Module 2: Assignment – 5

Creating a Table and Schema

```
hive> create table if not exists babynames
    > (id int, name string, year int, gender string, count int)
    > row format delimited fields terminated by ',' stored as textfile
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.146 seconds
```

Loading babaynames data in the table

```
hive> load data local inpath '/home/bitnami/babynames.csv' into table babynames;
Loading data to table aug.babynames
OK
Time taken: 2.16 seconds
```

Viewing loaded data in the table

```
hive> select * from babynames limit 10;
OK
1       Mary    1880    F       7065
2       Anna    1880    F       2604
3       Emma    1880    F       2003
4       Elizabeth       1880    F       1939
5       Minnie  1880    F       1746
6       Margaret        1880    F       1578
7       Ida     1880    F       1472
8       Alice   1880    F       1414
9       Bertha  1880    F       1320
10      Sarah   1880    F       1288
Time taken: 2.082 seconds, Fetched: 10 row(s)
```

1. Create an output with all the names starting with the letter, ‘G’

```
hive> select name from babynames
    > where name like 'G%'
    > order by name
    > limit 10;
Automatically selecting local only mode for query
Query ID = root_20211109113735_a2e1c39e-6730-479a-bab5-8c594af07b30
Total 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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2021-11-09 11:37:36,870 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1472139005_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 179515594 HDFS Write: 90806635 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Gaal
Gaal
Gaal
Gabbanelli
Gabbi
Gabbi
Gabbi
Gabbi
Gabbi
Gabbi
Time taken: 1.591 seconds, Fetched: 10 row(s)
```

2. Create an output with all the names having only two vowels

```
hive> select name from babynames
    > where lower(name) regexp '^[^aeiou]*[aeiou][^aeiou]*[aeiou][^aeiou]*$'
    > limit 10;
OK
Anna
Emma
Ida
Bertha
Sarah
Clara
Ella
Cora
Martha
Grace
Time taken: 0.137 seconds, Fetched: 10 row(s)
```

3. Create an output with names without any vowels, if any

```
hive> select name from babynames
    > where lower(name) not regexp '[aeiou]'
    > limit 10;
OK
Byrd
Lynn
Lynn
Wm
Byrd
Lynn
Wm
Byrd
Lynn
Wm
Time taken: 0.147 seconds, Fetched: 10 row(s)
```

4. Create an output with names containing the syllable, ‘red’

```
hive> select name from babynames
    > where name regexp 'red+'
    > limit 10;
OK
Mildred
Winifred
Freda
Fred
Winnifred
Freddie
Fred
Frederick
Alfred
Fredrick
Time taken: 0.117 seconds, Fetched: 10 row(s)
```

5. Create an output with the total count of each name, throughout all years

Running Map Reduce Job to view Output

```
hive> select name, count(name) as name_count
    > from babynames
    > group by name
    > order by name_count desc limit 10;
Automatically selecting local only mode for query
Query ID = root_20211109123241_573f5349-6d25-4577-8a7a-74a85192bc36
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)
2021-11-09 12:32:43,469 Stage-1 map = 0%,  reduce = 0%
2021-11-09 12:32:44,483 Stage-1 map = 100%,  reduce = 0%
2021-11-09 12:32:46,502 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1413403676_0004
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>
Selecting local mode for task: Stage-2
Job running in-process (local Hadoop)
2021-11-09 12:32:48,311 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local524700431_0005
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1602118478 HDFS Write: 93207384 SUCCESS
Stage-Stage-2:  HDFS Read: 1606928966 HDFS Write: 95622140 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Joseph  270
Johnnie 270
William 270
Jesse   270
Jessie  270
James   270
Jean    270
Leslie  270
Robert  270
Francis 270
Time taken: 6.387 seconds, Fetched: 10 row(s)
```

6. Find the most popular female name in 1918

```
hive> select name, gender, year, count
    > from babynames
    > where year = 1918
    > AND gender = "F"
    > group by name, gender, year, count
    > order by count desc limit 1;
```

Running Map Reduce Job to view Output

```
Automatically selecting local only mode for query
Query ID = root_20211109131057_6f5c8ed4-d4f4-450c-8f06-a80b8e15f704
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)
2021-11-09 13:10:59,122 Stage-1 map = 0%,  reduce = 0%
2021-11-09 13:11:00,128 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1333432011_0012
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>
Selecting local mode for task: Stage-2
Job running in-process (local Hadoop)
2021-11-09 13:11:02,031 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1100270413_0013
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 714609196 HDFS Write: 4991878 SUCCESS
Stage-Stage-2:  HDFS Read: 714620325 HDFS Write: 5007401 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Mary    F       1918    67370
Time taken: 4.572 seconds, Fetched: 1 row(s)
```

7. Find the most popular male name in 1945

```
hive> select name, gender, year, count
    > from babynames
    > where year = 1945
    > AND gender = "M"
    > group by name, gender, year, count
    > order by count desc limit 1;
```

Running Map Reduce Job to view Output

```
Automatically selecting local only mode for query
Query ID = root_20211109131251_766f482f-6287-4d38-882d-a2b77faad0d4
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)
2021-11-09 13:12:53,047 Stage-1 map = 0%,  reduce = 0%
2021-11-09 13:12:54,051 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local475655206_0014
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>
Selecting local mode for task: Stage-2
Job running in-process (local Hadoop)
2021-11-09 13:12:55,962 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local487184769_0015
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 803326186 HDFS Write: 5007641 SUCCESS
Stage-Stage-2:  HDFS Read: 803337317 HDFS Write: 5023166 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
James   M       1945    74450
Time taken: 4.576 seconds, Fetched: 1 row(s)
```