**1. System Configuration**

| **Machine** | **Role** | **RAM** | **Cores** | **Processes** |
| --- | --- | --- | --- | --- |
| 192.168.1.16 (vmware1) | **Master + Worker1** | **4GB** | 2 | **2** |
| 192.168.1.20 (vmware2) | **Worker2** | **4GB** | 2 | **2** |
| 192.168.1.21 (vmware3) | **Worker2** | **4GB** | **2** | **2** |

**2. Install Java (On Both Machines)**

Apache Spark requires **Java 11**. Install Java 11:

Do it on 2 machines

sudo apt update

sudo apt install openjdk-11-jdk -y

**3. Install Apache Spark 3.5.4 (On Both Machines)**

**3.1. Download and Extract Spark**

cd /opt

wget https://dlcdn.apache.org/spark/spark-3.5.4/spark-3.5.4-bin-hadoop3.tgz

tar -xvzf spark-3.5.4-bin-hadoop3.tgz

**3.2. Set Environment Variables**

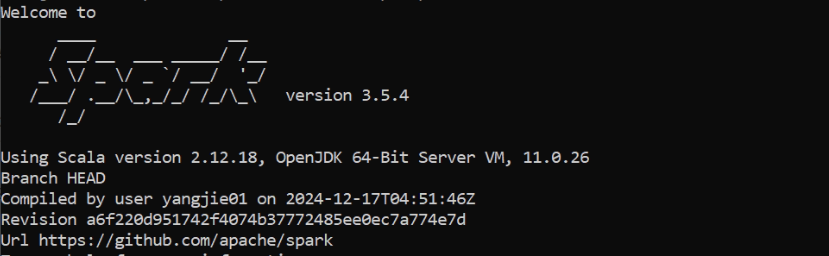
Add the following to ~/.bashrc:



source ~/.bashrc

Verify Spark installation:

spark-shell –version



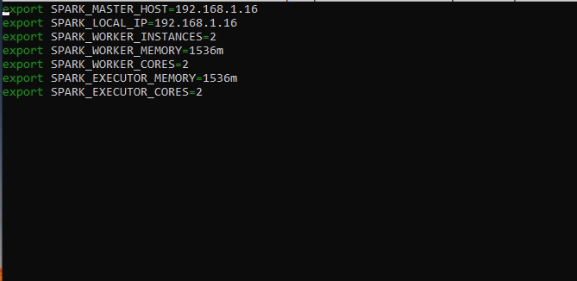
**4. Configure Spark**

**4.1. Configure Spark Master (On vmware1 - 192.168.1.16)**

Edit **spark-env.sh**:

nano $SPARK\_HOME/conf/spark-env.sh

Add the following:



**Explaination:**

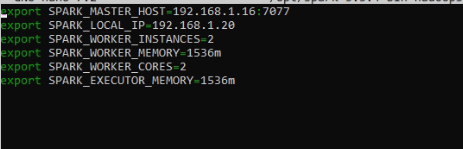
* **SPARK\_MASTER\_HOST=192.168.1.16**  
  Specifies the Spark Master’s IP address, allowing Workers to properly register and receive tasks.
* **SPARK\_LOCAL\_IP=192.168.1.16**  
  Defines the local IP for Spark to bind to, ensuring network traffic is routed correctly on this machine.
* **SPARK\_WORKER\_INSTANCES=2**  
  Launches 2 Worker processes on the same host, distributing tasks across processes for improved parallelism.
* **SPARK\_WORKER\_MEMORY=1536m (1.5GB)**  
  Allocates 1.5GB of memory per Worker, preventing memory bottlenecks and ensuring stable task execution.
* **SPARK\_WORKER\_CORES=2**  
  Assigns 2 CPU cores per Worker, enabling concurrent task processing and increasing throughput.
* **SPARK\_EXECUTOR\_MEMORY=1536m (1.5GB)**  
  Reserves 1.5GB of RAM for each Executor, providing enough memory for parallel task execution without causing out-of-memory errors.
* **SPARK\_EXECUTOR\_CORES=2**  
  Grants each Executor 2 CPU cores, enhancing parallel processing capabilities and reducing execution time.

**4.2. Configure Spark Worker (On vmware2 - 192.168.1.20)**

Edit **spark-env.sh**:

nano $SPARK\_HOME/conf/spark-env.sh

Add the following:



**Explaination:**

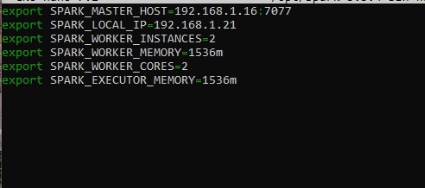
* **Same as Master configuration**, ensuring balanced resource allocation.

**4.3. Configure Spark Worker (On vmware2 - 192.168.1.21)**

Edit **spark-env.sh**:

nano $SPARK\_HOME/conf/spark-env.sh

Add the following:



**Explaination:**

**Same as Master configuration**, ensuring balanced resource allocation.

**4.3. Configure workers File**

Edit the **workers** file on **vmware1** and **vmware2**:

nano $SPARK\_HOME/conf/workers

vmware1

vmware2

vmware3

**5. Start Spark Cluster**

**5.1. Start Spark Master (On vmware1)**

$SPARK\_HOME/sbin/start-master.sh

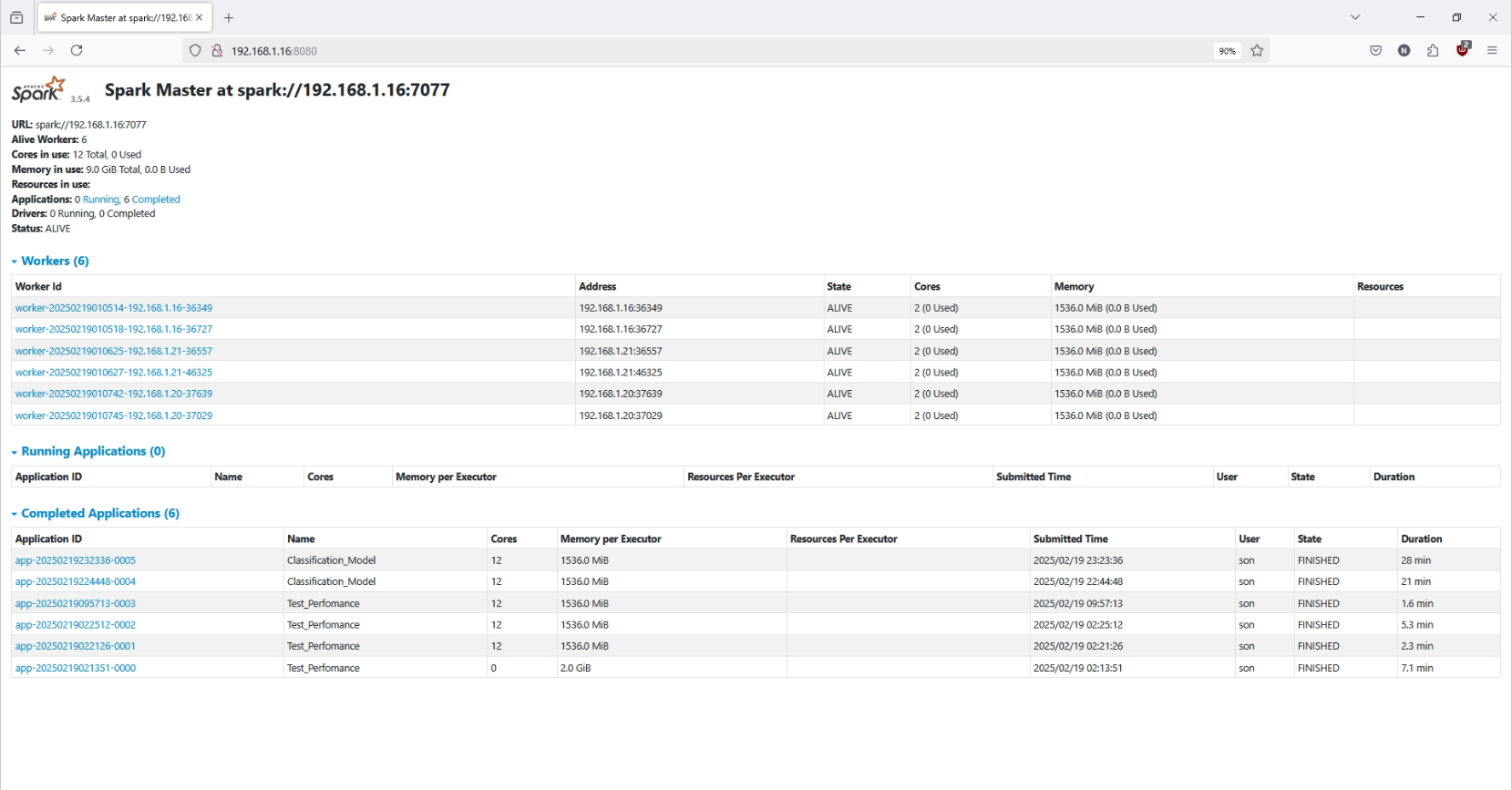
Verify the Master UI:  
👉 Open: **http://192.168.1.16:8080**

**5.2. Start Spark Workers**

On **vmware1** and **vmware2**:

$SPARK\_HOME/sbin/start-worker.sh spark://192.168.1.16:7077

Verify on **Master UI** (http://192.168.1.16:8080), you should see **2 workers**.



**7. Stop Spark Cluster**

bash

CopyEdit

stop-all.sh