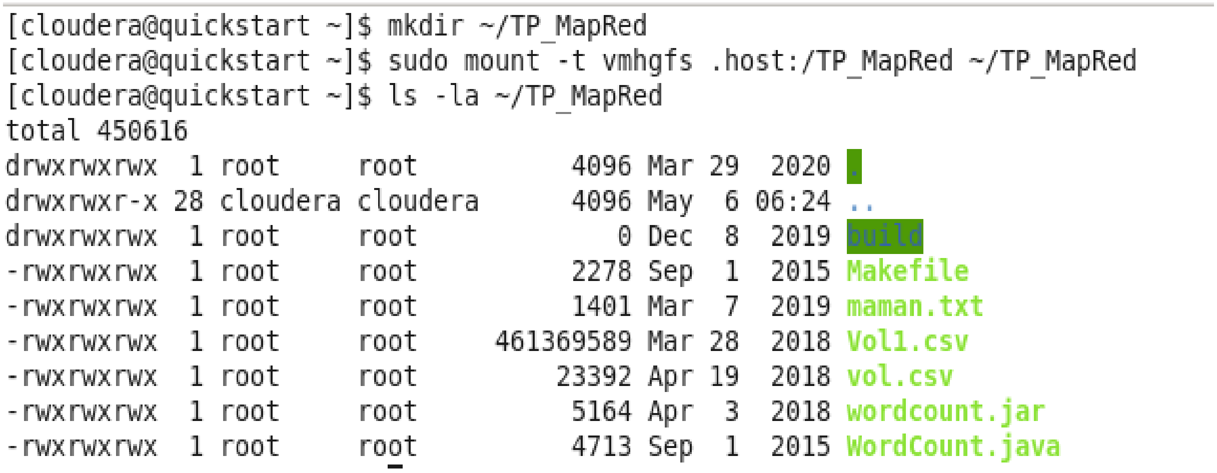
TP 02 – MapReduce : Implémentation et Exploration

Réaliser Par : **Soufiane Erraad ,** **Groupe : G8 /4IIR/Les Orangers**

Amine Eddahiri ,

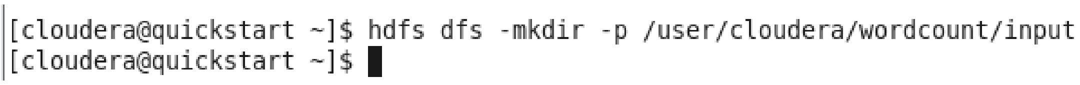
Larbi Faddani .

1. Décompresser le fichier TP_MapRed.rar fourni et copier le dossier Tp_MapRed dans le dossier local /home/cloudera de votre VM



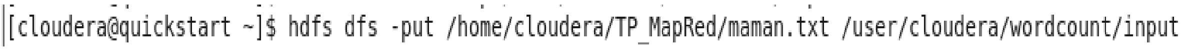
```
[cloudera@quickstart ~]$ mkdir ~/TP_MapRed
[cloudera@quickstart ~]$ sudo mount -t vmhgfs .host:/TP_MapRed ~/TP_MapRed
[cloudera@quickstart ~]$ ls -la ~/TP_MapRed
total 450616
drwxrwxrwx  1 root     root          4096 Mar 29  2020 .
drwxrwxr-x 28 cloudera cloudera      4096 May  6 06:24 ..
drwxrwxrwx  1 root     root             0 Dec  8  2019 build
-rwxrwxrwx  1 root     root          2278 Sep  1  2015 Makefile
-rwxrwxrwx  1 root     root          1401 Mar  7  2019 maman.txt
-rwxrwxrwx  1 root     root     461369589 Mar 28  2018 Vol1.csv
-rwxrwxrwx  1 root     root         23392 Apr 19  2018 vol.csv
-rwxrwxrwx  1 root     root          5164 Apr  3  2018 wordcount.jar
-rwxrwxrwx  1 root     root          4713 Sep  1  2015 WordCount.java
```

2. Créer le dossier HDFS : /user/cloudera/wordcount/input

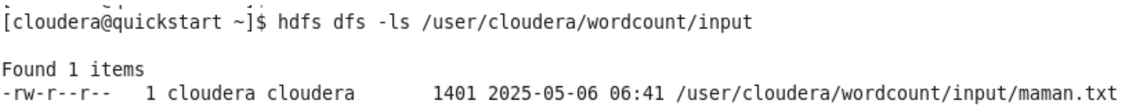


```
[cloudera@quickstart ~]$ hdfs dfs -mkdir -p /user/cloudera/wordcount/input
[cloudera@quickstart ~]$ 
```

3. Charger le fichier texte maman.txt dans le dossier HDFS /user/cloudera/wordcount/input



```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/TP_MapRed/maman.txt /user/cloudera/wordcount/input
```



```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/wordcount/input

Found 1 items
-rw-r--r--   1 cloudera cloudera       1401 2025-05-06 06:41 /user/cloudera/wordcount/input/maman.txt
```

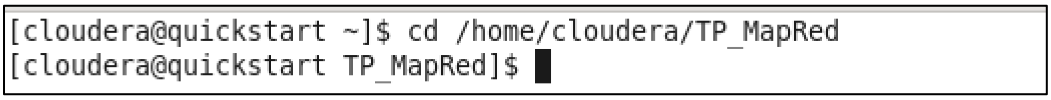
4. Exercices Pratiques MapReduce

PARTIE 1 : WordCount et WordTotal

4.1 Exécution du Job WordCount

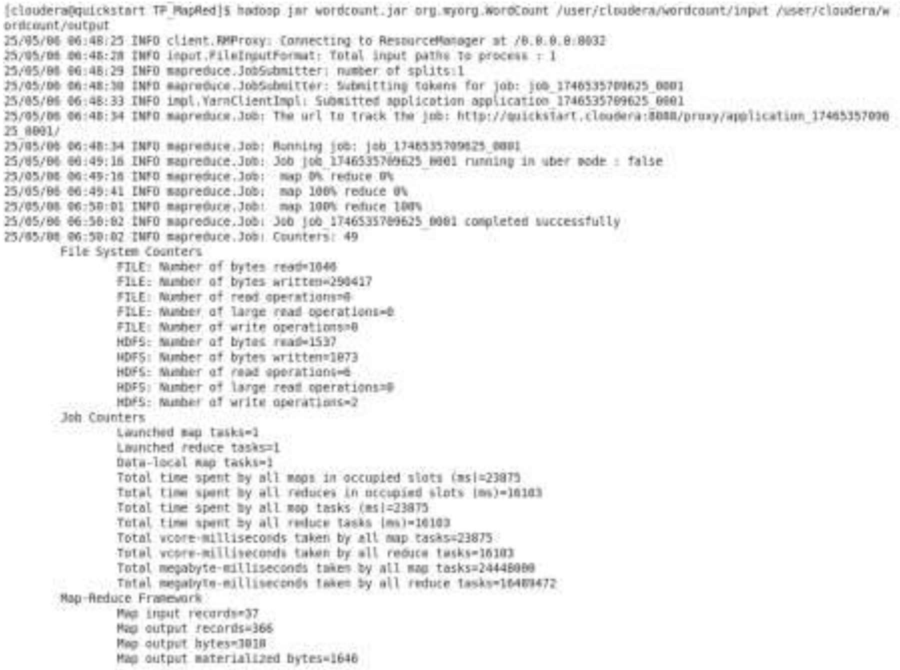
Le programme WordCount est l'exemple classique d'application MapReduce. Il compte les occurrences de chaque mot dans un ensemble de textes.

1. Se positionner dans le dossier du TP :

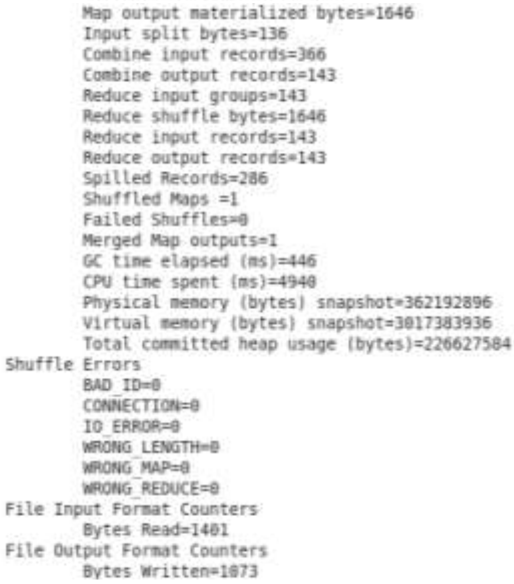


```
[cloudera@quickstart ~]$ cd /home/cloudera/TP_MapRed
[cloudera@quickstart TP_MapRed]$ 
```

2.Exécution de wordcount.jar dont la classe principale est WordCount du package org.myorg :

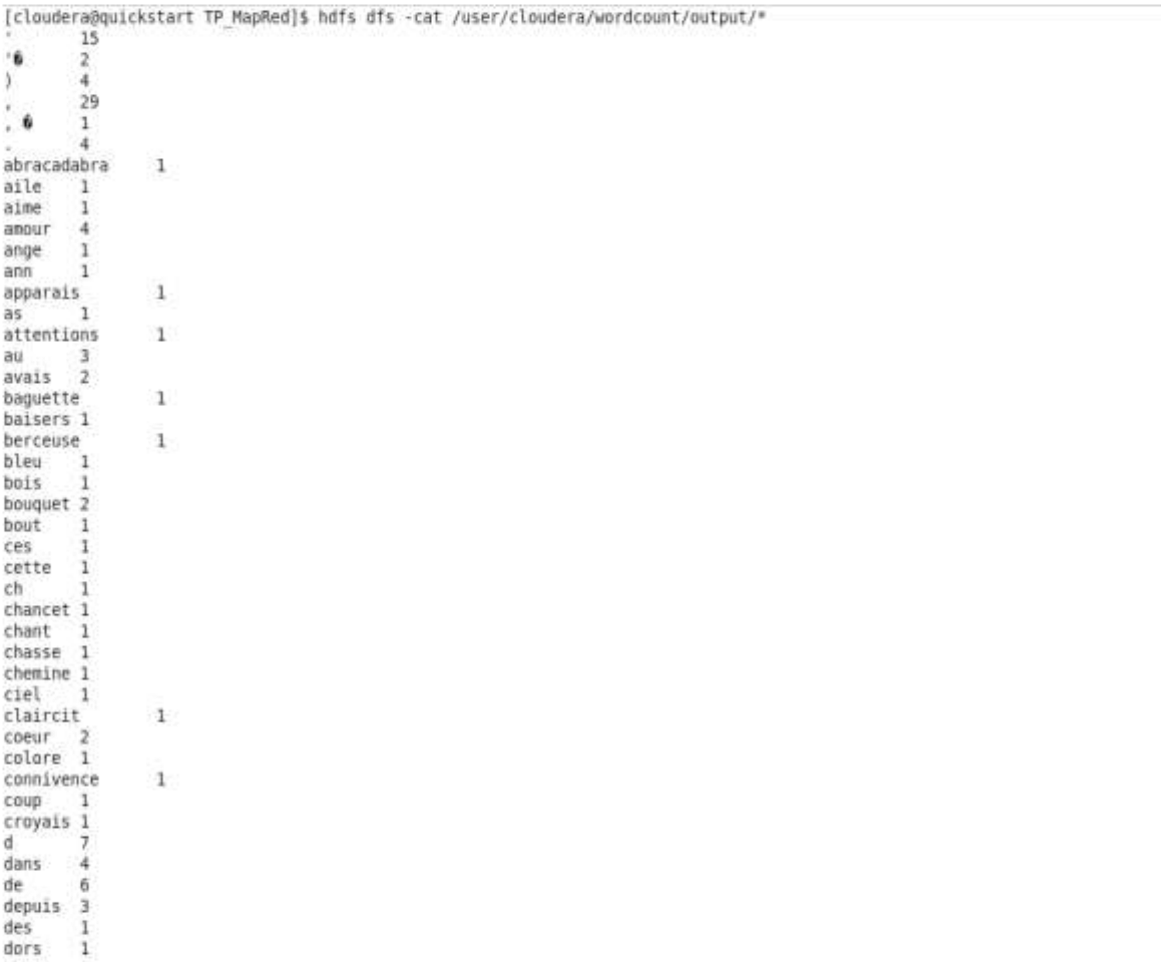


```
[cloudera@quickstart TP_MapRed]$ hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input /user/cloudera/w
ordcount/output
25/05/06 06:48:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 06:48:28 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 06:48:29 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 06:48:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0001
25/05/06 06:48:33 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0001
25/05/06 06:48:34 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_17465357096
25_0001/
25/05/06 06:48:34 INFO mapreduce.Job: Running job: job_1746535709625_0001
25/05/06 06:49:16 INFO mapreduce.Job: Job job_1746535709625_0001 running in uber mode : false
25/05/06 06:49:16 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 06:49:41 INFO mapreduce.Job:  map 100% reduce 0%
25/05/06 06:50:01 INFO mapreduce.Job:  map 100% reduce 100%
25/05/06 06:50:02 INFO mapreduce.Job: Job job_1746535709625_0001 completed successfully
25/05/06 06:50:02 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1646
                FILE: Number of bytes written=290417
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1537
                HDFS: Number of bytes written=1073
                HDFS: Number of read operations=6
                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)=23875
                Total time spent by all reduces in occupied slots (ms)=16103
                Total time spent by all map tasks (ms)=23875
                Total time spent by all reduce tasks (ms)=16103
                Total vcore-milliseconds taken by all map tasks=23875
                Total vcore-milliseconds taken by all reduce tasks=16103
                Total megabyte-milliseconds taken by all map tasks=24448000
                Total megabyte-milliseconds taken by all reduce tasks=16489472
        Map-Reduce Framework
                Map input records=37
                Map output records=366
                Map output bytes=3018
                Map output materialized bytes=1646
```



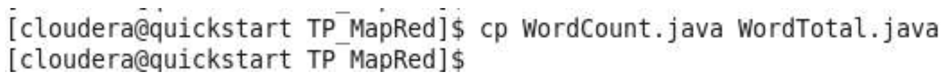
```
                Map output materialized bytes=1646
                Input split bytes=136
                Combine input records=366
                Combine output records=143
                Reduce input groups=143
                Reduce shuffle bytes=1646
                Reduce input records=143
                Reduce output records=143
                Spilled Records=286
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=446
                CPU time spent (ms)=4940
                Physical memory (bytes) snapshot=362192896
                Virtual memory (bytes) snapshot=3017383936
                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=1401
        File Output Format Counters
                Bytes Written=1073
```

3. Afficher le contenu des fichiers résultats dans HDFS :



```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/wordcount/output/*
'       15
'�      2
)       4
,       29
, �     1
.       4
abracadabra     1
aile    1
aime    1
amour   4
ange    1
ann     1
apparais        1
as      1
attentions      1
au      3
avais   2
baguette        1
baisers 1
berceuse        1
bleu    1
bois    1
bouquet 2
bout    1
ces     1
cette   1
ch      1
chancet 1
chant   1
chasse  1
chemine 1
ciel    1
claircit        1
coeur   2
colore  1
connivence      1
coup    1
croyais 1
d       7
dans    4
de      6
depuis  3
des     1
dors    1
```

4.2 Création du Programme WordTotal

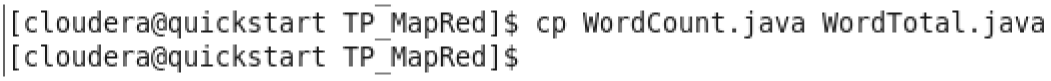


```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java WordTotal.java
[cloudera@quickstart TP_MapRed]$
```

Dans cette partie, nous allons adapter WordCount pour compter le nombre total de mots dans les

fichiers (et non pas le nombre d'occurrences de chaque mot).

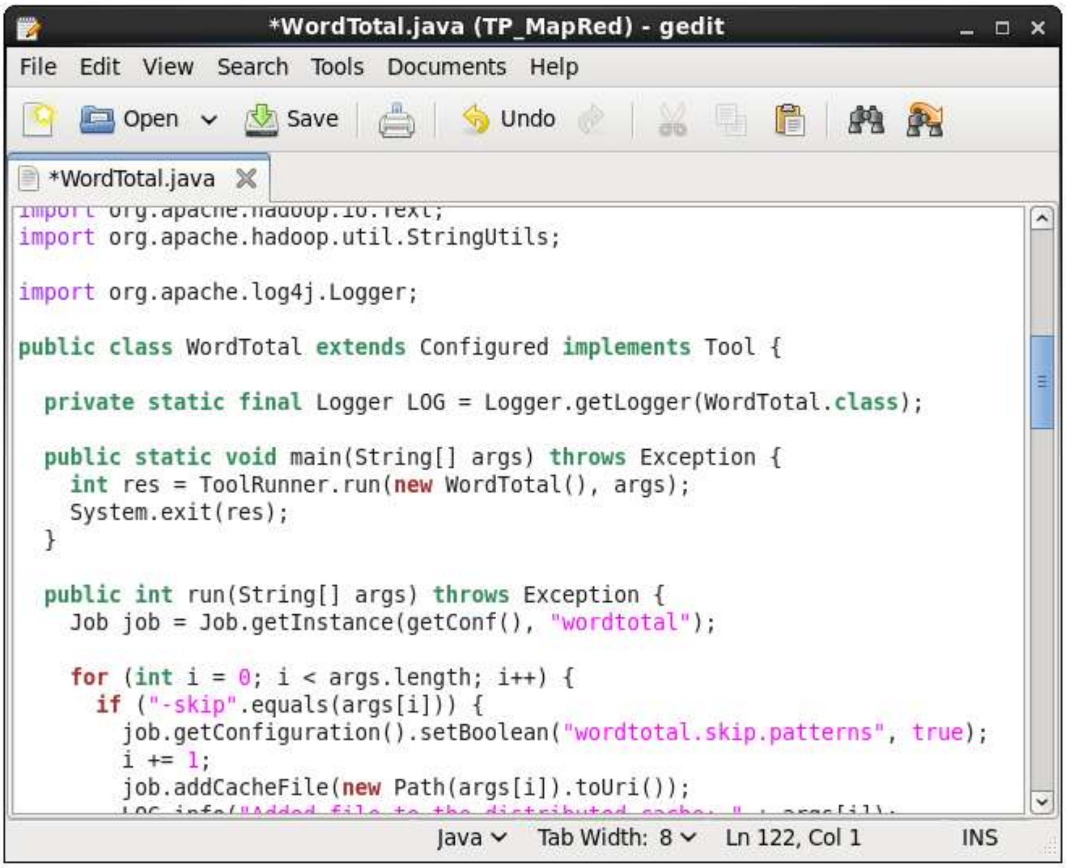
1. Dans le dossier Tp_MapRed, faire une copie de WordCount.java dans WordTotal.java



```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java WordTotal.java
[cloudera@quickstart TP_MapRed]$
```

2. Dans WordTotal.java, remplacer « WordCount » par « WordTotal » et « wordcount » par« wordtotal »

3. Modifier les méthodes Map et Reduce dans WordTotal.java pour calculer le nombre total de mots.



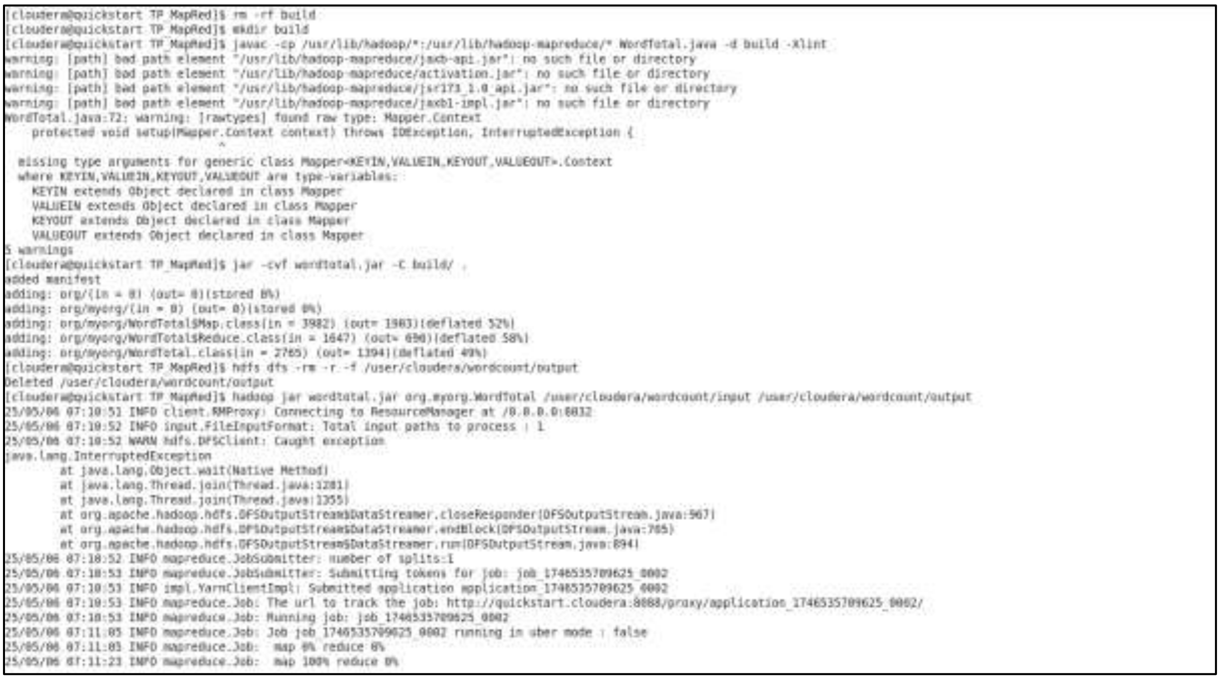
*WordTotal.java (TP_MapRed) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*WordTotal.java
Java
Tab Width: 8
Ln 122, Col 1
INS

```java
import org.apache.hadoop.io.Text;
import org.apache.hadoop.util.StringUtils;

import org.apache.log4j.Logger;

public class WordTotal extends Configured implements Tool {

  private static final Logger LOG = Logger.getLogger(WordTotal.class);

  public static void main(String[] args) throws Exception {
    int res = ToolRunner.run(new WordTotal(), args);
    System.exit(res);
  }

  public int run(String[] args) throws Exception {
    Job job = Job.getInstance(getConf(), "wordtotal");

    for (int i = 0; i < args.length; i++) {
      if ("-skip".equals(args[i])) {
        job.getConfiguration().setBoolean("wordtotal.skip.patterns", true);
        i += 1;
        job.addCacheFile(new Path(args[i]).toUri());
        LOG.info("Added file to the distributed cache: " + args[i]);
```

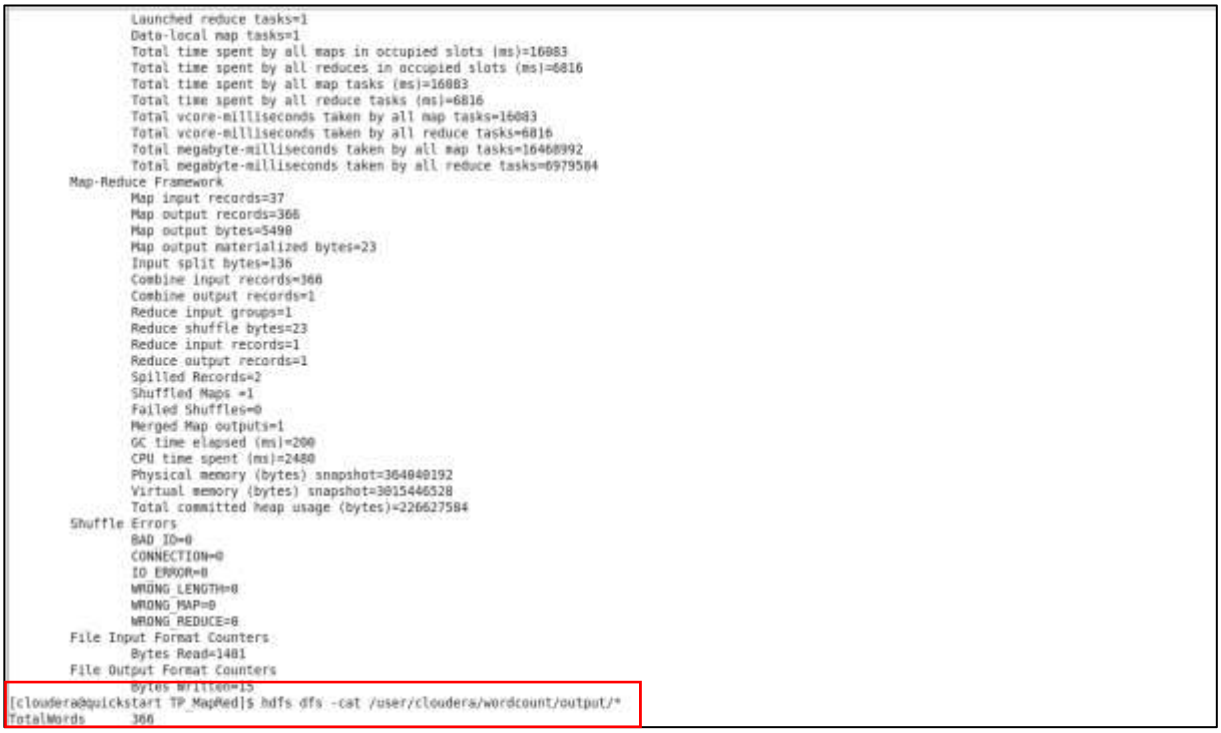
TP02 – MapReduce : Implémentation et Exploration

Indication : Tous les tuples produits par la tâche Map doivent avoir la même clé (par exemple: « Nombre de mots »)

4. Compiler le code source et exécuter votre Job :



```
[cloudera@quickstart TP_MapRed]$ rm -rf build
[cloudera@quickstart TP_MapRed]$ mkdir build
[cloudera@quickstart TP_MapRed]$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* WordTotal.java -d build -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no such file or directory
WordTotal.java:72: warning: [rawtypes] found raw type: Mapper.Context
    protected void setup(Mapper.Context context) throws IOException, InterruptedException {
                               ^
  missing type arguments for generic class Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>.Context
  where KEYIN,VALUEIN,KEYOUT,VALUEOUT are type-variables:
    KEYIN extends Object declared in class Mapper
    VALUEIN extends Object declared in class Mapper
    KEYOUT extends Object declared in class Mapper
    VALUEOUT extends Object declared in class Mapper
5 warnings
[cloudera@quickstart TP_MapRed]$ jar -cvf wordtotal.jar -C build/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordTotal$Map.class(in = 3982) (out= 1903)(deflated 52%)
adding: org/myorg/WordTotal$Reduce.class(in = 1647) (out= 690)(deflated 58%)
adding: org/myorg/WordTotal.class(in = 2765) (out= 1394)(deflated 49%)
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/wordcount/output
Deleted /user/cloudera/wordcount/output
[cloudera@quickstart TP_MapRed]$ hadoop jar wordtotal.jar org.myorg.WordTotal /user/cloudera/wordcount/input /user/cloudera/wordcount/output
25/05/06 07:10:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 07:10:52 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 07:10:52 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.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
25/05/06 07:10:52 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 07:10:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0002
25/05/06 07:10:53 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0002
25/05/06 07:10:53 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0002/
25/05/06 07:10:53 INFO mapreduce.Job: Running job: job_1746535709625_0002
25/05/06 07:11:05 INFO mapreduce.Job: Job job_1746535709625_0002 running in uber mode : false
25/05/06 07:11:05 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 07:11:23 INFO mapreduce.Job:  map 100% reduce 0%
```



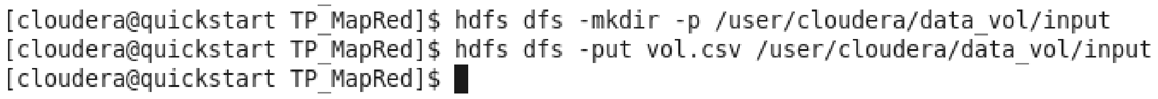
```
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=16083
                Total time spent by all reduces in occupied slots (ms)=6816
                Total time spent by all map tasks (ms)=16083
                Total time spent by all reduce tasks (ms)=6816
                Total vcore-milliseconds taken by all map tasks=16083
                Total vcore-milliseconds taken by all reduce tasks=6816
                Total megabyte-milliseconds taken by all map tasks=16468992
                Total megabyte-milliseconds taken by all reduce tasks=6979584
        Map-Reduce Framework
                Map input records=37
                Map output records=366
                Map output bytes=5490
                Map output materialized bytes=23
                Input split bytes=136
                Combine input records=366
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=23
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=200
                CPU time spent (ms)=2480
                Physical memory (bytes) snapshot=364040192
                Virtual memory (bytes) snapshot=3015446528
                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=1481
        File Output Format Counters
                Bytes Written=15
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/wordcount/output/*
TotalWords      366
```

PARTIE 2 : Analyse de Données de Vol

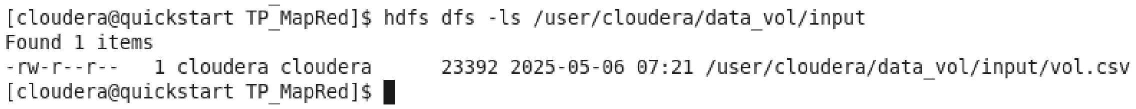
Dans cette partie, nous allons travailler avec un fichier CSV contenant des données de vols aériens.

4.3 Préparation des Données

1. Charger le fichier vol.csv dans le dossier HDFS à créer : /user/cloudera/data_vol/input



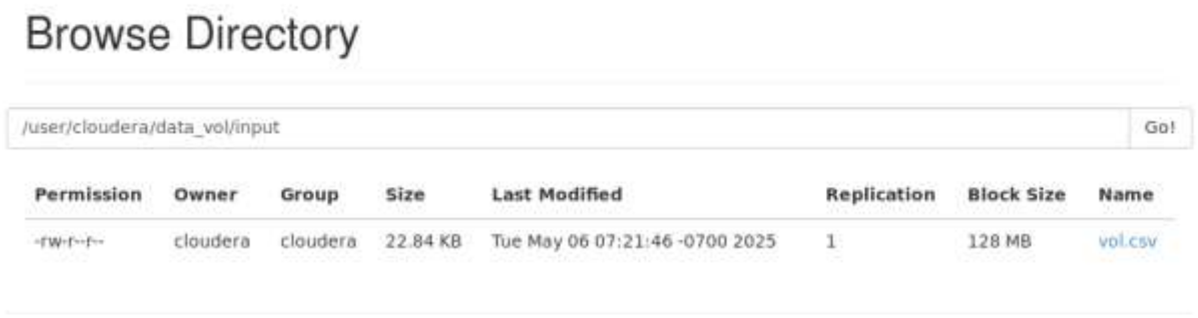
```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -mkdir -p /user/cloudera/data_vol/input
[cloudera@quickstart TP_MapRed]$ hdfs dfs -put vol.csv /user/cloudera/data_vol/input
[cloudera@quickstart TP_MapRed]$ 
```



```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -ls /user/cloudera/data_vol/input
Found 1 items
-rw-r--r--   1 cloudera cloudera      23392 2025-05-06 07:21 /user/cloudera/data_vol/input/vol.csv
[cloudera@quickstart TP_MapRed]$ 
```

2. Afficher le contenu du dossier HDFS :

3. Explorer l’arborescence HDFS via l’interface Web : http://localhost:50070 → Utilities



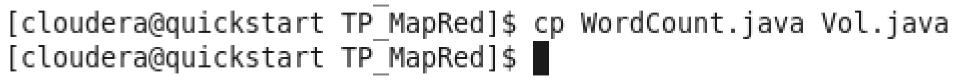
Browse Directory

/user/cloudera/data_vol/input

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	22.84 KB	Tue May 06 07:21:46 -0700 2025	1	128 MB	vol.csv

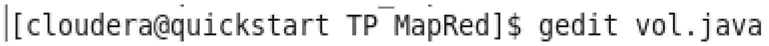
4.4 Analyse des Vols par Aéroport de Départ

1. Enregistrer WordCount.java sous le nom Vol.java

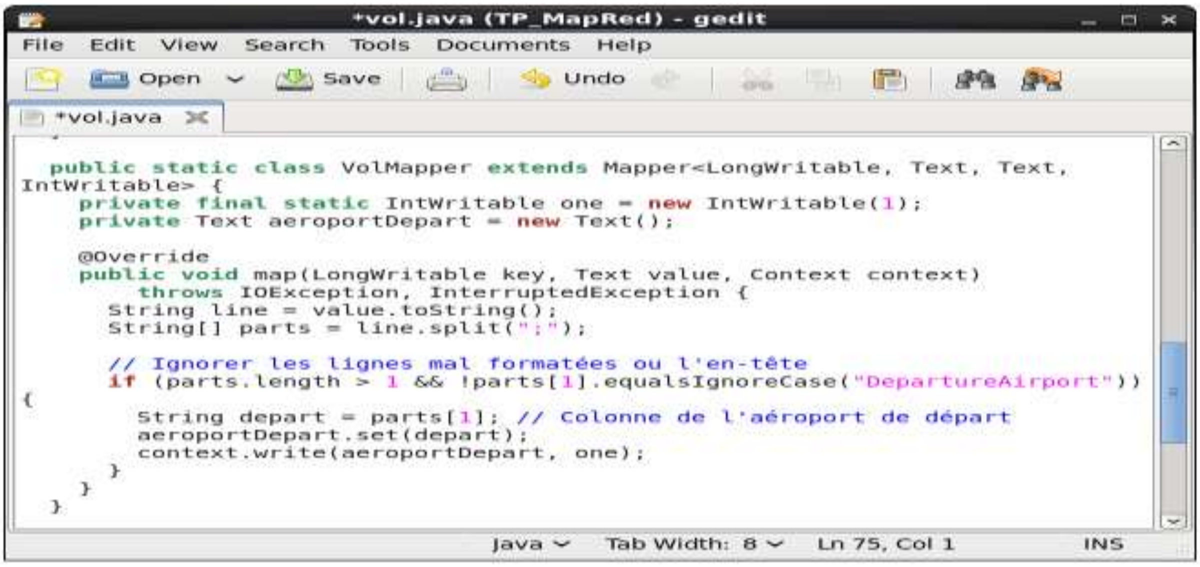


```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java Vol.java
[cloudera@quickstart TP_MapRed]$ 
```

2. Modifier Vol.java pour calculer le nombre de vols en partance de chaque aéroport.



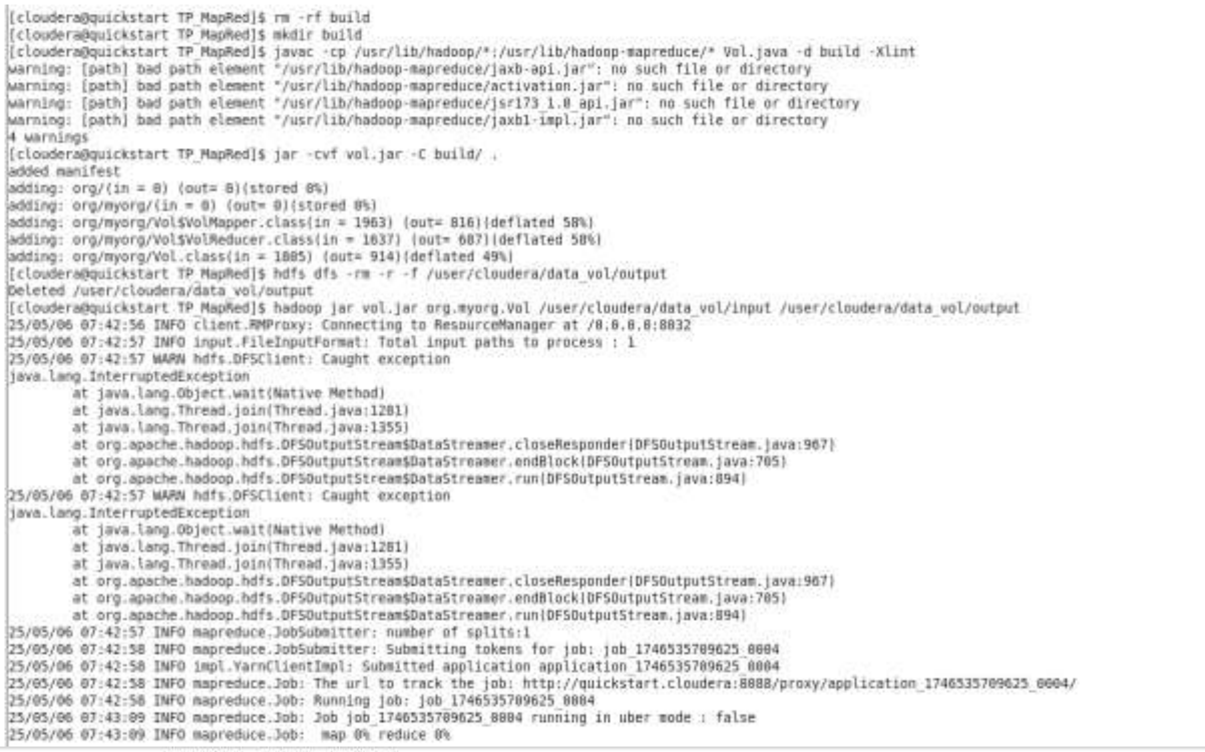
```
[cloudera@quickstart TP_MapRed]$ gedit vol.java
```



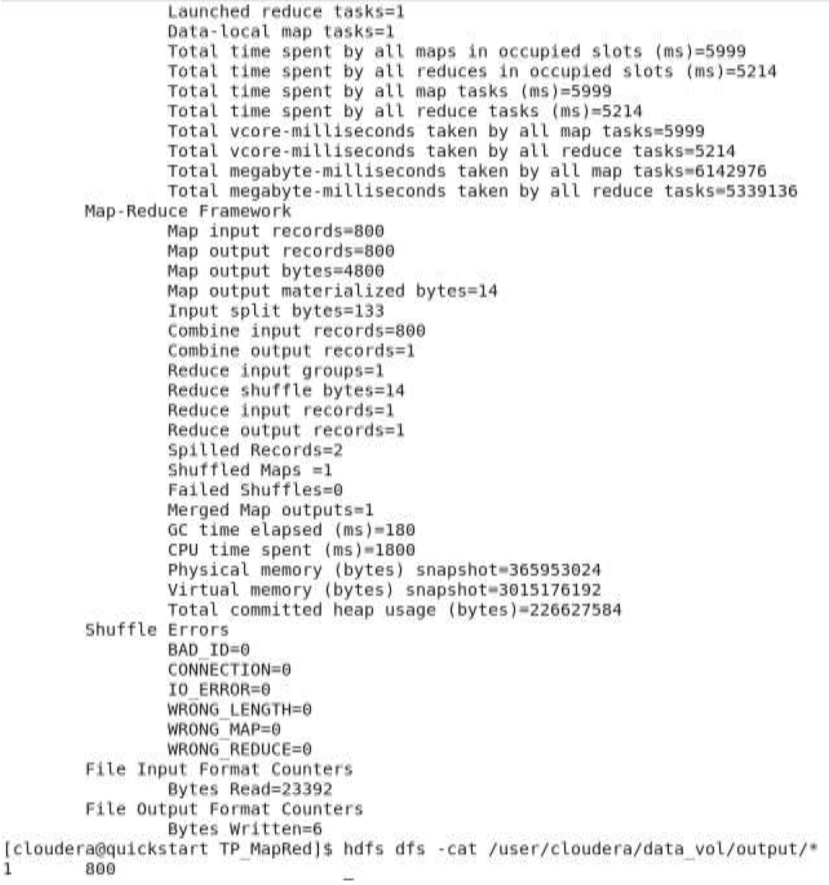
*vol.java (TP_MapRed) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
*vol.java

```
  public static class VolMapper extends Mapper<LongWritable, Text, Text,
IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text aeroportDepart = new Text();

    @Override
    public void map(LongWritable key, Text value, Context context)
      throws IOException, InterruptedException {
      String line = value.toString();
      String[] parts = line.split(";");

      // Ignorer les lignes mal formatées ou l'en-tête
      if (parts.length > 1 && !parts[1].equalsIgnoreCase("DepartureAirport"))
{
        String depart = parts[1]; // Colonne de l'aéroport de départ
        aeroportDepart.set(depart);
        context.write(aeroportDepart, one);
      }
    }
  }
```

3. Compiler et exécuter le programme :



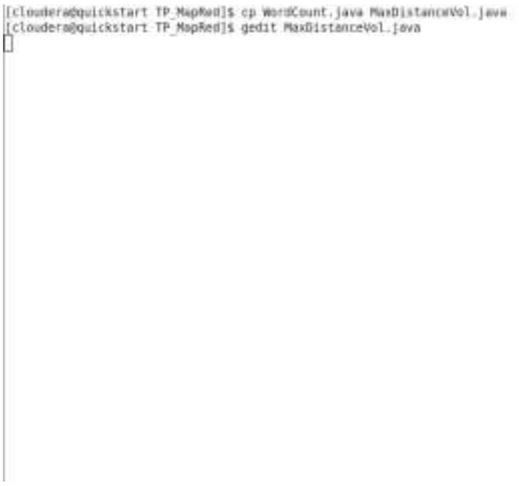
```
[cloudera@quickstart TP_MapRed]$ rm -rf build
[cloudera@quickstart TP_MapRed]$ mkdir build
[cloudera@quickstart TP_MapRed]$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* Vol.java -d build -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no such file or directory
4 warnings
[cloudera@quickstart TP_MapRed]$ jar -cvf vol.jar -C build/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/Vol$VolMapper.class(in = 1963) (out= 816)(deflated 58%)
adding: org/myorg/Vol$VolReducer.class(in = 1637) (out= 687)(deflated 58%)
adding: org/myorg/Vol.class(in = 1805) (out= 914)(deflated 49%)
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/data_vol/output
Deleted /user/cloudera/data_vol/output
[cloudera@quickstart TP_MapRed]$ hadoop jar vol.jar org.myorg.Vol /user/cloudera/data_vol/input /user/cloudera/data_vol/output
25/05/06 07:42:56 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 07:42:57 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 07:42:57 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.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
	at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
	at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
25/05/06 07:42:57 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
	at java.lang.Object.wait(Native Method)
	at java.lang.Thread.join(Thread.java:1281)
	at java.lang.Thread.join(Thread.java:1355)
	at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
	at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
	at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
25/05/06 07:42:57 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 07:42:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0004
25/05/06 07:42:58 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0004
25/05/06 07:42:58 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0004/
25/05/06 07:42:58 INFO mapreduce.Job: Running job: job_1746535709625_0004
25/05/06 07:43:09 INFO mapreduce.Job: Job job_1746535709625_0004 running in uber mode : false
25/05/06 07:43:09 INFO mapreduce.Job:  map 0% reduce 0%
```



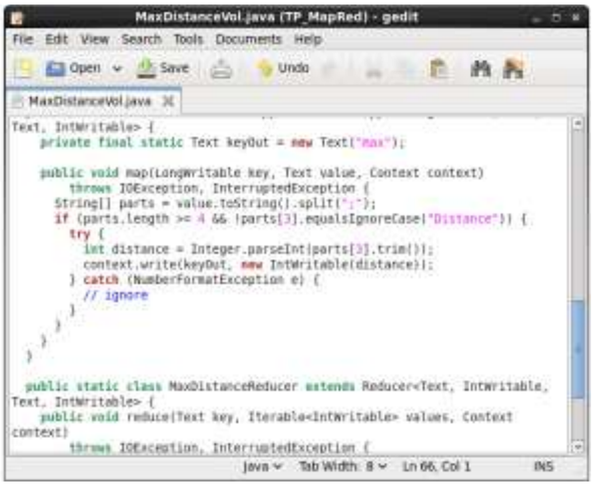
```
		Launched reduce tasks=1
		Data-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=5999
		Total time spent by all reduces in occupied slots (ms)=5214
		Total time spent by all map tasks (ms)=5999
		Total time spent by all reduce tasks (ms)=5214
		Total vcore-milliseconds taken by all map tasks=5999
		Total vcore-milliseconds taken by all reduce tasks=5214
		Total megabyte-milliseconds taken by all map tasks=6142976
		Total megabyte-milliseconds taken by all reduce tasks=5339136
	Map-Reduce Framework
		Map input records=800
		Map output records=800
		Map output bytes=4800
		Map output materialized bytes=14
		Input split bytes=133
		Combine input records=800
		Combine output records=1
		Reduce input groups=1
		Reduce shuffle bytes=14
		Reduce input records=1
		Reduce output records=1
		Spilled Records=2
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=180
		CPU time spent (ms)=1800
		Physical memory (bytes) snapshot=365953024
		Virtual memory (bytes) snapshot=3015176192
		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=23392
	File Output Format Counters
		Bytes Written=6
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/data_vol/output/*
1	800
```

4.5 Analyse de la Distance Maximale

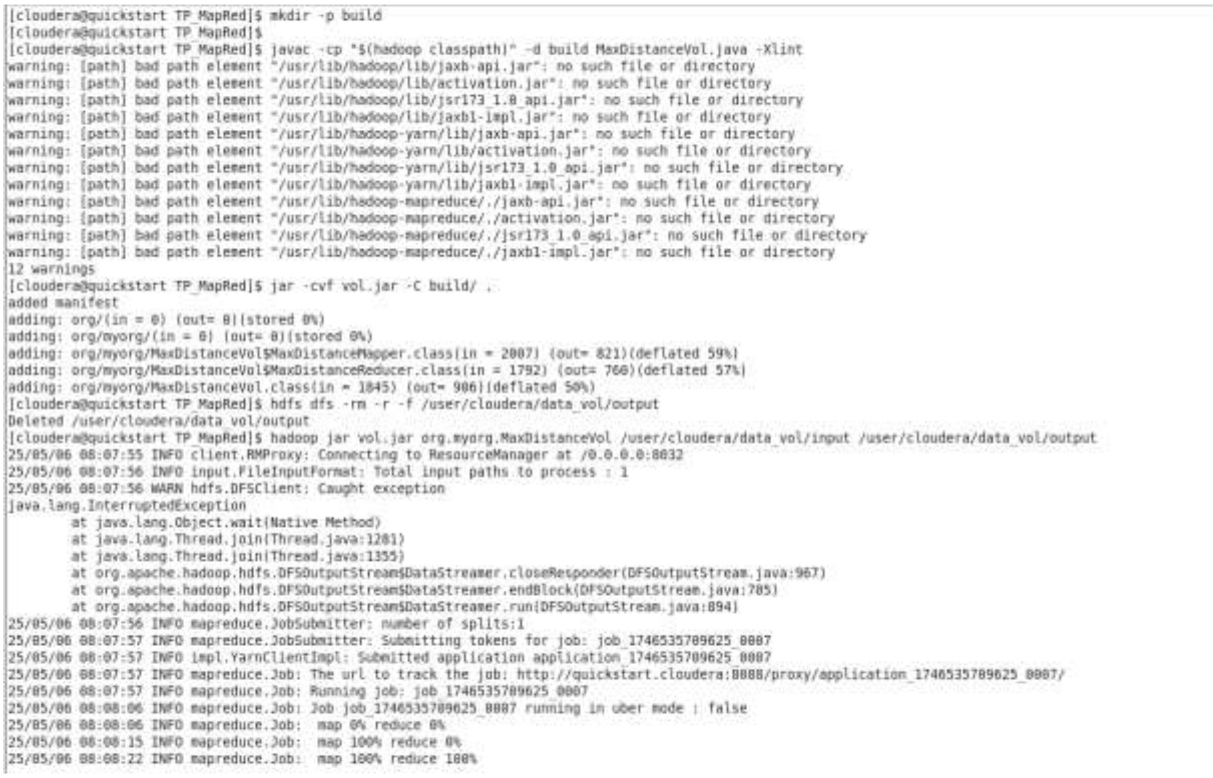
Créez une version modifiée de votre programme pour calculer la distance maximale de vol.



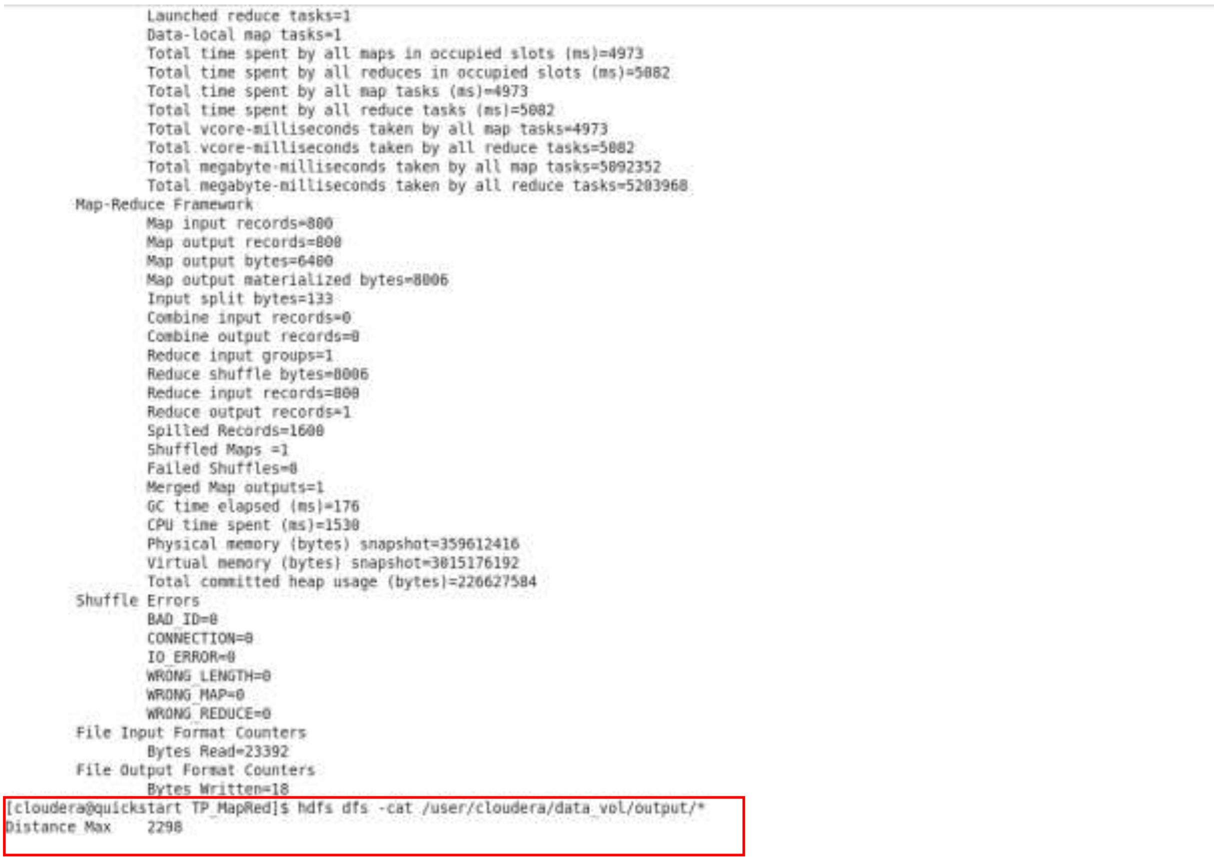
```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java MaxDistanceVol.java
[cloudera@quickstart TP_MapRed]$ gedit MaxDistanceVol.java
```



```
Text, IntWritable> {
    private final static Text keyOut = new Text("max");

    public void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException {
      String[] parts = value.toString().split(";");
      if (parts.length >= 4 && !parts[3].equalsIgnoreCase("Distance")) {
        try {
          int distance = Integer.parseInt(parts[3].trim());
          context.write(keyOut, new IntWritable(distance));
        } catch (NumberFormatException e) {
          // ignore
        }
      }
    }
  }

  public static class MaxDistanceReducer extends Reducer<Text, IntWritable,
Text, IntWritable> {
    public void reduce(Text key, Iterable<IntWritable> values, Context
context)
        throws IOException, InterruptedException {
```



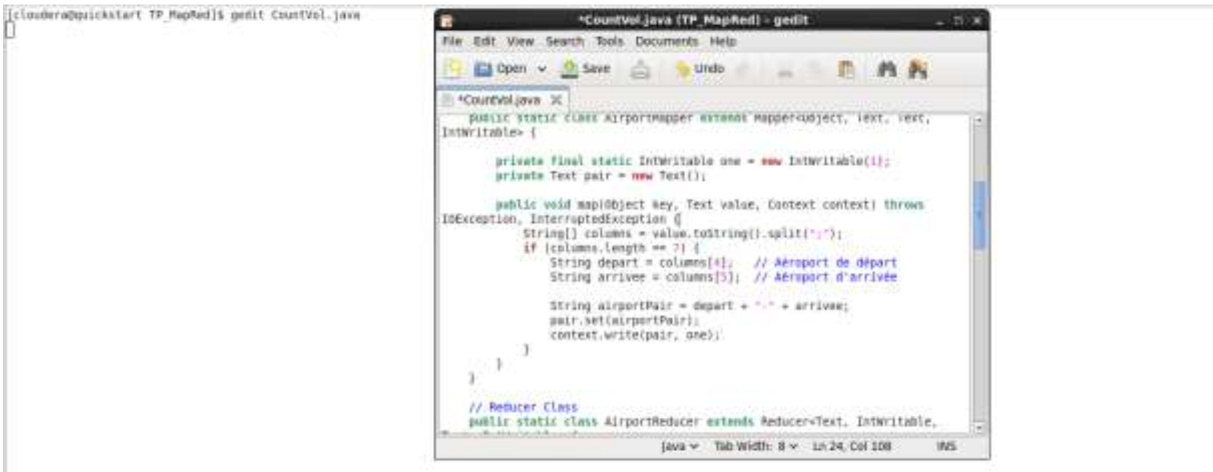
```
[cloudera@quickstart TP_MapRed]$ mkdir -p build
[cloudera@quickstart TP_MapRed]$
[cloudera@quickstart TP_MapRed]$ javac -cp "$(hadoop classpath)" -d build MaxDistanceVol.java -Xlint
warning: [path] bad path element "/usr/lib/hadoop/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jaxb1-impl.jar": no such file or directory
12 warnings
[cloudera@quickstart TP_MapRed]$ jar -cvf vol.jar -C build/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/MaxDistanceVol$MaxDistanceMapper.class(in = 2007) (out= 821)(deflated 59%)
adding: org/myorg/MaxDistanceVol$MaxDistanceReducer.class(in = 1792) (out= 760)(deflated 57%)
adding: org/myorg/MaxDistanceVol.class(in = 1845) (out= 906)(deflated 50%)
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/data_vol/output
Deleted /user/cloudera/data_vol/output
[cloudera@quickstart TP_MapRed]$ hadoop jar vol.jar org.myorg.MaxDistanceVol /user/cloudera/data_vol/input /user/cloudera/data_vol/output
25/05/06 08:07:55 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 08:07:56 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 08:07:56 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.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
25/05/06 08:07:56 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 08:07:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0007
25/05/06 08:07:57 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0007
25/05/06 08:07:57 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0007/
25/05/06 08:07:57 INFO mapreduce.Job: Running job: job_1746535709625_0007
25/05/06 08:08:06 INFO mapreduce.Job: Job job_1746535709625_0007 running in uber mode : false
25/05/06 08:08:06 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 08:08:15 INFO mapreduce.Job:  map 100% reduce 0%
25/05/06 08:08:22 INFO mapreduce.Job:  map 100% reduce 100%
```



```
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4973
                Total time spent by all reduces in occupied slots (ms)=5082
                Total time spent by all map tasks (ms)=4973
                Total time spent by all reduce tasks (ms)=5082
                Total vcore-milliseconds taken by all map tasks=4973
                Total vcore-milliseconds taken by all reduce tasks=5082
                Total megabyte-milliseconds taken by all map tasks=5092352
                Total megabyte-milliseconds taken by all reduce tasks=5203968
        Map-Reduce Framework
                Map input records=800
                Map output records=800
                Map output bytes=6400
                Map output materialized bytes=8006
                Input split bytes=133
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=8006
                Reduce input records=800
                Reduce output records=1
                Spilled Records=1600
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=176
                CPU time spent (ms)=1530
                Physical memory (bytes) snapshot=359612416
                Virtual memory (bytes) snapshot=3015176192
                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=23392
        File Output Format Counters
                Bytes Written=18
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/data_vol/output/*
Distance Max    2298
```

4.6 Analyse des Paires d’Aéroports

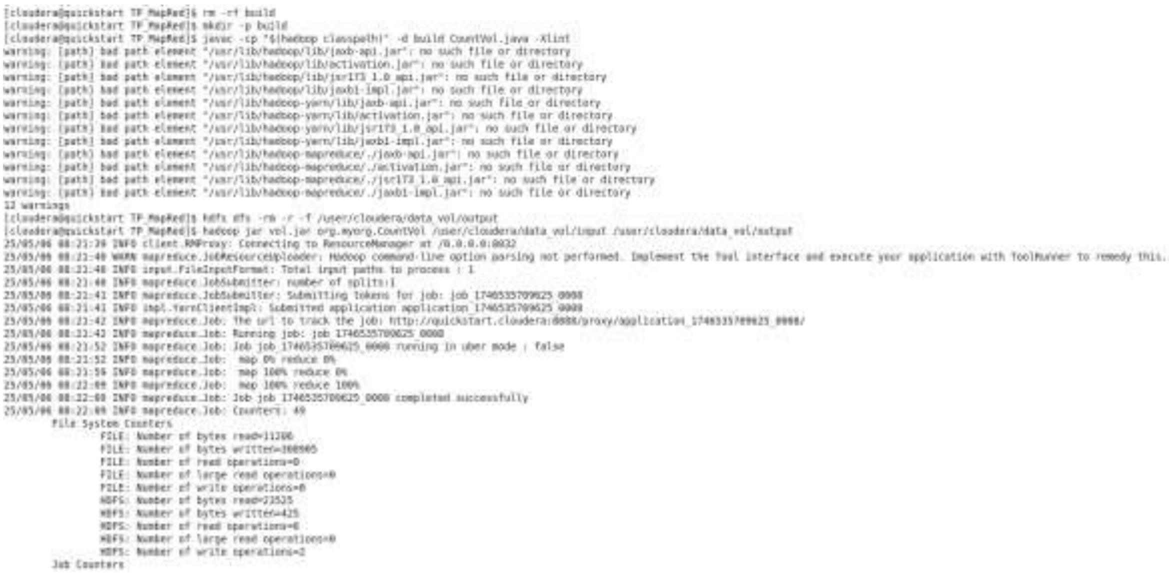
Modifiez votre programme pour compter le nombre de vols pour chaque paire d’aéroports, sans distinction entre départ et arrivée.



```
[cloudera@quickstart TP_MapRed]$ gedit CountVol.java
```

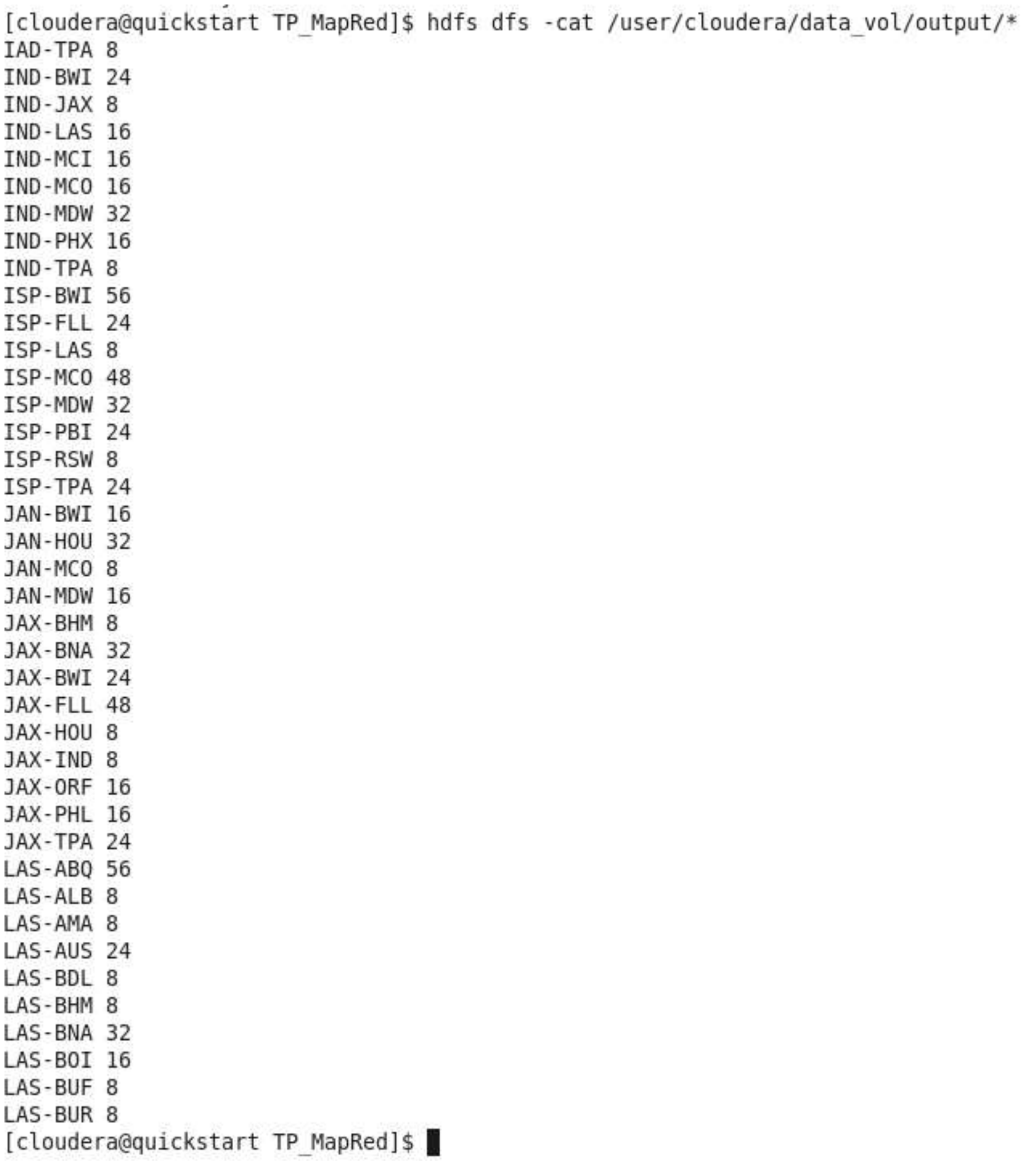
*CountVol.java (TP_MapRed) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*CountVol.java
Java
Tab Width: 8
Ln 24, Col 108
INS

```
    public static class AirportMapper extends Mapper<Object, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text pair = new Text();

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            String[] columns = value.toString().split(";");
            if (columns.length == 7) {
                String depart = columns[4];   // Aéroport de départ
                String arrivee = columns[5];  // Aéroport d'arrivée

                String airportPair = depart + "-" + arrivee;
                pair.set(airportPair);
                context.write(pair, one);
            }
        }
    }

    // Reducer Class
    public static class AirportReducer extends Reducer<Text, IntWritable,
```



```
[cloudera@quickstart TP_MapRed]$ rm -rf build
[cloudera@quickstart TP_MapRed]$ mkdir -p build
[cloudera@quickstart TP_MapRed]$ javac -cp "$(hadoop classpath)" -d build CountVol.java -Xlint
warning: [path] bad path element "/usr/lib/hadoop/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-yarn/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/./jaxb1-impl.jar": no such file or directory
12 warnings
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/data_vol/output
[cloudera@quickstart TP_MapRed]$ hadoop jar vol.jar org.myorg.CountVol /user/cloudera/data_vol/input /user/cloudera/data_vol/output
25/05/06 08:21:39 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 08:21:40 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
25/05/06 08:21:40 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 08:21:40 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 08:21:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0008
25/05/06 08:21:41 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0008
25/05/06 08:21:42 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0008/
25/05/06 08:21:42 INFO mapreduce.Job: Running job: job_1746535709625_0008
25/05/06 08:21:52 INFO mapreduce.Job: Job job_1746535709625_0008 running in uber mode : false
25/05/06 08:21:52 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 08:21:59 INFO mapreduce.Job:  map 100% reduce 0%
25/05/06 08:22:09 INFO mapreduce.Job:  map 100% reduce 100%
25/05/06 08:22:09 INFO mapreduce.Job: Job job_1746535709625_0008 completed successfully
25/05/06 08:22:09 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=11206
                FILE: Number of bytes written=308905
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=23525
                HDFS: Number of bytes written=425
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
```

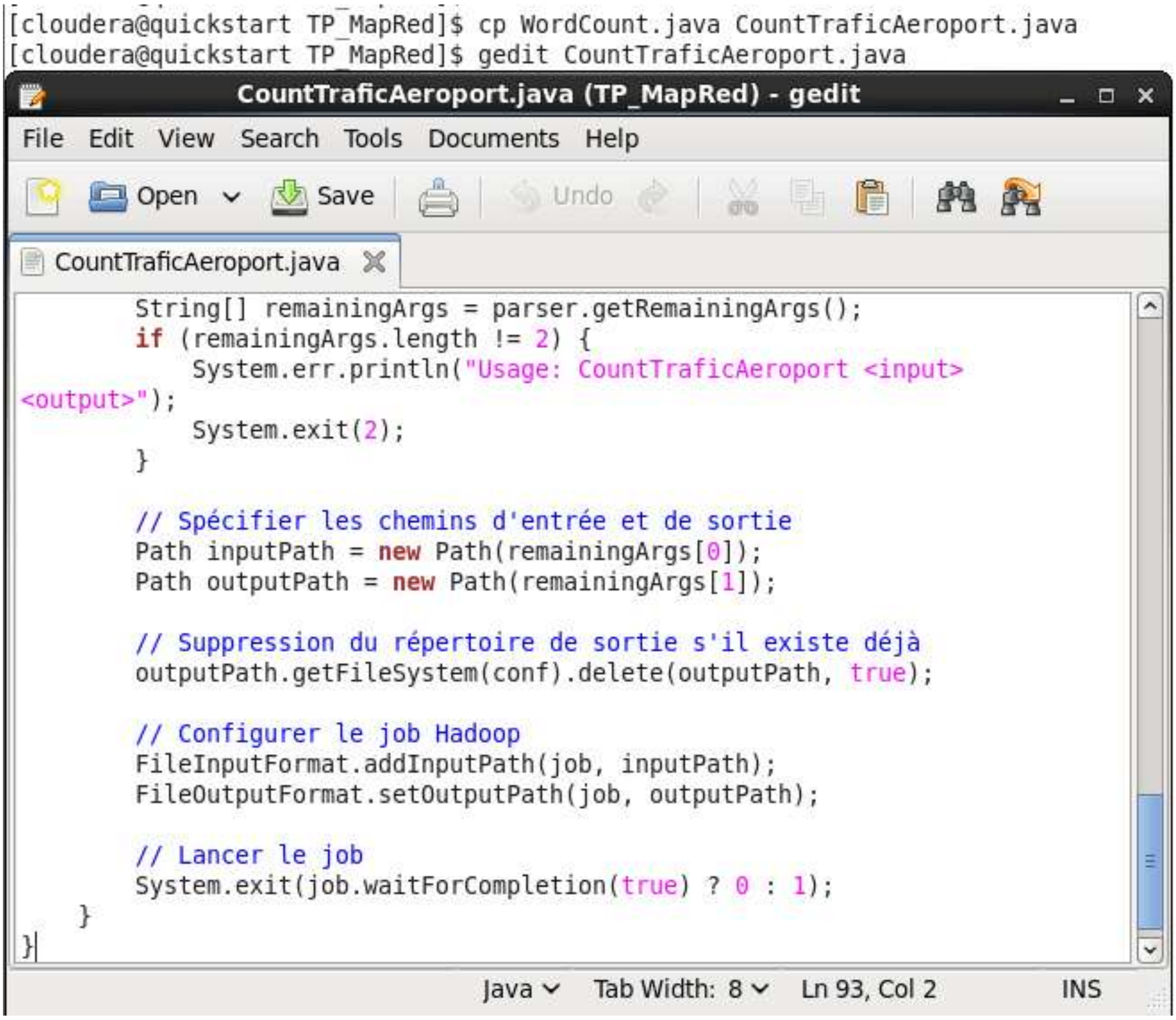
Resulta :



```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/data_vol/output/*
IAD-TPA 8
IND-BWI 24
IND-JAX 8
IND-LAS 16
IND-MCI 16
IND-MCO 16
IND-MDW 32
IND-PHX 16
IND-TPA 8
ISP-BWI 56
ISP-FLL 24
ISP-LAS 8
ISP-MCO 48
ISP-MDW 32
ISP-PBI 24
ISP-RSW 8
ISP-TPA 24
JAN-BWI 16
JAN-HOU 32
JAN-MCO 8
JAN-MDW 16
JAX-BHM 8
JAX-BNA 32
JAX-BWI 24
JAX-FLL 48
JAX-HOU 8
JAX-IND 8
JAX-ORF 16
JAX-PHL 16
JAX-TPA 24
LAS-ABQ 56
LAS-ALB 8
LAS-AMA 8
LAS-AUS 24
LAS-BDL 8
LAS-BHM 8
LAS-BNA 32
LAS-BOI 16
LAS-BUF 8
LAS-BUR 8
[cloudera@quickstart TP_MapRed]$ █
```

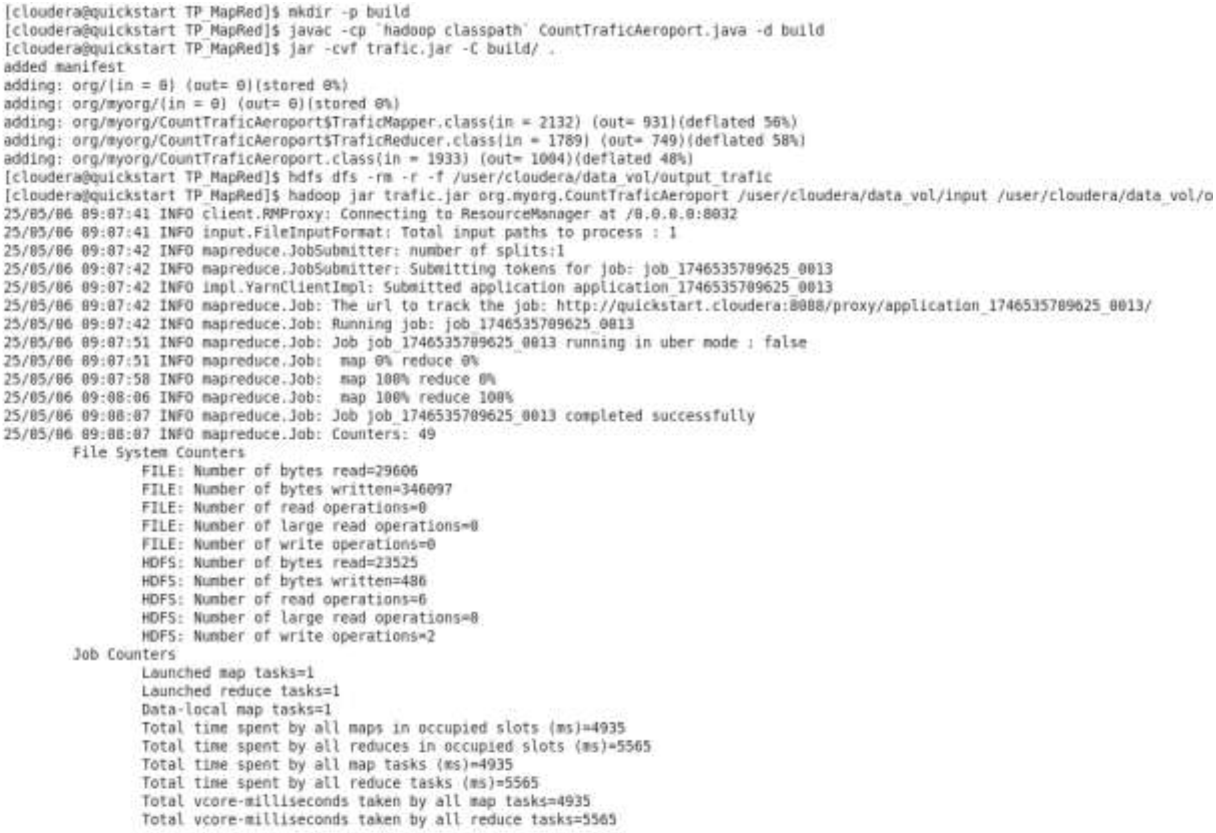
4.7 Analyse du Trafic par Aéroport

Modifiez votre programme pour compter les vols en partance et en arrivée pour chaque aéroport.

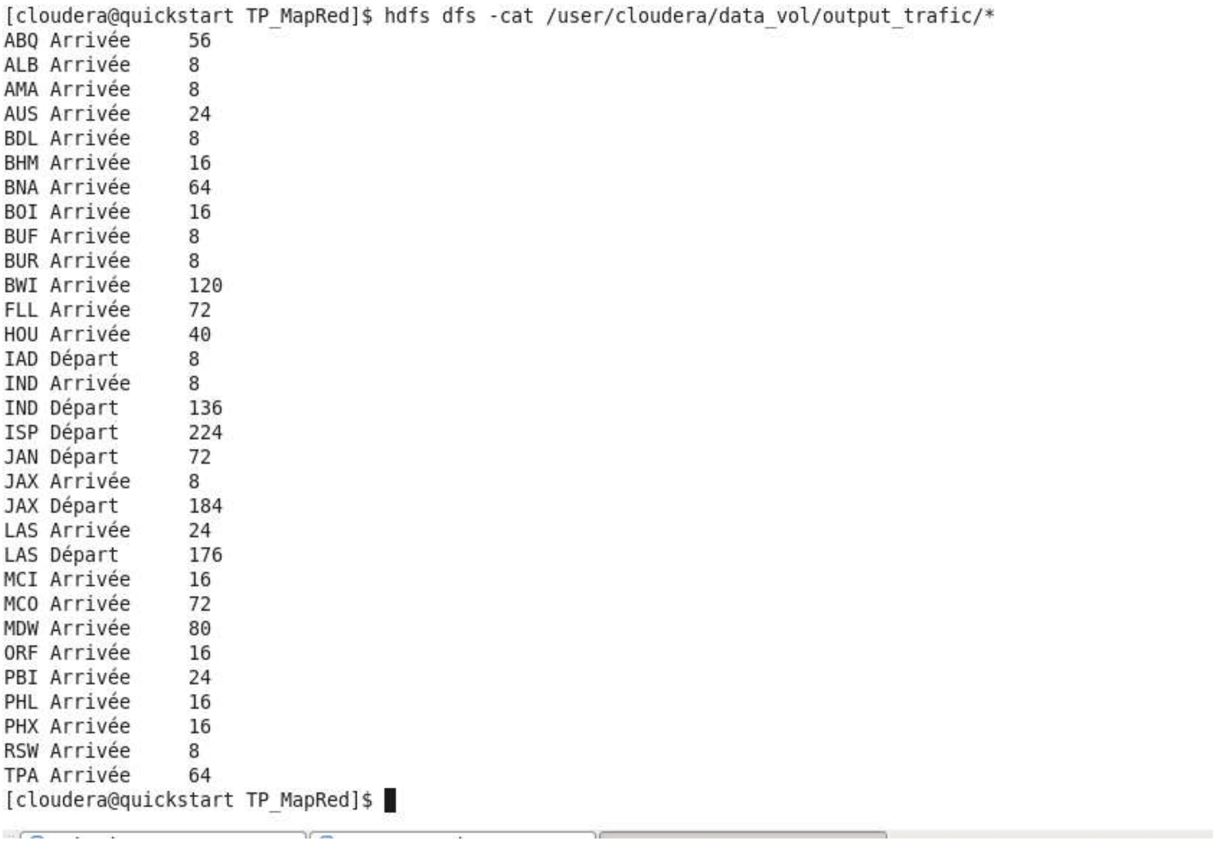


```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java CountTraficAeroport.java
[cloudera@quickstart TP_MapRed]$ gedit CountTraficAeroport.java
```

```
        String[] remainingArgs = parser.getRemainingArgs();
        if (remainingArgs.length != 2) {
            System.err.println("Usage: CountTraficAeroport <input>
<output>");
            System.exit(2);
        }

        // Spécifier les chemins d'entrée et de sortie
        Path inputPath = new Path(remainingArgs[0]);
        Path outputPath = new Path(remainingArgs[1]);

        // Suppression du répertoire de sortie s'il existe déjà
        outputPath.getFileSystem(conf).delete(outputPath, true);

        // Configurer le job Hadoop
        FileInputFormat.addInputPath(job, inputPath);
        FileOutputFormat.setOutputPath(job, outputPath);

        // Lancer le job
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```



```
[cloudera@quickstart TP_MapRed]$ mkdir -p build
[cloudera@quickstart TP_MapRed]$ javac -cp `hadoop classpath` CountTraficAeroport.java -d build
[cloudera@quickstart TP_MapRed]$ jar -cvf trafic.jar -C build/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/CountTraficAeroport$TraficMapper.class(in = 2132) (out= 931)(deflated 56%)
adding: org/myorg/CountTraficAeroport$TraficReducer.class(in = 1789) (out= 749)(deflated 58%)
adding: org/myorg/CountTraficAeroport.class(in = 1933) (out= 1004)(deflated 48%)
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/data_vol/output_trafic
[cloudera@quickstart TP_MapRed]$ hadoop jar trafic.jar org.myorg.CountTraficAeroport /user/cloudera/data_vol/input /user/cloudera/data_vol/o
25/05/06 09:07:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 09:07:41 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 09:07:42 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 09:07:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0013
25/05/06 09:07:42 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0013
25/05/06 09:07:42 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0013/
25/05/06 09:07:42 INFO mapreduce.Job: Running job: job_1746535709625_0013
25/05/06 09:07:51 INFO mapreduce.Job: Job job_1746535709625_0013 running in uber mode : false
25/05/06 09:07:51 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 09:07:58 INFO mapreduce.Job:  map 100% reduce 0%
25/05/06 09:08:06 INFO mapreduce.Job:  map 100% reduce 100%
25/05/06 09:08:07 INFO mapreduce.Job: Job job_1746535709625_0013 completed successfully
25/05/06 09:08:07 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=29606
                FILE: Number of bytes written=346097
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=23525
                HDFS: Number of bytes written=486
                HDFS: Number of read operations=6
                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)=4935
                Total time spent by all reduces in occupied slots (ms)=5565
                Total time spent by all map tasks (ms)=4935
                Total time spent by all reduce tasks (ms)=5565
                Total vcore-milliseconds taken by all map tasks=4935
                Total vcore-milliseconds taken by all reduce tasks=5565
```



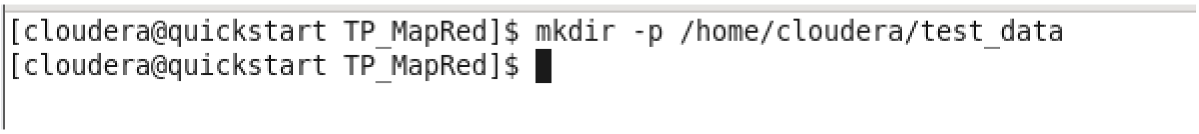
```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -cat /user/cloudera/data_vol/output_trafic/*
ABQ Arrivée     56
ALB Arrivée     8
AMA Arrivée     8
AUS Arrivée     24
BDL Arrivée     8
BHM Arrivée     16
BNA Arrivée     64
BOI Arrivée     16
BUF Arrivée     8
BUR Arrivée     8
BWI Arrivée     120
FLL Arrivée     72
HOU Arrivée     40
IAD Départ      8
IND Arrivée     8
IND Départ      136
ISP Départ      224
JAN Départ      72
JAX Arrivée     8
JAX Départ      184
LAS Arrivée     24
LAS Départ      176
MCI Arrivée     16
MCO Arrivée     72
MDW Arrivée     80
ORF Arrivée     16
PBI Arrivée     24
PHL Arrivée     16
PHX Arrivée     16
RSW Arrivée     8
TPA Arrivée     64
[cloudera@quickstart TP_MapRed]$ █
```

5. Optimisation des Performances

5.1 Exercice : Mesurer l’impact des Combiners

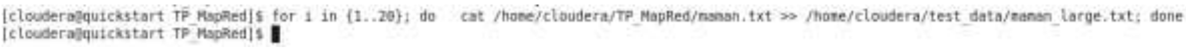
Dans cette activité, nous allons comparer les performances du programme WordCount avec et sans combiner.

1. Préparation des données pour le test :

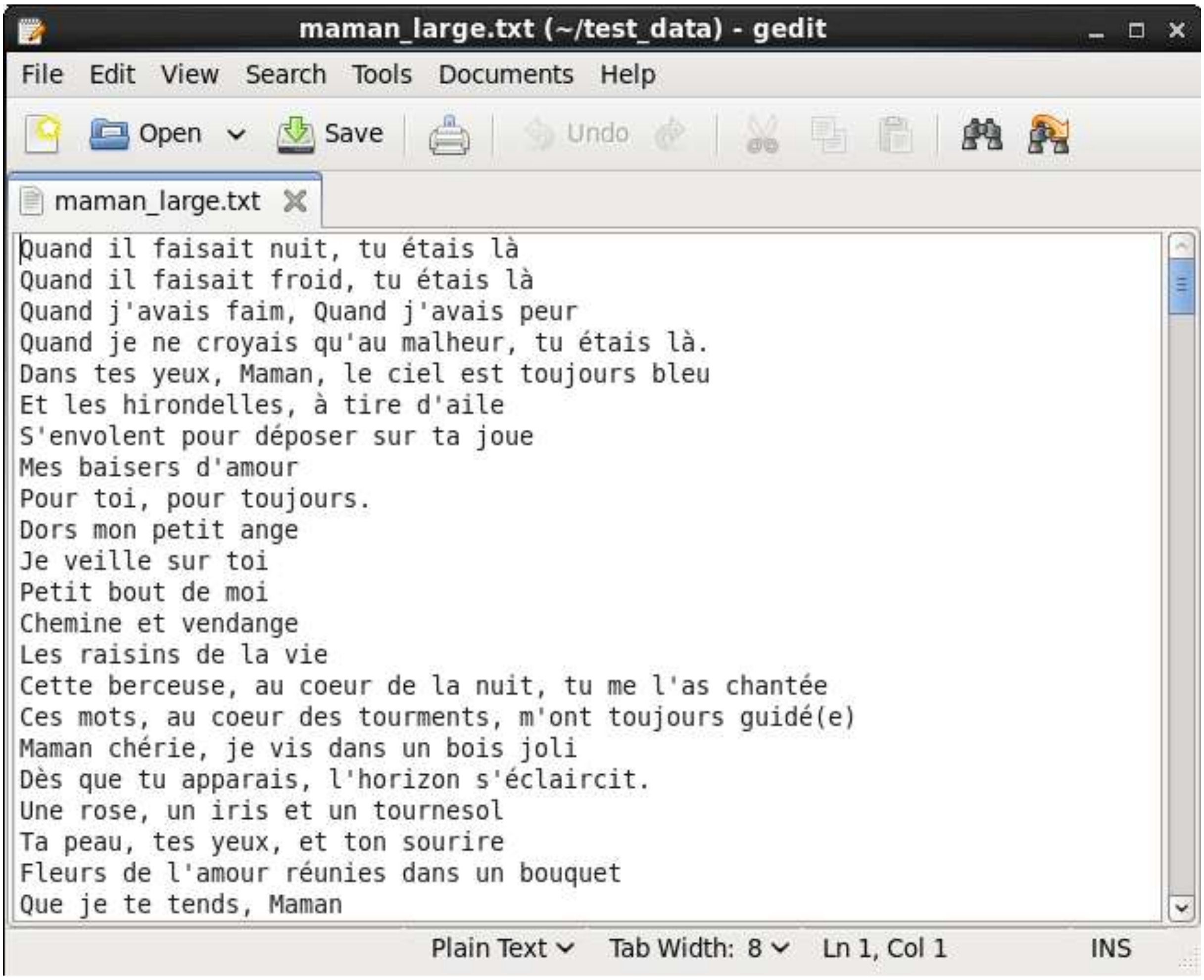


```
[cloudera@quickstart TP_MapRed]$ mkdir -p /home/cloudera/test_data
[cloudera@quickstart TP_MapRed]$ █
```

2. Test avec le Combiner (déjà activé dans WordCount.java) :

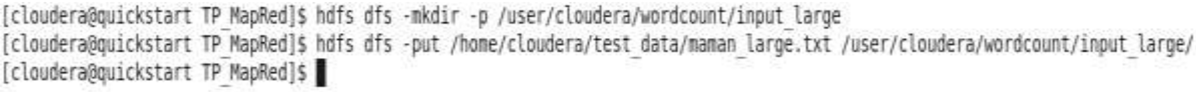


```
[cloudera@quickstart TP_MapRed]$ for i in {1..20}; do   cat /home/cloudera/TP_MapRed/maman.txt >> /home/cloudera/test_data/maman_large.txt; done
[cloudera@quickstart TP_MapRed]$ █
```

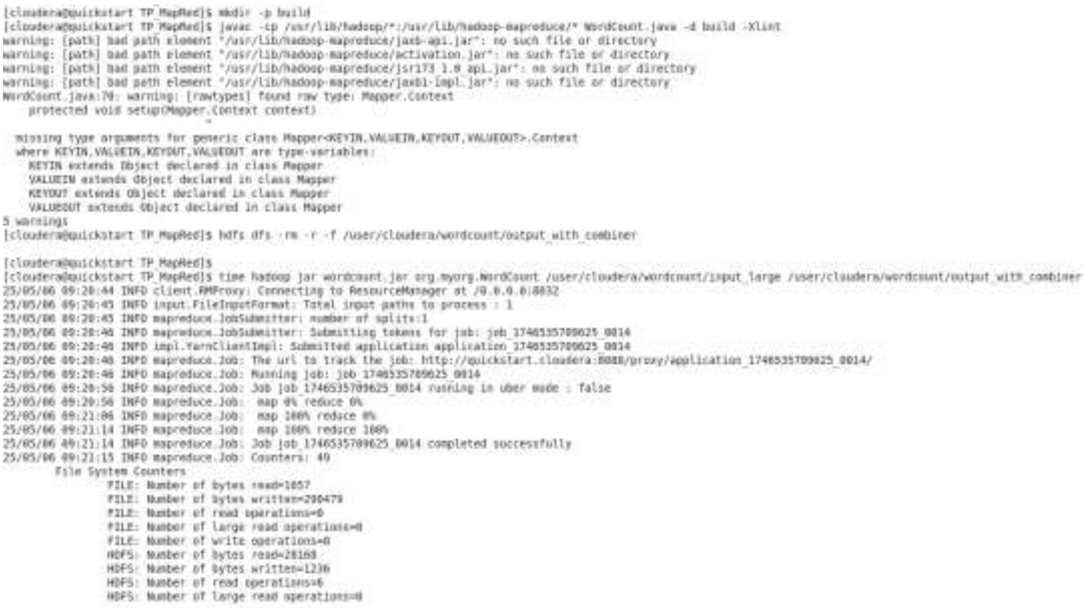


maman_large.txt (~/test_data) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
maman_large.txt
Plain Text
Tab Width: 8
Ln 1, Col 1
INS

```
Quand il faisait nuit, tu étais là
Quand il faisait froid, tu étais là
Quand j'avais faim, Quand j'avais peur
Quand je ne croyais qu'au malheur, tu étais là.
Dans tes yeux, Maman, le ciel est toujours bleu
Et les hirondelles, à tire d'aile
S'envolent pour déposer sur ta joue
Mes baisers d'amour
Pour toi, pour toujours.
Dors mon petit ange
Je veille sur toi
Petit bout de moi
Chemine et vendange
Les raisins de la vie
Cette berceuse, au coeur de la nuit, tu me l'as chantée
Ces mots, au coeur des tourments, m'ont toujours guidé(e)
Maman chérie, je vis dans un bois joli
Dès que tu apparais, l'horizon s'éclaircit.
Une rose, un iris et un tournesol
Ta peau, tes yeux, et ton sourire
Fleurs de l'amour réunies dans un bouquet
Que je te tends, Maman
```

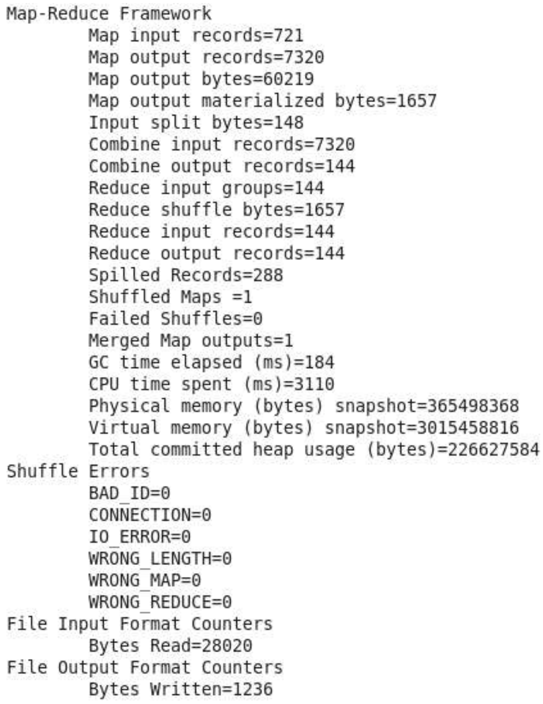


```
[cloudera@quickstart TP_MapRed]$ hdfs dfs -mkdir -p /user/cloudera/wordcount/input_large
[cloudera@quickstart TP_MapRed]$ hdfs dfs -put /home/cloudera/test_data/maman_large.txt /user/cloudera/wordcount/input_large/
[cloudera@quickstart TP_MapRed]$ █
```

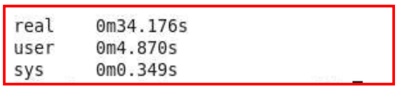


```
[cloudera@quickstart TP_MapRed]$ mkdir -p build
[cloudera@quickstart TP_MapRed]$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* WordCount.java -d build -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no such file or directory
WordCount.java:70: warning: [rawtypes] found raw type: Mapper.Context
    protected void setup(Mapper.Context context)
                               ^
  missing type arguments for generic class Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>.Context
  where KEYIN,VALUEIN,KEYOUT,VALUEOUT are type-variables:
    KEYIN extends Object declared in class Mapper
    VALUEIN extends Object declared in class Mapper
    KEYOUT extends Object declared in class Mapper
    VALUEOUT extends Object declared in class Mapper
5 warnings
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/wordcount/output_with_combiner

[cloudera@quickstart TP_MapRed]$
[cloudera@quickstart TP_MapRed]$ time hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input_large /user/cloudera/wordcount/output_with_combiner
25/05/06 09:20:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/05/06 09:20:45 INFO input.FileInputFormat: Total input paths to process : 1
25/05/06 09:20:45 INFO mapreduce.JobSubmitter: number of splits:1
25/05/06 09:20:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746535709625_0014
25/05/06 09:20:46 INFO impl.YarnClientImpl: Submitted application application_1746535709625_0014
25/05/06 09:20:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1746535709625_0014/
25/05/06 09:20:46 INFO mapreduce.Job: Running job: job_1746535709625_0014
25/05/06 09:20:56 INFO mapreduce.Job: Job job_1746535709625_0014 running in uber mode : false
25/05/06 09:20:56 INFO mapreduce.Job:  map 0% reduce 0%
25/05/06 09:21:06 INFO mapreduce.Job:  map 100% reduce 0%
25/05/06 09:21:14 INFO mapreduce.Job:  map 100% reduce 100%
25/05/06 09:21:14 INFO mapreduce.Job: Job job_1746535709625_0014 completed successfully
25/05/06 09:21:15 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1657
                FILE: Number of bytes written=290479
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=28168
                HDFS: Number of bytes written=1236
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
```

Time Avec Combiner



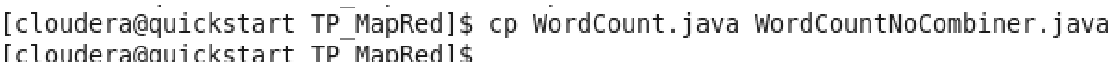
```
        Map-Reduce Framework
                Map input records=721
                Map output records=7320
                Map output bytes=60219
                Map output materialized bytes=1657
                Input split bytes=148
                Combine input records=7320
                Combine output records=144
                Reduce input groups=144
                Reduce shuffle bytes=1657
                Reduce input records=144
                Reduce output records=144
                Spilled Records=288
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=184
                CPU time spent (ms)=3110
                Physical memory (bytes) snapshot=365498368
                Virtual memory (bytes) snapshot=3015458816
                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=28020
        File Output Format Counters
                Bytes Written=1236
```



```
real    0m34.176s
user    0m4.870s
sys     0m0.349s
```

3. Test sans Combiner :

#Copier le fichier source

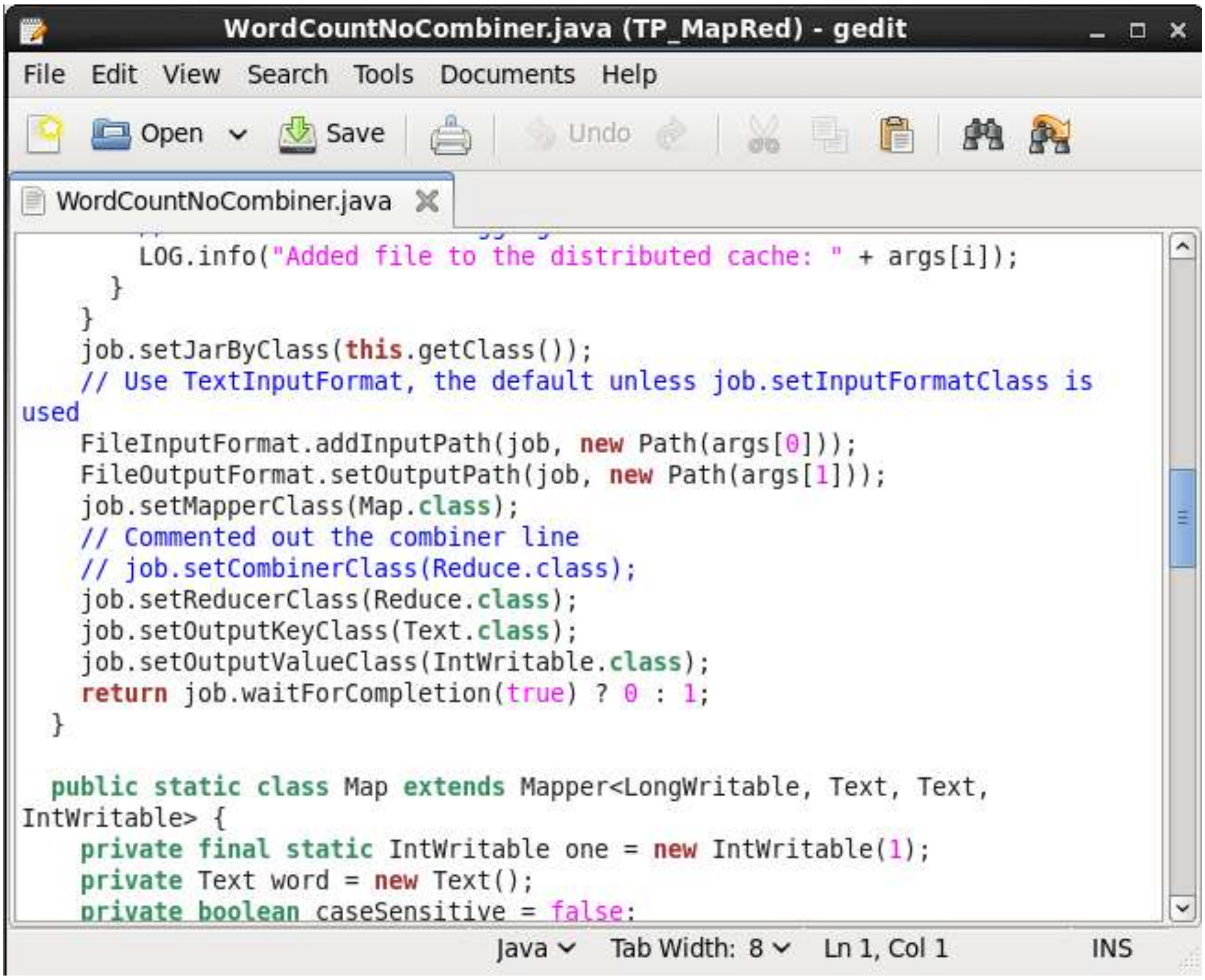


```
[cloudera@quickstart TP_MapRed]$ cp WordCount.java WordCountNoCombiner.java
[cloudera@quickstart TP_MapRed]$
```

# Modifier le fichier pour retirer le combiner

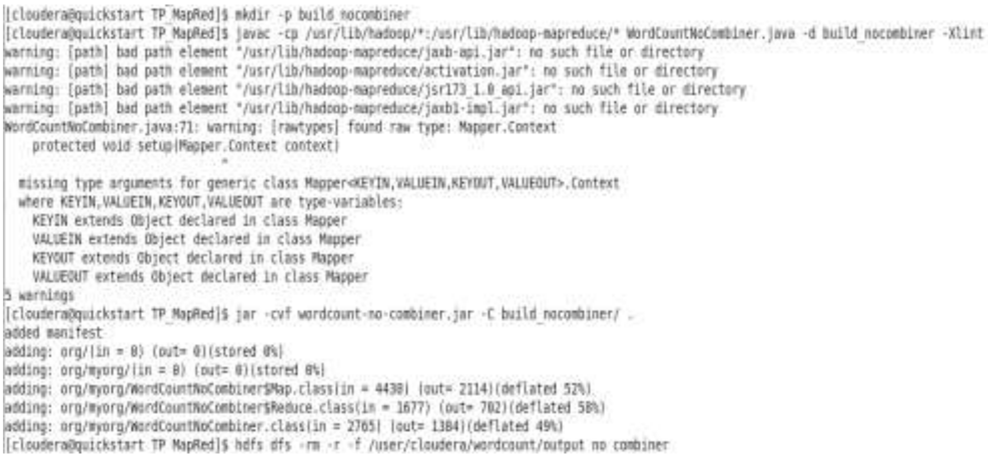
# Ouvrez le fichier et commentez la ligne job.setCombinerClass(Reduce.class);

Renommez aussi la classe en WordCountNoCombiner


WordCountNoCombiner.java (TP_MapRed) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
WordCountNoCombiner.java
Java
Tab Width: 8
Ln 1, Col 1
INS

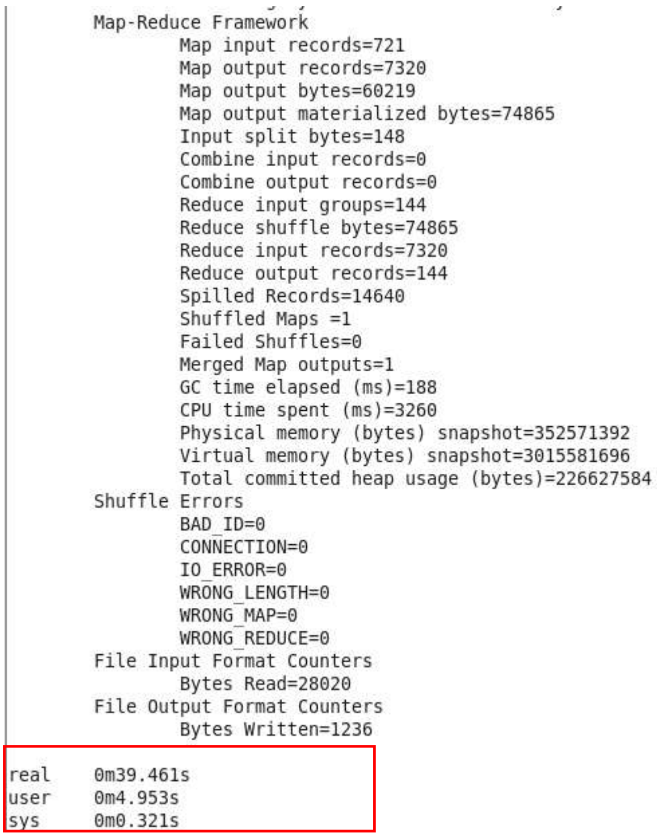
```
        LOG.info("Added file to the distributed cache: " + args[i]);
      }
    }
    job.setJarByClass(this.getClass());
    // Use TextInputFormat, the default unless job.setInputFormatClass is
used
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.setMapperClass(Map.class);
    // Commented out the combiner line
    // job.setCombinerClass(Reduce.class);
    job.setReducerClass(Reduce.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    return job.waitForCompletion(true) ? 0 : 1;
  }

  public static class Map extends Mapper<LongWritable, Text, Text,
IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    private boolean caseSensitive = false;
```

Exécution :



```
[cloudera@quickstart TP_MapRed]$ mkdir -p build_nocombiner
[cloudera@quickstart TP_MapRed]$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* WordCountNoCombiner.java -d build_nocombiner -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no such file or directory
WordCountNoCombiner.java:71: warning: [rawtypes] found raw type: Mapper.Context
    protected void setup(Mapper.Context context)
                               ^
  missing type arguments for generic class Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>.Context
  where KEYIN,VALUEIN,KEYOUT,VALUEOUT are type-variables:
    KEYIN extends Object declared in class Mapper
    VALUEIN extends Object declared in class Mapper
    KEYOUT extends Object declared in class Mapper
    VALUEOUT extends Object declared in class Mapper
5 warnings
[cloudera@quickstart TP_MapRed]$ jar -cvf wordcount-no-combiner.jar -C build_nocombiner/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCountNoCombiner$Map.class(in = 4430) (out= 2114)(deflated 52%)
adding: org/myorg/WordCountNoCombiner$Reduce.class(in = 1677) (out= 702)(deflated 58%)
adding: org/myorg/WordCountNoCombiner.class(in = 2765) (out= 1384)(deflated 49%)
[cloudera@quickstart TP_MapRed]$ hdfs dfs -rm -r -f /user/cloudera/wordcount/output_no_combiner
```

Test Sans Combiner



```
        Map-Reduce Framework
                Map input records=721
                Map output records=7320
                Map output bytes=60219
                Map output materialized bytes=74865
                Input split bytes=148
                Combine input records=0
                Combine output records=0
                Reduce input groups=144
                Reduce shuffle bytes=74865
                Reduce input records=7320
                Reduce output records=144
                Spilled Records=14640
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=188
                CPU time spent (ms)=3260
                Physical memory (bytes) snapshot=352571392
                Virtual memory (bytes) snapshot=3015581696
                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=28020
        File Output Format Counters
                Bytes Written=1236

real    0m39.461s
user    0m4.953s
sys     0m0.321s
```

1. **Quelle version s'est exécutée plus rapidement et pourquoi ?**

Généralement, **avec combiner** est plus rapide, car moins de données sont envoyées au réseau (réduction du shuffle).

2. **Quel est le taux de réduction de données obtenu grâce au combiner ?**

Calcul = (nombre de paires sans combiner - avec combiner) / sans combiner * 100%

3. **Pourquoi peut-on utiliser la classe Reduce comme combiner ?**

Parce que l'opération de réduction (somme des occurrences) est **associative et commutative**, ce qui est une condition pour les combiners.