

Big Data Frameworks CSE3120
Lab – 5 Map Reduce Join Experiment

Name: Naveen Nidadavolu
Roll No: 22MIA1049

Map Reduce Join

Aim: To perform a join operation using MapReduce on two datasets—cus.txt and trs.txt—to combine bo with their respective marks based on ID.

Procedure

1. Create input text files and upload them to HDFS.
2. Run the MapReduce join program to process the input files.
3. The Mapper reads and tags records from both files based on ID.
4. The Reducer merges records with the same ID to produce the final joined output.
5. The output is stored in HDFS.
6. Retrieve and display the results.

Program

```
import java.io.*;

import java.net.URI;

import java.util.HashMap;

import java.util.Map;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.filecache.DistributedCache;

import org.apache.hadoop.io.*;

import org.apache.hadoop.mapreduce.*;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;


public class msp2 {


    public static class JoinMapper extends Mapper<LongWritable, Text, Text, Text> {

        private Map<String, String> userMap = new HashMap<>();
```

```
private Text outputKey = new Text();
```

```
private Text outputValue = new Text();
```

@Override

protected void setup(Context context) throws IOException, InterruptedException {

 // Read the distributed cache file

 Configuration conf = context.getConfiguration();

 Path[] cacheFiles = DistributedCache.getLocalCacheFiles(conf);

 if (cacheFiles != null && cacheFiles.length > 0) {

 BufferedReader reader = new BufferedReader(new FileReader(cacheFiles[0].toString()));

 String line;

 while ((line = reader.readLine()) != null) {

 String[] parts = line.split(",");

 userMap.put(parts[0], parts[1]); // UserID -> UserName

 }

 reader.close();

 System.err.println("deb"+ userMap.size() + "users.");

 } else {

 throw new IOException("dis cah is missing");}

 }

@Override

protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

 String[] transaction = value.toString().split(",");

 String userId = transaction[0];

 String amount = transaction[1];

 if (userMap.containsKey(userId)) {

 outputKey.set(userMap.get(userId)); // UserName

```
        outputValue.set(amount);

        context.write(outputKey, outputValue);
    }

}
```

```
public static void main(String[] args) throws Exception {

    if(args.length<3){

        System.err.println("err");

        System.exit(1);

    }

    Configuration conf = new Configuration();

    DistributedCache.addCacheFile(new URI(args[2]),conf);

    Job job = new Job(conf);

    job.setJarByClass(msp2.class);

    job.setMapperClass(JoinMapper.class);

    job.setOutputKeyClass(Text.class);

    job.setOutputValueClass(Text.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));

    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    // Add users.txt to the Distributed Cache

    job.setNumReduceTasks(0); // No reducer required for mapside join

    System.exit(job.waitForCompletion(true) ? 0 : 1);

}
```

}

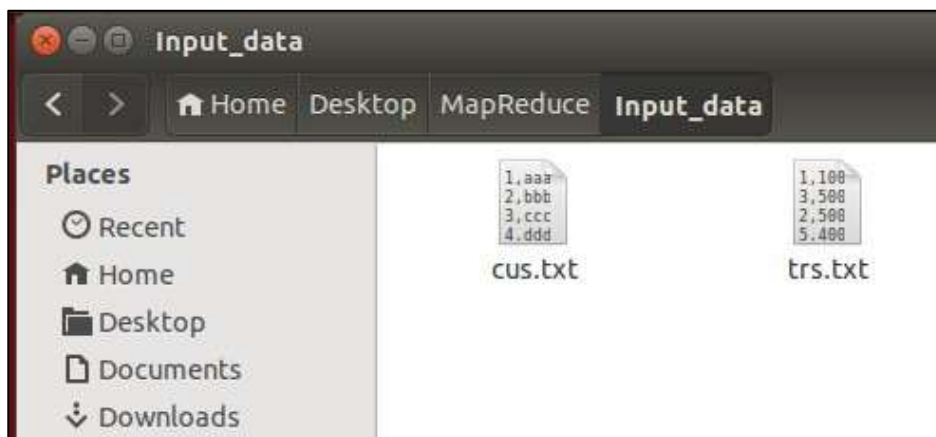
Input files

- File 1 (cus.txt): ID Name
- File 2 (trs.txt): ID Price

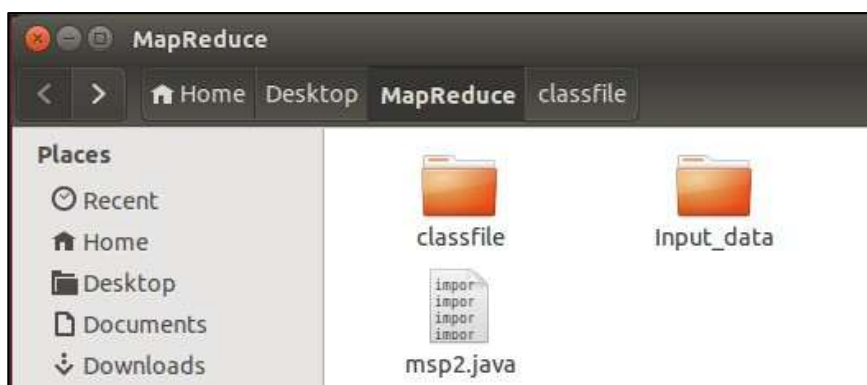
1. Download the text and program files, and place in a **separate folder in desktop**
2. Open the terminal and run basic commands to start the Hadoop services and check its version.



3. In the desktop, create a folder, For ex: here the folder name is given as **MapReduce**. Inside this folder create an **Input_data** folder and place the input text files (**cus.txt** and **trs.txt**).



4. Inside the MapReduce folder create an empty folder name classfile.



5. Refer the screenshots below

```
ponny@ubuntu:~$ start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-namenode-ubuntu.out
localhost: starting datanode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-jobtracker-ubuntu.out
localhost: starting tasktracker, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-tasktracker-ubuntu.out
ponny@ubuntu:~$ jps
13155 Jps
12871 SecondaryNameNode
12954 JobTracker
12539 NameNode
13116 TaskTracker
12704 DataNode
ponny@ubuntu:~$ javac -version
javac 1.7.0_201
ponny@ubuntu:~$ export HADOOP
ponny@ubuntu:~$ export HADOOP_CLASSPATH=$(hadoop classpath)
Warning: $HADOOP_HOME is deprecated.
```

```
ponny@ubuntu:~$ echo $HADOOP_CLASSPATH
/home/ponny/hadoop/libexec/../conf:/usr/lib/jvm/java-7-openjdk-i386/lib/tools.jar:/home/ponny/hadoop/libexec/...:/home/ponny/hadoop/libexec/.../hadoop-core-1.0.4.jar:/home/ponny/hadoop/libexec/.../lib/asm-3.2.jar:/home/ponny/hadoop/libexec/.../lib/aspectjrt-1.6.5.jar:/home/ponny/hadoop/libexec/.../lib/aspectjtools-1.6.5.jar:/home/ponny/hadoop/libexec/.../lib/commons-beanutils-1.7.0.jar:/home/ponny/hadoop/libexec/.../lib/commons-beanutils-core-1.8.0.jar:/home/ponny/hadoop/libexec/.../lib/commons-cli-1.2.jar:/home/ponny/hadoop/libexec/.../lib/commons-codec-1.4.jar:/home/ponny/hadoop/libexec/.../lib/commons-collections-3.2.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-configuration-1.6.jar:/home/ponny/hadoop/libexec/.../lib/commons-daemon-1.0.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-digester-1.8.jar:/home/ponny/hadoop/libexec/.../lib/commons-el-1.0.jar:/home/ponny/hadoop/libexec/.../lib/commons-httpclient-3.0.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-io-2.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-lang-2.4.jar:/home/ponny/hadoop/libexec/.../lib/commons-logging-1.1.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-logging-api-1.0.4.jar:/home/ponny/hadoop/libexec/.../lib/commons-math-2.1.jar:/home/ponny/hadoop/libexec/.../lib/commons-net-1.4.1.jar:/home/ponny/hadoop/libexec/.../lib/core-3.1.1.jar:/home/ponny/hadoop/libexec/.../lib/hadoop-capacity-scheduler-1.0.4.jar:/home/ponny/hadoop/libexec/.../lib/hadoop-fairscheduler-1.0.4.jar:/home/ponny/hadoop/libexec/.../lib/hadoop-thriftfs-1.0.4.jar:/home/ponny/hadoop/libexec/.../lib/hsqldb-1.8.0.10.jar:/home/ponny/hadoop/libexec/.../lib/jackson-core-asl-1.8.8.jar:/home/ponny/hadoop/libexec/.../lib/jackson-mapper-asl-1.8.8.jar:/home/ponny/hadoop/libexec/.../lib/jasper-compiler-5.5.12.jar:/home/ponny/hadoop/libexec/.../lib/jasper-runtime-5.5.12.jar:/home/ponny/hadoop/libexec/.../lib/jdeb-0.8.jar:/home/ponny/hadoop/libexec/.../lib/jersey-core-1.8.jar:/home/ponny/hadoop/libexec/.../lib/jersey-json-1.8.jar:/home/ponny/hadoop/libexec/.../lib/jersey-server-1.8.jar:/home/ponny/hadoop/libexec/.../lib/jets3t-0.6.1.jar:/home/ponny/hadoop/libexec/.../lib/jetty-6.1.26.jar:/home/ponny/h
```



```

t-1.4.1.jar:/home/ponny/hadoop/libexec/./lib/core-3.1.1.jar:/home/ponny/hadoop/libexec/./lib/hadoop-capacity-scheduler-1.0.4.jar:/home/ponny/hadoop/libexec/./lib/hadoop-fairscheduler-1.0.4.jar:/home/ponny/hadoop/libexec/./lib/hadoop-thriftfs-1.0.4.jar:/home/ponny/hadoop/libexec/./lib/hsqlldb-1.8.0.10.jar:/home/ponny/hadoop/libexec/./lib/jackson-core-asl-1.8.8.jar:/home/ponny/hadoop/libexec/./lib/jackson-mapper-asl-1.8.8.jar:/home/ponny/hadoop/libexec/./lib/jasper-compiler-5.5.12.jar:/home/ponny/hadoop/libexec/./lib/jasper-runtime-5.5.12.jar:/home/ponny/hadoop/libexec/./lib/jdeb-0.8.jar:/home/ponny/hadoop/libexec/./lib/jersey-core-1.8.jar:/home/ponny/hadoop/libexec/./lib/jersey-json-1.8.jar:/home/ponny/hadoop/libexec/./lib/jersey-server-1.8.jar:/home/ponny/hadoop/libexec/./lib/jets3t-0.6.1.jar:/home/ponny/hadoop/libexec/./lib/jetty-6.1.26.jar:/home/ponny/hadoop/libexec/./lib/jetty-util-6.1.26.jar:/home/ponny/hadoop/libexec/./lib/jsch-0.1.42.jar:/home/ponny/hadoop/libexec/./lib/junit-4.5.jar:/home/ponny/hadoop/libexec/./lib/kfs-0.2.2.jar:/home/ponny/hadoop/libexec/./lib/log4j-1.2.15.jar:/home/ponny/hadoop/libexec/./lib/mockito-all-1.8.5.jar:/home/ponny/hadoop/libexec/./lib/oro-2.0.8.jar:/home/ponny/hadoop/libexec/./lib/servlet-api-2.5-20081211.jar:/home/ponny/hadoop/libexec/./lib/slf4j-api-1.4.3.jar:/home/ponny/hadoop/libexec/./lib/slf4j-log4j12-1.4.3.jar:/home/ponny/hadoop/libexec/./lib/xmlenc-0.52.jar:/home/ponny/hadoop/libexec/./lib/jsp-2.1/jsp-2.1.jar:/home/ponny/hadoop/libexec/./lib/jsp-2.1/jsp-api-2.1.jar
ponny@ubuntu:~$ hadoop fs -mkdir /mapsidejoin
Warning: $HADOOP_HOME is deprecated.

mkdir: org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create directory /mapsidejoin. Name node is in safe mode.
ponny@ubuntu:~$ hadoop dfsadmin -safemode leave
Warning: $HADOOP_HOME is deprecated.

Safe mode is OFF
ponny@ubuntu:~$ hadoop fs -mkdir /mapsidejoin
Warning: $HADOOP_HOME is deprecated.

```

```

ponny@ubuntu:~$ hadoop fs -mkdir /mapsidejoin/Input
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~$ hadoop fs -put '/home/ponny/Desktop/MapReduce/Input_data/cus.txt' /mapsidejoin/Input
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~$ hadoop fs -put '/home/ponny/Desktop/MapReduce/Input_data/trs.txt' /mapsidejoin/Input
Warning: $HADOOP_HOME is deprecated.

```

localhost:50075/browseDirectory.jsp?dir=/mapsidejoin&namenodeInfoPort=50070

Contents of directory /mapsidejoin

Goto :

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
Input	dir				2025-03-10 11:01	rw-r--r--	ponny	supergroup

localhost:50075/browseDirectory.jsp?dir=%2Fmapsidejoin%2FInput&namenodeInfoPort=50070

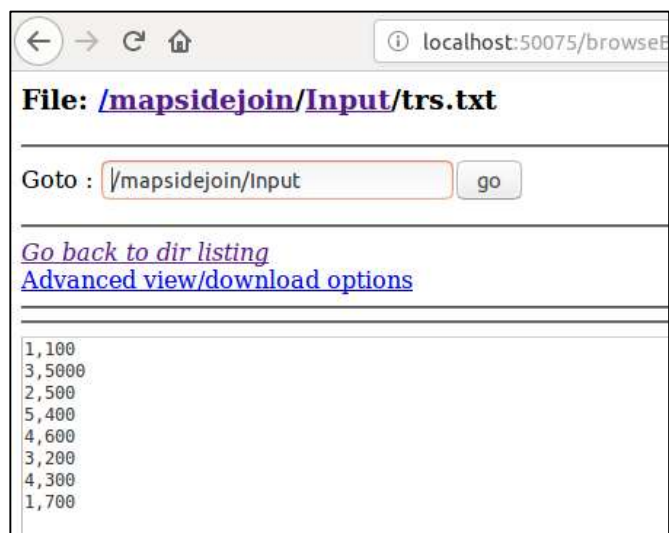
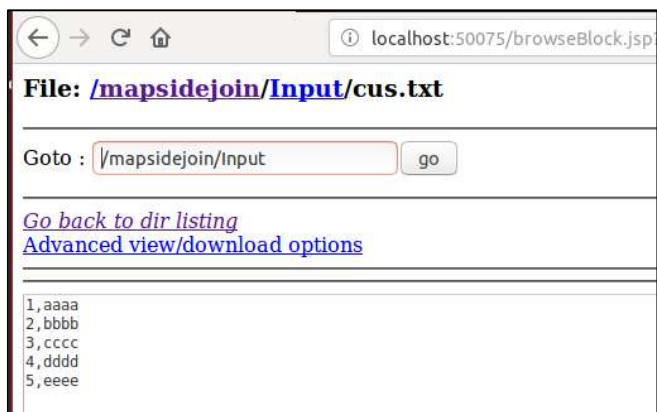
Contents of directory /mapsidejoin/Input

Goto :

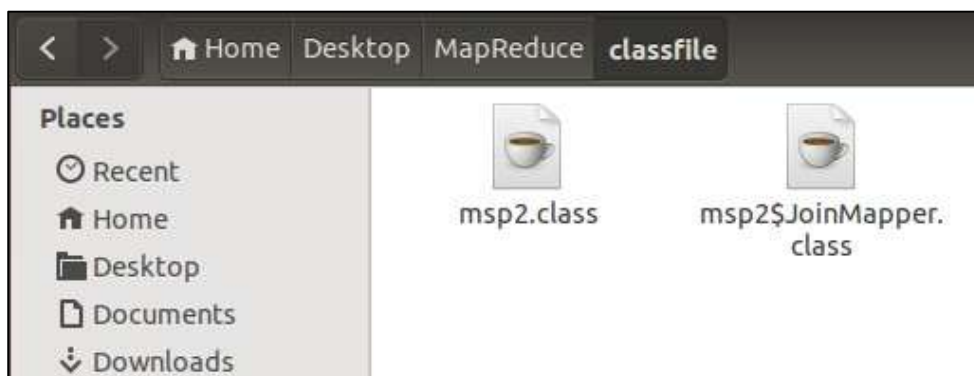
[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
cus.txt	file	0.03 KB	1	64 MB	2025-03-10 11:01	rw-r--r--	ponny	supergroup
trs.txt	file	0.05 KB	1	64 MB	2025-03-10 11:01	rw-r--r--	ponny	supergroup

[Go back to DFS home](#)



```
ponny@ubuntu:~/Desktop$ cd MapReduce
ponny@ubuntu:~/Desktop/MapReduce$ javac -classpath ${HADOOP_CLASSPATH} -d /home/ponny/Desktop/MapReduce/classfile /home/ponny/Desktop/MapReduce/msp2.java
ponny@ubuntu:~/Desktop/MapReduce$ jar -cvf mapsidejar.jar -C classfile/ .
added manifest
adding: msp2.class(in = 1560) (out= 874)(deflated 43%)
adding: msp2$JoinMapper.class(in = 3402) (out= 1452)(deflated 57%)
```




```

ponny@ubuntu:~/Desktop$ cd MapReduce
ponny@ubuntu:~/Desktop/MapReduce$ javac -classpath ${HADOOP_CLASSPATH} -d /home/ponny/Desktop/MapReduce/classfile /home/ponny/Desktop/MapReduce/msp2.java
ponny@ubuntu:~/Desktop/MapReduce$ jar -cvf mapsidejar.jar -C classfile/ .
added manifest
adding: msp2.class(in = 1560) (out= 874)(deflated 43%)
adding: msp2$JoinMapper.class(in = 3402) (out= 1452)(deflated 57%)
ponny@ubuntu:~/Desktop/MapReduce$ hadoop jar mapsidejar.jar msp2 /mapsidejoin/Input/trs.txt /mapsidejoin/Output /mapsidejoin/Input/cus.txt
Warning: $HADOOP_HOME is deprecated.

25/03/10 11:02:29 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
25/03/10 11:02:29 INFO input.FileInputFormat: Total input paths to process : 1
25/03/10 11:02:29 INFO util.NativeCodeLoader: Loaded the native-hadoop library
25/03/10 11:02:29 WARN snappy.LoadSnappy: Snappy native library not loaded
25/03/10 11:02:29 INFO mapred.JobClient: Running job: job_202503101059_0001
25/03/10 11:02:30 INFO mapred.JobClient: map 0% reduce 0%
25/03/10 11:02:47 INFO mapred.JobClient: map 100% reduce 0%
25/03/10 11:02:52 INFO mapred.JobClient: Job complete: job_202503101059_0001
25/03/10 11:02:52 INFO mapred.JobClient: Counters: 19
25/03/10 11:02:52 INFO mapred.JobClient:   Job Counters
25/03/10 11:02:52 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=13306
25/03/10 11:02:52 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
25/03/10 11:02:52 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
25/03/10 11:02:52 INFO mapred.JobClient:     Launched map tasks=1
25/03/10 11:02:52 INFO mapred.JobClient:     Data-local map tasks=1
25/03/10 11:02:52 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCE=0
25/03/10 11:02:52 INFO mapred.JobClient:   File Output Format Counters
25/03/10 11:02:52 INFO mapred.JobClient:     Bytes Written=73
25/03/10 11:02:52 INFO mapred.JobClient:   FileSystemCounters
25/03/10 11:02:52 INFO mapred.JobClient:     HDFS_BYTES_READ=162
25/03/10 11:02:52 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=21661
25/03/10 11:02:52 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=73
25/03/10 11:02:52 INFO mapred.JobClient:   File Input Format Counters
25/03/10 11:02:52 INFO mapred.JobClient:     Bytes Read=49
25/03/10 11:02:52 INFO mapred.JobClient:   Map-Reduce Framework
25/03/10 11:02:52 INFO mapred.JobClient:     Map input records=8
25/03/10 11:02:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot=45903872
25/03/10 11:02:52 INFO mapred.JobClient:     Spilled Records=0
25/03/10 11:02:52 INFO mapred.JobClient:     CPU time spent (ms)=80
25/03/10 11:02:52 INFO mapred.JobClient:     Total committed heap usage (bytes)=16252928
25/03/10 11:02:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=384790528
25/03/10 11:02:52 INFO mapred.JobClient:     Map output records=8
25/03/10 11:02:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=113
ponny@ubuntu:~/Desktop/MapReduce$ █

```

←
→
↺
🏠
localhost:50075/browseDirectory.jsp?dir=/mapsidejoin&namenodeinfoPort=50070

Contents of directory /mapsidejoin

Goto :

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
Input	dir				2025-03-10 11:01	rwxr-xr-x	ponny	supergroup
Output	dir				2025-03-10 11:02	rwxr-xr-x	ponny	supergroup

[Go back to DFS home](#)

localhost:50075/browseDirectory.jsp?dir=%2Fmapsidejoin%2FOutput&namenodeinfoPort=50070

Contents of directory [/mapsidejoin/Output](#)

Goto :

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	1	64 MB	2025-03-10 11:02	rw-r--r--	ponny	supergroup
_logs	dir				2025-03-10 11:02	rw-r--r--	ponny	supergroup
part-m-00000	file	0.07 KB	1	64 MB	2025-03-10 11:02	rw-r--r--	ponny	supergroup

[Go back to DFS home](#)

Output:

localhost:50075/browseBlo

File: [/mapsidejoin/Output/part-m-00000](#)

Goto :

[Go back to dir listing](#)
[Advanced view/download options](#)

aaaa	100
cccc	5000
bbbb	500
eeee	400
dddd	600
cccc	200
dddd	300
aaaa	700

Result: Successfully performed a join operation using MapReduce, merging customer details with their transactions based on ID.