SMIT R PATEL
19162121031
SEM 5
PRACTICAL 14
HIVE

AIM- To execute queries and work in Hive.

Exercise-

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Content

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity:

Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

7. Organic_carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. Trihalomethanes:

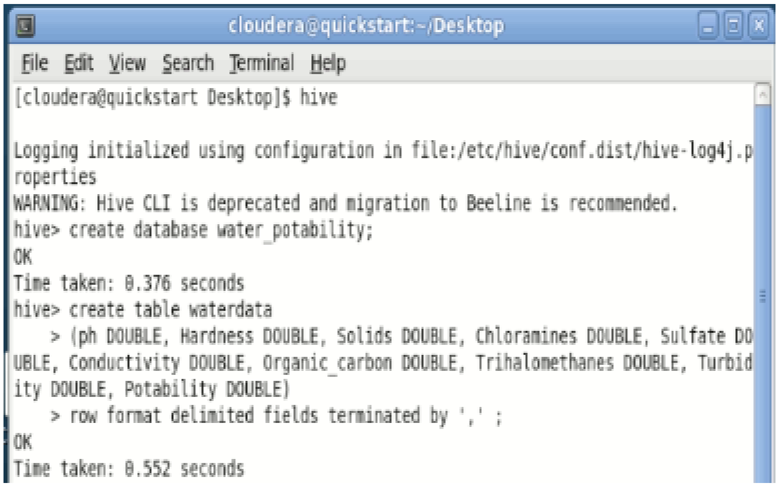
THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. Potability:

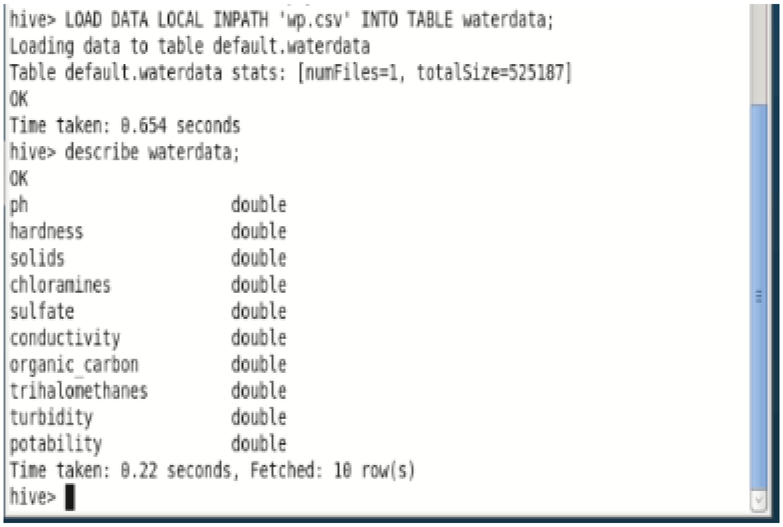
Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.



cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
[cloudera@quickstart Desktop]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database water_potability;
OK
Time taken: 0.376 seconds
hive> create table waterdata
    > (ph DOUBLE, Hardness DOUBLE, Solids DOUBLE, Chloramines DOUBLE, Sulfate DO
UBLE, Conductivity DOUBLE, Organic_carbon DOUBLE, Trihalomethanes DOUBLE, Turbid
ity DOUBLE, Potability DOUBLE)
    > row format delimited fields terminated by ',' ;
OK
Time taken: 0.552 seconds
```

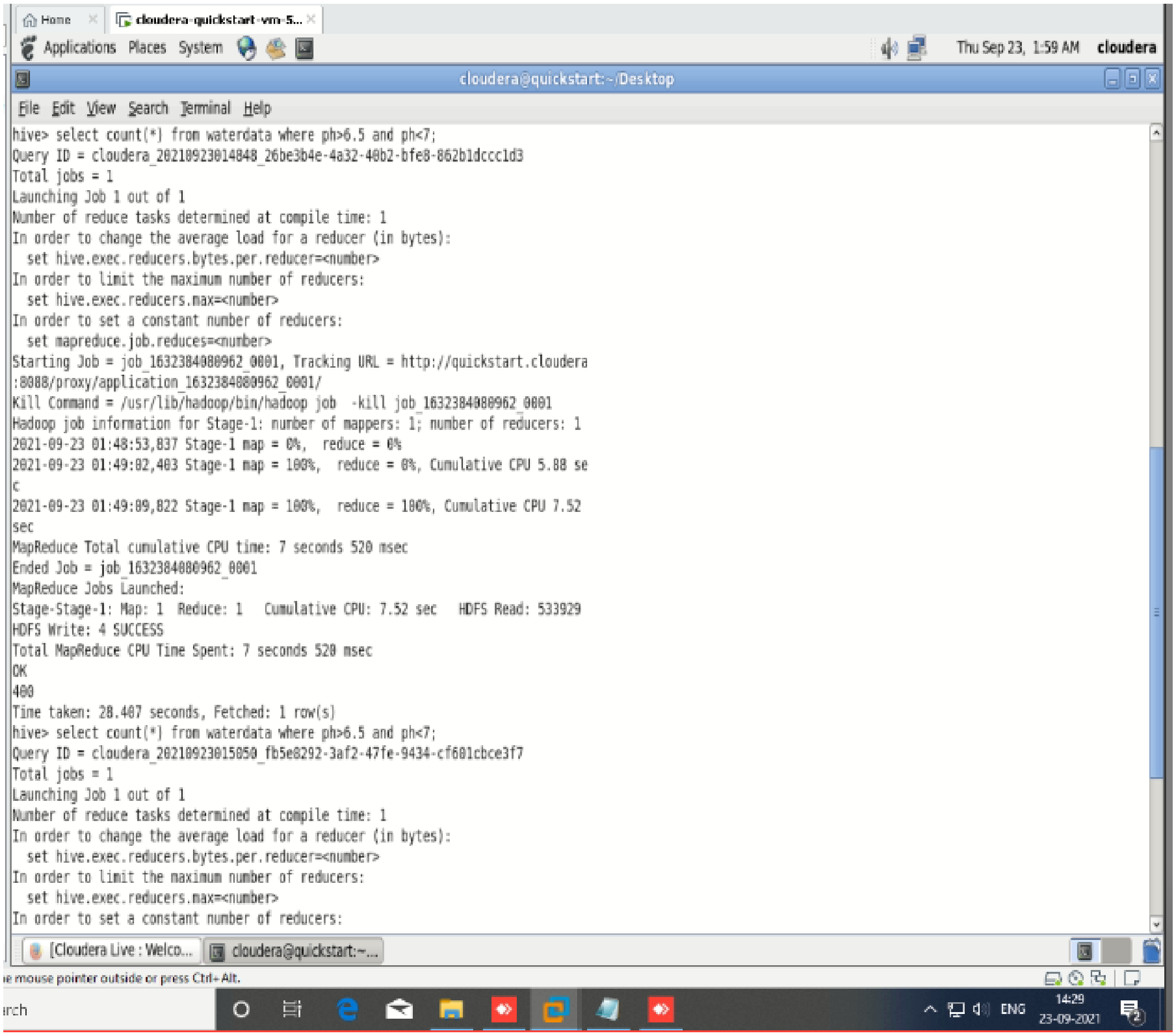


```
hive> LOAD DATA LOCAL INPATH 'wp.csv' INTO TABLE waterdata;
Loading data to table default.waterdata
Table default.waterdata stats: [numFiles=1, totalSize=525187]
OK
Time taken: 0.654 seconds
hive> describe waterdata;
OK
ph                      double
hardness                double
solids                  double
chloramines             double
sulfate                 double
conductivity            double
organic_carbon          double
trihalomethanes         double
turbidity               double
potability              double
Time taken: 0.22 seconds, Fetched: 10 row(s)
hive> 
```

Tasks:

1. Find out how many entries exist for the pH of 6.5-7 in the dataset.

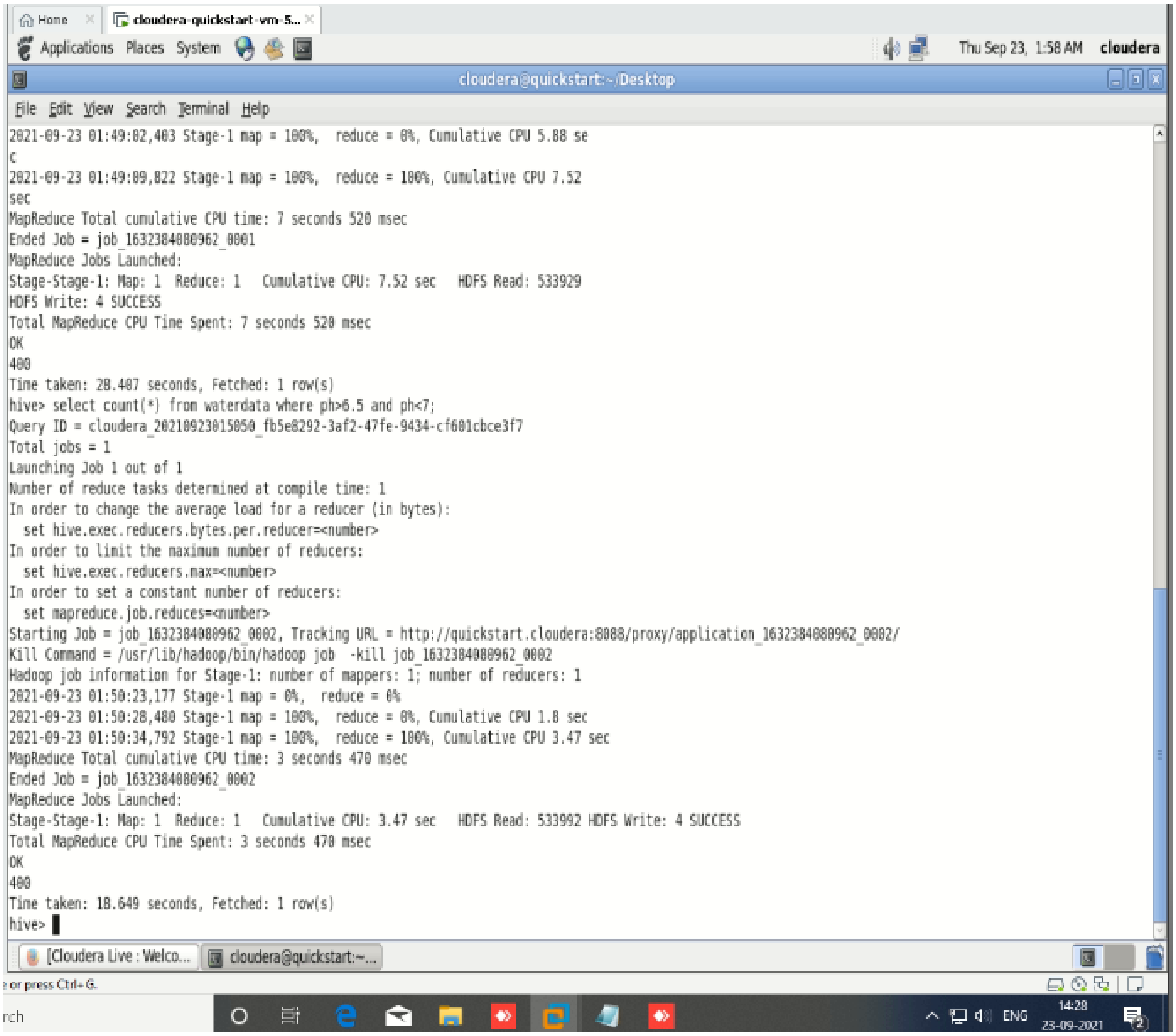
   Command :-
   hive> select count(*) from waterdata where ph>6.5 and ph<7;



Home
cloudera-quickstart-vm-5...
Applications Places System
Thu Sep 23, 1:59 AM cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
hive> select count(*) from waterdata where ph>6.5 and ph<7;
Query ID = cloudera_20210923014848_26be3b4e-4a32-40b2-bfe8-862b1dccc1d3
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>
Starting Job = job_1632384080962_0001, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1632384080962_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632384080962_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 01:48:53,837 Stage-1 map = 0%,  reduce = 0%
2021-09-23 01:49:02,403 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.88 se
c
2021-09-23 01:49:09,822 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.52
sec
MapReduce Total cumulative CPU time: 7 seconds 520 msec
Ended Job = job_1632384080962_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.52 sec   HDFS Read: 533929
HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 520 msec
OK
400
Time taken: 28.407 seconds, Fetched: 1 row(s)
hive> select count(*) from waterdata where ph>6.5 and ph<7;
Query ID = cloudera_20210923015050_fb5e8292-3af2-47fe-9434-cf601cbce3f7
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:
```

[Cloudera Live : Welco...
cloudera@quickstart:~...
e mouse pointer outside or press Ctrl+Alt.
rch
ENG
14:29
23-09-2021

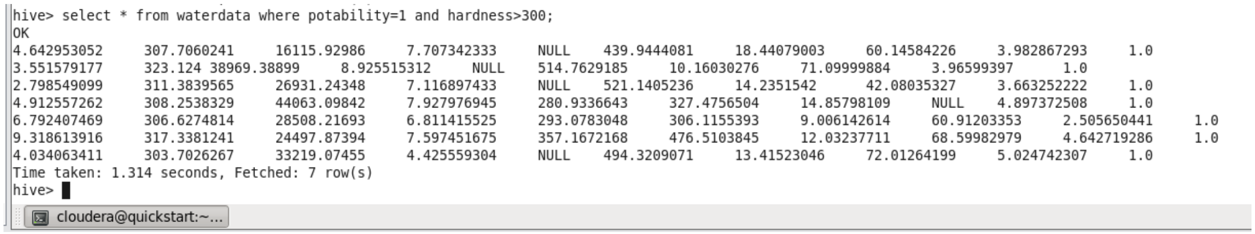

Home
cloudera-quickstart-vm-5...
Applications Places System
Thu Sep 23, 1:58 AM cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
2021-09-23 01:49:02,403 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.88 se
c
2021-09-23 01:49:09,822 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.52
sec
MapReduce Total cumulative CPU time: 7 seconds 520 msec
Ended Job = job_1632384080962_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.52 sec   HDFS Read: 533929
HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 520 msec
OK
400
Time taken: 28.407 seconds, Fetched: 1 row(s)
hive> select count(*) from waterdata where ph>6.5 and ph<7;
Query ID = cloudera_20210923015050_fb5e8292-3af2-47fe-9434-cf601cbce3f7
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>
Starting Job = job_1632384080962_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1632384080962_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632384080962_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 01:50:23,177 Stage-1 map = 0%,  reduce = 0%
2021-09-23 01:50:28,480 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.8 sec
2021-09-23 01:50:34,792 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.47 sec
MapReduce Total cumulative CPU time: 3 seconds 470 msec
Ended Job = job_1632384080962_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.47 sec   HDFS Read: 533992 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 470 msec
OK
400
Time taken: 18.649 seconds, Fetched: 1 row(s)
hive>
```

2. Check whether columns exist where the water is potable yet Hardness is above 300.

   Command:-

   Hive> select * from waterdata where potability=1 and hardness>300;

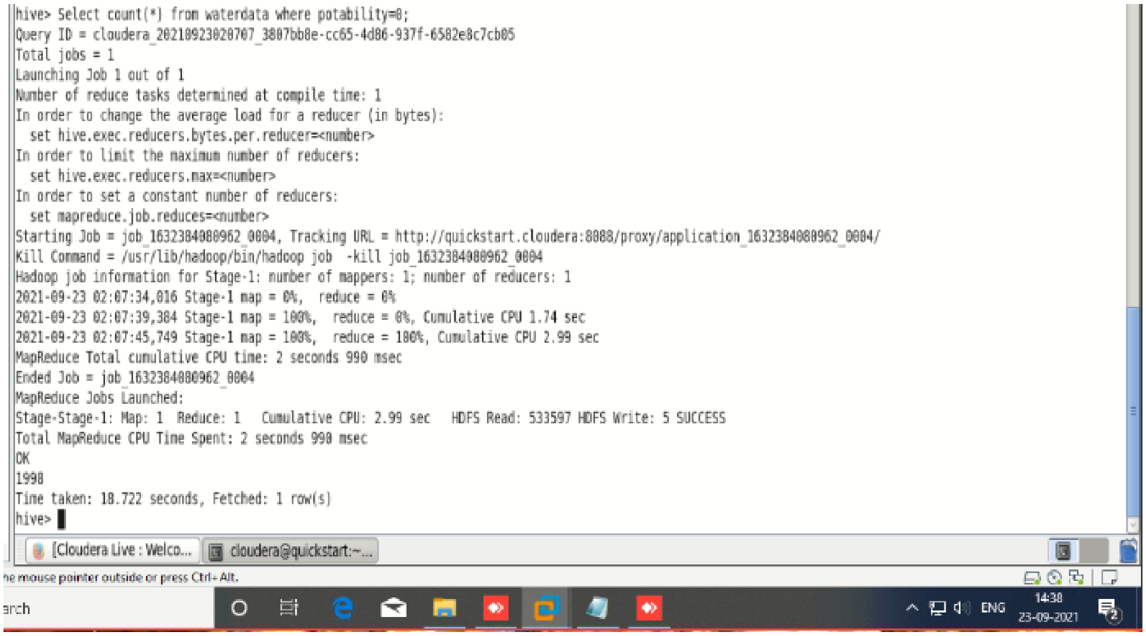


```
hive> select * from waterdata where potability=1 and hardness>300;
OK
4.642953052     307.7060241     16115.92986     7.707342333     NULL    439.9444081     18.44079003     60.14584226     3.982867293     1.0
3.551579177     323.124 38969.38899     8.925515312     NULL    514.7629185     10.16030276     71.09999884     3.96599397      1.0
2.798549099     311.3839565     26931.24348     7.116897433     NULL    521.1405236     14.2351542      42.08035327     3.663252222     1.0
4.912557262     308.2538329     44063.09842     7.927976945     280.9336643     327.4756504     14.85798109     NULL    4.897372508     1.0
6.792407469     306.6274814     28508.21693     6.811415525     293.0783048     306.1155393     9.006142614     60.91203353     2.505650441     1.0
9.318613916     317.3381241     24497.87394     7.597451675     357.1672168     476.5103845     12.03237711     68.59982979     4.642719286     1.0
4.034063411     303.7026267     33219.07455     4.425559304     NULL    494.3209071     13.41523046     72.01264199     5.024742307     1.0
Time taken: 1.314 seconds, Fetched: 7 row(s)
hive> █
```

3. Find out how many rows are given for Potability=0 and Potability= 1.

   Command :-

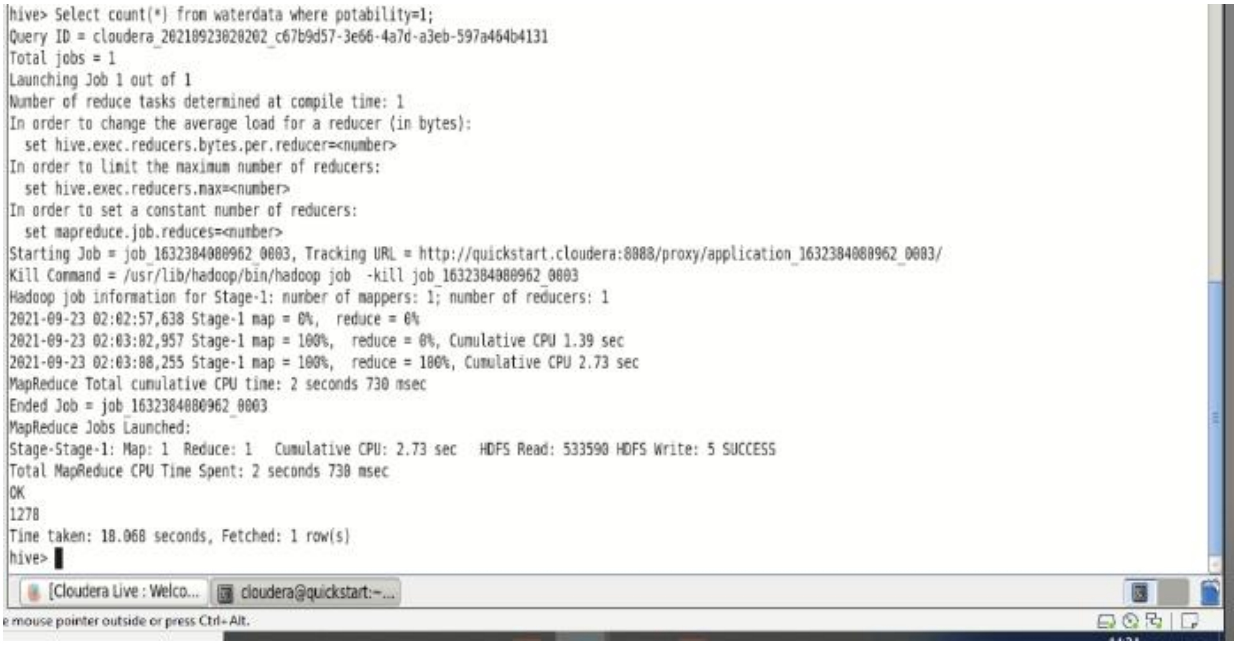
   Hive> Select count(*) from waterdata where potability=0;



```
hive> Select count(*) from waterdata where potability=0;
Query ID = cloudera_20210923020707_3807bb8e-cc65-4d86-937f-6582e8c7cb05
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>
Starting Job = job_1632384080962_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1632384080962_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632384080962_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 02:07:34,016 Stage-1 map = 0%,  reduce = 0%
2021-09-23 02:07:39,384 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.74 sec
2021-09-23 02:07:45,749 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.99 sec
MapReduce Total cumulative CPU time: 2 seconds 990 msec
Ended Job = job_1632384080962_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.99 sec   HDFS Read: 533597 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 990 msec
OK
1998
Time taken: 18.722 seconds, Fetched: 1 row(s)
hive> █
```

[Cloudera Live : Welco...
cloudera@quickstart:~...
he mouse pointer outside or press Ctrl+Alt.
arch
ENG
14:38
23-09-2021

Hive> Select count(*) from waterdata where potability=1;



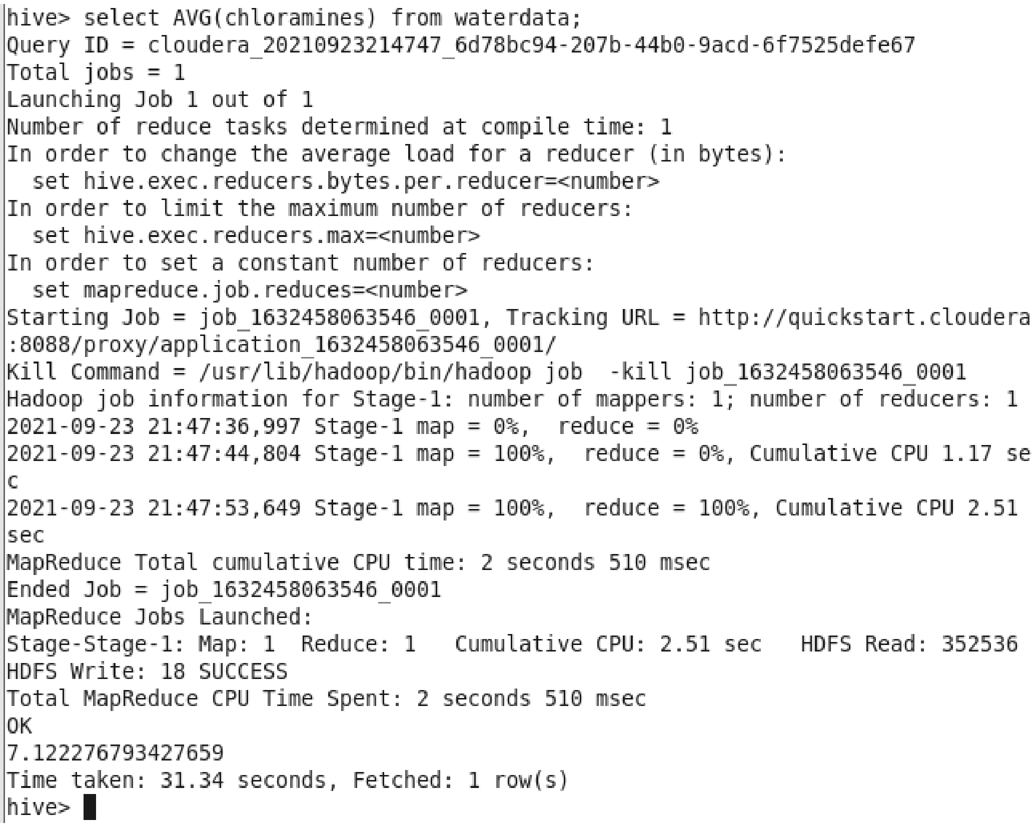
```
hive> Select count(*) from waterdata where potability=1;
Query ID = cloudera_20210923020202_c67b9d57-3e66-4a7d-a3eb-597a464b4131
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1632384080962_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1632384080962_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632384080962_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 02:02:57,638 Stage-1 map = 0%,  reduce = 0%
2021-09-23 02:03:02,957 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.39 sec
2021-09-23 02:03:08,255 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.73 sec
MapReduce Total cumulative CPU time: 2 seconds 730 msec
Ended Job = job_1632384080962_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.73 sec   HDFS Read: 533590 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 730 msec
OK
1278
Time taken: 18.068 seconds, Fetched: 1 row(s)
hive>
```

[Cloudera Live : Welco...
cloudera@quickstart:~...
e mouse pointer outside or press Ctrl+Alt.

4. What is the average Chloramine value in the dataset?

Command:-

hive> select AVG(chloramines) from waterdata;

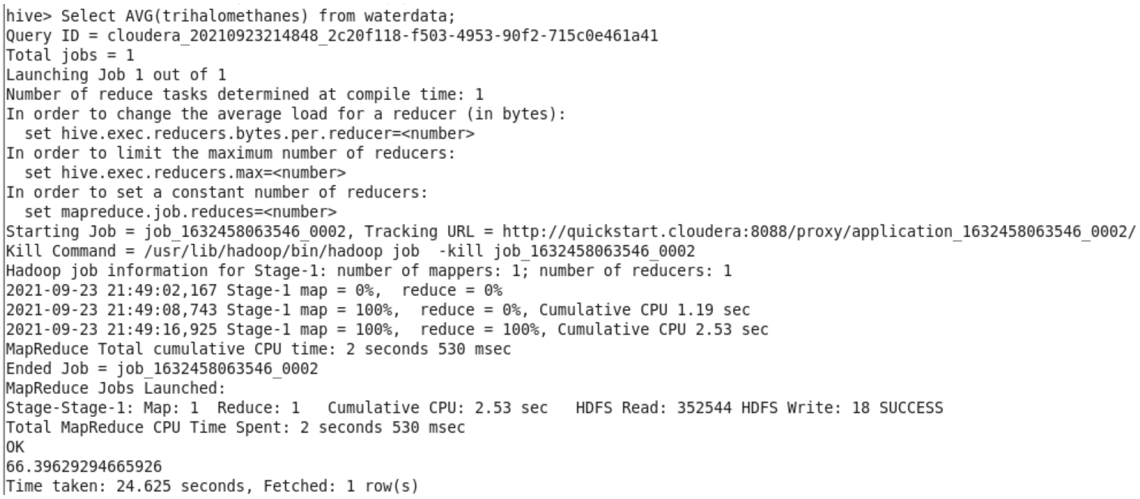


```
hive> select AVG(chloramines) from waterdata;
Query ID = cloudera_20210923214747_6d78bc94-207b-44b0-9acd-6f7525defe67
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1632458063546_0001, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1632458063546_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632458063546_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 21:47:36,997 Stage-1 map = 0%,  reduce = 0%
2021-09-23 21:47:44,804 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.17 se
c
2021-09-23 21:47:53,649 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.51
sec
MapReduce Total cumulative CPU time: 2 seconds 510 msec
Ended Job = job_1632458063546_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.51 sec   HDFS Read: 352536
HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 510 msec
OK
7.122276793427659
Time taken: 31.34 seconds, Fetched: 1 row(s)
hive> 
```

5. Calculate the average value of Trihalomethanes present in the dataset.

Command:-

Hive> Select AVG(trihalomethanes) from waterdata;

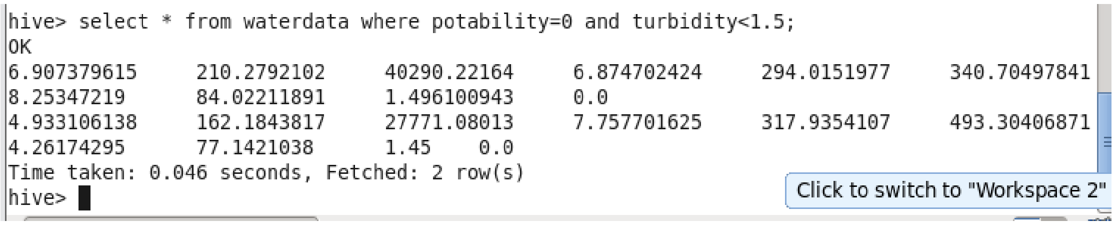


```
hive> Select AVG(trihalomethanes) from waterdata;
Query ID = cloudera_20210923214848_2c20f118-f503-4953-90f2-715c0e461a41
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1632458063546_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1632458063546_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632458063546_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-23 21:49:02,167 Stage-1 map = 0%,  reduce = 0%
2021-09-23 21:49:08,743 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.19 sec
2021-09-23 21:49:16,925 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.53 sec
MapReduce Total cumulative CPU time: 2 seconds 530 msec
Ended Job = job_1632458063546_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.53 sec   HDFS Read: 352544 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 530 msec
OK
66.39629294665926
Time taken: 24.625 seconds, Fetched: 1 row(s)
```

6. Display rows where potability is 0 and turbidity is less than 1.5.

Command:-

Hive> select * from waterdata where potability=0 and turbidity<1.5;



```
hive> select * from waterdata where potability=0 and turbidity<1.5;
OK
6.907379615     210.2792102     40290.22164     6.874702424     294.0151977     340.70497841
8.25347219      84.02211891     1.496100943     0.0
4.933106138     162.1843817     27771.08013     7.757701625     317.9354107     493.30406871
4.26174295      77.1421038      1.45    0.0
Time taken: 0.046 seconds, Fetched: 2 row(s)
hive> 
```