

Bigdata Assignment 1.6

We have a dataset of sales of different TV sets across different locations.

Records look like:

Samsung|Optima|14|Madhya Pradesh|132401|14200

The fields are arranged like:

Company Name|Product Name|Size in inches|State|Pin Code|Price

There are some invalid records which contain 'NA' in either Company Name or Product Name.

1. Write a Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Answer

Driver code-

```
package mapreduce.task.assignment;
```

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
```

```
public class Task {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Assignment1.6");
        job.setJarByClass(Task.class);
        //Key is LongWritable as it is the index
        job.setMapOutputKeyClass(LongWritable.class);
        //Value is Text as the whole string is required
        job.setMapOutputValueClass(Text.class);

        //Key is LongWritable as it is the index
        job.setOutputKeyClass(LongWritable.class);
        //Value is Text as the whole string is required
```

```

        job.setOutputValueClass(Text.class);
        job.setMapperClass(TaskMapper.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

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

        /*
        Path out=new Path(args[1]);
        out.getFileSystem(conf).delete(out);
        */

        job.waitForCompletion(true);
    }
}

```

Mapper Code -

```
package mapreduce.task.assignment;
```

```
import java.io.IOException;
```

```
import org.apache.hadoop.io.LongWritable;
```

```
import org.apache.hadoop.io.Text;
```

```
import org.apache.hadoop.mapreduce.*;
```

```
public class TaskMapper extends Mapper<LongWritable, Text,LongWritable,Text> {
```

```
    @Override
```

```
    public void map(LongWritable key, Text value, Context context)
```

```
        throws IOException, InterruptedException {
```

```
        String[] lineArray = value.toString().split("\\|");
```

```
        String company = lineArray[0];
```

```
        String product = lineArray[1];
```

```
        //Checking if company name or product name must not equal to NA
```

```
        if(!(company.equals("NA")||product.equals("NA")))
```

```
        {
```

```
            context.write(key,value);
```

```
        }
```

```
}
}
```

```
[acadgild@localhost A 1.6]$ ls
Assignment 1.6  assignment1.6.jar  television.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost A 1.6]$ hadoop fs -put assignment1.6.jar /user/acadgild/hadoop
18/05/18 13:33:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
^[[A[acadgild@localhost A 1.6]$ hadoop fs -put television.txt /user/acadgild/hadoop
18/05/18 13:33:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost A 1.6]$ hadoop fs -ls /user/acadgild/hadoop
18/05/18 13:33:59 WARN util.NativeCodeLoader: Unable to load native-hadoop libfary for your platform... using builtin-java cl
asses where applicable
Found 4 items
-rw-r--r-- 1 acadgild supergroup 2800 2018-05-18 13:33 /user/acadgild/hadoop/assignment1.6.jar
-rwxrwx--- 1 acadgild supergroup 168 2018-05-17 13:31 /user/acadgild/hadoop/max-temp.txt
-rw-r--r-- 1 acadgild supergroup 733 2018-05-18 13:33 /user/acadgild/hadoop/television.txt
-rw-r--r-- 1 acadgild supergroup 227 2018-05-17 01:02 /user/acadgild/hadoop/word-count.txt
[acadgild@localhost A 1.6]$ hadoop jar assignment1.6.jar mapreduce.task.assignment.Task /user/acadgild/hadoop/television.txt
/user/acadgild/hadoop/assignment1.6_output
18/05/18 13:34:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/05/18 13:34:12 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/05/18 13:34:13 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
18/05/18 13:34:14 INFO input.FileInputFormat: Total input paths to process : 1
18/05/18 13:34:14 INFO mapreduce.JobSubmitter: number of splits:1
18/05/18 13:34:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1526626806481_0008
18/05/18 13:34:14 INFO impl.YarnClientImpl: Submitted application application_1526626806481_0008
18/05/18 13:34:14 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1526626806481_0008/
18/05/18 13:34:14 INFO mapreduce.Job: Running job: job_1526626806481_0008
18/05/18 13:34:23 INFO mapreduce.Job: Job job_1526626806481_0008 running in uber mode : false
18/05/18 13:34:23 INFO mapreduce.Job: map 0% reduce 0%
18/05/18 13:34:28 INFO mapreduce.Job: map 100% reduce 0%
18/05/18 13:34:34 INFO mapreduce.Job: map 100% reduce 100%
18/05/18 13:34:34 INFO mapreduce.Job: Job job_1526626806481_0008 completed successfully
18/05/18 13:34:34 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=812
  FILE: Number of bytes written=217429
```

Execution Steps:-

- television.txt was put to hdfs.
- After the driver code and mapping code , a jar file of was generated.
- assignmen1.6.jar file(mapping code) was put to hdfs.
- Mapping task was performed by the following command – **hadoop jar assignment1.6.jar mapreduce.task.assignment.task /user/acadgild/hadoop/television.txt /user/acadgild/hadoop/assignment1.6_output**
- To see the content of the output following command was used - **hadoop cat /user/acadgild/hadoop/assignment1.6_output/part-r-00000**

```
Applications Places System acadgild@localhost:~/RITESH/A_1.6
Virtual memory (bytes) snapshot=4163575808
Total committed heap usage (bytes)=301465600
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=733
File Output Format Counters
Bytes Written=706
[acadgild@localhost A_1.6]$ hadoop fs -ls /user/acadgild/hadoop/assignment1.6_output
18/05/18 13:39:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2018-05-18 13:34 /user/acadgild/hadoop/assignment1.6_output/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 706 2018-05-18 13:34 /user/acadgild/hadoop/assignment1.6_output/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost A_1.6]$ hadoop fs -cat /user/acadgild/hadoop/assignment1.6_output/part-r-00000
18/05/18 13:39:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
0 Samsung|Optima|14|Madhya Pradesh|132401|14200
47 Onida|Lucid|18|Uttar Pradesh|232401|16200
90 Akai|Decent|16|Kerala|922401|12200
126 Lava|Attention|20|Assam|454601|24200
164 Zen|Super|14|Maharashtra|619082|9200
202 Samsung|Optima|14|Madhya Pradesh|132401|14200
249 Onida|Lucid|18|Uttar Pradesh|232401|16200
292 Onida|Decent|14|Uttar Pradesh|232401|16200
369 Lava|Attention|20|Assam|454601|24200
407 Zen|Super|14|Maharashtra|619082|9200
445 Samsung|Optima|14|Madhya Pradesh|132401|14200
532 Samsung|Decent|16|Kerala|922401|12200
571 Lava|Attention|20|Assam|454601|24200
609 Samsung|Super|14|Maharashtra|619082|9200
651 Samsung|Super|14|Maharashtra|619082|9200
693 Samsung|Super|14|Maharashtra|619082|9200
[acadgild@localhost A_1.6]$
```

Output

As in the above screen shot ,it is evident that the content of the output file does not contain any records whose company name or product name equals “NA”.