Spark作业

目录：

题目1

使用RDD API实现带词频的倒排索引

倒排索引（Inverted index），也被称为反向索引。它是文档检索系统中最常用的数据结构。被广泛地应用于全文搜索引擎。
例子如下，被索引的文件为（0，1，2代表文件名）

0. "it is what it is"
1. "what is it"
2. "it is a banana"
   我们就能得到下面的反向文件索引：
   "a": {2}
   "banana": {2}
   "is": {0, 1, 2}
   "it": {0, 1, 2}
   "what": {0, 1}
   再加上词频为：
   "a": {(2,1)}
   "banana": {(2,1)}
   "is": {(0,2), (1,1), (2,1)}
   "it": {(0,2), (1,1), (2,1)}
   "what": {(0,1), (1,1)}

源码地址

https://github.com/wanghuan2054/geektime/tree/master/0815SparkAPI

开发环境

1. scala 2.12.12
2. spark 3.1.2
3. hadoop 3.1.3
4. jdk 1.8.182

核心思路

1. 读取指定目录下所有文本文件列表，使用wholeTextFiles方法，可以返回文档名和整篇文档内容的KV对

sc.textFile（）和sc.wholeTextFiles()都支持精确匹配文件和正则匹配文件

sc.textFiles(path) 能将path 里的所有文件内容读出，以文件中的每一行作为一条记录的方式。

wholeTextFiles允许你读取文件夹下所有的文件，比如多个小的文本文件， 返回[(k,v),(k1,v1),....] k代表文件名，v代表整个文件内容。

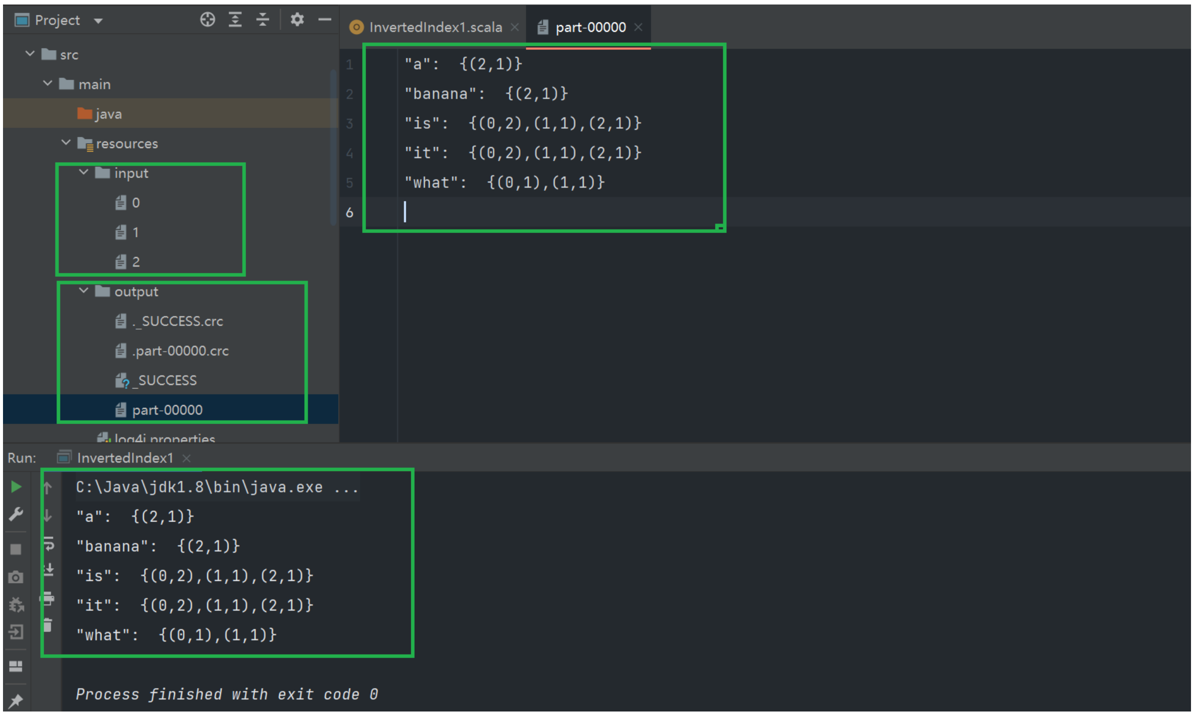
用textFile时，它的partition的数量是与文件夹下的文件数量（实例中用3个文件）相关，一个文件就是一个partition(既然3个文件就是：partition=3，特别提醒：这里的测试数据是几十byte，较小，如果每个文件较大，要根据相应切分原则切分)。
wholeTextFiles的partition数量是根据用户指定或者文件大小来（文件内的数据量少）确定，与hdfs目录下的文件数量无关！ 所以wholeTextFile通常用于读取许多小文件的需求。

2. 从文件全路径名K中， 提取出名字（正则匹配/取出文件名）
3. 将文档内容中的换行符替换成空格，考虑到windows和linux平台换行符不同， 及其它一些特殊的stop words替换清洗
4. 对文档内容切分并压平 ，组装 , 分组聚合， 并按照文档ID 升序排列
5. 将结果格式化成目标输出格式，为避免小文件，采用coalesce(1)缩减分区（coalesce会比repartition减少shuffle次数），并写入最终输出目录中单个文件中。

POM文件

```
<?xml version="1.0" encoding="UTF-8"?>
<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>org.example</groupId>
    <artifactId>SparkAPI</artifactId>
    <version>1.0-SNAPSHOT</version>

    <properties>
        <maven.compiler.source>8</maven.compiler.source>
        <maven.compiler.target>8</maven.compiler.target>
        <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
        <scala.version>2.12</scala.version>
        <scala.binary.version>2.12.12</scala.binary.version>
        <spark.version>3.1.2</spark.version>
        <hadoop.version>3.2</hadoop.version>
    </properties>

    <dependencies>
        <dependency>
            <groupId>org.apache.spark</groupId>
            <artifactId>spark-core_${scala.version}</artifactId>
            <version>${spark.version}</version>
        </dependency>
    </dependencies>

    <build>
        <plugins>
        <!-- 该插件将scala代码编译成class文件 -->
        <plugin>
            <groupId>net.alchim31.maven</groupId>
            <artifactId>scala-maven-plugin</artifactId>
            <version>3.2.2</version>
            <executions>
                <execution>
                    <goals>
                        <goal>compile</goal>
                        <goal>testCompile</goal>
                    </goals>
                </execution>
            </executions>
        </plugin>
        </plugins>
    </build>

</project>
```

本地运行



```
"a":  {(2,1)}
"banana":  {(2,1)}
"is":  {(0,2),(1,1),(2,1)}
"it":  {(0,2),(1,1),(2,1)}
"what":  {(0,1),(1,1)}
```

```
C:\Java\jdk1.8\bin\java.exe ...
"a":  {(2,1)}
"banana":  {(2,1)}
"is":  {(0,2),(1,1),(2,1)}
"it":  {(0,2),(1,1),(2,1)}
"what":  {(0,1),(1,1)}

Process finished with exit code 0
```

集群运行

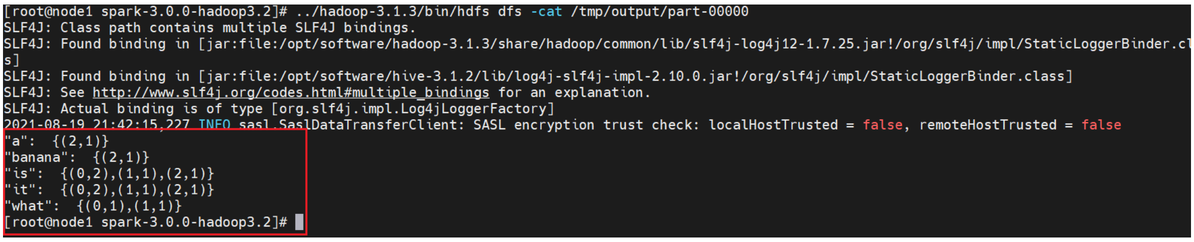
Input文件上传HDFS

```
[root@node1 hadoop-3.1.3]# bin/hdfs dfs -put /home/hadoop/input/* /tmp/input/

# 验证上传是否成功
[root@node1 hadoop-3.1.3]# bin/hdfs dfs -ls /tmp/input/
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/software/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/software/hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 4 items
-rw-r--r--   3 root supergroup         16 2021-08-18 15:57 /tmp/input/0
-rw-r--r--   3 root supergroup         10 2021-08-18 15:57 /tmp/input/1
-rw-r--r--   3 root supergroup         14 2021-08-18 15:57 /tmp/input/2
```

spark-submit提交jar包

```
#  hdfs 路径
[root@node1 hadoop-3.1.3]# bin/spark-submit --master yarn --deploy-mode client --class com.geektime.InvertedIndex1 /home/hadoop/SparkAPI-1.0-SNAPSHOT.jar hdfs:///tmp/input/ hdfs:///tmp/output/

# 本地路径
[root@node1 hadoop-3.1.3]#  bin/spark-submit --master yarn --deploy-mode client --class com.geektime.InvertedIndex1 /home/hadoop/SparkAPI-1.0-SNAPSHOT.jar file:///home/hadoop/input/ file:///home/hadoop/output/
```

运行结果



```
[root@node1 spark-3.0.0-hadoop3.2]# ../hadoop-3.1.3/bin/hdfs dfs -cat /tmp/output/part-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/software/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.cl
s]
SLF4J: Found binding in [jar:file:/opt/software/hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2021-08-19 21:42:15,227 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
"a":  {(2,1)}
"banana":  {(2,1)}
"is":  {(0,2),(1,1),(2,1)}
"it":  {(0,2),(1,1),(2,1)}
"what":  {(0,1),(1,1)}
[root@node1 spark-3.0.0-hadoop3.2]#
```

题目2

Distcp的spark实现
使用Spark实现Hadoop 分布式数据传输工具DistCp (distributed copy)，只要求实现最基础的copy功能，对于-update、-diff、-p不做要求

POM

```
<?xml version="1.0" encoding="UTF-8"?>
<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>org.example</groupId>
    <artifactId>SparkAPI</artifactId>
    <version>1.0-SNAPSHOT</version>

    <properties>
        <maven.compiler.source>8</maven.compiler.source>
        <maven.compiler.target>8</maven.compiler.target>
        <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
        <scala.version>2.12</scala.version>
        <scala.binary.version>2.12.12</scala.binary.version>
        <spark.version>3.1.2</spark.version>
        <hadoop.version>3.1.3</hadoop.version>
    </properties>

    <dependencies>
        <dependency>
            <groupId>org.apache.spark</groupId>
            <artifactId>spark-core_${scala.version}</artifactId>
            <version>${spark.version}</version>
```

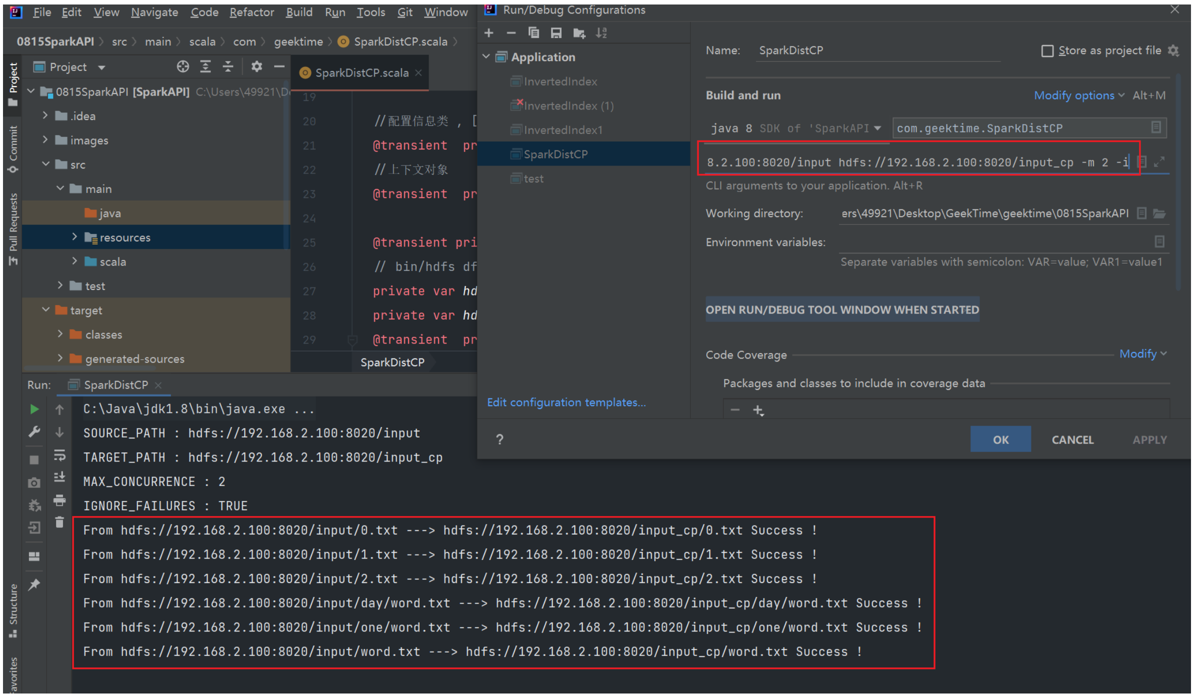
```xml
        </dependency>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-client</artifactId>
            <version>${hadoop.version}</version>
        </dependency>
    </dependencies>

    <build>
        <plugins>
        <!-- 该插件将scala代码编译成class文件 -->
        <plugin>
            <groupId>net.alchim31.maven</groupId>
            <artifactId>scala-maven-plugin</artifactId>
            <version>3.2.2</version>
            <executions>
                <execution>
                    <goals>
                        <goal>compile</goal>
                        <goal>testCompile</goal>
                    </goals>
                </execution>
            </executions>
        </plugin>
        </plugins>
    </build>

</project>
```

源码地址

https://github.com/wanghuan2054/geektime/tree/master/0815SparkAPI

本地运行

IDEA 中连接HDFS 进行DistCp 分布式拷贝



```
//配置信息类 , [
@transient  pr
//上下文对象
@transient  pr

@transient pri
// bin/hdfs df
private var hd
private var hd
@transient  pr
```

Name: SparkDistCP

□ Store as project file

Build and run — Modify options Alt+M

java 8 SDK of 'SparkAPI | com.geektime.SparkDistCP

8.2.100:8020/input hdfs://192.168.2.100:8020/input_cp -m 2 -i

CLI arguments to your application. Alt+R

Working directory: ers\49921\Desktop\GeekTime\geektime\0815SparkAPI

Environment variables:

Separate variables with semicolon: VAR=value; VAR1=value1

OPEN RUN/DEBUG TOOL WINDOW WHEN STARTED

Code Coverage — Modify

Packages and classes to include in coverage data

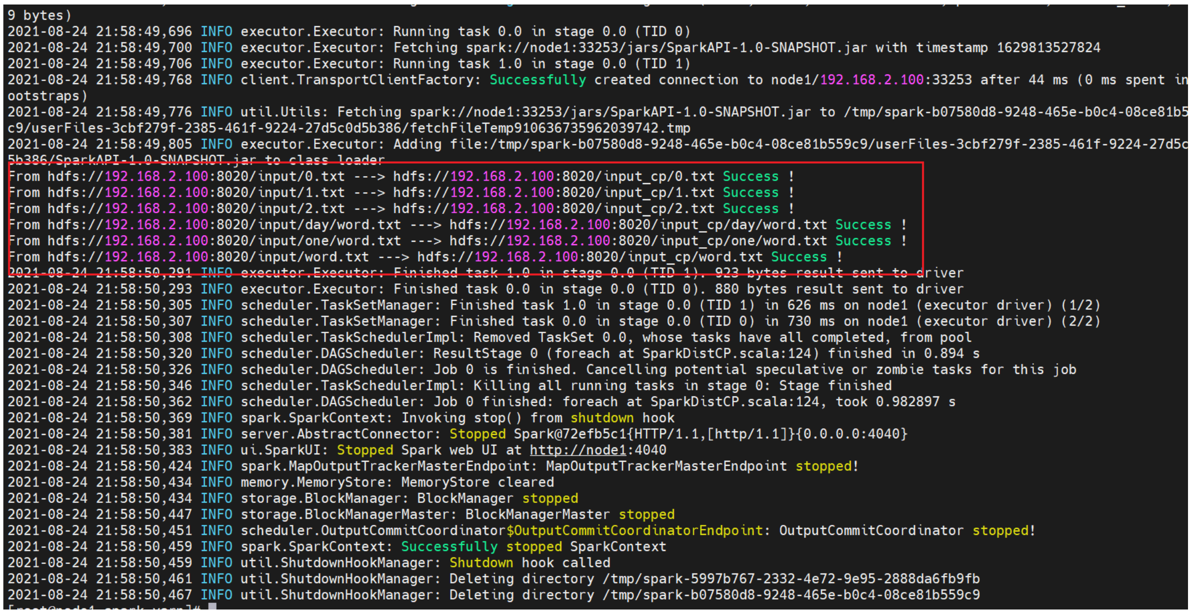
```
C:\Java\jdk1.8\bin\java.exe ...
SOURCE_PATH : hdfs://192.168.2.100:8020/input
TARGET_PATH : hdfs://192.168.2.100:8020/input_cp
MAX_CONCURRENCE : 2
IGNORE_FAILURES : TRUE
From hdfs://192.168.2.100:8020/input/0.txt ---> hdfs://192.168.2.100:8020/input_cp/0.txt Success !
From hdfs://192.168.2.100:8020/input/1.txt ---> hdfs://192.168.2.100:8020/input_cp/1.txt Success !
From hdfs://192.168.2.100:8020/input/2.txt ---> hdfs://192.168.2.100:8020/input_cp/2.txt Success !
From hdfs://192.168.2.100:8020/input/day/word.txt ---> hdfs://192.168.2.100:8020/input_cp/day/word.txt Success !
From hdfs://192.168.2.100:8020/input/one/word.txt ---> hdfs://192.168.2.100:8020/input_cp/one/word.txt Success !
From hdfs://192.168.2.100:8020/input/word.txt ---> hdfs://192.168.2.100:8020/input_cp/word.txt Success !
```

集群运行

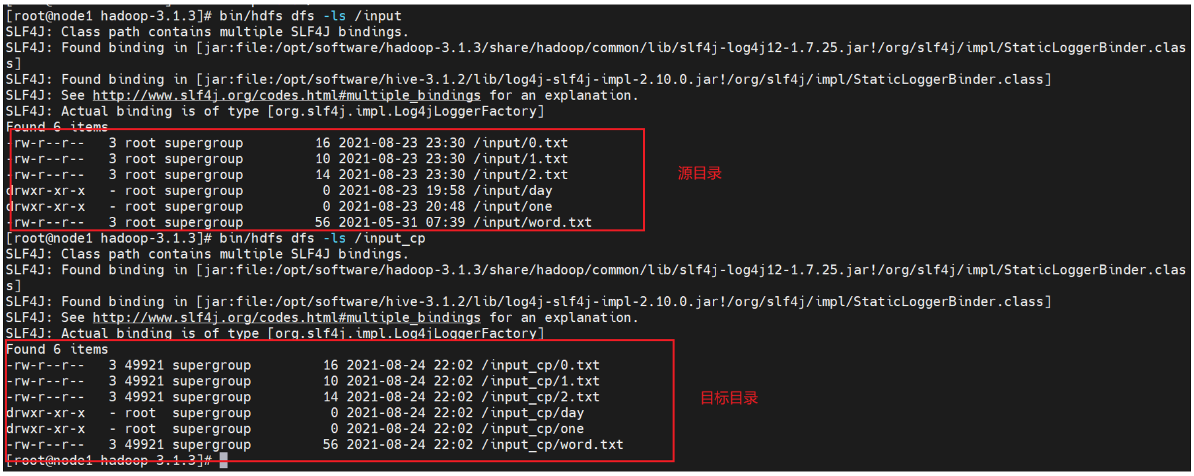
spark-submit提交jar包

```
#  hdfs 路径
[root@node1 spark-yarn]# bin/spark-submit --master yarn --deploy-mode client --class com.geektime.SparkDistCP /home/hadoop/SparkAPI-1.0-SNAPSHOT.jar hdfs://192.168.2.100:8020/input hdfs://192.168.2.100:8020/input_cp -m 2 -i
```

运行结果



```
9 bytes)
2021-08-24 21:58:49,696 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
2021-08-24 21:58:49,700 INFO executor.Executor: Fetching spark://node1:33253/jars/SparkAPI-1.0-SNAPSHOT.jar with timestamp 1629813527824
2021-08-24 21:58:49,706 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
2021-08-24 21:58:49,768 INFO client.TransportClientFactory: Successfully created connection to node1/192.168.2.100:33253 after 44 ms (0 ms spent in
ootstraps)
2021-08-24 21:58:49,776 INFO util.Utils: Fetching spark://node1:33253/jars/SparkAPI-1.0-SNAPSHOT.jar to /tmp/spark-b07580d8-9248-465e-b0c4-08ce81b5
c9/userFiles-3cbf279f-2385-461f-9224-27d5c0d5b386/fetchFileTemp9106367359620397 42.tmp
2021-08-24 21:58:49,805 INFO executor.Executor: Adding file:/tmp/spark-b07580d8-9248-465e-b0c4-08ce81b559c9/userFiles-3cbf279f-2385-461f-9224-27d5c
5b386/SparkAPI-1.0-SNAPSHOT.jar to class loader
From hdfs://192.168.2.100:8020/input/0.txt ---> hdfs://192.168.2.100:8020/input_cp/0.txt Success !
From hdfs://192.168.2.100:8020/input/1.txt ---> hdfs://192.168.2.100:8020/input_cp/1.txt Success !
From hdfs://192.168.2.100:8020/input/2.txt ---> hdfs://192.168.2.100:8020/input_cp/2.txt Success !
From hdfs://192.168.2.100:8020/input/day/word.txt ---> hdfs://192.168.2.100:8020/input_cp/day/word.txt Success !
From hdfs://192.168.2.100:8020/input/one/word.txt ---> hdfs://192.168.2.100:8020/input_cp/one/word.txt Success !
From hdfs://192.168.2.100:8020/input/word.txt ---> hdfs://192.168.2.100:8020/input_cp/word.txt Success !
2021-08-24 21:58:50,291 INFO executor.Executor: Finished task 1.0 in stage 0.0 (TID 1). 923 bytes result sent to driver
2021-08-24 21:58:50,293 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 880 bytes result sent to driver
2021-08-24 21:58:50,305 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 626 ms on node1 (executor driver) (1/2)
2021-08-24 21:58:50,307 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 730 ms on node1 (executor driver) (2/2)
2021-08-24 21:58:50,308 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
2021-08-24 21:58:50,320 INFO scheduler.DAGScheduler: ResultStage 0 (foreach at SparkDistCP.scala:124) finished in 0.894 s
2021-08-24 21:58:50,326 INFO scheduler.DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
2021-08-24 21:58:50,346 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
2021-08-24 21:58:50,362 INFO scheduler.DAGScheduler: Job 0 finished: foreach at SparkDistCP.scala:124, took 0.982897 s
2021-08-24 21:58:50,369 INFO spark.SparkContext: Invoking stop() from shutdown hook
2021-08-24 21:58:50,381 INFO server.AbstractConnector: Stopped Spark@72efb5c1{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2021-08-24 21:58:50,383 INFO ui.SparkUI: Stopped Spark web UI at http://node1:4040
2021-08-24 21:58:50,424 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2021-08-24 21:58:50,434 INFO memory.MemoryStore: MemoryStore cleared
2021-08-24 21:58:50,434 INFO storage.BlockManager: BlockManager stopped
2021-08-24 21:58:50,447 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
2021-08-24 21:58:50,451 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
2021-08-24 21:58:50,459 INFO spark.SparkContext: Successfully stopped SparkContext
2021-08-24 21:58:50,459 INFO util.ShutdownHookManager: Shutdown hook called
2021-08-24 21:58:50,461 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-5997b767-2332-4e72-9e95-2888da6fb9fb
2021-08-24 21:58:50,467 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-b07580d8-9248-465e-b0c4-08ce81b559c9
[root@node1 spark-yarn]#
```



```
[root@node1 hadoop-3.1.3]# bin/hdfs dfs -ls /input
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/software/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.clas
s]
SLF4J: Found binding in [jar:file:/opt/software/hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 6 items
-rw-r--r--   3 root supergroup         16 2021-08-23 23:30 /input/0.txt
-rw-r--r--   3 root supergroup         10 2021-08-23 23:30 /input/1.txt
-rw-r--r--   3 root supergroup         14 2021-08-23 23:30 /input/2.txt
drwxr-xr-x   - root supergroup          0 2021-08-23 19:58 /input/day
drwxr-xr-x   - root supergroup          0 2021-08-23 20:48 /input/one
-rw-r--r--   3 root supergroup         56 2021-05-31 07:39 /input/word.txt
[root@node1 hadoop-3.1.3]# bin/hdfs dfs -ls /input_cp
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/software/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.clas
s]
SLF4J: Found binding in [jar:file:/opt/software/hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 6 items
-rw-r--r--   3 49921 supergroup         16 2021-08-24 22:02 /input_cp/0.txt
-rw-r--r--   3 49921 supergroup         10 2021-08-24 22:02 /input_cp/1.txt
-rw-r--r--   3 49921 supergroup         14 2021-08-24 22:02 /input_cp/2.txt
drwxr-xr-x   - root  supergroup          0 2021-08-24 22:02 /input_cp/day
drwxr-xr-x   - root  supergroup          0 2021-08-24 22:02 /input_cp/one
-rw-r--r--   3 49921 supergroup         56 2021-08-24 22:02 /input_cp/word.txt
[root@node1 hadoop-3.1.3]#
```