|  |  |  |
| --- | --- | --- |
| **Name : Shreya Singh** | **Class/Roll No. : 55/D16AD** | **Grade :** |

**Title of Experiment :** To study Hadoop Ecosystem and to demonstrate Basic Hadoop Commands.

**Objective of Experiment :**

Acquire a foundational understanding of the Hadoop ecosystem and its components, focusing on basic Hadoop commands for effective data management and processing.

**Outcome of Experiment :**

We successfully installed Hadoop Eco-system and executed

basic Hadoop commands on it.

**Problem Statement :**

Learn how to use Hadoop for managing and processing big data

by mastering essential commands and concepts

**Description / Theory :**

Hadoop:

Hadoop is an open-source framework designed for processing and managing large

volumes of data across distributed computing clusters. It was initially developed by the Apache Software Foundation and is widely used in the field of big data analytics.

Purpose and Use:

Hadoop is used to process and store massive amounts of data in a scalable and reliable manner. It is particularly well-suited for handling unstructured or semi-structured data,

such as text, images, videos, and log files. Hadoop is employed for tasks like batch

processing, data warehousing, data transformation, and more. It's commonly used in

industries like finance, healthcare, e-commerce, and social media to extract valuable

insights from vast datasets.

Features:

Distributed Storage: Hadoop's HDFS (Hadoop Distributed File System) breaks data

into blocks and stores copies across multiple nodes in a cluster, ensuring redundancy

and fault tolerance.

Distributed Processing: The MapReduce programming model allows data

processing tasks to be divided into smaller tasks that are distributed across cluster

nodes, enabling parallel processing and efficient resource utilization.

Scalability: Hadoop can easily scale by adding more nodes to the cluster, allowing it

to handle growing data volumes without significant changes to the architecture.

Fault Tolerance: Hadoop maintains data redundancy, so even if a node or hardware

fails, the system can still function without data loss. It automatically replicates data

across nodes.

Flexibility: Hadoop can process structured, semi-structured, and unstructured data.

This versatility makes it suitable for various data types.

Ecosystem: Hadoop has a rich ecosystem of tools and frameworks, including Hive for

querying data using a SQL-like language, Pig for scripting data processing, and Spark

for in-memory processing, which enhances its capabilities.

Cost-Effectiveness: Hadoop can run on commodity hardware, making it more costeffective than traditional data processing systems that require specialized hardware.

Open Source: Being open-source, Hadoop is freely available for use, modification,

and distribution, which has contributed to its widespread adoption.

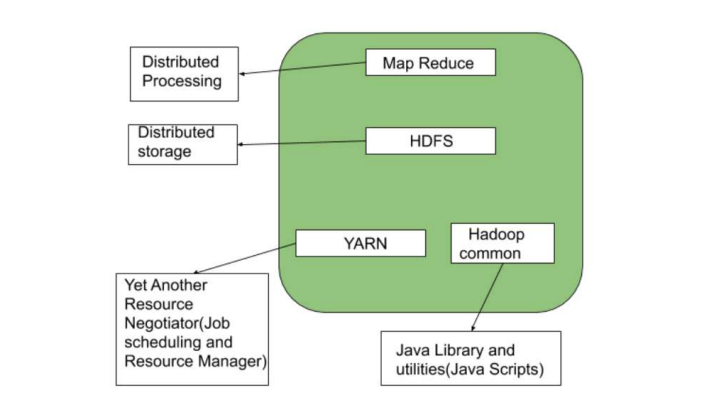
Data Locality: Hadoop's data processing paradigm strives to process data on the nodes

where it resides. This reduces network traffic and enhances performance.

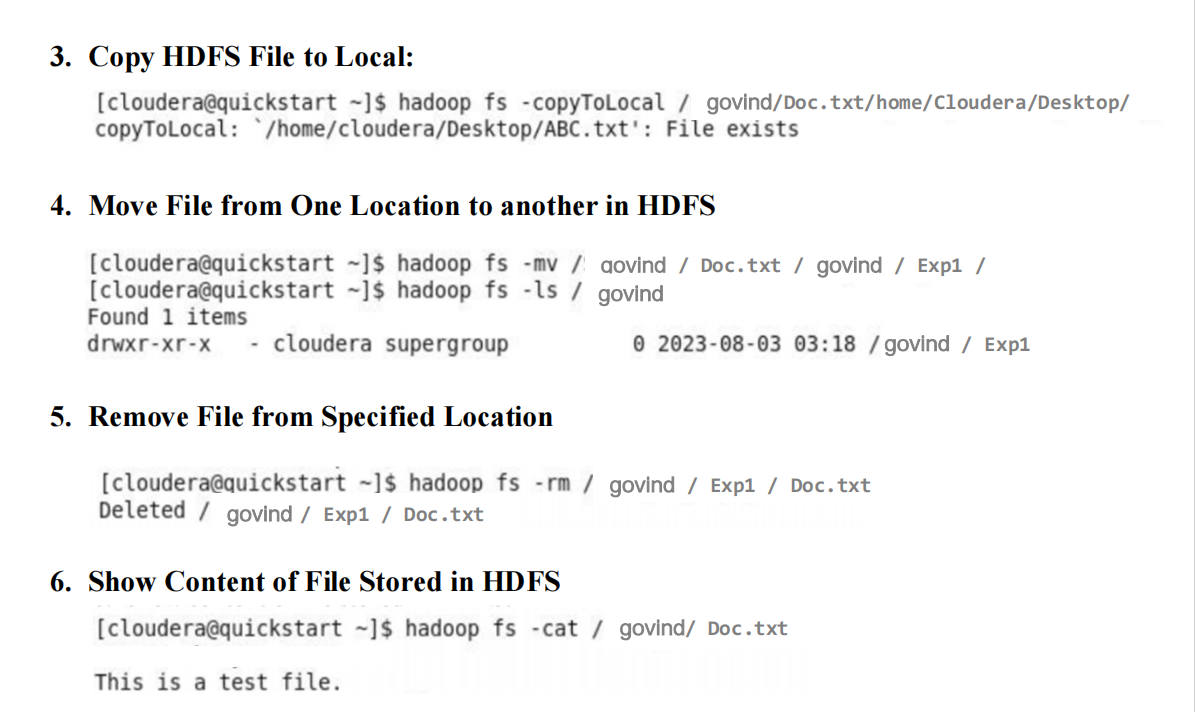
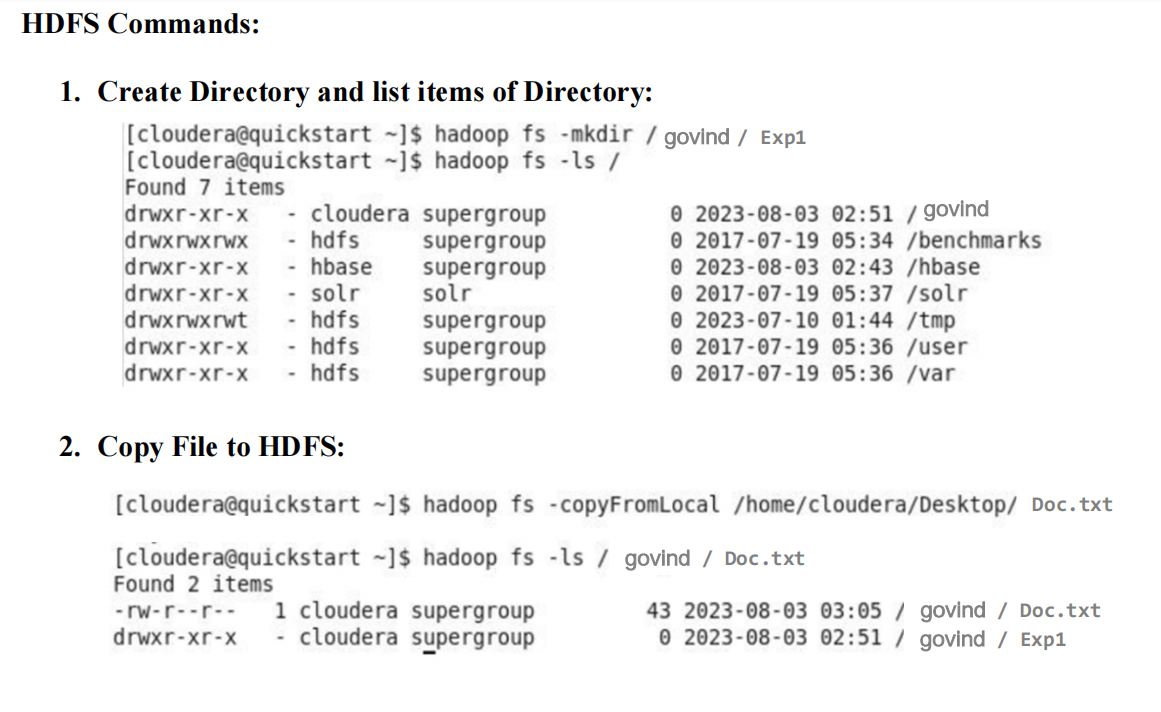
Highly Parallel Processing: Hadoop's MapReduce model and its successors enable

distributed data processing, facilitating efficient parallelism to speed up computation.

**Hadoop Architecture:**



**Output**:



**Results and Discussions :**

We learnt about Hadoop’s important features and structure. We practiced basic Hadoop

commands, like making folders, moving files, and checking data. These actions showed us how

Hadoop can handle big data tasks effectively.