Cluster Setup Documentation

Team : 122AD0045, 122AD0043, 122AD0018

---

1. Technologies Used

- **Apache Spark**: Distributed processing engine for large-scale data analytics.
- **Hadoop HDFS**: Distributed storage system for handling tweet datasets.
- **Python/PySpark**: For implementing NLP, ML models, and Spark jobs.
- **Jupyter Notebook**: Interactive environment for prototyping and analysis.
- **OpenCage Geocoding API**: For converting text locations to coordinates.
- **Folium/Plotly**: Visualization libraries for spatial and sentiment analysis.

---

2. Cluster Configuration

Hardware Setup:

- **Master Node**: 1 PC (4 CPU cores, 4GB RAM, 25GB SSD).
- **Slave Nodes**: 2 PCs (each with 4 CPU cores, 4GB RAM, 25GB SSD).

Software Stack:

- **OS**: Ubuntu 20.04 LTS.
- **Java**: OpenJDK 8 (required for Spark/Hadoop).
- **Spark**: v3.2.1 (configured in standalone mode).
- **Hadoop**: v3.3.1 (for HDFS storage).

Network:

- Static IPs assigned to all nodes.
- Password-less SSH configured between master and slaves.

---

3. Steps to Set Up the Cluster

A. Prerequisites

1. **Install Java**:

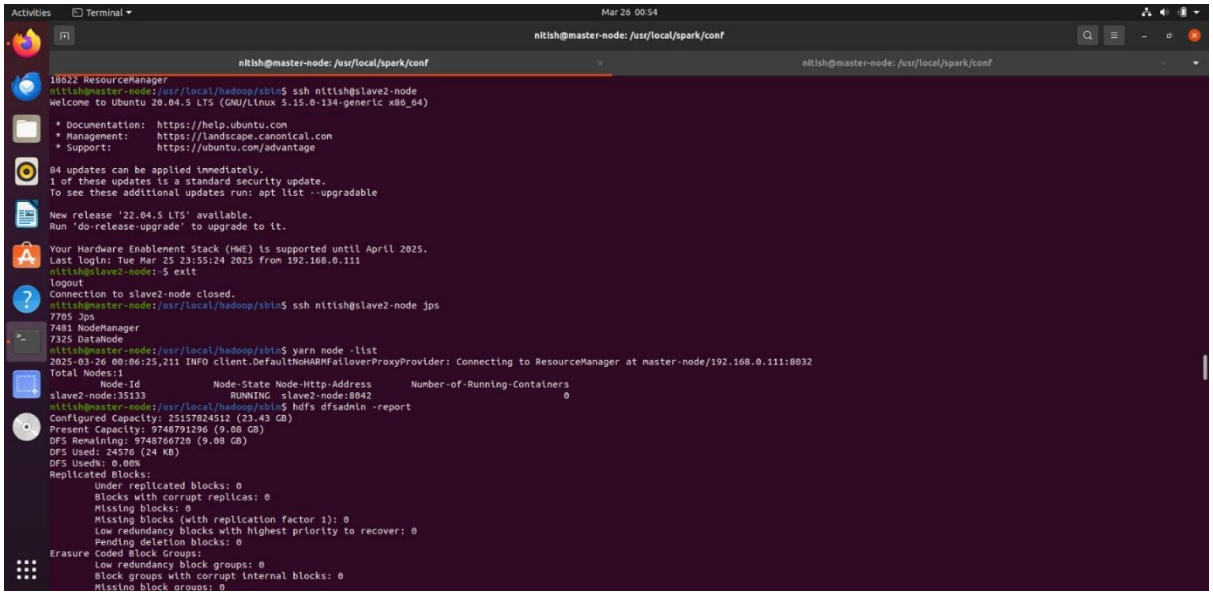
sudo apt update && sudo apt install openjdk-8-jdk

2. **Configure SSH**:

ssh-keygen -t rsa # Generate keys on master

ssh-copy-id slave1 # Copy keys to slaves

ssh-copy-id slave2



B. Install Hadoop & Spark

1. **Download and Extract**:

wget https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz

tar -xvzf hadoop-3.3.1.tar.gz

wget https://downloads.apache.org/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz

tar -xvzf spark-3.2.1-bin-hadoop3.2.tgz

2. **Configure Environment Variables**:

Add to ~/.bashrc:

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export HADOOP_HOME=/path/to/hadoop-3.3.1

export SPARK_HOME=/path/to/spark-3.2.1-bin-hadoop3.2

export PATH=$PATH:$HADOOP_HOME/bin:$SPARK_HOME/bin

C. Configure Spark & Hadoop

1. **Spark Config**:
   - Edit $SPARK_HOME/conf/spark-env.sh:

export SPARK_MASTER_HOST=<master-node-IP>

export SPARK_WORKER_CORES=4

export SPARK_WORKER_MEMORY=8g

- Update $SPARK_HOME/conf/slaves:

slave1

slave2

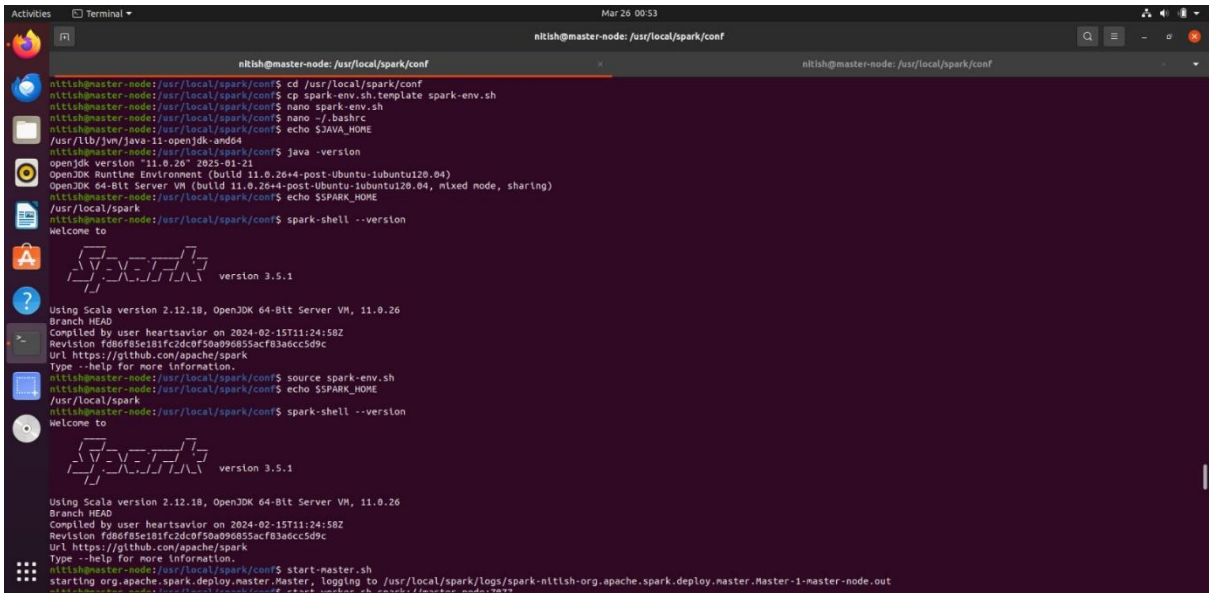
2. **Hadoop Config**:
   - Core-site.xml: Set HDFS URI.
   - Hdfs-site.xml: Configure replication factor (2)

D. Start the Cluster

1. **Launch Hadoop HDFS**:

hdfs namenode -format # First-time setup

start-dfs.sh



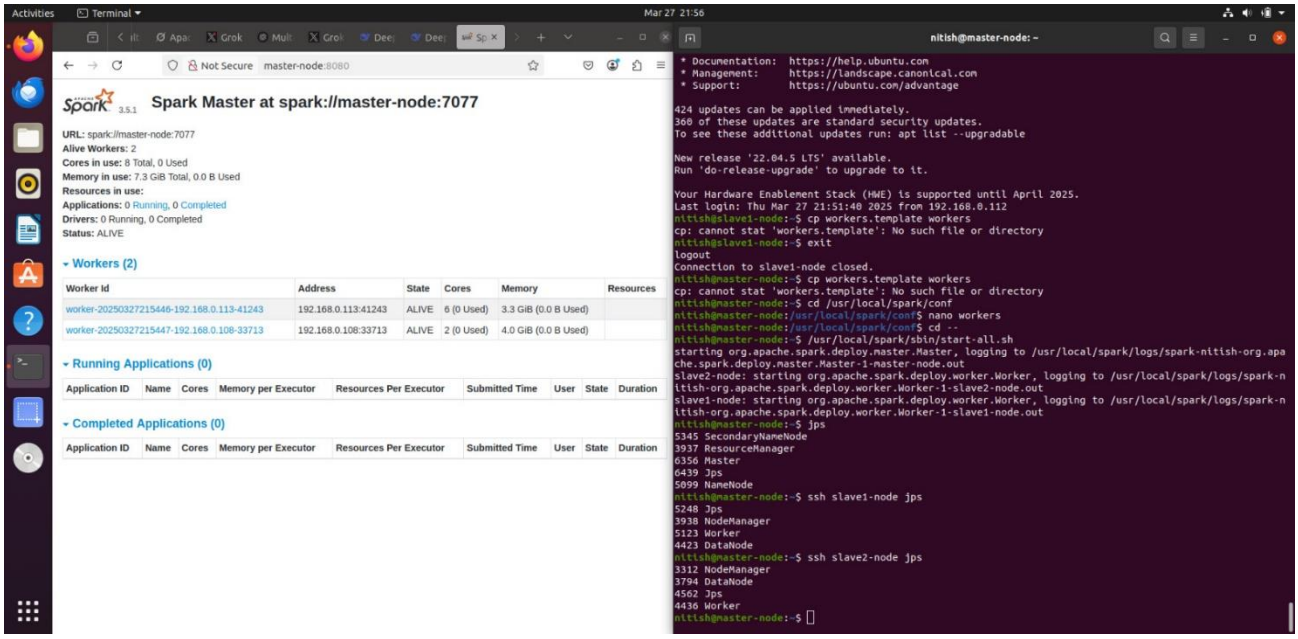
2. **Start Spark Cluster**:

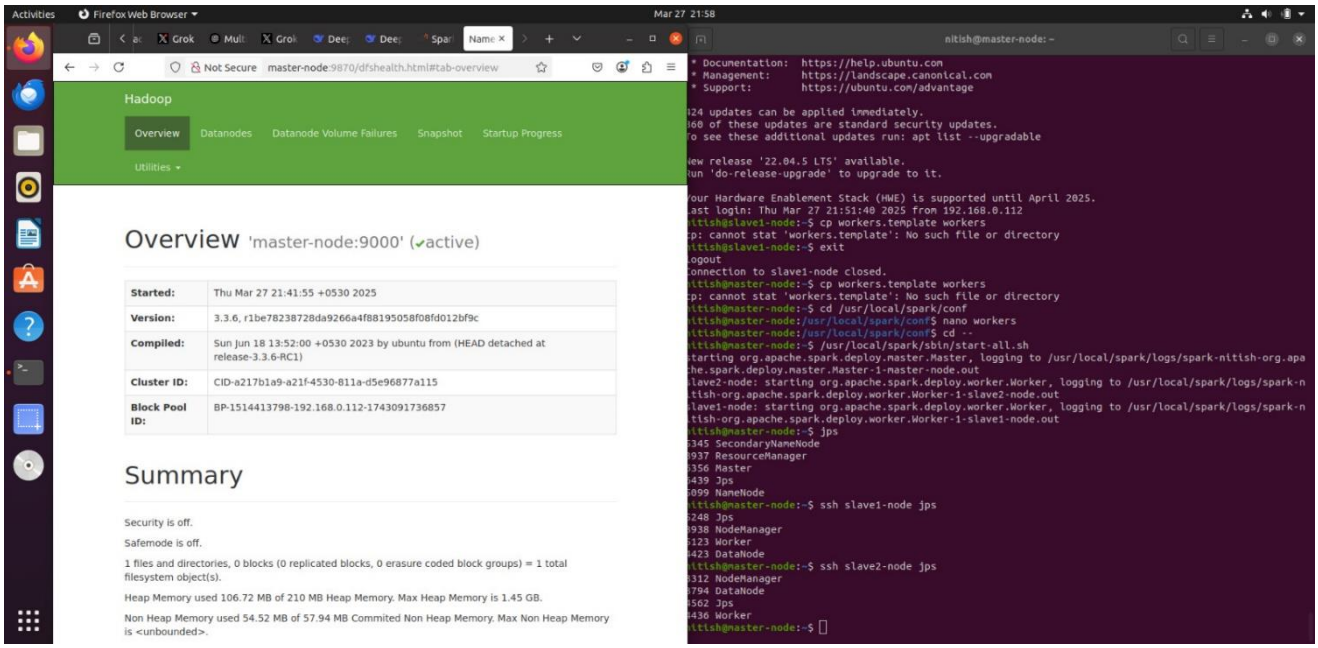
$SPARK_HOME/sbin/start-master.sh

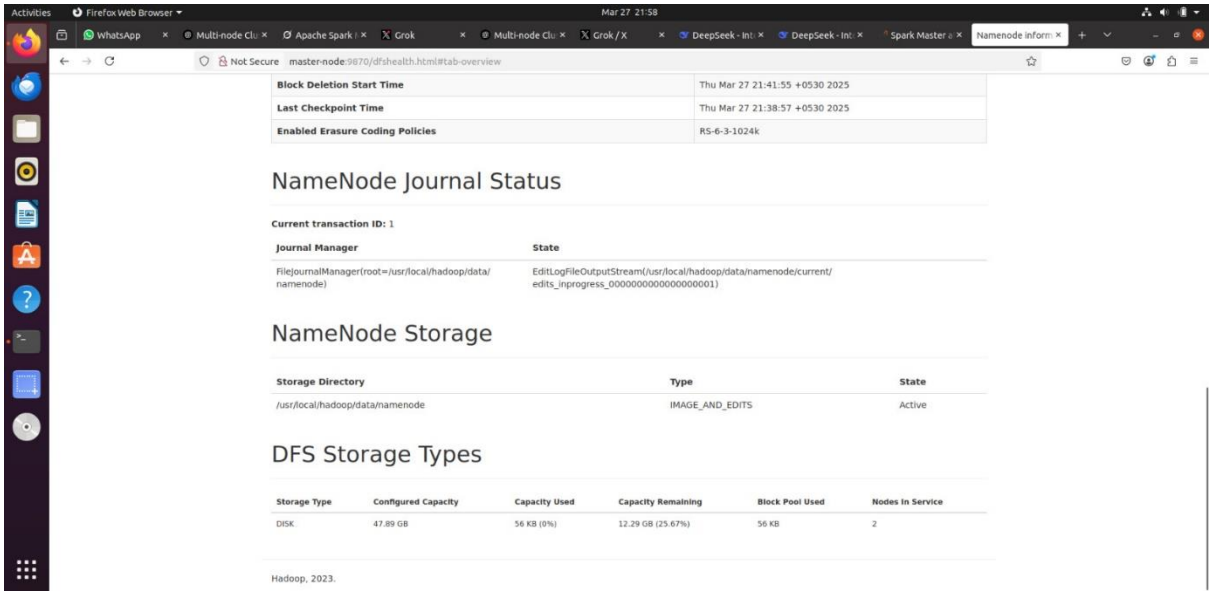
$SPARK_HOME/sbin/start-workers.sh

E. Verify Setup

- **Spark UI**: Access http://<master-IP>:8080 to view active workers.
- **HDFS UI**: Check http://<master-IP>:9870.


Activities
Terminal
Mar 27 21:56
Not Secure master-node:8080
Spark 3.5.1
Spark Master at spark://master-node:7077
URL: spark://master-node:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 7.3 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE
Workers (2)
Worker Id | Address | State | Cores | Memory | Resources
worker-20250327215446-192.168.0.113-41243 | 192.168.0.113:41243 | ALIVE | 6 (0 Used) | 3.3 GiB (0.0 B Used)
worker-20250327215447-192.168.0.108-33713 | 192.168.0.108:33713 | ALIVE | 2 (0 Used) | 4.0 GiB (0.0 B Used)
Running Applications (0)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
Completed Applications (0)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
nitish@master-node: ~
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
424 updates can be applied immediately.
360 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable
New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Thu Mar 27 21:51:40 2025 from 192.168.0.112
nitish@slave1-node:~$ cp workers.template workers
cp: cannot stat 'workers.template': No such file or directory
nitish@slave1-node:~$ exit
logout
Connection to slave1-node closed.
nitish@master-node:~$ cp workers.template workers
cp: cannot stat 'workers.template': No such file or directory
nitish@master-node:~$ cd /usr/local/spark/conf
nitish@master-node:/usr/local/spark/conf$ nano workers
nitish@master-node:/usr/local/spark/conf$ cd --
nitish@master-node:~$ /usr/local/spark/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.master.Master-1-master-node.out
slave2-node: starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.worker.Worker-1-slave2-node.out
slave1-node: starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.worker.Worker-1-slave1-node.out
nitish@master-node:~$ jps
5345 SecondaryNameNode
3937 ResourceManager
6356 Master
6439 Jps
5099 NameNode
nitish@master-node:~$ ssh slave1-node jps
5248 Jps
3938 NodeManager
5123 Worker
4423 DataNode
nitish@master-node:~$ ssh slave2-node jps
3312 NodeManager
3794 DataNode
4562 Jps
4436 Worker
nitish@master-node:~$


Activities
Firefox Web Browser
Mar 27 21:58
Not Secure master-node:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'master-node:9000' (active)
Started: Thu Mar 27 21:41:55 +0530 2025
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-a217b1a9-a21f-4530-811a-d5e96877a115
Block Pool ID: BP-1514413798-192.168.0.112-1743091736857
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 106.72 MB of 210 MB Heap Memory. Max Heap Memory is 1.45 GB.
Non Heap Memory used 54.52 MB of 57.94 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
nitish@master-node: ~
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
424 updates can be applied immediately.
360 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable
New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Thu Mar 27 21:51:40 2025 from 192.168.0.112
nitish@slave1-node:~$ cp workers.template workers
cp: cannot stat 'workers.template': No such file or directory
nitish@slave1-node:~$ exit
logout
Connection to slave1-node closed.
nitish@master-node:~$ cp workers.template workers
cp: cannot stat 'workers.template': No such file or directory
nitish@master-node:~$ cd /usr/local/spark/conf
nitish@master-node:/usr/local/spark/conf$ nano workers
nitish@master-node:/usr/local/spark/conf$ cd --
nitish@master-node:~$ /usr/local/spark/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.master.Master-1-master-node.out
slave2-node: starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.worker.Worker-1-slave2-node.out
slave1-node: starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark/logs/spark-nitish-org.apache.spark.deploy.worker.Worker-1-slave1-node.out
nitish@master-node:~$ jps
5345 SecondaryNameNode
3937 ResourceManager
6356 Master
6439 Jps
5099 NameNode
nitish@master-node:~$ ssh slave1-node jps
5248 Jps
3938 NodeManager
5123 Worker
4423 DataNode
nitish@master-node:~$ ssh slave2-node jps
3312 NodeManager
3794 DataNode
4562 Jps
4436 Worker
nitish@master-node:~$



---

4. Issues Encountered & Resolutions

Issue	Resolution
Slave nodes not connecting	Ensured SSH keys were copied correctly and firewall rules allowed port 8080/7077.
Out-of-memory errors	Reduced SPARK_WORKER_MEMORY to 6g and optimized partition sizes.
HDFS permission errors	Ran hdfs dfs -chmod -R 777 / for development (not recommended for production).
Geocoding API rate limits	Implemented caching for repeated location queries.
Jupyter-Spark integration fails	Used findspark.init() and set PYSPARK_PYTHON explicitly.