**Exp No: 9** **Date:**

# HADOOP

## SET UP A SINGLE HADOOP CLUSTER AND SHOW THE PROCESS USING WEB UI

**AIM:**

To set-up one node Hadoop cluster.

**PROCEDURE:**

1. System Update
2. Install Java
3. Add a dedicated Hadoop user
4. Install SSH and setup SSH certificates
5. Check if SSH works
6. Install Hadoop
7. Modify Hadoop config files
8. Format Hadoop filesystem
9. Start Hadoop
10. Check Hadoop through web UI
11. Stop Hadoop

**THEORY**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from a single server to thousands of machines, each offering local computation and storage.

**HADOOP ARCHITECTURE**

Hadoop framework includes following four modules:

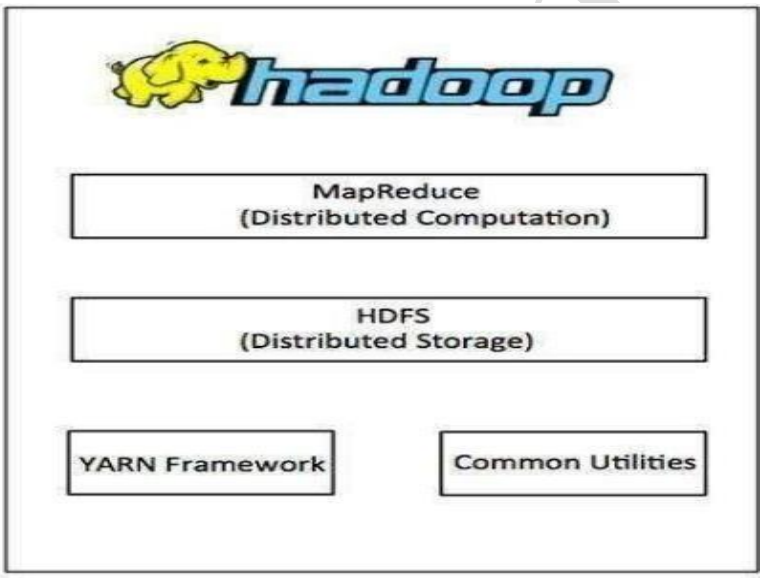
Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contain the necessary Java files and scripts required to start Hadoop.

Hadoop YARN: This is a framework for job scheduling and cluster resource management.

Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.

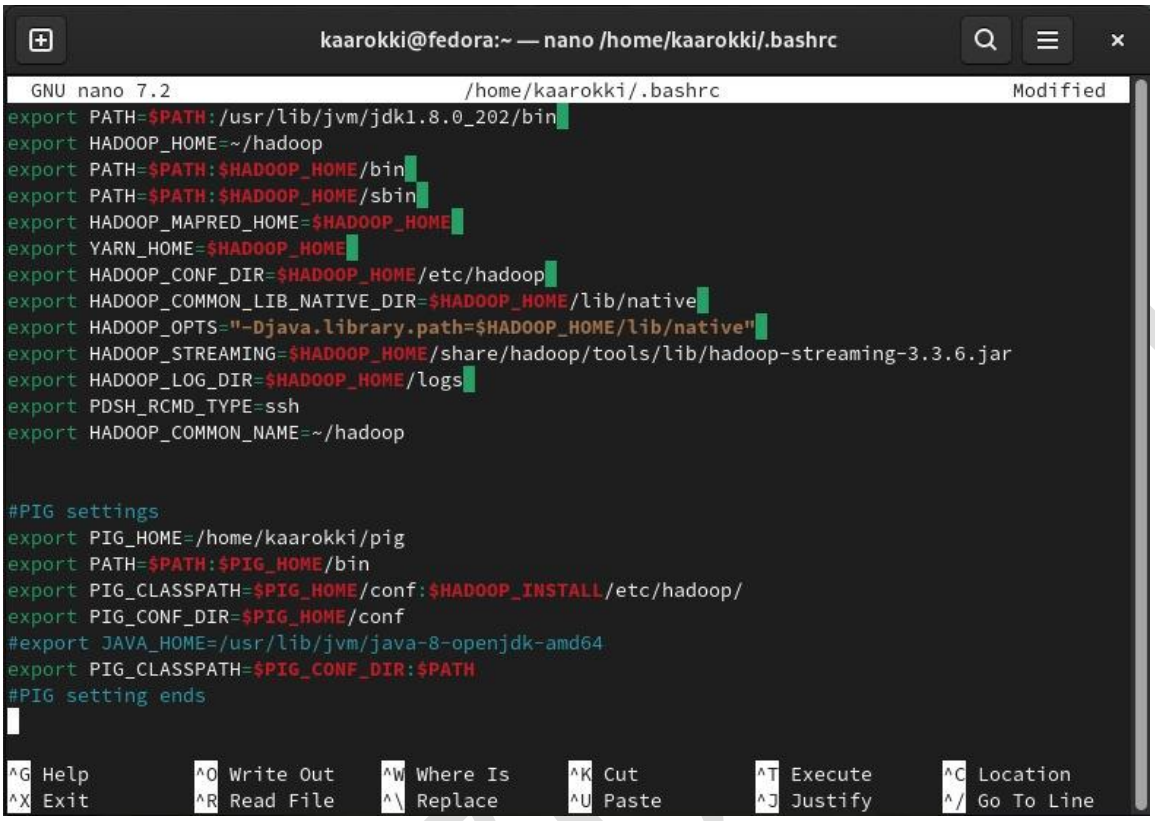
Hadoop MapReduce: This is a YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.

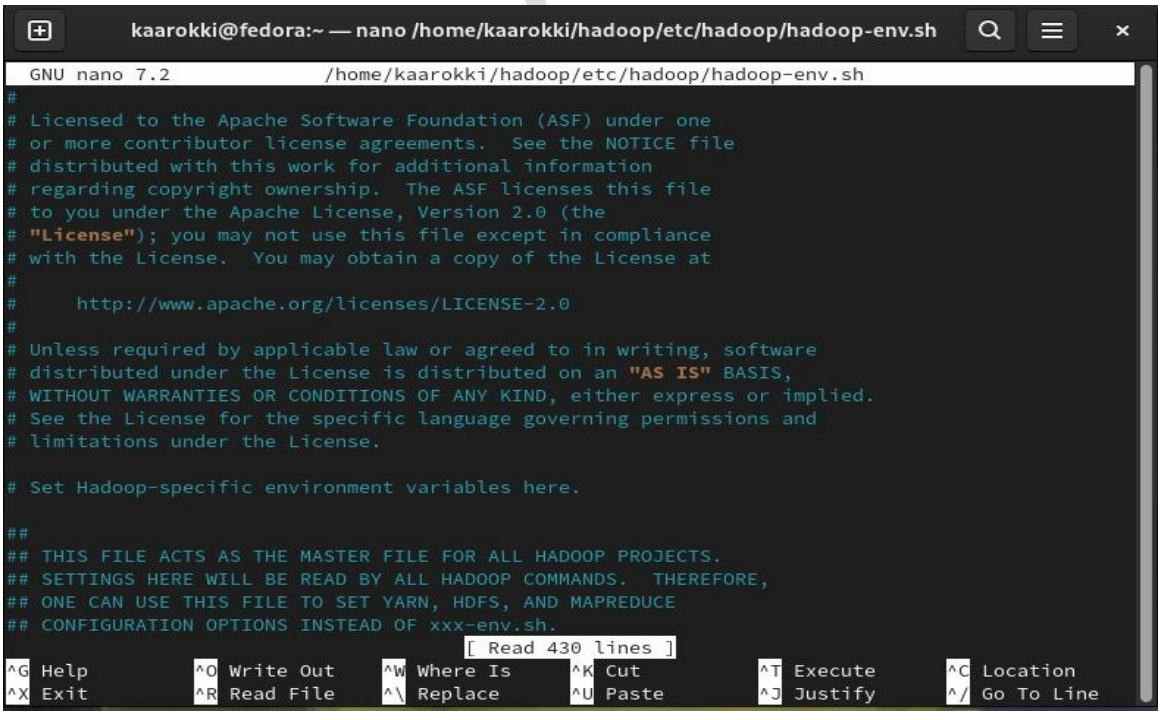


**PROCEDURE**

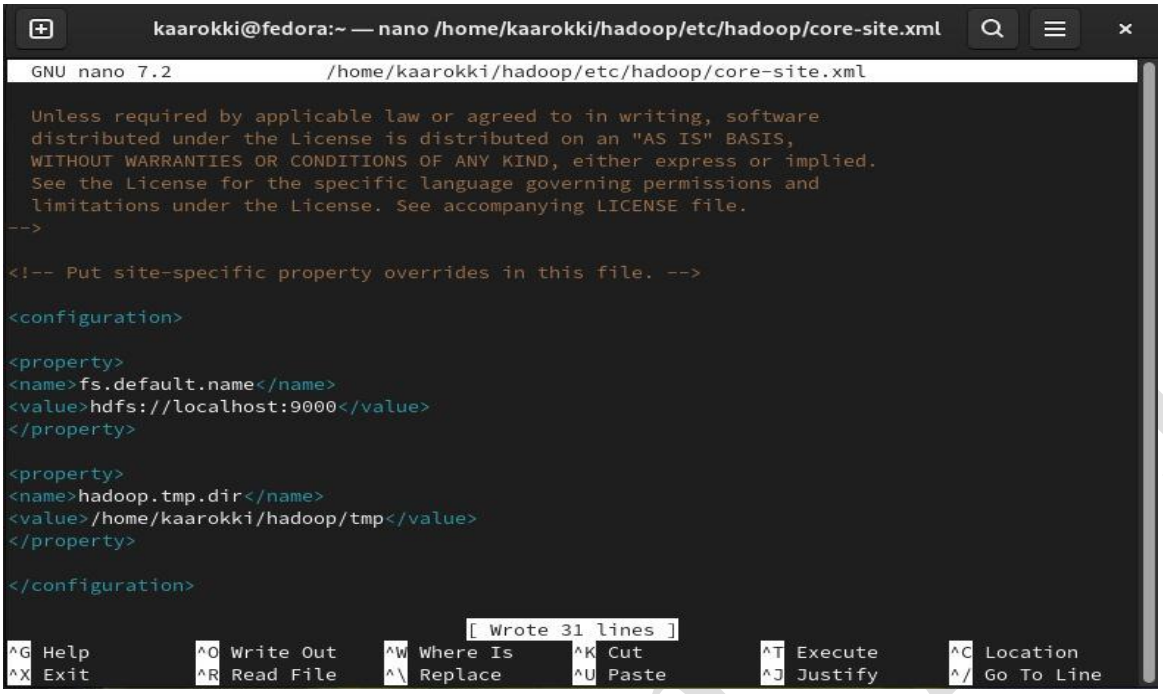
**$ nano ~/.bashrc**



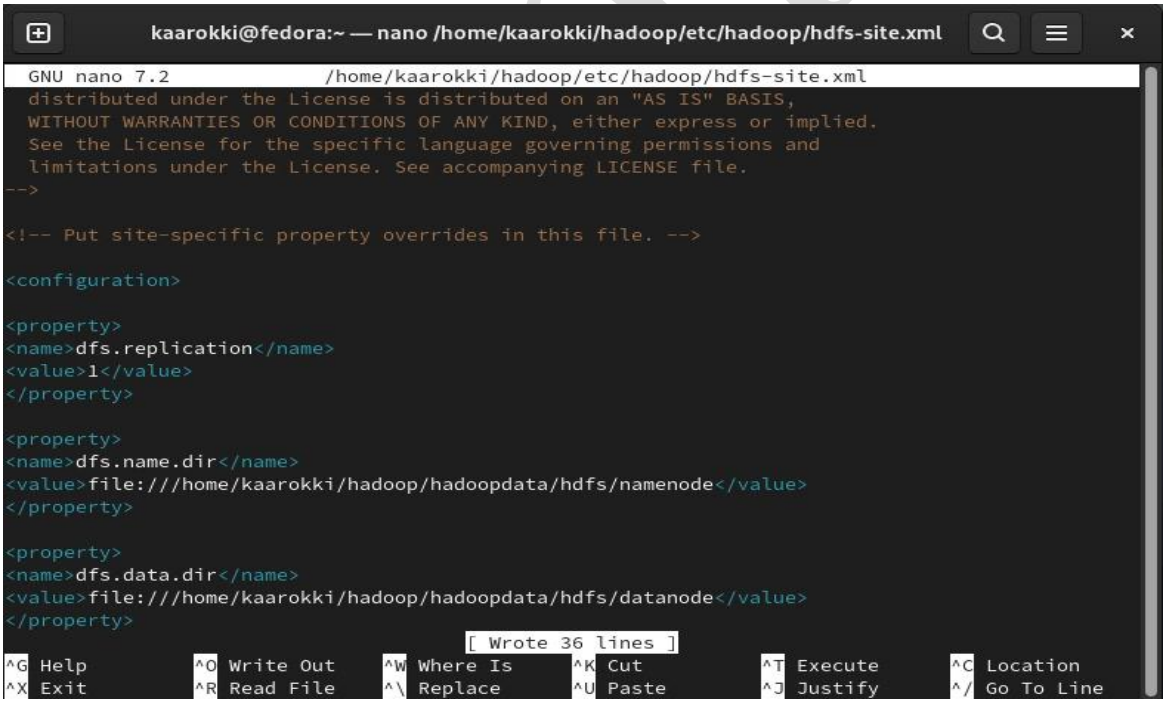

**$ nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh**



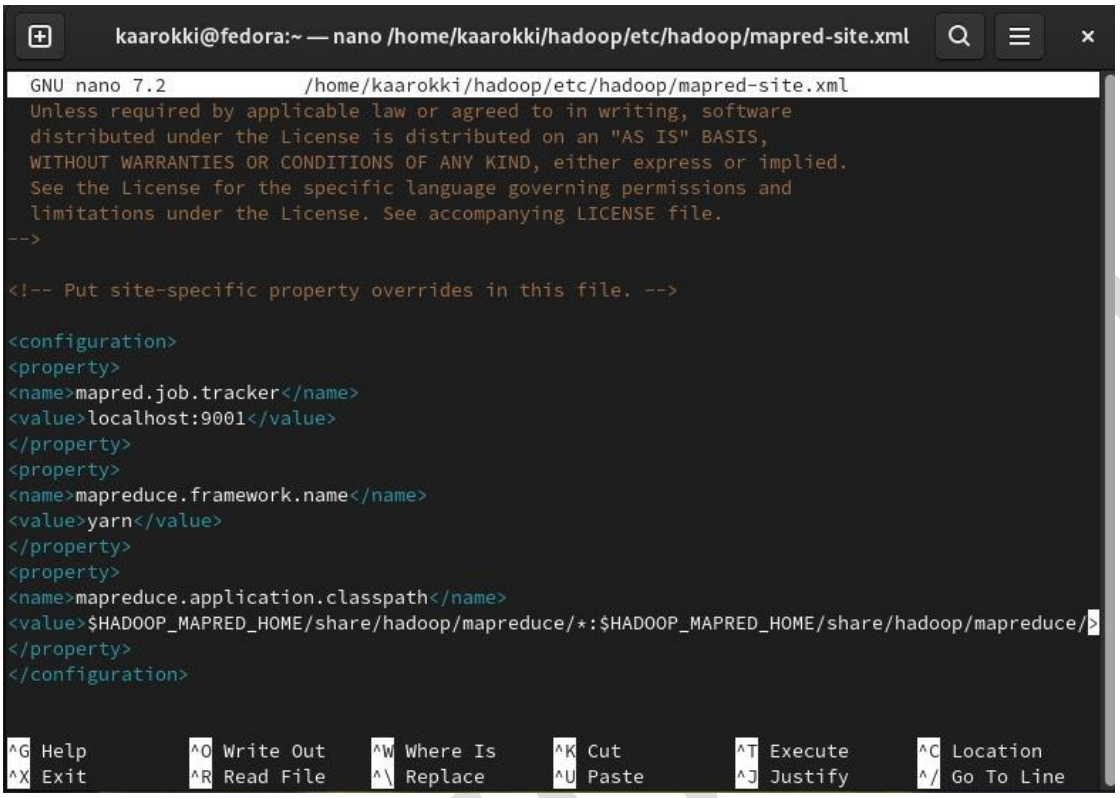
**$nano $HADOOP_HOME/etc/hadoop/core-site.xml**



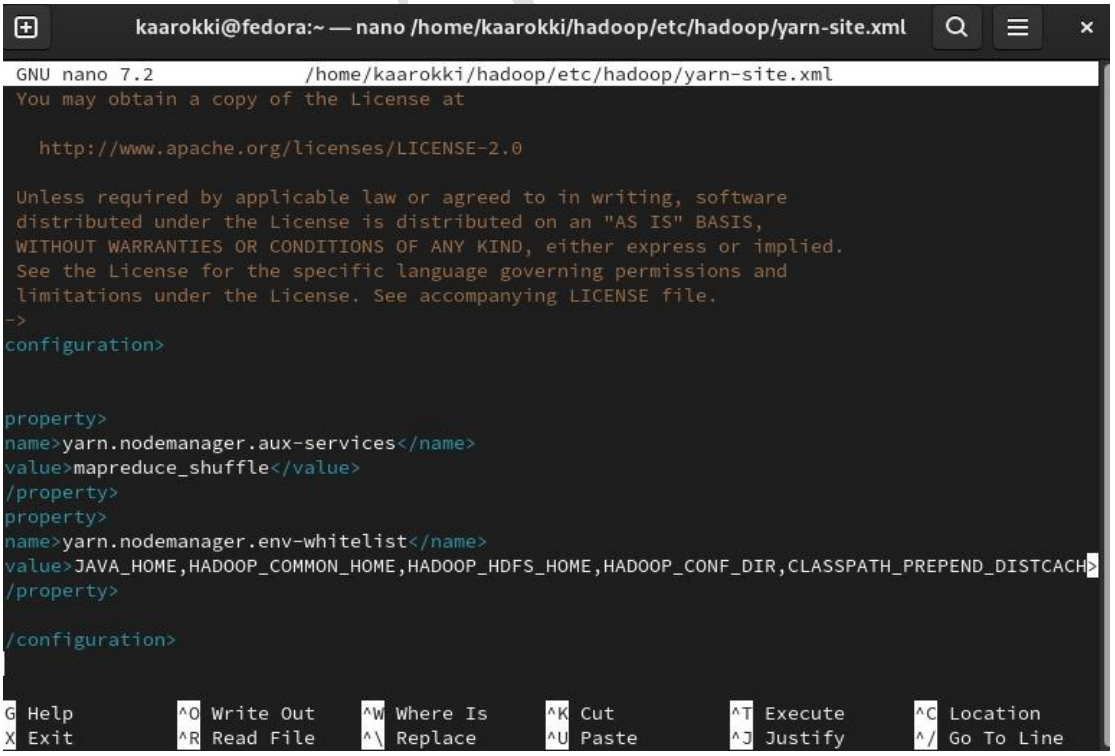

**$nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml**



**$nano $HADOOP_HOME/etc/hadoop/mapred-site.xml**




**$nano $HADOOP_HOME/etc/hadoop/yarn-site.xml**



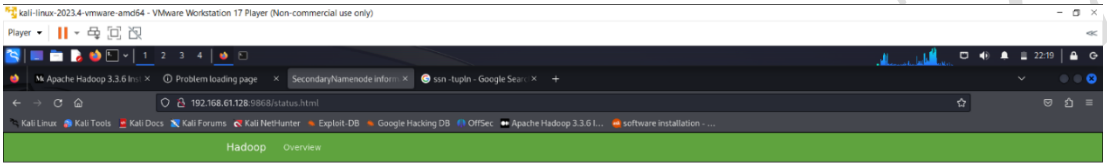
**$ start-all.sh**

```
kaarokki@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kaarokki in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: ssh: connect to host localhost port 22: Connection refused
Starting datanodes
localhost: ssh: connect to host localhost port 22: Connection refused
Starting secondary namenodes [fedora]
fedora: ssh: connect to host fedora port 22: Connection refused
Starting resourcemanager
Starting nodemanagers
localhost: ssh: connect to host localhost port 22: Connection refused
```

**$ jps**

```
kaarokki@fedora:~$ jps
3456 ResourceManager
3750 Jps
```

**localhost:9870**




## Overview

| | |
|---|---|
| **Version** | 3.3.6 |
| **Compiled** | 2023-06-18T08:22Z by ubuntu from (HEAD detached at release-3.3.6-RC1) |
| **NameNode Address** | localhost:9000 |
| **Started** | Wed Aug 14 21:51:32 -0400 2024 |
| **Last Checkpoint** | Never |
| **Checkpoint Period** | 3600 seconds |
| **Checkpoint Transactions** | 1000000 |

Checkpoint Image URI

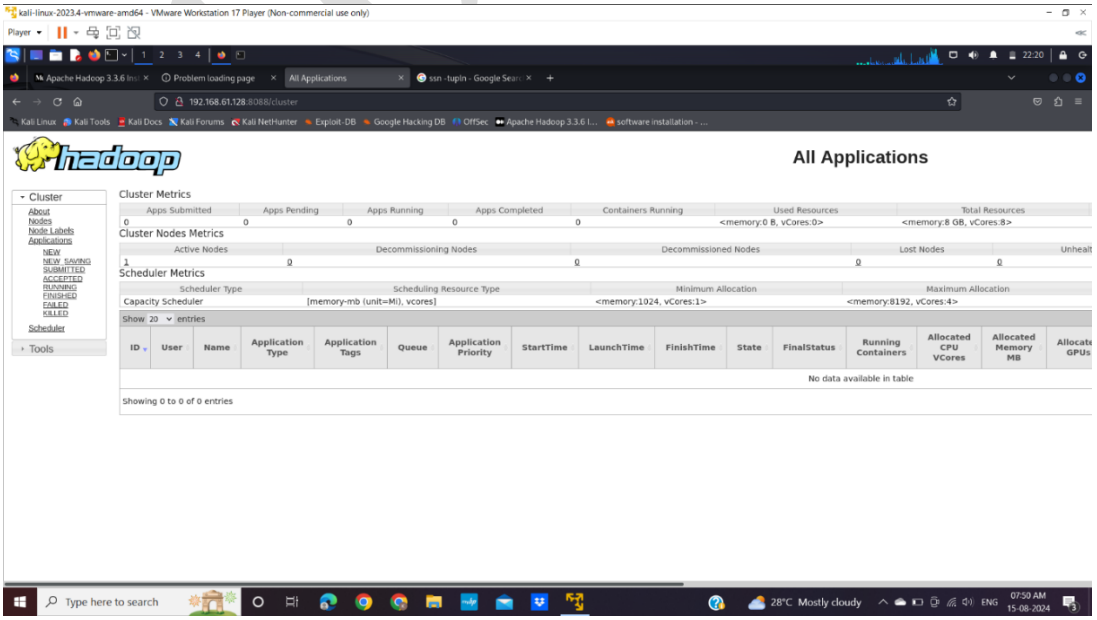
- file:///tmp/hadoop-kali/dfs/namesecondary

Checkpoint Editlog URI

- file:///tmp/hadoop-kali/dfs/namesecondary

Hadoop, 2023.


**localhost:8088**



**RESULT:**

Thus, Hadoop has been successfully installed.