“大数据工程”课程实验报告

题目：Spark 编程实践	学号姓名：21377061 范春	日期：2024 年 5 月 23 日

实验环境：
虚拟机软件：VMvare
Linux 系统：Ubuntu 操作系统
编程语言：Spark Shell 和 Java

实验内容与完成情况：

1、将数据集上传到 HDFS

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -mkdir week13
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -ls .
Found 7 items
drwxr-xr-x   - hadoop supergroup          0 2024-03-05 22:00 input
drwxr-xr-x   - hadoop supergroup          0 2024-03-05 22:34 output
drwxr-xr-x   - hadoop supergroup          0 2024-03-29 00:05 result1
drwxr-xr-x   - hadoop supergroup          0 2024-03-29 00:21 result2
drwxr-xr-x   - hadoop supergroup          0 2024-03-12 21:16 test
drwxr-xr-x   - hadoop supergroup          0 2024-05-10 09:49 week11
drwxr-xr-x   - hadoop supergroup          0 2024-05-23 13:24 week13
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -put /home/hadoop/桌面/export.csv week13
2024-05-23 13:25:28,721 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remot
eHostTrusted = false
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -ls week13
Found 1 items
-rw-r--r--   1 hadoop supergroup     109949 2024-05-23 13:25 week13/export.csv
```

2、在本机安装和部署 Spark 平台

```
hadoop@u-virtual-machine:~$ sudo tar -zxf /home/hadoop/桌面/spark-2.4.0-bin-with
out-hadoop.tgz -C /usr/local/
[sudo] hadoop 的密码：
hadoop@u-virtual-machine:~$ cd /usr/local
hadoop@u-virtual-machine:/usr/local$ sudo mv ./spark-2.4.0-bin-without-hadoop/ .
/spark
hadoop@u-virtual-machine:/usr/local$ sudo chown -R hadoop:hadoop ./spark
hadoop@u-virtual-machine:/usr/local$ cd /usr/local/spark
hadoop@u-virtual-machine:/usr/local/spark$ cp ./conf/spark-env.sh.template ./con
f/spark-env.sh
hadoop@u-virtual-machine:/usr/local/spark$ gedit ./conf/spark-env.sh
hadoop@u-virtual-machine:/usr/local/spark$ bin/run-example SparkPi 2>&1 | grep "
Pi is"
Pi is roughly 3.1458157290786453
hadoop@u-virtual-machine:/usr/local/spark$
```

3、启动 Spark Shell

```
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.spark.deploy.JavaMainApplication.start(SparkApplication.scala:52)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scal
a:849)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:167)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:195)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:86)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:924)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:933)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)

Spark context Web UI available at http://192.168.150.128:4040
Spark context available as 'sc' (master = local[*], app id = local-1716441612042).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_162)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

4、在 Spark Shell 中读取 HDFS 上的上述文件，然后统计出文件的行数

```
scala> val textFile = sc.textFile("hdfs://localhost:9000/user/hadoop/week13/export.csv")
textFile: org.apache.spark.rdd.RDD[String] = hdfs://localhost:9000/user/hadoop/week13/export.csv MapPartitionsRDD[5] at
textFile at <console>:24

scala> val lineCount = textFile.count()
lineCount: Long = 1001
```

5、Java 编程实践

(1) 安装 Maven

```
hadoop@u-virtual-machine:~$ sudo unzip /home/hadoop/桌面/apache-maven-3.6.3-bin.
zip -d /usr/local
[sudo] hadoop 的密码:
Archive:  /home/hadoop/桌面/apache-maven-3.6.3-bin.zip
   creating: /usr/local/apache-maven-3.6.3/
   creating: /usr/local/apache-maven-3.6.3/lib/
   creating: /usr/local/apache-maven-3.6.3/boot/
   creating: /usr/local/apache-maven-3.6.3/lib/jansi-native/
   creating: /usr/local/apache-maven-3.6.3/lib/jansi-native/freebsd32/
   creating: /usr/local/apache-maven-3.6.3/lib/jansi-native/freebsd64/
```

```
hadoop@u-virtual-machine:~$ cd /usr/local
hadoop@u-virtual-machine:/usr/local$ sudo mv apache-maven-3.6.3/ ./maven
hadoop@u-virtual-machine:/usr/local$ sudo chown -R hadoop ./maven
hadoop@u-virtual-machine:/usr/local$
```

(2) Java 应用程序代码

```java
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

public class SparkLineCount {
    public static void main(String[] args) {
        String inputFile = "hdfs://localhost:9000/user/hadoop/week13/export.csv";

        SparkConf conf = new SparkConf().setAppName("Spark Line Count");
        JavaSparkContext sc = new JavaSparkContext(conf);

        JavaRDD<String> textFile = sc.textFile(inputFile);
        long lineCount = textFile.count();

        System.out.println("文件中有 " + lineCount + " 行");

        sc.close();
    }
}

```

打开(O) pom.xml ~/sparkapp2

```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.example</groupId>
    <artifactId>my-spark-app</artifactId>
    <version>1.0-SNAPSHOT</version>

    <dependencies>
        <!-- Spark Core dependency -->
        <dependency>
            <groupId>org.apache.spark</groupId>
            <artifactId>spark-core_2.12</artifactId>
            <version>3.1.2</version>
        </dependency>

        <!-- Spark SQL dependency (if needed) -->
        <dependency>
            <groupId>org.apache.spark</groupId>
            <artifactId>spark-sql_2.12</artifactId>
            <version>3.1.2</version>
        </dependency>
    </dependencies>

    <build>
        <plugins>
            <plugin>
                <groupId>org.apache.maven.plugins</groupId>
                <artifactId>maven-compiler-plugin</artifactId>
                <version>3.8.1</version>
                <configuration>
                    <source>1.8</source>
                    <target>1.8</target>
                </configuration>
            </plugin>
            <plugin>
                <groupId>org.apache.maven.plugins</groupId>
                <artifactId>maven-shade-plugin</artifactId>
                <version>3.2.4</version>
                <executions>
```

(3) 打包

```
hadoop@u-virtual-machine:/usr/local$ cd ~
hadoop@u-virtual-machine:~$ mkdir -p ./sparkapp2/src/main/java
hadoop@u-virtual-machine:~$ gedit ./sparkapp2/src/main/java/SparkLineCount.java
hadoop@u-virtual-machine:~$ cd ~/sparkapp2
hadoop@u-virtual-machine:~/sparkapp2$ gedit pom.xml
hadoop@u-virtual-machine:~/sparkapp2$ find .
.
./src
./src/main
./src/main/java
./src/main/java/SparkLineCount.java
./pom.xml
hadoop@u-virtual-machine:~/sparkapp2$
```

```
hadoop@u-virtual-machine:~/sparkapp2$ /usr/local/maven/bin/mvn  package
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< cn.edu.xmu:simple-project >----------------------
[INFO] Building Simple Project 1.0
[INFO] --------------------------------[ jar ]---------------------------------
Downloading from aliyunmaven: https://maven.aliyun.com/repository/public/org/apache/maven
/plugins/maven-jar-plugin/3.1.0/maven-jar-plugin-3.1.0.pom
Downloaded from aliyunmaven: https://maven.aliyun.com/repository/public/org/apache/maven/
plugins/maven-jar-plugin/3.1.0/maven-jar-plugin-3.1.0.pom (6.6 kB at 12 kB/s)
Downloading from aliyunmaven: https://maven.aliyun.com/repository/public/org/apache/maven
/plugins/maven-plugins/31/maven-plugins-31.pom
```

```
[INFO] Building jar: /home/hadoop/sparkapp2/target/simple-project-1.0.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  5.703 s
[INFO] Finished at: 2024-05-23T17:24:25+08:00
[INFO] ------------------------------------------------------------------------
```

通过以上结果我们可以看出打包成功。

(4) 通过 spark-submit 运行程序

```
hadoop@u-virtual-machine:~/sparkapp2$ /usr/local/spark/bin/spark-submit --class
"SparkLineCount" ~/sparkapp2/target/simple-project-1.0.jar
```

```
文件中有 1001 行
```