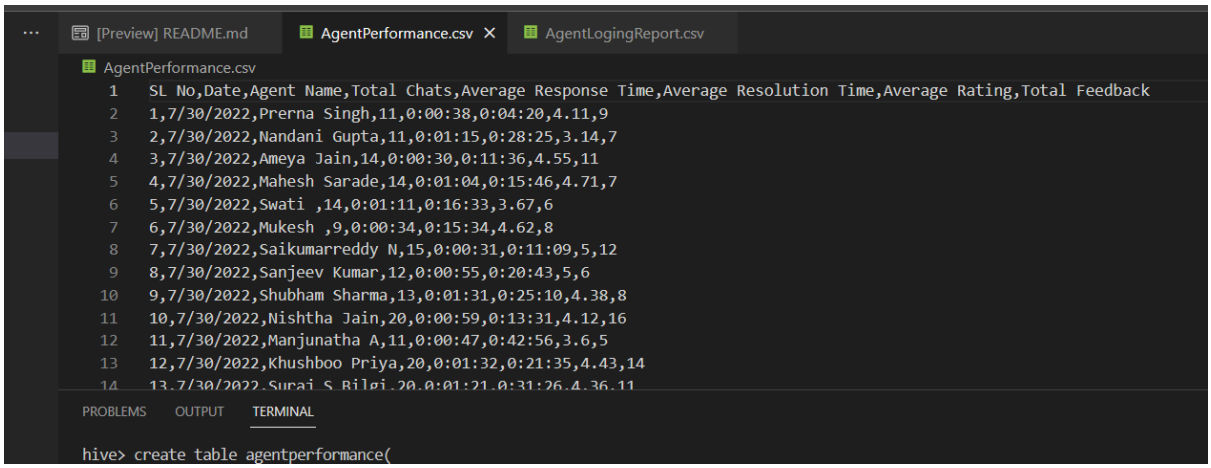
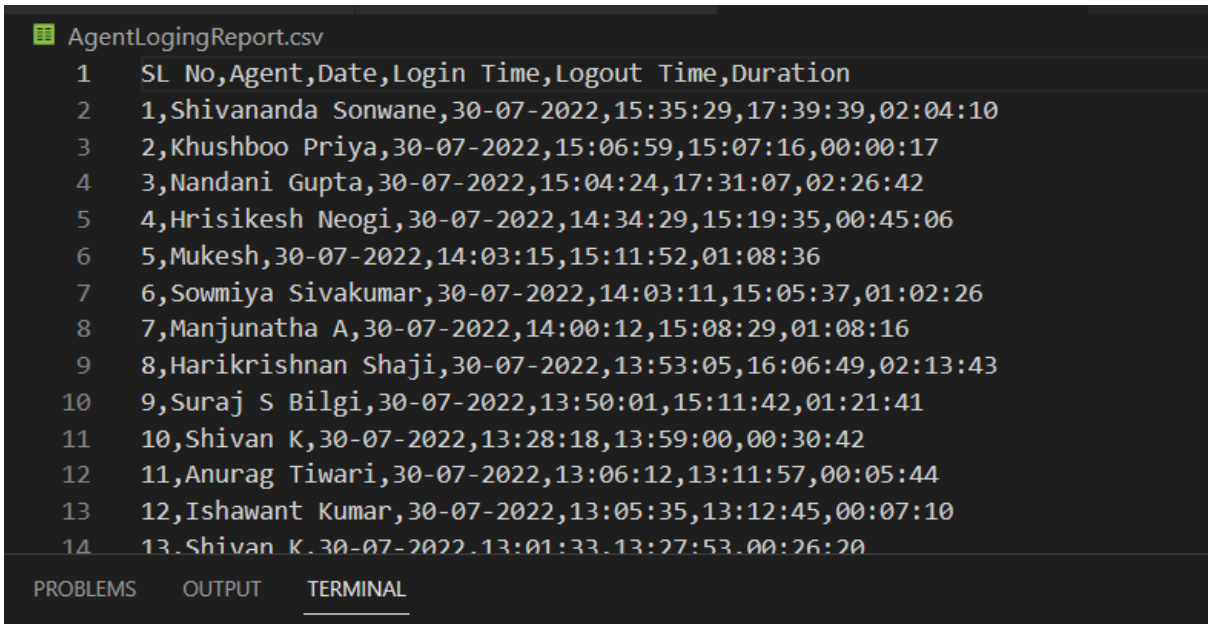
HIVE-Mini Project-1

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

```
hive> create table agentreport(
    > sl int,agent string,loging_date string,login_time string,logout_time string,duration string)
    > row format delimited
    > fields terminated by ','
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.093 seconds
```

```
hive> create table agentperformance(
    > sl int,performance_date string,name string,a_response_t string,a_resolution_t string,average_rating float,feedback int)
    > row format delimited
    > fields terminated by ','
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.065 seconds
```

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

```
abc@46150eb321a4:~/workspace$ hadoop fs -put '/config/workspace/AgentPerformance.csv' /tmp
2023-02-25 16:00:14,463 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
abc@46150eb321a4:~/workspace$ hadoop fs -ls /tmp
Found 3 items
-rw-r--r--   1 abc supergroup      56352 2023-02-25 15:59 /tmp/AgentLogingReport.csv
-rw-r--r--   1 abc supergroup     109853 2023-02-25 16:00 /tmp/AgentPerformance.csv
drwx-wx-wx   - abc supergroup          0 2023-02-25 14:59 /tmp/hive
```

3. List of all agents' names.

```
hive> select agent from agentreport;
OK
agent
Shivananda Sonwane
Khushboo Priya
Nandani Gupta
Hrisikesh Neogi
Mukesh
Sowmiya Sivakumar
Manjunatha A
Harikrishnan Shaji
Suraj S Bilgi
Shivan K
Anurag Tiwari
Ishawant Kumar
Shivan K
Shubham Sharma
Shivan K
Prerna Singh
Shivan K
```

```
hive> select name from agentperformance limit 50;
OK
name
Prerna Singh
Nandani Gupta
Ameya Jain
Mahesh Sarade
Swati
Mukesh
Saikumarreddy N
Sanjeev Kumar
Shubham Sharma
Nishtha Jain
Manjunatha A
Khushboo Priya
Suraj S Bilgi
```

4. Find out agent average rating.

```
hive> select name,avg(average_rating) from agentperformance
    > group by name;
Query ID = abc_20230225161623_a9c98472-f3fb-4075-9271-c1728102dd03
```

```
OK
name    _c1
Abhishek        0.0
Aditya  0.0
Aditya Shinde   1.8003333409627278
Aditya_iot      2.3453333377838135
Amersh  0.0
Ameya Jain      2.21966667175293
Anirudh         0.6449999968210857
Ankit Sharma    0.0
Ankitjha        0.26666666666666666
Anurag Tiwari   0.18333333333333332
Aravind         2.1813333511352537
Ashad Nasim     0.16666666666666666
Ashish  0.0
Ayushi Mishra   3.481999969482422
Bharath         2.9836666584014893
Boktiar Ahmed Bappy     3.567999982833862
```

5. Total working days for each agents

```
hive> select agent,count(loging_date) as total_workingdays from agentreport
    > group by agent;
```

```
agent   total_workingdays
Aditya Shinde   1
Aditya_iot      9
Amersh  4
Ameya Jain      10
Ankitjha        4
Anurag Tiwari   37
Aravind 10
Ayushi Mishra   18
Bharath 9
Boktiar Ahmed Bappy     17
Chaitra K Hiremath      13
Deepranjan Gupta        58
Dibyanshu       208
Harikrishnan Shaji      23
Hrisikesh Neogi 37
Hyder Abbas     2
Ineuron Intelligence    1
```

6. Total query that each agent have taken

```
hive> select name,sum(chat) as total_query from agentperformance
    > group by name;
```

```
name    total_query
Abhishek        0
Aditya  0
Aditya Shinde   277
Aditya_iot      231
Amersh  0
Ameya Jain      322
Anirudh         81
Ankit Sharma    0
Ankitjha        5
Anurag Tiwari   4
Aravind         366
Ashad Nasim     18
Ashish  0
Ayushi Mishra   514
Bharath         369
Boktiar Ahmed Bappy     452
Chaitra K Hiremath      64
Deepranjan Gupta        493
Dibyanshu       1
Harikrishnan Shaji      381
```

7. Total Feedback that each agent have received

```
hive> select name,sum(feedback) as total_query from agentperformance
    > group by name;
```

```
Sanjeev Kumar   311
Sanjeevan       0
Saurabh Shukla  8
Shiva Srivastava        46
Shivan K        243
Shivan_S        4
Shivananda Sonwane      263
Shubham Sharma  300
Sowmiya Sivakumar       141
Spuri   0
Sudhanshu Kumar 2
Suraj S Bilgi   15
Swati   302
Tarun   6
Uday Mishra     0
Vasanth P       0
Vivek   20
Wasim   284
Zeeshan         335
Time taken: 23.581 seconds, Fetched: 70 row(s)
```

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

```
hive> select name from agentperformance where average_rating between 3.5 and 4;
OK
```

```
name
Swati
Manjunatha A
Boktiar Ahmed Bappy
Prateek _iot
Nandani Gupta
Jaydeep Dixit
Mahesh Sarade
Zeeshan
Hrisikesh Neogi
Muskan Garg
Khushboo Priya
Wasim
Jawala Prakash
Shiva Srivastava
```

9. Agent name who have rating less than 3.5

```
hive> select name from agentperformance where average_rating>3.5;
OK
```

```
name
Prerna Singh
Ameya Jain
Mahesh Sarade
Swati
Mukesh
Saikumarreddy N
Sanjeev Kumar
Shubham Sharma
Nishtha Jain
Manjunatha A
Khushboo Priya
Suraj S Bilgi
```

10. Agent name who have rating more than 4.5

```
hive> select name from agentperformance where average_rating>4.5;
OK
```

```
name
Ameya Jain
Mahesh Sarade
Mukesh
Saikumarreddy N
Sanjeev Kumar
Harikrishnan Shaji
Sowmiya Sivakumar
Boktiar Ahmed Bappy
Shivananda Sonwane
Ishawant Kumar
Deepranjan Gupta
Shivananda Sonwane
Muskan Garg
Aditya_iot
```

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

```
hive> select name,count(feedback) from agentperformance where average_rating>4.5
    > group by name;
```

```
Aditya Shinde   7
Aditya_iot      6
Ameya Jain      8
Anirudh         1
Ankitjha        1
Aravind         11
Ayushi Mishra   8
Bharath         17
Boktiar Ahmed Bappy     8
Chaitra K Hiremath      2
Deepranjan Gupta        7
Harikrishnan Shaji      7
Hrisikesh Neogi 11
Ishawant Kumar  12
Jawala Prakash  5
Jayant Kumar    2
Jaydeep Dixit   13
```

12. average weekly response time for each agent

```
hive> SELECT
    >     name,
    >     AVG(a_response_t) AS Avg_Response_Time_Weekly
    > FROM
    >     agentperformance
    > GROUP BY
    >     name,performance_date;
```

```
Mahesh Sarade   0.0
Maitry  0.0
Maneesh         0.0
Manjunatha A    0.0
Mithun S        0.0
Mukesh  0.0
Mukesh Rao      0.0
Muskan Garg     0.0
Nandani Gupta   0.0
Nishtha Jain    0.0
```

13. average weekly resolution time for each agents

```
hive> SELECT
    >     name,
    >     AVG(a_resolution_t) AS Avg_Resolution_Time_Weekly
    > FROM
    >     agentperformance
    > GROUP BY
    >     name,performance_date;
```

```
Prerna Singh    0.0
Rishav Dash     0.0
Rohan   0.0
Saif Khan       0.0
Saikumarreddy N 0.0
Samprit         0.0
Sandipan Saha   0.0
Sanjeev Kumar   0.0
Sanjeevan       0.0
Saurabh Shukla  0.0
```

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

```
hive> SELECT name, COUNT(DISTINCT chat) AS num_feedback_chats
    > FROM agentperformance
    > GROUP BY name;
```

```
Abhishek        1
Aditya  1
Aditya Shinde   12
Aditya_iot      13
Amersh  1
Ameya Jain      14
Anirudh         8
Ankit Sharma    1
Ankitjha        3
Anurag Tiwari   3
Aravind         14
Ashad Nasim     3
Ashish  1
Ayushi Mishra   14
Bharath         12
Boktiar Ahmed Bappy     22
```

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

```
hive> SELECT Agent, DATE_TRUNC('week',Date) AS week_start,
    >        SUM(Duration) AS contribution_hours
    > FROM agentreport
    > GROUP BY Agent, week_start;
```

```
Dibyanshu       00:02:00
Shivan K        00:18:10
Harikrishnan Shaji      00:02:10
Khushboo Priya  00:01:51
Khushboo Priya  04:18:13
Deepranjan Gupta        02:26:51
Dibyanshu       00:02:00
Prabir Kumar Satapathy  01:37:21
Nishtha Jain    00:01:48
Nishtha Jain    04:28:10
Dibyanshu       00:02:00
Dibyanshu       00:02:00
Rishav Dash     02:08:42
Rishav Dash     04:23:38
Prabir Kumar Satapathy  00:53:04
Prabir Kumar Satapathy  00:01:50
Shivan K        00:41:13
Shivan K        00:01:49
```

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.

Left join

```
hive> insert overwrite local directory '/config/workspace/left_join_op.csv'
    > row format delimited
    > fields terminated by ','
    > select ap.name,ar.agent from agentperformance ap
    > left join agentreport ar
    > on ap.name=ar.agent limit 5;
```

```
11      Prerna Singh
11      Prerna Singh
11      Prerna Singh
11      Prerna Singh
11      Prerna Singh
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
11      Nandani Gupta
14      Ameya Jain
```

Inner Join

```
hive> insert overwrite local directory '/config/workspace/Output.csv'
    > row format delimited
    > fields terminated by ','
    > select ap.name,ar.login_time from agentperformance ap
    > inner join agentreport ar
    > on ap.name=ar.agent limit 5;
```

```
OK
ap.name ar.login_time
Prerna Singh    12:32:28
Prerna Singh    17:47:06
Prerna Singh    15:08:22
Prerna Singh    12:08:23
Prerna Singh    13:11:06
Time taken: 79.497 seconds, Fetched: 5 row(s)
```

Right Join

```
hive> insert overwrite local directory '/config/workspace/Output.csv'
    > row format delimited
    > fields terminated by ','
    > select ap.sl,ar.logout_time from agentperformance ap
    > right join agentreport ar
    > on ap.name=ar.agent limit 5;
```

```
ap.sl   ar.logout_time
69      17:39:39
73      17:39:39
214     17:39:39
285     17:39:39
360     17:39:39
Time taken: 260.569 seconds, Fetched: 5 row(s)
```

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

Dynamic Partition

```
hive> create table agentreport_dynamic2(
    > sl int,agent string,login_time string,logout_time string,duration string)
    > partitioned by (loging_date string);
OK
Time taken: 0.271 seconds
```

```
hive> insert overwrite table agentreport_dynamic2 partition(loging_date)
    > select sl,agent,loging_date,login_time,logout_time,duration from agentreport;
```

Static Partition

```
hive> create table agentreport_static(
    > sl int,loging_date string,login_time string,logout_time string,duration string)
    > partitioned by (agent string);
OK
```

```
hive> insert overwrite table agentreport_static2 partition(agent='Shivan K')
    > select sl,loging_date,login_time,logout_time,duration from agentreport where agent='Shivan K';
```

```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-03-04 15:54:18,105 Stage-1 map = 0%,  reduce = 0%
2023-03-04 15:54:49,336 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 45.06 sec
2023-03-04 15:55:04,741 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 52.5 sec
MapReduce Total cumulative CPU time: 52 seconds 500 msec
Ended Job = job_1677925174679_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost/user/hive/warehouse/agentreport_static/agent=Shivan K/.hive-staging_hive_2023-03-04_15-53-25_643_6214502962
940448936-1/-ext-10000
Loading data to table default.agentreport_static partition (agent=Shivan K)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 52.5 sec   HDFS Read: 75679 HDFS Write: 2650 SUCCESS
Total MapReduce CPU Time Spent: 52 seconds 500 msec
OK
Time taken: 106.763 seconds
```

```
hive> select *from agentreport_static;
OK
10      30-07-2022      13:28:18        13:59:00        00:30:42        Shivan K
13      30-07-2022      13:01:33        13:27:53        00:26:20        Shivan K
15      30-07-2022      12:34:27        12:40:37        00:06:10        Shivan K
17      30-07-2022      12:26:15        12:28:15        00:02:00        Shivan K
18      30-07-2022      11:54:57        12:13:38        00:18:40        Shivan K
24      30-07-2022      09:22:39        11:54:03        02:31:23        Shivan K
101     29-07-2022      13:57:28        15:05:42        01:08:14        Shivan K
119     29-07-2022      10:13:53        13:23:36        03:09:42        Shivan K
122     29-07-2022      09:37:02        10:11:39        00:34:36        Shivan K
192     28-07-2022      16:27:08        19:44:43        03:17:34        Shivan K
194     28-07-2022      16:14:24        16:16:24        00:02:00        Shivan K
205     28-07-2022      14:39:46        15:16:17        00:36:31        Shivan K
208     28-07-2022      14:07:23        14:29:24        00:22:00        Shivan K
248     28-07-2022      02:22:50        02:35:50        00:13:00        Shivan K
422     26-07-2022      15:15:04        16:26:35        01:11:30        Shivan K
423     26-07-2022      15:15:04        15:16:52        00:01:47        Shivan K
```

Bucketing

```
hive> create table bucket_table1 (sl int,agent string,loging_date string,login_time string,logout_time string,duration string)
    > clustered by(sl)
    > sorted by(sl)
    > into 2 buckets;
OK
Time taken: 0.288 seconds
```

```
hive> insert overwrite table bucket_table1 select * from agentreport_static;
```

```
OK
10      30-07-2022      13:28:18        13:59:00        00:30:42        Shivan K
18      30-07-2022      11:54:57        12:13:38        00:18:40        Shivan K
24      30-07-2022      09:22:39        11:54:03        02:31:23        Shivan K
122     29-07-2022      09:37:02        10:11:39        00:34:36        Shivan K
192     28-07-2022      16:27:08        19:44:43        03:17:34        Shivan K
194     28-07-2022      16:14:24        16:16:24        00:02:00        Shivan K
208     28-07-2022      14:07:23        14:29:24        00:22:00        Shivan K
248     28-07-2022      02:22:50        02:35:50        00:13:00        Shivan K
422     26-07-2022      15:15:04        16:26:35        01:11:30        Shivan K
456     26-07-2022      11:28:38        11:29:16        00:00:38        Shivan K
520     25-07-2022      14:33:13        14:54:03        00:20:50        Shivan K
524     25-07-2022      13:24:29        14:23:52        00:59:22        Shivan K
526     25-07-2022      12:47:41        13:21:02        00:33:21        Shivan K
532     25-07-2022      11:52:24        12:24:55        00:32:30        Shivan K
540     25-07-2022      09:21:30        10:14:51        00:53:20        Shivan K
784     22-07-2022      11:13:53        11:36:54        00:23:00        Shivan K
786     22-07-2022      10:55:29        11:13:39        00:18:10        Shivan K
802     22-07-2022      09:34:52        09:36:41        00:01:49        Shivan K
13      30-07-2022      13:01:33        13:27:53        00:26:20        Shivan K
15      30-07-2022      12:34:27        12:40:37        00:06:10        Shivan K
17      30-07-2022      12:26:15        12:28:15        00:02:00        Shivan K
101     29-07-2022      13:57:28        15:05:42        01:08:14        Shivan K
119     29-07-2022      10:13:53        13:23:36        03:09:42        Shivan K
205     28-07-2022      14:39:46        15:16:17        00:36:31        Shivan K
423     26-07-2022      15:15:04        15:16:52        00:01:47        Shivan K
473     26-07-2022      09:18:43        09:37:06        00:18:22        Shivan K
535     25-07-2022      10:39:20        11:52:07        01:12:46        Shivan K
565     24-07-2022      17:34:03        18:04:47        00:30:44        Shivan K
589     24-07-2022      14:03:38        15:47:47        01:44:09        Shivan K
595     24-07-2022      12:12:49        13:54:51        01:42:02        Shivan K
599     24-07-2022      11:14:26        12:07:57        00:53:31        Shivan K
603     24-07-2022      09:20:18        11:13:07        01:52:49        Shivan K
689     23-07-2022      09:10:59        15:14:10        06:03:11        Shivan K
763     22-07-2022      13:07:17        15:14:18        02:07:01        Shivan K
767     22-07-2022      12:41:02        13:06:13        00:25:11        Shivan K
801     22-07-2022      09:34:52        10:16:06        00:41:13        Shivan K
```