

PROBLEM1:

1/

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.Random;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

public class ab {

    public static class TransMapper extends Mapper <Object, Text, Text, Text>
    {
        public void map(Object key, Text value, Context context)
        throws IOException, InterruptedException
        {
            String record = value.toString().trim();
            String[] parts = record.split(",");
            String date = parts[1];
            String month = date.substring(0,3);
            String id= parts[2];
            String amount = parts[3];
            context.write(new Text(month), new Text(id+" "+amount));
        }
    }

    public static class TransReducer extends Reducer <Text, Text, Text, Text>
    {
        private String maxMonth;
        private double maxCost;
        public void reduce(Text key, Iterable<Text> values, Context context)
        throws IOException, InterruptedException
        {
            List<String> ids =new ArrayList<String>();
            double total = 0.0;
            int count=0;

            for (Text t : values)
            {
                String part[]=t.toString().trim().split(" ");
                String id=part[0];
                String amount=part[1];
                total += Float.parseFloat(amount);
                if(!ids.contains(id))
                {
                    ids.add(id);

                    count +=1;
                }
            }
        }
    }
}
```

```

    }
    if (total > maxCost) {
        maxCost = total;
        maxMonth = key.toString();
    }
    context.write(key, new Text(count+ " "+Double.toString(total)));
}

public void cleanup(Context context
    ) throws IOException, InterruptedException {
context.write(new Text(maxMonth), new Text(" is the month with highest cost (" + maxCost + ")"));
}
}

public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "Trans analysis 1");
job.setJarByClass(ab.class);
job.setMapperClass(TransMapper.class);
job.setReducerClass(TransReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
//job.setNumReduceTasks(0);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);

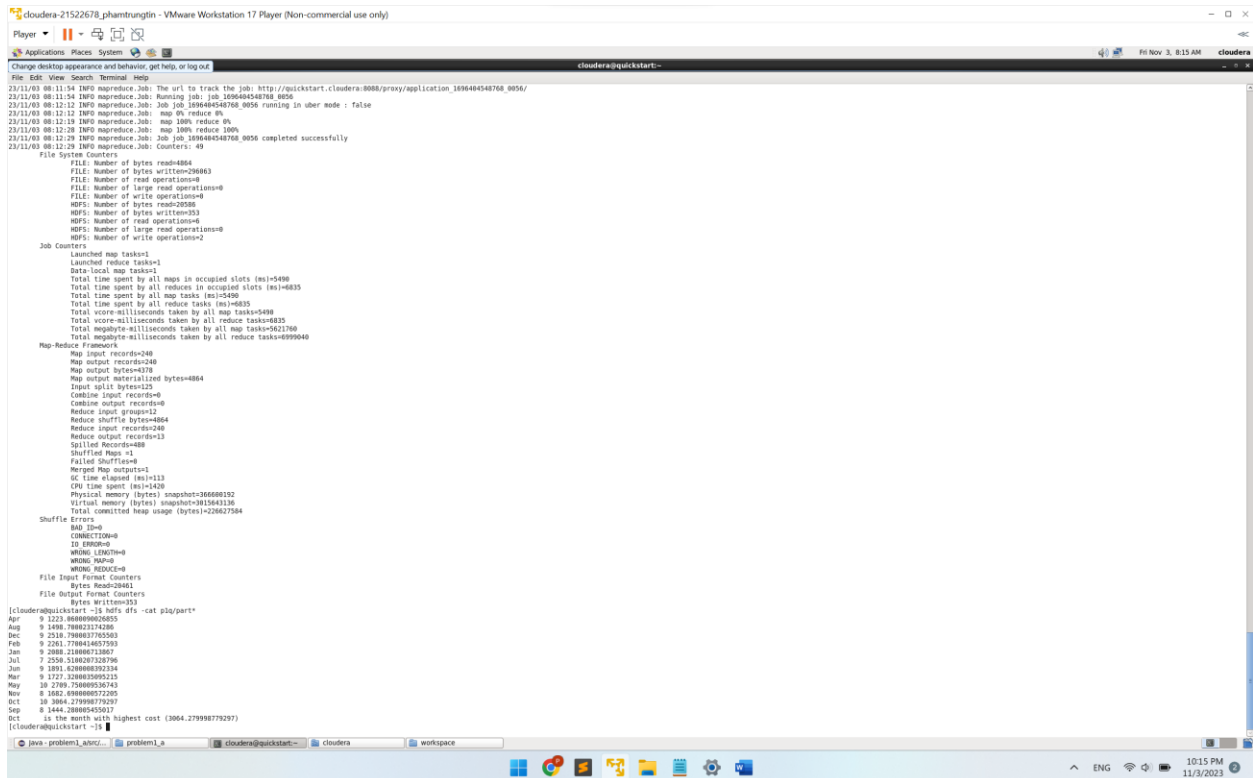
}
}

```

```

cloudera-21522678_phamtrungtin - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
Feb 9 713.8760124925983
Jun 9 1848.8688015648279
Jul 7 709.8308088836121
Jun 9 759.2399826848085
Mar 9 856.709914538781
May 10 1848.3189189188342
Nov 8 712.64885588315
Oct 10 738.6499928428191
Sep 8 1832.85998783803
[cloudera@quickstart ~]$ hadoop jar workspace/problem1 a/ab.jar ab trans4811.txt plq
23/11/83 08:11:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/11/83 08:11:43 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/11/83 08:11:46 INFO input.FileInputFormat: Total input paths to process : 1
23/11/83 08:11:47 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:894)
23/11/83 08:11:47 INFO mapreduce.JobSubmitter: number of splits:1
23/11/83 08:11:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1696484548768_0056
23/11/83 08:11:52 INFO impl.YarnClientImpl: Submitted application application_1696484548768_0056
23/11/83 08:11:54 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1696484548768_0056/
23/11/83 08:11:54 INFO mapreduce.Job: Running job: job_1696484548768_0056
23/11/83 08:12:12 INFO mapreduce.Job: Job job_1696484548768_0056 running in user mode : false
23/11/83 08:12:12 INFO mapreduce.Job: map 0% reduce 0%
23/11/83 08:12:19 INFO mapreduce.Job: map 100% reduce 0%
23/11/83 08:12:20 INFO mapreduce.Job: map 100% reduce 100%

```



2/

```
import java.io.*;
import java.util.*;
import java.net.URI;
```

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.Random;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
```

```
public class pb2 {
```

```
public static class TransMapper extends Mapper<Object, Text, Text, Text>
{
    // user map to keep the userID-username
    private Map<Integer, String> userMap = new HashMap<>();
```

```

public void setup(Context context) throws IOException,
    InterruptedException
{
    try (BufferedReader br = new BufferedReader(new FileReader("cus.txt"))) {
        String line;
        while ((line = br.readLine()) != null) {
            String columns[] = line.split(",");
            String id = columns[0];
            String name = columns[1];
            String tn=name;
            userMap.put(Integer.parseInt(id),tn);
        }
    } catch (IOException e) {
        e.printStackTrace();
    }
}

public void map(Object key, Text value, Context context)
throws IOException, InterruptedException
{
    String record = value.toString().trim();
    String[] parts = record.split(",");
    String id= parts[2];
    String date = parts[1];
    String month = date.substring(0,3);
    String nt = userMap.get(Integer.parseInt(id));
    context.write(new Text(month), new Text(nt));
}
}

```

```

public static class TransReducer extends Reducer <Text, Text, Text, Text>
{
    public void reduce(Text key, Iterable<Text> values, Context context)
    throws IOException, InterruptedException
    {
        List<String> name =new ArrayList<String>();
        Map<String, Integer> userCount = new HashMap<>();
        int dem=0;
        for (Text t : values)
        {
            String part[]=t.toString().trim().split(" ");
            String ten = part[0];
            if (!name.contains(ten)) {
                name.add(ten);
                userCount.put(ten,1);
            }
            else {
                userCount.put(ten, userCount.get(ten)+1);
            }
        }
        context.write(key, new Text(userCount + " "));
    }
}

```

```

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

    Configuration conf = new Configuration();

```

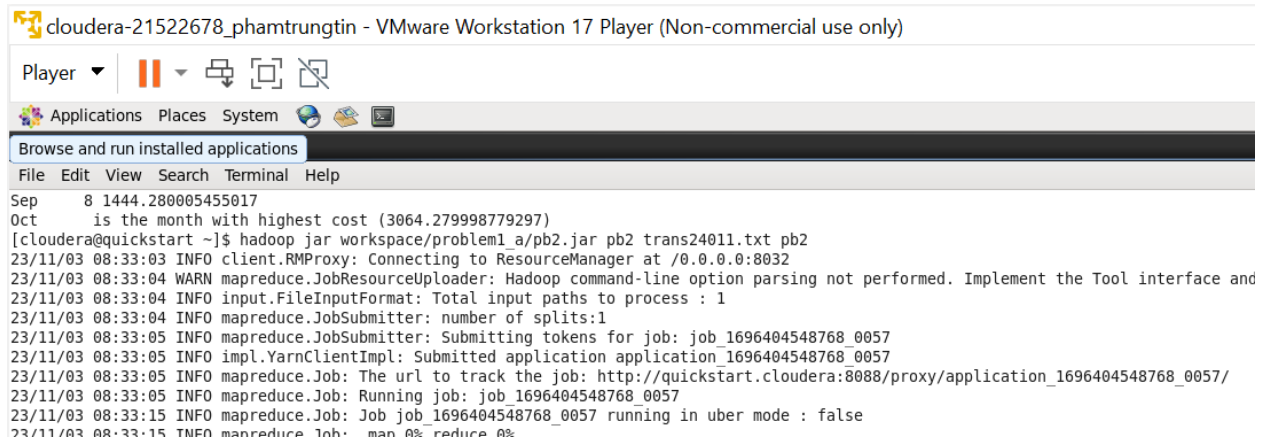
```

Job job = new Job(conf, "MapsideJoin");
job.setJarByClass(pb2.class);
job.setMapperClass(TransMapper.class);
job.setReducerClass(TransReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
// Setting reducer to zero
//job.setNumReduceTasks(0);
try {

    job.addCacheFile(new URI("hdfs://localhost:8020/icache/cus.txt"));
}
catch (Exception e) {
    System.out.println("File Not Added");
    System.exit(1);
}

FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}

```



The screenshot shows a VMware Workstation 17 Player window titled "cloudera-21522678_phamtrungtin - VMware Workstation 17 Player (Non-commercial use only)". The interface includes a top toolbar with icons for Player, Full Screen, and other controls. Below the toolbar is a menu bar with "Applications", "Places", and "System". A "Browse and run installed applications" button is visible. The main area is a terminal window with a menu bar (File, Edit, View, Search, Terminal, Help) and the following output:

```

Sep  8 1444.280005455017
Oct   is the month with highest cost (3064.279998779297)
[cloudera@quickstart ~]$ hadoop jar workspace/problem1_a/pb2.jar pb2 trans24011.txt pb2
23/11/03 08:33:03 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/11/03 08:33:04 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
23/11/03 08:33:04 INFO input.FileInputFormat: Total input paths to process : 1
23/11/03 08:33:04 INFO mapreduce.JobSubmitter: number of splits:1
23/11/03 08:33:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1696404548768_0057
23/11/03 08:33:05 INFO impl.YarnClientImpl: Submitted application application_1696404548768_0057
23/11/03 08:33:05 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1696404548768_0057/
23/11/03 08:33:05 INFO mapreduce.Job: Running job: job_1696404548768_0057
23/11/03 08:33:15 INFO mapreduce.Job: Job job_1696404548768_0057 running in uber mode : false
23/11/03 08:33:15 INFO mapreduce.Job: map 0% reduce 0%

```

```
File Edit View Search Terminal mwp
23/11/03 08:33:45 INFO impl.VarImplImpl: Submitted application application 1896484548768 0057
23/11/03 08:33:45 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/pxxy/application/1896484548768_0057/
23/11/03 08:33:45 INFO mapreduce.Job: Running job: job_1896484548768_0057
23/11/03 08:33:45 INFO mapreduce.Job: Job job_1896484548768_0057 running in uber mode : false
23/11/03 08:33:45 INFO mapreduce.Job: map 0% reduce 0%
23/11/03 08:33:45 INFO mapreduce.Job: map 100% reduce 0%
23/11/03 08:33:45 INFO mapreduce.Job: map 100% reduce 100%
23/11/03 08:33:45 INFO mapreduce.Job: Job job_1896484548768_0057 completed successfully
23/11/03 08:33:45 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=3177
  FILE: Number of bytes written=24653
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=30580
  HDFS: Number of bytes written=1162
  HDFS: Number of read operations=0
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=5288
  Total time spent by all reducers in occupied slots (ms)=6686
  Total time spent by all map tasks (ms)=5288
  Total time spent by all reduce tasks (ms)=6686
  Total vcore-millisecods taken by all map tasks=5288
  Total vcore-millisecods taken by all reduce tasks=6686
  Total megabyte-millisecods taken by all map tasks=5488720
  Total megabyte-millisecods taken by all reduce tasks=6846464
Map-Reduce Framework
  Map input records=248
  Map output records=248
  Map output bytes=2481
  Map output materialized bytes=3177
  Input split bytes=125
  Combine input records=0
  Reduce input groups=12
  Reduce shuffle bytes=3177
  Reduce input records=248
  Reduce output records=12
  Spilled Records=248
  Shuffled Maps=1
  Failed Shuffles=1
  Merged Map outputs=1
  GC time elapsed (ms)=136
  CPU time spent (ms)=1390
  Physical memory (bytes) mapped=330800992
  Virtual memory (bytes) swapped=3815614464
  Total committed heap usage (bytes)=226627584
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=30580
File Output Format Counters
  Bytes Written=1162
[cloudera@quickstart ~]$ hdfs dfs -cat /opt/part*
Apr (Sherrin4, Gretchen2, Paige2, Patrick2, Elian1, Karen2, Kristina2, Malcolam1, Dolares2, Hazel2)
Aug (Gretchen2, Paige2, Patrick2, Elian1, Karen1, Kristina2, Malcolam1, Hazel2, Dolares4)
Dec (Sherrin2, Gretchen2, Paige2, Patrick2, Elian1, Kristina1, Hazel2, Malcolam1, Dolares2)
Feb (Gretchen2, Paige2, Patrick2, Elian1, Karen2, Kristina2, Malcolam2, Dolares2, Hazel2)
Jul (Sherrin2, Gretchen2, Paige2, Patrick2, Elian1, Kristina1, Hazel2, Malcolam2, Dolares2)
Jun (Sherrin2, Gretchen2, Paige2, Patrick2, Karen2, Elian4, Hazel2)
Mar (Sherrin4, Gretchen2, Paige2, Elian1, Karen1, Kristina2, Dolares2, Hazel1, Malcolam3)
May (Sherrin2, Gretchen2, Paige2, Karen2, Elian2, Kristina1, Dolares2, Hazel2, Malcolam1)
Nov (Sherrin2, Gretchen2, Paige2, Karen1, Kristina4, Hazel2, Malcolam3, Dolares2)
Oct (Sherrin2, Gretchen2, Paige2, Patrick2, Elian2, Karen2, Kristina2, Malcolam1, Hazel2, Dolares2)
Sep (Gretchen2, Paige2, Karen2, Elian4, Kristina1, Hazel2, Malcolam1, Dolares2)
[cloudera@quickstart ~]$
```

PROBLEM 2:

```
import java.io.BufferedReader;

import java.io.FileReader;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.Mapper.Context;
```

```
import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.*;
import java.util.*;
import java.net.URI;

import java.io.IOException;
import java.util.ArrayList;
import java.util.HashMap;
import java.util.List;
import java.util.Map;
import java.util.Random;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

public class b2 {
    public static class CustsMapper extends Mapper <Object, Text, Text, Text>
    {
        public void map(Object key, Text value, Context context)
        throws IOException, InterruptedException
        {
```

```

        String record = value.toString();
        String[] columns= record.split(",");
        String id = columns[0];
        String pro = columns[4];
        String name = columns[1];
        String k=id;
        context.write(new Text(k), new Text("cust " + name+','+pro));
    }
}

public static class tranMapper extends Mapper <Object, Text, Text, Text>
{
    public void map(Object key, Text value, Context context)
        throws IOException, InterruptedException
    {
        String record = value.toString();
        String[] columns= record.split(",");
        context.write(new Text(columns[2]), new Text("tran " + columns[0]));
    }
}

```

```

public static class ReduceJoinReducer extends Reducer <Text, Text, Text, Text>
{
    private Map<String, String> userMap = new HashMap<>();

    public void setup(Context context) throws IOException,
        InterruptedException
    {
        try (BufferedReader br = new BufferedReader(new FileReader("pro.txt"))) {
            String line;

```



```

        while ((line = br.readLine()) != null) {
            String columns[] = line.split(",");
            userMap.put(columns[0],columns[1]);
        }
    } catch (NumberFormatException ex) {
        System.out.println("not a number");
    }
}

public void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException
{
    String name = "";
    String pro = "";
    String sar="";
    int count1 = 0;
    int count2 = 0;
    int count3 = 0;
    for (Text t : values)
    {

        String parts[] = t.toString().split(" ");
        if (parts[0].equals("cust"))
        {
            String[] col =parts[1].toString().trim().split(",");
            name=col[0];
            pro=col[1];
            sar=userMap.get(pro);
        }
        else if (parts[0].equals("tran"))

```

```

{
    count2+=1;
}
}
count3=count1+count2;
if(count2<12 && count2>0){
    if(Integer.parseInt(sar)>70000){
context.write(new Text(name), new Text(count2+" "+pro+" "+sar));
    }

}
}
}
}

```

```

public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "Reduce-side join");
job.setJarByClass(b2.class);
job.setReducerClass(ReduceJoinReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);

MultipleInputs.addInputPath(job, new Path(args[0]),TextInputFormat.class, CustsMapper.class);
MultipleInputs.addInputPath(job, new Path(args[1]),TextInputFormat.class, tranMapper.class);

Path outputPath = new Path(args[2]);
try {

    job.addCacheFile(new URI("hdfs://localhost:8020/icache/pro.txt"));
}
}

```

```

    }

    catch (Exception e) {

        System.out.println("File Not Added");

        System.exit(1);

    }

    FileOutputFormat.setOutputPath(job, outputPath);

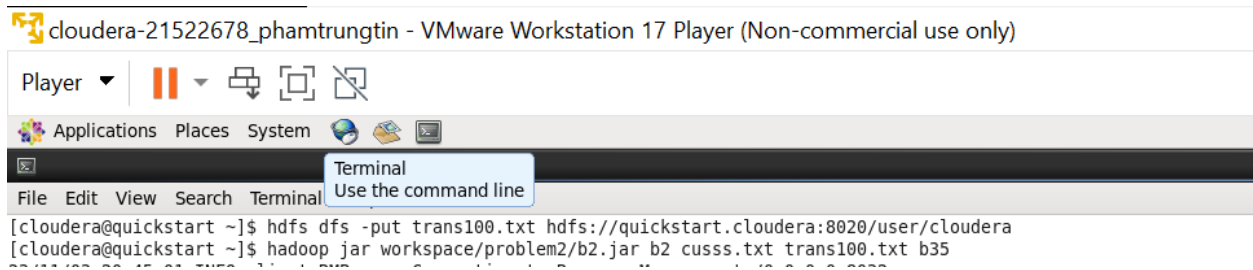
    outputPath.getFileSystem(conf).delete(outputPath);

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

}

}

```



```

cloudera-21522678_phamtrungtin - VMware Workstation 17 Player (Non-commercial use only)

Player
Applications Places System
Terminal
Use the command line
File Edit View Search Terminal
[cloudera@quickstart ~]$ hdfs dfs -put trans100.txt hdfs://quickstart.cloudera:8020/user/cloudera
[cloudera@quickstart ~]$ hadoop jar workspace/problem2/b2.jar b2 cusss.txt trans100.txt b35

```

cloudera-21522678.phamtrungtin - VMware Workstation 17 Player (Non-commercial use only)

Player

Applications Places System

cloudera@quickstart:~\$

```
File Edit View Search Terminal Help
at java.lang.Object.wait(Native Method)
at java.lang.Thread.join(Thread.java:1281)
at java.lang.Thread.join(Thread.java:1355)
at org.apache.hadoop.hdfs.DFSOutputStream.readDataStreamer.closeResponder(DFSOutputStream.java:967)
at org.apache.hadoop.hdfs.DFSOutputStream.readDataStreamer.run(DFSOutputStream.java:894)
23/11/03 20:45:42 INFO mapreduce.JobSubmitter: number of splits:2
23/11/03 20:45:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1696404548768_0099
23/11/03 20:45:43 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8086/pxxy/application_1696404548768_0099/
23/11/03 20:45:43 INFO mapreduce.Job: Running job: job_1696404548768_0099
23/11/03 20:45:12 INFO mapreduce.Job: Job job_1696404548768_0099 running in uber mode : false
23/11/03 20:45:12 INFO mapreduce.Job: map 0% reduce 0%
23/11/03 20:45:26 INFO mapreduce.Job: map 50% reduce 0%
23/11/03 20:45:27 INFO mapreduce.Job: map 100% reduce 0%
23/11/03 20:45:34 INFO mapreduce.Job: map 100% reduce 100%
23/11/03 20:45:35 INFO mapreduce.Job: Job job_1696404548768_0099 completed successfully
23/11/03 20:45:36 INFO mapreduce.Job: Counter: 49
File System Counters
FILE: Number of bytes read=5583
FILE: Number of bytes written=44889
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=21744
HDFS: Number of bytes written=74
HDFS: Number of read operations=9
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=21713
Total time spent by all reducers in occupied slots (ms)=5527
Total time spent by all map tasks (ms)=21713
Total time spent by all reduce tasks (ms)=5527
Total vcore-milliseconds taken by all map tasks=21713
Total vcore-milliseconds taken by all reduce tasks=5527
Total megabyte-milliseconds taken by all map tasks=22236168
Total megabyte-milliseconds taken by all reduce tasks=5039648
Map-Reduce Framework
Map input records=259
Map output records=259
Map output materialized bytes=5589
Input split bytes=0
Combine input records=0
Combine output records=0
Reduce input groups=19
Reduce shuffle bytes=5589
Reduce input records=259
Reduce output records=3
Spilled Records=18
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=230
CPU time spent (ms)=1840
Physical memory (bytes) snapshot=688317440
Virtual memory (bytes) snapshot=4519579648
Total committed heap usage (bytes)=392372224
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
MRIO_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=74
[cloudera@quickstart ~]$ hdfs dfs -cat hls/part*
Karen 10 Lawyer 120800
Patrick 30 Veterinarian 200300
Elise 6 Pilot 97000
[cloudera@quickstart ~]$
```

cloudera

10:46 AM 11/4/2023