

빅데이터 플랫폼 머신러닝 개발을 위한

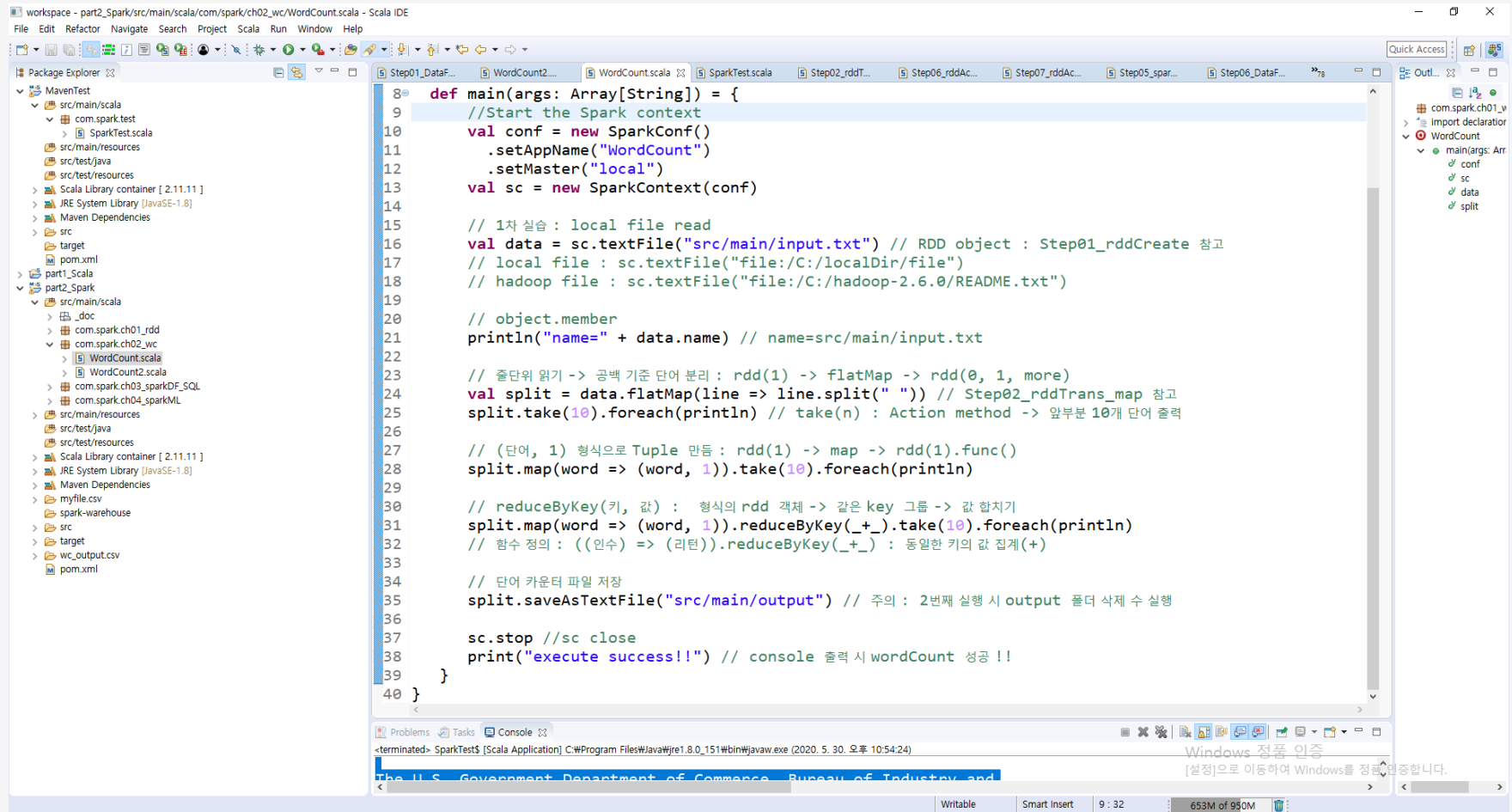
Spark RDD App & Hadoop

작성자 : 김진성

1. WordCount.scala 파일 작성

클래스 импорт

def main() 함수 작성



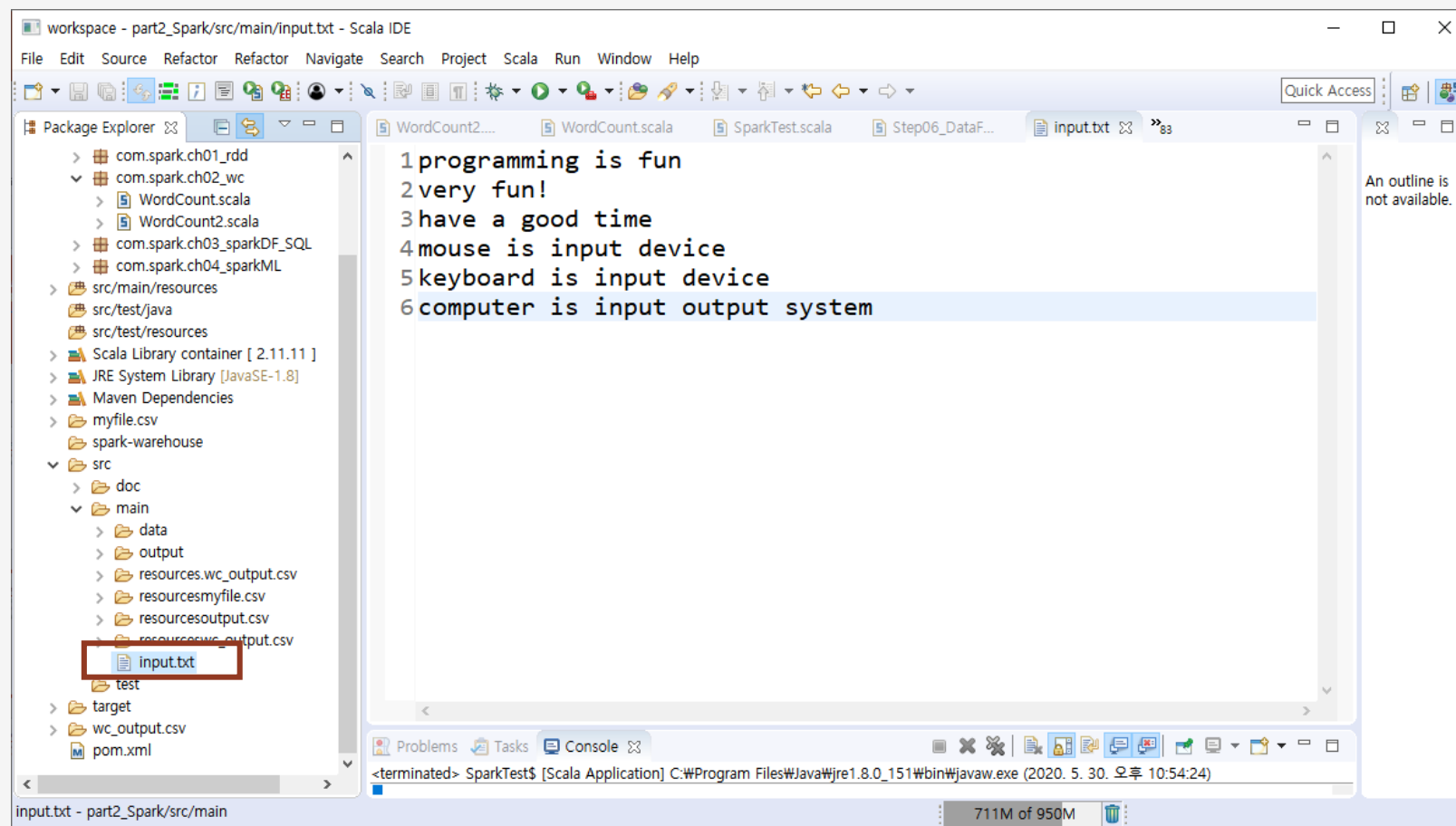
```
8 def main(args: Array[String]) = {
9     //Start the Spark context
10    val conf = new SparkConf()
11        .setAppName("WordCount")
12        .setMaster("local")
13    val sc = new SparkContext(conf)
14
15    // 1차 실습 : local file read
16    val data = sc.textFile("src/main/input.txt") // RDD object : Step01_rddCreate 참고
17    // local file : sc.textFile("file:/C:/localDir/file")
18    // hadoop file : sc.textFile("file:/C:/hadoop-2.6.0/README.txt")
19
20    // object.member
21    println("name=" + data.name) // name=src/main/input.txt
22
23    // 줄단위 읽기 -> 공백 기준 단어 분리 : rdd(1) -> flatMap -> rdd(0, 1, more)
24    val split = data.flatMap(line => line.split(" ")) // Step02_rddTrans_map 참고
25    split.take(10).foreach(println) // take(n) : Action method -> 앞부분 10개 단어 출력
26
27    // (단어, 1) 형식으로 Tuple 만들 : rdd(1) -> map -> rdd(1).func()
28    split.map(word => (word, 1)).take(10).foreach(println)
29
30    // reduceByKey(키, 값) : 형식의 rdd 객체 -> 같은 key 그룹 -> 값 합치기
31    split.map(word => (word, 1)).reduceByKey(_+_).take(10).foreach(println)
32    // 함수 정의 : ((인수) => (리턴)).reduceByKey(_+_) : 동일한 키의 값 집계(+)
33
34    // 단어 카운터 파일 저장
35    split.saveAsTextFile("src/main/output") // 주의 : 2번째 실행 시 output 폴더 삭제 수 실행
36
37    sc.stop //sc close
38    print("execute success!!") // console 출력 시 wordCount 성공 !!
39 }
40 }
```

2. input.txt 파일 배치

src/main 폴더에

프로그램에 이용되는

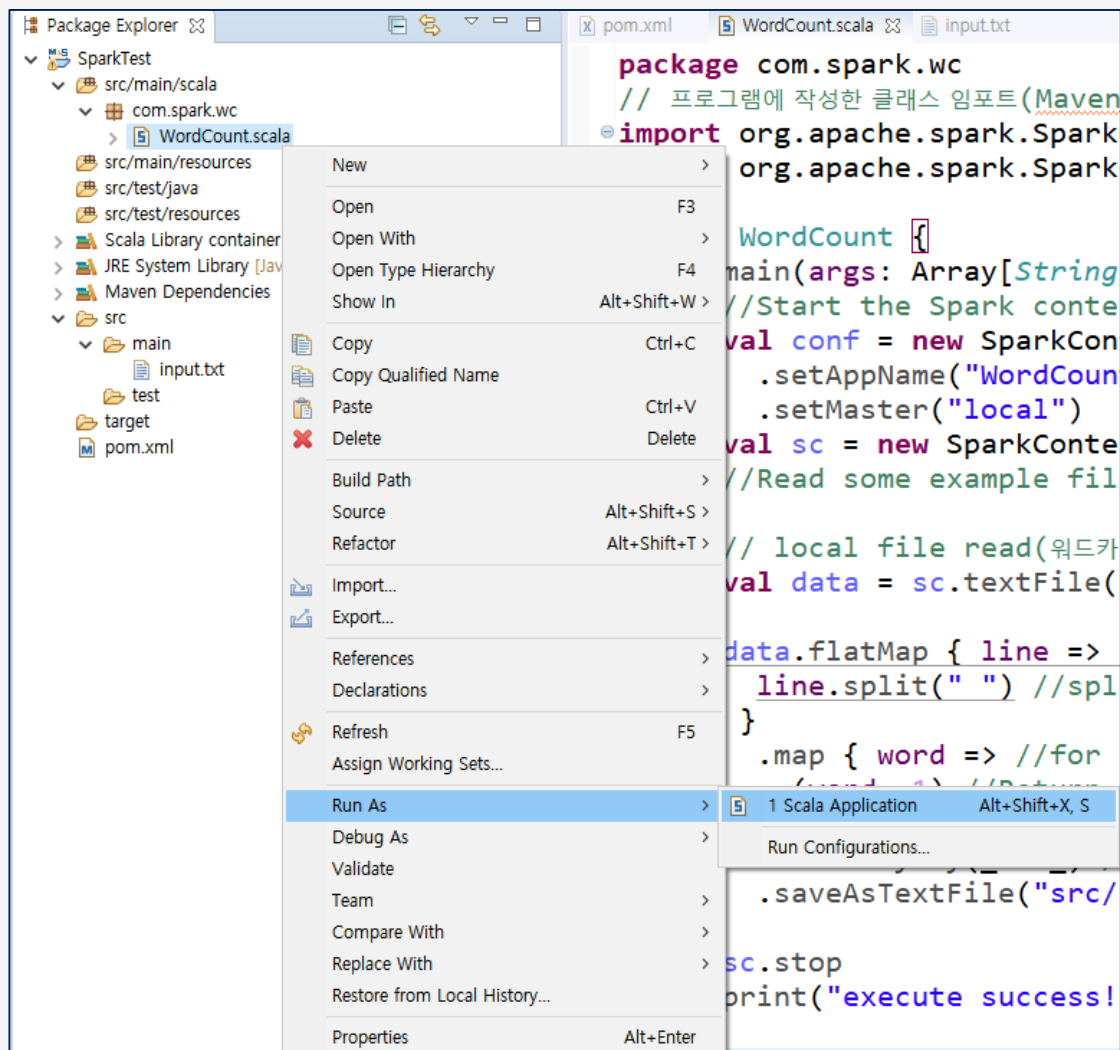
Input.txt 파일 배치



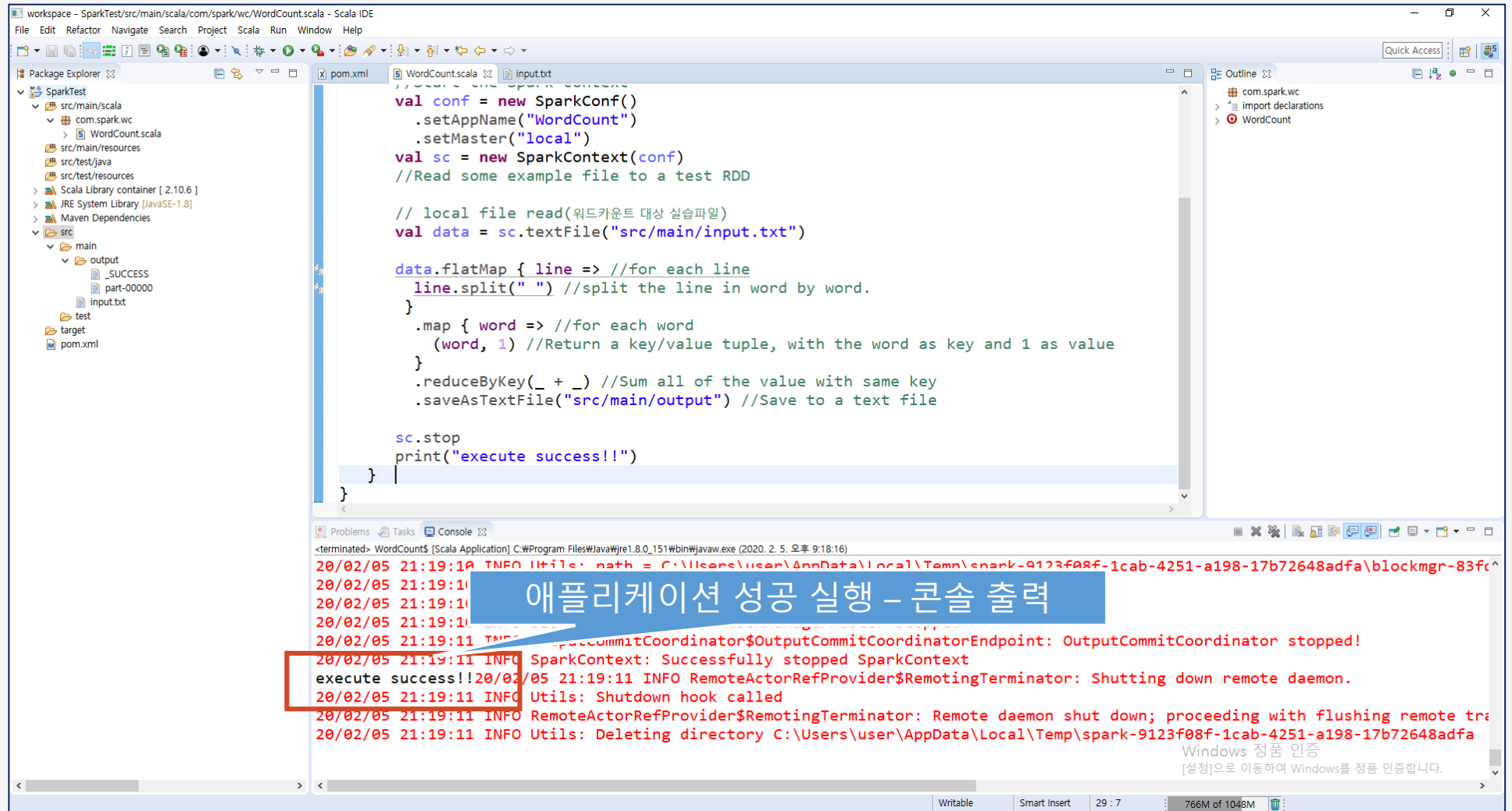
3. Scala Application 실행

WordCount.scala 파일에서

Run Aa -> 1.Scala Application 선택



4. Scala Application 실행 : 콘솔 출력 결과



The screenshot displays an IDE window titled "workspace - SparkTest/src/main/scala/com/spark/wc/WordCount.scala - Scala IDE". The main editor shows the Scala code for WordCount.scala, which reads a file, processes it with Spark, and saves the output. The Package Explorer on the left shows the project structure, including the src/main directory and its subdirectories. The Console window at the bottom shows the execution output, which includes the message "execute success!!" highlighted by a red box. A blue callout box with the text "애플리케이션 성공 실행 - 콘솔 출력" points to this message. The console also shows various INFO messages from the Spark framework, including the shutdown of the remote daemon and the deletion of the temporary directory.

```
val conf = new SparkConf()
    .setAppName("WordCount")
    .setMaster("local")
val sc = new SparkContext(conf)
//Read some example file to a test RDD

// local file read(워드카운트 대상 실습파일)
val data = sc.textFile("src/main/input.txt")

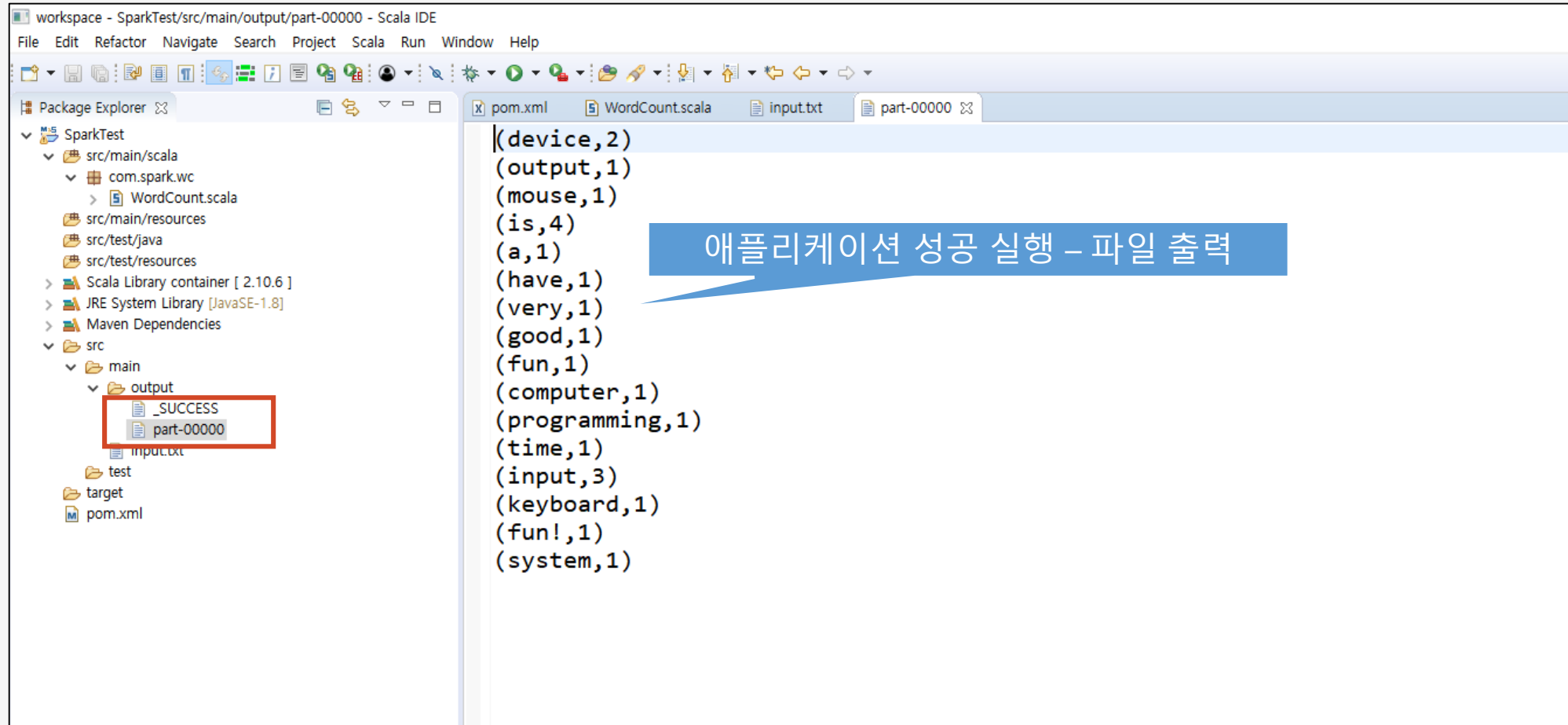
data.flatMap { line => //for each line
    line.split(" ") //split the line in word by word.
}
    .map { word => //for each word
        (word, 1) //Return a key/value tuple, with the word as key and 1 as value
    }
    .reduceByKey(_ + _) //Sum all of the value with same key
    .saveAsTextFile("src/main/output") //Save to a text file

sc.stop
print("execute success!!")
}
```

20/02/05 21:19:10 INFO Utils: path = C:\Users\user\AppData\Local\Temp\spark-9123f08f-1cab-4251-a198-17b72648adfa\blockmgr-83fc
20/02/05 21:19:11 INFO SparkContext: Successfully stopped SparkContext
execute success!! 20/02/05 21:19:11 INFO RemoteActorRefProvider\$RemotingTerminator: Shutting down remote daemon.
20/02/05 21:19:11 INFO Utils: Shutdown hook called
20/02/05 21:19:11 INFO RemoteActorRefProvider\$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote tra
20/02/05 21:19:11 INFO Utils: Deleting directory C:\Users\user\AppData\Local\Temp\spark-9123f08f-1cab-4251-a198-17b72648adfa

Windows 제품 인증
[설정]으로 이동하여 Windows를 제품 인증합니다.

5. Scala Application 실행 : 파일 출력 결과



❖ 애플리케이션 재 실행 시 기존 output 폴더 삭제 후 실행

6. Hadoop + Spark Application 실행

1) Hadoop 기동

```
C:\Users\User>
C:\Users\User>start-all.cmd
This script is deprecated. Instead use start-hadoop.cmd
'C:\Program'은(는) 내부 또는 외부 명령, 실행 가능 파일 또는
배치 파일이 아닙니다.
'C:\Program'은(는) 내부 또는 외부 명령, 실행 가능 파일 또는
배치 파일이 아닙니다.
starting yarn daemons
'C:\Program'은(는) 내부 또는 외부 명령, 실행 가능 파일 또는
배치 파일이 아닙니다.
C:\Users\User>
```

Apache Hadoop Distribution - hadoop datanode

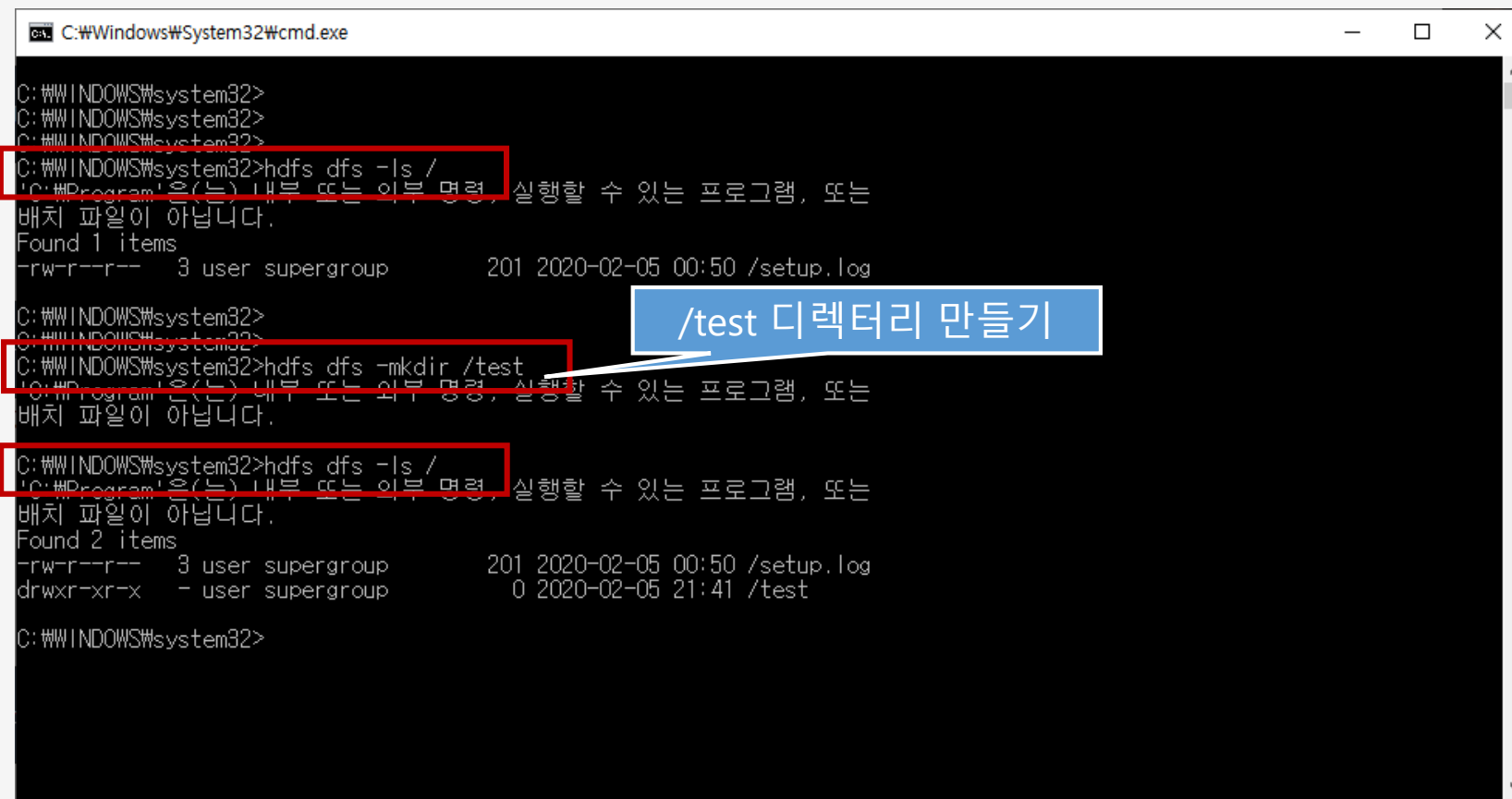
Apache Hadoop Distribution - hadoop namenode

Apache Hadoop Distribution - yarn resourcemanager

Apache Hadoop Distribution - yarn nodemanager

```
20/2/5 00:05:59 INFO http.HttpRequestLog: Http request log for http.requests.nodemanager is not defined
20/2/5 00:05:59 INFO http.HttpServer2: Added global filter 'safety' (class=org.apache.hadoop.http.HttpServer2$SafetyFilter)
20/2/5 00:05:59 INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context node
20/2/5 00:05:59 INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context static
20/2/5 00:05:59 INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context logs
20/2/5 00:05:59 INFO http.HttpServer2: adding path spec: /node/*
20/2/5 00:05:59 INFO http.HttpServer2: adding path spec: /ws/*
20/2/5 00:05:59 INFO http.HttpServer2: Jetty bound to port 8042
20/2/5 00:05:59 INFO mortbay.log: jetty-6.1.26
20/2/5 00:05:59 INFO mortbay.log: Extract jar:file:/C:/hadoop-2.6.0/share/hadoop/yarn/hadoop-yarn-common-2.6.0.jar!/webapps/node to C:\Users\User\AppData\Local\Temp\Jetty_0_0_0_8042_node_19tj0xwebapp
20/2/5 00:06:00 INFO mortbay.log: Started HttpServer2$SelectChannelConnectorWithSafeStartup@0.0.0.0:8042
20/2/5 00:06:00 INFO webapp.WebApps: Web app /node started at 8042
20/2/5 00:06:00 INFO webapp.WebApps: Registered webapp guice modules
20/2/5 00:06:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031
20/2/5 00:06:01 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
20/2/5 00:06:01 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers: []
20/2/5 00:06:01 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id 1284112573
20/2/5 00:06:01 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id 832695464
20/2/5 00:06:01 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as LAPTOP-7JT8HEBE:56723 with total resource of <memory:8192, vCores:8>
20/2/5 00:06:01 INFO nodemanager.NodeStatusUpdaterImpl: Notifying ContainerManager to unblock new container-requests
```

2) data 디렉터리 만들기



```
C:\Windows\System32\cmd.exe
C:\Windows\system32>
C:\Windows\system32>
C:\Windows\system32>
C:\Windows\system32>hdfs dfs -ls /
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 1 items
-rw-r--r--  3 user supergroup      201 2020-02-05 00:50 /setup.log

C:\Windows\system32>
C:\Windows\system32>
C:\Windows\system32>hdfs dfs -mkdir /test
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.

C:\Windows\system32>hdfs dfs -ls /
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 2 items
-rw-r--r--  3 user supergroup      201 2020-02-05 00:50 /setup.log
drwxr-xr-x  - user supergroup      0 2020-02-05 21:41 /test

C:\Windows\system32>
```

/test 디렉터리 만들기

3) data file 올리기

```
C:\Windows\System32\cmd.exe
C:\Windows\system32>
C:\Windows\system32>cd C:\hadoop-2.6.0
C:\hadoop-2.6.0>dir
C 드라이브의 볼륨에는 이름이 없습니다.
볼륨 할당 번호: F0AC-9393
C:\hadoop-2.6.0 디렉터리

2020-02-04 오후 11:58 <DIR> .
2020-02-04 오후 11:58 <DIR> ..
2020-02-04 오후 11:19 <DIR> bin
2020-02-04 오후 11:19 <DIR> etc
2020-02-04 오후 11:19 <DIR> include
2020-02-04 오후 11:19 <DIR> libexec
2015-01-20 오전 12:59      15,429 LICENSE.txt
2020-02-05 오전 12:01 <DIR> logs
2015-01-20 오전 12:59      101 NOTICE.txt
2015-01-20 오전 12:59    1,366 README.txt
2020-02-04 오후 11:19 <DIR> sbin
2020-02-04 오후 11:19 <DIR> share
                3개 파일      16,896 바이트
                9개 디렉터리 77,579,915,264 바이트 남음
C:\hadoop-2.6.0>hdfs dfs -put NOTICE.txt /test
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
C:\hadoop-2.6.0>hdfs dfs -ls /test
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 1 items
-rw-r--r--  3 user supergroup      101 2020-02-05 21:44 /test/NOTICE.txt
C:\hadoop-2.6.0>
```

디렉터리 이동

업로드 파일 확인

HDFS file upload

file upload 확인

7. Hadoop + Spark Application 실행

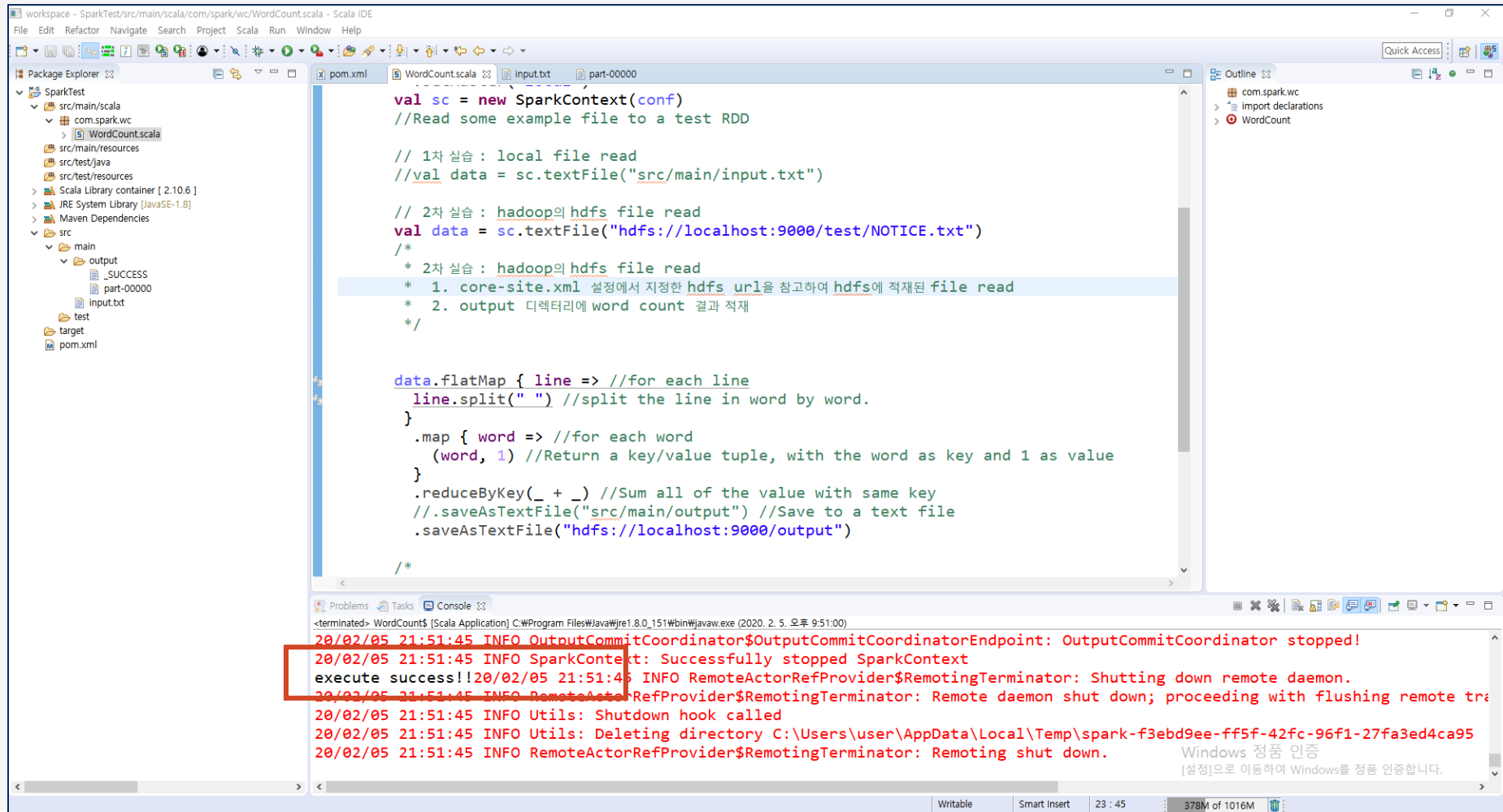
1) Scala Application 수정

```
17 //val data = sc.textFile("src/main/input.txt")
18
19 // 2차 실습 : hadoop의 hdfs file read
20 val data = sc.textFile("hdfs://localhost:9000/test/NOTICE.txt")
21 /*
22  * 실행 전 준비 : hadoop의 hdfs 실행 : ppt 참고
23  *
24  * 2차 실습 : hadoop의 hdfs file read
25  * 1. core-site.xml 설정에서 지정한 hdfs url을 참고하여 hdfs에 적재된 file read
26  * 2. output 디렉터리에 word count 결과 적재
27  */
28
29
30 data.flatMap { line => //for each line
31   line.split(" ") //split the line in word by word.
32 }
33 .map { word => //for each word
34   (word, 1) //Return a key/value tuple,
35 }
36 .reduceByKey(_ + _) //Sum all of the value with
37
38 //saveAsTextFile("src/main/output") //Save to a text file
39 .saveAsTextFile("hdfs://localhost:9000/output")
40
41 /*
```

HDFS에서 읽어올 텍스트 파일

HDFS에 word count 결과 저장 디렉터리

2) Scala Application 실행 : 콘솔 결과



The screenshot displays the Scala IDE interface. The main editor shows the `WordCount.scala` file with the following code:

```
val sc = new SparkContext(conf)
//Read some example file to a test RDD

// 1차 실행 : local file read
//val data = sc.textFile("src/main/input.txt")

// 2차 실행 : hadoop의 hdfs file read
val data = sc.textFile("hdfs://localhost:9000/test/NOTICE.txt")
/*
 * 2차 실행 : hadoop의 hdfs file read
 * 1. core-site.xml 설정에서 지정한 hdfs url을 참고하여 hdfs에 적재된 file read
 * 2. output 디렉터리에 word count 결과 적재
 */

data.flatMap { line => //for each line
  line.split(" ") //split the line in word by word.
}
.map { word => //for each word
  (word, 1) //Return a key/value tuple, with the word as key and 1 as value
}
.reduceByKey(_ + _) //Sum all of the value with same key
//.saveAsTextFile("src/main/output") //Save to a text file
.saveAsTextFile("hdfs://localhost:9000/output")

/*
```

The console output at the bottom shows the execution results:

```
<terminated> WordCount$ (Scala Application) C:\Program Files\Java\jre1.8.0_151\bin\javaw.exe (2020. 2. 5. 오후 9:51:00)
20/02/05 21:51:45 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/02/05 21:51:45 INFO SparkContext: Successfully stopped SparkContext
execute success!!20/02/05 21:51:45 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
20/02/05 21:51:45 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote tra
20/02/05 21:51:45 INFO Utils: Shutdown hook called
20/02/05 21:51:45 INFO Utils: Deleting directory C:\Users\user\AppData\Local\Temp\spark-f3ebd9ee-fff5f-42fc-96f1-27fa3ed4ca95
20/02/05 21:51:45 INFO RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
```

The console output is partially highlighted with a red box, and the status bar at the bottom indicates the application is running on a Windows system.

3) Scala Application 실행 : HDFS 결과

```
C:\Windows\System32\cmd.exe
배치 파일이 아닙니다.

C:\hadoop-2.6.0>hdfs dfs -ls /test
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 1 items
-rw-r--r--  3 user supergroup      101 2020-02-05 21:44 /test/NOTICE.txt

C:\hadoop-2.6.0>hdfs dfs -ls /
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 3 items
drwxr-xr-x  - user supergroup      0 2020-02-05 21:51 /output
-rw-r--r--  3 user supergroup    201 2020-02-05 00:50 /setup.log
drwxr-xr-x  - user supergroup      0 2020-02-05 21:44 /test

C:\hadoop-2.6.0>hdfs dfs -ls /output
'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
Found 2 items
-rw-r--r--  3 user supergroup      0 2020-02-05 21:51 /output/_SUCCESS
-rw-r--r--  3 user supergroup   145 2020-02-05 21:51 /output/part-00000

C:\hadoop-2.6.0>
```

/output 디렉터리 확인

Word count 결과 파일 확인

4) Scala Application 실행 : HDFS 결과

```
C:\Windows\System32\cmd.exe
C:\hadoop-2.6.0>
C:\hadoop-2.6.0>hdfs dfs -cat /output/part-00000
C:\Program (는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.
(includes,1)
(Software,1)
(The,1)
((http://www.apache.org/),,1)
(Apache,1)
(developed,1)
(This,1)
(by,1)
(software,1)
(product,1)
(Foundation,1)
C:\hadoop-2.6.0>
```