Step 4: Select Installation Options

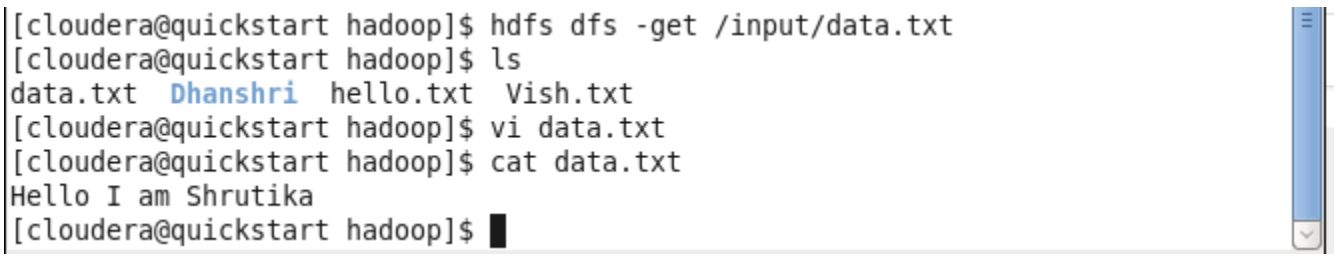
Step 5: Network Interface Warning

Step 6: Install VirtualBox

Step 7: Finish the Installation (installed cloudera quickstart vm in oracle virtualbox)


cloudera@quickstart:~/hadoop
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ ls
cloudera-manager enterprise-deployment.json parcels Templates
cm_api.py express-deployment.json Pictures Videos
Desktop hadoop Public Vishakha
Documents kerberos Shrutika Vishakha.txt
Downloads lib Shweta workspace
eclipse Music ShwetaC.txt
[cloudera@quickstart ~]$ hdfs dfs -ls
[cloudera@quickstart ~]$ cd hadoop
[cloudera@quickstart hadoop]$ ls
data.txt Dhanshri Vish.txt
[cloudera@quickstart hadoop]$ vi hello.txt
[cloudera@quickstart hadoop]$ ls
data.txt Dhanshri hello.txt Vish.txt
[cloudera@quickstart hadoop]$ cat hello.txt
Hello I am Shrutika
[cloudera@quickstart hadoop]$


cloudera@quickstart:~/hadoop
File Edit View Search Terminal Help
[cloudera@quickstart hadoop]$ hdfs dfs -mkdir /shrutika
[cloudera@quickstart hadoop]$ hdfs dfs -put hello.txt /shrutika/data.txt
[cloudera@quickstart hadoop]$ dfs -cat /shrutika/data.txt
bash: dfs: command not found
[cloudera@quickstart hadoop]$ hdfs dfs -cat /shrutika/data.txt
Hello I am Shrutika
[cloudera@quickstart hadoop]$ ls
data.txt Dhanshri hello.txt Vish.txt
[cloudera@quickstart hadoop]$ hdfs dfs -mkdir /input
mkdir: `/input': File exists
[cloudera@quickstart hadoop]$ hdfs dfs -cp /shrutika/data.txt /input/data.txt
[cloudera@quickstart hadoop]$ hdfs dfs -cat /input/data.txt
Hello I am Shrutika
[cloudera@quickstart hadoop]$ hdfs dfs -get /input/data.txt
get: `data.txt': File exists
[cloudera@quickstart hadoop]$ ls
data.txt Dhanshri hello.txt Vish.txt
[cloudera@quickstart hadoop]$ cat data.txt


[cloudera@quickstart hadoop]$ hdfs dfs -get /input/data.txt
[cloudera@quickstart hadoop]$ ls
data.txt Dhanshri hello.txt Vish.txt
[cloudera@quickstart hadoop]$ vi data.txt
[cloudera@quickstart hadoop]$ cat data.txt
Hello I am Shrutika
[cloudera@quickstart hadoop]$

CONCLUSION:

In this practical, we successfully implemented the Cloudera Data Platform (CDP) in a distributed environment and analyzed the functionality of Cloudera Manager for efficient cluster management. We deployed, configured, and monitored Hadoop services using Cloudera Manager, gaining insights into resource allocation and performance tracking. Additionally, we explored basic Hadoop commands via the Cloudera terminal, performing essential tasks such as file system navigation and job execution, which reinforced our understanding of Hadoop's distributed file system and job management capabilities. This hands-on experience provided a comprehensive view of Big Data management using Cloudera's ecosystem.

DISCUSSION AND VIVA VOCE:

1. What are the key components of Cloudera Data Platform (CDP), and how do they interact to support Big Data workloads in a distributed environment?
2. How does Cloudera Manager provide real-time monitoring and performance insights for Hadoop clusters?
3. What are the basic Hadoop commands used for file operations in HDFS, and how can they be executed in the Cloudera terminal?
4. Explain the process of adding a new node to an existing Cloudera Hadoop cluster. How does Cloudera Manager simplify this process?
5. How does Cloudera Manager facilitate the backup and recovery of data in a Hadoop cluster?

REFERENCE:

- https://www.cloudera.com/
- https://www.youtube.com/watch?v=jT1q5YQ2cpw.
- https://education.cloudera.com/store/3086306-introducing-cloudera-data-analyst-training

Observation book: (3)	Viva-Voce (3)	Quality of Submission and timely Evaluation (4)
Total:		**Sign with date:**