

Session 4: MR -INTRODUCTION

Assignment 1

BigData – Assignment-4 (MapReduce)

Input File: Television.txt

```
1 Samsung|Optima|14|Madhya Pradesh|132401|14200
2 Onida|Lucid|18|Uttar Pradesh|232401|16200
3 Akai|Decent|16|Kerala|922401|12200
4 Lava|Attention|20|Assam|454601|24200
5 Zen|Super|14|Maharashtra|619082|9200
6 Samsung|Optima|14|Madhya Pradesh|132401|14200
7 Onida|Lucid|18|Uttar Pradesh|232401|16200
8 Onida|Decent|14|Uttar Pradesh|232401|16200
9 Onida|NA|16|Kerala|922401|12200
10 Lava|Attention|20|Assam|454601|24200
11 Zen|Super|14|Maharashtra|619082|9200
12 Samsung|Optima|14|Madhya Pradesh|132401|14200
13 NA|Lucid|18|Uttar Pradesh|232401|16200
14 Samsung|Decent|16|Kerala|922401|12200
15 Lava|Attention|20|Assam|454601|24200
16 Samsung|Super|14|Maharashtra|619082|9200
17 Samsung|Super|14|Maharashtra|619082|9200
18 Samsung|Super|14|Maharashtra|619082|9200
```

Task – 1:

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

Program:

Mapper program: ProductSalesMapper.java

```
package com.company.sales;

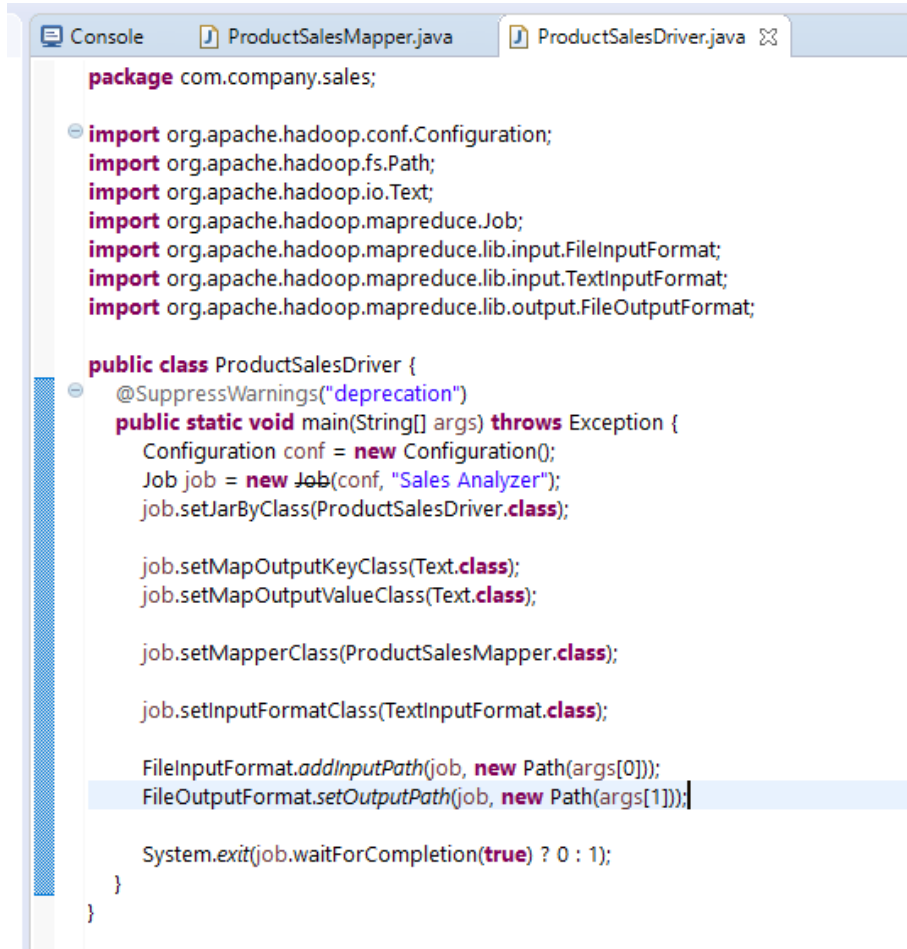
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

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

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        //StringTokenizer tokens = new StringTokenizer(value.toString().trim(), "|");
        String[] tokens = value.toString().split("|");
        if(tokens[0] != "NA" && tokens[1] != "NA") {
            context.write(new Text("Summary"), value);
        }
    }
}
```

BigData – Assignment-4 (MapReduce)

Driver program: ProductSalesDriver.java



```
package com.company.sales;

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

public class ProductSalesDriver {
    @SuppressWarnings("deprecation")
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Sales Analyzer");
        job.setJarByClass(ProductSalesDriver.class);

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

        job.setMapperClass(ProductSalesMapper.class);

        job.setInputFormatClass(TextInputFormat.class);

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

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

The JAR file is created for the project. Then JAR file is copied to VM box. Then input file is also moved to HDFS path `"/mapreduce/television.txt"`.

Then the below command is executed to trigger the driver program by passing input file path and output file path as arguments to the program.

```
hadoop jar /home/acadgild/workspace/Mapreduce/ProductSalesAnalyzer.jar /mapreduce/television.txt /mapreduce/sales-out
```

BigData – Assignment-4 (MapReduce)

Output:

```
acadgild@localhost:~/install/hadoop/hadoop-2.6.5/bin
File Edit View Search Terminal Help
Total time spent by all reduce tasks (ms)=6109
Total vcore-milliseconds taken by all map tasks=5480
Total vcore-milliseconds taken by all reduce tasks=6109
Total megabyte-milliseconds taken by all map tasks=5611520
Total megabyte-milliseconds taken by all reduce tasks=6255616
Map-Reduce Framework
  Map input records=18
  Map output records=16
  Map output bytes=774
  Map output materialized bytes=812
  Input split bytes=111
  Combine input records=0
  Combine output records=0
  Reduce input groups=1
  Reduce shuffle bytes=812
  Reduce input records=16
  Reduce output records=16
  Spilled Records=32
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=175
  CPU time spent (ms)=1570
  Physical memory (bytes) snapshot=299028480
  Virtual memory (bytes) snapshot=4117905408
  Total committed heap usage (bytes)=170004480
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=774
[acadgild@localhost bin]$
```

As per the execution summary, the input file has 18 records of which, 2 records have "NA". Those invalid records are filtered and only 16 records are output.

```
Bytes Written=774
[acadgild@localhost bin]$ ./hadoop fs -cat /mapreduce/sales-output/part-r-00000
19/02/18 23:28:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... us
ing builtin-java classes where applicable
Summary Samsung|Super|14|Maharashtra|619082|9200
Summary Samsung|Super|14|Maharashtra|619082|9200
Summary Samsung|Super|14|Maharashtra|619082|9200
Summary Lava|Attention|20|Assam|454601|24200
Summary Samsung|Decent|16|Kerala|922401|12200
Summary Samsung|Optima|14|Madhya Pradesh|132401|14200
Summary Zen|Super|14|Maharashtra|619082|9200
Summary Lava|Attention|20|Assam|454601|24200
Summary Onida|Decent|14|Uttar Pradesh|232401|16200
Summary Onida|Lucid|18|Uttar Pradesh|232401|16200
Summary Samsung|Optima|14|Madhya Pradesh|132401|14200
Summary Zen|Super|14|Maharashtra|619082|9200
Summary Lava|Attention|20|Assam|454601|24200
Summary Akai|Decent|16|Kerala|922401|12200
Summary Onida|Lucid|18|Uttar Pradesh|232401|16200
Summary Samsung|Optima|14|Madhya Pradesh|132401|14200
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost bin]$
```

BigData – Assignment-4 (MapReduce)

Task – 2:

Programs:

CompanySalesMapper.java

```
Console CompanySalesMapper.java CompanySalesReducer.java CompanySalesDriver.java

package com.company.sales;

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

public class CompanySalesMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        //StringTokenizer tokens = new StringTokenizer(value.toString().trim(), "|");
        String[] tokens = value.toString().split("\\|");
        if (!tokens[0].equalsIgnoreCase("NA") && !tokens[1].equalsIgnoreCase("NA")) {
            context.write(new Text(tokens[0]), new IntWritable(1));
        }
    }
}
```

CompanySalesReducer.java

```
Console CompanySalesMapper.java CompanySalesReducer.java CompanySalesDriver.java

package com.company.sales;

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

public class CompanySalesReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text company, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable value : values) {
            sum += value.get();
        }
        context.write(company, new IntWritable(sum));
    }
}
```

BigData – Assignment-4 (MapReduce)

CompanySalesDriver.java

```
Console CompanySalesMapper.java CompanySalesReducer.java CompanySalesDriver.java
package com.company.sales;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
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;

// TASK-2 : Write a Map Reduce program to calculate the total units sold for each Company.
public class CompanySalesDriver {
    @SuppressWarnings("deprecation")
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Company Sales Analyzer");

        // Configure Mapper, Reducer and Driver classes
        job.setJarByClass(CompanySalesDriver.class);
        job.setMapperClass(CompanySalesMapper.class);
        job.setReducerClass(CompanySalesReducer.class);

        // Setup mapper output key and value types
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);

        // Setup input file format
        job.setInputFormatClass(TextInputFormat.class);

        // Configure input and output file paths
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
    }
}
```

Execution:

```
hadoop jar /home/acadgild/workspace/Mapreduce/CompanySalesAnalyzer.jar
/mapreduce/television.txt /mapreduce/companysales-out
```

BigData – Assignment-4 (MapReduce)

Output:

```
acadgild@localhost:~/install/hadoop/hadoop-2.6.5/bin
File Edit View Search Terminal Help
3972 JobHistoryServer
3498 NodeManager
2942 NameNode
[acadgild@localhost sbin]$ cd ..
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost hadoop-2.6.5]$ cd bin
[acadgild@localhost bin]$ ./hadoop fs -ls /mapreduce/companysales-out
19/02/19 00:10:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2019-02-19 00:09 /mapreduce/companysales-out/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 38 2019-02-19 00:09 /mapreduce/companysales-out/part-r-000000
[acadgild@localhost bin]$ ./hadoop fs -cat /mapreduce/companysales-out/part-r-000000
19/02/19 00:11:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Akai 1
Lava 3
Onida 3
Samsung 7
Zen 2
[acadgild@localhost bin]$
```

Task – 3:

Programs:

CountrySalesMapper.java

```
Console CountrySalesMapper.java CountrySalesReducer.java CountrySalesDriver.java

package com.company.sales;

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

public class CountrySalesMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String[] tokens = value.toString().split("\\\\");
        if (!tokens[0].equalsIgnoreCase("NA") && !tokens[1].equalsIgnoreCase("NA")) {
            // Filter records of company as Onida
            if (tokens[0].equalsIgnoreCase("Onida")) {
                context.write(new Text(tokens[3]), new IntWritable(1));
            }
        }
    }
}
```

BigData – Assignment-4 (MapReduce)

CountrySalesReducer.java

```
package com.company.sales;

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

public class CountrySalesReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text company, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable value : values) {
            sum += value.get();
        }
        context.write(company, new IntWritable(sum));
    }
}
```

CountrySalesDriver.java

```
package com.company.sales;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
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;

// TASK-3 : Write a Map Reduce program to calculate the total units sold in each state for Onida company.
public class CountrySalesDriver {
    @SuppressWarnings("deprecation")
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Company Sales Analyzer");

        // Configure Mapper, Reducer and Driver classes
        job.setJarByClass(CountrySalesDriver.class);
        job.setMapperClass(CountrySalesMapper.class);
        job.setReducerClass(CountrySalesReducer.class);

        // Setup mapper output key and value types
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);

        // Setup input file format
        job.setInputFormatClass(TextInputFormat.class);

        // Configure input and output file paths
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
    }
}
```


BigData – Assignment-4 (MapReduce)

Execution:

```
hadoop jar /home/acadgild/workspace/Mapreduce/CountrySalesAnalyzer.jar /mapreduce/television.txt  
/mapreduce/countrysales-out
```

Output:

```
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost bin]$ hadoop fs -ls /mapreduce/countrysales-out  
19/02/19 21:34:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform  
... using builtin-java classes where applicable  
Found 2 items  
-rw-r--r-- 1 acadgild supergroup          0 2019-02-19 21:33 /mapreduce/countrysales-out/_SUCCESS  
-rw-r--r-- 1 acadgild supergroup       16 2019-02-19 21:33 /mapreduce/countrysales-out/part-r-000000  
[acadgild@localhost bin]$ hadoop fs -cat /mapreduce/countrysales-out/part-r-000000  
19/02/19 21:34:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform  
... using builtin-java classes where applicable  
Uttar Pradesh      3  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost bin]$
```