Máster en Big Data

Tecnologías de Almacenamiento

5. Hands-On: Desarrollo MapReduce Avanzado

Realizado por
Oscar Tenesaca

Índice

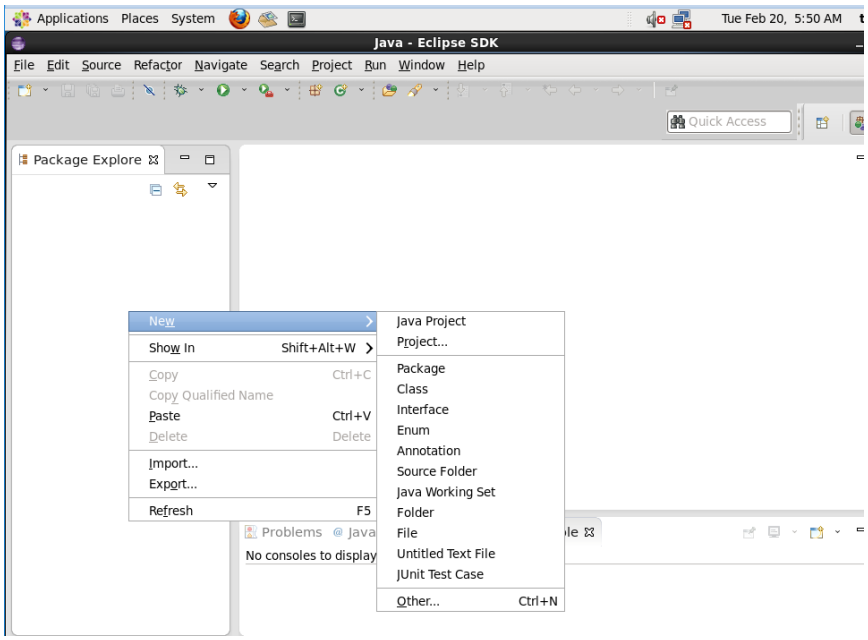
1. Introducción

El objetivo de este Hands-On es poner en práctica conceptos avanzados en el desarrollo de Jobs de MapReduce

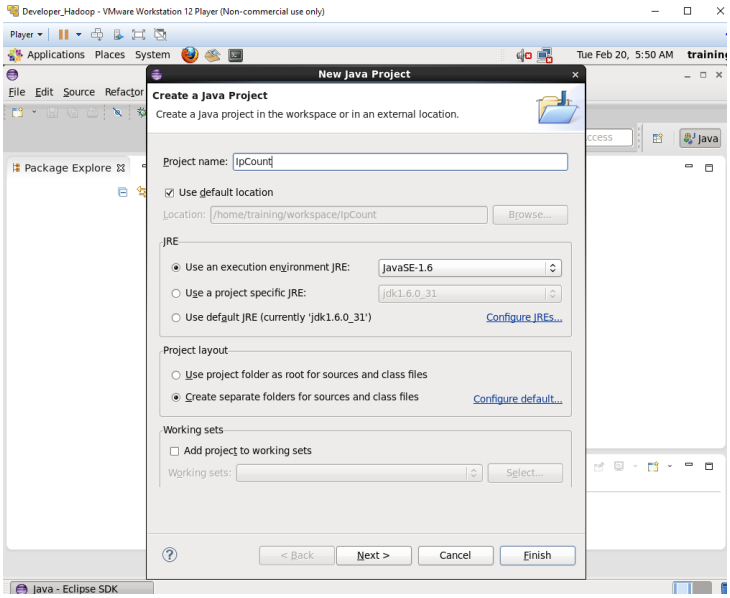
2. Entorno de desarrollo

Para realizar el desarrollo lo haremos mediante el IDE Eclipse de la máquina virtual importada en ejercicios anteriores.

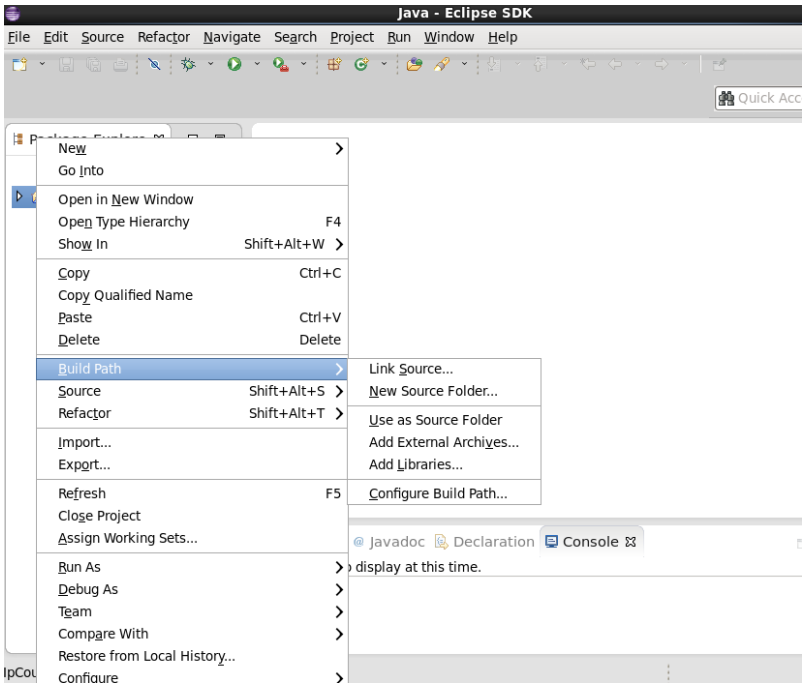
Para crear un nuevo proyecto, haremos click derecho sobre el package explorer New → Java Project



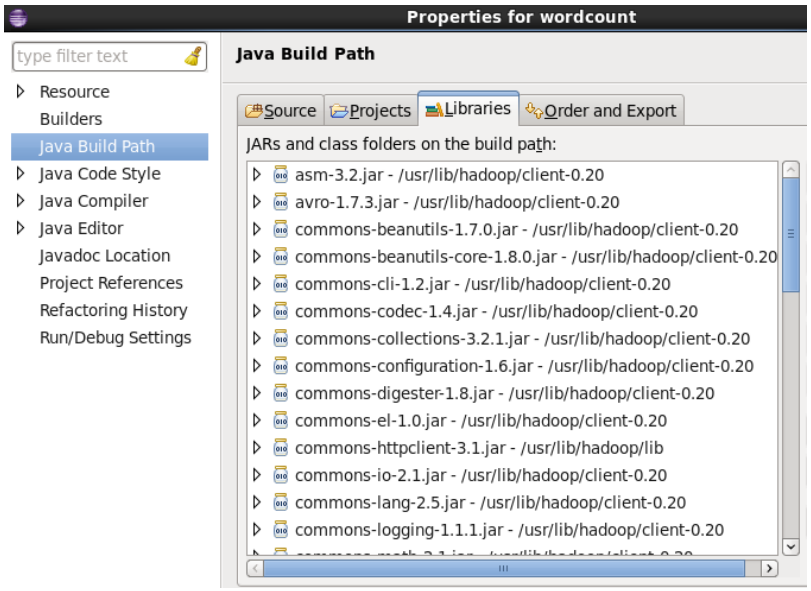
Introducimos el nombre del proyecto y click en Finish



Importamos manualmente las lilbrerías necesarias haciendo click derecho sobre el proyecto que acabamos de crear y seleccionamos Build Path → Configure Build Path



En la pestaña de libraries, seleccionamos Add Extertnal Jars e importamos todo el contenido de la carpeta /usr/lib/hadoop/client-0.20/



3. Tool Runner y parámetros

Desarrollar y ejecutar el siguiente MapReduce:

Aprovechando el ejercicio del Hands-On anterior (**AvarageWordLength)** realizar las siguientes modificaciones:

- La clase driver use ToolRunner

- Modificar el Mapper para referenciar una variable booleana llamada caseSensitive. Si esta variable es true, el mapper no diferenciara entre mayúsculas ni minúsculas, si es false, hará una conversión de todas las letras a minúscula.

Código fuente

IpDriver.java | IpReduce.java | IpMapper.java

```java
package practic06;

import org.apache.hadoop.fs.Path;

public class IpDriver extends Configured implements Tool {

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

        Configuration conf = new Configuration();
        int exitCode = ToolRunner.run(conf, new IpDriver(), args);
        System.exit(exitCode);
    }

    @Override
    public int run(String[] args) throws Exception {

        if (args.length != 2) {
            System.out.printf("Usage: IpDriver <input dir> <output dir>\n");
            System.exit(-1);
        }

        Job job = new Job(getConf());
        job.setJarByClass(IpDriver.class);
        job.setJobName("Average Word Length");
        FileInputFormat.setInputPaths(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.setMapperClass(IpMapper.class);
        job.setReducerClass(IpReduce.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(DoubleWritable.class);

        boolean success = job.waitForCompletion(true);
        return(success ? 0 : 1);
    }
}
```

IpDriver.java | IpReduce.java | IpMapper.java

```java
package practic06;

import java.io.IOException;

public class IpReduce  extends Reducer<Text, IntWritable, Text, DoubleWritable> {

    @Override
    public void reduce (Text key, Iterable<IntWritable> values, Context context)
            throws IOException, InterruptedException {

        double wordCount = 0;
        double wordSum = 0;

        for(IntWritable value : values) {
            wordCount += value.get();
            wordSum++;
        }

        double wordAverage = (double) wordCount / wordSum;
        context.write(key, new DoubleWritable(wordAverage));
    }
}
```

IpDriver.java | IpReduce.java | IpMapper.java

```java
package practic06;

import java.io.IOException;

public class IpMapper extends Mapper<LongWritable, Text, Text, IntWritable> {

    public boolean myParam = false;

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

        for (String word : line.split("\\W+")) {
            if (word.length() > 0) {

                String letter;

                if (myParam) {
                    letter = word.substring(0, 1);
                } else {
                    letter = word.substring(0, 1).toLowerCase();
                }

                context.write(new Text(letter), new IntWritable(word.length()));
            }
        }
    }

    @Override
    public void setup(Context context) {
        Configuration conf = context.getConfiguration();
        myParam = conf.getBoolean("myParam", false);
    }
}
```

Resultados

Creación y ejecución del nuestro *.jar con nuestro parámetro false

```
File  Edit  View  Search  Terminal  Help
[training@localhost src]$ javac -classpath `hadoop classpath` practic06/IpMapper.java practic06/IpReduce.java practic06/IpDriver.java
[training@localhost src]$ jar cf MyMR.jar practic06/IpMapper.class practic06/IpReduce.class practic06/IpDriver.class
[training@localhost src]$ hadoop jar MyMR.jar practic06.IpDriver -DmyParam=false /user/training/shakespeare /user/training/result_false
24/04/10 15:44:17 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
24/04/10 15:44:17 INFO input.FileInputFormat: Total input paths to process : 4
24/04/10 15:44:17 INFO mapred.JobClient: Running job: job_202403061449_0077
24/04/10 15:44:18 INFO mapred.JobClient:  map 0% reduce 0%
24/04/10 15:44:30 INFO mapred.JobClient:  map 50% reduce 0%
24/04/10 15:44:40 INFO mapred.JobClient:  map 75% reduce 0%
24/04/10 15:44:41 INFO mapred.JobClient:  map 100% reduce 0%
24/04/10 15:44:45 INFO mapred.JobClient:  map 100% reduce 100%
24/04/10 15:44:47 INFO mapred.JobClient: Job complete: job_202403061449_0077
24/04/10 15:44:47 INFO mapred.JobClient: Counters: 32
24/04/10 15:44:47 INFO mapred.JobClient:   File System Counters
24/04/10 15:44:47 INFO mapred.JobClient:     FILE: Number of bytes read=15029594
24/04/10 15:44:47 INFO mapred.JobClient:     FILE: Number of bytes written=23710648
24/04/10 15:44:47 INFO mapred.JobClient:     FILE: Number of read operations=0
24/04/10 15:44:47 INFO mapred.JobClient:     FILE: Number of large read operations=0
24/04/10 15:44:47 INFO mapred.JobClient:     FILE: Number of write operations=0
24/04/10 15:44:47 INFO mapred.JobClient:     HDFS: Number of bytes read=5284706
24/04/10 15:44:47 INFO mapred.JobClient:     HDFS: Number of bytes written=589
24/04/10 15:44:47 INFO mapred.JobClient:     HDFS: Number of read operations=9
24/04/10 15:44:47 INFO mapred.JobClient:     HDFS: Number of large read operations=0
24/04/10 15:44:47 INFO mapred.JobClient:     HDFS: Number of write operations=1
24/04/10 15:44:47 INFO mapred.JobClient:   Job Counters
24/04/10 15:44:47 INFO mapred.JobClient:     Launched map tasks=4
24/04/10 15:44:47 INFO mapred.JobClient:     Launched reduce tasks=1
24/04/10 15:44:47 INFO mapred.JobClient:     Data-local map tasks=4
24/04/10 15:44:47 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=40768
24/04/10 15:44:47 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=13862
24/04/10 15:44:47 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
24/04/10 15:44:47 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
24/04/10 15:44:47 INFO mapred.JobClient:   Map-Reduce Framework
24/04/10 15:44:47 INFO mapred.JobClient:     Map input records=173126
24/04/10 15:44:47 INFO mapred.JobClient:     Map output records=964453
24/04/10 15:44:47 INFO mapred.JobClient:     Map output bytes=5786718
24/04/10 15:44:47 INFO mapred.JobClient:     Input split bytes=475
24/04/10 15:44:47 INFO mapred.JobClient:     Combine input records=0
24/04/10 15:44:47 INFO mapred.JobClient:     Combine output records=0
24/04/10 15:44:47 INFO mapred.JobClient:     Reduce input groups=35
24/04/10 15:44:47 INFO mapred.JobClient:     Reduce shuffle bytes=7715648
24/04/10 15:44:47 INFO mapred.JobClient:     Reduce input records=964453
24/04/10 15:44:47 INFO mapred.JobClient:     Reduce output records=35
24/04/10 15:44:47 INFO mapred.JobClient:     Spilled Records=2843147
24/04/10 15:44:47 INFO mapred.JobClient:     CPU time spent (ms)=7190
24/04/10 15:44:47 INFO mapred.JobClient:     Physical memory (bytes) snapshot=975867904
24/04/10 15:44:47 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=3618029568
24/04/10 15:44:47 INFO mapred.JobClient:     Total committed heap usage (bytes)=657342464
[training@localhost src]$
```

```
[training@localhost ~]$ hdfs dfs -cat /user/training/result_false/part-r-00000
1       1.02
2       1.0588235294117647
3       1.0
4       1.5
5       1.5
6       1.5
7       1.0
8       1.5
9       1.0
a       3.275899648342265
b       4.43676859192148
c       6.204073527743107
d       4.306200411401704
e       5.307238813182565
f       4.87378966665806
g       5.163681818181818
h       3.966131770968778
i       2.1290417039114753
j       5.148983570036202
k       4.622562809195848
l       4.454545454545454
m       3.990697595029614
n       3.749964544036307
o       2.8046206567868732
p       6.209215222076958
q       5.852795739825028
r       5.854965809182676
s       4.48601978893492
t       3.772103219434822
u       4.588696504410324
v       5.540627750073336
w       4.373096283946263
x       3.1650485436893203
y       3.51717399473882
z       5.053333333333334
[training@localhost ~]$
```

Ejecución del nuestro *.jar con nuestro parámetro true

```
File  Edit  View  Search  Terminal  Help
training@localhost src]$ javac -classpath `hadoop classpath` practic06/IpMapper.java practic06/IpReduce.java practic06/IpDriver.java
training@localhost src]$ jar cf MyMR.jar practic06/IpMapper.class practic06/IpReduce.class practic06/IpDriver.class
training@localhost src]$ hadoop jar MyMR.jar practic06.IpDriver -DmyParam=true /user/training/shakespeare /user/training/result_true
14/04/10 15:49:02 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
14/04/10 15:49:02 INFO input.FileInputFormat: Total input paths to process : 4
14/04/10 15:49:02 INFO mapred.JobClient: Running job: job_202403061449_0078
14/04/10 15:49:03 INFO mapred.JobClient:  map 0% reduce 0%
14/04/10 15:49:14 INFO mapred.JobClient:  map 50% reduce 0%
14/04/10 15:49:24 INFO mapred.JobClient:  map 100% reduce 0%
14/04/10 15:49:28 INFO mapred.JobClient:  map 100% reduce 100%
14/04/10 15:49:29 INFO mapred.JobClient: Job complete: job_202403061449_0078
14/04/10 15:49:30 INFO mapred.JobClient: Counters: 32
14/04/10 15:49:30 INFO mapred.JobClient:   File System Counters
14/04/10 15:49:30 INFO mapred.JobClient:     FILE: Number of bytes read=15029594
14/04/10 15:49:30 INFO mapred.JobClient:     FILE: Number of bytes written=23710638
14/04/10 15:49:30 INFO mapred.JobClient:     FILE: Number of read operations=0
14/04/10 15:49:30 INFO mapred.JobClient:     FILE: Number of large read operations=0
14/04/10 15:49:30 INFO mapred.JobClient:     FILE: Number of write operations=0
14/04/10 15:49:30 INFO mapred.JobClient:     HDFS: Number of bytes read=5284706
14/04/10 15:49:30 INFO mapred.JobClient:     HDFS: Number of bytes written=1076
14/04/10 15:49:30 INFO mapred.JobClient:     HDFS: Number of read operations=9
14/04/10 15:49:30 INFO mapred.JobClient:     HDFS: Number of large read operations=0
14/04/10 15:49:30 INFO mapred.JobClient:     HDFS: Number of write operations=1
14/04/10 15:49:30 INFO mapred.JobClient:   Job Counters
14/04/10 15:49:30 INFO mapred.JobClient:     Launched map tasks=4
14/04/10 15:49:30 INFO mapred.JobClient:     Launched reduce tasks=1
14/04/10 15:49:30 INFO mapred.JobClient:     Data-local map tasks=4
14/04/10 15:49:30 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=40152
14/04/10 15:49:30 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=13306
14/04/10 15:49:30 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/04/10 15:49:30 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
14/04/10 15:49:30 INFO mapred.JobClient:   Map-Reduce Framework
14/04/10 15:49:30 INFO mapred.JobClient:     Map input records=173126
14/04/10 15:49:30 INFO mapred.JobClient:     Map output records=964453
14/04/10 15:49:30 INFO mapred.JobClient:     Map output bytes=5786718
14/04/10 15:49:30 INFO mapred.JobClient:     Input split bytes=475
14/04/10 15:49:30 INFO mapred.JobClient:     Combine input records=0
14/04/10 15:49:30 INFO mapred.JobClient:     Combine output records=0
14/04/10 15:49:30 INFO mapred.JobClient:     Reduce input groups=60
14/04/10 15:49:30 INFO mapred.JobClient:     Reduce shuffle bytes=7715648
14/04/10 15:49:30 INFO mapred.JobClient:     Reduce input records=964453
14/04/10 15:49:30 INFO mapred.JobClient:     Reduce output records=60
14/04/10 15:49:30 INFO mapred.JobClient:     Spilled Records=2843147
14/04/10 15:49:30 INFO mapred.JobClient:     CPU time spent (ms)=7120
14/04/10 15:49:30 INFO mapred.JobClient:     Physical memory (bytes) snapshot=986595328
14/04/10 15:49:30 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=3618029568
14/04/10 15:49:30 INFO mapred.JobClient:     Total committed heap usage (bytes)=657342464
training@localhost src]$
```

```
[training@localhost ~]$ hdfs dfs -cat /user/training/result_true/part-r-00000
1       1.02
2       1.0588235294117647
3       1.0
4       1.5
5       1.5
6       1.5
7       1.0
8       1.5
9       1.0
A       3.891394576646375
B       5.139302507836991
C       6.629694233531706
D       5.201834862385321
E       5.514263685427911
F       5.255528255528255
G       5.809792180345192
H       4.42107243650047
I       1.4526860926284046
J       4.984008528784648
K       4.657106838953672
L       5.115881561238224
M       5.44646530258742
N       3.9848387785607517
O       2.8794768365725463
P       6.505740766357726
Q       5.5216426193118755
R       5.929275069461985
S       5.293126010314833
T       3.959143714919723
U       5.325
V       5.194537815126051
W       4.464014043300176
X       3.1650485436893203
Y       3.4432244242099626
Z       6.1
a       3.0776554817818575
b       4.245396808453862
c       6.041441229514624
d       4.146387533448764
e       5.182465923172243
f       4.778552071234998
g       4.938916799411837
h       3.8777881295555434
i       2.7292957500654507
j       5.329446064139941
k       4.607202914798206
l       4.272777716124736
m       3.7182168186423508
n       3.7032013944985334
o       2.7875536480686693
p       6.10748861047836
q       6.025462962962963
r       5.829150579150579
s       4.327014649237208
```

4. Combiner

Desarrollar y ejecutar el siguiente MapReduce:

Añadir un combiner al proyecto **IpCount** realizado en el Hands-On anterior

Código fuente

IpDriver.java | IpMapper.java | IpReduce.java

```java
package practic07;

import org.apache.hadoop.fs.Path;

public class IpDriver extends Configured implements Tool {

  @Override
  public int run(String[] args) throws Exception {

    if (args.length != 2) {
      System.out.printf("Usage: IpDriver <input dir> <output dir>\n");
      return -1;
    }

    Job job = new Job(getConf());
    job.setJarByClass(IpDriver.class);
    job.setJobName("Word Count Driver");
    FileInputFormat.setInputPaths(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.setMapperClass(IpMapper.class);
    job.setReducerClass(IpReduce.class);
    job.setCombinerClass(IpReduce.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    if (job.getCombinerClass() == null) {
      throw new Exception("No hay combinacion");
    }

    boolean success = job.waitForCompletion(true);
    return success ? 0 : 1;
  }

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

IpDriver.java | *IpMapper.java | IpReduce.java

```java
package practic07;

import java.io.IOException;

public class IpMapper extends Mapper<LongWritable, Text, Text, IntWritable> {

    @Override
    public void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException {

      String line = value.toString();

      for (String word : line.split("\\W+")) {
        if (word.length() > 0) {
          context.write(new Text(word), new IntWritable(1));
        }
      }

    }
  }
```

IpDriver.java | *IpMapper.java | IpReduce.java

```java
package practic07;

import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class IpReduce extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {

      int counter = 0;

      for (IntWritable value : values) {
        counter += value.get();
      }

      context.write(key, new IntWritable(counter));

    }
  }
```

Ejecución del nuestro *.jar con nuestro parámetro true

```
[training@localhost src]$ javac -classpath `hadoop classpath` practic07/IpMapper.java practic07/IpReduce.java practic07/IpDriver.java
[training@localhost src]$ jar cf MyMR.jar practic07/IpMapper.class practic07/IpReduce.class practic07/IpDriver.class
[training@localhost src]$ hadoop jar MyMR.jar practic07.IpDriver  /user/training/weblog/access_log /user/training/result_combiner
24/04/10 15:56:08 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
24/04/10 15:56:08 INFO input.FileInputFormat: Total input paths to process : 1
24/04/10 15:56:09 INFO mapred.JobClient: Running job: job_202403061449_0079
24/04/10 15:56:10 INFO mapred.JobClient:  map 0% reduce 0%
24/04/10 15:56:25 INFO mapred.JobClient:  map 3% reduce 0%
24/04/10 15:56:28 INFO mapred.JobClient:  map 5% reduce 0%
24/04/10 15:56:31 INFO mapred.JobClient:  map 8% reduce 0%
24/04/10 15:56:34 INFO mapred.JobClient:  map 12% reduce 0%
24/04/10 15:56:36 INFO mapred.JobClient:  map 14% reduce 0%
24/04/10 15:56:39 INFO mapred.JobClient:  map 17% reduce 0%
24/04/10 15:56:43 INFO mapred.JobClient:  map 20% reduce 0%
24/04/10 15:56:46 INFO mapred.JobClient:  map 23% reduce 0%
24/04/10 15:56:48 INFO mapred.JobClient:  map 25% reduce 0%
24/04/10 15:57:02 INFO mapred.JobClient:  map 27% reduce 8%
24/04/10 15:57:05 INFO mapred.JobClient:  map 30% reduce 8%
24/04/10 15:57:08 INFO mapred.JobClient:  map 32% reduce 8%
24/04/10 15:57:11 INFO mapred.JobClient:  map 35% reduce 8%
24/04/10 15:57:14 INFO mapred.JobClient:  map 38% reduce 8%
24/04/10 15:57:17 INFO mapred.JobClient:  map 41% reduce 8%
24/04/10 15:57:20 INFO mapred.JobClient:  map 44% reduce 8%
24/04/10 15:57:23 INFO mapred.JobClient:  map 47% reduce 8%
24/04/10 15:57:26 INFO mapred.JobClient:  map 50% reduce 8%
24/04/10 15:57:32 INFO mapred.JobClient:  map 50% reduce 16%
24/04/10 15:57:38 INFO mapred.JobClient:  map 53% reduce 16%
24/04/10 15:57:41 INFO mapred.JobClient:  map 56% reduce 16%
24/04/10 15:57:44 INFO mapred.JobClient:  map 59% reduce 16%
24/04/10 15:57:47 INFO mapred.JobClient:  map 62% reduce 16%
24/04/10 15:57:50 INFO mapred.JobClient:  map 64% reduce 16%
24/04/10 15:57:53 INFO mapred.JobClient:  map 67% reduce 16%
24/04/10 15:57:56 INFO mapred.JobClient:  map 70% reduce 16%
24/04/10 15:57:59 INFO mapred.JobClient:  map 73% reduce 16%
24/04/10 15:58:01 INFO mapred.JobClient:  map 75% reduce 16%
24/04/10 15:58:02 INFO mapred.JobClient:  map 75% reduce 25%
24/04/10 15:58:12 INFO mapred.JobClient:  map 80% reduce 25%
24/04/10 15:58:15 INFO mapred.JobClient:  map 84% reduce 25%
24/04/10 15:58:18 INFO mapred.JobClient:  map 88% reduce 25%
24/04/10 15:58:21 INFO mapred.JobClient:  map 93% reduce 25%
24/04/10 15:58:24 INFO mapred.JobClient:  map 96% reduce 25%
24/04/10 15:58:26 INFO mapred.JobClient:  map 96% reduce 29%
24/04/10 15:58:27 INFO mapred.JobClient:  map 99% reduce 29%
24/04/10 15:58:28 INFO mapred.JobClient:  map 100% reduce 29%
24/04/10 15:58:29 INFO mapred.JobClient:  map 100% reduce 33%
24/04/10 15:58:32 INFO mapred.JobClient:  map 100% reduce 100%
24/04/10 15:58:34 INFO mapred.JobClient: Job complete: job_202403061449_0079
24/04/10 15:58:34 INFO mapred.JobClient: Counters: 32
24/04/10 15:58:34 INFO mapred.JobClient:   File System Counters
24/04/10 15:58:34 INFO mapred.JobClient:     FILE: Number of bytes read=63166036
24/04/10 15:58:34 INFO mapred.JobClient:     FILE: Number of bytes written=67849881
24/04/10 15:58:34 INFO mapred.JobClient:     FILE: Number of read operations=0
24/04/10 15:58:34 INFO mapred.JobClient:     FILE: Number of large read operations=0
24/04/10 15:58:34 INFO mapred.JobClient:     FILE: Number of write operations=0
24/04/10 15:58:34 INFO mapred.JobClient:     HDFS: Number of bytes read=504971132
```

```
[training@localhost ~]$ hdfs dfs -cat /user/training/result_combiner/part-r-00000

0       280141
00      269842
0000    2095670
000000  60
0000000 2
000001  190
000001_thumb    162
000006  181
000006_thumb    161
000011  180
000011_thumb    162
000017  180
000017_thumb    161
000021  109
000021_thumb    81
000041  187
000041_thumb    161
000054  955
000054_thumb    708
000083  51
000083_thumb    49
0001    2505
000143  936
000143_thumb    696
000159  938
000159_thumb    690
000163  2196
000163_thumb    683
000181  856
000181_thumb    679
0002    88
0003    138
0004    32
00040Xc 174
00040Xc_thumb   130
00040Xj 169
00040Xj_thumb   133
0005    2389
0005_thumb      459
0006    3899
0007    605
0007_thumb      462
0008    7399
0008_thumb      461
0009    5644
0009_thumb      447
001     595
0010    11563
0010BNb 194
0010BNb_thumb   133
0011    6112
0011_thumb      451
0012    1940
```

5. Partitioner

Desarrollar y ejecutar el siguiente MapReduce:

Aprovechando el proyecto original **IpCount** realizar los cambios pertinentes para escribir un Job con múltiples reducers e implementar un partitioner que redirija la salida según el mes del año hacia un reducer concreto.

Es decir, en total habrán 12 reducers (uno para cada mes del año) y el partitioner será el encargado de redirigir esa clave/valor hacia el reducer correcto.

La salida final consistirá en 12 ficheros, uno para cada mes del año, y contendrán el número de veces que se ha repetido la ip en ese mes del año.

Solución:

Input: `96.7.4.14 - - [24/Apr/2011:04:20:11 -0400] "GET /cat.jpg HTTP/1.1" 200 12433`

Output key: `96.7.4.14`

Output value: `Apr`

Código fuente

IpDriver.java | IpMapper.java | IpReduce.java | IpReducerMonth.java

```java
package practic08;

import org.apache.hadoop.fs.Path;

public class IpDriver {

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

        if (args.length != 2) {
          System.out.printf("Usage: IpDriver <input dir> <output dir>\n");
          System.exit(-1);
        }

        Job job = new Job();
        job.setJarByClass(IpDriver.class);
        job.setJobName("Process Logs");

        FileInputFormat.setInputPaths(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.setMapperClass(IpMapper.class);
        job.setReducerClass(IpReduce.class);

        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(Text.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        job.setNumReduceTasks(12);
        job.setPartitionerClass(IpReducerMonth.class);

        boolean success = job.waitForCompletion(true);
        System.exit(success ? 0 : 1);
      }
    }
```

IpDriver.java | IpMapper.java | IpReduce.java | IpReducerMonth.java

```java
package practic08;

import java.io.IOException;

public class IpMapper extends Mapper<LongWritable, Text, Text, Text> {

    public static List<String> months = Arrays.asList("Jan", "Feb", "Mar",
            "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec");

    @Override
    public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

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

        if (fields.length > 3) {

            String ip = fields[0];

            String[] dtFields = fields[3].split("/");
            if (dtFields.length > 1) {
                String theMonth = dtFields[1];

                if (months.contains(theMonth))
                    context.write(new Text(ip), new Text(theMonth));
            }
        }
    }
}
```

IpDriver.java | IpMapper.java | IpReduce.java | IpReducerMonth.java

```java
package practic08;

import java.io.IOException;

public class IpReduce extends Reducer<Text, Text, Text, IntWritable> {

    @Override
    public void reduce(Text key, Iterable<Text> values, Context context)
            throws IOException, InterruptedException {

        int count = 0;

        for (@SuppressWarnings("unused")
        Text value : values) {

            count++;
        }

        context.write(key, new IntWritable(count));
    }
}
```

IpDriver.java | IpMapper.java | IpReduce.java | IpReducerMonth.java

```java
package practic08;

import java.util.HashMap;


public class IpReducerMonth<K2, V2> extends Partitioner<Text, Text> implements
Configurable {

private Configuration configuration;
HashMap<String, Integer> months = new HashMap<String, Integer>();

@Override
public void setConf(Configuration configuration) {
this.configuration = configuration;
months.put("Jan", 0);
months.put("Feb", 1);
months.put("Mar", 2);
months.put("Apr", 3);
months.put("May", 4);
months.put("Jun", 5);
months.put("Jul", 6);
months.put("Aug", 7);
months.put("Sep", 8);
months.put("Oct", 9);
months.put("Nov", 10);
months.put("Dec", 11);
}

@Override
public Configuration getConf() {
return configuration;
}

public int getPartition(Text key, Text value, int numReduceTasks) {
return (int) (months.get(value.toString()));
}
}
```

Ejecución del nuestro *.jar con nuestro parámetro true

```
[training@localhost src]$ javac -classpath `hadoop classpath` practic08/IpMapper.java practic08/IpReduce.java practic08/IpDriver.java practic08/IpReducerMonth.java
[training@localhost src]$ jar cf MyMR.jar practic08/IpMapper.class practic08/IpReduce.class practic08/IpDriver.class practic08/IpReducerMonth.class
[training@localhost src]$ hadoop jar MyMR.jar practic08.IpDriver /user/training/weblog/access_log /user/training/result_partitioner
24/04/10 15:32:46 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
24/04/10 15:32:46 INFO input.FileInputFormat: Total input paths to process : 1
24/04/10 15:32:47 INFO mapred.JobClient: Running job: job_202403061449_0076
24/04/10 15:32:48 INFO mapred.JobClient:  map 0% reduce 0%
24/04/10 15:33:02 INFO mapred.JobClient:  map 25% reduce 0%
24/04/10 15:33:18 INFO mapred.JobClient:  map 38% reduce 1%
24/04/10 15:33:21 INFO mapred.JobClient:  map 50% reduce 1%
24/04/10 15:33:24 INFO mapred.JobClient:  map 50% reduce 2%
24/04/10 15:33:33 INFO mapred.JobClient:  map 72% reduce 2%
24/04/10 15:33:34 INFO mapred.JobClient:  map 75% reduce 2%
24/04/10 15:33:36 INFO mapred.JobClient:  map 75% reduce 4%
24/04/10 15:33:44 INFO mapred.JobClient:  map 87% reduce 4%
24/04/10 15:33:45 INFO mapred.JobClient:  map 100% reduce 4%
24/04/10 15:33:48 INFO mapred.JobClient:  map 100% reduce 11%
24/04/10 15:33:50 INFO mapred.JobClient:  map 100% reduce 16%
24/04/10 15:33:59 INFO mapred.JobClient:  map 100% reduce 33%
24/04/10 15:34:08 INFO mapred.JobClient:  map 100% reduce 50%
24/04/10 15:34:17 INFO mapred.JobClient:  map 100% reduce 66%
24/04/10 15:34:26 INFO mapred.JobClient:  map 100% reduce 83%
24/04/10 15:34:35 INFO mapred.JobClient:  map 100% reduce 100%
24/04/10 15:34:37 INFO mapred.JobClient: Job complete: job_202403061449_0076
24/04/10 15:34:38 INFO mapred.JobClient: Counters: 32
24/04/10 15:34:38 INFO mapred.JobClient:   File System Counters
24/04/10 15:34:38 INFO mapred.JobClient:     FILE: Number of bytes read=177776776
24/04/10 15:34:38 INFO mapred.JobClient:     FILE: Number of bytes written=269642168
24/04/10 15:34:38 INFO mapred.JobClient:     FILE: Number of read operations=0
24/04/10 15:34:38 INFO mapred.JobClient:     FILE: Number of large read operations=0
24/04/10 15:34:38 INFO mapred.JobClient:     FILE: Number of write operations=0
24/04/10 15:34:38 INFO mapred.JobClient:     HDFS: Number of bytes read=504971132
24/04/10 15:34:38 INFO mapred.JobClient:     HDFS: Number of bytes written=5893437
24/04/10 15:34:38 INFO mapred.JobClient:     HDFS: Number of read operations=16
24/04/10 15:34:38 INFO mapred.JobClient:     HDFS: Number of large read operations=0
24/04/10 15:34:38 INFO mapred.JobClient:     HDFS: Number of write operations=12
24/04/10 15:34:38 INFO mapred.JobClient:   Job Counters
24/04/10 15:34:38 INFO mapred.JobClient:     Launched map tasks=8
24/04/10 15:34:38 INFO mapred.JobClient:     Launched reduce tasks=12
24/04/10 15:34:38 INFO mapred.JobClient:     Data-local map tasks=8
24/04/10 15:34:38 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=111371
24/04/10 15:34:38 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=185982
24/04/10 15:34:38 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
24/04/10 15:34:38 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
24/04/10 15:34:38 INFO mapred.JobClient:   Map-Reduce Framework
24/04/10 15:34:38 INFO mapred.JobClient:     Map input records=4477843
24/04/10 15:34:38 INFO mapred.JobClient:     Map output records=4477817
24/04/10 15:34:38 INFO mapred.JobClient:     Map output bytes=79640812
24/04/10 15:34:38 INFO mapred.JobClient:     Input split bytes=928
24/04/10 15:34:38 INFO mapred.JobClient:     Combine input records=0
24/04/10 15:34:38 INFO mapred.JobClient:     Combine output records=0
24/04/10 15:34:38 INFO mapred.JobClient:     Reduce input groups=370578
24/04/10 15:34:38 INFO mapred.JobClient:     Reduce shuffle bytes=88597022
24/04/10 15:34:38 INFO mapred.JobClient:     Reduce input records=4477817
24/04/10 15:34:38 INFO mapred.JobClient:     Reduce output records=370578
```

Verificamos que tengamos todos los archivos

```
[training@localhost ~]$ hdfs dfs -ls /user/training/result_partitioner
Found 14 items
-rw-r--r--   1 training supergroup          0 2024-04-10 15:34 /user/training/result_partitioner/_SUCCESS
drwxr-xr-x   - training supergroup          0 2024-04-10 15:32 /user/training/result_partitioner/_logs
-rw-r--r--   1 training supergroup     151256 2024-04-10 15:33 /user/training/result_partitioner/part-r-00000
-rw-r--r--   1 training supergroup     452255 2024-04-10 15:33 /user/training/result_partitioner/part-r-00001
-rw-r--r--   1 training supergroup    1273871 2024-04-10 15:33 /user/training/result_partitioner/part-r-00002
-rw-r--r--   1 training supergroup     318451 2024-04-10 15:33 /user/training/result_partitioner/part-r-00003
-rw-r--r--   1 training supergroup     423827 2024-04-10 15:34 /user/training/result_partitioner/part-r-00004
-rw-r--r--   1 training supergroup     415929 2024-04-10 15:34 /user/training/result_partitioner/part-r-00005
-rw-r--r--   1 training supergroup     487404 2024-04-10 15:34 /user/training/result_partitioner/part-r-00006
-rw-r--r--   1 training supergroup     732348 2024-04-10 15:34 /user/training/result_partitioner/part-r-00007
-rw-r--r--   1 training supergroup     356677 2024-04-10 15:34 /user/training/result_partitioner/part-r-00008
-rw-r--r--   1 training supergroup     493539 2024-04-10 15:34 /user/training/result_partitioner/part-r-00009
-rw-r--r--   1 training supergroup     509767 2024-04-10 15:34 /user/training/result_partitioner/part-r-00010
-rw-r--r--   1 training supergroup     278113 2024-04-10 15:34 /user/training/result_partitioner/part-r-00011
[training@localhost ~]$
```

Resultado

```
File  Edit  View  Search  Terminal  Help
[training@localhost ~]$  hdfs dfs -cat /user/training/result_partitioner/part-r-00000

10.1.100.199    35
10.1.103.179    1
10.1.109.144    8
10.1.110.64     1
10.1.118.242    1
10.1.133.90     1
10.1.139.212    1
10.1.148.72     2
10.1.156.219    2
10.1.158.223    1
10.1.171.161    21
10.1.181.142    415
10.1.183.134    1
10.1.186.241    2
10.1.187.27     15
10.1.190.237    18
10.1.204.192    1
10.1.212.51     17
10.1.212.93     5
10.1.219.14     3
10.1.223.119    22
10.1.227.158    1
10.1.229.62     7
10.1.232.31     1423
10.1.27.4       1
10.1.30.248     22
10.1.36.126     1
10.1.39.73      1
10.1.42.238     3
10.1.57.2       1
10.1.6.32       22
10.1.62.196     2
10.1.63.183     2
10.1.64.145     3
10.1.72.184     1
10.1.79.2       8
10.1.82.146     1
10.1.84.242     1
10.1.85.125     2
10.1.86.124     6
10.1.91.33      2
10.1.94.160     21
10.10.107.193   60
10.10.113.115   1
10.10.116.110   1
10.10.118.62    2
10.10.126.5     62
10.10.135.149   14
10.10.14.81     1
10.10.143.169   1
10.10.145.144   45
10.10.161.161   20
10.10.163.156   1
```