Session 3

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 3 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 3 – Assignment 2** |
|  |  |
| **Start Date:** | 02/09/2017 |
|  |  |
|  |  |

# 

# Contents

[Contents 2](#_Toc493042266)

[Change History 3](#_Toc493042267)

[1. Problem Statement 4](#_Toc493042268)

[2. The input file (television.txt) 4](#_Toc493042269)

[3. Solutions 5](#_Toc493042270)

[4. Running Job 8](#_Toc493042271)

[5. Check for Results 9](#_Toc493042272)

[6. View Results 9](#_Toc493042273)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 02/09/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

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.

2. Write a Map Reduce program to calculate the total units sold for each Company.

3. Write a Map Reduce program to calculate the total units sold in each state for Onida Company.

# The input file (television.txt)

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

Onida|Lucid|18|Uttar Pradesh|232401|16200

Akai|Decent|16|Kerala|922401|12200

Lava|Attention|20|Assam|454601|24200

Zen|Super|14|Maharashtra|619082|9200

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

Onida|Lucid|18|Uttar Pradesh|232401|16200

Onida|Decent|14|Uttar Pradesh|232401|16200

Onida|NA|16|Kerala|922401|12200

Lava|Attention|20|Assam|454601|24200

Zen|Super|14|Maharashtra|619082|9200

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

NA|Lucid|18|Uttar Pradesh|232401|16200

Samsung|Decent|16|Kerala|922401|12200

Lava|Attention|20|Assam|454601|24200

Samsung|Super|14|Maharashtra|619082|9200

Samsung|Super|14|Maharashtra|619082|9200

Samsung|Super|14|Maharashtra|619082|9200

# Solutions 3.1

The required jar file was created in Eclipse. There is a Driver and a mapper no reducer is required.

Driver TvCleanDriver

Mapper TvCleanMapper

The file and the jar file are copied to hdfs using hdfs fs –copyFromLocal.

**TVCleanDriver**

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.io.Text;

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;

@SuppressWarnings("unused")

public class TV {

public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException

{

//Check if input parameters provided appropriately

if(args==null || args.length!=2)

{

System.err.println("Wrong input parameters provided");

System.exit(-1);

}

//Instantiate configuration object

Configuration conf = new Configuration();

//Instantiate job object

@SuppressWarnings("deprecation")

Job job = new Job(conf,"Remove records with NA");

job.setJarByClass(TV.class);

/\*

\* Input path for data

\*/

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

/\*

\* Output path for results

\*/

Path outputPath = new Path(args[1]);

FileOutputFormat.setOutputPath(job, outputPath);

//Delete output directory if already existing will fail if already exists

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

//Set mapper class to mapper TVMapper

job.setMapperClass(TVMapper.class);

//Set input and output format class

job.setInputFormatClass(TextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.class);

//Set output key'value class types

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(Text.class);

//To set the job with no reducer

job.setNumReduceTasks(0);

//Execute the job and wait until completion and then exit

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

}

}

**TvCleanMapper**

import java.io.IOException;

import java.util.StringTokenizer;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

// Set data types

public class TVMapper extends Mapper<LongWritable,Text,Text,Text>

{

private final static String DELIMITER = "|";

private final static String NA = "NA";

@Override

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

{

//Input format : Company|Product|Size|State|ZIP|Price

String strValue = value.toString();

System.out.println("Current value is "+strValue);

StringTokenizer tokenizer = new StringTokenizer(strValue, DELIMITER);

String strCompanyName = tokenizer.nextToken();

String strProductName = tokenizer.nextToken();

//Check if CompanyName or Product Name is 'NA'

if(!strCompanyName.equalsIgnoreCase(NA) && !strProductName.equalsIgnoreCase(NA))

{

context.write(value,null);

System.out.println("Put value "+strValue+" in context");

}

}

}

# Running Job

[acadgild@localhost Documents]$ yarn jar TVCleaner.jar television.txt tvclean

17/09/03 13:29:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/09/03 13:29:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

17/09/03 13:29:33 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

17/09/03 13:29:34 INFO input.FileInputFormat: Total input paths to process : 1

17/09/03 13:29:34 INFO mapreduce.JobSubmitter: number of splits:1

17/09/03 13:29:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1504029604882\_0016

17/09/03 13:29:34 INFO impl.YarnClientImpl: Submitted application application\_1504029604882\_0016

17/09/03 13:29:34 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1504029604882\_0016/

17/09/03 13:29:34 INFO mapreduce.Job: Running job: job\_1504029604882\_0016

17/09/03 13:29:41 INFO mapreduce.Job: Job job\_1504029604882\_0016 running in uber mode : false

17/09/03 13:29:41 INFO mapreduce.Job: map 0% reduce 0%

17/09/03 13:29:47 INFO mapreduce.Job: map 100% reduce 0%

17/09/03 13:29:48 INFO mapreduce.Job: Job job\_1504029604882\_0016 completed successfully

17/09/03 13:29:48 INFO mapreduce.Job: Counters: 30

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=105766

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=848

HDFS: Number of bytes written=646

HDFS: Number of read operations=5

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=1

Data-local map tasks=1

Total time spent by all maps in occupied slots (ms)=3622

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=3622

Total vcore-seconds taken by all map tasks=3622

Total megabyte-seconds taken by all map tasks=3708928

Map-Reduce Framework

Map input records=18

Map output records=16

Input split bytes=115

Spilled Records=0

Failed Shuffles=0

Merged Map outputs=0

GC time elapsed (ms)=58

CPU time spent (ms)=650

Physical memory (bytes) snapshot=159854592

Virtual memory (bytes) snapshot=2079461376

Total committed heap usage (bytes)=147849216

File Input Format Counters

Bytes Read=733

File Output Format Counters

Bytes Written=646

# Check for Results

[acadgild@localhost Documents]$ hadoop fs -ls /user/acadgild/tvclean

17/09/02 19:46:58 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 2017-09-02 19:44 /user/acadgild/tvclean/\_SUCCESS

-rw-r--r-- 1 acadgild supergroup 646 2017-09-02 19:44 /user/acadgild/tvclean/part-m-00000

Note m - indicates it is a map job r would indicate a reduced job.

# View Results

[acadgild@localhost Documents]$ hadoop fs -cat tvclean/part-m-00000

17/09/02 19:47:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

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

Onida|Lucid|18|Uttar Pradesh|232401|16200

Akai|Decent|16|Kerala|922401|12200

Lava|Attention|20|Assam|454601|24200

Zen|Super|14|Maharashtra|619082|9200

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

Onida|Lucid|18|Uttar Pradesh|232401|16200

Onida|Decent|14|Uttar Pradesh|232401|16200

Lava|Attention|20|Assam|454601|24200

Zen|Super|14|Maharashtra|619082|9200

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

Samsung|Decent|16|Kerala|922401|12200

Lava|Attention|20|Assam|454601|24200

Samsung|Super|14|Maharashtra|619082|9200

Samsung|Super|14|Maharashtra|619082|9200

Samsung|Super|14|Maharashtra|619082|9200