Spark Installation

Tags	Installation
URL	None
發表時間	@October 14, 2021

Installation Platform

- Ubuntu 20.04
- Java 11
- Spark 3.1.2 with hadoop 3.2 download: https://www.apache.org/dyn/closer.lua/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz
- Scala 3.0.2: https://github.com/lampepfl/dotty/releases/tag/3.0.2

Installation of Scala

Installing Scala

Follow the below given steps for installing Scala.

https://github.com/lampepfl/dotty/releases/tag/3.0.2

Extract the Scala tar file

Type the following command for extracting the Scala tar file.

```
$ tar xvf scala-2.11.6.tgz
```

Move Scala software files

Use the following commands for moving the Scala software files, to respective directory **(/usr/local/scala)**.

```
$ su –
Password:
cd /home/Hadoop/Downloads/
mv scala-2.11.6 /usr/local/scala
# exit
```

Set PATH for Scala

Use the following command for setting PATH for Scala.

```
$ export PATH = $PATH:/usr/local/scala/bin
```

Verifying Scala Installation

After installation, it is better to verify it. Use the following command for verifying Scala installation.

```
$scala -version
```

If Scala is already installed on your system, you get to see the following response −

```
Scala code runner version 2.11.6 -- Copyright 2002-2013, LAMP/EPFL
```

Steps of installation of Spark

Following steps:

```
# update the package manager information
sudo apt update && upgrade

# install java
sudo apt install openjdk-11-jdk

# download spark
wget https://dlcdn.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

# extract spark package
tar zxvf [your spark package name]

# move spark directory to /opt/spark
sudo mv [your spark directory name] /opt/spark

# add your binary path file of spark to the .bashrc (add environment variable)
sudo nano ~/.bashrc
```

In the `.bashrc` file, please export your path of binary file of spark:

```
...
# add three lines below
export JAVA_HOME=[your java binary file of path]
export SPARK_HOME=/opt/spark
export PATH=$PATH:$JAVA_HOME/bin:$SPARK_HOME/sbin
```

Finally, activate your file changed on environment variable.

`source ~/.bashrc`

In additional, you can change the hostname for you to identify well.

`hostnamectl set-hostname [your host name]`

Successful Running

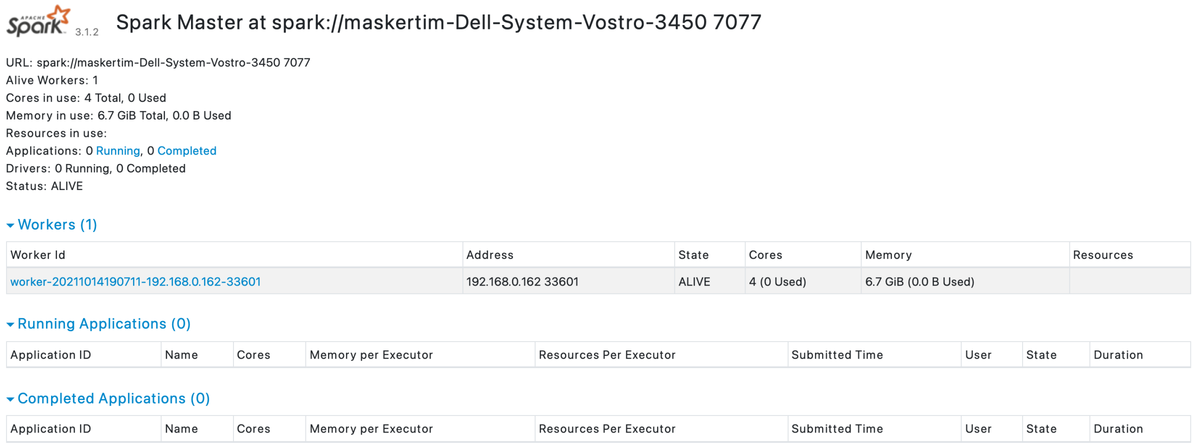
Start your master node as following command:

`start-master`

Start your worker node as following command:

`start-worker [master url address]`

This is a spark Web Management GUI. Default url is `http://[your ip]:8080`



3.1.2 Spark Master at spark://maskertim-Dell-System-Vostro-3450 7077

URL: spark://maskertim-Dell-System-Vostro-3450 7077
Alive Workers: 1
Cores in use: 4 Total, 0 Used
Memory in use: 6.7 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20211014190711-192.168.0.162-33601	192.168.0.162 33601	ALIVE	4 (0 Used)	6.7 GiB (0.0 B Used)	

▾ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

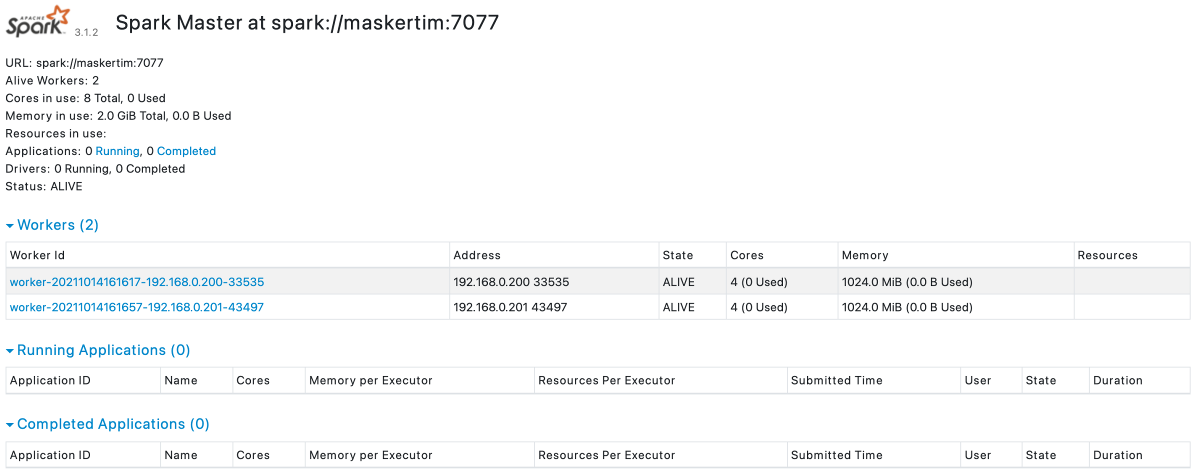
▾ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Cluster with two raspberry pi:

`start-worker [master url address (e.g., spark://192.168.0.250:7077)]`

Worker Web GUI is default `http://[worker ip]:8081`



3.1.2 Spark Master at spark://maskertim:7077

URL: spark://maskertim:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 2.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (2)

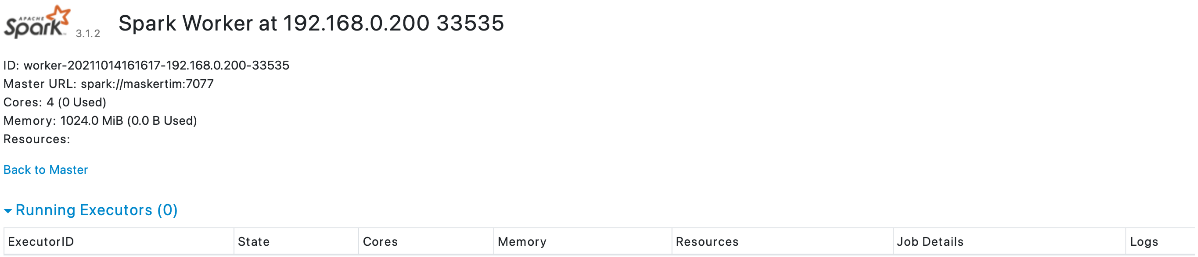
Worker Id	Address	State	Cores	Memory	Resources
worker-20211014161617-192.168.0.200-33535	192.168.0.200 33535	ALIVE	4 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20211014161657-192.168.0.201-43497	192.168.0.201 43497	ALIVE	4 (0 Used)	1024.0 MiB (0.0 B Used)	

▾ 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



Spark 3.1.2 Spark Worker at 192.168.0.200 33535

ID: worker-20211014161617-192.168.0.200-33535
Master URL: spark://maskertim:7077
Cores: 4 (0 Used)
Memory: 1024.0 MiB (0.0 B Used)
Resources:

Back to Master

▾ Running Executors (0)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs

Reference

Download page of Spark:

Downloads | Apache Spark

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12. Preview releases, as the name suggests, are

http://spark.apache.org/downloads.html

Video tutorial of installation of Spark

How to Install Apache Spark on Linux Ubuntu 20.04 easily?

Spark #ApacheSpark #LinuxIn this video, I want to show you How to Install Apache Spark on Linux Ubuntu 20.04 easily.

https://www.youtube.com/watch?v=XajAVPMZxeU