# POC Table of Contents

# POC #1 Map Reduce 1

Input Dataset: https://nycopendata.socrata.com/Business/2012-NYC-Farmers-Market-List/b7kx-qikm

Given the area and the day we are trying to find how many times the particular market is open.

Driver Code:

```
import org.apache.hadoop.mapreduce.lib.input.KeyValueTextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class NYCFarmersMarketApexMain {

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

        Job job = new Job();

        job.setJarByClass(NYCFarmersMarketApexMain.class);

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

        job.setMapperClass(NYCFarmersMarketMapper.class);
        job.setReducerClass(NYCFarmersMarketReducer.class);

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

        // Reducer related code below.
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

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

    }

}
```

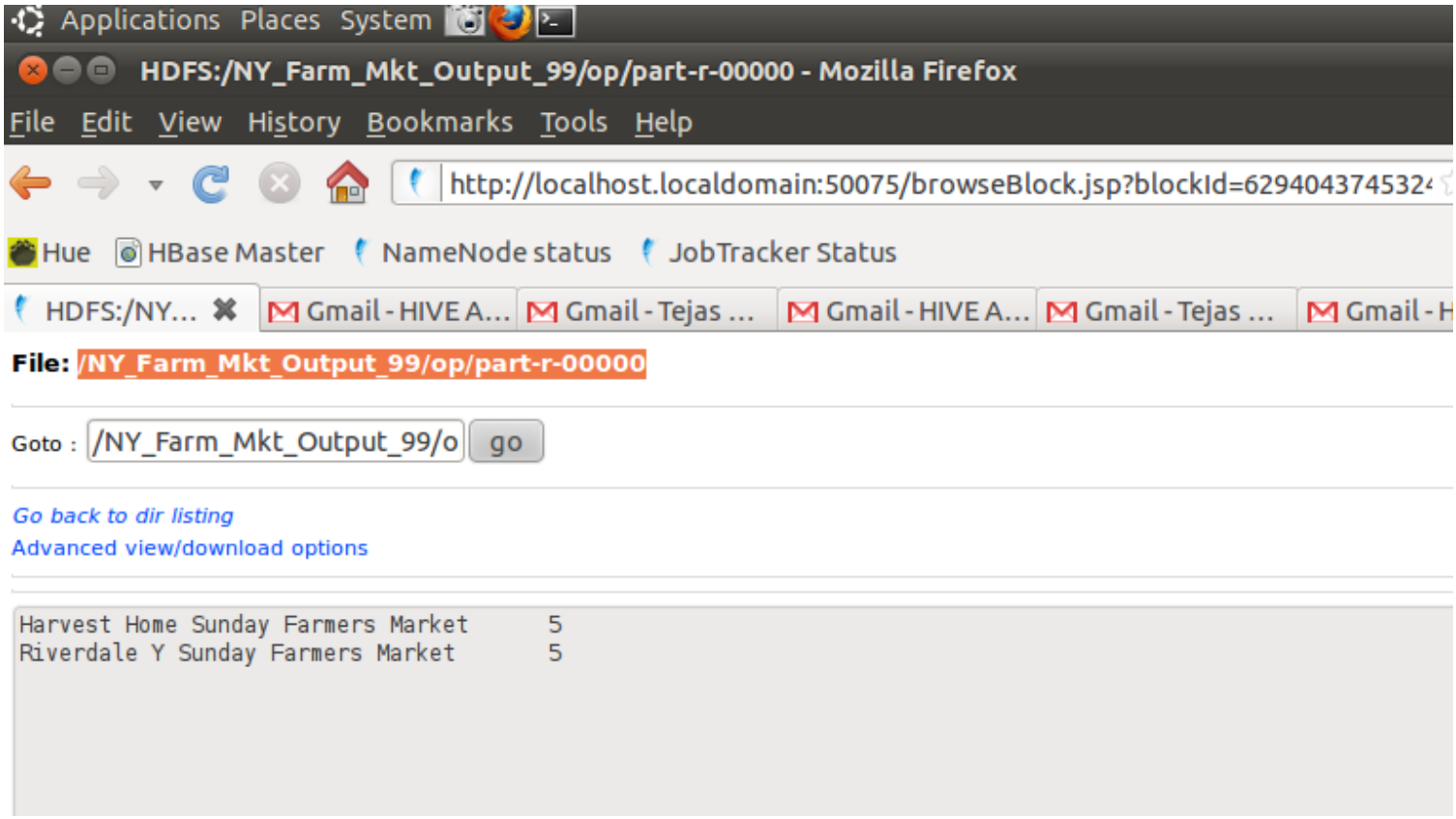
Mapper Code:

```java
public class NYCFarmersMarketMapper extends
        Mapper<LongWritable, Text, Text, IntWritable> {

    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

        Text mapOutPutKey = new Text();
        IntWritable mapOutPutValue = new IntWritable();

        String line = value.toString();
        String lineHeaders[] = line.split(",");// (str.split(","));

        for (int i = 0; i < lineHeaders.length; i++) {
            String area = lineHeaders[0];
            String market = lineHeaders[1];
            String day = lineHeaders[2];
            String EBT = lineHeaders[3];
            String gateNumber = lineHeaders[4];

            if (area.equals("Bronx") && day.equals("Sunday")) {
                context.write(new Text(market), new IntWritable(1));

            }
            // mapOutPutKey = new Text("");
            // mapOutPutValue = new IntWritable(1);
            // context.write(mapOutPutKey, mapOutPutValue);

        }

    }
```

Reducer Code:

```java
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class NYCFarmersMarketReducer extends
        Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    public void reduce(Text keyReducer, Iterable<IntWritable> valueStore,
            Context context) throws IOException, InterruptedException {

        int count = 0;
        int sum = 0;

        while (valueStore.iterator().hasNext()) {

            IntWritable i = valueStore.iterator().next();
            sum = sum + i.get();
            count++;
        }
        context.write(keyReducer, new IntWritable(sum));
    }
}
```

Output:



Ubuntu Output for Map Reduce POC #1

```
cloudera@cloudera-vm:~$ hadoop jar NYCFarmersMarket.jar /NY_Farm_Mkt_IP/ip.txt /NY_Farm_Mkt_Output_99/op
14/07/25 21:22:49 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Too
l for the same.
14/07/25 21:22:50 INFO input.FileInputFormat: Total input paths to process : 1
14/07/25 21:22:50 INFO mapred.JobClient: Running job: job_201407182325_0034
14/07/25 21:22:51 INFO mapred.JobClient:  map 0% reduce 0%
14/07/25 21:22:56 INFO mapred.JobClient:  map 100% reduce 0%
14/07/25 21:23:05 INFO mapred.JobClient:  map 100% reduce 100%
14/07/25 21:23:05 INFO mapred.JobClient: Job complete: job_201407182325_0034
14/07/25 21:23:05 INFO mapred.JobClient: Counters: 22
14/07/25 21:23:05 INFO mapred.JobClient:   Job Counters
14/07/25 21:23:05 INFO mapred.JobClient:     Launched reduce tasks=1
14/07/25 21:23:05 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4173
14/07/25 21:23:05 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
14/07/25 21:23:05 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/07/25 21:23:05 INFO mapred.JobClient:     Launched map tasks=1
14/07/25 21:23:05 INFO mapred.JobClient:     Data-local map tasks=1
14/07/25 21:23:05 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=9479
14/07/25 21:23:05 INFO mapred.JobClient:   FileSystemCounters
14/07/25 21:23:05 INFO mapred.JobClient:     FILE_BYTES_READ=411
14/07/25 21:23:05 INFO mapred.JobClient:     HDFS_BYTES_READ=7710
14/07/25 21:23:05 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=107054
14/07/25 21:23:05 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=73
14/07/25 21:23:05 INFO mapred.JobClient:   Map-Reduce Framework
14/07/25 21:23:05 INFO mapred.JobClient:     Reduce input groups=2
14/07/25 21:23:05 INFO mapred.JobClient:     Combine output records=0
14/07/25 21:23:05 INFO mapred.JobClient:     Map input records=137
14/07/25 21:23:05 INFO mapred.JobClient:     Reduce shuffle bytes=411
14/07/25 21:23:05 INFO mapred.JobClient:     Reduce output records=2
14/07/25 21:23:05 INFO mapred.JobClient:     Spilled Records=20
14/07/25 21:23:05 INFO mapred.JobClient:     Map output bytes=385
14/07/25 21:23:05 INFO mapred.JobClient:     Combine input records=0
14/07/25 21:23:05 INFO mapred.JobClient:     Map output records=10
14/07/25 21:23:05 INFO mapred.JobClient:     SPLIT_RAW_BYTES=103
14/07/25 21:23:05 INFO mapred.JobClient:     Reduce input records=10
cloudera@cloudera-vm:~$
```

# POC #2 Map Reduce 2

Input Dataset: https://data.cityofnewyork.us/Business/DCA-Current-Licensees/spgx-ssye

After processing the Map Reduce job the output id split in 3 different reducers depending on the name of locality.

Mapper and Partitioner Code:

```
import org.apache.hadoop.mapred.TextOutputFormat;

public class NYCLicense {

    public static class NYCLicenseMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
        @Override
        public void map(LongWritable key, Text value,
                OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
            Text mapOutPutKey = new Text();
            IntWritable mapOutPutValue = new IntWritable();
            String line = value.toString();
            String lineHeaders[] = line.split(",");// (str.split(","));
            for (int i = 0; i < lineHeaders.length; i++) {
                String businessName = lineHeaders[0];
                String licenseNumb = lineHeaders[1];
                String building = lineHeaders[2];
                String street = lineHeaders[3];
                String city = lineHeaders[4];
                String zip = lineHeaders[5];                //String county = lineHeaders[6];           //String phone =
                if (city.equals("Bronx") || zip.equals("11209")) {
                    output.collect(new Text(businessName), new IntWritable(1));
                }
            }
        }       // Output types of Mapper should be same as IP arguments of Partitioner
```

```
        public static class NYCLicensePartitioner implements Partitioner<Text, IntWritable> {
            @Override
            public int getPartition(Text key, IntWritable value,int numPartitions) {
                String myKey = key.toString();
                if (myKey.equals("GWNINC")) {
                    return 0; // Reducer Number 1
                }
                if (myKey.equals("RITEAIDOFNEWYORKINC")) {
                    return 1; // Reducer Number 2
                } else {
                    return 2; // Reducer Number 3
                }
            }

            @Override
            public void configure(JobConf arg0) {       // Gives you a new instance of JobConf if you want to change Job Configurations

            }
        }

```

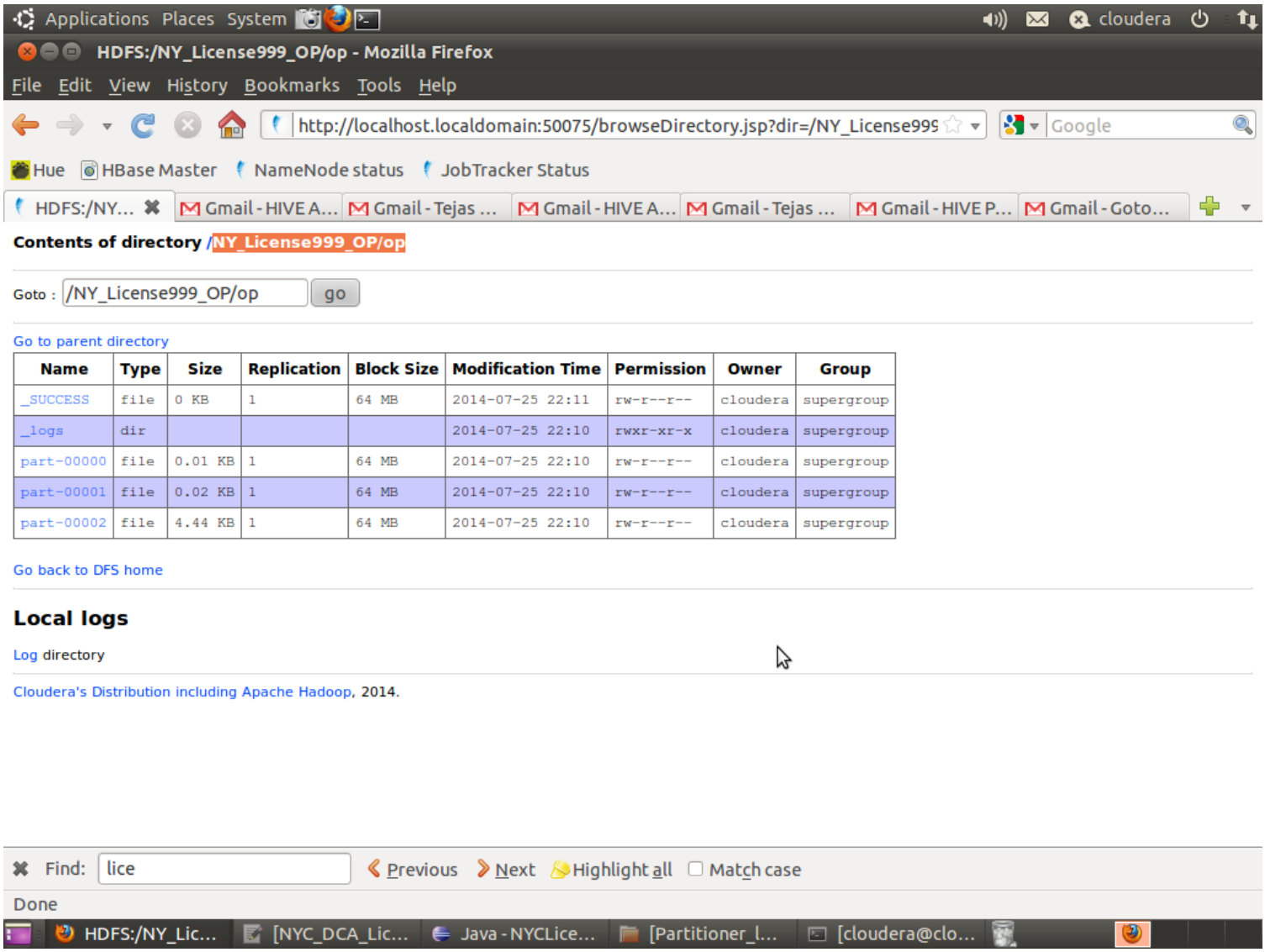
Reducer Code:

```
64          public static class NYCLicenseReducer extends MapReduceBase implements
65                  Reducer<Text, IntWritable, Text, IntWritable> {
66
67              @Override
68              public void reduce(Text key, Iterator<IntWritable> valuesLoop,
69                      OutputCollector<Text, IntWritable> output, Reporter reporter)
70                      throws IOException {
71
72                  int sum = 0;
73                  while (valuesLoop.hasNext()) {
74                      sum += valuesLoop.next().get();
75                      // sum = sum + 1;
76                  }
77
78                  output.collect(key, new IntWritable(sum));
79              }
80          }
81
```

Driver Code:

```
82          public static void main(String[] args) throws IOException {
83
84              JobConf jobConf = new JobConf(NYCLicense.class);
85              jobConf.setJobName("NYCLicense");
86
87              // Forcing program to run 3 reducers
88              jobConf.setNumReduceTasks(3);
89
90              jobConf.setMapperClass(NYCLicenseMapper.class);
91              jobConf.setCombinerClass(NYCLicenseReducer.class);
92              jobConf.setReducerClass(NYCLicenseReducer.class);
93              jobConf.setPartitionerClass(NYCLicensePartitioner.class);
94
95              jobConf.setOutputKeyClass(Text.class);
96              jobConf.setOutputValueClass(IntWritable.class);
97
98              jobConf.setInputFormat(TextInputFormat.class);
99              jobConf.setOutputFormat(TextOutputFormat.class);
100
101             FileInputFormat.setInputPaths(jobConf, new Path(args[0]));
102             FileOutputFormat.setOutputPath(jobConf, new Path(args[1]));
103
104             JobClient.runJob(jobConf);
105
106         }
107     }
108 }
109
```

**part-00000, part-00001 and part-00002**




| Name | Type | Size | Replication | Block Size | Modification Time | Permission | Owner | Group |
|---|---|---|---|---|---|---|---|---|
| _SUCCESS | file | 0 KB | 1 | 64 MB | 2014-07-25 22:11 | rw-r--r-- | cloudera | supergroup |
| _logs | dir | | | | 2014-07-25 22:10 | rwxr-xr-x | cloudera | supergroup |
| part-00000 | file | 0.01 KB | 1 | 64 MB | 2014-07-25 22:10 | rw-r--r-- | cloudera | supergroup |
| part-00001 | file | 0.02 KB | 1 | 64 MB | 2014-07-25 22:10 | rw-r--r-- | cloudera | supergroup |
| part-00002 | file | 4.44 KB | 1 | 64 MB | 2014-07-25 22:10 | rw-r--r-- | cloudera | supergroup |

Map Reduce POC 2 Ubuntu Terminal Output where reducers are 3.

```
Applications Places System                                   cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
-rw-r--r--  1 cloudera cloudera   3381046 2014-07-25 22:06 NYCL.jar
cloudera@cloudera-vm:~$ hadoop jar NYCL.jar  /NY_License88/NYC_DCA_License.txt  /NY_License999_OP/op
14/07/25 22:10:18 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Too
l for the same.
14/07/25 22:10:19 INFO mapred.FileInputFormat: Total input paths to process : 1
14/07/25 22:10:20 INFO mapred.JobClient: Running job: job_201407182325_0037
14/07/25 22:10:21 INFO mapred.JobClient:  map 0% reduce 0%
14/07/25 22:10:33 INFO mapred.JobClient:  map 100% reduce 0%
14/07/25 22:10:49 INFO mapred.JobClient:  map 100% reduce 33%
14/07/25 22:10:50 INFO mapred.JobClient:  map 100% reduce 66%
14/07/25 22:11:01 INFO mapred.JobClient:  map 100% reduce 100%
14/07/25 22:11:01 INFO mapred.JobClient: Job complete: job_201407182325_0037
14/07/25 22:11:01 INFO mapred.JobClient: Counters: 23
14/07/25 22:11:01 INFO mapred.JobClient:   Job Counters
14/07/25 22:11:01 INFO mapred.JobClient:     Launched reduce tasks=3
14/07/25 22:11:01 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=21215
14/07/25 22:11:01 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
14/07/25 22:11:01 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/07/25 22:11:01 INFO mapred.JobClient:     Launched map tasks=2
14/07/25 22:11:01 INFO mapred.JobClient:     Data-local map tasks=2
14/07/25 22:11:01 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=42846
14/07/25 22:11:01 INFO mapred.JobClient:   FileSystemCounters
14/07/25 22:11:01 INFO mapred.JobClient:     FILE_BYTES_READ=5641
14/07/25 22:11:01 INFO mapred.JobClient:     HDFS_BYTES_READ=3912731
14/07/25 22:11:01 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=277661
14/07/25 22:11:01 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=4580
14/07/25 22:11:01 INFO mapred.JobClient:   Map-Reduce Framework
14/07/25 22:11:01 INFO mapred.JobClient:     Reduce input groups=200
14/07/25 22:11:01 INFO mapred.JobClient:     Combine output records=211
14/07/25 22:11:01 INFO mapred.JobClient:     Map input records=58175
14/07/25 22:11:01 INFO mapred.JobClient:     Reduce shuffle bytes=5659
14/07/25 22:11:01 INFO mapred.JobClient:     Reduce output records=200
14/07/25 22:11:01 INFO mapred.JobClient:     Spilled Records=422
14/07/25 22:11:01 INFO mapred.JobClient:     Map output bytes=43898
14/07/25 22:11:01 INFO mapred.JobClient:     Map input bytes=3909275
14/07/25 22:11:01 INFO mapred.JobClient:     Combine input records=1781
14/07/25 22:11:01 INFO mapred.JobClient:     Map output records=1781
14/07/25 22:11:01 INFO mapred.JobClient:     SPLIT_RAW_BYTES=204
14/07/25 22:11:01 INFO mapred.JobClient:     Reduce input records=211
HDFS:/NY_Lic...  [NYC_DCA_Lic...  Java - NYCLice...  [Partitioner_l...  cloudera@clo...
```

# POC #3 Hive 1

Input https://data.cityofnewyork.us/Education/SAT-College-Board-2010-School-Level-Results/zt9s-n5aj

Code in Terminal.

```
Applications Places System                                  cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
"YORK EARLY COLLEGE ACADEMY"
"YOUNG ADULT BOROUGH CENTER AT ART0 AND BU0INE00 H0"
"YOUNG ADULT BOROUGH CENTER AT LOUI0 D. BRANDEI0 H0"
"YOUNG ADULT BOROUGH CNTR CHRI0TOPHER COLUMBU0 H0"
"YOUNG WOMEN'0 LEADER0HIP "
Time taken: 19.476 seconds
hive> DESCRIBE  Hive_SAT_NYCY;
OK
id      string
name    string
numberofstudents        int
readingav       int
mathsav int
writingav       int
Time taken: 0.12 seconds
hive> SELECT id , name ,numberofstudents,readingav,mathsav FROM Hive_SAT_NYCY ORDER BY numberofstudents DESC LIMIT 5;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201407182325_0020, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407182325_0020
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407182325_0020
2014-07-19 05:24:30,464 Stage-1 map = 0%,  reduce = 0%
2014-07-19 05:24:35,608 Stage-1 map = 100%,  reduce = 0%
2014-07-19 05:24:45,763 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201407182325_0020
OK
"13K430"        "BROOKLYN TECHNICAL HIGH "      1277    587     659
"26Q430"        "FRANCI0 LEWI0 HIGH "   934     468     539
"26Q415"        "BENJAMIN N. CARDOZO HIGH "     888     480     545
"02M475"        "0TUYVE0ANT HIGH "      832     679     735
"22K405"        "MIDWOOD HIGH " 824     478     519
Time taken: 20.203 seconds
hive>
[Gmail - Tejas ...  Hive_SAT_Res...  cloudera@clo...  [HIVE_NYC_SAT]  Take Screenshot
```

```
Applications Places System                                  cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
    >
    > INSERT OVERWRITE DIRECTORY '/Hive_SAT_NYC_OrderBy' SELECT id , name ,numberofstudents,readingav,mathsav FROM Hive_SAT
 NYCY ORDER BY numberofstudents DESC LIMIT 4;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201407182325_0021, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407182325_0021
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407182325_0021
2014-07-19 05:45:45,243 Stage-1 map = 0%,  reduce = 0%
2014-07-19 05:45:50,300 Stage-1 map = 50%,  reduce = 0%
2014-07-19 05:45:51,320 Stage-1 map = 100%,  reduce = 0%
2014-07-19 05:46:02,498 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201407182325_0021
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201407182325_0022, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407182325_0022
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407182325_0022
2014-07-19 05:46:07,080 Stage-2 map = 0%,  reduce = 0%
2014-07-19 05:46:10,189 Stage-2 map = 100%,  reduce = 0%
2014-07-19 05:46:19,455 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201407182325_0022
Moving data to: /Hive_SAT_NYC_OrderBy
4 Rows loaded to /Hive_SAT_NYC_OrderBy
OK
Time taken: 39.99 seconds
hive>
Gmail - Compo...  [ParkingLot.tx...  cloudera@clo...  [HIVE_NYC_SAT]  HIVE_NYC_SAT
```

Data loaded in HDFS in 40 seconds.

--

Map Reduce Job Details:

Applications Places System cloudera

Hadoop job_201407182325_0020 on localhost - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost:50030/jobdetails.jsp?jobid=job_201407182325_0020

Hue HBase Master NameNode status JobTracker Status

HDFS:/ Pig Latin ... Gmail - HI... Gmail - HI... Gmail - T... Gmail - HI... Gmail - T... Gmail - HI... Hado...

# Hadoop job_201407182325_0020 on localhost

**User:** root
**Job Name:** SELECT id , name ,numberofstudents,readi...5(Stage-1)
**Job File:** hdfs://localhost/var/lib/hadoop-0.20/cache/mapred/mapred/staging/root/.staging/job_201407182325_0020/job.xml
**Submit Host:** cloudera-vm
**Submit Host Address:** 127.0.1.1
**Job-ACLs: All users are allowed**
**Job Setup:** Successful
**Status:** Succeeded
**Started at:** Sat Jul 19 05:24:27 PDT 2014
**Finished at:** Sat Jul 19 05:24:46 PDT 2014
**Finished in:** 18sec
**Job Cleanup:** Successful

| Kind | % Complete | Num Tasks | Pending | Running | Complete | Killed | Failed/Killed Task Attempts |
|---|---|---|---|---|---|---|---|
| map | 100.00% | 2 | 0 | 0 | 2 | 0 | 0 / 0 |
| reduce | 100.00% | 1 | 0 | 0 | 1 | 0 | 0 / 0 |

| | Counter | Map | Reduce | Total |
|---|---|---|---|---|
| org.apache.hadoop.hive.ql.exec.Operator$ProgressCounter | CREATED_FILES | 0 | 1 | 1 |
| | SLOTS_MILLIS_MAPS | 0 | 0 | 13,396 |
| | Launched reduce tasks | 0 | 0 | 1 |

Done

Hadoop job_2... [ParkingLot.tx... cloudera@clo... [HIVE_NYC_SAT] [HIVE_NYC_SAT]

Applications Places System cloudera

Hadoop job_201407182325_0021 on localhost - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost:50030/jobdetails.jsp?jobid=job_201407182325_0021

Hue HBase Master NameNode status JobTracker Status

# Hadoop job_201407182325_0021 on localhost

**User:** root
**Job Name:** INSERT OVERWRITE DIRECTORY '/Hive_SAT_NY...4(Stage-1)
**Job File:** hdfs://localhost/var/lib/hadoop-0.20/cache/mapred/mapred/staging/root/.staging/job_201407182325_0021/job.xml
**Submit Host:** cloudera-vm
**Submit Host Address:** 127.0.1.1
**Job-ACLs: All users are allowed**
**Job Setup:** Successful
**Status:** Succeeded
**Started at:** Sat Jul 19 05:45:42 PDT 2014
**Finished at:** Sat Jul 19 05:46:03 PDT 2014
**Finished in:** 21sec
**Job Cleanup:** Successful

| Kind | % Complete | Num Tasks | Pending | Running | Complete | Killed | Failed/Killed Task Attempts |
|---|---|---|---|---|---|---|---|
| map | 100.00% | 2 | 0 | 0 | 2 | 0 | 0 / 0 |
| reduce | 100.00% | 1 | 0 | 0 | 1 | 0 | 0 / 0 |

| | Counter | Map | Reduce | Total |
|---|---|---|---|---|
| org.apache.hadoop.hive.ql.exec.Operator$ProgressCounter | CREATED_FILES | 0 | 1 | 1 |
| | SLOTS_MILLIS_MAPS | 0 | 0 | 13,987 |
| | Launched reduce tasks | 0 | 0 | 1 |

Done

# POC #4 PIG 1

Input Dataset: https://data.cityofnewyork.us/Business/NYC-Jobs/kpav-sd4t

Code Base 1:

```
New York City Jobs SCHEMA for PIG:

Agency:chararray  Count:int  Title:chararray  Level:chararray  Salary:int  Frequency:chararray  Location:chararray
--
grunt> A_NYjobs = LOAD '/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt' using PigStorage('\t') as
(agency:chararray,count:int,title:)chararray,level:chararray,salary:int,frequency:chararray,location:chararray);
grunt> DESCRIBE A_NYjobs;
A_NYjobs: {agency: chararray,count: int,title: chararray,level: chararray,salary: int,frequency: chararray,location: chararray}
grunt> grp_jobs_By_Agency = GROUP A_NYjobs by agency;
grunt> DESCRIBE grp_jobs_By_Agency;
grp_jobs_By_Agency: {group: chararray,A_NYjobs: {agency: chararray,count: int,title: chararray,level: chararray,salary: int,frequency: chararray,location:
chararray}}
grunt>
--

grp_jobs_By_Title

grunt> DESCRIBE A_NYjobs;
A_NYjobs: {agency: chararray,count: int,title: chararray,level: chararray,salary: int,frequency: chararray,location: chararray}
grunt> grp_jobs_By_Title = GROUP A_NYjobs by title;
grunt> DESCRIBE grp_jobs_By_Title;
grp_jobs_By_Title: {group: chararray,A_NYjobs: {agency: chararray,count: int,title: chararray,level: chararray,salary: int,frequency: chararray,location:
chararray}}
grunt> cloudera@cloudera-vm:~$
```

Code Base 2:

```

IN PIG Group By and Filter are like SELECT of My SQL

grunt> grp_NYCJOBS_location = GROUP A_NYC_jobs BY location;
grunt> A_NYC_jobs = LOAD '/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt' using PigStorage('\t') as
(agency:chararray,count:int,title                                                DESCRIBE A_NYC_jobs;
A_NYC_jobs: {agency: chararray,count: int,title: chararray,level: chararray,salary: int,frequency: chararray,location: chararray}
grunt> filter_NYC_jobs_freq = FILTER A_NYC_jobs BY frequency=='Hourly';
grunt> DESCRIBE filter_NYC_jobs_freq;

--

grp_filter_NYC_jobs_freq_BY_level = GROUP filter_NYC_jobs_freq by level;
--
grunt> result_Loop_NYCjobs = FOREACH grp_filter_NYC_jobs_freq_BY_level GENERATE  SUM(filter_NYC_jobs_freq.count);
grunt> DESCRIBE result_Loop_NYCjobs;
result_Loop_NYCjobs: {long}

-
STORE result_Loop_NYCjobs INTO '/NYC_JOBS_Loop_SUM_filter_count';
```

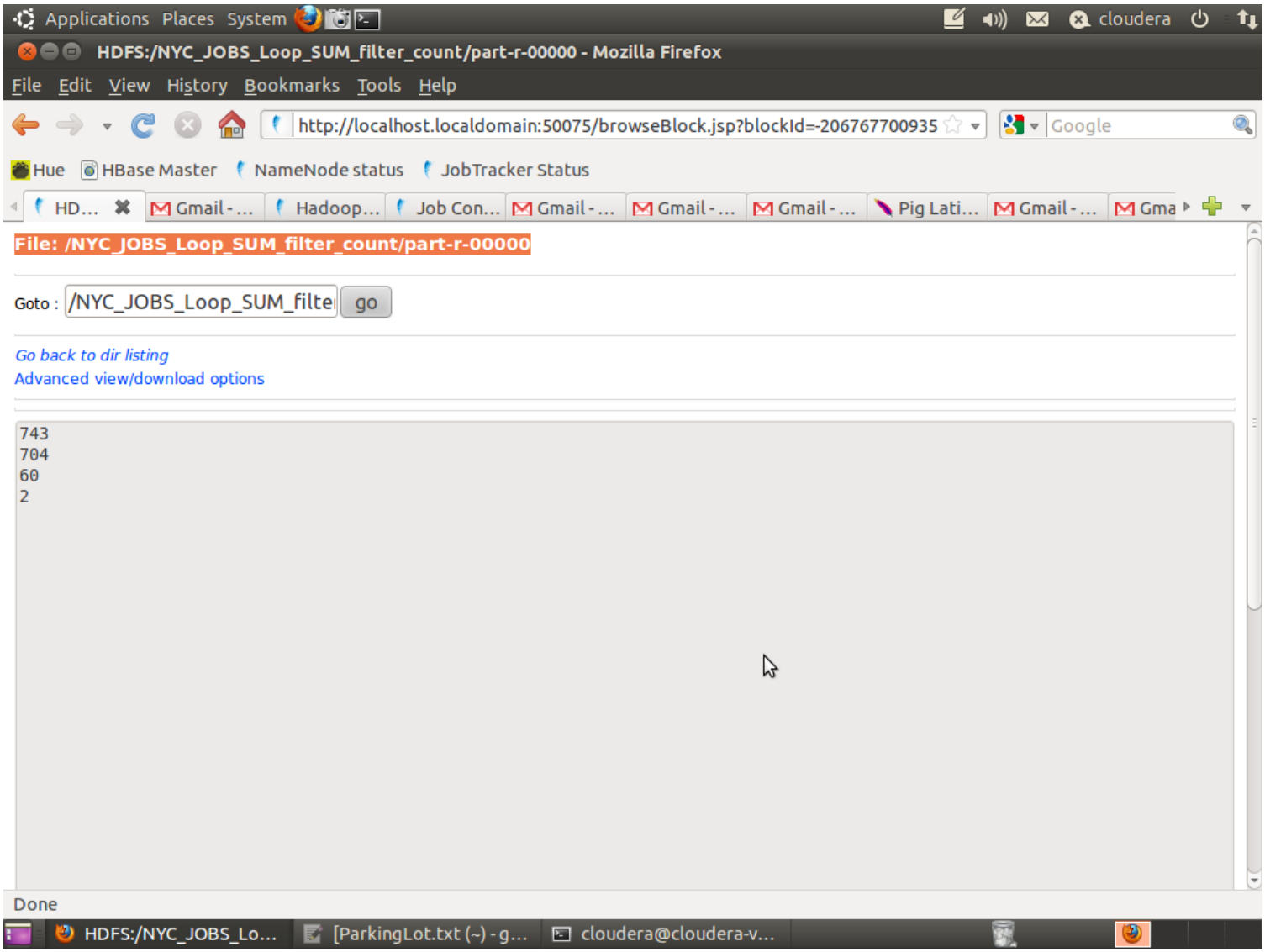
Success message of Query in Grunt Shell

```
ormation at: http://localhost:50030/jobdetails.jsp?jobid=job_201407031019_0073
2014-07-04 19:28:16,417 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% compl
ete
2014-07-04 19:28:23,077 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% comp
lete
2014-07-04 19:28:41,823 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% com
plete
2014-07-04 19:28:41,823 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-04 19:28:12     2014-07-04 19:28:41     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   AliasF
eature  Outputs
job_201407031019_0073   1       1       3       3       3       11      11      11      A_NYC_jobs,filter_NYC_jobs_freq,grp_fi
lter_NYC_jobs_freq_BY_level,result_Loop_NYCjobs GROUP_BY,COMBINER       /NYC_JOBS_Loop_SUM_filter_count,

Input(s):
Successfully read 1795 records (188680 bytes) from: "/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt"

Output(s):
Successfully stored 4 records (13 bytes) in: "/NYC_JOBS_Loop_SUM_filter_count"

Counters:
Total records written : 4
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407031019_0073


2014-07-04 19:28:41,842 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

```
2014-07-04 18:23:54,820 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More
 information at: http://localhost:50030/jobdetails.jsp?jobid=job_201407031019_0070
2014-07-04 18:24:04,219 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50%
complete
2014-07-04 18:24:20,013 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100%
 complete
2014-07-04 18:24:20,015 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-04 18:23:48     2014-07-04 18:24:20     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   Al
ias     Feature Outputs
job_201407031019_0070   1       1       4       4       4       12      12      12      A_NYjobs,grp_jobs_By_Agency     GR
OUP_BY  /NYCITY_JOBS_OP_PIG,

Input(s):
Successfully read 1795 records (188680 bytes) from: "/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt"

Output(s):
Successfully stored 39 records (191089 bytes) in: "/NYCITY_JOBS_OP_PIG"

Counters:
Total records written : 39
Total bytes written : 191089
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407031019_0070


2014-07-04 18:24:20,030 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
grunt>
```

```
2014-07-04 18:48:11,618 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More
 information at: http://localhost:50030/jobdetails.jsp?jobid=job_201407031019_0071
2014-07-04 18:48:17,547 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50%
complete
2014-07-04 18:48:36,202 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100%
 complete
2014-07-04 18:48:36,206 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-04 18:48:05     2014-07-04 18:48:36     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   Al
ias     Feature Outputs
job_201407031019_0071   1       1       3       3       3       12      12      12      A_NYC_jobs,grp_NYCJOBS_location GR
OUP_BY  /NYC_JOBS_LOCATION,

Input(s):
Successfully read 1795 records (188680 bytes) from: "/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt"

Output(s):
Successfully stored 138 records (193708 bytes) in: "/NYC_JOBS_LOCATION"

Counters:
Total records written : 138
Total bytes written : 193708
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407031019_0071


2014-07-04 18:48:36,224 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
grunt>
```

```
Applications Places System                                  cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
 information at: http://localhost:50030/jobdetails.jsp?jobid=job_201407031019_0072
2014-07-04 19:11:55,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50%
complete
2014-07-04 19:12:10,628 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100%
 complete
2014-07-04 19:12:10,629 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-04 19:11:40     2014-07-04 19:12:10     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   Al
ias     Feature Outputs
job_201407031019_0072   1       1       6       6       6       12      12      12      A_NYC_jobs,filter_NYC_jobs_freq,gr
p_filter_NYC_jobs_freq_BY_level GROUP_BY        /NYCjobs_OP_FREQ_Level,

Input(s):
Successfully read 1795 records (188680 bytes) from: "/NYCITY_JOBS_IP/NYC_Jobs_Tabbed.txt"

Output(s):
Successfully stored 4 records (16222 bytes) in: "/NYCjobs_OP_FREQ_Level"

Counters:
Total records written : 4
Total bytes written : 16222
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407031019_0072


2014-07-04 19:12:10,660 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
grunt>
Gmail - PIG Ha...  [ParkingLot.tx...  cloudera@clo...  newyorkcityjo...  [NYC_Jobs_CV...
```


Applications Places System
cloudera
HDFS:/NYC_JOBS_Loop_SUM_filter_count/part-r-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075/browseBlock.jsp?blockId=-206767700935
Google
Hue
HBase Master
NameNode status
JobTracker Status
HD...
Gmail - ...
Hadoop...
Job Con...
Gmail - ...
Gmail - ...
Gmail - ...
Pig Lati...
Gmail - ...
Gma
File: /NYC_JOBS_Loop_SUM_filter_count/part-r-00000
Goto :
/NYC_JOBS_Loop_SUM_filte
go
Go back to dir listing
Advanced view/download options
743
704
60
2
Done
HDFS:/NYC_JOBS_Lo...
[ParkingLot.txt (~) - g...
cloudera@cloudera-v...

# POC #5 PIG 2

Input Data: https://edureka.wistia.com/medias/jdwwsfa91j 80 MB with 10 lakh records and renamed to Sports_Authority text file.

PIG query to find unique store locations with the Ubuntu terminal output.

```
Applications Places System                                   cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
2014-07-19 00:33:36,277 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
- 47% complete
2014-07-19 00:33:39,374 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
- 49% complete
2014-07-19 00:34:14,773 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
- 100% complete
2014-07-19 00:34:14,774 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-19 00:32:25     2014-07-19 00:34:14     DISTINCT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTim
e       Alias   Feature Outputs
job_201407182325_0004   2       1       63      50      56      41      41      41      A_Sport_Auth77,LOOP_Sport_Au
th77    DISTINCT        /Sports_Authority_Unique_City_Locations,

Input(s):
Successfully read 1000000 records (88333615 bytes) from: "/Sports_Authority_IP/sa_txns.txt"

Output(s):
Successfully stored 108 records (1053 bytes) in: "/Sports_Authority_Unique_City_Locations"

Counters:
Total records written : 108
Total bytes written : 1053
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407182325_0004


2014-07-19 00:34:14,780 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
- Success!
grunt>
cloudera@cloudera-... Pig Cookbook - Moz... [cloudera] [ParkingLot.txt (~) - ...
```

Limit query to 500 only and store in HDFS

```
Applications Places System                                  cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
lete
2014-07-19 00:50:21,277 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% com
plete
2014-07-19 00:50:21,278 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-19 00:47:52     2014-07-19 00:50:21     ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   AliasF
eature  Outputs
job_201407182325_0005   2       0       36      18      27      0       0       0       A_Sport_Auth77  MAP_ONLY
job_201407182325_0006   2       1       14      12      13      21      21      21      Sport_Auth_By_scor      SAMPLER
job_201407182325_0007   2       1       36      27      31      17      17      17      Sport_Auth_By_scor      ORDER_BY,COMBI
NER     /Sports_Authority_Score_500_records_Only,

Input(s):
Successfully read 1000000 records (88333615 bytes) from: "/Sports_Authority_IP/sa_txns.txt"

Output(s):
Successfully stored 500 records (42748 bytes) in: "/Sports_Authority_Score_500_records_Only"

Counters:
Total records written : 500
Total bytes written : 42748
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407182325_0005   ->      job_201407182325_0006,
job_201407182325_0006   ->      job_201407182325_0007,
job_201407182325_0007


2014-07-19 00:50:21,297 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
cloudera@clo...  [Pig Cookboo...  [cloudera]  [ParkingLot.tx...  [Sports_Autho...
```

Aggregate Query in PIG terminal output

```
Applications Places System                                  cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
lete
2014-07-19 01:27:09,787 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 55% comp
lete
2014-07-19 01:27:12,804 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 58% comp
lete
2014-07-19 01:27:28,274 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% com
plete
2014-07-19 01:27:28,275 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-19 01:26:23     2014-07-19 01:27:28     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   AliasF
eature  Outputs
job_201407182325_0009   2       1       42      21      32      29      29      29      A_Sport_Auth77,Sport_Auth_Aggregate,Sp
ort_Auth_Count  GROUP_BY,COMBINER       /Sport_Auth_Aggregate,

Input(s):
Successfully read 1000000 records (88333615 bytes) from: "/Sports_Authority_IP/sa_txns.txt"

Output(s):
Successfully stored 125 records (2269 bytes) in: "/Sport_Auth_Aggregate"

Counters:
Total records written : 125
Total bytes written : 2269
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407182325_0009


2014-07-19 01:27:28,280 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
cloudera@clo...  Pig UDF Manu...  [cloudera]  [ParkingLot.tx...  [Sports_Autho...
```

**PIG was used for partitioning and 5 reducers were obtained and output was stored in HDFS**

```
Applications Places System                                   cloudera
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
lete
2014-07-19 01:05:15,564 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 89% comp
lete
2014-07-19 01:05:32,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 98% comp
lete
2014-07-19 01:05:44,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% com
plete
2014-07-19 01:05:44,104 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
0.20.2-cdh3u0   0.8.0-cdh3u0    cloudera        2014-07-19 01:03:19     2014-07-19 01:05:44     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinReduceTime   AvgReduceTime   AliasF
eature  Outputs
job_201407182325_0008   2       5       64      40      52      49      21      33      A_Sport_Auth77,Sport_Auth_parallel_red
ucer5   GROUP_BY        /Sport_Auth_Five_Parallel_reducer,

Input(s):
Successfully read 1000000 records (88333615 bytes) from: "/Sports_Authority_IP/sa_txns.txt"

Output(s):
Successfully stored 15 records (89715627 bytes) in: "/Sport_Auth_Five_Parallel_reducer"

Counters:
Total records written : 15
Total bytes written : 89715627
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 5
Total records proactively spilled: 270273

Job DAG:
job_201407182325_0008


2014-07-19 01:05:44,117 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
cloudera@clo...  Pig Cookbook ...  [cloudera]  [ParkingLot.tx...  [Sports_Autho...
```

Applications Places System cloudera

HDFS:/Sport_Auth_Five_Parallel_reducer - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseDirectory.jsp?dir=/Sport_Auth_Fi

ILLUSTARate describe

Hue HBase Master NameNode status JobTracker Status

HDFS:/Sport_Auth_Five_P... Pig Cookbook Testing and Diagnostics

**Contents of directory /Sport_Auth_Five_Parallel_reducer**

Goto : /Sport_Auth_Five_Parallel_re go

Go to parent directory

| Name | Type | Size | Replication | Block Size | Modification Time | Permission | Owner | Group |
|---|---|---|---|---|---|---|---|---|
| _logs | dir | | | | 2014-07-19 01:03 | rwxr-xr-x | cloudera | supergroup |
| part-r-00000 | file | 19.21 MB | 1 | 64 MB | 2014-07-19 01:04 | rw-r--r-- | cloudera | supergroup |
| part-r-00001 | file | 10.14 MB | 1 | 64 MB | 2014-07-19 01:04 | rw-r--r-- | cloudera | supergroup |
| part-r-00002 | file | 23.99 MB | 1 | 64 MB | 2014-07-19 01:05 | rw-r--r-- | cloudera | supergroup |
| part-r-00003 | file | 11.53 MB | 1 | 64 MB | 2014-07-19 01:05 | rw-r--r-- | cloudera | supergroup |
| part-r-00004 | file | 20.69 MB | 1 | 64 MB | 2014-07-19 01:05 | rw-r--r-- | cloudera | supergroup |

Done

[cloudera@clo... HDFS:/Sport_... [cloudera] [ParkingLot.tx... [Sports_Autho...

Applications Places System cloudera

HDFS:/Sport_Auth_Five_Parallel_reducer/part-r-00000 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseBlock.jsp?blockId=243512252578 ILLUSTARate describe

Hue HBase Master NameNode status JobTracker Status

HDFS:/Sport_Auth_Five_P... Pig Cookbook Testing and Diagnostics

**File: /Sport_Auth_Five_Parallel_reducer/part-r-00000**

Goto : /Sport_Auth_Five_Parallel_re go

*Go back to dir listing*

Advanced view/download options

View Next chunk

```
Puzzles {(00861711,04-07-2011,4003414,141.88,Puzzles,Jigsaw Puzzles,Coral Springs,Florida,credit),
(00883199,07-19-2011,4005646,54.57,Puzzles,Jigsaw Puzzles,Chattanooga,Tennessee,credit),
(00862784,01-18-2011,4005644,91.75,Puzzles,Jigsaw Puzzles,Sunnyvale,California,credit),
(00829168,11-03-2011,4003543,103.11,Puzzles,Mechanical Puzzles,Orange,California,credit),
(00979420,02-22-2011,4004352,59.92,Puzzles,Mechanical Puzzles,Portland,Oregon,credit),
(00844613,04-01-2011,4001722,66.54,Puzzles,Mechanical Puzzles,Lowell,Massachusetts,credit),
(00788667,02-20-2011,4001477,184.25,Puzzles,Jigsaw Puzzles,San Jose,California,credit),
(00842145,05-11-2011,4006837,100.53,Puzzles,Jigsaw Puzzles,Pasadena,California,credit),
(00809268,07-01-2011,4004253,180.08,Puzzles,Mechanical Puzzles,Scottsdale,Arizona,credit),
(00895281,08-12-2011,4004136,82.68,Puzzles,Mechanical Puzzles,New Orleans,Louisiana,credit),
(00947649,01-08-2011,4003264,99.37,Puzzles,Jigsaw Puzzles,Columbia,Missouri,credit),
(00788658,11-30-2011,4007851,182.0,Puzzles,Jigsaw Puzzles,Milwaukee,Wisconsin,credit),
(00976075,05-24-2011,4006182,104.98,Puzzles,Jigsaw Puzzles,Jackson,Mississippi,credit),
(00944649,08-23-2011,4005259,191.19,Puzzles,Mechanical Puzzles,Columbia,South Carolina,credit),
(00904954,04-25-2011,4000122,170.26,Puzzles,Jigsaw Puzzles,Cambridge,Massachusetts,credit),
(00780008,04-12-2011,4009868,178.16,Puzzles,Jigsaw Puzzles,Chattanooga,Tennessee,credit),
(00866343,12-02-2011,4008640,5.39,Puzzles,Jigsaw Puzzles,Bellevue,Washington,cash),
(00844595,11-07-2011,4003296,45.21,Puzzles,Jigsaw Puzzles,Everett,Washington,cash),
(00957508,06-07-2011,4003496,132.97,Puzzles,Mechanical Puzzles,Paterson,New Jersey,credit),
(00834987,09-30-2011,4001473,185.01,Puzzles,Jigsaw Puzzles,Milwaukee,Wisconsin,credit),
(00932970,07-22-2011,4009088,145.1,Puzzles,Jigsaw Puzzles,Durham,North Carolina,credit),
```

Done

[cloudera@clo... HDFS:/Sport_... [cloudera] [ParkingLot.tx... [Sports_Autho...

Applications Places System cloudera

HDFS:/Sport_Auth_Five_Parallel_reducer/part-r-00001 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseBlock.jsp?blockId=-170249775809 ILLUSTARate describe

Hue HBase Master NameNode status JobTracker Status

HDFS:/Sport_Auth_Five_P... Pig Cookbook Testing and Diagnostics

**File: /Sport_Auth_Five_Parallel_reducer/part-r-00001**

Goto : /Sport_Auth_Five_Parallel_re go

*Go back to dir listing*
Advanced view/download options

View Next chunk

```
Jumping {(00922028,09-08-2011,4009551,198.42,Jumping,Trampolines,Coral Springs,Florida,credit),
(00937540,04-21-2011,4008999,75.26,Jumping,Pogo Sticks,St. Petersburg,Florida,credit),
(00812477,03-26-2011,4001154,33.9,Jumping,Pogo Sticks,Sacramento,California,cash),
(00934052,12-15-2011,4002809,11.34,Jumping,Trampolines,Hampton  ,Virginia,cash),
(00922021,05-21-2011,4002070,79.63,Jumping,Pogo Sticks,Stamford,Connecticut,credit),
(00782450,01-27-2011,4004865,128.8,Jumping,Trampolines,Indianapolis ,Indiana,credit),
(00987639,06-10-2011,4001068,193.51,Jumping,Jumping Stilts,Sunnyvale,California,credit),
(00833878,03-18-2011,4001800,195.15,Jumping,Pogo Sticks,Charleston,South Carolina,credit),
(00872594,03-09-2011,4000047,65.33,Jumping,Bungee Jumping,Montgomery,Alabama,credit),
(00771434,01-08-2011,4000262,163.04,Jumping,Bungee Jumping,Gilbert,Arizona,credit),
(00793862,03-21-2011,4009855,148.11,Jumping,Trampoline Accessories,St. Louis  ,Missouri,credit),
(00922009,04-14-2011,4002606,170.86,Jumping,Trampolines,Everett,Washington,credit),
(00851125,03-06-2011,4005925,161.02,Jumping,Trampolines,Clarksville,Tennessee,credit),
(00845314,02-25-2011,4006307,61.27,Jumping,Trampolines,Austin,Texas,credit),
(00872598,07-10-2011,4005269,142.18,Jumping,Trampolines,San Francisco,California,credit),
(00921990,06-13-2011,4007282,75.85,Jumping,Trampoline Accessories,Hampton  ,Virginia,credit),
(00825613,10-24-2011,4007899,125.16,Jumping,Trampoline Accessories,Jackson,Mississippi,credit),
(00921983,06-29-2011,4003029,173.17,Jumping,Trampoline Accessories,Sacramento,California,credit),
(00959784,07-27-2011,4000342,52.6,Jumping,Jumping Stilts,Orange,California,credit),
(00995188,12-05-2011,4002359,72.9,Jumping,Bungee Jumping,Boise,Idaho,credit),
(00903062,09-14-2011,4004544,119.63,Jumping,Jumping Stilts,Westminster,Colorado,credit),
```

Done

[cloudera@clo... HDFS:/Sport_... [cloudera] [ParkingLot.tx... [Sports_Autho...

Applications Places System cloudera

HDFS:/Sport_Auth_Five_Parallel_reducer/part-r-00002 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseBlock.jsp?blockId=-506039318520 ILLUSTARate describe

Hue HBase Master NameNode status JobTracker Status

HDFS:/Sport_Auth_Five_P... Pig Cookbook Testing and Diagnostics

**File: /Sport_Auth_Five_Parallel_reducer/part-r-00002**

Goto : /Sport_Auth_Five_Parallel_re go

*Go back to dir listing*
Advanced view/download options

View Next chunk

```
Dancing {(00852474,02-06-2011,4000960,153.68,Dancing,Ballet Bars,El Paso,Texas,credit),
(00796723,07-23-2011,4006235,59.48,Dancing,Ballet Bars,Chicago,Illinois,credit),
(00801307,09-18-2011,4001666,148.98,Dancing,Ballet Bars,Centennial,Colorado,credit),
(00919676,11-30-2011,4003676,85.09,Dancing,Ballet Bars,Oklahoma City,Oklahoma,credit),
(00942137,08-09-2011,4000396,181.33,Dancing,Ballet Bars,Midland,Texas,credit),
(00853484,08-08-2011,4008980,46.67,Dancing,Ballet Bars,Portland,Oregon,cash),
(00992963,09-30-2011,4009157,187.84,Dancing,Ballet Bars,Boston,Massachusetts,credit),
(00814775,03-21-2011,4001166,174.02,Dancing,Ballet Bars,Denver  ,Colorado,credit),
(00902207,02-25-2011,4002667,90.54,Dancing,Ballet Bars,Jacksonville ,Florida,credit),
(00862506,07-11-2011,4008656,38.57,Dancing,Ballet Bars,Cincinnati,Ohio,credit),
(00787075,05-24-2011,4008359,135.48,Dancing,Ballet Bars,San Antonio,Texas,credit),
(00806511,04-21-2011,4008474,157.16,Dancing,Ballet Bars,St. Louis  ,Missouri,credit),
(00839255,01-26-2011,4006419,137.6,Dancing,Ballet Bars,Jacksonville ,Florida,credit),
(00916530,08-20-2011,4006857,149.36,Dancing,Ballet Bars,Pasadena,Texas,credit),
(00833917,11-01-2011,4000135,174.58,Dancing,Ballet Bars,New Orleans,Louisiana,credit),
(00946028,10-23-2011,4001894,44.52,Dancing,Ballet Bars,Vancouver,Washington,credit),
(00866044,07-02-2011,4001823,165.73,Dancing,Ballet Bars,Atlanta,Georgia,credit),
(00953543,07-01-2011,4000049,42.4,Dancing,Ballet Bars,Charleston,South Carolina,cash),
(00791528,11-20-2011,4003289,48.64,Dancing,Ballet Bars,West Valley City,Utah,credit),
(00838628,02-15-2011,4007355,87.94,Dancing,Ballet Bars,Baltimore,Maryland,credit),
(00762214,09-16-2011,4008557,43.53,Dancing,Ballet Bars,Phoenix,Arizona,credit),
```

Done

[cloudera@clo... HDFS:/Sport_... [cloudera] [ParkingLot.tx... [Sports_Autho...

Applications Places System cloudera

HDFS:/Sport_Auth_Five_Parallel_reducer/part-r-00003 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseBlock.jsp?blockId=407872684443 ILLUSTARate describe

Hue HBase Master NameNode status JobTracker Status

HDFS:/Sport_Auth_Five_P... Pig Cookbook Testing and Diagnostics

**File: /Sport_Auth_Five_Parallel_reducer/part-r-00003**

Goto : /Sport_Auth_Five_Parallel_re go

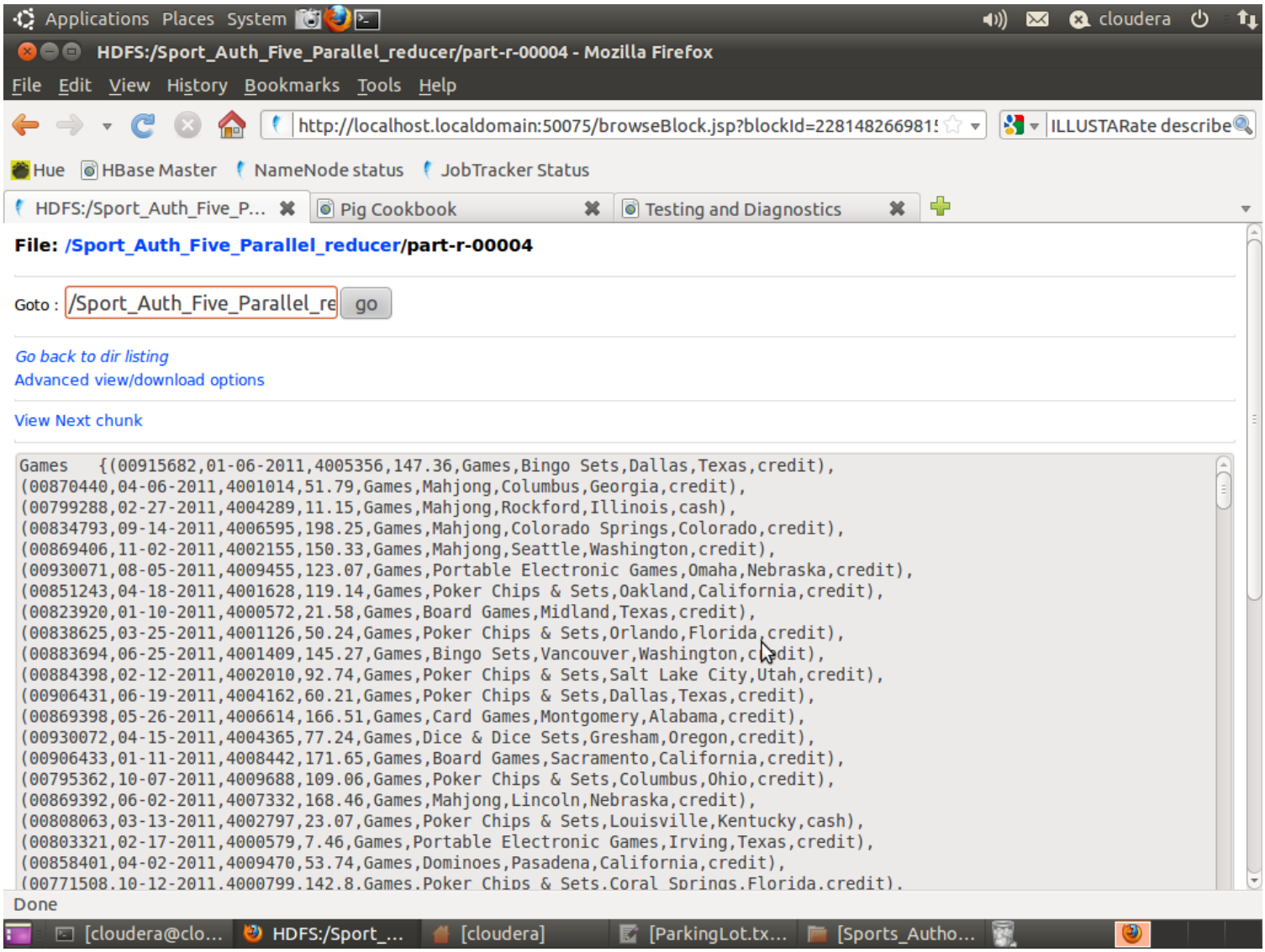
*Go back to dir listing*
Advanced view/download options

View Next chunk

```
Water Sports    {(00817664,02-01-2011,4006676,155.79,Water Sports,Wetsuits,Centennial,Colorado,credit),
(00918996,11-16-2011,4007074,133.79,Water Sports,Wetsuits,St. Louis  ,Missouri,credit),
(00995067,05-17-2011,4009482,123.43,Water Sports,Water Polo,Phoenix,Arizona,credit),
(00998608,07-06-2011,4007965,48.59,Water Sports,Bodyboarding,Charleston,South Carolina,credit),
(00817661,05-19-2011,4000454,16.3,Water Sports,Towed Water Sports,Minneapolis,Minnesota,cash),
(00884983,03-12-2011,4008757,67.46,Water Sports,Boating,Buffalo,New York,credit),
(00959489,03-21-2011,4004702,136.99,Water Sports,Swimming,Stamford,Connecticut,credit),
(00931468,04-06-2011,4000008,126.79,Water Sports,Bodyboarding,Indianapolis ,Indiana,credit),
(00959493,05-11-2011,4005977,82.05,Water Sports,Surfing,Jackson,Mississippi,credit),
(00861319,05-17-2011,4004463,121.36,Water Sports,Water Polo,Orlando,Florida,credit),
(00919002,12-21-2011,4001124,186.19,Water Sports,Scuba Diving & Snorkeling,Westminster,Colorado,credit),
(00828009,09-20-2011,4009335,49.97,Water Sports,Scuba Diving & Snorkeling,Buffalo,New York,credit),
(00959503,10-27-2011,4009535,92.24,Water Sports,Boating,Bellevue,Washington,credit),
(00794111,12-05-2011,4000135,44.04,Water Sports,Towed Water Sports,Memphis,Tennessee,credit),
(00959507,06-20-2011,4000269,139.68,Water Sports,Windsurfing,Chicago,Illinois,credit),
(00835670,07-20-2011,4007100,40.94,Water Sports,Water Polo,Clarksville,Tennessee,cash),
(00959512,04-05-2011,4004028,188.69,Water Sports,Boating,Madison,Wisconsin,credit),
(00767012,10-22-2011,4002879,73.76,Water Sports,Water Polo,Los Angeles,California,credit),
(00959517,05-03-2011,4004362,29.23,Water Sports,Windsurfing,Newark,New Jersey,credit),
(00828005,07-19-2011,4002043,120.65,Water Sports,Swimming,Salt Lake City,Utah,credit),
(00817646,03-07-2011,4000430,68.91,Water Sports,Windsurfing,Houston,Texas,credit),
```

Done

[cloudera@clo... HDFS:/Sport_... [cloudera] [ParkingLot.tx... [Sports_Autho...


Applications Places System
cloudera
HDFS:/Sport_Auth_Five_Parallel_reducer/part-r-00004 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075/browseBlock.jsp?blockId=228148266981
ILLUSTARate describe
Hue
HBase Master
NameNode status
JobTracker Status
HDFS:/Sport_Auth_Five_P...
Pig Cookbook
Testing and Diagnostics
File: /Sport_Auth_Five_Parallel_reducer/part-r-00004
Goto : /Sport_Auth_Five_Parallel_re
go
Go back to dir listing
Advanced view/download options
View Next chunk
Games {(00915682,01-06-2011,4005356,147.36,Games,Bingo Sets,Dallas,Texas,credit),
(00870440,04-06-2011,4001014,51.79,Games,Mahjong,Columbus,Georgia,credit),
(00799288,02-27-2011,4004289,11.15,Games,Mahjong,Rockford,Illinois,cash),
(00834793,09-14-2011,4006595,198.25,Games,Mahjong,Colorado Springs,Colorado,credit),
(00869406,11-02-2011,4002155,150.33,Games,Mahjong,Seattle,Washington,credit),
(00930071,08-05-2011,4009455,123.07,Games,Portable Electronic Games,Omaha,Nebraska,credit),
(00851243,04-18-2011,4001628,119.14,Games,Poker Chips & Sets,Oakland,California,credit),
(00823920,01-10-2011,4000572,21.58,Games,Board Games,Midland,Texas,credit),
(00838625,03-25-2011,4001126,50.24,Games,Poker Chips & Sets,Orlando,Florida,credit),
(00883694,06-25-2011,4001409,145.27,Games,Bingo Sets,Vancouver,Washington,credit),
(00884398,02-12-2011,4002010,92.74,Games,Poker Chips & Sets,Salt Lake City,Utah,credit),
(00906431,06-19-2011,4004162,60.21,Games,Poker Chips & Sets,Dallas,Texas,credit),
(00869398,05-26-2011,4006614,166.51,Games,Card Games,Montgomery,Alabama,credit),
(00930072,04-15-2011,4004365,77.24,Games,Dice & Dice Sets,Gresham,Oregon,credit),
(00906433,01-11-2011,4008442,171.65,Games,Board Games,Sacramento,California,credit),
(00795362,10-07-2011,4009688,109.06,Games,Poker Chips & Sets,Columbus,Ohio,credit),
(00869392,06-02-2011,4007332,168.46,Games,Mahjong,Lincoln,Nebraska,credit),
(00808063,03-13-2011,4002797,23.07,Games,Poker Chips & Sets,Louisville,Kentucky,cash),
(00803321,02-17-2011,4000579,7.46,Games,Portable Electronic Games,Irving,Texas,credit),
(00858401,04-02-2011,4009470,53.74,Games,Dominoes,Pasadena,California,credit),
(00771508,10-12-2011,4000799,142.8,Games,Poker Chips & Sets,Coral Springs,Florida,credit),
Done
[cloudera@clo...
HDFS:/Sport_...
[cloudera]
[ParkingLot.tx...
[Sports_Autho...

# POC #6 HBase

```
18  public class Rating {
19
20      public static void main(String ratings []) throws IOException{
21
22          String ratingsFile = "/home/cloudera/Desktop/Books/RatingSample.txt";
23          String line;
24          int row = 0;
25
26          Configuration configuration = HBaseConfiguration.create();
27          // Ratings_TB must be created in Ubuntu Terminal before this code runs in Eclipse.
28          HTable hBaseTable = new HTable(configuration, "Ratings_TB");
29          BufferedReader bufferedReader = new BufferedReader(new FileReader(ratingsFile));
30          // cf_ratings must be created in Ubuntu Terminal before this code runs in Eclipse.
31
32          while ((line = bufferedReader.readLine()) != null) {
33              row++;
34              String value[] = line.split(",");// Similar to Map method of Mapper Class
35              String rowid = Integer.toString(row);
36              Put p = new Put(Bytes.toBytes(rowid));
37              p.add(Bytes.toBytes("cf_ratings"), Bytes.toBytes("Rating"),Bytes.toBytes(value[0]))
38              p.add(Bytes.toBytes("cf_ratings"), Bytes.toBytes("ISBN"),Bytes.toBytes(value[1]));
39              p.add(Bytes.toBytes("cf_ratings"), Bytes.toBytes("Price"),Bytes.toBytes(value[2]));
40              hBaseTable.put(p);
41
42              //System.out.println("Value : " + value[0]);
43              //System.out.println("Value : " + value[1]);
44              //System.out.println("Value : " + value[2]);
45
46          }
47      }
```

# Output of input below.

```
Applications Places System                                  cloudera
root@cloudera-vm: /home/cloudera
File  Edit  View  Search  Terminal  Help
88                          column=cf_ratings:Rating, timestamp=1406973016318, value="276796"
89                          column=cf_ratings:ISBN, timestamp=1406973016319, value="0006379702"
89                          column=cf_ratings:Price, timestamp=1406973016319, value="5"
89                          column=cf_ratings:Rating, timestamp=1406973016319, value="276798"
9                           column=cf_ratings:ISBN, timestamp=1406973016153, value="038550120X"
9                           column=cf_ratings:Price, timestamp=1406973016153, value="7"
9                           column=cf_ratings:Rating, timestamp=1406973016153, value="276744"
90                          column=cf_ratings:ISBN, timestamp=1406973016320, value="3423084049"
90                          column=cf_ratings:Price, timestamp=1406973016320, value="0"
90                          column=cf_ratings:Rating, timestamp=1406973016320, value="276798"
91                          column=cf_ratings:ISBN, timestamp=1406973016321, value="3442131340"
91                          column=cf_ratings:Price, timestamp=1406973016321, value="7"
91                          column=cf_ratings:Rating, timestamp=1406973016321, value="276798"
92                          column=cf_ratings:ISBN, timestamp=1406973016322, value="3442437407"
92                          column=cf_ratings:Price, timestamp=1406973016322, value="0"
92                          column=cf_ratings:Rating, timestamp=1406973016322, value="276798"
93                          column=cf_ratings:ISBN, timestamp=1406973016328, value="3446202102"
93                          column=cf_ratings:Price, timestamp=1406973016328, value="0"
93                          column=cf_ratings:Rating, timestamp=1406973016328, value="276798"
94                          column=cf_ratings:ISBN, timestamp=1406973016329, value="3453073398"
94                          column=cf_ratings:Price, timestamp=1406973016329, value="0"
94                          column=cf_ratings:Rating, timestamp=1406973016329, value="276798"
95                          column=cf_ratings:ISBN, timestamp=1406973016330, value="3453115783"
95                          column=cf_ratings:Price, timestamp=1406973016330, value="0"
95                          column=cf_ratings:Rating, timestamp=1406973016330, value="276798"
96                          column=cf_ratings:ISBN, timestamp=1406973016331, value="3499134004"
96                          column=cf_ratings:Price, timestamp=1406973016331, value="0"
96                          column=cf_ratings:Rating, timestamp=1406973016331, value="276798"
97                          column=cf_ratings:ISBN, timestamp=1406973016335, value="349915398X"
97                          column=cf_ratings:Price, timestamp=1406973016335, value="0"
97                          column=cf_ratings:Rating, timestamp=1406973016335, value="276798"
98                          column=cf_ratings:ISBN, timestamp=1406973016336, value="3548603203"
98                          column=cf_ratings:Price, timestamp=1406973016336, value="6"
98                          column=cf_ratings:Rating, timestamp=1406973016336, value="276798"
99                          column=cf_ratings:ISBN, timestamp=1406973016337, value="3764501383"
99                          column=cf_ratings:Price, timestamp=1406973016337, value="0"
99                          column=cf_ratings:Rating, timestamp=1406973016337, value="276798"
203 row(s) in 2.7230 seconds
HBase Comma...  [ParkingLot.tx...  root@clouder...  [Books]  Java - HbaseB...
```

---

```java
17 public class User {
18
19     public static void main(String[] userIP) throws IOException{
20
21         String ratingsFile = "/home/cloudera/Desktop/Books/User.txt";
22         String line;
23         int row = 0;
24
25         Configuration configuration = HBaseConfiguration.create();
26         // User_TB  must be created in Ubuntu Terminal before this code runs in Eclipse.
27         HTable hBaseTable = new HTable(configuration, "Users_TABLE");
28         BufferedReader bufferedReader = new BufferedReader(new FileReader(ratingsFile));
29         // cf_user must be created in Ubuntu Terminal before this code runs in Eclipse.
30
31         while ((line = bufferedReader.readLine()) != null) {
32             row++;
33             String value[] = line.split(",");// Similar to Map method of Mapper Class
34             String rowid = Integer.toString(row);
35             Put p = new Put(Bytes.toBytes(rowid));
36             p.add(Bytes.toBytes("cf_user"), Bytes.toBytes("Serial_Id"),Bytes.toBytes(value[0]));
37             p.add(Bytes.toBytes("cf_user"), Bytes.toBytes("Location"),Bytes.toBytes(value[1]));
38             p.add(Bytes.toBytes("cf_user"), Bytes.toBytes("Cost"),Bytes.toBytes(value[2]));
39             hBaseTable.put(p);
40
41
42         }
```

# Output of 30 MB input file after (approx) 50 minutes of processing loaded 2.5 lakh records

## http://snk.to/f-cdxkl7r2

```
Applications Places System                                   cloudera
root@cloudera-vm: /home/cloudera
File Edit View Search Terminal Help
 99989                          column=cf_user:Cost, timestamp=1406974111071, value="45"
 99989                          column=cf_user:Location, timestamp=1406974111071, value="lyon rhone alpes france"
 99989                          column=cf_user:Serial_Id, timestamp=1406974111071, value="99989"
 9999                           column=cf_user:Cost, timestamp=1406974026819, value="4"
 9999                           column=cf_user:Location, timestamp=1406974026819, value="beaverton oregon usa"
 9999                           column=cf_user:Serial_Id, timestamp=1406974026819, value="9999"
 99990                          column=cf_user:Cost, timestamp=1406974111078, value="16"
 99990                          column=cf_user:Location, timestamp=1406974111078, value="marl nordrhein-westfalen germany"
 99990                          column=cf_user:Serial_Id, timestamp=1406974111078, value="99990"
 99991                          column=cf_user:Cost, timestamp=1406974111079, value="20"
 99991                          column=cf_user:Location, timestamp=1406974111079, value="boonton township new jersey usa"
 99991                          column=cf_user:Serial_Id, timestamp=1406974111079, value="99991"
 99992                          column=cf_user:Cost, timestamp=1406974111079, value="55"
 99992                          column=cf_user:Location, timestamp=1406974111079, value="darwin northern territory australia"
 99992                          column=cf_user:Serial_Id, timestamp=1406974111079, value="99992"
 99993                          column=cf_user:Cost, timestamp=1406974111080, value="35"
 99993                          column=cf_user:Location, timestamp=1406974111080, value="munich bayern germany"
 99993                          column=cf_user:Serial_Id, timestamp=1406974111080, value="99993"
 99994                          column=cf_user:Cost, timestamp=1406974111081, value="4"
 99994                          column=cf_user:Location, timestamp=1406974111081, value="hambergen lower saxony germany"
 99994                          column=cf_user:Serial_Id, timestamp=1406974111081, value="99994"
 99995                          column=cf_user:Cost, timestamp=1406974111081, value="41"
 99995                          column=cf_user:Location, timestamp=1406974111081, value="lancaster california usa"
 99995                          column=cf_user:Serial_Id, timestamp=1406974111081, value="99995"
 99996                          column=cf_user:Cost, timestamp=1406974111082, value="43"
 99996                          column=cf_user:Location, timestamp=1406974111082, value="toronto ontario canada"
 99996                          column=cf_user:Serial_Id, timestamp=1406974111082, value="99996"
 99997                          column=cf_user:Cost, timestamp=1406974111083, value="33"
 99997                          column=cf_user:Location, timestamp=1406974111083, value="fredericton new brunswick canada"
 99997                          column=cf_user:Serial_Id, timestamp=1406974111083, value="99997"
 99998                          column=cf_user:Cost, timestamp=1406974111084, value="22"
 99998                          column=cf_user:Location, timestamp=1406974111084, value="cambridge england united kingdom"
 99998                          column=cf_user:Serial_Id, timestamp=1406974111084, value="99998"
 99999                          column=cf_user:Cost, timestamp=1406974111085, value="46"
 99999                          column=cf_user:Location, timestamp=1406974111085, value="hamburg hamburg germany"
 99999                          column=cf_user:Serial_Id, timestamp=1406974111085, value="99999"
243226 row(s) in 2936.5360 seconds

hbase(main):028:0>
HBase Comma...  [ParkingLot.tx...  root@clouder...  [Books]  Java - HbaseB...
```