

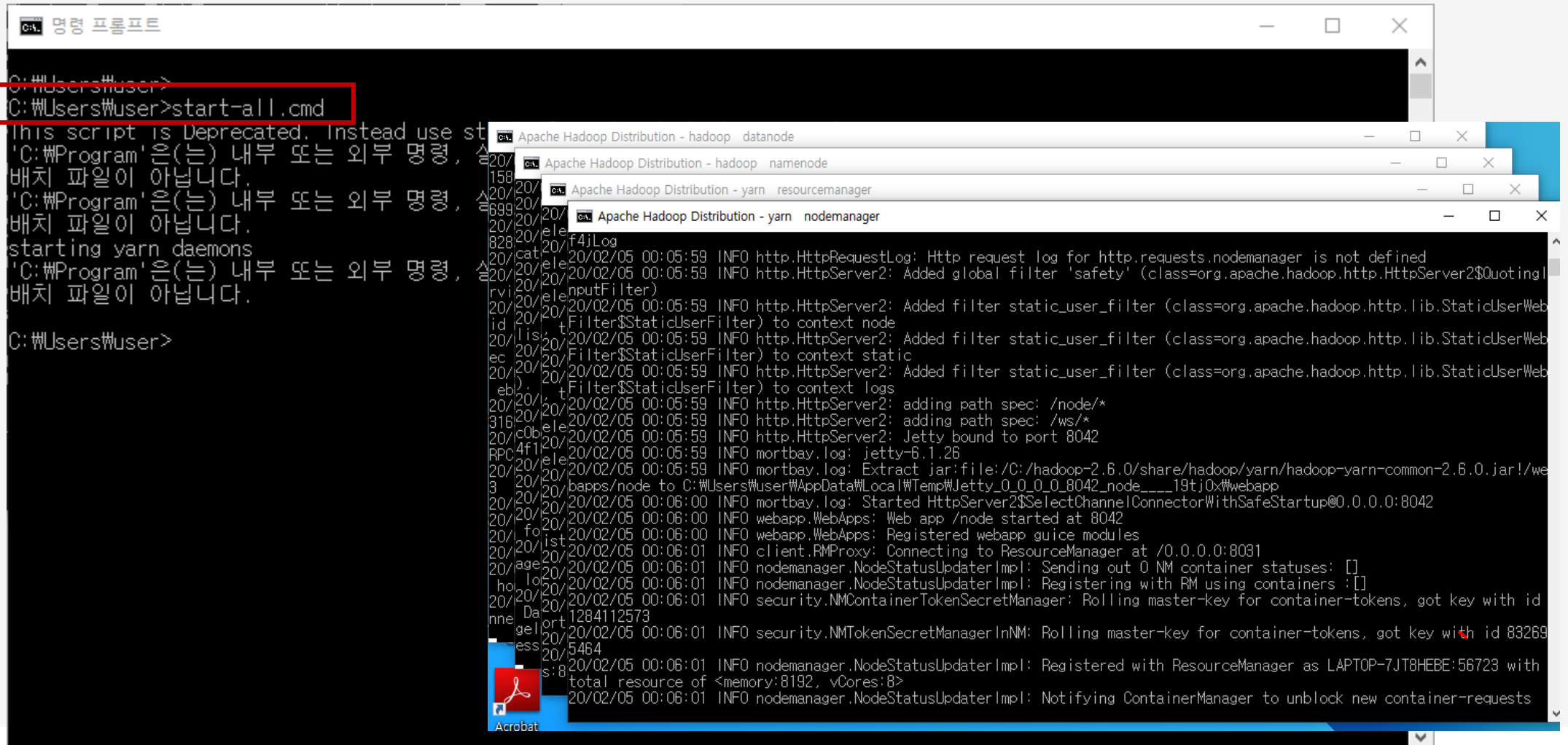
빅데이터 플랫폼 머신러닝 개발을 위한

Spark application & Hadoop 연동

작성자 : 김진성

1. Hadoop 기동 및 파일 업로드

1) Hadoop 기동



The screenshot shows a Windows desktop environment with several open windows. The primary window is a command prompt titled "명령 프롬프트" (Command Prompt) with the following text:

```
C:\Users\User>
C:\Users\User>start-all.cmd
```

The text "This script is Deprecated. Instead use st" is partially visible. Below this, there are several lines of Korean text: "'C:\Program'은(는) 내부 또는 외부 명령, 실행할 수 있는 배치 파일이 아닙니다." ("'C:\Program' is not an internal or external command, batch file or program. "). This is followed by "starting yarn daemons" and another line of the same Korean error message. The prompt ends with "C:\Users\User>".

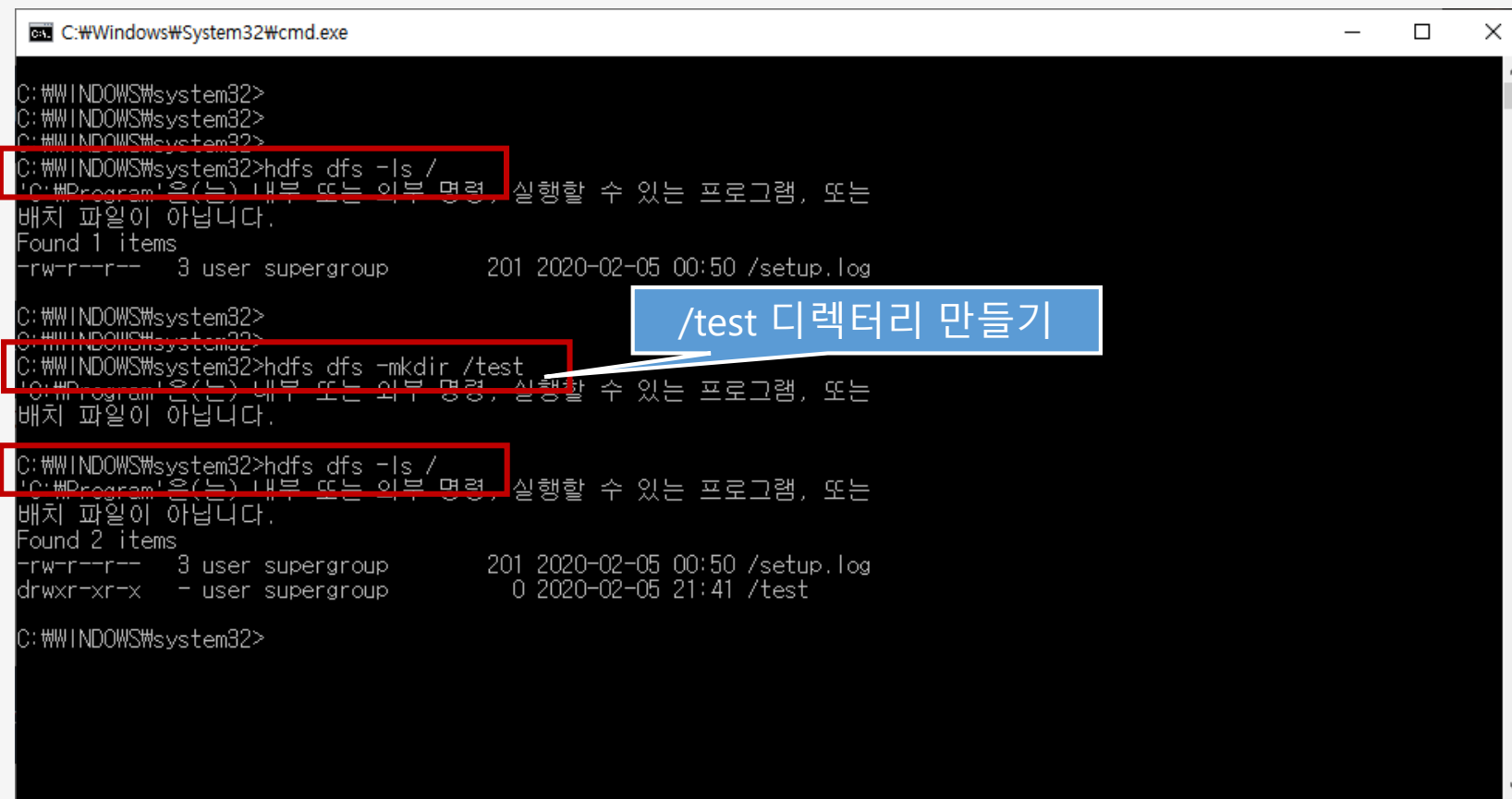
Overlaid on the command prompt are four windows from the "Apache Hadoop Distribution" showing the startup logs for different services:

- hadoop datanode**
- hadoop namenode**
- yarn resourcemanager**
- yarn nodemanager**

The **yarn nodemanager** window displays a detailed log of the startup process, including:

- INFO http.HttpRequestLog: Http request log for http.requests.nodemanager is not defined
- INFO http.HttpServer2: Added global filter 'safety' (class=org.apache.hadoop.http.HttpServer2\$SafetyFilter)
- INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context node
- INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context static
- INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.http.lib.StaticUserWebFilter) to context logs
- INFO http.HttpServer2: adding path spec: /node/*
- INFO http.HttpServer2: adding path spec: /ws/*
- INFO http.HttpServer2: Jetty bound to port 8042
- INFO mortbay.log: jetty-6.1.26
- 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_19tj0x\webapp
- INFO mortbay.log: Started HttpServer2\$SelectChannelConnectorWithSafeStartup@0.0.0.0:8042
- INFO webapp.WebApps: Web app /node started at 8042
- INFO webapp.WebApps: Registered webapp guice modules
- INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8031
- INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
- INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers: []
- INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id 1284112573
- INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id 832695464
- INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as LAPTOP-7JT8HEBE:56723 with total resource of <memory:8192, vCores:8>
- 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

2. SparkTest 어플리케이션 수정

```
package com.spark.test

// Maven에서 제공하는 library 추가
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext

object SparkTest {
  def main(args: Array[String]) = {

    // 1. SparkContext object 생성
    val conf = new SparkConf()
      .setAppName("SparkTest")
      .setMaster("local")

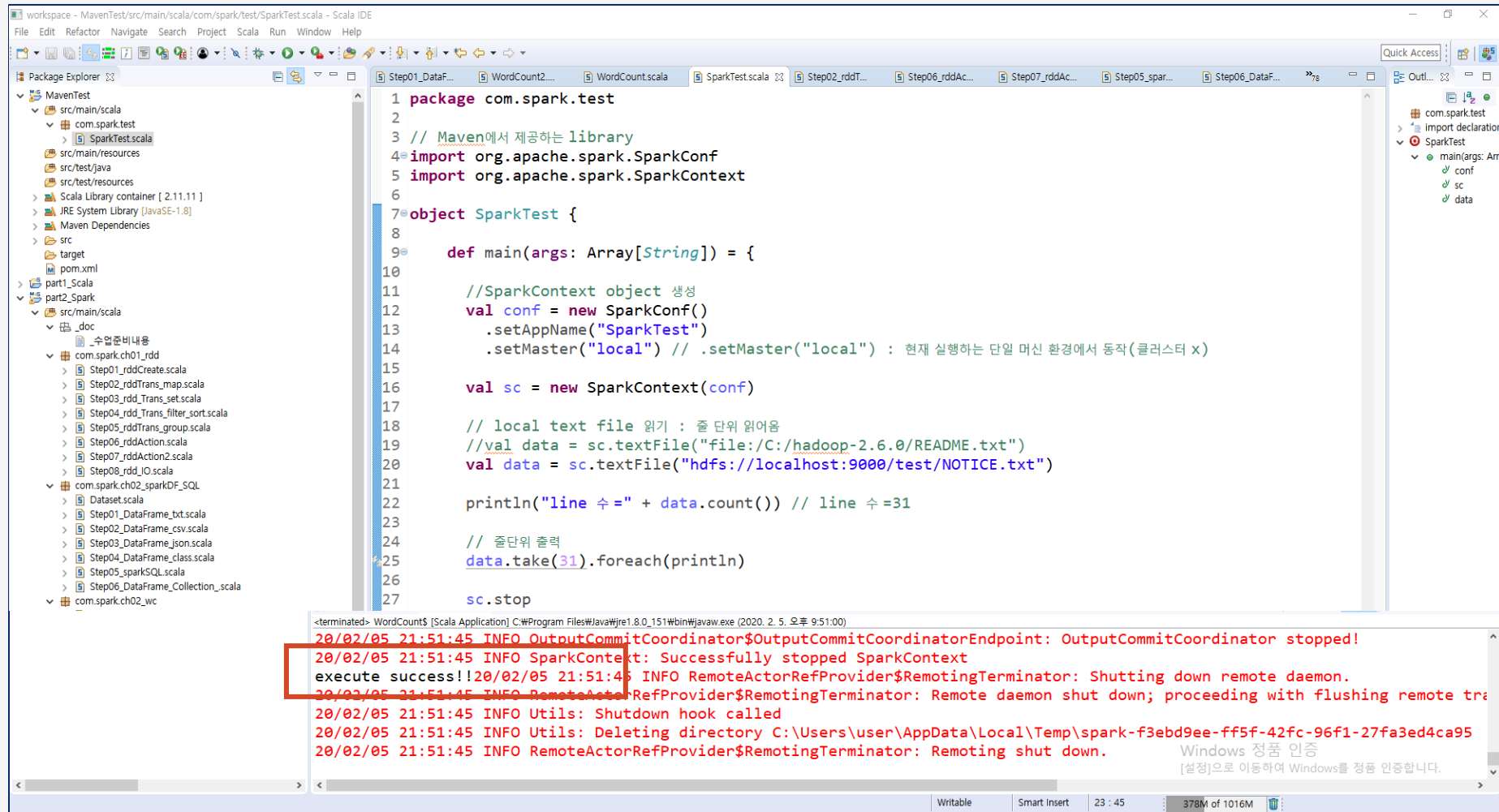
    val sc = new SparkContext(conf)

    // 2. Local or Hadoop file 읽기
    //val data = sc.textFile("file:/C:/hadoop-2.6.0/README.txt")
    val data = sc.textFile("hdfs://localhost:9000/test/NOTICE.txt")
    println("line 수 =" + data.count()) // line 수 =31

    // 3. 줄단위 출력
    data.take(31).foreach(println)
    sc.stop // 객체 닫기
  }
}
```

HDFS 변경

3. Spark Application 실행



The screenshot shows an IDE window titled "workspace - MavenTest/src/main/scala/com/spark/test/SparkTest.scala - Scala IDE". The left sidebar displays the "Package Explorer" with a project structure including "MavenTest", "src/main/scala", "src/main/resources", "src/test/java", "src/test/resources", "Scala Library container [2.11.11]", "JRE System Library [JavaSE-1.8]", "Maven Dependencies", "src", "target", "pom.xml", "part1_Scala", "part2_Spark", "src/main/scala", and "com.spark.ch02_wc". The main editor displays the following Scala code:

```
1 package com.spark.test
2
3 // Maven에서 제공하는 library
4 import org.apache.spark.SparkConf
5 import org.apache.spark.SparkContext
6
7 object SparkTest {
8
9     def main(args: Array[String]) = {
10
11         //SparkContext object 생성
12         val conf = new SparkConf()
13             .setAppName("SparkTest")
14             .setMaster("local") // .setMaster("local") : 현재 실행하는 단일 머신 환경에서 동작 (클러스터 x)
15
16         val sc = new SparkContext(conf)
17
18         // local text file 읽기 : 줄 단위 읽어옴
19         //val data = sc.textFile("file:/C:/hadoop-2.6.0/README.txt")
20         val data = sc.textFile("hdfs://localhost:9000/test/NOTICE.txt")
21
22         println("line 수 =" + data.count()) // line 수 =31
23
24         // 줄단위 출력
25         data.take(31).foreach(println)
26
27         sc.stop
28     }
29 }
```

The right sidebar shows the "Outline" view with a tree structure: "com.spark.test", "import declaration", "SparkTest", "main(args: Array[String])", "conf", "sc", and "data".

The bottom status bar shows the execution logs for "WordCount\$ [Scala Application] C:\Program Files\Java\jre1.8.0_151\bin\javaw.exe (2020. 2. 5. 오후 9:51:00)". The logs include the following messages:

```
<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$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote tra
20/02/05 21:51:45 INFO Utils: Shutdown hook called
20/02/05 21:51:45 INFO Utils: Deleting directory C:\Users\user\AppData\Local\Temp\spark-f3ebd9ee-ff5f-42fc-96f1-27fa3ed4ca95
20/02/05 21:51:45 INFO RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
```

The status bar at the bottom indicates "Writable", "Smart Insert", "23 : 45", and "378M of 1016M".