Task 1:

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

Data:

```
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
```

Program:

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Mapper.Context;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;


public class TVAssignMent1 {

public static class TVMapper extends Mapper<LongWritable,Text,NullWritable,Text>{

private final static NullWritable okey = NullWritable.get();
        //private Text word = new Text();
```

```
            public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
                System.out.println("START#map()");
                String tokens[] = value.toString().split("\\|");
                System.out.println(" KEY : "+ key.toString()+ "|| value : "+value.toString());

                if(tokens.length==6) {
                        if(!(tokens[0].equals("NA")) || !(tokens[1].equals("NA"))) {
                                context.write(okey, value);
                                System.out.println(value.toString());
                        }
                }

                System.out.println("END#map()");
            }
}

            public static class TVMapperWithCounter extends
Mapper<LongWritable,Text,NullWritable,Text> {
                        private final static NullWritable okey = NullWritable.get();

                        enum InvalidRecords{CompanyInvalidId, ProductInvalidId, DataMissing};

                        public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
                                System.out.println("START#TVMapperWithCounter#map()");
                                String[] tokens = value.toString().split("\\|");

                                System.out.println("DATA ---> KEY : "+ key.toString()+ "|| value :
"+value.toString());

                                if(tokens.length == 6) {
                                        if(tokens[0].equals("NA")) {

context.getCounter(InvalidRecords.CompanyInvalidId).increment(1);
                                        }else if(tokens[1].equals("NA")) {

context.getCounter(InvalidRecords.ProductInvalidId).increment(1);
                                        }else {
                                                System.out.println("REFINED DATA--> KEY : "+ key.toString()+
"|| value : "+value.toString());
                                                context.write(okey, value);
                                        }

                                }else {
                                        context.getCounter(InvalidRecords.DataMissing).increment(1);
                                }
                                System.out.println("END#TVMapperWithCounter#map()");
```

```
		}
	}

	@SuppressWarnings("deprecation")
	public static void main(String[] args) throws Exception {
		//create an instance of Configuration object
		Configuration conf = new Configuration();
		conf.addResource(new Path("/home/acadgild/install/hadoop/hadoop-
2.6.5/etc/hadoop/core-site.xml"));
		conf.addResource(new Path("/home/acadgild/install/hadoop/hadoop-
2.6.5/etc/hadoop/hdfs-site.xml"));

		//creatkeye an instance of FileSystem that holds Filesystem namespace
		FileSystem fs = FileSystem.get(conf);
		//variables to hold path of input file and output directory
		String inPath;
		String outPath;


		System.out.println("Usage: wordcount <input file> <output dir>");
		System.out.println("Using default file: WordCount.txt");

		inPath = "/user/Hadoop/TVAssignment/Input/television.txt";
		outPath = "/user/Hadoop/TVAssignment/Output/";

		//inPath = "/home/acadgild/Desktop/MyDocument/television.txt";
		//outPath = "/home/acadgild/Desktop/MyDocument/Output/";

		//create an instance of job
		try {
			Job job = new Job(conf, "TVASSIGNMENT");
			job.setJarByClass(TVAssignMent1.class);
			//job.setMapperClass(TVMapper.class);
			job.setMapperClass(TVMapperWithCounter.class);
			job.setNumReduceTasks(0); // as no reducer class we are using we need  to set
here value zero
			job.setMapOutputKeyClass(NullWritable.class);
			job.setMapOutputValueClass(Text.class);
			job.setOutputKeyClass(NullWritable.class);
			job.setOutputValueClass(Text.class);


			FileInputFormat.addInputPath(job, new Path(inPath));

			if (fs.exists(new Path(outPath))) {
				fs.delete(new Path(outPath), true);
			}
			FileOutputFormat.setOutputPath(job, new Path(outPath));
```

```
                System.exit(job.waitForCompletion(true) ? 0 : 1);
            }catch(Exception e) {
                System.out.println(e);
            }
        }
}
```

To compile the program in eclipse need to add all the jars like Hadoop Mapreduce jars, Hadoop Commons Jar, Hadoop Hdfc jars, Hadoop Yarn client jar

Out Put of the Program:

```
2018-10-31 23:14:19,025 WARN  [main] util.NativeCodeLoader (NativeCodeLoader.java:<clinit>(62)) - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Usage: wordcount <input file> <output dir>
Using default file: WordCount.txt
2018-10-31 23:14:21,103 INFO  [main] Configuration.deprecation (Configuration.java:warnOnceIfDeprecated(1129)) - session.id is deprecated. Instead, use dfs.metrics.session-id
2018-10-31 23:14:21,112 INFO  [main] jvm.JvmMetrics (JvmMetrics.java:init(76)) - Initializing JVM Metrics with processName=JobTracker, sessionId=
2018-10-31 23:14:21,847 WARN  [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(64)) - Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2018-10-31 23:14:21,866 WARN  [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(171)) - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2018-10-31 23:14:21,896 INFO  [main] input.FileInputFormat (FileInputFormat.java:listStatus(281)) - Total input paths to process : 1
2018-10-31 23:14:22,249 INFO  [main] mapreduce.JobSubmitter (JobSubmitter.java:submitJobInternal(199)) - number of splits:1
2018-10-31 23:14:22,575 INFO  [main] mapreduce.JobSubmitter (JobSubmitter.java:printTokens(288)) - Submitting tokens for job: job_local212922866_0001
2018-10-31 23:14:22,960 INFO  [main] mapreduce.Job (Job.java:submit(1301)) - The url to track the job: http://localhost:8080/
2018-10-31 23:14:22,961 INFO  [main] mapreduce.Job (Job.java:monitorAndPrintJob(1346)) - Running job: job_local212922866_0001
2018-10-31 23:14:22,974 INFO  [Thread-19] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(471)) - OutputCommitter set in config null
2018-10-31 23:14:22,997 INFO  [Thread-19] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(489)) - OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2018-10-31 23:14:23,145 INFO  [Thread-19] mapred.LocalJobRunner (LocalJobRunner.java:runTasks(448)) - Waiting for map tasks
2018-10-31 23:14:23,147 INFO  [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:run(224)) - Starting task: attempt_local212922866_0001_m_000000_0
```

2018-10-31 23:14:23,257 INFO [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:initialize(587)) - Using ResourceCalculatorProcessTree : [ ]
2018-10-31 23:14:23,270 INFO [LocalJobRunner Map Task Executor #0] mapred.MapTask (MapTask.java:runNewMapper(753)) - Processing split: hdfs://localhost:8020/user/Hadoop/TVAssignment/Input/television.txt:0+733
START#TVMapperWithCounter#map()
DATA ---> KEY : 0|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
REFINED DATA--> KEY : 0|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 47|| value : Onida|Lucid|18|Uttar Pradesh|232401|16200
REFINED DATA--> KEY : 47|| value : Onida|Lucid|18|Uttar Pradesh|232401|16200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 90|| value : Akai|Decent|16|Kerala|922401|12200
REFINED DATA--> KEY : 90|| value : Akai|Decent|16|Kerala|922401|12200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 126|| value : Lava|Attention|20|Assam|454601|24200
REFINED DATA--> KEY : 126|| value : Lava|Attention|20|Assam|454601|24200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 164|| value : Zen|Super|14|Maharashtra|619082|9200
REFINED DATA--> KEY : 164|| value : Zen|Super|14|Maharashtra|619082|9200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 202|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
REFINED DATA--> KEY : 202|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 249|| value : Onida|Lucid|18|Uttar Pradesh|232401|16200
REFINED DATA--> KEY : 249|| value : Onida|Lucid|18|Uttar Pradesh|232401|16200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 292|| value : Onida|Decent|14|Uttar Pradesh|232401|16200
REFINED DATA--> KEY : 292|| value : Onida|Decent|14|Uttar Pradesh|232401|16200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 336|| value : Onida|NA|16|Kerala|922401|12200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 369|| value : Lava|Attention|20|Assam|454601|24200
REFINED DATA--> KEY : 369|| value : Lava|Attention|20|Assam|454601|24200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 407|| value : Zen|Super|14|Maharashtra|619082|9200
REFINED DATA--> KEY : 407|| value : Zen|Super|14|Maharashtra|619082|9200
END#TVMapperWithCounter#map()

START#TVMapperWithCounter#map()
DATA ---> KEY : 445|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
REFINED DATA--> KEY : 445|| value : Samsung|Optima|14|Madhya Pradesh|132401|14200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 492|| value : NA|Lucid|18|Uttar Pradesh|232401|16200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 532|| value : Samsung|Decent|16|Kerala|922401|12200
REFINED DATA--> KEY : 532|| value : Samsung|Decent|16|Kerala|922401|12200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 571|| value : Lava|Attention|20|Assam|454601|24200
REFINED DATA--> KEY : 571|| value : Lava|Attention|20|Assam|454601|24200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 609|| value : Samsung|Super|14|Maharashtra|619082|9200
REFINED DATA--> KEY : 609|| value : Samsung|Super|14|Maharashtra|619082|9200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 651|| value : Samsung|Super|14|Maharashtra|619082|9200
REFINED DATA--> KEY : 651|| value : Samsung|Super|14|Maharashtra|619082|9200
END#TVMapperWithCounter#map()
START#TVMapperWithCounter#map()
DATA ---> KEY : 693|| value : Samsung|Super|14|Maharashtra|619082|9200
REFINED DATA--> KEY : 693|| value : Samsung|Super|14|Maharashtra|619082|9200
END#TVMapperWithCounter#map()
2018-10-31 23:14:23,617 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:statusUpdate(591)) -
2018-10-31 23:14:23,967 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1367)) - Job job_local212922866_0001 running in uber mode : false
2018-10-31 23:14:23,971 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1374)) - map 0% reduce 0%
2018-10-31 23:14:24,056 INFO [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:done(1001)) - Task:attempt_local212922866_0001_m_000000_0 is done. And is in the process of committing
2018-10-31 23:14:24,086 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:statusUpdate(591)) -
2018-10-31 23:14:24,089 INFO [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:commit(1162)) - Task attempt_local212922866_0001_m_000000_0 is allowed to commit now
2018-10-31 23:14:24,118 INFO [LocalJobRunner Map Task Executor #0] output.FileOutputCommitter (FileOutputCommitter.java:commitTask(439)) - Saved output of task 'attempt_local212922866_0001_m_000000_0' to hdfs://localhost:8020/user/Hadoop/TVAssignment/Output/_temporary/0/task_local212922866_0001_m_000000
2018-10-31 23:14:24,131 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:statusUpdate(591)) - map

2018-10-31 23:14:24,132 INFO [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:sendDone(1121)) - Task 'attempt_local212922866_0001_m_000000_0' done.
2018-10-31 23:14:24,132 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:run(249)) - Finishing task: attempt_local212922866_0001_m_000000_0
2018-10-31 23:14:24,132 INFO [Thread-19] mapred.LocalJobRunner (LocalJobRunner.java:runTasks(456)) - map task executor complete.
2018-10-31 23:14:24,973 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1374)) - map 100% reduce 0%
2018-10-31 23:14:24,974 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1385)) - Job job_local212922866_0001 completed successfully
2018-10-31 23:14:25,001 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1392)) - Counters: 25

- File System Counters
  - FILE: Number of bytes read=188
  - FILE: Number of bytes written=271426
  - FILE: Number of read operations=0
  - FILE: Number of large read operations=0
  - FILE: Number of write operations=0
  - HDFS: Number of bytes read=733
  - HDFS: Number of bytes written=646
  - HDFS: Number of read operations=8
  - HDFS: Number of large read operations=0
  - HDFS: Number of write operations=4
- Map-Reduce Framework
  - Map input records=18
  - Map output records=16
  - Input split bytes=132
  - Spilled Records=0
  - Failed Shuffles=0
  - Merged Map outputs=0
  - GC time elapsed (ms)=17
  - CPU time spent (ms)=0
  - Physical memory (bytes) snapshot=0
  - Virtual memory (bytes) snapshot=0
  - Total committed heap usage (bytes)=32571392
- TVAssignMent1$TVMapperWithCounter$InvalidRecords
  - CompanyInvalidId=1
  - ProductInvalidId=1
- File Input Format Counters
  - Bytes Read=733
- File Output Format Counters
  - Bytes Written=646