

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

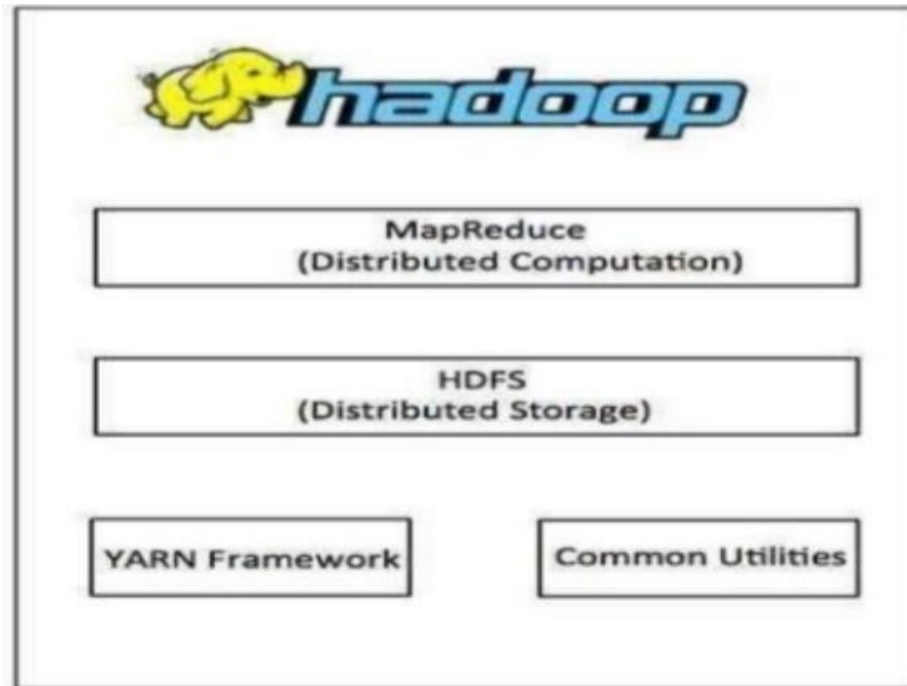
Hadoop framework includes following four modules:

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.



```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.22631.4169]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd/

C:\>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\>
```

```
C:\>jps
22624 Jps
24224 ResourceManager
12164 NameNode
26948 NodeManager
1612 DataNode
```

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview 'localhost:9000' (✔active)

| | |
|----------------|--|
| Started: | Mon Sep 16 12:26:15 +0530 2024 |
| Version: | 3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c |
| Compiled: | Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1) |
| Cluster ID: | CID-5d1cd15f-2cbc-4ded-aefc-9d1d8350c3f2 |
| Block Pool ID: | BP-590392694-192.168.1.5-1724249318101 |

Summary

Security is off.

Safemode is off.

79 files and directories, 24 blocks (24 replicated blocks, 0 erasure coded block groups) = 103 total filesystem object(s).

Heap Memory used 163.88 MB of 379.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 51.34 MB of 52.75 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

| | |
|--|--|
| Configured Capacity: | 217.09 GB |
| Configured Remote Capacity: | 0 B |
| DFS Used: | 42.05 MB (0.02%) |
| Non DFS Used: | 176.49 GB |
| DFS Remaining: | 40.56 GB (18.68%) |
| Block Pool Used: | 42.05 MB (0.02%) |
| DataNodes usages% (Min/Median/Max/stdDev): | 0.02% / 0.02% / 0.02% / 0.00% |
| Live Nodes | 1 (Decommissioned: 0, In Maintenance: 0) |
| Dead Nodes | 0 (Decommissioned: 0, In Maintenance: 0) |

NameNode Journal Status

Current transaction ID: 433

| Journal Manager | State |
|--|--|
| FileJournalManager(root=C:\hadoop-3.3.6\data\namenode) | EditLogFileOutputStream(C:\hadoop-3.3.6\data\namenode\current\edits_inprogress_000000000000000433) |

NameNode Storage

| Storage Directory | Type | State |
|-------------------------------|-----------------|--------|
| C:\hadoop-3.3.6\data\namenode | IMAGE_AND_EDITS | Active |

DFS Storage Types

| Storage Type | Configured Capacity | Capacity Used | Capacity Remaining | Block Pool Used | Nodes In Service |
|--------------|---------------------|------------------|--------------------|-----------------|------------------|
| DISK | 217.09 GB | 42.05 MB (0.02%) | 40.56 GB (18.68%) | 42.05 MB | 1 |

Hadoop, 2023.

RESULT:

Thus the set up of single hadoop cluster and show the process using web UI is completed successfully.