BigData And Hadoop

Assignment 1 of Session 4

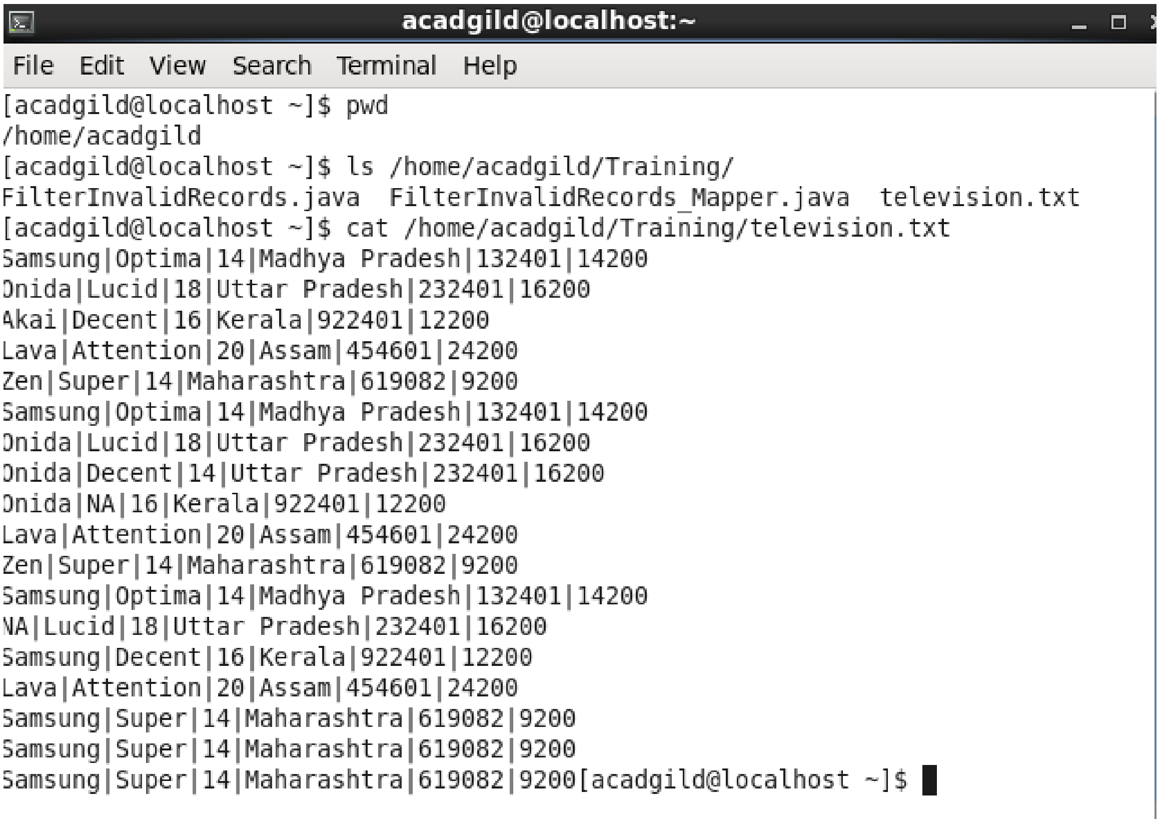
Problem Statement :
Given a dataset of sales of different TV sets across different locations, with records in the format :
Company Name|Product Name|Size in inches|State|Pin Code|Price
Eg : Samsung|Optima|14|Madhya Pradesh|132401|14200
There are some invalid records which contain 'NA' in either Company Name or Product Name.
Write a Map Reduce program to filter out the invalid records.

Solution

Code files are as follows :
Mapper class : FilterInvalidRecords_Mapper.java
Driver class : FilterInvalidRecords.java

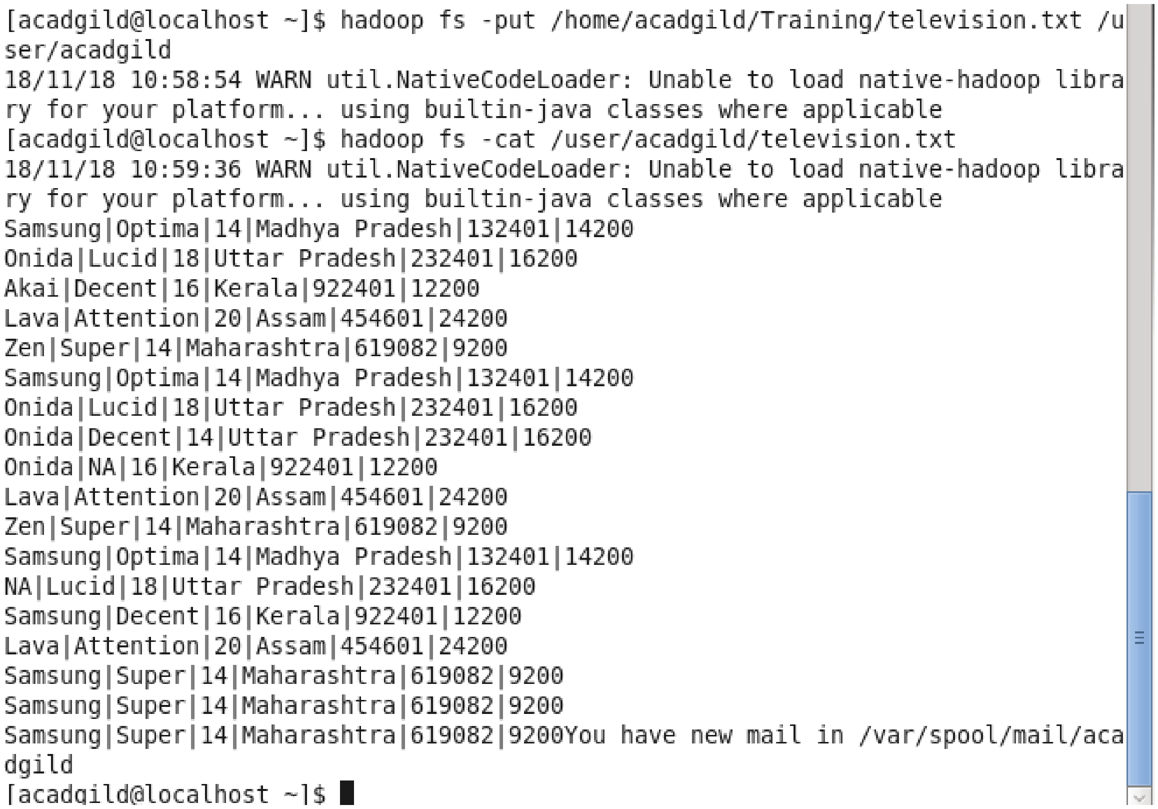
Snapshots of the output are as follows :

- Input file present in '/home/acadgild/Training'. Its name is television.txt



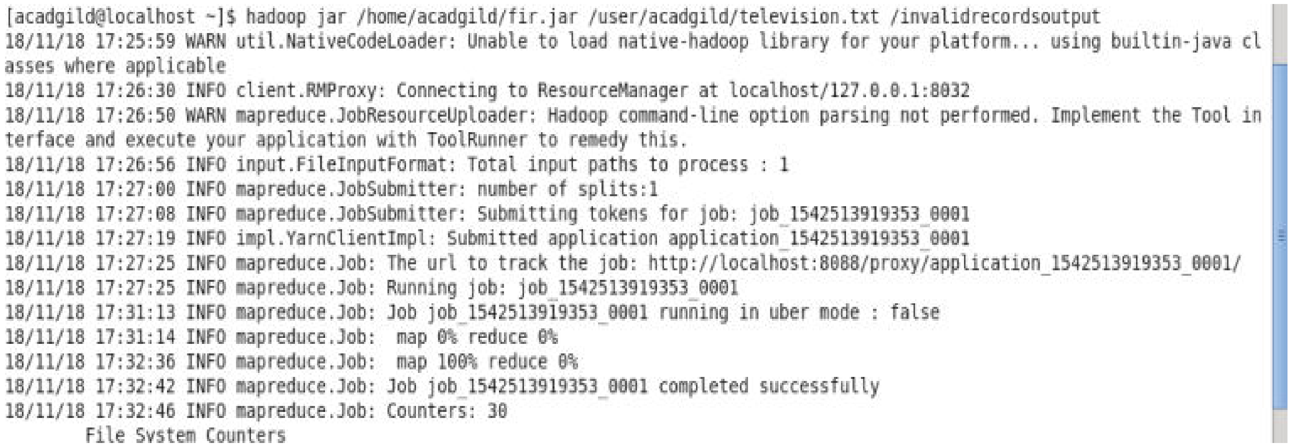
```
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ ls /home/acadgild/Training/
FilterInvalidRecords.java  FilterInvalidRecords_Mapper.java  television.txt
[acadgild@localhost ~]$ cat /home/acadgild/Training/television.txt
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar 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[acadgild@localhost ~]$ █
```

- Command executed to put television.txt to hdfs system



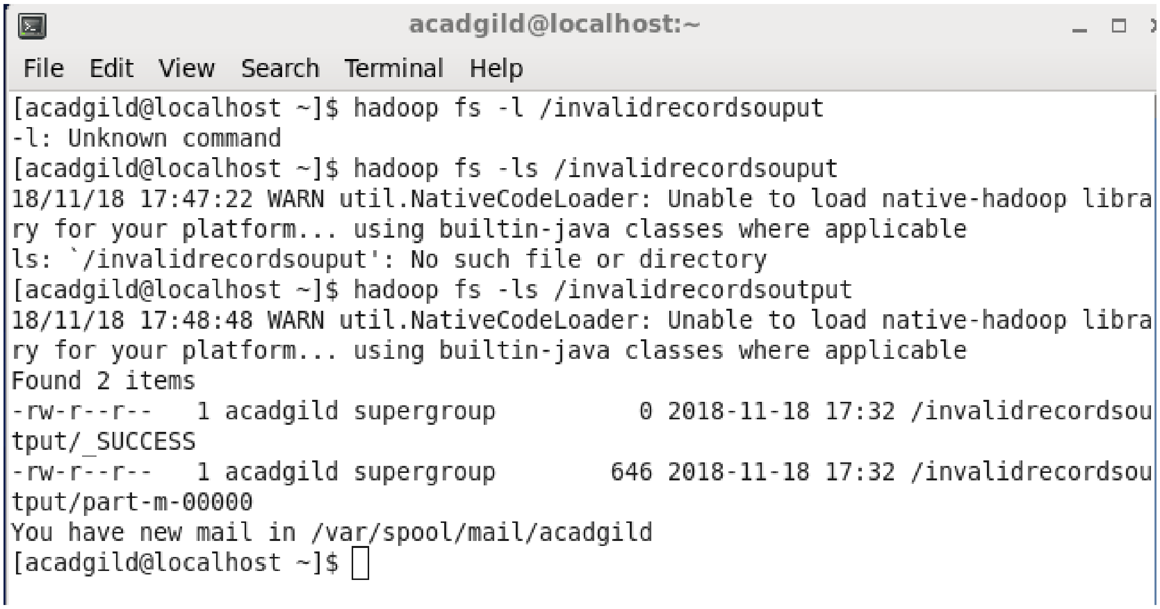
```
[acadgild@localhost ~]$ hadoop fs -put /home/acadgild/Training/television.txt /u
ser/acadgild
18/11/18 10:58:54 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/television.txt
18/11/18 10:59:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry 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
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|9200You have new mail in /var/spool/mail/aca
dgild
[acadgild@localhost ~]$ 
```

- Executing the map-reduce program on Hadoop. Jar is placed locally at /home/acadgild/. Its name is fir.jar



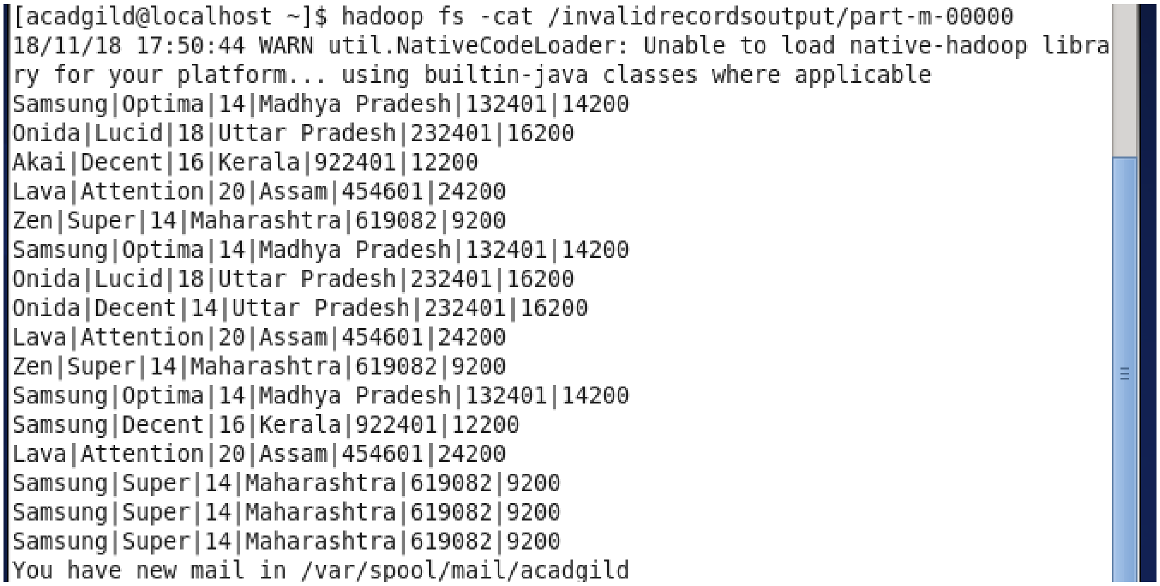
```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/fir.jar /user/acadgild/television.txt /invalidrecordsoutput
18/11/18 17:25:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/11/18 17:26:30 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/11/18 17:26:50 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
18/11/18 17:26:56 INFO input.FileInputFormat: Total input paths to process : 1
18/11/18 17:27:00 INFO mapreduce.JobSubmitter: number of splits:1
18/11/18 17:27:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1542513919353_0001
18/11/18 17:27:19 INFO impl.YarnClientImpl: Submitted application application_1542513919353_0001
18/11/18 17:27:25 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1542513919353_0001/
18/11/18 17:27:25 INFO mapreduce.Job: Running job: job_1542513919353_0001
18/11/18 17:31:13 INFO mapreduce.Job: Job job_1542513919353_0001 running in uber mode : false
18/11/18 17:31:14 INFO mapreduce.Job:  map 0% reduce 0%
18/11/18 17:32:36 INFO mapreduce.Job:  map 100% reduce 0%
18/11/18 17:32:42 INFO mapreduce.Job: Job job_1542513919353_0001 completed successfully
18/11/18 17:32:46 INFO mapreduce.Job: Counters: 30
        File System Counters
```

- Executed MR program filtered out the invalid records i.e. records having NA



```
[acadgild@localhost ~]$ hadoop fs -l /invalidrecordsouput
-l: Unknown command
[acadgild@localhost ~]$ hadoop fs -ls /invalidrecordsouput
18/11/18 17:47:22 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
ls: `/invalidrecordsouput': No such file or directory
[acadgild@localhost ~]$ hadoop fs -ls /invalidrecordsoutput
18/11/18 17:48:48 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-11-18 17:32 /invalidrecordsou
tput/_SUCCESS
-rw-r--r--   1 acadgild supergroup        646 2018-11-18 17:32 /invalidrecordsou
tput/part-m-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

- Output file Content:



```
[acadgild@localhost ~]$ hadoop fs -cat /invalidrecordsoutput/part-m-00000
18/11/18 17:50:44 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry 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
You have new mail in /var/spool/mail/acadgild
```