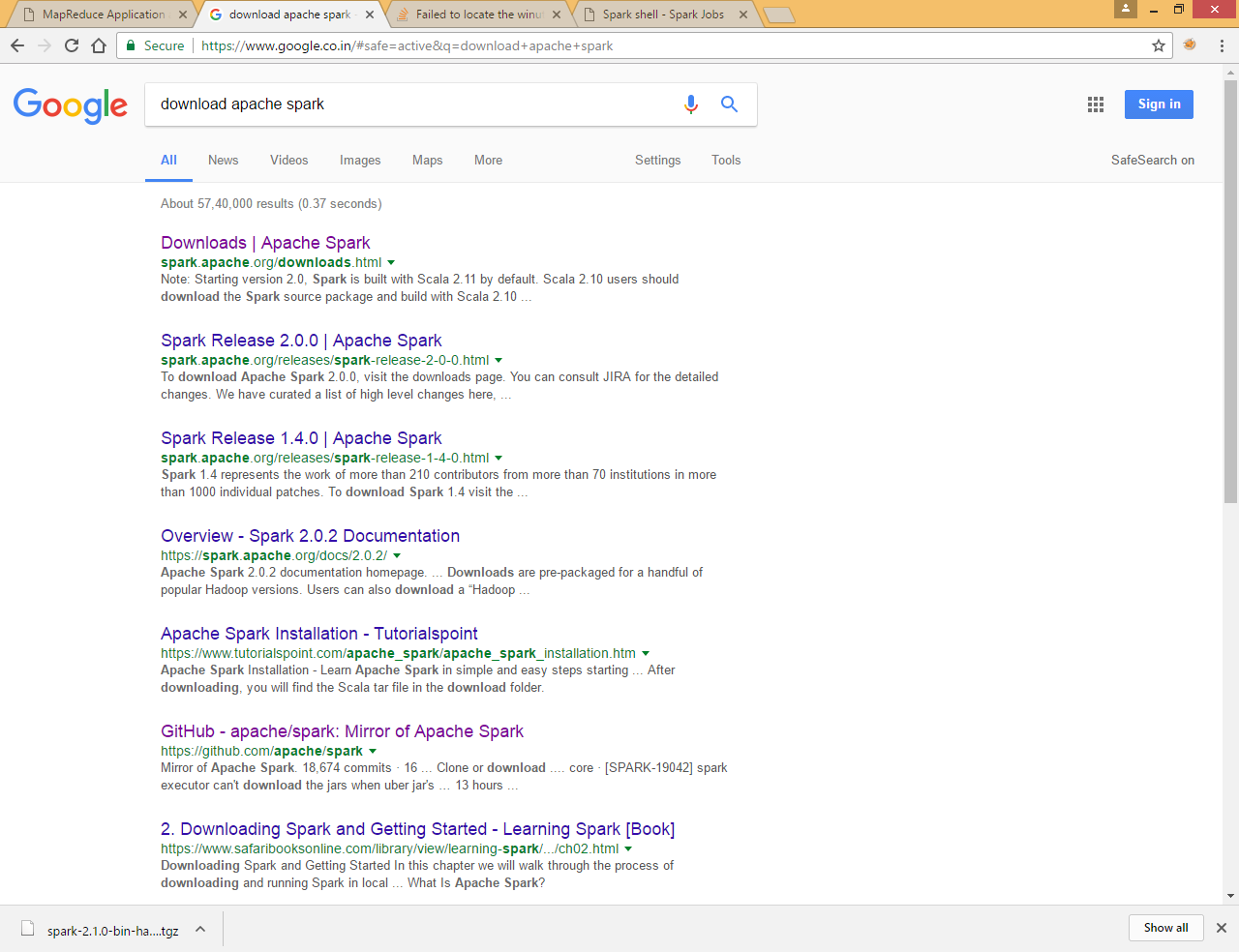
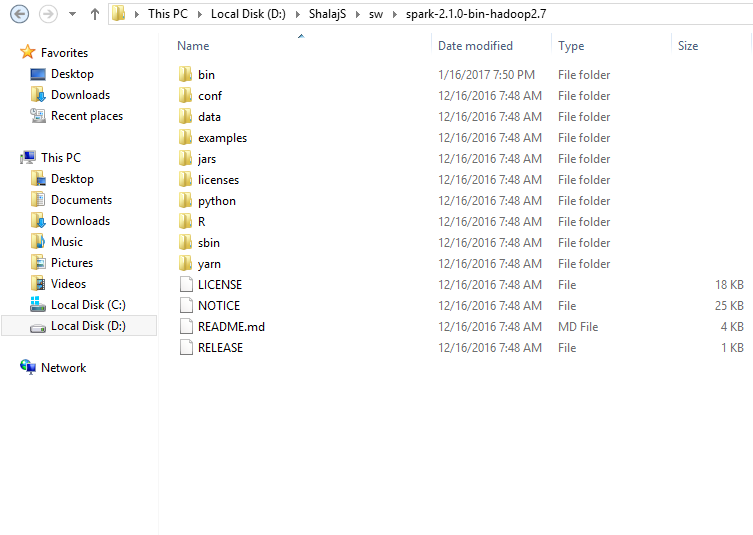
# Download and Configure Spark

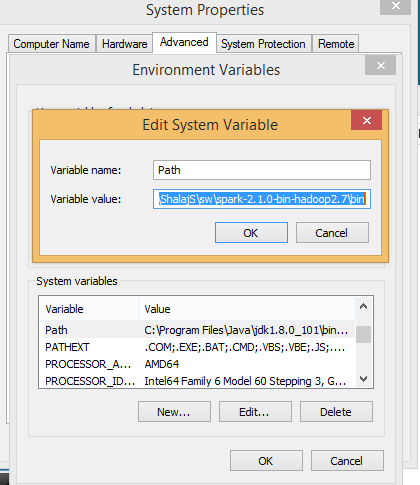




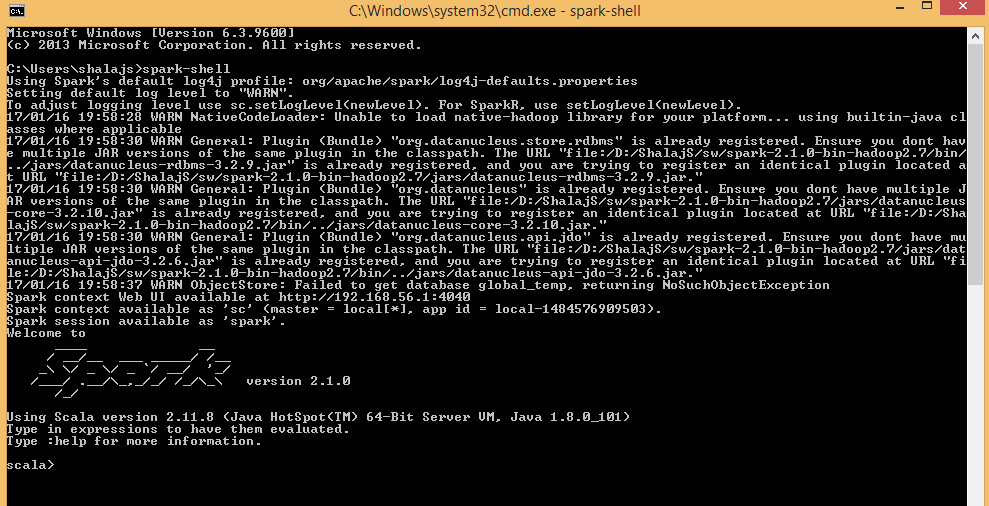
Extract file



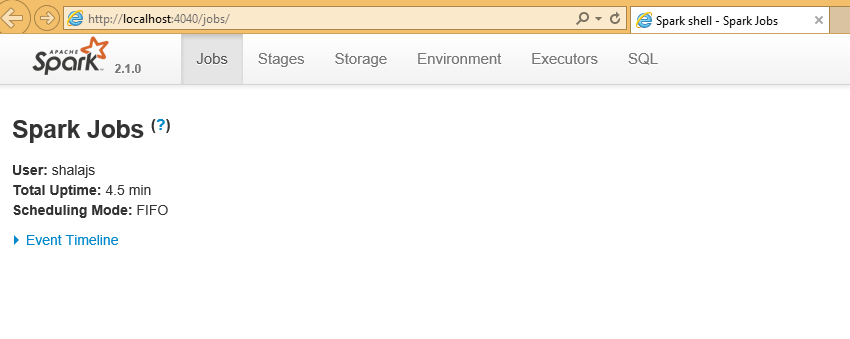
Add bin folder in path environment variable



Open command prompt and run spark-shell command



You can access web UI <http://localhost:4040>



# Prerequisites in windows

To run spark application on eclipse you need to have java, scala and maven installed on computer,

Download winutils.exe binary from <https://github.com/steveloughran/winutils> repository.

You should select the version of Hadoop the Spark distribution was compiled with, e.g. use hadoop-2.7.1 for Spark 2

Save winutils.exe binary to a directory of your choice, e.g. c:\hadoop\bin.

Set HADOOP\_HOME to reflect the directory with winutils.exe (without bin).

|  |
| --- |
| set HADOOP\_HOME=c:\hadoop |

Set PATH environment variable to include %HADOOP\_HOME%\bin as follows:

|  |
| --- |
| set PATH=%HADOOP\_HOME%\bin;%PATH% |

Or you can set these variables in environment variables so that any window program can use them

As a verification step, execute the following line on spark-shell to display the content of a DataFrame:

scala> (0 to 1).toDS.show

+-----+

|value|

+-----+

| 0|

| 1|

+-----+



Now you can create maven project in eclipse for spark

Use pom.xml

|  |
| --- |
| <project xmlns=*"http://maven.apache.org/POM/4.0.0"* xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*  xsi:schemaLocation=*"http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd"*>  <modelVersion>4.0.0</modelVersion>  <groupId>com.spark.tutorial</groupId>  <artifactId>sample-spark-project</artifactId>  <version>0.0.1-SNAPSHOT</version>  <dependencies>  <dependency>  <!-- Spark dependency -->  <groupId>org.apache.spark</groupId>  <artifactId>spark-core\_2.11</artifactId>  <version>1.6.0</version>  </dependency>  </dependencies>  </project> |

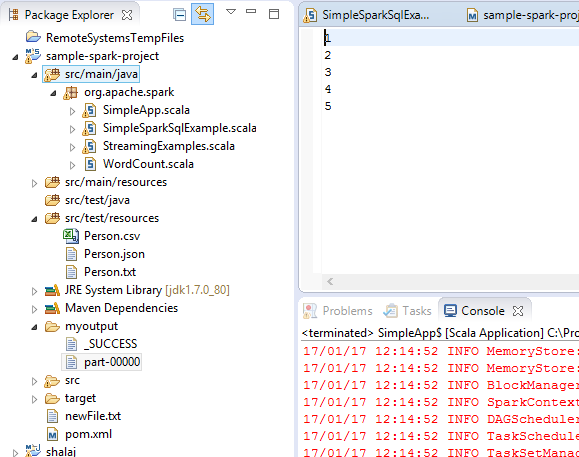
|  |
| --- |
| **package** org.apache.spark  **import** org.apache.spark.SparkConf  **import** org.apache.spark.SparkContext  **object** SimpleApp {  **def** main(args: Array[*String*]) {  **val** conf = **new** SparkConf().setAppName("scala spark").setMaster(args(0))  **val** sc = **new** SparkContext(conf)  **val** i = List(1, 2, 3, 4, 5)  **val** dataRDD = sc.parallelize(i)  dataRDD.saveAsTextFile(args(1))  }  } |

Try to run this program by giving below arguments

|  |
| --- |
| local myoutput |

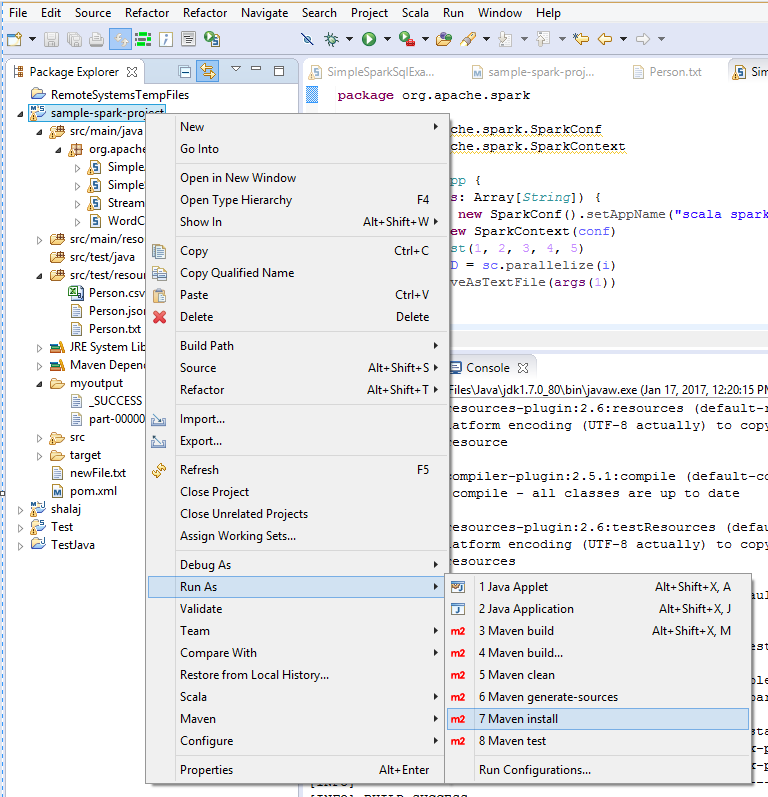
You can provide local[2] in place of local to process it with 2 processors , in that case you will get two part files

Output

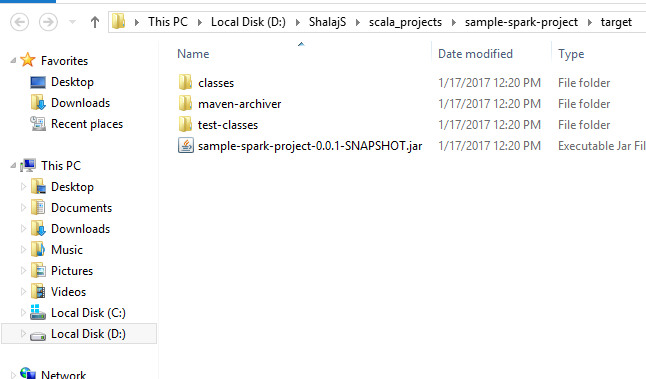


You can create jar file and run the same program in spark environment

Right click on project and click maven install to create jar file

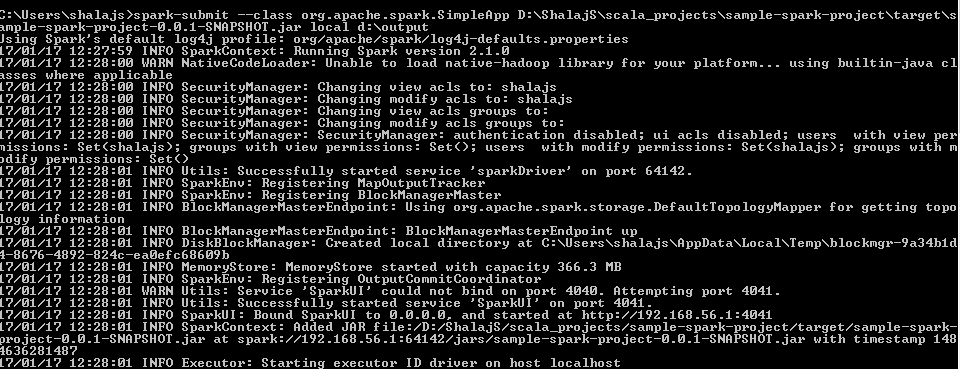


Go to target folder and check the jar file



Now run following command in command prompt

|  |
| --- |
| spark-submit --class org.apache.spark.SimpleApp D:\ShalajS\scala\_projects\sample-spark-project\target\sample-spark-project-0.0.1-SNAPSHOT.jar local d:\output |



Check the output folder

