Spark 安裝介紹

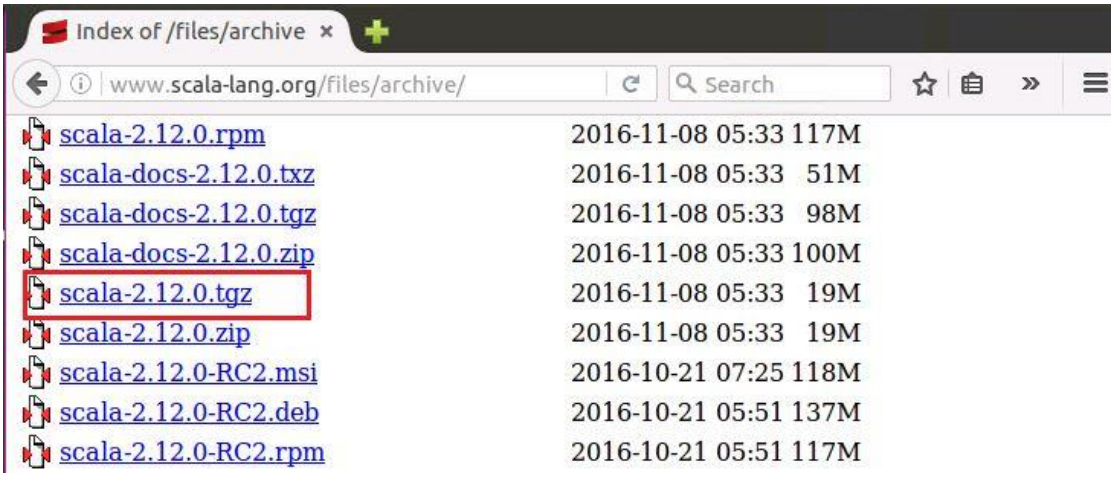
Spark Cluster Manager 可以執行在下列模式：

1. 本機執行(Local Machine)：於本機執行，適合入門學習，測試用。
2. Spark Standalone cluster：由 Spark 提供的 cluster 管理模式，若沒有架設 Hadoop Multi Node cluster，可以用本模式操作 HDFS。
3. Hadoop YARN：於 YARN 上執行，由 YARN 進行多台機器的資源管理。
4. 雲端執行：針對更大型規模的計算工作，可以將 Spark 程式在雲端執行，如 AWS 的 EC2 平台。

本文將教導本機執行和 Spark Standalone cluster 和 YARN 的安裝和執行方式。

1. Scala 安裝

- Spark 可以用 python、Java 等多種語言執行，本文選擇 scala 為主
- 下載 Scala，可於網址看到不同版本的 Scala



- 執行 wget http://www.scala-lang.org/files/archive/scala-2.12.0.tgz

- 解壓縮 Scala，輸入 tar xvf scala-2.12.0.tgz

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ tar xvf scala-2.12.0.tgz
scala-2.12.0/
scala-2.12.0/man/
scala-2.12.0/man/man1/
scala-2.12.0/man/man1/scala.1
scala-2.12.0/man/man1/scalap.1
scala-2.12.0/man/man1/fsc.1
scala-2.12.0/man/man1/scaladoc.1
scala-2.12.0/man/man1/scalac.1
scala-2.12.0/bin/
scala-2.12.0/bin/scalac
```

- 搬移至/usr/local 下，輸入 sudo mv scala-2.12.0 /usr/local/scala

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo mv scala-2.12.0 /usr/local/scala
[sudo] password for hduser:
hduser@hadoopmaster:~$
```

- 編輯~/.bashrc，輸入 sudo gedit ~/.bashrc

```
Export SCALA_HOME=/usr/local/scala
Export PATH=$PATH:$SCALA_HOME/bin
```

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo gedit ~/.bashrc
Open ▾   *.bashrc ~/   Save
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
export JAVA_LIBRARY_PATH=$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH
#Hadoop Variables
#SCALA Variables
export SCALA_HOME=/usr/local/scala
export PATH=$PATH:$SCALA_HOME/bin
#SCALA Variables
```

- 讓~/.bashrc 生效，輸入 source ~/.bashrc

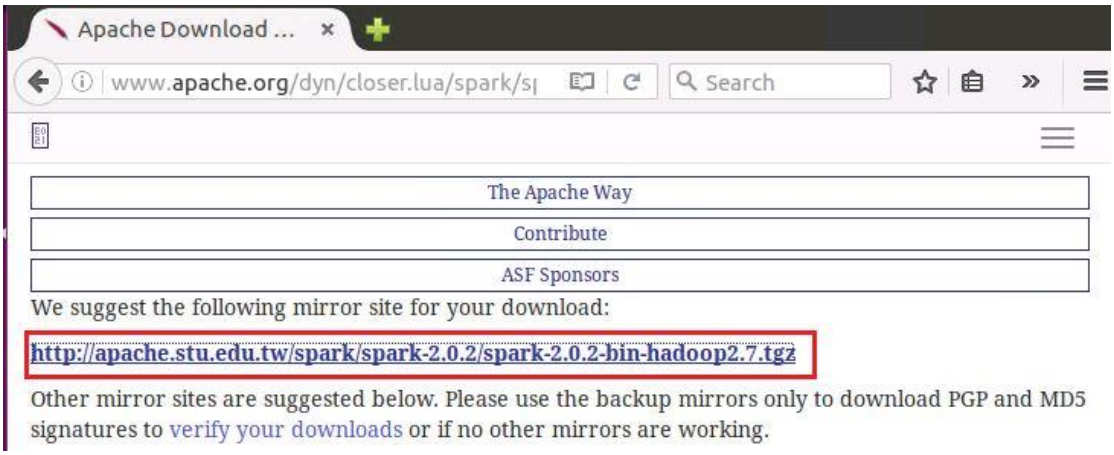
```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ source ~/.bashrc
hduser@hadoopmaster:~$ ~
```

- 到此即可執行 Scala，輸入 scala，測試輸入程式執行

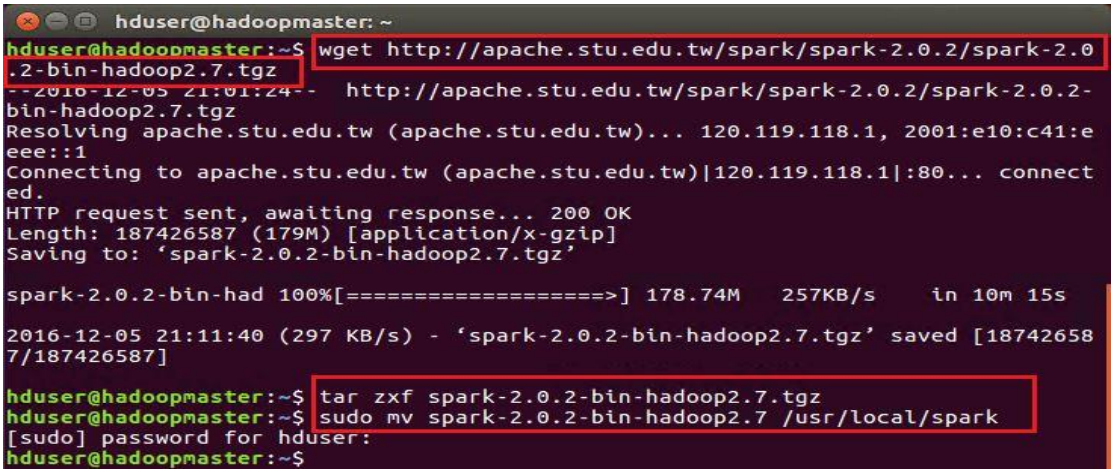
```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ scala
Welcome to Scala 2.12.0 (OpenJDK 64-Bit Server VM, Java 1.8.0_111).
Type in expressions for evaluation. Or try :help.

scala> 1+1
res0: Int = 2

scala> :q
hduser@hadoopmaster:~$
```

2. 安裝 Spark

- 到 Spark 網址下載 Spark，注意需配合 Hadoop 版本來選擇 Spark 版本



- 下載 Spark，輸入 wget http://apache.stu.edu.tw/spark/spark-2.0.2/spark-2.0.2-bin-hadoop2.7.tgz
- 解壓縮，輸入 tar zxf spark-2.0.2-bin-hadoop2.7.tgz
- 搬移至/usr/local/spark 下，輸入 sudo mv spark-2.0.2-bin-hadoop2.7 /usr/local/spark



- 編輯~/.bashrc，輸入 sudo gedit ~/.bashrc

```
Export SPARK_HOME=/usr/local/spark
Export PATH=$PATH:$SPARK_HOME/bin
```

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo gedit ~/.bashrc
Open ▾   *.bashrc ~/   Save
#Hadoop Variables
#SCALA Variables
export SCALA_HOME=/usr/local/scala
export PATH=$PATH:$SCALA_HOME/bin
#SCALA Variables
#SPARK Variables
export SPARK_HOME=/usr/local/spark
export PATH=$PATH:$SPARK_HOME/bin
#SPARK Variables
```

- 讓設定生效，輸入 source ~/.bashrc

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ source ~/.bashrc
hduser@hadoopmaster:~$
```

- 啟動 spark-shell，輸入 spark-shell

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/12/05 21:19:21 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/12/05 21:19:26 WARN SparkContext: Use an existing SparkContext, some configur
ation may not take effect.
Spark context Web UI available at http://192.168.59.137:4040
Spark context available as 'sc' (master = local[*], app id = local-1480943965187
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- 設定 spark-shell 互動介面的顯示訊息，因為預設會顯示過多訊息，影响閱讀。
  i. cd /usr/local/spark/conf
  ii. cp log4j.properties.template log4j.properties
  iii. 編輯 log4j.properties，輸入 sudo gedit log4j.properties

```
hduser@hadoopmaster: /usr/local/spark/conf
hduser@hadoopmaster:~$ cd /usr/local/spark/conf
hduser@hadoopmaster:/usr/local/spark/conf$ cp log4j.properties.template log4j.pr
operties
hduser@hadoopmaster:/usr/local/spark/conf$ sudo gedit log4j.properties
```

```
Open ▾   *log4j.properties
         /usr/local/spark/conf   Save
# limitations under the License.
#

# Set everything to be logged to the console
log4j.rootCategory=WARN, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %
m%n
```

- 再次進入 spark-shell，輸入 spark-shell，會發現少了一些訊息

```
hduser@hadoopmaster: /usr/local/spark/conf
hduser@hadoopmaster:/usr/local/spark/conf$ spark-shell
16/12/05 21:28:31 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/12/05 21:28:36 WARN SparkContext: Use an existing SparkContext, some configur
ation may not take effect.
Spark context Web UI available at http://192.168.59.137:4040
Spark context available as 'sc' (master = local[*], app id = local-1480944514757
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

3. 本機執行 Spark

- 啟動虛擬機 HadoopMaster、HadoopSlave1、HadoopSlave2
- 啟動 Hadoop，於 HadoopMaster 輸入 start-all.sh

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [hadoopmaster]
hadoopmaster: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser
-namenode-hadoopmaster.out
hadoopslave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser
-datanode-hadoopslave1.out
hadoopslave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser
-datanode-hadoopslave2.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hd
user-secondarynamenode-hadoopmaster.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resource
manager-hadoopmaster.out
hadoopslave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduse
r-nodemanager-hadoopslave1.out
hadoopslave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduse
r-nodemanager-hadoopslave2.out
hduser@hadoopmaster:~$
```

- 進入 spark-shell，輸入 spark-shell --master local[4]

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ spark-shell --master local[4]
16/12/09 19:44:30 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/12/09 19:44:36 WARN SparkContext: Use an existing SparkContext, some configur
ation may not take effect.
Spark context Web UI available at http://192.168.59.137:4040
Spark context available as 'sc' (master = local[4], app id = local-1481283874802
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.
```

- 測試讀取本機檔案，輸入
  i. val textFile=sc.textFile("file:/usr/local/spark/README.md")
  ii. textFile.count

```
scala> val textFile=sc.textFile("file:/usr/local/spark/README.md")
textFile: org.apache.spark.rdd.RDD[String] = file:/usr/local/spark/README.md Map
PartitionsRDD[1] at textFile at <console>:24

scala> textFile.count
res0: Long = 99
```

- 讀取 HDFS 的檔案(假設在 HDFS 上有一個/user/hduser/test/README.txt 的檔案)
  i. val

textFile=sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/README.txt")

ii. textFile.count

```
scala> val textFile = sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/REA
DME.txt")
textFile: org.apache.spark.rdd.RDD[String] = hdfs://hadoopmaster:9000/user/hduse
r/test/README.txt MapPartitionsRDD[5] at textFile at <console>:24

scala> textFile.count()
res2: Long = 31

scala>
```

4. 在 Spark standalone cluster 環境執行

- 自樣本建立 spark-env.sh 檔案，輸入 cp /usr/local/spark/conf/spark-env.sh.template /usr/local/spark/conf/spark-env.sh

```
hduser@hadoopmaster:~$ cp /usr/local/spark/conf/spark-env.sh.template  /usr/loca
l/spark/conf/spark-env.sh
```

- 設定 spark-env.sh，設定每個 worker 的資源分配，輸入 sudo gedit /usr/local/spark/conf/spark-env.sh(注意：每個 worker 的記憶體不得低於 1G，否則無法運行)

```
export SPARK_MASTER_IP=hadoopmaster
export SPARK_WORKER_CORE=1
export SPARK_WORKER_MEMORY=1g
export SPARK_WORKER_INSTANCES=2
```

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo gedit /usr/local/spark/conf/spark-env.sh
Open ▾   *spark-env.sh  /usr/local/spark/conf   Save
# - SPARK_NICENESS       The scheduling priority for daemons. (Default: 0)
#export SPARK_MASTER_IP=hadoopmaster
export SPARK_MASTER_IP=hadoopmaster
export SPARK_WORKER_CORE=1
export SPARK_WORKER_MEMORY=1g
export SPARK_WORKER_INSTANCES=2
```

- 將 hadoopmaster 的 Spark 程式複製到 HadoopSlave1，輸入以下指令

i. ssh hadoopslave1

ii. sudo mkdir /usr/local/spark

iii. sudo chown hduser:hduser /usr/local/spark

iv. exit

```
hduser@hadoopmaster:~$ ssh hadoopslave1
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-31-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

196 packages can be updated.
4 updates are security updates.

*** System restart required ***
Last login: Thu Dec  8 20:24:01 2016 from 192.168.59.137
hduser@hadoopslave1:~$ sudo mkdir /usr/local/spark
[sudo] password for hduser:
hduser@hadoopslave1:~$ sudo chown hduser:hduser /usr/local/spark
hduser@hadoopslave1:~$ exit
logout
Connection to hadoopslave1 closed.
hduser@hadoopmaster:~$ sudo scp -r /usr/local/spark hduser@hadoopslave1:/usr/loc
al
The authenticity of host 'hadoopslave1 (192.168.59.134)' can't be established.
ECDSA key fingerprint is SHA256:l5HfVz2GKon2xpmavQSLRqfvPdxuogiqaF/Xjx5XV3E.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'hadoopslave1,192.168.59.134' (ECDSA) to the list of
known hosts.
hduser@hadoopslave1's password:
NOTICE                                        100%   24KB  24.2KB/s   00:00
hello.txt                                     100%   13     0.0KB/s   00:00
userlib-0.1.zip                               100%  668     0.7KB/s   00:00
```

- 仿上面步驟將 Spark 複製到 HadoopSlave2
- 編輯 slaves 檔案，設定 Spark Standalone cluster 有那些伺服器，輸入 sudo gedit /usr/local/spark/conf/slaves

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo gedit /usr/local/spark/conf/slaves
Open ▾   *slaves  /usr/local/spark/conf   Save
hadoopslave1
hadoopslave2
```

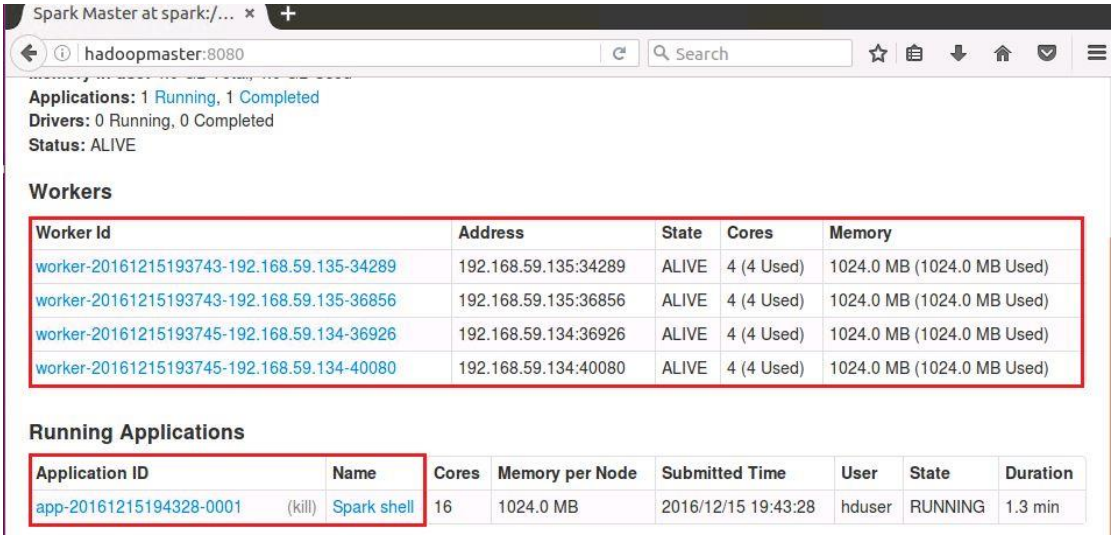
- 啟動 Spark Standalone cluster，輸入/usr/local/spark/sbin/start-all.sh

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ /usr/local/spark/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /usr/local/spark/logs
/spark-hduser-org.apache.spark.deploy.master.Master-1-hadoopmaster.out
hadoopslave2: starting org.apache.spark.deploy.worker.Worker, logging to /usr/lo
cal/spark/logs/spark-hduser-org.apache.spark.deploy.worker.Worker-1-hadoopslave2
.out
hadoopslave1: starting org.apache.spark.deploy.worker.Worker, logging to /usr/lo
cal/spark/logs/spark-hduser-org.apache.spark.deploy.worker.Worker-1-hadoopslave1
.out
hadoopslave2: starting org.apache.spark.deploy.worker.Worker, logging to /usr/lo
cal/spark/logs/spark-hduser-org.apache.spark.deploy.worker.Worker-2-hadoopslave2
.out
hadoopslave1: starting org.apache.spark.deploy.worker.Worker, logging to /usr/lo
cal/spark/logs/spark-hduser-org.apache.spark.deploy.worker.Worker-2-hadoopslave1
.out
hduser@hadoopmaster:~$
```

- 可由上圖得知，共啟動了 4 個 worker
- 執行 Spark-shell，輸入 spark-shell --master spark://hadoopmaster:7077

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ spark-shell --master spark://hadoopmaster:7077
16/12/09 20:13:14 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/12/09 20:13:16 WARN SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '2').
This is deprecated in Spark 1.0+.

Please instead use:
 - ./spark-submit with --num-executors to specify the number of executors
 - Or set SPARK_EXECUTOR_INSTANCES
 - spark.executor.instances to configure the number of instances in the spark co
nfig.

16/12/09 20:13:34 WARN SparkContext: Use an existing SparkContext, some configur
ation may not take effect.
Spark context Web UI available at http://192.168.59.137:4040
Spark context available as 'sc' (master = spark://hadoopmaster:7077, app id = ap
p-20161209201330-0001).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- 查看 Spark Standalone WebUI，於瀏覽器輸入 http://hadoopmaster:8080



- 讀取本地檔案
  i. val textFile=sc.textFile("file:/usr/local/spark/README.md")
  ii. textFile.count

```
scala> val textFile=sc.textFile("file:/usr/local/spark/README.md")
textFile: org.apache.spark.rdd.RDD[String] = file:/usr/local/spark/README.md MapPartitionsRDD
[1] at textFile at <console>:24

scala> textFile.count
res0: Long = 99
```

- 讀取 HDFS 的檔案(假設在 HDFS 上有一個/user/hduser/test/README.txt

的檔案)

i. val textFile=sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/README.txt")

ii. textFile.count

```
scala> val textFile=sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/README.txt")
textFile: org.apache.spark.rdd.RDD[String] = hdfs://hadoopmaster:9000/user/hduser/test/README
.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> textFile.count
res0: Long = 31
```

- 停止 Spark Standalone cluster，輸入 /usr/local/spark/sbin/stop-all.sh

```
hduser@hadoopmaster:~$ /usr/local/spark/sbin/stop-all.sh
hadoopslave1: stopping org.apache.spark.deploy.worker.Worker
hadoopslave2: stopping org.apache.spark.deploy.worker.Worker
hadoopslave1: stopping org.apache.spark.deploy.worker.Worker
hadoopslave2: stopping org.apache.spark.deploy.worker.Worker
stopping org.apache.spark.deploy.master.Master
```

5. 在 Hadoop YARN 執行 spark

- 執行 YARN 需指定 HADOOP_CONF_DIR 參數，輸入 sudo gedit ~/.bashrc，加入下列參數

```
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
```

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ sudo gedit ~/.bashrc
Open ▾   .bashrc ~/   Save
fi

#Hadoop Variables
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
```

- 讓~/.bashrc 生效，輸入 source ~/.bashrc
- 在 YARN 上執行 spark-shell，輸入/usr/local/spark/bin/spark-shell --master yarn --deploy-mode client

```
hduser@hadoopmaster: ~
hduser@hadoopmaster:~$ /usr/local/spark/bin/spark-shell --master yarn --deploy-m
ode client
```

```
Spark context Web UI available at http://192.168.59.137:4040
Spark context available as 'sc' (master = yarn, app id = application_14827350680
75_0002).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_111)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- 測試讀取本機資料，輸入
  i. val textFile=sc.textFile("file:/usr/local/spark/README.md")
  ii. textFile.count

```
scala> val textFile=sc.textFile("file:/usr/local/spark/README.md")
textFile: org.apache.spark.rdd.RDD[String] = file:/usr/local/spark/README.md Map
PartitionsRDD[1] at textFile at <console>:24

scala> textFile.count
res0: Long = 99
```

- 測試讀取 HDFS 資料，輸入
  i. val textFile=sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/README.txt")
  ii. textFile.count

```
scala> val textFile=sc.textFile("hdfs://hadoopmaster:9000/user/hduser/test/READM
E.txt")
textFile: org.apache.spark.rdd.RDD[String] = hdfs://hadoopmaster:9000/user/hduse
r/test/README.txt MapPartitionsRDD[3] at textFile at <console>:24

scala> textFile.count
res1: Long = 31
```

- 到 Hadoop Web 介面查看