

2015

# BAY AREA BIKE SHARING



Snehal Bhilare

INFO7250:Engineering Big-Data Systems

12/17/2015

## Index

---

● Project Description	2
● Dataset	2
● Techniques included	3
● Analytics	3
 <u>MapReduce Java:</u>	
1. Total no of stations per landmark	3
2. Average bike rented per day per station (Chaining)	7
 <u>PIG:</u>	
3. Total no of stations per landmark	15
4. Avg time between stations ordered by trip count (composite key)	16
5. Bike usage per hour (UDF)	18
 <u>Hive:</u>	
6. Creating and Loading data to HIVE	21
7. Ratio of subscriber to customer in trip data	23
8. Top 10 zip of customer type by total_trip	24
9. Most used terminal with total trip count	28
 <u>Hbase:</u>	
10. Load data to hbase	30
11. Retrive value from hbase	35



### PROJECT DESCRIPTION

The “Bay Area Bike Share” is the region’s bike sharing system with 700 bikes and 70 stations across the region, with locations in San Francisco, Redwood City, Mountain View, Palo Alto, and San Jose. It provides the transportation facilities to the residents with ease of use affordable service.



To get started with this service you just need to get the annual or 3 days membership. Enter your ride code or membership key and get going with your trip.



### FILE LIST

- 1) 201508\_status\_data.csv – approx. 37 million records of bike and dock availability by minute per station.
- 2) 201508\_station\_data.csv – 70 records – station ID, name, latitude, longitude, dockcount, city, installation date
- 3) 201508\_trip\_data.csv – approx. 354,000 records of individual trips.
- 4) 201508\_weather\_data.csv – 1,825 records of daily weather by city.

Link: [https://s3.amazonaws.com/babs-open-data/babs\\_open\\_data\\_year\\_2.zip](https://s3.amazonaws.com/babs-open-data/babs_open_data_year_2.zip)

### **Techniques included:**

- AWS EC2
- Hadoop
- HDFS
- JAVA MAPREDUCE
- CHAINING
- PIG
- UDF
- HIVE
- HBASE
- SPRING MVC
- HTML
- CSS
- BOOTSTRAP
- JAVA SCRIPT
- JSON
- D3JS
- TABLUE
- MAPS



### **CONFIGURATION**

Pseudo distributed mode AWS EC2 (Manuaaly configured ):

Data inside local : /usr/local/data/      Data inside hdfs: /input/data      hbase data : hbase://trip\_data

### **JAVA MAPREDUCE:**



----- Analysis : 1 -----

**Analysis:** Total no of stations per landmark from station data.

**Script on ec2:** /usr/local/java\_script/Station\_Per\_Landmark.jar

## Output on ec2: /map\_reduce\_javaCode/station\_per\_landmark/part-r-00000

Run:

```
[ec2-user@ip-172-31-33-207 ~]$ cd /usr/local/lib/hadoop-2.7.1/bin  
[ec2-user@ip-172-31-33-207 ~]$ chmod 777 Station_Per_Landmark.jar  
[ec2-user@ip-172-31-33-207 ~]$ java -scriptFile SHADOOP_HOME/bin  
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop jar /usr/local/java_script/Station_Per_Landmark.jar  
hdfs://localhost:9000/input/data/201508_station_data.csv  
/map_reduce_javaCode/station_per_landmark  
15/12/13 07:06:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id  
15/12/13 07:06:49 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=  
15/12/13 07:06:49 INFO InputFormat: Total input paths to process : 1  
15/12/13 07:06:49 INFO mapreduce.JobSubmission: number of splits:1  
15/12/13 07:06:49 INFO mapreduce.JobsSubmitter: Submitting tokens for job: job_local239289535_0001  
15/12/13 07:06:49 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
15/12/13 07:06:49 INFO mapreduce.Job: Running job: job_local239289535_0001  
15/12/13 07:06:49 INFO mapreduce.Job: Job complete: job_local239289535_0001  
15/12/13 07:06:49 INFO mapreduce.Job: OutputCommitter: FileOutputCommitter Algorithm version is 1  
15/12/13 07:06:49 INFO mapreduce.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter  
15/12/13 07:06:49 INFO mapreduce.LocalJobRunner: Waiting for map tasks  
15/12/13 07:06:49 INFO mapred.LocalJobRunner: Starting task: attempt_local239289535_0001_m_000000_0  
15/12/13 07:06:49 INFO mapred.Task: Using ResourceCalculatorProcessTree : []  
15/12/13 07:06:49 INFO mapred.Task: Processing split: hdfs://localhost:9000/input/data/201508_station_data.csv:0+5214  
15/12/13 07:06:49 INFO mapred.MapTask: (EQUATOR) 0 kv1 26214396 (104857584)  
15/12/13 07:06:49 INFO mapred.mapred.Task: mapreduce.task.io.sort.mb: 100  
15/12/13 07:06:49 INFO mapred.mapred.MapTask: soft limit at 83886080  
15/12/13 07:06:49 INFO mapred.mapred.MapTask: bufstart = 0; buffend = 104857600  
15/12/13 07:06:49 INFO mapred.mapred.MapTask: kvstart = 26214396; length = 6553600  
15/12/13 07:06:49 INFO mapred.mapred.MapTask: MapTask: map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer  
15/12/13 07:06:49 INFO mapred.mapred.MapTask: mapred.LocalJobRunner:  
15/12/13 07:06:49 INFO mapred.MapTask: Starting flush of map output  
15/12/13 07:06:49 INFO mapred.MapTask: Spilling map output  
15/12/13 07:06:49 INFO mapred.MapTask: bufstart = 0; buffend = 1153; bufvoid = 104857600  
15/12/13 07:06:49 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214120(104856480); length = 277/6553600  
15/12/13 07:06:50 INFO mapred.mapred.MapTask: Finished split 0  
15/12/13 07:06:50 INFO mapred.Task: Task:attempt_local239289535_0001_m_000000_0 is done. And is in the process of committing  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: map  
15/12/13 07:06:50 INFO mapred.Task: Task 'attempt_local239289535_0001_m_000000_0' done.  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: Finishing task: attempt_local239289535_0001_m_000000_0  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: map task executor complete.  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: Waiting for reduce tasks  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: Starting task: attempt_local239289535_0001_r_000000_0  
15/12/13 07:06:50 INFO mapred.Task: Using ResourceCalculatorProcessTree : []  
15/12/13 07:06:50 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7c490620  
15/12/13 07:06:50 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=333971456, maxSingleShuffleLimit=83492864, mergeThreshold=220421168, ioSortFactor=10, memToMemMergeOutputsThreshold=10  
15/12/13 07:06:50 INFO reduce.EventFetcher: attempt_local239289535_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events  
15/12/13 07:06:50 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local239289535_0001_m_000000_0 decomp: 92 len: 96 to MEMORY  
15/12/13 07:06:50 INFO reduce.InMemoryMapOutput: Read 92 bytes from map-output for attempt_local239289535_0001_m_000000_0  
15/12/13 07:06:50 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 92, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->92  
15/12/13 07:06:50 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
```

```
[ec2-user@ip-172-31-33-207 ~]$ ./task_local239289535_0001_r_000000  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: reduce > reduce  
15/12/13 07:06:50 INFO mapred.Task: Task 'attempt_local239289535_0001_r_000000_0' done.  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: reduce task executor complete.  
15/12/13 07:06:50 INFO mapred.LocalJobRunner: reduce task executor complete.  
15/12/13 07:06:50 INFO mapred.Job: Job job_local239289535_0001 running in uber mode : false  
15/12/13 07:06:50 INFO mapreduce.Job: map 10% reduce 100%  
15/12/13 07:06:50 INFO mapreduce.Job: Job job_local239289535_0001 completed successfully  
15/12/13 07:06:50 INFO mapreduce.Job: Counters : 35  
File System Counters  
FILE: Number of bytes read=44278  
FILE: Number of bytes written=604038  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=10428  
HDFS: Number of bytes written=72  
HDFS: Number of read operations=13  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=4  
Map-Reduce Framework  
Map input records=70  
Map output records=70  
Map output bytes=1153  
Map output materialized bytes=96  
Input split bytes=121  
Combine input records=70  
Combine output records=5  
Reduce input groups=5  
Reduce shuffle bytes=96  
Reduce input records=5  
Reduce output records=5  
Spilled Records=10  
Shuffled Maps =1  
Failed Shuffles=0  
Merged Map outputs=1  
GC time elapsed (ms)=11  
Total committed heap usage (bytes)=535298048  
Shuffle Errors  
BAD_ID=0  
CONNECTION=0  
IO_ERROR=0  
WRONG_PARTITION=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=5214  
File Output Format Counters  
Bytes Written=72  
[ec2-user@ip-172-31-33-207 bin]$
```

**Output:**

```
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -ls /map_reduce_javaCode
Found 1 items
drwxr-xr-x - ec2-user supergroup          0 2015-12-13 07:06 /map_reduce_javaCode/station_per_landmark
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -ls /map_reduce_javaCode/station_per_landmark/
Found 2 items
-rw-r--r-- 3 ec2-user supergroup          0 2015-12-13 07:06 /map_reduce_javaCode/station_per_landmark/_SUCCESS
-rw-r--r-- 3 ec2-user supergroup         72 2015-12-13 07:06 /map_reduce_javaCode/station_per_landmark/part-r-00000
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -cat /map_reduce_javaCode/station_per_landmark/part-r-00000
Mountain View    7
Palo Alto      5
Redwood City   7
San Francisco  35
San Jose       16
[ec2-user@ip-172-31-33-207 bin]$
```

Mountain View 7

Palo Alto 5

Redwood City 7

San Francisco 35

San Jose 16

**JAVA code:**

Station Per Landmark.java

```
/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
package Station;

import java.io.File;
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
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;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

/**
 *
 * @author ubuntu
 */
public class Station_Per_Landmark extends Configured implements Tool{

    public static class WordCountMapper
        extends Mapper<LongWritable, Text, Text, IntWritable> {
        private static IntWritable one = new IntWritable(1);
        private Text wordText = new Text();
        public void map(LongWritable key, Text line, Context context) throws IOException, InterruptedException {
            String []words=line.toString().split(",");//this will split the words by spaces and tabs[^\\w]
instead
            wordText.set(words[5]);
        }
    }
}
```

```
        context.write(wordText, one);
    }

}

public static class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context)
                      throws IOException, InterruptedException {
        int total = 0;
        for (IntWritable x : values) {
            total+= x.get();
        }
        result.set(total);
        context.write(key, result);
    }
}

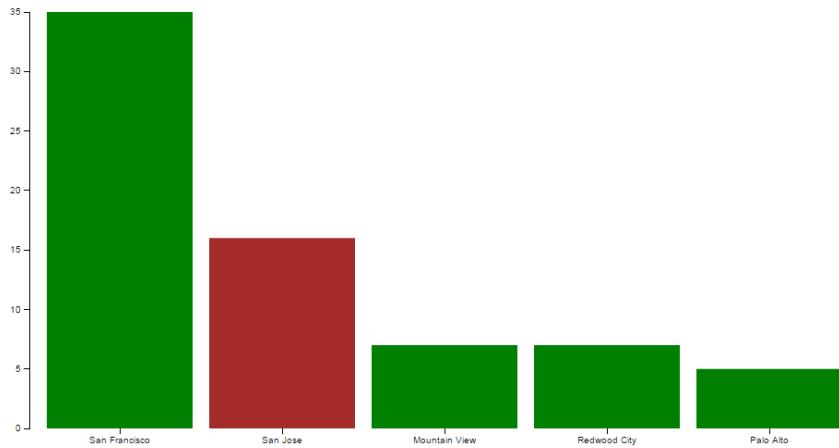
public int run(String[] args) throws Exception {
    Configuration conf = getConf();
    Job job = new Job(conf, "access count");
    job.setJarByClass(Station_Per_Landmark.class);
    final File f = new
File(Station_Per_Landmark.class.getProtectionDomain().getCodeSource().getLocation().getPath());
    // String inFiles = f.getAbsolutePath().replace("/build/classes", "") + "/src/inFiles/";
    String inFiles ="hdfs://localhost:9000/input/data/201508_station_data.csv";
    String outFiles = "/map_reduce_javaCode/station_per_landmark";
    //use the arguments instead if provided.
    if (args.length > 1) {
        inFiles = args[0];
        outFiles = args[1];
    }
    System.out.println(inFiles);
    System.out.println(outFiles);
    Path in = new Path(inFiles);
    Path out = new Path(outFiles);
    FileInputFormat.setInputPaths(job, in);
    FileOutputFormat.setOutputPath(job, out);
    job.setMapperClass(WordCountMapper.class);
    job.setCombinerClass(WordCountReducer.class);
    job.setReducerClass(WordCountReducer.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    System.exit(job.waitForCompletion(true) ? 0 : 1);
    return 0;
}

public static void main(String[] args) throws Exception {
    int res = ToolRunner.run(new Configuration(), new Station_Per_Landmark(), args);
    System.exit(res);
}
}
```

**Graph:**



**Station Per Landmark**



---

**Analysis : 2 JAVA Chaining**

---

**Analysis:** Average bike rented per day per station. Used java chaining map reduce to generate composite key as “station\_id + day of week”.

**Result:** Business is rarely affected by the day of the week.

**Jar path :** /usr/local/java\_script/bikes\_perDay.jar

**Output path:** /map\_reduce\_javaCode/bikes\_perDay\_perStation/part-r-00000

## Run:

```

ec2-user@ip-172-31-33-207:/usr/local/lib/hadoop-2.7.1/bin
12393 SecondaryNameNode
10662 Jps
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop jar /usr/local/java_script/bikes_perDay.jar
hdfs://localhost:9000/input/data/201508_status_data.csv
map reduce javaCode/bikes_perDay_perStation
15/12/16 01:50:11 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/12/16 01:50:11 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/12/16 01:50:12 INFO mapreduce.JobSubmitter: number of splits:9
15/12/16 01:50:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2004975357_0001
15/12/16 01:50:12 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/12/16 01:50:12 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/12/16 01:50:12 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/12/16 01:50:12 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
15/12/16 01:50:12 INFO mapred.LocalJobRunner: Waiting for map tasks
15/12/16 01:50:12 INFO mapred.LocalJobRunner: Starting task attempt_local2004975357_0001_m_000000_0
15/12/16 01:50:12 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
15/12/16 01:50:12 INFO mapred.Task: Processing split: hdfs://localhost:9000/input/data/201508_status_data.csv:0+134217728
15/12/16 01:50:12 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/12/16 01:50:12 INFO mapred.MapTask: mapred.task.io.sort.mb: 100
15/12/16 01:50:12 INFO mapred.MapTask: soft limit at 83886080
15/12/16 01:50:12 INFO mapred.MapTask: bufstart = 0; bufend = 104857600
15/12/16 01:50:12 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/12/16 01:50:13 INFO mapreduce.Job: map_local2004975357_0001 running in uber mode : false
15/12/16 01:50:13 INFO mapreduce.Job: map 0% reduce 0%
15/12/16 01:50:18 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:18 INFO mapred.LocalJobRunner: map 1% reduce 0%
15/12/16 01:50:21 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:22 INFO mapred.LocalJobRunner: map 2% reduce 0%
15/12/16 01:50:24 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:25 INFO mapred.LocalJobRunner: map 3% reduce 0%
15/12/16 01:50:25 INFO mapred.MapTask: Spilling map output
15/12/16 01:50:25 INFO mapred.MapTask: bufstart = 0; bufend = 48887970; bufwind = 104857600
15/12/16 01:50:25 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 17464868(69859472); length = 8749529/6553600
15/12/16 01:50:25 INFO mapred.MapTask: 57718066 kvi 14429512(57718048)
15/12/16 01:50:27 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:28 INFO mapred.LocalJobRunner: map 4% reduce 0%
15/12/16 01:50:30 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:33 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:39 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:39 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:39 INFO mapred.MapTask: Finished spill 0
15/12/16 01:50:39 INFO mapred.MapTask: (RESET) equator 57718066 kv 14429512(57718048) kvi 12255932(49023728)
15/12/16 01:50:42 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:43 INFO mapred.LocalJobRunner: map 5% reduce 0%
15/12/16 01:50:45 INFO mapred.LocalJobRunner: map > map
15/12/16 01:50:46 INFO mapred.LocalJobRunner: map 6% reduce 0%
15/12/16 01:50:46 INFO mapreduce.Job: map 6% reduce 0%
File System Counters
FILE: Number of bytes read=550332
FILE: Number of bytes written=3521768
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=7006534968
HDFS: Number of bytes written=31299
HDFS: Number of read operations=141
HDFS: Number of large read operations=12
HDFS: Number of write operations=12
Map-Reduce Framework
  Input split records=36647622
  Map output records=36647622
  Map output bytes=818860110
  Map output materialized bytes=36456
  Input split bytes=1080
  Combine input records=36648272
  Combine output records=1195
  Reduce input groups=497
  Reduce shuffle bytes=36456
  Reduce input records=545
  Reduce output records=497
  Spilled Records=1740
  Shuffled Maps=0
  Failed Shuffles=0
  Merged Map outputs=9
  GC time elapsed (ms)=3033
  Total committed heap usage (bytes)=3890741248
Shuffle Errors
  BAD ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1087274652
File Output Format Counters
  Bytes Written=31299
[ec2-user@ip-172-31-33-207 bin]$ 
```

## Output:

```

Bytes Written=31299
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -ls /map_reduce_javaCode/bikes_perDay_perStation
Found 2 items
-rw-r--r-- 3 ec2-user supergroup          0 2015-12-16 01:54 /map_reduce_javaCode/bikes_perDay_perStation/_SUCCESS
-rw-r--r-- 3 ec2-user supergroup 31299 2015-12-16 01:54 /map_reduce_javaCode/bikes_perDay_perStation/part-r-00000
[ec2-user@ip-172-31-33-207 bin]$ 
```

```
[ec2-user@ip-172-31-33-207 ~]$ ./hadoop fs -ls /map_reduce_javaCode/bikes_perDay_perStation
Found 2 items
-rw-r--r-- 3 ec2-user supergroup 0 2015-12-16 01:54 /map_reduce_javaCode/bikes_perDay_perStation/_SUCCESS
-rw-r--r-- 3 ec2-user supergroup 31299 2015-12-16 01:54 /map_reduce_javaCode/bikes_perDay_perStation/part-r-00000
[ec2-user@ip-172-31-33-207 ~]$ ./hadoop fs -cat /map_reduce_javaCode/bikes_perDay_perStation/part-r-00000
2:Friday 13
2:Monday 13
2:Saturday 13
2:Sunday 14
2:Thursday 13
2:Wednesday 13
3:Friday 5
3:Monday 5
3:Saturday 5
3:Sunday 5
3:Thursday 6
3:Tuesday 6
3:Wednesday 6
4:Friday 5
4:Monday 5
4:Saturday 5
4:Sunday 5
4:Thursday 5
4:Tuesday 5
4:Wednesday 6
5:Friday 11
5:Monday 11
5:Saturday 11
5:Sunday 12
5:Thursday 11
5:Tuesday 11
5:Wednesday 11
6:Friday 8
6:Monday 7
6:Saturday 8
6:Sunday 7
6:Wednesday 7
6:Tuesday 7
6:Thursday 7
7:Friday 5
7:Monday 5
7:Saturday 5
7:Sunday 5
7:Tuesday 5
7:Wednesday 5
8:Friday 7
8:Monday 7
8:Saturday 7
```

```
[ec2-user@ip-172-31-33-207 ~]$ hadoop jar /usr/local/lib/hadoop-2.7.1/bin  
76:Friday 10  
76:Monday 9  
76:Tuesday 10  
76:Sunday 10  
76:Thursday 10  
76:Tuesday 10  
76:Wednesday 10  
77:Friday 13  
77:Monday 13  
77:Saturday 13  
77:Sunday 14  
77:Thursday 14  
77:Tuesday 12  
77:Wednesday 13  
78:Friday 8  
80:Saturday 9  
80:Sunday 8  
80:Thursday 8  
80:Tuesday 8  
80:Wednesday 8  
82:Friday 8  
82:Monday 8  
82:Saturday 8  
82:Sunday 8  
82:Thursday 8  
82:Tuesday 8  
82:Wednesday 8  
83:Friday 9  
83:Monday 7  
83:Saturday 8  
83:Sunday 7  
83:Thursday 8  
83:Tuesday 8  
83:Wednesday 8  
84:Friday 7  
84:Monday 7  
84:Saturday 6  
84:Sunday 7  
84:Thursday 7  
84:Tuesday 6  
84:Wednesday 7  
84:Sunday 8  
84:Thursday 8  
84:Tuesday 8  
84:Wednesday 7  
[ec2-user@ip-172-31-33-207 bin]$
```

## JAVA Code:

station perDay.java

```
/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
package bikes_perDayStation;
```

```
import java.io.DataInput;
import java.io.DataOutput;
import java.io.IOException;
import org.apache.hadoop.io.WritableComparable;

/**
 *
 * @author ubuntu
 */
public class station_perDay implements WritableComparable<station_perDay>
{

    private int station_id;
    private String day;

    public String getDay() {
        return day;
    }

    public void setDay(String day) {
        this.day = day;
    }

    public int getStation_id() {
        return station_id;
    }

    public void setStation_id(int station_id) {
        this.station_id = station_id;
    }

    @Override
    public void write(DataOutput out) throws IOException {
        out.writeInt(station_id);
        out.writeChars(day);

    }

    @Override
    public void readFields(DataInput in) throws IOException {

        station_id = in.readInt();
        day= in.readLine();
    }

    @Override
    public int compareTo(station_perDay t) {

        int result= (int)(this.getStation_id()-t.getStation_id());
        if (0 == result) {
            result = day.compareTo(t.day);
        }
        return result;

    }

    @Override
    public String toString() {
        return station_id+":"+day;
    }

}
```

main.java

```
/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
package bikes_perDayStation;

import java.io.File;
import java.io.IOException;
import java.text.DateFormat;
import java.text.ParseException;
import java.text.SimpleDateFormat;

import java.util.Date;
import java.util.logging.Level;
import java.util.logging.Logger;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
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;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

/**
 *
 * @author ubuntu
 */
public class main extends Configured implements Tool{

    public static class WordCountMapper
        extends Mapper<LongWritable, Text, station_perDay, IntWritable> {
        private station_perDay wordText = new station_perDay();

        public void map(LongWritable key, Text line, Context context) throws IOException, InterruptedException {
            String []words=line.toString().split(",");//this will split the words by spaces and tabs[^\\w]
instead
            if (words[0].equals("")||words[0].isEmpty()||words[0].equals("station_id")||
                words[3].equals("")||words[3].isEmpty()||words[3].equals("time")||
                words[2].equals("")||words[2].isEmpty()||words[2].equals("docks_available")) return;
            wordText.setStation_id(Integer.parseInt(words[0]));
            String day;

            try {
                day = getDayFromDate(words[3]);
                wordText.setDay(day);
            } catch (ParseException ex) {
                Logger.getLogger(main.class.getName()).log(Level.SEVERE, null, ex);
            }

            IntWritable result;
            result = new IntWritable(Integer.parseInt(words[2]));

            context.write(wordText, result);
        }

        private String getDayFromDate(String word) throws ParseException {
            word = word.substring(1, word.length()-1);

            Date yourDate= new SimpleDateFormat("yyyy-mm-dd HH:mm:ss").parse(word);
            DateFormat format2=new SimpleDateFormat("EEEE");
            String finalDay=format2.format(yourDate);

            return finalDay;
        }
    }
}
```

```
        }

    }

    public static class WordCountReducer extends Reducer<station_perDay, IntWritable, station_perDay,
IntWritable> {

        public void reduce(station_perDay key, Iterable<IntWritable> values,
                           Context context)
                throws IOException, InterruptedException {

            int result =0;
            int count =0;
            for (IntWritable value : values) {

                result += value.get();
                count++;

            }

            IntWritable total = new IntWritable(result/count);

            context.write(key, total);

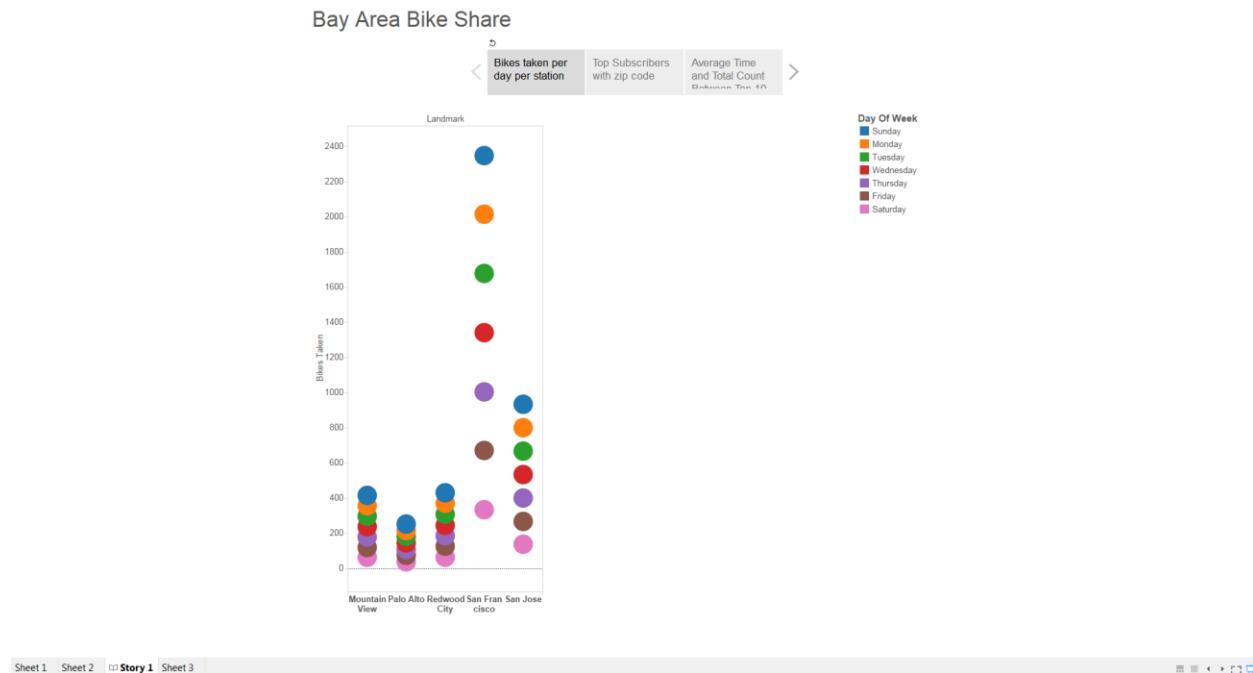
        }
    }

    public int run(String[] args) throws Exception {
        Configuration conf = getConf();
        Job job = new Job(conf, "access count");
        job.setJarByClass(main.class);
        final File f = new File(main.class.getProtectionDomain().getCodeSource().getLocation().getPath());
        //String inFile="/usr/local/data/201508_status_data.csv";
        //String outFile=f.getAbsolutePath().replace("/build/classes", "") + "/src/outFiles/WorWdCount";
        String inFile ="hdfs://localhost:9000/input/data/201508_status_data.csv";
        String outFile = "/map_reduce_javaCode/bikes_perDay_perStation";

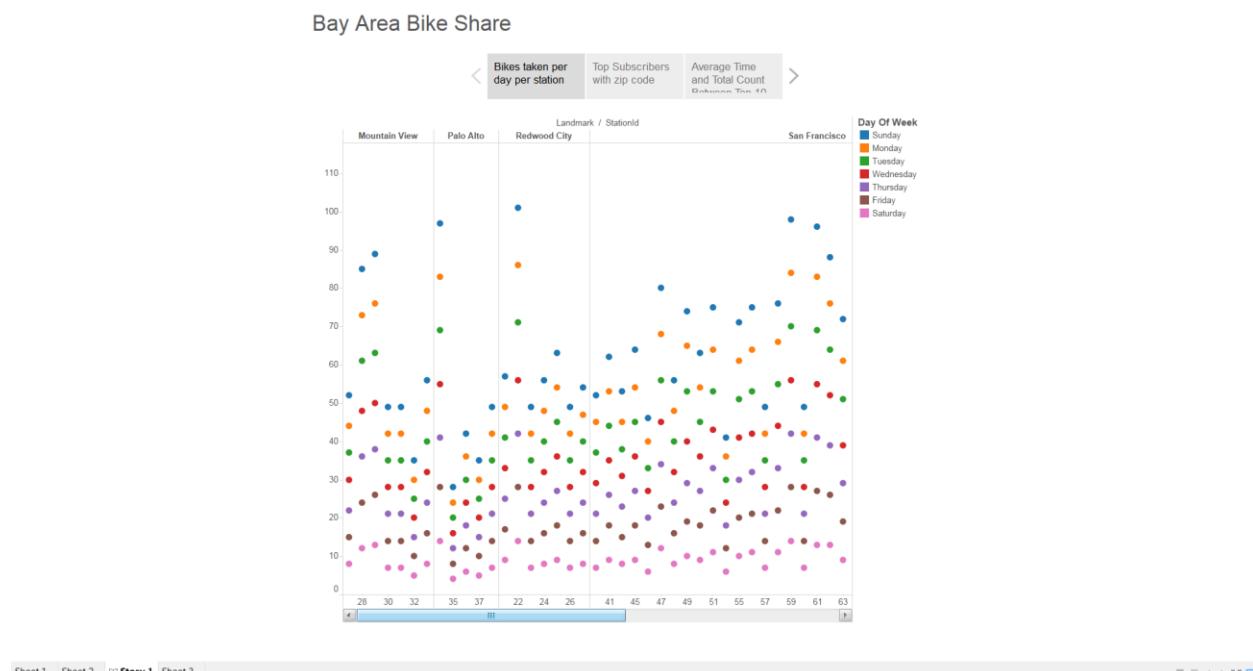
        if (args.length > 1) {
            inFile = args[0];
            outFile = args[1];
        }
        System.out.println(inFile);
        System.out.println(outFile);
        Path in = new Path(infile);
        Path out = new Path(outFile);
        FileInputFormat.setInputPaths(job, in);
        FileOutputFormat.setOutputPath(job, out);
        job.setMapperClass(WordCountMapper.class);
        job.setCombinerClass(WordCountReducer.class);
        job.setReducerClass(WordCountReducer.class);
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        job.setOutputKeyClass(station_perDay.class);
        job.setOutputValueClass(IntWritable.class);
        System.exit(job.waitForCompletion(true) ? 0 : 1);
        return 0;
    }

    public static void main(String[] args) throws Exception {
        int res = ToolRunner.run(new Configuration(), new main(), args);
        System.exit(res);
    }
}
```

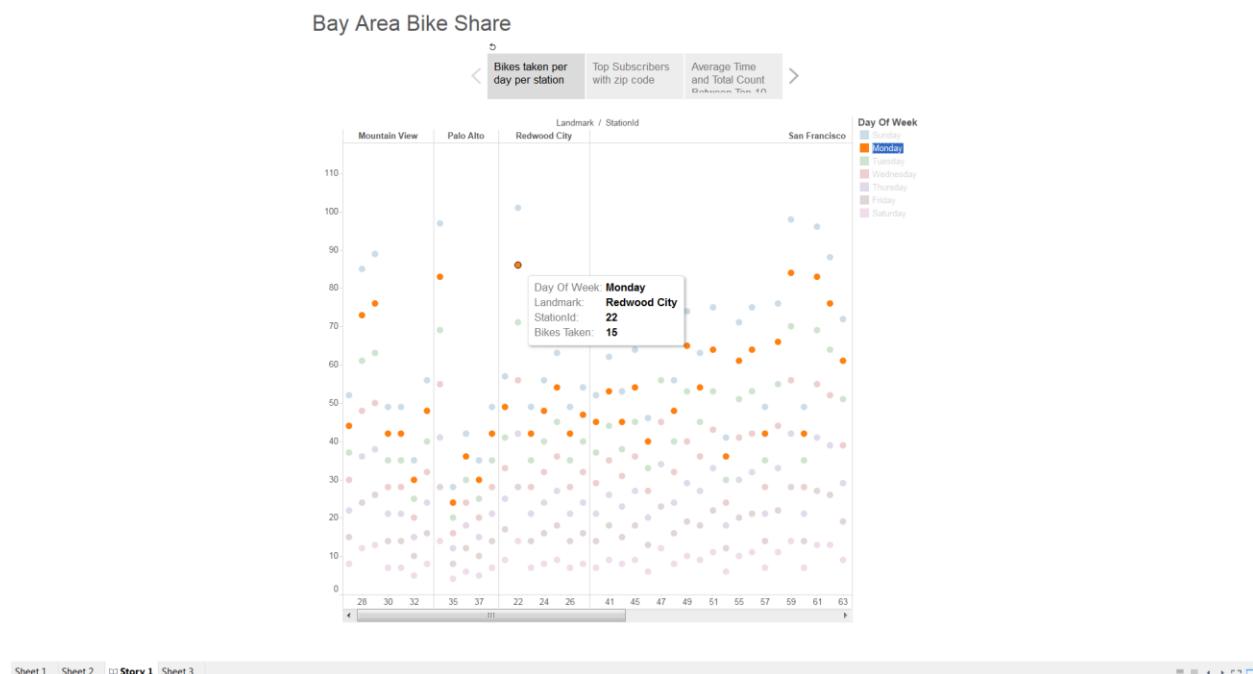
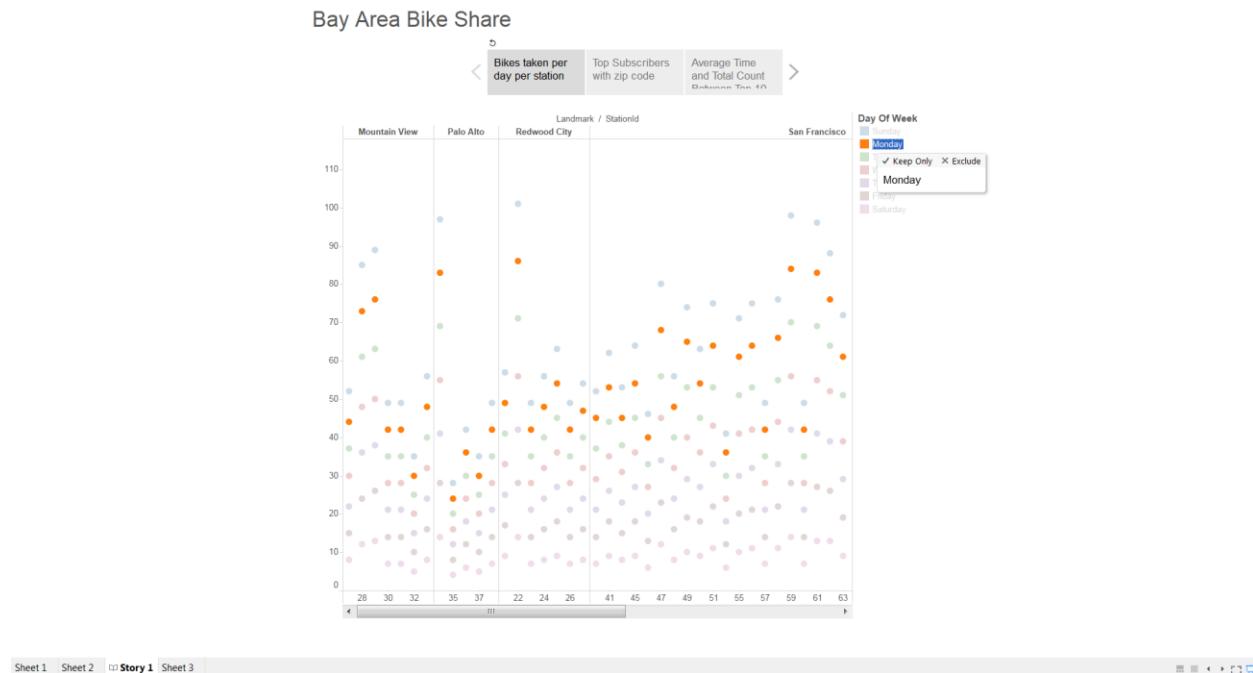
**Graph:**



**Drill down landmark**



Filter by day:





**PIG:**

----- **Analysis : 3** -----

**Analysis:** Total no of stations per landmark from station data.

**pig script:** /usr/local/pig\_script/station\_per\_landmark.pig

**pig output:** /pig\_mapreduce/station\_per\_landmark/part-r-00000

**Script:**

```
station_data = LOAD 'hdfs:/input/data/201508_station_data.csv' USING PigStorage(',') AS  
(station_id:int,name:chararray,lat:double,long:double,dockcount:int,landmark:chararray,installation:dat  
etime);
```

B = group station\_data by landmark;

C = FOREACH B generate group , COUNT(station\_data) as noOfStation;

D = ORDER C by noOfStation DESC;

Store D into 'hdfs:/pig\_mapreduce/station\_per\_landmark';

**Output:**

```
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -ls /pig_mapreduce/station_per_landmark/  
Found 2 items  
-rw-r--r-- 3 ec2-user supergroup 0 2015-12-13 05:54 /pig_mapreduce/station_per_landmark/_SUCCESS  
-rw-r--r-- 3 ec2-user supergroup 72 2015-12-13 05:54 /pig_mapreduce/station_per_landmark/part-r-00000  
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -cat /pig_mapreduce/station_per_landmark/part-r-00000  
San Francisco 35  
San Jose 16  
Mountain View 7  
Redwood City 7  
Palo Alto 5  
[ec2-user@ip-172-31-33-207 bin]$
```

San Francisco 35

San Jose 16

Mountain View 7

Redwood City 7

Palo Alto 5

**Graph:**



----- Analysis : 4 -----

**Analysis:** Estimated time between two stations (average time to reach from one station to other) and total no of trips in between them ordered by trip\_count. Used start and end station as key with CONCAT function

**pig script:** /usr/local/pig\_script/Sorted\_totalTrip\_estimatedDuration.pig

**pig output:** hdfs://pig\_mapreduce/Sorted\_stationPairAnalysis/part-r-00000

**Script:**

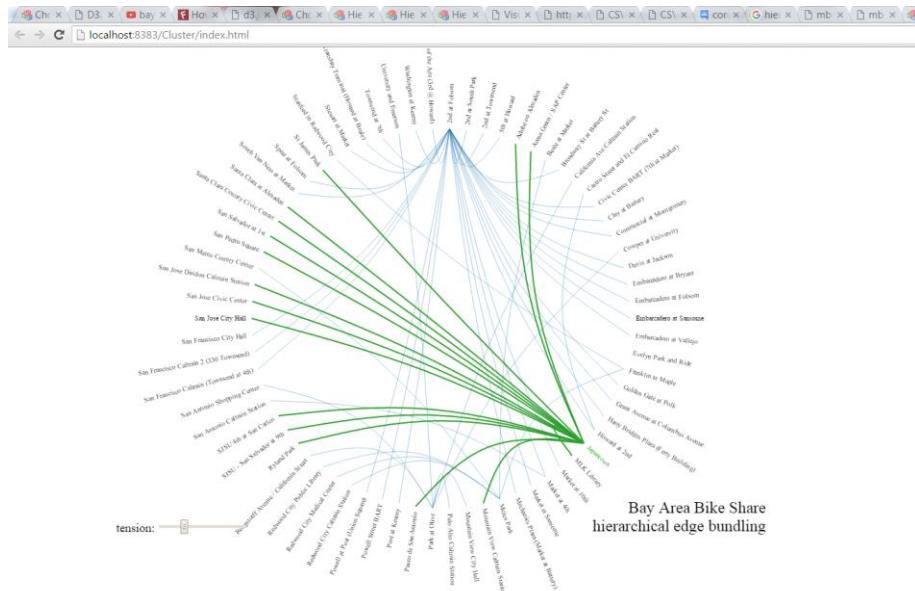
```
trip_data = LOAD 'hdfs:/input/data/201508_trip_data.csv' USING PigStorage(',') AS  
(trip_id:int,duration:int,start_date:chararray,start_station:chararray,start_terminal:int,end_date:chararray,end_station:chararray,end_terminal:int,bike:int,subscriber:chararray,zip:int);  
  
group_data = foreach trip_data generate CONCAT(start_station,' ',end_station)as stationPair,duration;  
  
station_pairs = group group_data by stationPair;  
  
per_pair_analysis = FOREACH station_pairs generate group , COUNT(group_data) as trip_count,  
AVG(group_data.duration) as estimated_time;  
  
sorted_per_pair_analysis = ORDER per_pair_analysis by trip_count DESC;  
  
Store sorted_per_pair_analysis into 'hdfs:/pig_mapreduce/Sorted_stationPairAnalysis';
```

## Output:

```
ec2-user@ip-172-31-207:~/usr/local/lib/hadoop-2.7.1/bin
Palo Alto      5
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -ls /pig_mapreduce/Sorted_stationPairAnalysis/
Found 2 items
-rw-r--r--  3 ec2-user supergroup   0 2015-12-13 06:21 /pig_mapreduce/Sorted_stationPairAnalysis/_SUCCESS
-rw-r--r--  3 ec2-user supergroup 116167 2015-12-13 06:21 /pig_mapreduce/Sorted_stationPairAnalysis/part-r-00000
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -cat /pig_mapreduce/Sorted_stationPairAnalysis/part-r-00000
San Francisco Caltrain 2 (330 Townsend) , Townsend at 7th      3748    282.13794023479187
Harry Bridges Plaza (Ferry Building) , Embarcadero at Sansome     3145    1110.8642289348172
2nd at Townsend , Harry Bridges Plaza (Ferry Building)      2973    554.713084265052
Townsend at 7th , San Francisco Caltrain 2 (330 Townsend)      2734    275.23372348207755
Harry Bridges Plaza (Ferry Building) , 2nd at Townsend      2640    645.450757575756
Embarcadero at Polson , San Francisco Caltrain (Townsend at 4th) 2439    777.539975399754
Stewart at Market , 2nd at Townsend      2356    551.893039049236
Embarcadero at Sansome , Stewart at Market      2330    486.0321884120173
Townsend at 7th , San Francisco Caltrain (Townsend at 4th)      2194    252.86661313869613
Temporary Transbay Terminal (Howard at Beale) , San Francisco Caltrain (Townsend at 4th)      2184    653.2047985347985
San Francisco Caltrain (Townsend at 4th) , Harry Bridges Plaza (Ferry Building)      2091    800.6934461109517
San Francisco Caltrain 2 (330 Townsend) , Powell Street BART      2074    559.0838958534233
Powell Street BART , San Francisco Caltrain 2 (330 Townsend)      2042    472.78403525954946
Stewart at Market , San Francisco Caltrain (Townsend at 4th)      1967    712.467208947636
San Francisco Caltrain (Townsend at 4th) , Temporary Transbay Terminal (Howard at Beale)      1937    765.0
San Francisco Caltrain 2 (330 Townsend) , 5th at Howard      1865    395.3324396762842
Harry Bridges Plaza (Ferry Building) , San Francisco Caltrain (Townsend at 4th)      1856    898.7957974137931
Townsend at 7th , Civic Center BART (7th at Market)      1844    555.0379609544468
Market at 10th , San Francisco Caltrain (Townsend at 4th)      1815    633.2314049586777
Market at 10th , San Francisco Caltrain 2 (330 Townsend)      1806    620.6434108527131
Embarcadero at Sansome , Harry Bridges Plaza (Ferry Building)      1798    792.3648496331479
2nd at South Park , Market at Sansome      1790    694.5094572057039
Civic Center BART (7th at Market) , Townsend at 7th      1765    600.8645892351275
2nd at Townsend , Market at Market      1662    577.2268235138387
Market at Sansome , 2nd at South Park      1636    332.44009779951
Stewart at Market , Embarcadero at Sansome      1614    733.011524163568
Market at 4th , San Francisco Caltrain (Townsend at 4th)      1583    514.9001895135818
San Francisco Caltrain 2 (330 Townsend) , Market at 10th      1566    815.0849297573435
San Francisco Caltrain (Townsend at 4th) , Embarcadero at Polson      1530    630.9039215686274
San Francisco Caltrain (Townsend at 4th) , Stewart at Market      1431    731.5590496156534
Mountain View Caltrain Station , Mountain View City Hall      1419    377.1374207188161
5th at Howard , San Francisco Caltrain 2 (330 Townsend)      1389    503.8768898488121
Market at 10th , Market at 4th      1389    471.89992800575993
Howard at 2nd , San Francisco Caltrain (Townsend at 4th)      1322    490.303328290469
Mountain View City Hall , Mountain View Caltrain Station      1308    456.7874617737003
San Jose Diridon Station , San Jose Civic Center Almaden      1302    329.84537662331
People at Market , San Francisco Caltrain (Townsend at 4th)      1215    735.615139395062
Santa Clara at Almaden , San Jose Diridon Caltrain Station      1214    786.5296540362439
Grant Avenue at Columbus Avenue , Market at Sansome      1209    714.920595334987
Market at Sansome , San Francisco Caltrain (Townsend at 4th)      1208    761.9387417218543
Market at 4th , Market at 10th      1205    687.6174273858921
San Francisco Caltrain (Townsend at 4th) , Market at Sansome      1203    1050.8645054031588
San Francisco Caltrain (Townsend at 4th) , Townsend at 7th      1198    303.8408001335559
San Francisco Caltrain (Townsend at 4th) , Howard at 2nd      1189    499.8435660218671
[ec2-user@ip-172-31-33-207 bin]$
```

```
ec2-user@ip-172-31-207:~/usr/local/lib/hadoop-2.7.1/bin
Mezes Park , Palo Alto Caltrain Station 2      2517.5
Castro Street and El Camino Real , Park at Olive 2      51631.5
St James Park , San Jose Civic Center 2      360.5
Park at Olive , Castro Street and El Camino Real 2      2453.5
San Antonio Shopping Center , Evelyn Park and Ride 2      1934.0
University and Emerson , San Mateo County Center 2      1918.5
Redwood City Caltrain Station , Park at Olive 2      2346.5
Santa Clara County Civic Center , SJU and San Salvador at 9th 2      1260.5
San Antonio Caltrain Station , Franklin at Maple 1      9493.0
Franklin at Maple , Palo Alto Caltrain Station 1      2761.0
Palo Alto Caltrain Station , Franklin at Maple 1      3120.0
Redwood City Public Library , Palo Alto Caltrain Station 1      1983.0
San Francisco Caltrain (Townsend at 4th) , San Antonio Caltrain Station 1      28226.0
Redwood City Caltrain Station , California Ave Caltrain Station 1      2457.0
MLK Library , Market at 4th 1      29942.0
San Antonio Caltrain Station , San Jose Civic Center 1      11083.0
Stanford in Redwood City , California Ave Caltrain Station 1      2872.0
Palo Alto Caltrain Station , Mountain View City Hall 1      3044.0
San Mateo County Center , Cooper at University 1      12731.0
Redwood City Caltrain Station , Mountain View Caltrain Station 1      16744.0
Oxford in Redwood City , Cooper at University 1      4694.0
San Antonio Caltrain Station , Evelyn Park and Ride 1      1185.0
San Antonio Caltrain Station , Cooper at University 1      2114.0
San Mateo County Center , Park at Olive 1      1860.0
Redwood City Medical Center , Mezes Park 1      6321.0
Redwood City Public Library , Mezes Park 1      551.0
Powell Street BART , San Antonio Shopping Center 1      6162.0
Park at Olive , Mountain View City Hall 1      2818.0
Castro Street and El Camino Real , Cooper at University 1      7530.0
University and Emerson , Redwood City Caltrain Station 1      1806.0
Market at 4th , Stanford in Redwood City 1      23661.0
Rylan Park , SJU and San Salvador at 9th 1      855.0
Cooper at University , San Antonio Shopping Center 1      2946.0
Cooper at University , Evelyn Park and Ride 1      11334.0
University and Emerson , Castro Street and El Camino Real 1      4150.0
San Antonio Shopping Center , Cooper at University 1      1631.0
Castro Street and El Camino Real , California Ave Caltrain Station 1      1645.0
Paseo de San Antonio , Santa Clara County Civic Center 1      30751.0
Cooper at University , Stamford in Redwood City 1      2106.0
Beale at Market , Mezes Park 1      20586.0
Palo Alto Caltrain Station , Redwood City Public Library 1      2845.0
Mezes Park , California Ave Caltrain Station 1      2300.0
University and Emerson , Franklin at Maple 1      1776.0
Cooper at University , San Antonio Caltrain Station 1      2430.0
Park at Olive , San Francisco Caltrain (Townsend at 4th) 1      15940.0
San Francisco Caltrain (Townsend at 4th) , San Jose County Civic Center 1      973.0
Mountain View Caltrain Station , Coop at University 1      5067.0
Evelyn Park and Ride , Arena Green / SAP Center 1      27409.0
Evelyn Park and Ride , San Antonio / SAP Center 1      859.0
[ec2-user@ip-172-31-33-207 bin]$
```

## Graph:



## Analysis : 5 UDF

## Analysis:

Timely (Hourly) analysis of total docks available (average) on particular stations (station id =2): tells how many bikes are taken from station. Can be used to analyze peak time when customer uses bike services.

Used UDF to convert date to each hour.

**Result:** Maximum bike taken at 9 and 10 am

**UDF Jar:** /usr/local/java\_script/Hourly\_TripStatus/sne.jar

**UDF Java Class:** /usr/local/java\_script/Hourly\_TripStatus/ConvertDate.class

**UDF script For (station id =2):** /usr/local/java\_script/Hourly\_TripStatus/bikeUsage\_station2.pig

**Output** at hdfs:/pig\_mapreduce/Hourly\_TripStatus\_Station2

## Script:

-- myscript.pig

```
REGISTER /usr/local/java script/Hourly TripStatus/sne.jar;
```

```
status_dummy_data = LOAD 'hdfs:/input/data/201508_status_data.csv' USING PigStorage(',') AS (station_id:int,bikes_avail:int,dock_avail:int,time:chararray);
```

```
status_data = FILTER status_dummy_data BY station_id ==2;  
C = FOREACH status_data GENERATE sne.ConvertDate(time) as date, dock_avail;  
D = group C by date;  
E = FOREACH D generate group, AVG(C.dock_avail) as total_rented_bikes;  
STORE E INTO 'hdfs:/pig_mapreduce/Hourly_TripStatus_Station2';
```

**UDF class to get hours from a date:**

```
- ConvertDate.java-----  
  
package sne;  
  
import java.io.IOException;  
  
import org.apache.pig.EvalFunc;  
  
import org.apache.pig.data.Tuple;  
  
  
public class ConvertDate extends EvalFunc<String>  
{  
  
    public String exec(Tuple input) throws IOException {  
  
        if (input == null || input.size() == 0)  
            return null;  
  
        try{  
  
            String str = (String)input.get(0);  
  
            str = str.substring(11,str.length()-7);  
  
            str = str+" Hour ";  
  
            return str;  
  
        }catch(Exception e){  
  
            throw new IOException("Caught exception processing input row ", e);  
        }  
    }  
}
```

## Output:

```

ec2-user@ip-172-31-33-207:~/usr/local/java/script/Hourly_TripStatus
2015-12-14 04:39:05,604 [pool-6-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local1588941011_0001_r_000001_0 is done. And is in the process of committing
2015-12-14 04:39:05,608 [pool-6-thread-1] INFO org.apache.hadoop.mapred.Task - 9 / 9 copied.
2015-12-14 04:39:05,608 [pool-6-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Task attempt_local1588941011_0001_r_000001_0 is allowed to commit now
2015-12-14 04:39:05,613 [pool-6-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local1588941011_0001_r_000001_0' to hdfs://localhost:9000/pig/mapreduce/Hourly_TripStatus_Station2/_temporary0/_task_local1588941011_0001_r_000001
2015-12-14 04:39:05,614 [pool-6-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce > reduce
2015-12-14 04:39:05,614 [pool-6-thread-1] INFO org.apache.hadoop.mapred.Task - Task attempt_local1588941011_0001_r_000001_0 is done.
2015-12-14 04:39:05,614 [Thread-20] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1588941011_0001_r_000001_0
2015-12-14 04:39:05,614 [Thread-20] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2015-12-14 04:39:05,614 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,539 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,543 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2015-12-14 04:39:10,544 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,592 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2015-12-14 04:39:10,594 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimpleFigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.7.1 0.14.0-SNAPSHOT ec2-user 2015-12-14 04:37:53 2015-12-14 04:39:10 GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_local1588941011_0001 9 2 n/a n/a n/a n/a n/a n/a C,D,E,status_data,status_dummy_data GROUP_BY,COMBINER hdfs:/pig_mapreduce/Hourly_TripStatus_Station2

Input(s):
Successfully read 36647622 records (8167083040 bytes) from: "hdfs:/input/data/201508_status_data.csv"

Output(s):
Successfully stored 24 records (73274457 bytes) in: "hdfs:/pig_mapreduce/Hourly_TripStatus_Station2"

Counters:
Total records written : 24
Total bytes written : 73274457
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1588941011_0001

2015-12-14 04:39:10,595 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,596 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,597 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-14 04:39:10,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2015-12-14 04:39:10,639 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 19 seconds and 317 milliseconds (7931 ms)
[ec2-user@ip-172-31-33-207 Hourly_TripStatus]$ 
```

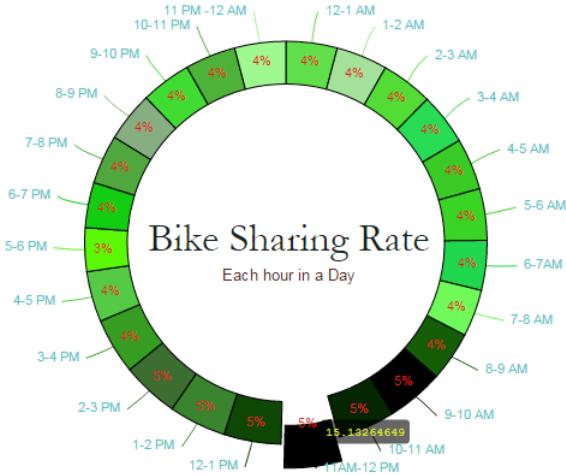


```

ec2-user@ip-172-31-33-207:~/usr/local/lib/hadoop-2.7.1/bin
(13, 13 Hour ) 8.377458603970359
(13, 15 Hour ) 8.358326068003487
(13, 17 Hour ) 8.76652305435987
(13, 19 Hour ) 8.753499472210748
(13, 20 Hour ) 8.753499472210748
(13, 22 Hour ) 8.665766656796318
(14, 01 Hour ) 10.871609305210007
(14, 03 Hour ) 10.891666666666667
(14, 05 Hour ) 10.890781700783075
(14, 07 Hour ) 10.83904360571638
(14, 09 Hour ) 10.7718711333131947
(14, 10 Hour ) 10.602714600146735
(14, 12 Hour ) 10.41851631504279
(14, 14 Hour ) 10.321049258377586
(14, 16 Hour ) 10.254235731668121
(14, 18 Hour ) 10.373986718571102
(14, 21 Hour ) 10.491063244729606
(14, 23 Hour ) 10.423191920000002
(15, 00 Hour ) 8.210851543516135
(15, 03 Hour ) 8.210851543516135
(15, 05 Hour ) 8.211017996977606
(16, 07 Hour ) 8.338631366801026
(16, 09 Hour ) 8.718023921910087
(16, 10 Hour ) 8.753439104915627
(16, 12 Hour ) 8.72614525651
(16, 14 Hour ) 8.677350302142464
(16, 16 Hour ) 8.750264015795032
(16, 18 Hour ) 8.644607281886679
(16, 21 Hour ) 8.385563703024747
(16, 23 Hour ) 8.262280058651026
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -rm -r /tmp
15/12/14 04:30:42 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /tmp
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -rm -r /pig/mapreduce/Hourly_TripStatus
15/12/14 04:30:50 INFO fs.TrashPolicyDefault: Namenode Trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /pig/mapreduce/Hourly_TripStatus
[ec2-user@ip-172-31-33-207 bin]$ ./hadoop fs -cat /pig_mapreduce/Hourly_TripStatus_Station2/part-r-00000
01 Hour 13.477989442276796
03 Hour 13.5385989010989
05 Hour 13.547053166643769
07 Hour 12.354159032612678
09 Hour 14.822418770908758
10 Hour 15.219071218795888
12 Hour 15.025056218795936
14 Hour 14.9324133975666
15 Hour 13.2561366800578
18 Hour 11.881474696588047
21 Hour 13.2005957823684692
23 Hour 13.414406158357771
[ec2-user@ip-172-31-33-207 bin]$ 
```



**Graph:**



**HIVE:**



----- Analysis : 6 -----

**Creating and Loading data to HIVE:**

Copy data at hdfs : ./hadoop fs -put /usr/local/data/ /input\_TO\_HIVE

**Status table:**

Create table to hive :

```
create table status_data(station_id int, bikes_available int, docks_available int, time String) row format  
delimited fields terminated by ',' stored as textfile;
```

Load data: LOAD DATA INPATH 'hdfs:/input\_TO\_HIVE/data/201508\_status\_data.csv' OVERWRITE INTO  
TABLE status\_data;

Select data : select \* from status\_data;

```
Time taken: 0.2 seconds
hive> create table status_data(station_id int, bikes_available int, docks_available int, time timestamp) row format delimited fields terminated by ',' stored as textfile;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDITask. AlreadyExistsException(message:Table status_data already exists)
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/data/201508_status_data.csv' OVERWRITE INTO TABLE status_data;
FAILED: SemanticException Line 1:17 Invalid path ''hdfs:/input_TO_HIVE/data/201508_status_data.csv'': No files matching path hdfs://localhost:9000/input_TO_HIVE/data/201508_status_data.csv
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/data/201508_status_data.csv' OVERWRITE INTO TABLE status_data;
Loading data to table default.status_data
Table default.status_data stats: [numFiles=1, numRows=0, totalSize=1087241884, rawDataSize=0]
OK
Time taken: 0.326 seconds
hive> select * from status_data
> ;
Query ID = root_20151214061212_33cd63dd-1fa4-48d1-878d-bc733d4e7222
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
Hadoop job information for Stage-1: number of mappers: 0; number of reducers: 0
2015-12-14 06:12:03,732 Stage-1 map = 0%,  reduce = 0%
hive>
```

### Weather table:

```
./hadoop fs -cp /input/data/201508_weather_data.csv /input_TO_HIVE
```

```
create table weather_data(PDT String,MaxTemperatureF int,MeanTemperatureF int,MinTemperatureF int,MaxDew int,MeanDew int,MinDew int,MaxHumidity int, MeanHumidity int, MinHumidity int, MaxSea int, MeanSea int, MinSea int, MaxVisibilityMiles int, MeanVisibilityMiles int, MinVisibilityMiles int, MaxWind int, MeanWind int, MaxGust int,PrecipitationIn int, CloudCover int, Events int, WindDirDegrees int,Zip int) row format delimited fields terminated by ',' stored as textfile;
```

```
LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_weather_data.csv' OVERWRITE INTO TABLE weather_data;
```

```
Time taken: 2.23 seconds, Fetched: 1 row(s)
hive> create table weather_data(PDT String,MaxTemperatureF int,MeanTemperatureF int,MinTemperatureF int,MaxDew int,MeanDew int,MinDew int,MaxHumidity int, MeanHumidity int, MinHumidity int, MaxSea int, MeanSea int, MinSea int, MaxVisibilityMiles int, MeanVisibilityMiles int, MinVisibilityMiles int, MaxWind int, MeanWind int, MaxGust int,PrecipitationIn int, CloudCover int, Events int, WindDirDegrees int,Zip int) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.253 seconds
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/data/201508_weather_data.csv' OVERWRITE INTO TABLE weather_data;
FAILED: SemanticException Line 1:17 Invalid path ''hdfs:/input_TO_HIVE/data/201508_weather_data.csv'': No files matching path hdfs://localhost:9000/input_TO_HIVE/data/201508_weather_data.csv
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_weather_data.csv' OVERWRITE INTO TABLE weather_data;
Loading data to table default.weather_data
Table default.weather_data stats: [numFiles=1, numRows=0, totalSize=158238, rawDataSize=0]
OK
Time taken: 0.372 seconds
hive>
```

### Trip\_data

```
./hadoop fs -cp /input/data/201508_trip_data.csv /input_TO_HIVE
```

```
create table trip_data(Trip_ID int,Duration int,Start_Date String,Start_Station String,Start_Terminal String,End_Date String,End_Station String,End_Terminal String,Bike String,Subscriber_Type String,Zip int) row format delimited fields terminated by ',' stored as textfile;
```

```
LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_trip_data.csv' OVERWRITE INTO TABLE trip_data;
```

```
hive> create table trip_data(Trip_ID int,Duration int,Start_Date String,Start_Station String,Start_Terminal String,End_Date String,End_Station String,End_Terminal String,Bike String,Subscriber_Type String,Zip int) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.052 seconds
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_trip_data.csv' OVERWRITE INTO TABLE trip_data;
FAILED: SemanticException Line 1:17 Invalid path ''hdfs:/input_TO_HIVE/201508_trip_data.csv'': No files matching path hdfs://localhost:9000/input_TO_HIVE/201508_trip_data.csv
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_trip_data.csv' OVERWRITE INTO TABLE trip_data;
FAILED: SemanticException Line 1:17 Invalid path ''hdfs:/input_TO_HIVE/201508_trip_data.csv'': No files matching path hdfs://localhost:9000/input_TO_HIVE/201508_trip_data.csv
hive> LOAD DATA INPATH 'hdfs:/input_TO_HIVE/201508_trip_data.csv' OVERWRITE INTO TABLE trip_data;
Loading data to table default.trip_data
Table default.trip_data stats: [numFiles=1, numRows=0, totalSize=43012526, rawDataSize=0]
OK
Time taken: 0.207 seconds
hive>
```

----- Analysis : 7 -----

**Ratio of subscriber to customer in trip data:**

```
hive> Select subscriber_type , COUNT(*) as trip_count from trip_data group by subscriber_type order by trip_count desc;
Query ID = ec2-user_20151215052005_84452742-3d78-4cec-9b41-736c5f3cc956
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:20:07,136 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1529005141_0007
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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:20:08,300 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1787768458_0008
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 946308340 HDFS Write: 0 SUCCESS
Stage-Stage-2: HDFS Read: 946308340 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Subscriber      310217
Customer       43935
Time taken: 2.384 seconds, Fetched: 2 row(s)
```

**Query :**

```
Select subscriber_type , COUNT(*) as trip_count from trip_data group by subscriber_type order by trip_count desc;
```

**Output:**

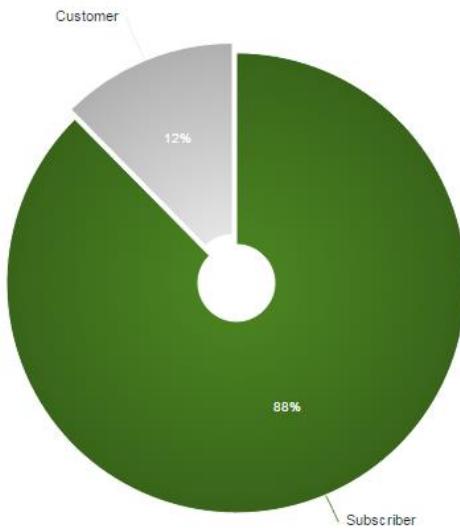
Subscriber 310217

Customer 43935

**Graph:**

← → C [localhost:8383/Test/js/Subscriber\\_CustomerRatio.html](http://localhost:8383/Test/js/Subscriber_CustomerRatio.html)

Customer To Subscriber Ratio



----- Analysis : 8 -----

**Top 10 zip of customer type by total\_trip:**

```
Time taken: 0.363 seconds, Fetched: 10 row(s)
hive> Select zip, subscriber_type , COUNT(*) as trip_count from trip data group by zip,subscriber_type  order by trip_count desc limit 10;
Query ID = ec2-user_20151215052447_99ea8c0a-2329-4bfd-bcfa-3e4e9389460c
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:24:48,460 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2059511318_0011
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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:24:49,612 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1342277069_0012
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1118358444 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 1118358444 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
94107  Subscriber      45819
94105  Subscriber      19915
94133  Subscriber      15868
94103  Subscriber      14446
94111  Subscriber      10697
94102  Subscriber      9696
NULL    Customer        6929
94109  Subscriber      6029
95112  Subscriber      4488
94403  Subscriber      4124
Time taken: 2.363 seconds, Fetched: 10 row(s)
```

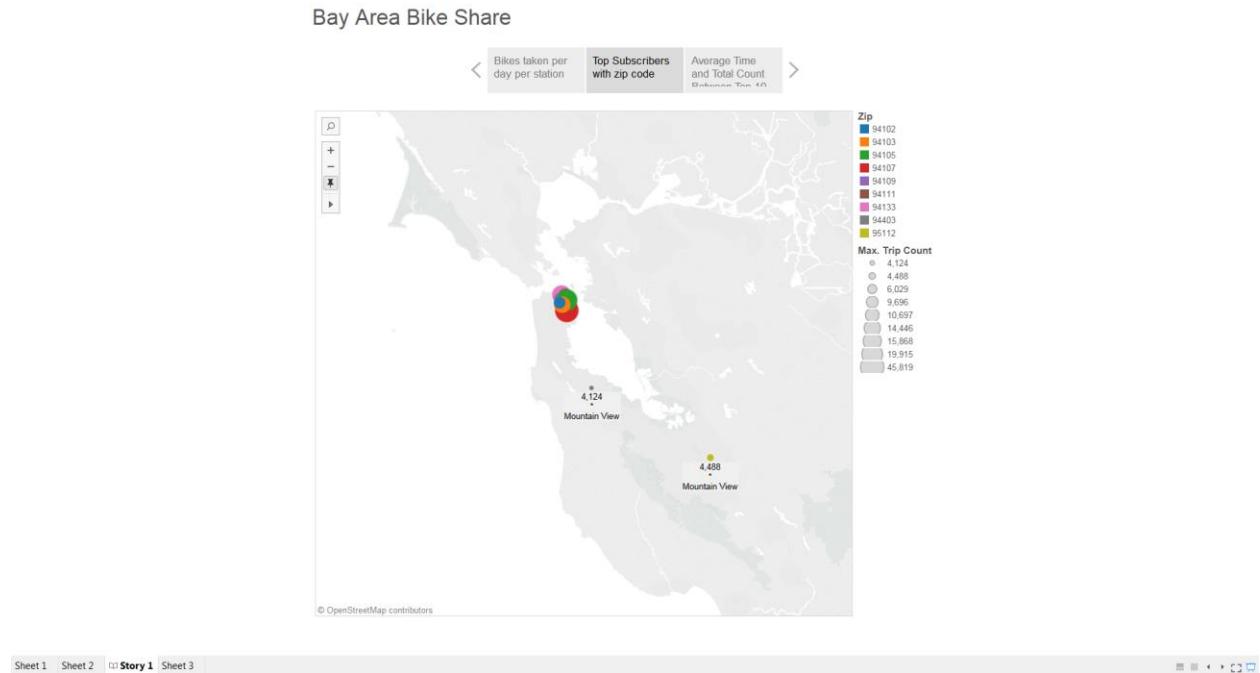
**Query:**

```
Select zip, subscriber_type , COUNT(*) as trip_count from trip_data group by zip,subscriber_type order by trip_count desc limit 10;
```

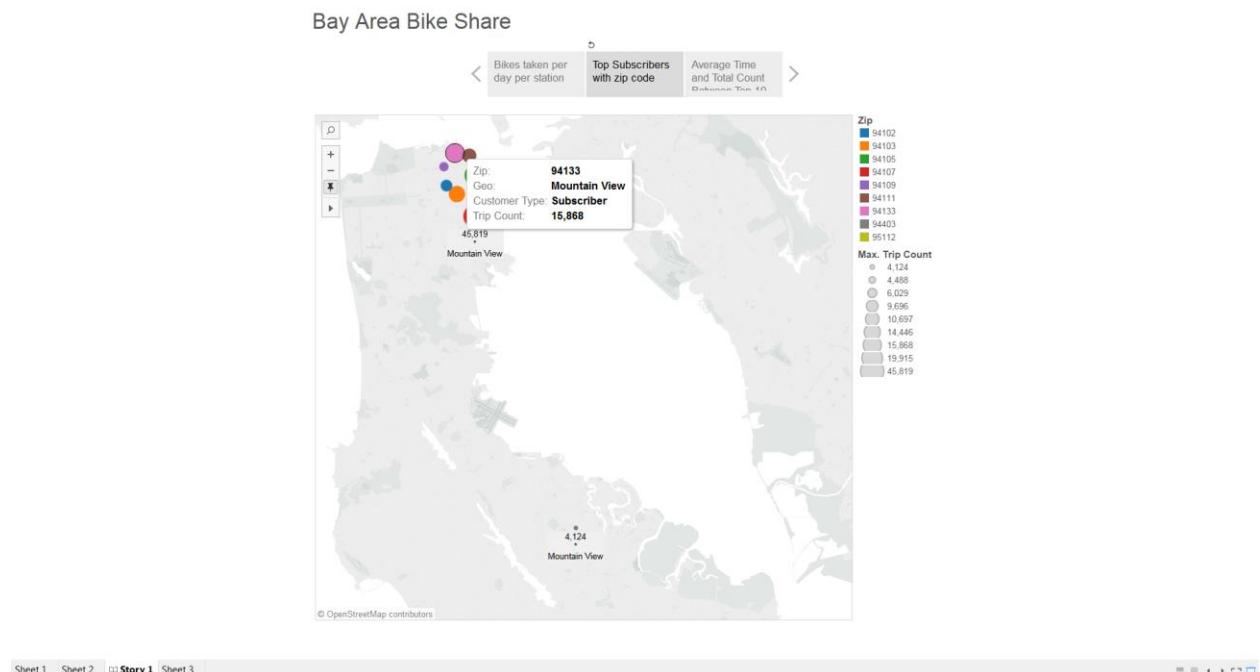
**Output:**

94107	Subscriber	45819
94105	Subscriber	19915
94133	Subscriber	15868
94103	Subscriber	14446
94111	Subscriber	10697
94102	Subscriber	9696
NULL	Customer	6929
94109	Subscriber	6029
95112	Subscriber	4488
94403	Subscriber	4124

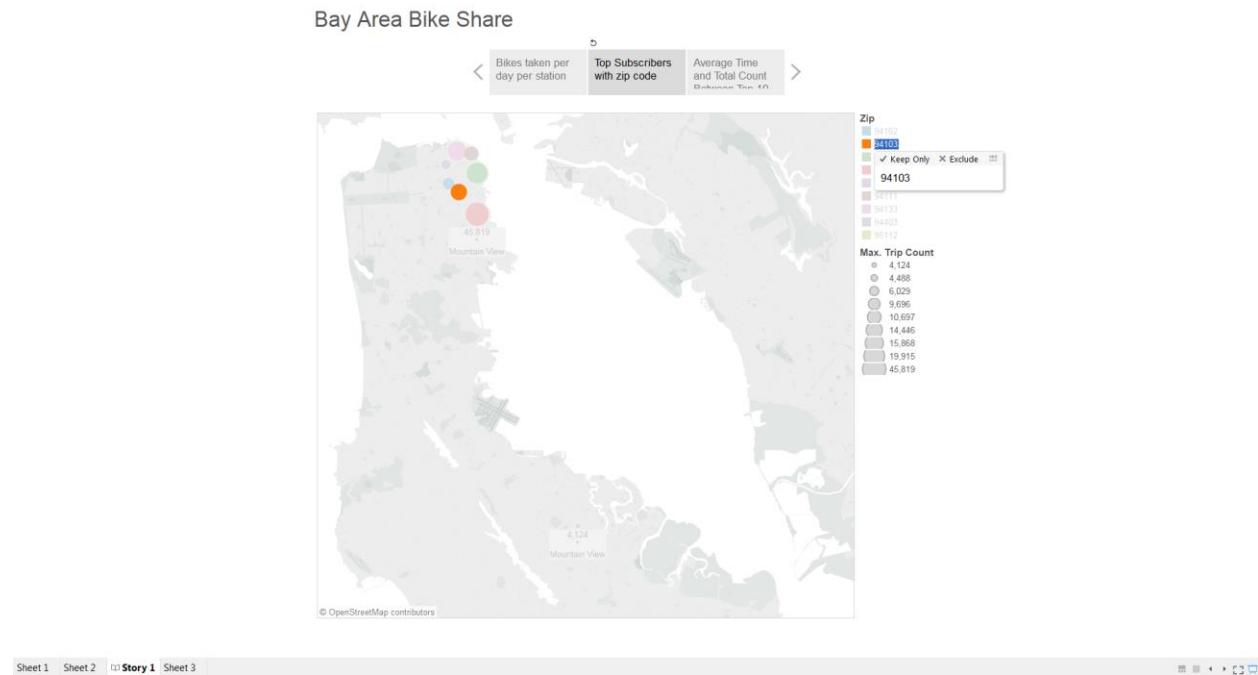
**Graph:**



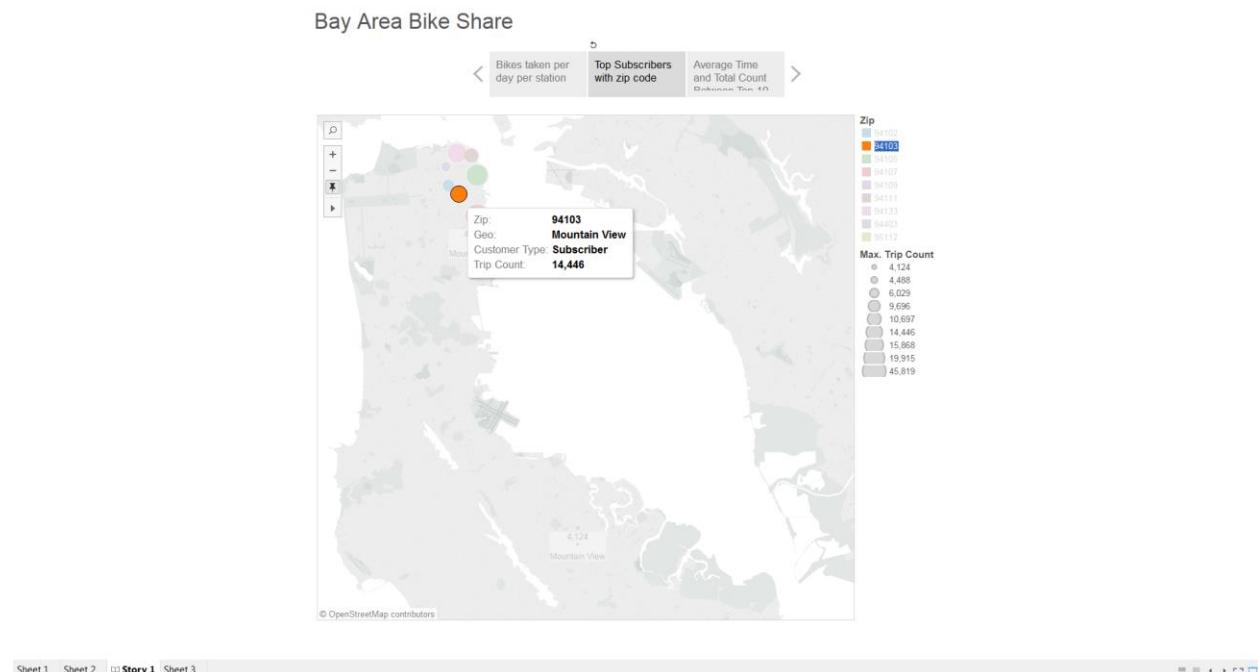
Zoom



Select Zip:

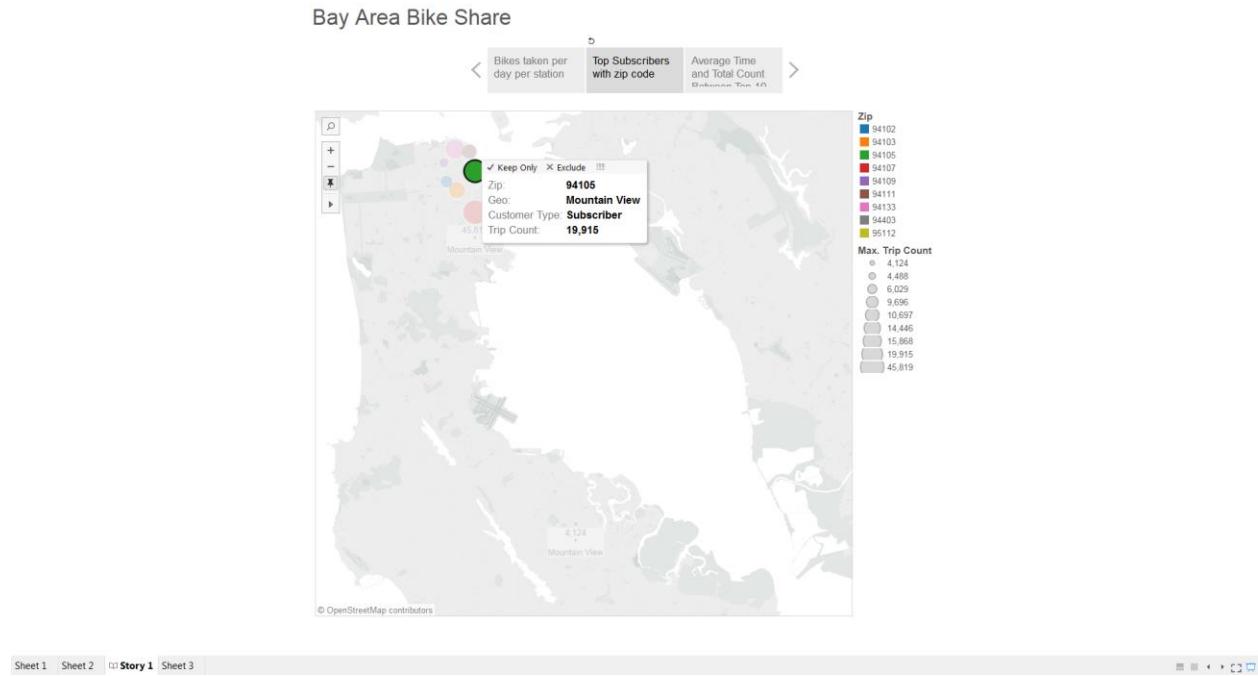


Sheet 1 Sheet 2 Story 1 Sheet 3



Sheet 1 Sheet 2 Story 1 Sheet 3

Joined with station table for landmark name:



### ----- Analysis : 9 -----

#### Analysis:

Most used start terminal with total trip count

#### Query:

```
Select start_station , COUNT(*) as trip_count from trip_data group by start_station order by trip_count desc limit 10;
```

**SS:**

```
hive> Select start_station , COUNT(*) as trip_count from trip_data group by start_station order by trip_count desc limit 10;
Query ID = ec2-user_20151215050516_665f4201-eb68-4db3-93b2-3cf1f3eb7639
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:05:17,657 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1414676948_0003
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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2015-12-15 05:05:18,850 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1200427644_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 688200416 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 688200416 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
San Francisco Caltrain (Townsend at 4th)      26304
San Francisco Caltrain 2 (330 Townsend) 21758
Harry Bridges Plaza (Ferry Building)      17255
Temporary Transbay Terminal (Howard at Beale) 14436
Embarcadero at Sansome 14158
2nd at Townsend 14026
Townsend at 7th 13752
Steuart at Market      13687
Market at 10th 11885
Market at Sansome      11431
Time taken: 2.44 seconds, Fetched: 10 row(s)
```

**Output:**

San Francisco Caltrain (Townsend at 4th) 26304

San Francisco Caltrain 2 (330 Townsend) 21758

Harry Bridges Plaza (Ferry Building) 17255

Temporary Transbay Terminal (Howard at Beale) 14436

Embarcadero at Sansome 14158

2nd at Townsend 14026

Townsend at 7th 13752

Steuart at Market 13687

Market at 10th 11885

Market at Sansome 11431



----- Analysis : 10 Load data into HBASE using PIG script -----

**Start HBASE:**

```
CMPS          7.4G  2.2G  3.3G  50% /dev/simr
[ec2-user@ip-172-31-33-207 bin]$ jps
12078 NameNode
12214 DataNode
12560 ResourceManager
11322 HQuorumPeer
11431 HMaster
13039 JobHistoryServer
12672 NodeManager
11568 HRegionServer
13364 Main
12393 SecondaryNameNode
13560 Jps
[ec2-user@ip-172-31-33-207 bin]$
```

**Create table 'trip\_data':**

```
create 'trip_data','trip_info','start_details','end_details','customer_details'
```

```
hbase(main):004:0> create 'trip_data','trip_info','start_details','end_details','customer_details'
0 row(s) in 0.2280 seconds

=> Hbase::Table - trip_data
hbase(main):005:0> list
TABLE
trip_data
1 row(s) in 0.0070 seconds

=> ["trip_data"]
hbase(main):006:0> describe'trip_data'
Table trip_data is ENABLED
trip_data
COLUMN FAMILIES DESCRIPTION
(NAME => 'customer_details', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true')
(NAME => 'end_details', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true')
(NAME => 'start_details', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true')
(NAME => 'trip_info', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true')

4 row(s) in 0.0210 seconds

hbase(main):007:0>
```

**Insert data to hbase using “/usr/local/pig\_script/loadToHbase.pig” script:**

trip table:-----

```
trip_data = LOAD 'hdfs:/input/data/201508_trip_data.csv' USING PigStorage(',') AS
(trip_id,int,duration:int,start_date:chararray,start_station:chararray,start_terminal:int,
end_date:chararray,end_station:chararray,end_terminal:int,bike:int,subscriber:chararray,zip:int);
```

STORE trip\_data INTO 'hbase://trip\_data' USING

```
org.apache.pig.backend.hadoop.hbase.HBaseStorage()  
'trip_info:duration,  
start_details:start_date,  
start_details:start_station,  
start_details:start_terminal,  
end_details:end_date,  
end_details:end_station,  
end_details:end_terminal,  
customer_details:bike,  
customer_details:subscriber,  
customer_details:zip','loadKey true');
```

**Run pig script and its output:**

```
[ec2-user@ip-172-31-33-207 pig_script]$ ll  
[ec2-user@ip-172-31-33-207 pig_script]$ ll  
total 16  
-rwxrwxrwx 1 root      root      637 Dec 16 02:46 loadToHbase.pig  
-rwxrwxrwx 1 ec2-user  ec2-user   716 Dec 13 06:14 Sorted_totalTrip_estimatedDuration.pig  
-rwxrwxrwx 1 ec2-user  ec2-user   394 Dec 13 05:50 station_per_landmark.pig  
-rwxrwxrwx 1 ec2-user  ec2-user   631 Dec 13 06:13 totalTripbetweenStations.pig  
[ec2-user@ip-172-31-33-207 pig_script]$ pig loadToHbase.pig
```

```

[ec2-user@ip-172-31-33-207 ~]$ ./pig_script
-rwxrwxrwx 1 ec2-user ec2-user 716 Dec 13 06:14 Sorted_totalTrip_estimatedDuration.pig
-rwxrwxrwx 1 ec2-user ec2-user 394 Dec 13 05:58 station_per_landmark.pig
-rwxrwxrwx 1 ec2-user ec2-user 631 Dec 13 06:13 totalTripbetweenStations.pig
[ec2-user@ip-172-31-33-207 ~]$ pig script1.pig loadToHbase.pig
[5/12/16 02:47:17] INFO org.apache.pig.LoadToHBase: Trying ExecType : LOCAL
[5/12/16 02:47:17] INFO org.apache.pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
[5/12/16 02:47:17] WARN org.Main: Need write permission in the directory: /usr/local/lib/pig_log to create log file.
2015-12-16 02:47:17,290 [main] INFO org.apache.pig.Main - Apache Pig version 0.14.0-SNAPSHOT (r: unknown) compiled Dec 10 2015, 03:09:26
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/hadoop-2.7.1/share/hadoop/common/lib/slf4j-1.7.10.jar!/org/slf4j.impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/base-1.0.2/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2015-12-16 02:47:18,140 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2015-12-16 02:47:18,140 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:18,140 [main] INFO org.apache.hadoop.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2015-12-16 02:47:18,065 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:18,211 [main] INFO org.apache.hadoop.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:19,342 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:19,358 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2015-12-16 02:47:19,387 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:19,396 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2015-12-16 02:47:19,436 [main] INFO org.apache.newplan.logical.optimizer.LogicalPlanOptimizer - (RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSplitter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter])
2015-12-16 02:47:19,604 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2015-12-16 02:47:19,630 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:19,650 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2015-12-16 02:47:19,650 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2015-12-16 02:47:19,650 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-12-16 02:47:19,662 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId=
2015-12-16 02:47:19,683 [main] INFO org.apache.pig.tools.pigstats.MRScriptState - Pig script settings are added to the job
2015-12-16 02:47:19,706 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Map reduce.markrset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markrset.buf
fer.percent
2015-12-16 02:47:19,712 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markrset.buffer.percent is not set, set to default 0.3
2015-12-16 02:47:19,712 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2015-12-16 02:47:19,714 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2015-12-16 02:47:19,917 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/lib/pig-0.14.0/pig-0.14.0-SNAPSHOT-core-h2.jar
ar to DistributedCache through /tmp/temp-2048781721/tmp1183873083/pig-0.14.0-SNAPSHOT-core-h2.jar
2015-12-16 02:47:19,952 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/lib/hadoop-2.7.1/share/hadoop/common/lib/qua
va-1.0.2.jar to DistributedCache through /tmp/temp-2048781721/tmp2090095509/guava-11.0.2.jar
2015-12-16 02:47:19,952 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - protobuffer-Java-2.5.0.jar to DistributedCache through /tmp/temp-2048781721/tmp1286630549/protobuf-Java-2.5.0.jar
2015-12-16 02:47:19,993 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/lib/hbase-client-1.0.2.jar to
DistributedCache through /tmp/temp-2048781721/tmp-667299687/hbase-client-1.0.2.jar
2015-12-16 02:47:20,027 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/lib/pig-0.14.0/lib/hbase-protocol-1.0.2.jar to
DistributedCache through /tmp/temp-2048781721/tmp372924671/hbase-protocol-1.0.2.jar

```

9:48 PM  
12/15/2015

```

[ec2-user@ip-172-31-33-207 ~]$ ./pig_script
2015-12-16 02:47:42,941 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 49% complete
2015-12-16 02:47:42,941 [main] INFO org.apache.hadoop.mapred.MapTask - Running jobs are [job_local1913663099_0001]
2015-12-16 02:47:42,971 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task attempt local1913663099_0001_m_000000 is done. And is in the process of committing
2015-12-16 02:47:42,978 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task attempt local1913663099_0001_m_000000 is done.
2015-12-16 02:47:42,978 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task attempt local1913663099_0001_m_000000* done.
2015-12-16 02:47:42,978 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task attempt_local1913663099_0001_m_000000
2015-12-16 02:47:42,979 [Thread-38] INFO org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2015-12-16 02:47:42,979 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:42,979 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:46,450 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2015-12-16 02:47:46,450 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2015-12-16 02:47:46,451 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:46,479 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2015-12-16 02:47:46,481 [main] INFO org.apache.pig.tools.pigstats.SimpleFigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.7.1 0.14.0-SNAPSHOT ec2-user 2015-12-16 02:47:19 2015-12-16 02:47:46 UNKNOWN

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_local1913663099_0001 1 0 n/a n/a n/a 0 0 0 trip_data MAP_ONLY hbase://trip_data,
Input(s):
Successfully read 354152 records (6332571 bytes) from: "hdfs://input/data/201508_trip_data.csv"
Output(s):
Successfully stored 354152 records (20313045 bytes) in: "hbase://trip_data"
Counters:
Total records written : 354152
Total bytes written : 20313045
Spillable Memory Manager spill count : 0
total bags proactively spilled: 0
total records proactively spilled: 0
Job DAG:
job_local1913663099_0001

2015-12-16 02:47:46,482 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:46,483 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:46,484 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2015-12-16 02:47:46,490 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning TOO_LARGE_FOR_INT 2 time(s).
2015-12-16 02:47:46,490 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 6899 time(s)
2015-12-16 02:47:46,490 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2015-12-16 02:47:46,507 [main] INFO org.apache.Main - Pig script completed in 29 seconds and 361 milliseconds (29361 ms)
[ec2-user@ip-172-31-33-207 ~]$
```

9:49 PM  
12/15/2015

**Verify 'trip\_data' hbase table:**

```
ec2-user@ip-172-31-33-207:/usr/local/lib/hbase-1.0.2/bin
hbase(main):007:0> status 'trip_data'
1 servers, 0 dead, 3.0000 average load
hbase(main):008:0> count 'trip_data'
Current count: 1000, row: 434407
Current count: 2000, row: 435815
Current count: 3000, row: 437162
Current count: 4000, row: 438574
Current count: 5000, row: 439834
Current count: 6000, row: 441155
Current count: 7000, row: 442517
Current count: 8000, row: 443948
Current count: 9000, row: 445368
Current count: 10000, row: 446760
Current count: 11000, row: 448090
Current count: 12000, row: 449433
Current count: 13000, row: 450840
Current count: 14000, row: 452164
Current count: 15000, row: 453558
Current count: 16000, row: 454880
Current count: 17000, row: 456277
Current count: 18000, row: 457659
Current count: 19000, row: 459003
Current count: 20000, row: 460451
Current count: 21000, row: 461856
Current count: 22000, row: 463289
Current count: 23000, row: 464717
Current count: 24000, row: 466160
Current count: 25000, row: 467633
Current count: 26000, row: 469012
Current count: 27000, row: 470417
Current count: 28000, row: 471792
Current count: 29000, row: 473229
Current count: 30000, row: 474674
Current count: 31000, row: 476087
Current count: 32000, row: 477570
Current count: 33000, row: 478969
Current count: 34000, row: 480410
Current count: 35000, row: 481729
Current count: 36000, row: 483050
Current count: 37000, row: 484773
Current count: 38000, row: 486400
Current count: 39000, row: 488007
Current count: 40000, row: 489479
Current count: 41000, row: 490903
Current count: 42000, row: 492302
Current count: 43000, row: 493763
Current count: 44000, row: 495171
Current count: 45000, row: 496608
9:53 PM
12/15/2015
```

```
ec2-user@ip-172-31-33-207:/usr/local/lib/hbase-1.0.2/bin
Current count: 309000, row: 854188
Current count: 310000, row: 855468
Current count: 311000, row: 856836
Current count: 312000, row: 858138
Current count: 313000, row: 859470
Current count: 314000, row: 860802
Current count: 315000, row: 862098
Current count: 316000, row: 863366
Current count: 317000, row: 864700
Current count: 318000, row: 865953
Current count: 319000, row: 867272
Current count: 320000, row: 868591
Current count: 321000, row: 869927
Current count: 322000, row: 871185
Current count: 323000, row: 872495
Current count: 324000, row: 873816
Current count: 325000, row: 875192
Current count: 326000, row: 876536
Current count: 327000, row: 877852
Current count: 328000, row: 879232
Current count: 329000, row: 880572
Current count: 330000, row: 881906
Current count: 331000, row: 883203
Current count: 332000, row: 884538
Current count: 333000, row: 885871
Current count: 334000, row: 887161
Current count: 335000, row: 888495
Current count: 336000, row: 889803
Current count: 337000, row: 891167
Current count: 338000, row: 892386
Current count: 339000, row: 893700
Current count: 340000, row: 895005
Current count: 341000, row: 896275
Current count: 342000, row: 897644
Current count: 343000, row: 898957
Current count: 344000, row: 900247
Current count: 345000, row: 901517
Current count: 346000, row: 902749
Current count: 347000, row: 904045
Current count: 348000, row: 905396
Current count: 349000, row: 906710
Current count: 350000, row: 907967
Current count: 351000, row: 909270
Current count: 352000, row: 910545
Current count: 353000, row: 911876
Current count: 354000, row: 913222
354152 row(s) in 19.4190 seconds
=> 354152
hbase(main):009:0>
9:53 PM
12/15/2015
```

## Scan table:

```
ec2-user@ip-172-31-33-207:/usr/local/lib/hbase-1.0.2/bin
hbase(main):013:0> scan 'trip_data',(LIMIT=>10)
ROW                                     COLUMN+CELL
432947                                     column:customer.details:bike, timestamp=1450234062806, value=318
432947                                     column:customer.details:subscriber, timestamp=1450234062806, value=Customer
432947                                     column:customer.details:zip, timestamp=1450234062806, value=32
432947                                     column:end.details:end_date, timestamp=1450234062806, value=9/1/2014 0:15
432947                                     column:end.details:end_station, timestamp=1450234062806, value=5th at Howard
432947                                     column:end.details:end_terminal, timestamp=1450234062806, value=57
432947                                     column:start.details:start_date, timestamp=1450234062806, value=9/1/2014 0:05
432947                                     column:start.details:start_station, timestamp=1450234062806, value=South Van Ness at Market
432947                                     column:start.details:start_terminal, timestamp=1450234062806, value=66
432948                                     column:customer.details:bike, timestamp=1450234062806, value=569
432948                                     column:customer.details:subscriber, timestamp=1450234062806, value=Subsriber
432948                                     column:customer.details:zip, timestamp=1450234062806, value=461
432949                                     column:customer.details:bike, timestamp=1450234062806, value=Customer
432949                                     column:customer.details:zip, timestamp=1450234062806, value=32
432948                                     column:end.details:end_date, timestamp=1450234062806, value=9/1/2014 0:15
432948                                     column:end.details:end_station, timestamp=1450234062806, value=5th at Howard
432948                                     column:end.details:end_terminal, timestamp=1450234062806, value=57
432948                                     column:start.details:start_date, timestamp=1450234062806, value=9/1/2014 0:05
432948                                     column:start.details:start_station, timestamp=1450234062806, value=South Van Ness at Market
432948                                     column:start.details:start_terminal, timestamp=1450234062806, value=66
432948                                     column:trip.info:duration, timestamp=1450234062806, value=568
432949                                     column:customer.details:bike, timestamp=1450234062806, value=466
432949                                     column:customer.details:subscriber, timestamp=1450234062806, value=Customer
432949                                     column:customer.details:zip, timestamp=1450234062806, value=32
432949                                     column:end.details:end_date, timestamp=1450234062806, value=9/1/2014 0:14
432949                                     column:end.details:end_station, timestamp=1450234062806, value=5th at Howard
432949                                     column:end.details:end_terminal, timestamp=1450234062806, value=57
432949                                     column:start.details:start_date, timestamp=1450234062806, value=9/1/2014 0:05
432949                                     column:start.details:start_station, timestamp=1450234062806, value=South Van Ness at Market
432949                                     column:start.details:start_terminal, timestamp=1450234062806, value=66
432949                                     column:trip.info:duration, timestamp=1450234062806, value=538
432950                                     column:customer.details:bike, timestamp=1450234062806, value=259
432950                                     column:customer.details:subscriber, timestamp=1450234062806, value=Customer
432950                                     column:customer.details:zip, timestamp=1450234062806, value=4100
432950                                     column:end.details:end_date, timestamp=1450234062806, value=9/1/2014 5:08
432950                                     column:end.details:end_station, timestamp=1450234062806, value=Golden Francisco Caltrain (Townsend at 4th)
432950                                     column:end.details:end_terminal, timestamp=1450234062806, value=70
432950                                     column:start.details:start_date, timestamp=1450234062806, value=9/1/2014 3:16
432950                                     column:start.details:start_station, timestamp=1450234062806, value=Harry Bridges Plaza (Ferry Building)
432950                                     column:start.details:start_terminal, timestamp=1450234062806, value=50
432950                                     column:trip.info:duration, timestamp=1450234062806, value=672
432951                                     column:customer.details:bike, timestamp=1450234062806, value=335
432951                                     column:customer.details:subscriber, timestamp=1450234062806, value=Subscriber
432951                                     column:customer.details:zip, timestamp=1450234062806, value=94118
432951                                     column:end.details:end_date, timestamp=1450234062806, value=9/1/2014 4:32
432951                                     column:end.details:end_station, timestamp=1450234062806, value=Townsend at 7th
432951                                     column:end.details:end_terminal, timestamp=1450234062806, value=65
432951                                     column:start.details:start_date, timestamp=1450234062806, value=9/1/2014 4:21
```

```
ec2-user@ip-172-31-33-207:/usr/local/lib/hbase-1.0.2/bin
432951 column=start_details:start_terminal, timestamp=1450234062806, value=39
432951 column=trip_info:duration, timestamp=1450234062806, value=619
432952 column=customer_details:bike, timestamp=1450234062806, value=292
432952 column=customer_details:zip, timestamp=1450234062806, value=Subscriber
432952 column=customer_details:zip, timestamp=1450234062806, value=94102
432952 column=end_details:end_date, timestamp=1450234062806, value=9/1/2014 5:03
432952 column=end_details:end_station, timestamp=1450234062806, value=Civic Center BART (7th at Market)
432952 column=end_details:end_terminal, timestamp=1450234062806, value=72
432952 column=start_details:start_date, timestamp=1450234062806, value=9/1/2014 4:59
432952 column=start_details:start_station, timestamp=1450234062806, value=South Van Ness at Market
432952 column=start_details:start_terminal, timestamp=1450234062806, value=66
432952 column=trip_info:duration, timestamp=1450234062806, value=240
432957 column=customer_details:bike, timestamp=1450234062806, value=561
432957 column=customer_details:zip, timestamp=1450234062806, value=Subscriber
432957 column=customer_details:zip, timestamp=1450234062806, value=94112
432957 column=end_details:end_date, timestamp=1450234062806, value=9/1/2014 6:00
432957 column=end_details:end_terminal, timestamp=1450234062806, value=Stewart at Market
432957 column=end_details:end_station, timestamp=1450234062806, value=74
432957 column=start_details:start_date, timestamp=1450234062806, value=9/1/2014 5:54
432957 column=start_details:start_station, timestamp=1450234062806, value=Yerba Buena Center of the Arts (3rd @ Howard)
432957 column=start_details:start_terminal, timestamp=1450234062806, value=68
432957 column=trip_info:duration, timestamp=1450234062806, value=398
432959 column=customer_details:bike, timestamp=1450234062806, value=617
432959 column=customer_details:zip, timestamp=1450234062806, value=94132
432959 column=customer_details:zip, timestamp=1450234062806, value=9/1/2014 7:05
432959 column=end_details:end_date, timestamp=1450234062806, value=Market at Sansome
432959 column=end_details:end_terminal, timestamp=1450234062806, value=7
432959 column=end_details:end_station, timestamp=1450234062806, value=9/1/2014 6:58
432959 column=start_details:start_station, timestamp=1450234062806, value=Market at 10th
432959 column=start_details:start_terminal, timestamp=1450234062806, value=67
432959 column=trip_info:duration, timestamp=1450234062806, value=441
432960 column=customer_details:bike, timestamp=1450234062806, value=56
432960 column=customer_details:subscriber, timestamp=1450234062806, value=Customer
432960 column=customer_details:zip, timestamp=1450234062806, value=95112
432960 column=end_details:end_date, timestamp=1450234062806, value=9/1/2014 8:38
432960 column=end_details:end_station, timestamp=1450234062806, value=Japantown
432960 column=end_details:end_terminal, timestamp=1450234062806, value=9
432960 column=start_details:start_date, timestamp=1450234062806, value=9/1/2014 7:03
432960 column=start_details:start_station, timestamp=1450234062806, value=Japantown
432960 column=start_details:start_terminal, timestamp=1450234062806, value=9
432960 column=trip_info:duration, timestamp=1450234062806, value=557
432960 column=customer_details:bike, timestamp=1450234062806, value=496
432964 column=customer_details:subscriber, timestamp=1450234062806, value=Subscriber
432964 column=customer_details:zip, timestamp=1450234062806, value=94105
432964 column=end_details:end_date, timestamp=1450234062806, value=9/1/2014 7:35
432964 column=end_details:end_station, timestamp=1450234062806, value=Embarcadero at Folsom
432964 column=end_details:end_terminal, timestamp=1450234062806, value=51
432964 column=start_details:start_date, timestamp=1450234062806, value=9/1/2014 7:32
432964 column=start_details:start_station, timestamp=1450234062806, value=Embarcadero at Bryant
```

----- Analysis : 11 Scan single data into HBASE using JAVA -----

```
Trip_data.java -----  
  
/*  
  
 * To change this license header, choose License Headers in Project Properties.  
  
 * To change this template file, choose Tools | Templates  
  
 * and open the template in the editor.  
  
 */  
  
package trip_package;  
  
import java.io.FileNotFoundException;  
  
import java.io.IOException;  
  
import java.util.Arrays;  
  
import org.apache.hadoop.conf.Configuration;  
  
import org.apache.hadoop.hbase.HBaseConfiguration;  
  
  
import org.apache.hadoop.hbase.client.Get;  
  
import org.apache.hadoop.hbase.client.HTable;  
  
import org.apache.hadoop.hbase.client.Result;  
  
  
/**  
  
 *  
  
 * @author ubuntu  
  
 */  
  
public class Trip_data {  
  
    /**  
  
     * @param args the command line arguments  
  
     * @throws java.io.IOException  
  
     * @throws java.io.FileNotFoundException  
  
     */  
  
  
    public static void main(String[] args) throws IOException ,FileNotFoundException{  
  
        try{  
    
```

```
Configuration config = HBaseConfiguration.create();

HTable htable = new HTable(config, "trip_data");

String rowKey ="913454";

Get get = new Get(rowKey.getBytes());

Result rs =htable.get(get);

byte[] row = rs.getValue(Bytes.toBytes("trip_info"), Bytes.toBytes("duration"));

System.out.println(row);

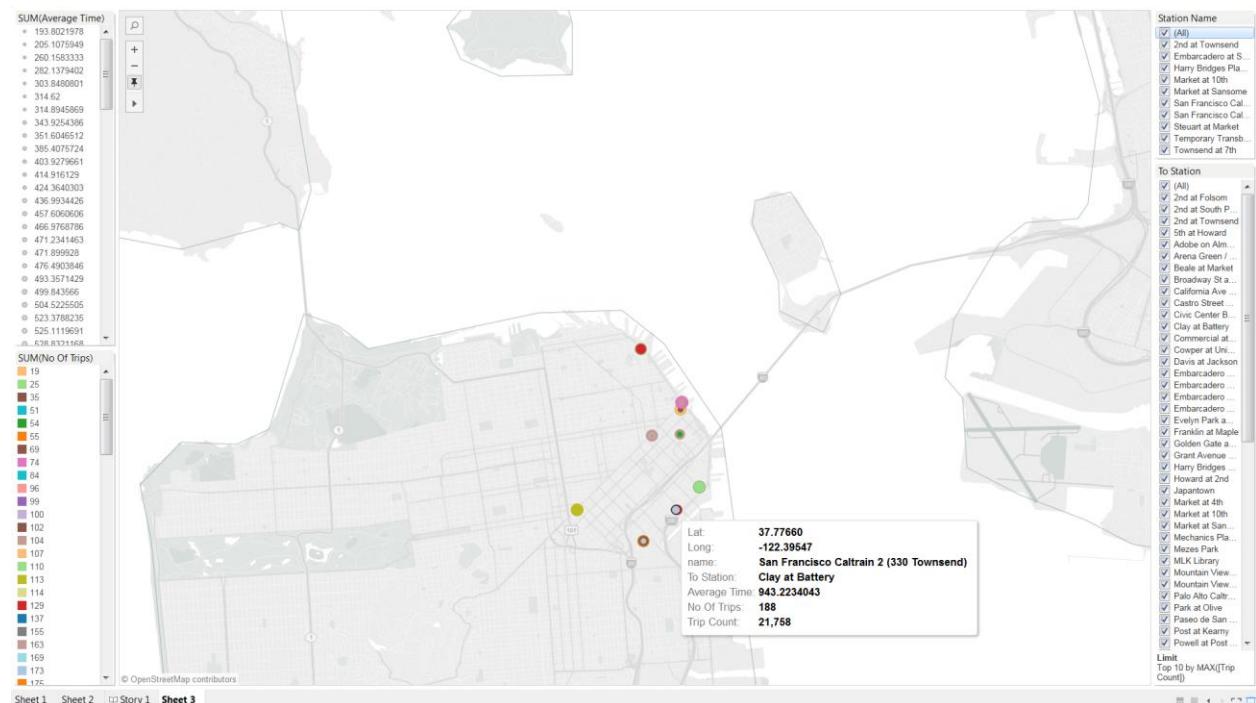
}catch (IOException e){

}

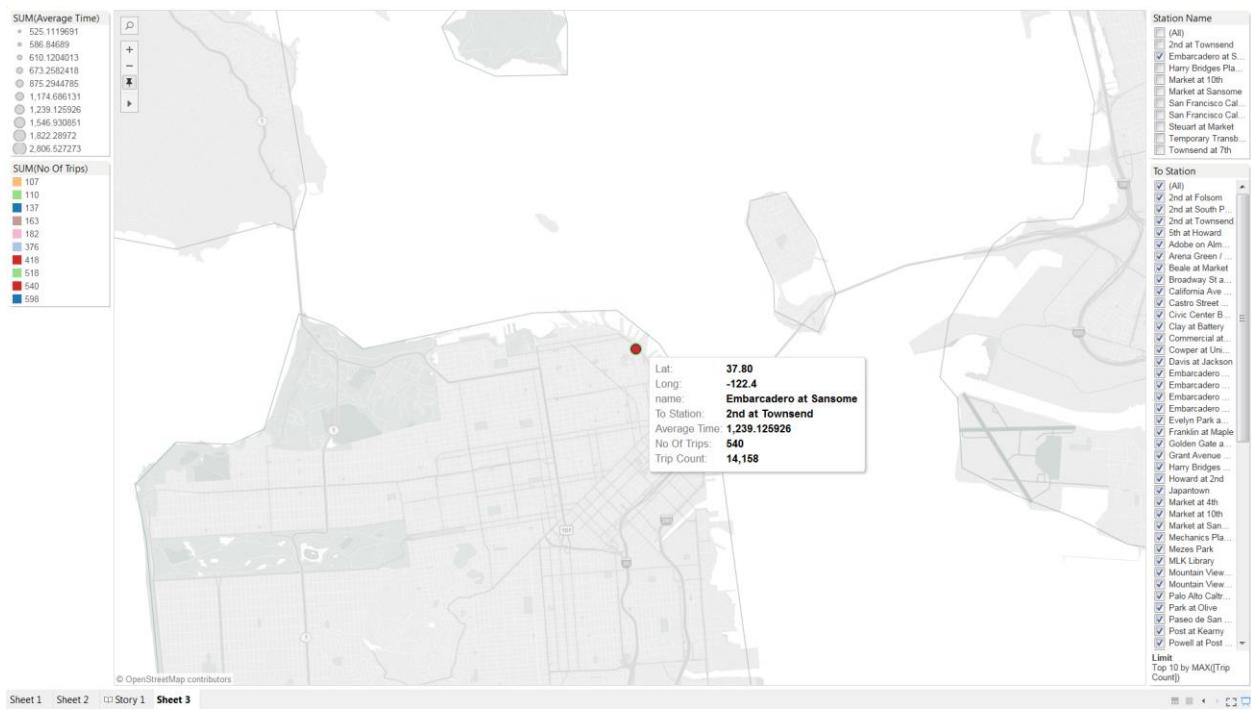
}
```

## **Graphs:**

## Top 10 stations:



### From A to B:



### One location multiple destinations:

