This is a real time dataset of the ineuron technical consultant team. You have to perform hive analysis on this given dataset.

Download Dataset 1 - https://drive.google.com/file/d/1WrG-9qv6atP-W3P_gYln1hHyFKRKMHP/view

Download Dataset 2 - https://drive.google.com/file/d/1-JIPCZ34dyN6k9CqJa-Y8yxIGq6vTVXU/view

Note: both files are csv files.

1. Create a schema based on the given dataset

note I chagned column name date to event_date because date keyword allready present in hive datatypes thats why it is not taking date name.

```
Create table AgentLogingReport

(

sr_no int,

Agent string,

Event_date date,

Login string,

Logout string,

Duration string

)

row format delimited

fields terminated by ','

tblproperties ("skip.header.line.count" = "1");
```

```
Create table AgentPerformance

(

sr_no int,

Event_date date,

Agent_Name string,

Total_charts int,

Avg_Response_Time string,
Avg_Resolution_Time string,
Avg_Rating float,
Total_Feedback int
)
row format delimited
fields terminated by ','
tblproperties ("skip.header.line.count" = "1");
```

2. Dump the data inside the hdfs in the given schema location.

data is present on root location(data should present on root location before we performing further processing on file)

```
hadoop fs -put AgentLogingReport.csv /
```

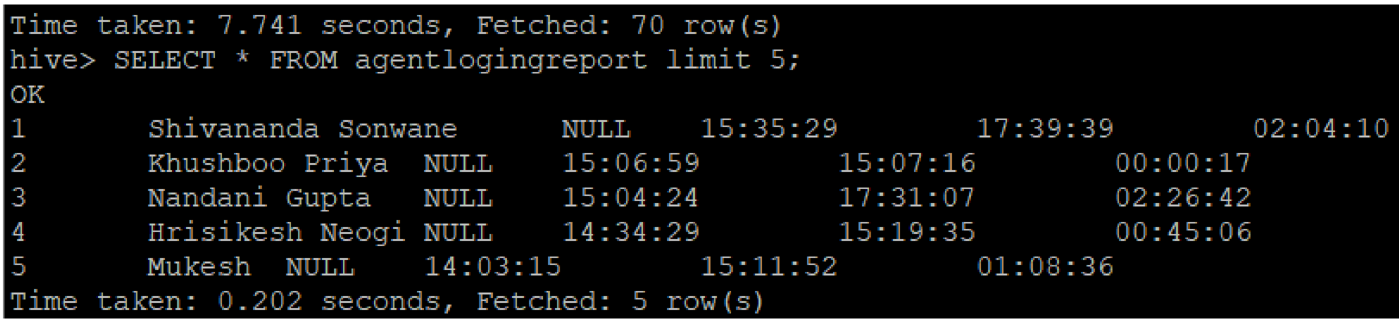
```
hadoop fs -put AgentPerformance.csv /
```

above this command help move data to the root location.

```
LOAD DATA INPATH '/AgentLogingReport.csv' INTO TABLE AgentLogingReport;
```

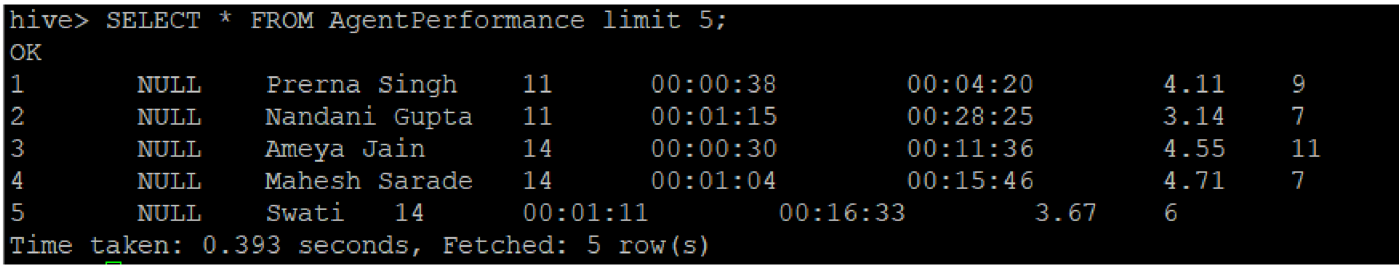
```
LOAD DATA INPATH '/AgentPerformance.csv' INTO TABLE AgentPerformance;
```

```
SELECT * FROM agentlogingreport limit 5;
```



```
Time taken: 7.741 seconds, Fetched: 70 row(s)
hive> SELECT * FROM agentlogingreport limit 5;
OK
1       Shivananda Sonwane      NULL    15:35:29        17:39:39        02:04:10
2       Khushboo Priya  NULL    15:06:59        15:07:16        00:00:17
3       Nandani Gupta   NULL    15:04:24        17:31:07        02:26:42
4       Hrisikesh Neogi NULL    14:34:29        15:19:35        00:45:06
5       Mukesh  NULL    14:03:15        15:11:52        01:08:36
Time taken: 0.202 seconds, Fetched: 5 row(s)
```

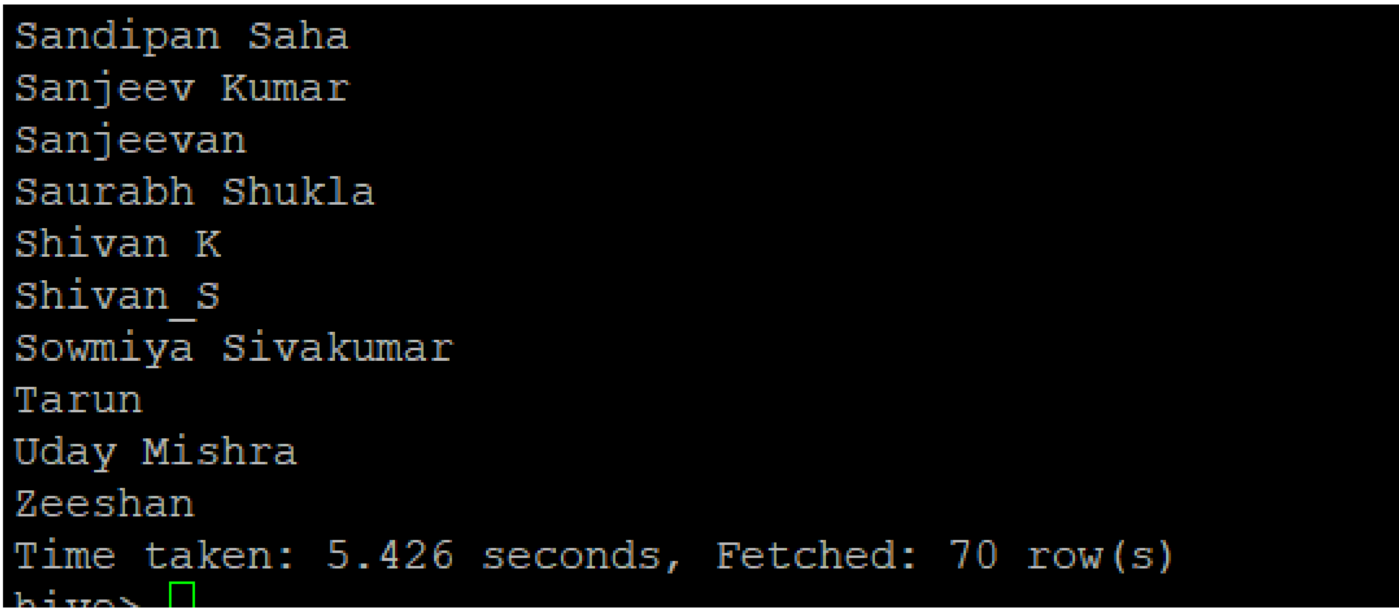
SELECT * FROM AgentPerformance limit 5;



```
hive> SELECT * FROM AgentPerformance limit 5;
OK
1       NULL    Prerna Singh    11      00:00:38        00:04:20        4.11    9
2       NULL    Nandani Gupta   11      00:01:15        00:28:25        3.14    7
3       NULL    Ameya Jain      14      00:00:30        00:11:36        4.55    11
4       NULL    Mahesh Sarade   14      00:01:04        00:15:46        4.71    7
5       NULL    Swati   14      00:01:11        00:16:33        3.67    6
Time taken: 0.393 seconds, Fetched: 5 row(s)
```

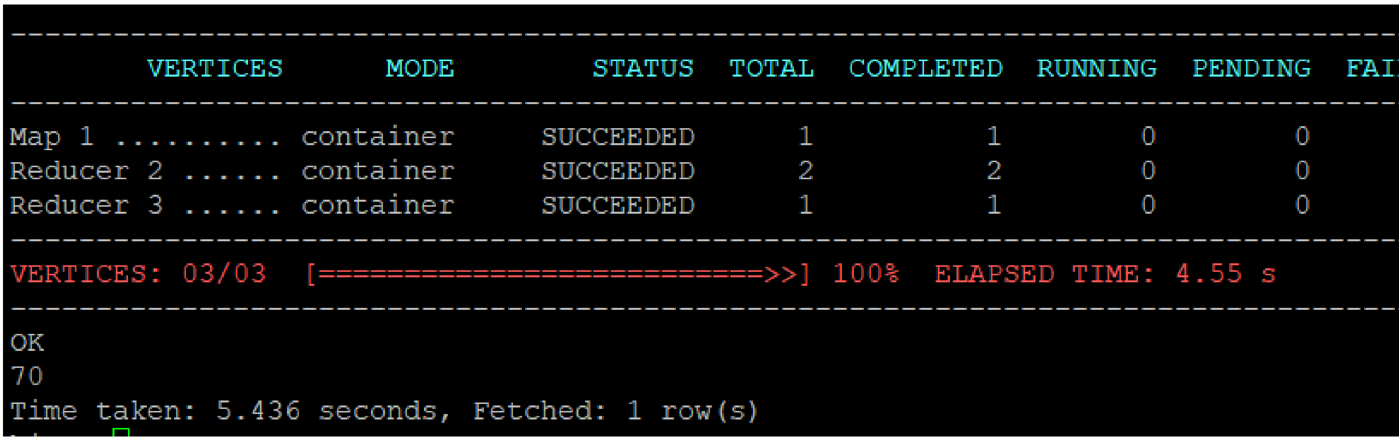
3. List all agents' names.

Hive> select distinct Agent_Name from AgentPerformance;



```
Sandipan Saha
Sanjeev Kumar
Sanjeevan
Saurabh Shukla
Shivan K
Shivan_S
Sowmiya Sivakumar
Tarun
Uday Mishra
Zeeshan
Time taken: 5.426 seconds, Fetched: 70 row(s)
hive>
```

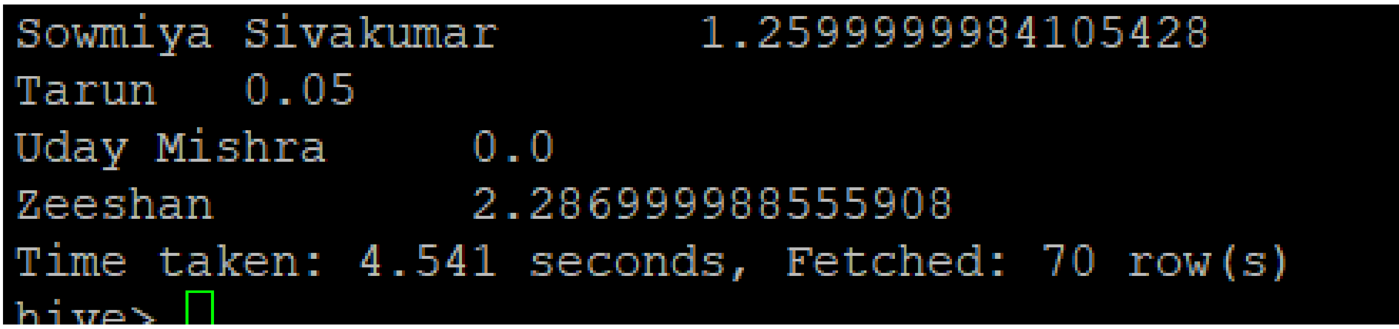
Hive> select count(distinct Agent_Name) from AgentPerformance;



```
----------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAI
----------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0
----------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.55 s
----------------------------------------------------------------------------------------
OK
70
Time taken: 5.436 seconds, Fetched: 1 row(s)
```

4. Find out agent average rating.

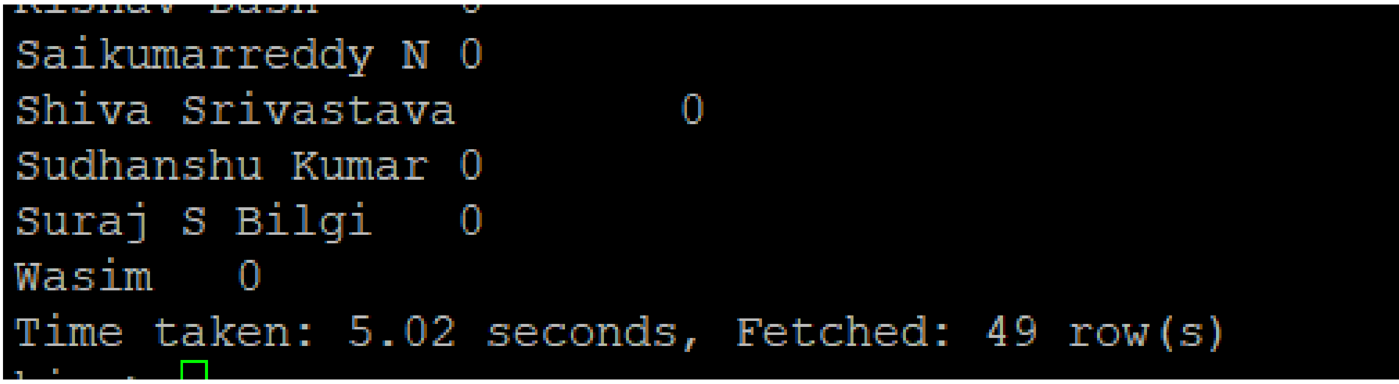
Hive> select Agent_name,avg(Avg_Rating) from AgentPerformance group by Agent_name;



```
Sowmiya Sivakumar       1.2599999984105428
Tarun   0.05
Uday Mishra     0.0
Zeeshan         2.286999988555908
Time taken: 4.541 seconds, Fetched: 70 row(s)
hive>
```

5. Total working days for each agent

Hive> select Agent,count(distinct Date) from AgentLogingReport group by Agent;



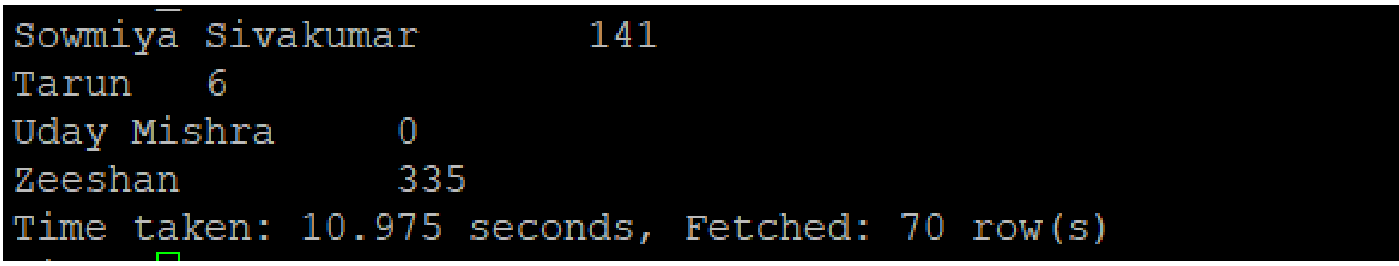
```
Saikumarreddy N 0
Shiva Srivastava        0
Sudhanshu Kumar 0
Suraj S Bilgi   0
Wasim   0
Time taken: 5.02 seconds, Fetched: 49 row(s)
```

6. Total query that each agent have taken

Hive> select Agent_name,sum(total_chats) from AgentPerformance group by Agent_name;

7. Total Feedback that each agent have received

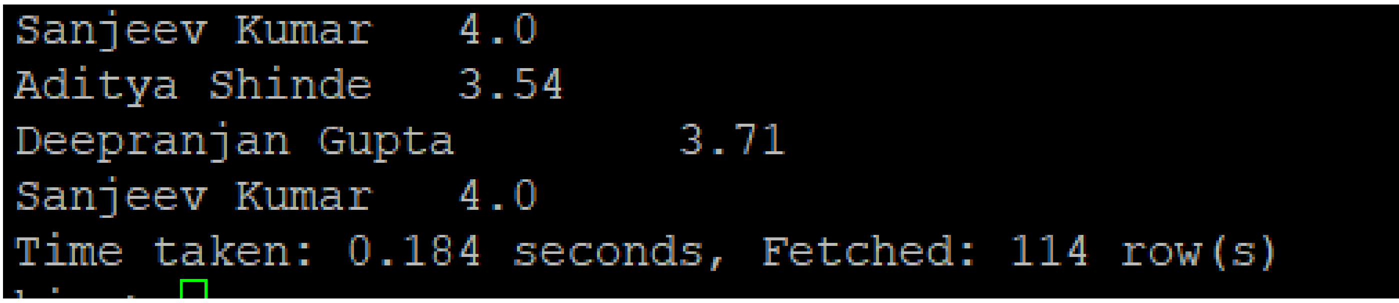
Hive> select Agent_name,sum(Total_Feedback) from AgentPerformance group by Agent_name;



```
Sowmiya Sivakumar       141
Tarun   6
Uday Mishra     0
Zeeshan         335
Time taken: 10.975 seconds, Fetched: 70 row(s)
```

8. Agent name who have average rating between 3.5 to 4

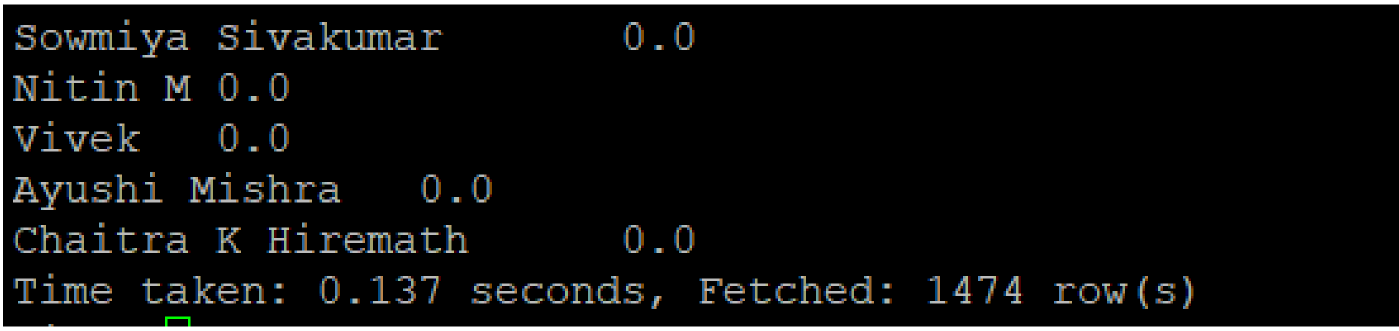
Hive> select Agent_name,Avg_Rating from AgentPerformance where Avg_Rating between 3.5 and 4;



```
Sanjeev Kumar   4.0
Aditya Shinde   3.54
Deepranjan Gupta        3.71
Sanjeev Kumar   4.0
Time taken: 0.184 seconds, Fetched: 114 row(s)
```

9. Agent name who have rating less than 3.5

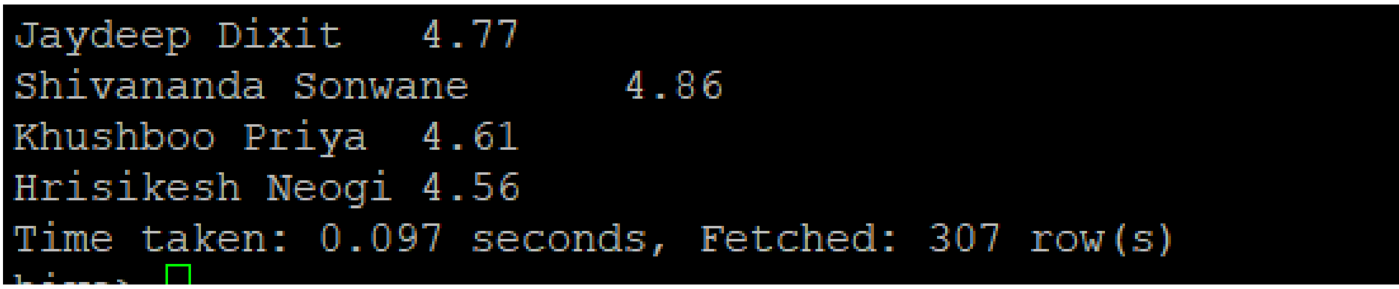
Hive> select Agent_name,Avg_Rating from AgentPerformance where Avg_Rating < 3.5;



```
Sowmiya Sivakumar       0.0
Nitin M 0.0
Vivek   0.0
Ayushi Mishra   0.0
Chaitra K Hiremath      0.0
Time taken: 0.137 seconds, Fetched: 1474 row(s)
```

10. Agent name who have rating more than 4.5

Hive> select Agent_name,Avg_Rating from AgentPerformance where Avg_Rating > 4.5;



```
Jaydeep Dixit   4.77
Shivananda Sonwane      4.86
Khushboo Priya  4.61
Hrisikesh Neogi 4.56
Time taken: 0.097 seconds, Fetched: 307 row(s)
```

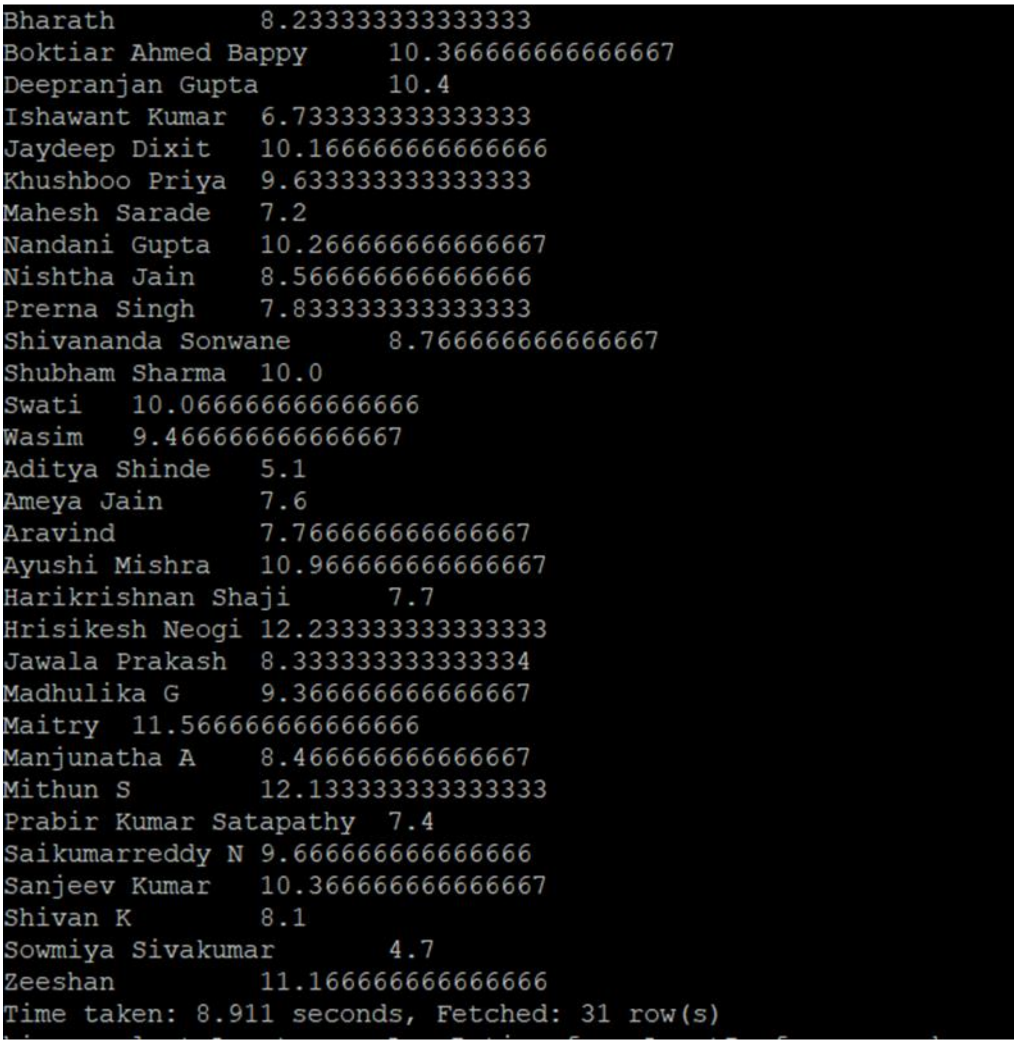
11. How many feedback agents have received more than 4.5 on average

SELECT Agent_name, AVG(Total_Feedback) as avg_feedback

FROM AgentPerformance

GROUP BY Agent_name

HAVING avg_feedback > 4.5;

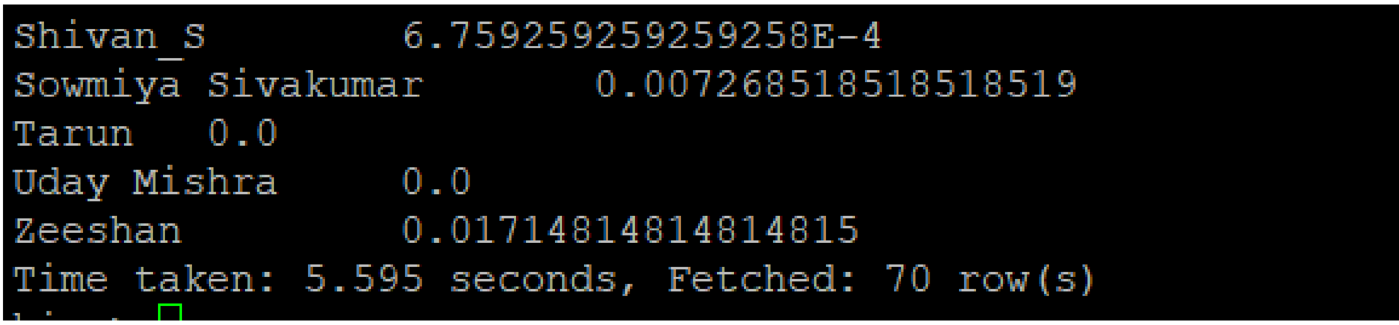


```
Bharath         8.233333333333333
Boktiar Ahmed Bappy     10.366666666666667
Deepranjan Gupta        10.4
Ishawant Kumar  6.733333333333333
Jaydeep Dixit   10.166666666666666
Khushboo Priya  9.633333333333333
Mahesh Sarade   7.2
Nandani Gupta   10.266666666666667
Nishtha Jain    8.566666666666666
Prerna Singh    7.833333333333333
Shivananda Sonwane      8.766666666666667
Shubham Sharma  10.0
Swati   10.066666666666666
Wasim   9.466666666666667
Aditya Shinde   5.1
Ameya Jain      7.6
Aravind         7.766666666666667
Ayushi Mishra   10.966666666666667
Harikrishnan Shaji      7.7
Hrisikesh Neogi 12.233333333333333
Jawala Prakash  8.333333333333334
Madhulika G     9.366666666666667
Maitry  11.566666666666666
Manjunatha A    8.466666666666667
Mithun S        12.133333333333333
Prabir Kumar Satapathy  7.4
Saikumarreddy N 9.666666666666666
Sanjeev Kumar   10.366666666666667
Shivan K        8.1
Sowmiya Sivakumar       4.7
Zeeshan         11.166666666666666
Time taken: 8.911 seconds, Fetched: 31 row(s)
```

12. average weekly response time for each agent.

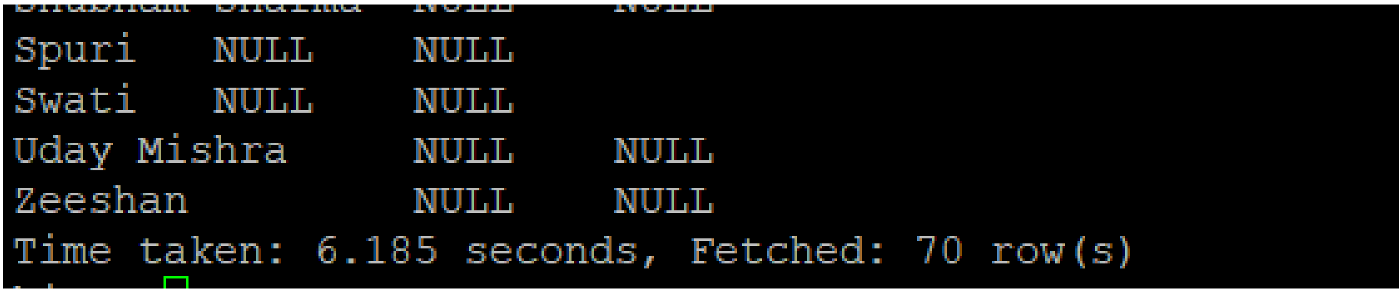
Hive> select s.agent_name,avg(col1[0]*3600+col1[1]*60+substr(col1[2],1,2))/3600 from(

select agent_name,split(Avg_Response_Time,':') as col1 from AgentPerformance)s group by s.agent_name;



```
Shivan_S        6.759259259259258E-4
Sowmiya Sivakumar       0.007268518518518519
Tarun   0.0
Uday Mishra     0.0
Zeeshan         0.017148148148148148
Time taken: 5.595 seconds, Fetched: 70 row(s)
```

Hive> select agent_name,avg(Avg_Response_Time)as Avg_Response_Time,weekofyear(Date) as weekly from AgentPerformance group by agent_name,weekofyear(Date);

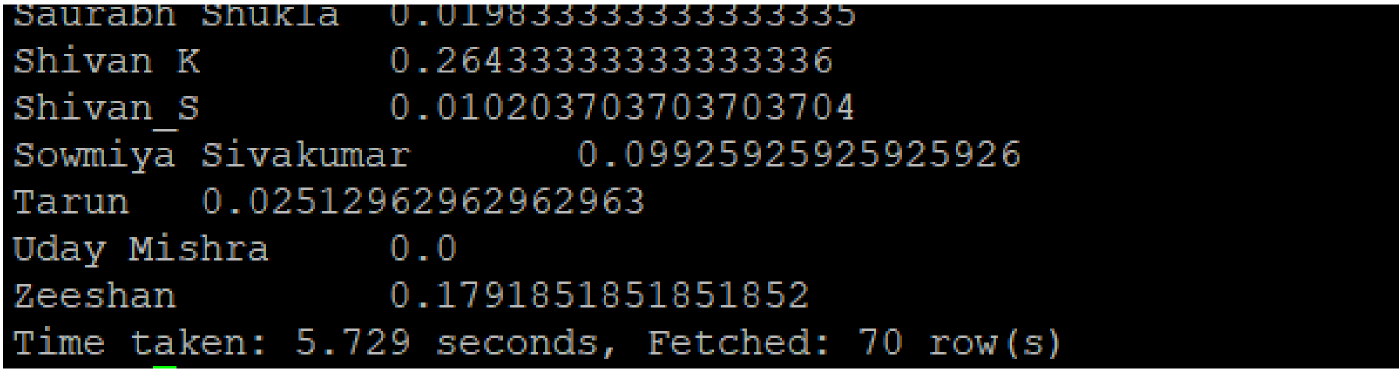


```
Shubham Sharma  NULL    NULL
Spuri   NULL    NULL
Swati   NULL    NULL
Uday Mishra     NULL    NULL
Zeeshan         NULL    NULL
Time taken: 6.185 seconds, Fetched: 70 row(s)
```

13. average weekly resolution time for each agents

Hive> select s.agent_name,avg(col1[0]*3600+col1[1]*60+substr(col1[2],1,2))/3600 from(

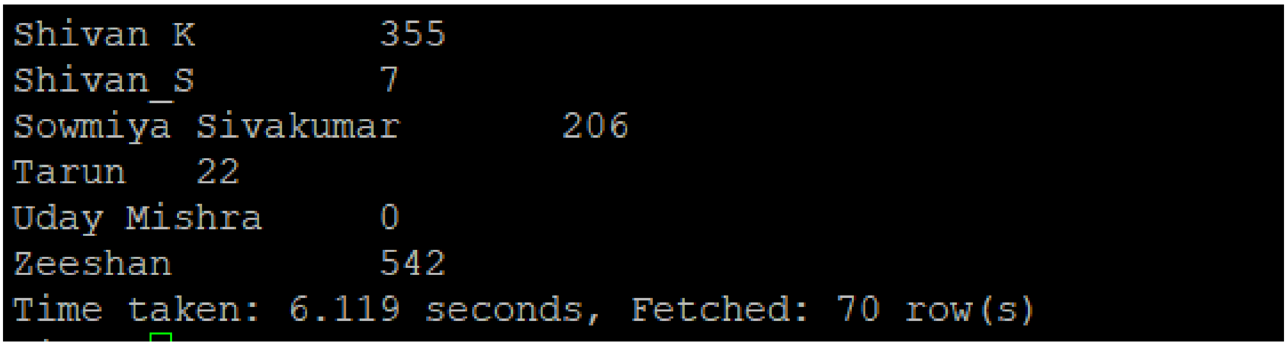
select agent_name,split(Avg_Resolution_Time,':') as col1 from AgentPerformance)s group by s.agent_name;



```
Saurabh Shukla  0.019833333333333335
Shivan K        0.26433333333333336
Shivan_S        0.010203703703703704
Sowmiya Sivakumar       0.09925925925925926
Tarun   0.02512962962962963
Uday Mishra     0.0
Zeeshan         0.17918518518518518
Time taken: 5.729 seconds, Fetched: 70 row(s)
```

14. Find the number of chat on which they have received a feedback

Hive> select agent_name,sum(Total_charts),Total_Feedback from AgentPerformance where Total_Feedback> 0 group by agent_name,Total_Feedback;

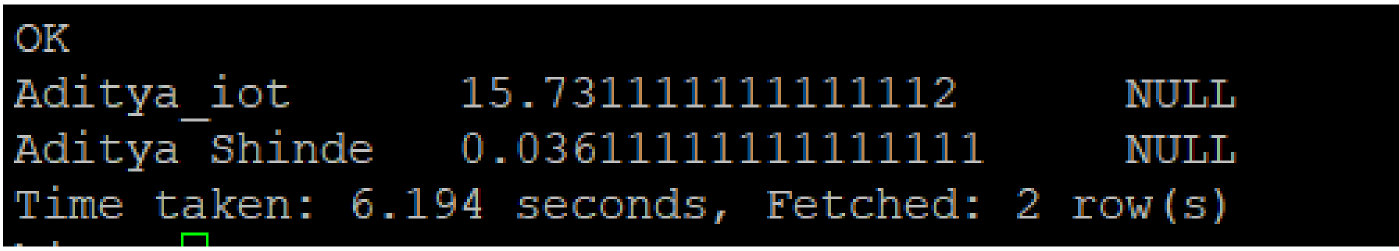


```
Shivan K        355
Shivan_S        7
Sowmiya Sivakumar       206
Tarun   22
Uday Mishra     0
Zeeshan         542
Time taken: 6.119 seconds, Fetched: 70 row(s)
```

15. Total contribution hour for each and every agents weekly basis

Hive> select s.agent,sum(col1[0]*3600+col1[1]*60+col1[2])/3600 timeInHour,s.weekly from(

select agent,split(duration,':') as col1 ,weekofyear(Event_Date) as weekly from AgentLogingReport)s group by s.agent,s.weekly limit 2;



```
OK
Aditya_iot      15.731111111111112      NULL
Aditya Shinde   0.036111111111111111    NULL
Time taken: 6.194 seconds, Fetched: 2 row(s)
```

16. Perform inner join, left join and right join based on the agent column and after joining the table export that data into your local system.

#inner join

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/hadoop/inner_join.csv'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT a.sr_no, a.Agent, a.Event_date, a.Login, a.Logout, a.Duration, b.Total_charts,
b.Avg_Response_Time, b.Avg_Resolution_Time, b.Avg_Rating, b.Total_Feedback

FROM AgentLogingReport a

JOIN AgentPerformance b ON a.Agent = b.Agent_Name;
```

#left join

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/hadoop/left_join.csv'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT a.sr_no, a.Agent, a.Event_date, a.Login, a.Logout, a.Duration, b.Total_charts,
b.Avg_Response_Time, b.Avg_Resolution_Time, b.Avg_Rating, b.Total_Feedback

FROM AgentLogingReport a

LEFT JOIN AgentPerformance b ON a.Agent = b.Agent_Name;
```

# Right join

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/hodoop/right_join.csv'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT a.sr_no, a.Agent, a.Event_date, a.Login, a.Logout, a.Duration, b.Total_charts,
b.Avg_Response_Time, b.Avg_Resolution_Time, b.Avg_Rating, b.Total_Feedback
```

```
FROM AgentLogingReport a

RIGHT JOIN AgentPerformance b ON a.Agent = b.Agent_Name;

hive -e 'select /*+ streamtable(a) */a.agent,a.date,a.Duration,b.Total_charts,b.Total_Feedback from challenge.AgentLogingReport a right join challenge.AgentPerformance b on a.agent = b.agent_name' > /home/cloudera/sidd/Challenge/mini_project_1/left_join.csv;
```

17. Perform partitioning on top of the agent column and then on top of that perform bucketing for each partitioning.

```
Create table AgentLogingReport_partitioned

(

sr_no int,

Event_Date date,

Login string,

Logout string,

Duration string

)partitioned by (Agent string)

CLUSTERED BY (Event_Date) sorted by (Event_Date) INTO 4 BUCKETS

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',';
```

```
hive> set hive.exec.dynamic.partition=true;

hive> set hive.exec.dynamic.partition.mode=nonstrict;

hive> insert into table AgentLogingReport_partitioned partition(Agent) select sr_no,Event_Date,Login,Logout,Duration,Agent from AgentLogingReport;

Hive> Create table AgentPerformance_partitioned

(

sr_no int,

Event_Date date,

Total_charts string,

Avg_Response_Time string,

Avg_Resolution_Time string,
Avg_Rating float,

Total_Feedback int

)partitioned by (Agent_name string)

CLUSTERED BY (Event_Date) sorted by (Event_Date) INTO 8 BUCKETS

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',';
```