Problem 1

1.1 Create a schema for the dataset in Hive. I have created a concrete structure describing all the required fields.

- **Source Code:**

```
CREATE TABLE july (
   `Source_IP` STRING,
   `Time_Stamp` STRING,
   `HTTP_Method` STRING,
   `Request_URL` STRING,
   `HTTP_Protocol` STRING,
   `Status_Code` STRING,
   `Response_Bytes` STRING
)
ROW FORMAT
SERDE ‘org.apache.hadoop.hive.serde2.RegexSerDe’
WITH SERDEPROPERTIES (
      “input.regex"= "(^.*) - - \\[(.*)\\] \\\"([A-Z]*|[^\\\" ]*) ([^HTTP]*) ([^\\\"]*)\\\" ([\\d]+) ([^-].*)",
      “output.format.string” = “%1$s %2$s %3$s %4$s %5$s %6$s %7$s”
);
$ hadoop fs -copyFromLocal july.txt /user/csds/input
$ hadoop fs -ls /user/csds/input
hive> LOAD DATA INPATH '/user/csds/input/july.txt' INTO TABLE july;
```

```
hive> [training@localhost csds-material]$ hadoop fs -copyFromLocal july.txt /user/csds/data
[training@localhost csds-material]$ hadoop fs -ls /user/csds/data
Found 1 items
-rw-r--r--   1 training supergroup 1121063431 2017-03-29 23:07 /user/csds/data/july.txt
[training@localhost csds-material]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Hive history file=/tmp/training/hive_job_log_training_201703292309_1157125705.txt
hive> DROP TABLE july;
OK
Time taken: 4.326 seconds
hive>  CREATE TABLE july(
    > `Source_IP` STRING,
    > `Time_stamp` STRING,
    > `HTTP_Method` STRING,
    > `Request_URL` STRING,
    > `HTTP_Protocol` STRING,
    > `Status_Code` STRING,
    > `Response_Bytes`STRING
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
    > WITH SERDEPROPERTIES
    > (
    > "input.regex"="(.*) - - \\[(.*)\\] \\"([A-Z]*| [^\"]*) ([^HTTP]*) ([^\"]*)\\" ([\\d-]+) (.*)",
    > "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s"
    > );
OK
Time taken: 0.362 seconds
hive> LOAD DATA INPATH '/user/csds/data/july.txt' INTO TABLE july;
Loading data to table default.july
OK
Time taken: 0.484 seconds
```

- **Improvement:**

```
Time taken: 0.638 seconds
hive> CREATE TABLE july2(
    > `Source_IP` STRING,
    > `Time_Stamp` STRING,
    > `HTTP_Method` STRING,
    > `Request_URL` STRING,
    > `HTTP_Protocol`STRING,
    > `Status_Code` STRING,
    > `Response_Bytes`STRING
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
    > WITH SERDEPROPERTIES
    > (
    > "input.regex"="(^.*) - - \\[(.*)\\] \\"(\\S+|[^\\" ]*)([^H]*)([^\\"]*)\\\
" ([\\d-]+) ([\\d-]+$)",
    > "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s"
    > );
```

- **Output:**

```
hive> select * from july2 limit 5;
OK
199.72.81.55    01/Jul/1995:00:00:01 -0400      GET      /history/apollo/       HT
TP/1.0  200     6245
unicomp6.unicomp.net    01/Jul/1995:00:00:06 -0400      GET      /shuttle/countdow
n/      HTTP/1.0        200     3985
199.120.110.21  01/Jul/1995:00:00:09 -0400      GET      /shuttle/missions/sts-73/
mission-sts-73.html     HTTP/1.0        200     4085
burger.letters.com      01/Jul/1995:00:00:11 -0400      GET      /shuttle/countdow
n/liftoff.html  HTTP/1.0        304     0
199.120.110.21  01/Jul/1995:00:00:11 -0400      GET      /shuttle/missions/sts-73/
sts-73-patch-small.gif  HTTP/1.0        200     4179
Time taken: 0.362 seconds
```

1.2 Find the number of 200 status code in the response in the month of August.

- **Code:**

hive>select count(*) from july where Status_Code = ”200” and Time_Stamp like “%Aug%”;

```
hive> select count(*) from july where Status_Code="200" and Time_Stamp like "%Aug%";
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_201703292000_0005, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201703292000_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201703292000_0005
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 1
2017-03-29 23:15:49,979 Stage-1 map = 0%,  reduce = 0%
2017-03-29 23:16:16,271 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:17,296 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:18,305 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:19,313 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:20,337 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:21,350 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:22,370 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:23,382 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:24,401 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:25,412 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:26,420 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:27,435 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
2017-03-29 23:16:28,446 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 7.6 sec
```

Output:

The number of 200 status code in the response in the month of August is 2797976.

```
2017-03-30 18:45:53,345 Stage-1 map = 80%,  reduce = 27%, Cumulative CPU 111.27 se
c
2017-03-30 18:45:54,354 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 117.3 se
c
2017-03-30 18:45:55,358 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 118.5 s
ec
2017-03-30 18:45:56,367 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 118.5 s
ec
2017-03-30 18:45:57,415 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 118.5 s
ec
MapReduce Total cumulative CPU time: 1 minutes 58 seconds 500 msec
Ended Job = job_201703301700_0006
MapReduce Jobs Launched:
Job 0: Map: 5  Reduce: 1   Cumulative CPU: 118.5 sec   HDFS Read: 0 HDFS Write: 0
SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 58 seconds 500 msec
OK
2797976
Time taken: 161.689 seconds
```

1.3 Find the number of unique source IPs that have made requests to the NASA server for the month of September.

If we assume all the entries are making requests to NASA server, then:

- **Code:**

hive>select count(distinct Source_IP) from july

>where Time_Stamp like “%Sep%”;

```
hive> select count(distinct Source_IP) from july2
    > where Time_Stamp like "%Sep%";
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_201703302330_0001, Tracking URL = http://0.0.0.0:50030/jobdet
ails.jsp?jobid=job_201703302330_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021
 -kill job_201703302330_0001
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 1
2017-03-30 23:37:37,764 Stage-1 map = 0%,  reduce = 0%
2017-03-30 23:38:03,059 Stage-1 map = 10%,  reduce = 0%
2017-03-30 23:38:18,242 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 20.17 se
c
2017-03-30 23:38:19,282 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 20.17 se
c
2017-03-30 23:38:20,298 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 20.17 se
```

- **Output:**

The number of unique source IPs are 81982.

```
2017-03-30 23:40:35,964 Stage-1 map = 100%,  reduce = 27%, Cumulative CPU 129.72
 sec
2017-03-30 23:40:37,061 Stage-1 map = 100%,  reduce = 27%, Cumulative CPU 129.72
 sec
2017-03-30 23:40:38,196 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 131.6
3 sec
2017-03-30 23:40:39,204 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 131.6
3 sec
2017-03-30 23:40:40,209 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 131.6
3 sec
2017-03-30 23:40:41,218 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 131.6
3 sec
MapReduce Total cumulative CPU time: 2 minutes 11 seconds 630 msec
Ended Job = job_201703302330_0001
MapReduce Jobs Launched:
Job 0: Map: 5  Reduce: 1   Cumulative CPU: 131.63 sec   HDFS Read: 0 HDFS Write:
 0 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 11 seconds 630 msec
OK
81982
Time taken: 192.982 seconds
hive>
```

1.4 Which was the most requested URL in the year 1995.

- **Code:**

```
hive>Select rs.Request_URL from
    >(
>select Request_URL, count(Request_URL) as score from july group by
Request_URL order by score DESC) rs
    >limit 1;
```

```
hive> select rs.Request_URL from
    > (
    > select Request_URL,count(Request_URL) as score from july group by Request_URL order by score DESC)rs
    > limit 1;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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_201703292000_0011, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201703292000_0011
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201703292000_0011
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 2
2017-03-30 00:44:23,116 Stage-1 map = 0%,  reduce = 0%
2017-03-30 00:44:52,282 Stage-1 map = 5%,  reduce = 0%
2017-03-30 00:44:55,301 Stage-1 map = 10%,  reduce = 0%
2017-03-30 00:45:07,357 Stage-1 map = 15%,  reduce = 0%
2017-03-30 00:45:17,404 Stage-1 map = 20%,  reduce = 0%
2017-03-30 00:45:25,458 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 41.83 sec
2017-03-30 00:45:26,465 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 41.83 sec
2017-03-30 00:45:27,473 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 41.83 sec
```

- **Output:**

The most requested URL in the year 1995 is /images/NASA-logosmall.gif

```
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_201703292000_0012, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201703292000_0012
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201703292000_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-30 00:47:51,189 Stage-2 map = 0%,  reduce = 0%
2017-03-30 00:47:56,212 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2017-03-30 00:47:57,215 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2017-03-30 00:47:58,221 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2017-03-30 00:47:59,228 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2017-03-30 00:48:00,232 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2017-03-30 00:48:01,241 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
2017-03-30 00:48:02,250 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
2017-03-30 00:48:03,255 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
MapReduce Total cumulative CPU time: 3 seconds 670 msec
Ended Job = job_201703292000_0012
MapReduce Jobs Launched:
Job 0: Map: 5  Reduce: 2   Cumulative CPU: 150.07 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 33 seconds 740 msec
OK
/images/NASA-logosmall.gif
Time taken: 225.114 seconds
```

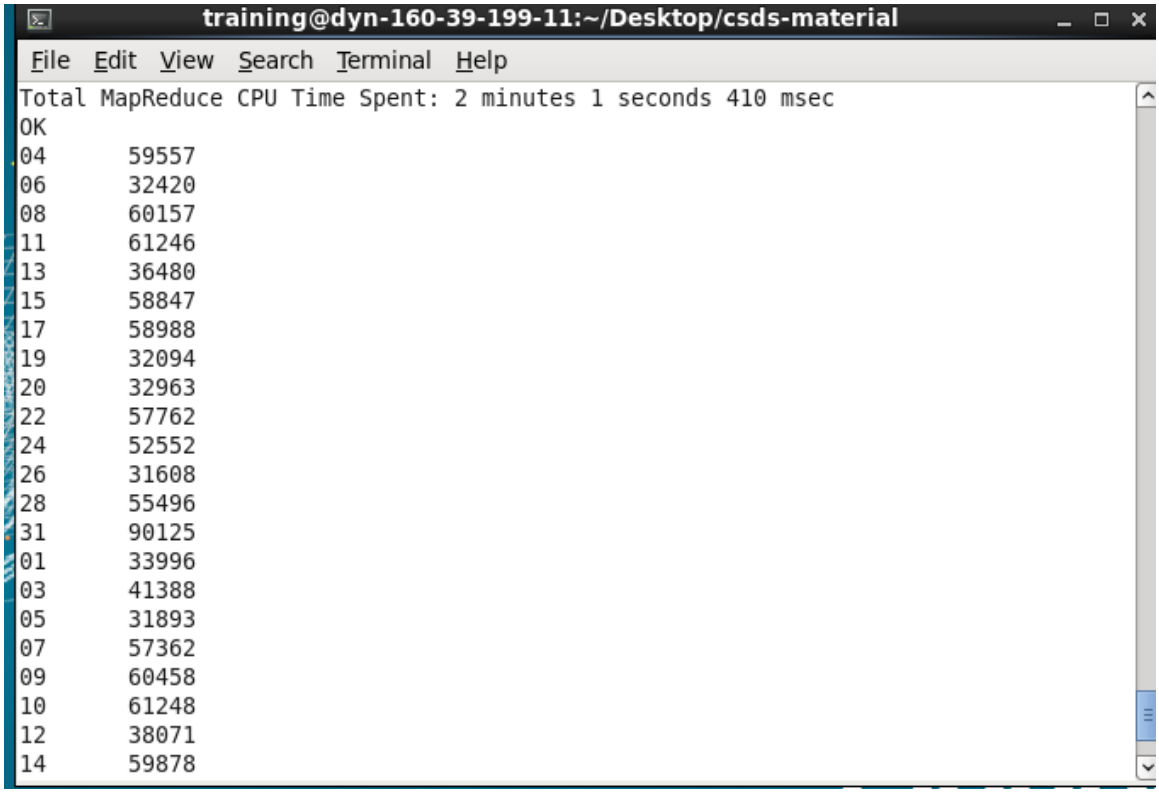
1.5 Make a histogram depicting the number of requests made in a day for every day in the month of October.

- **Code**

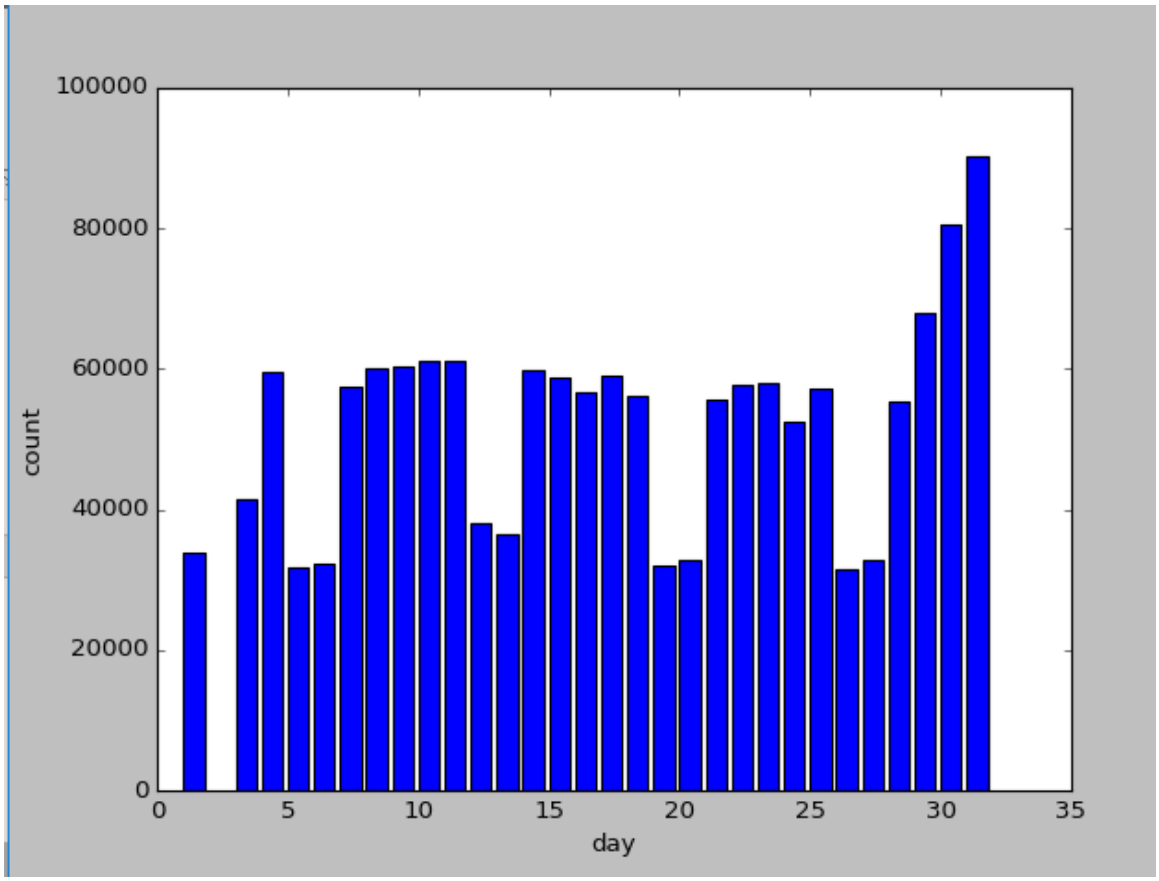
```
hive> select rs.day, count(rs.day) from
    > ( select split(Time_Stamp,'/')[0] as day from july
    > where Time_Stamp like "%Oct%" ) rs
    > group by rs.day;
```

```
hive> select rs.day,count(rs.day) from
    > (select split(Time_Stamp,'/')[0] as day from july
    > where Time_Stamp like "%Oct%") rs
    > group by rs.day;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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_201703300104_0006, Tracking URL = http://0.0.0.0:50030/jobdet
ails.jsp?jobid=job_201703300104_0006
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021
 -kill job_201703300104_0006
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 2
2017-03-30 01:37:49,681 Stage-1 map = 0%,  reduce = 0%
2017-03-30 01:38:12,784 Stage-1 map = 10%,  reduce = 0%
```

- **Output:**




```
hive>insert overwrite local directory '/home/training/Desktop/csds-material'
    > select rs.day, count(rs.day) from
    > ( select split(Time_Stamp,'/')[0] as day from july
    > where Time_Stamp like "%Oct%" ) rs
    > group by rs.day;
```


100000
80000
60000
40000
20000
0
count
0
5
10
15
20
25
30
35
day