CS522 – Big Data

SPARK PROJECT

Quoc Anh Khuong – 985391

Hien Vo – 985433

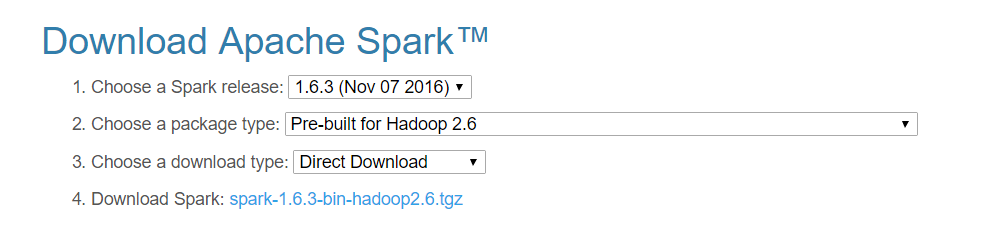
Maharishi University of Management

Department of Computer Science

**PART 5: SPART PROJECT**

# 1. INSTALL Apache Spark

1.1 Download Apache Spark at link: http://spark.apache.org/downloads.html



Spark version 1.6.3 already includes Scala 2.10.5 binaries.

1.2 Unzip Spark and setup Environment Variable

Unzip Spark:

$tar xfv spark-1.6.3-bin-hadoop2.6.tgz

Change Spark directory name and move to /usr/local

$mv spark-1.6.3-bin-hadoop2.6 spark

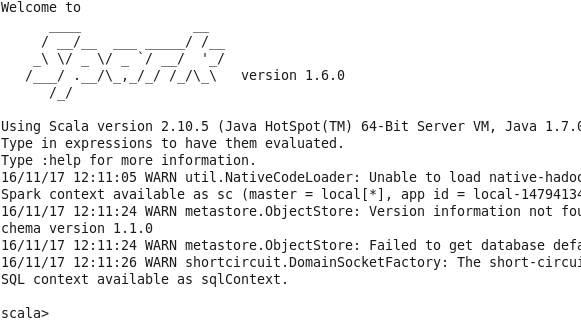
$sudo mv spark /usr/local

Add Spark directory path to environment variable

# In Centos Linux OS, add lines in file “ ~/.bash\_profile” (~ point to home directory of user)

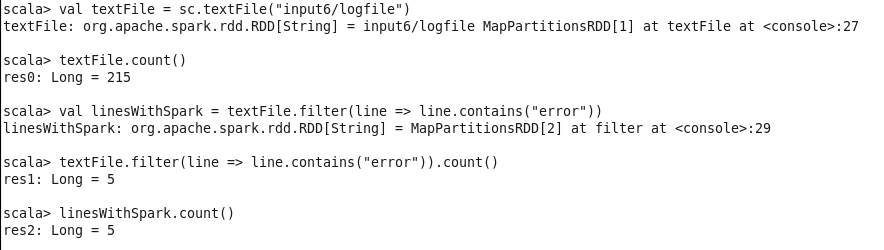


Try to run spark-shell by: $spark-shell



# 2. Run Spark Scala Example

2.1 Run Scala command at prompt of Spark-shell



2.2 Run Scala source file

Create “example.scala” file with content

|  |
| --- |
| val textFile = sc.textFile("input6/logfile")  textFile.count()  textFile.filter(line => line.contains("error")).count()  System.exit(0) |

Run “example.scala”: $spark-shell -i example.scala

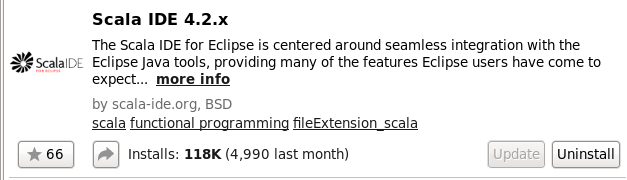
See the result:



3. Scala Spark Project to Analyze Apache Log in Eclipse

3.1 Install Scala IDE and Maven Plugins

- Go to Help -> Eclipse MarketPlace, search and install Scala IDE and Maven plugins to Eclipse



3.2 Create Spark project

- New Maven Project in Eclipse

- Add dependencies for Scala and Spark as below



* Change src/main/java -> src/main/scala
* Add new source file scala “LogProcessor.scala” as follows

New -> Others -> Scala Wizards -> Scala Object

* After finish code, export jar file “sparkproject.jar”
* Run sparkproject.jar as command:

$spark-submit --class "spark.project.LogProcessor" --master local[4] sparkproject.jar input6/apache\_access.log