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

**# creating table**

**Hive> Create table AgentLogingReport**

**(**

**sr\_no int,**

**Agent string,**

**Date date,**

**Login string,**

**Logout string,**

**Duration string**

**)**

**row format delimited**

**fields terminated by ','**

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

**Hive> Create table AgentPerformance**

**(**

**sr\_no int,**

**Date date,**

**Agent\_Name string,**

**Total\_charts string,**

**Avg\_Response\_Time string,**

**Avg\_Resolution\_Time string,**

**Avg\_Rating float,**

**Total\_Feedback int**

**)**

**row format delimited**

**fields terminated by ','**

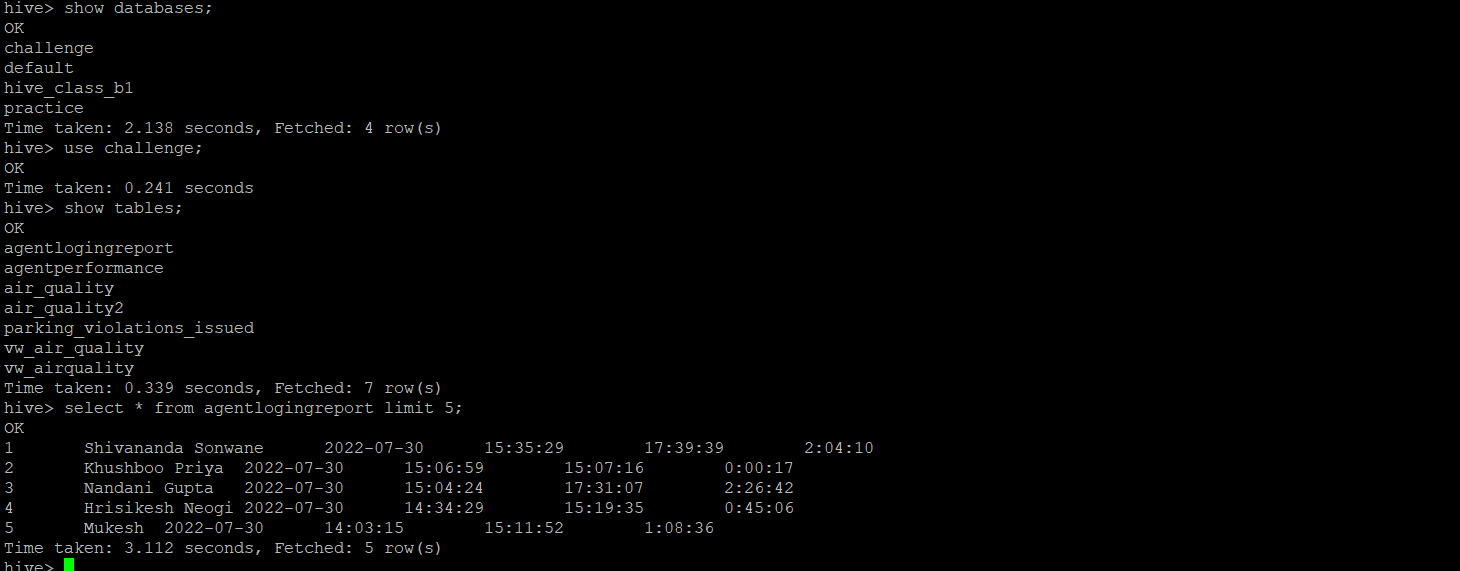
**tblproperties ("skip.header.line.count" = "1");**

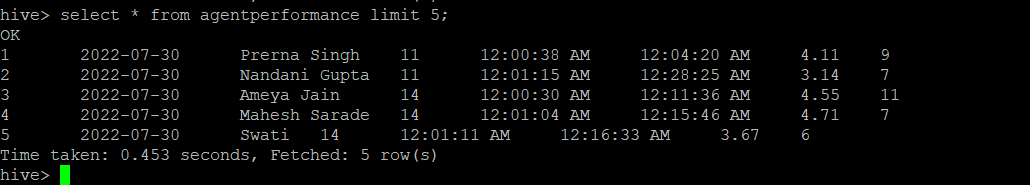
**hive> alter table AgentPerformance change Total\_charts Total\_charts int;**

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

**Load data local inpath’/home/cloudera/sidd/Challenge/mini \_project\_1/AgentLogingReport.csv’ into table AgentLogingReport;**

**Load data local inpath’/home/cloudera/sidd/Challenge/mini \_project\_1/ AgentPerformance.csvinto table AgentPerformance;**

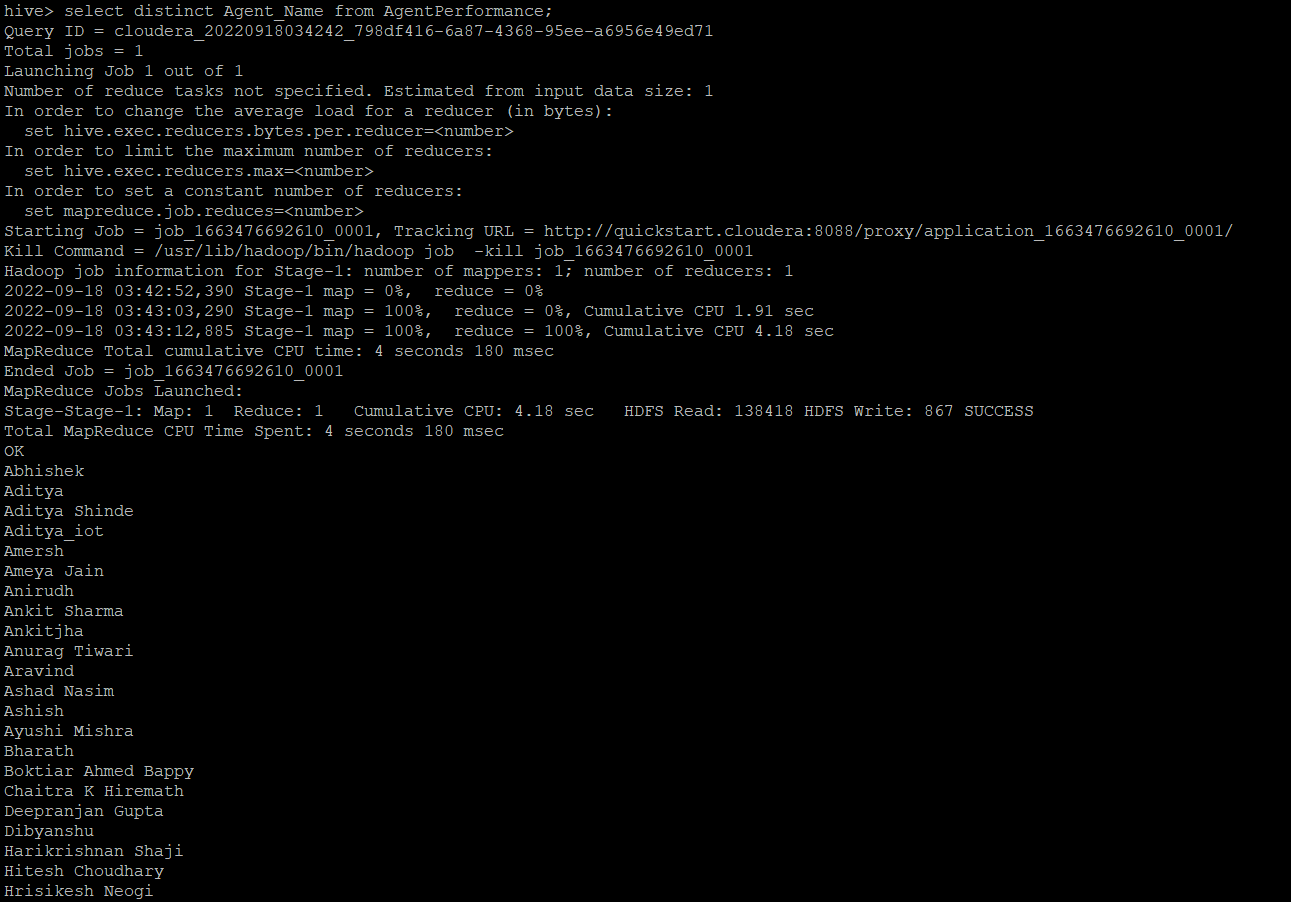


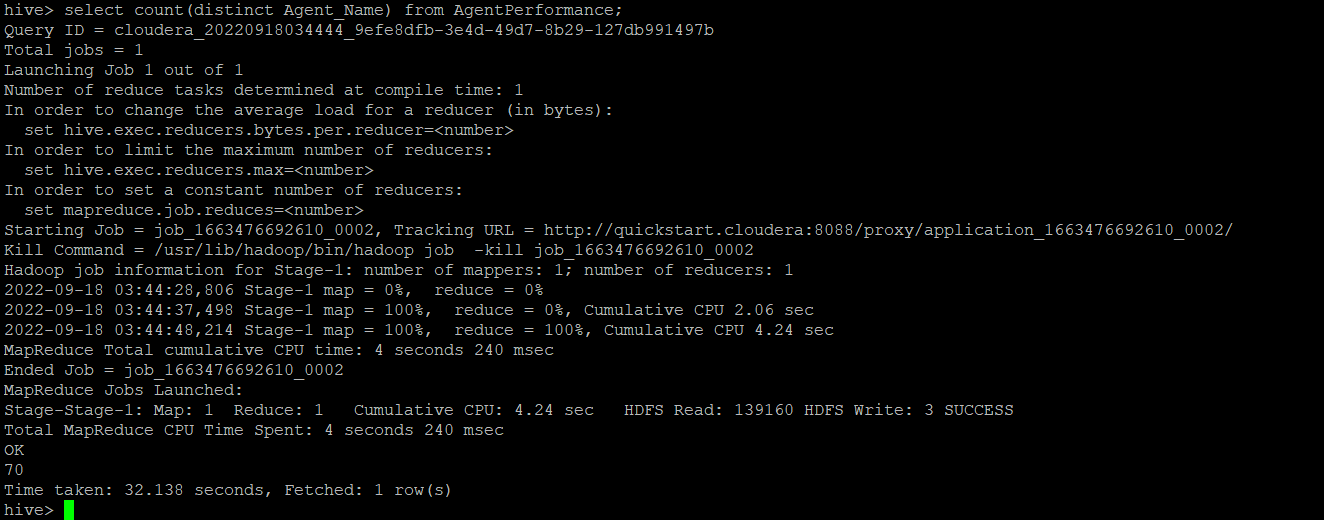


3. List of all agents' names.

**Hive> select distinct Agent\_Name from AgentPerformance;**

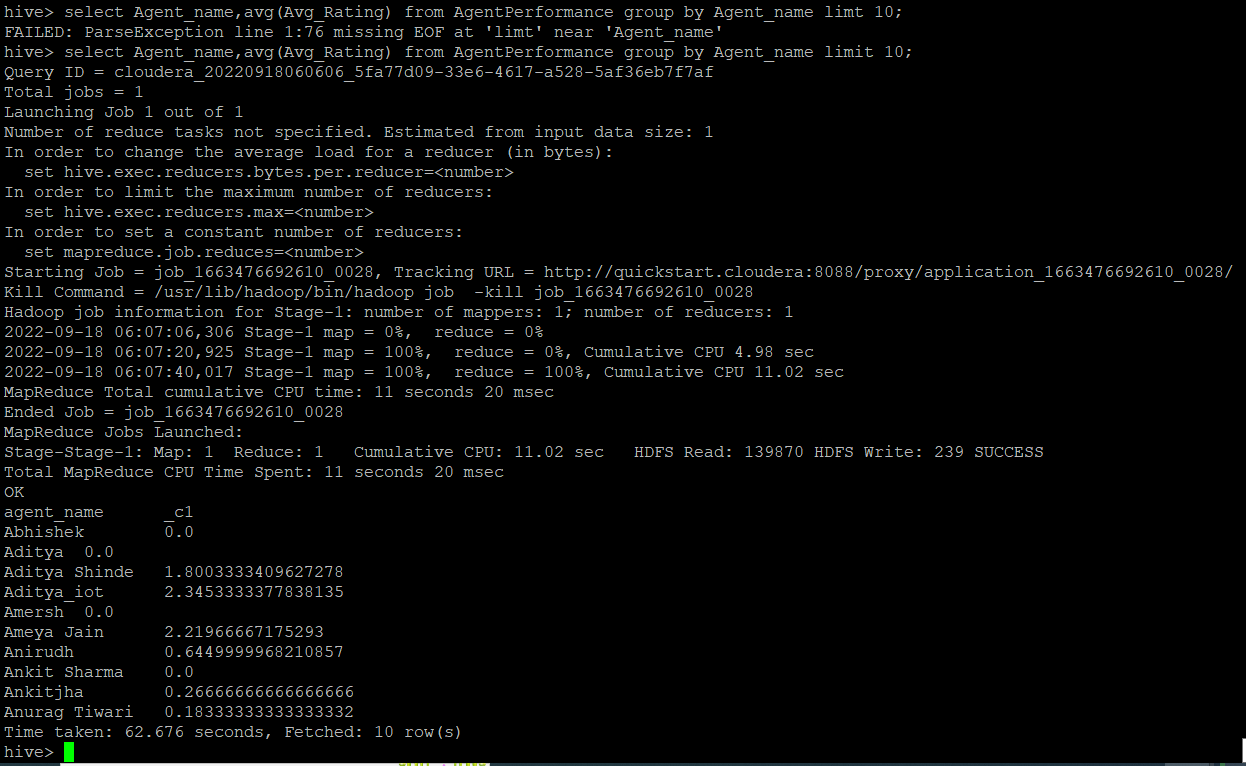
**Hive> select count(distinct Agent\_Name) from AgentPerformance;**





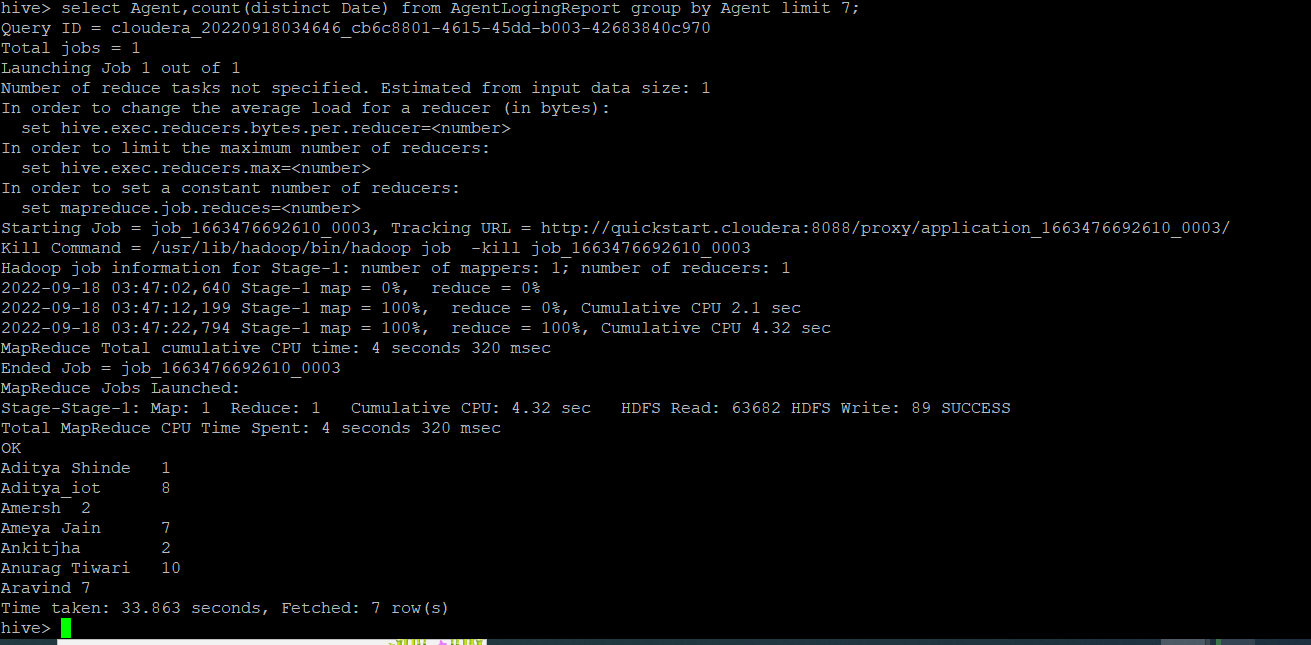
4. Find out agent average rating.

**Hive> select Agent\_name,avg(Avg\_Rating) from AgentPerformance group by Agent\_name;**



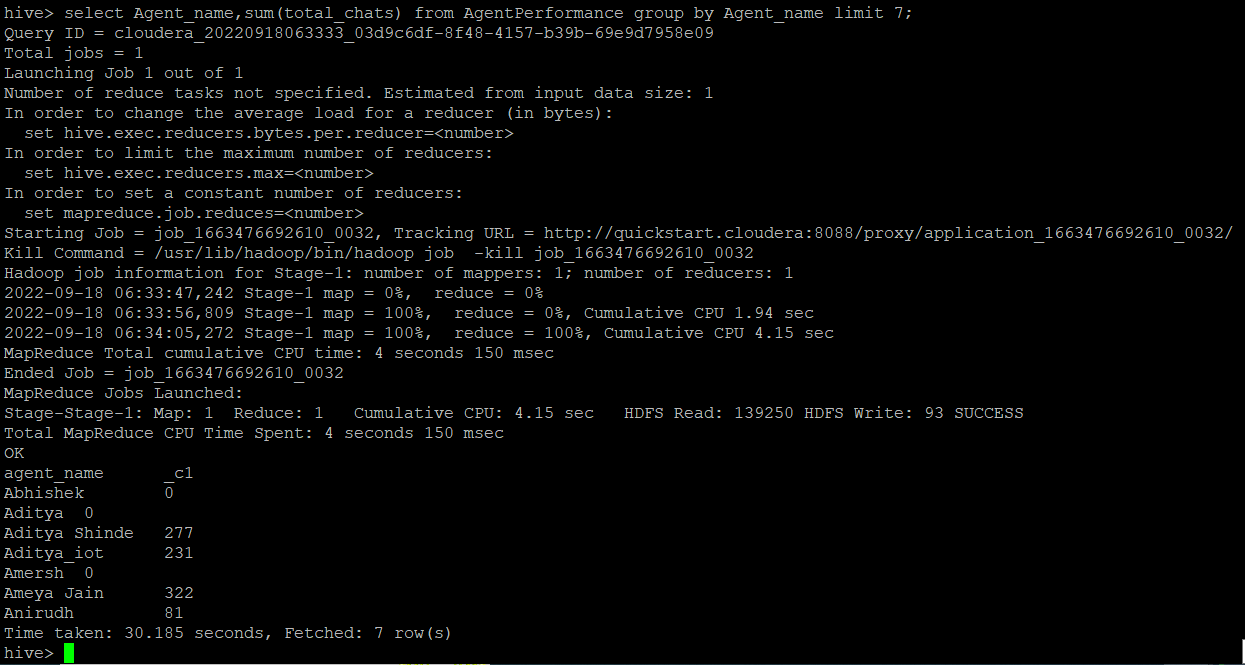
5. Total working days for each agents

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



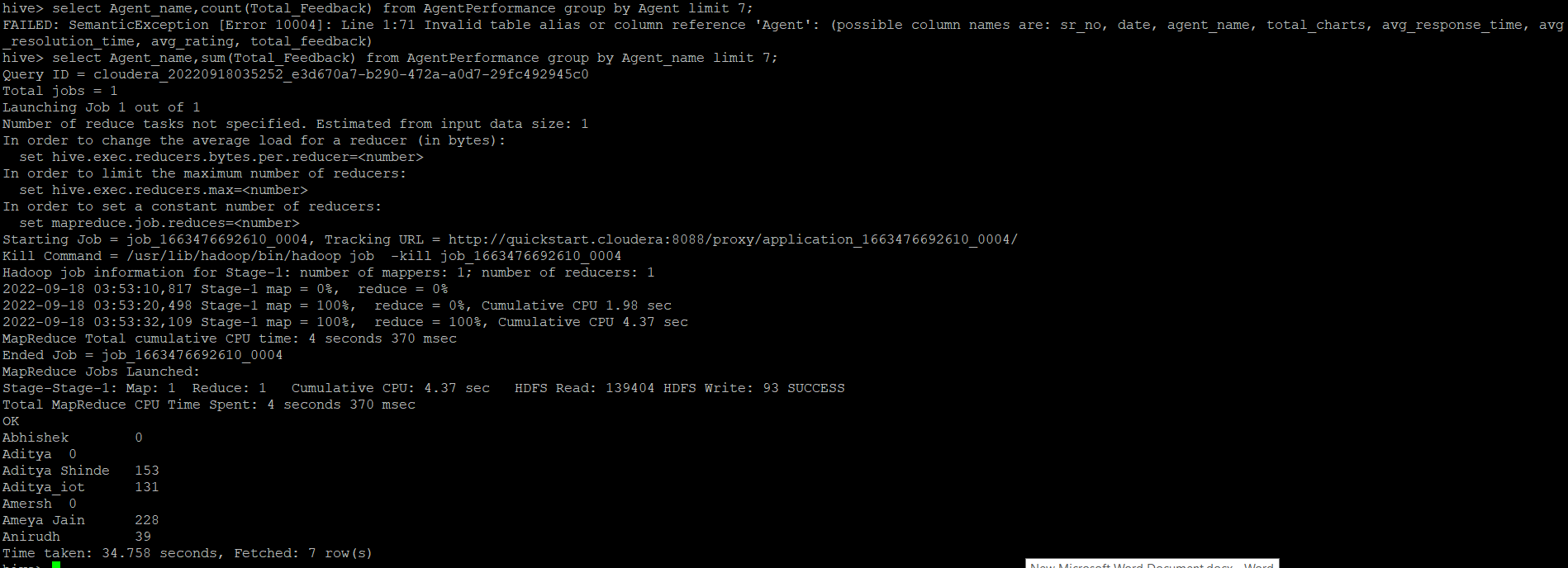
6. Total query that each agent have taken

**Hive> select Agent\_name,sum(total\_chats) from AgentPerformance group by Agent\_name;**



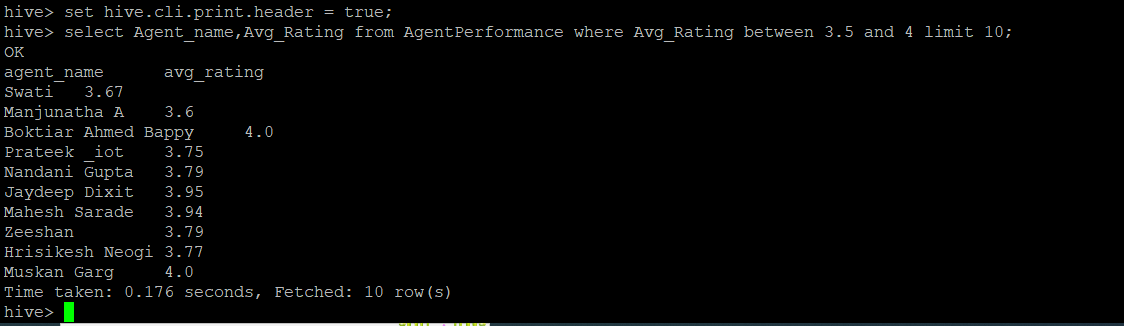
7. Total Feedback that each agent have received

**Hive> select Agent\_name,sum(Total\_Feedback) from AgentPerformance group by Agent\_name;**



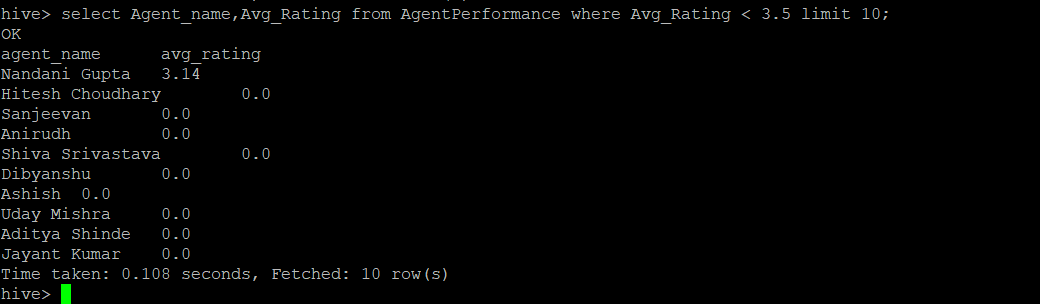
8. Agent name who have average rating between 3.5 to 4

**Hive> select Agent\_name,Avg\_Rating from AgentPerformance where Avg\_Rating between 3.5 and 4;**



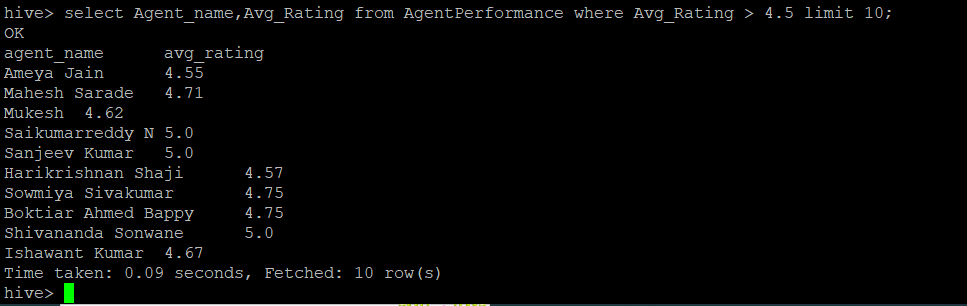
9. Agent name who have rating less than 3.5

**Hive> select Agent\_name,Avg\_Rating from AgentPerformance where Avg\_Rating < 3.5;**



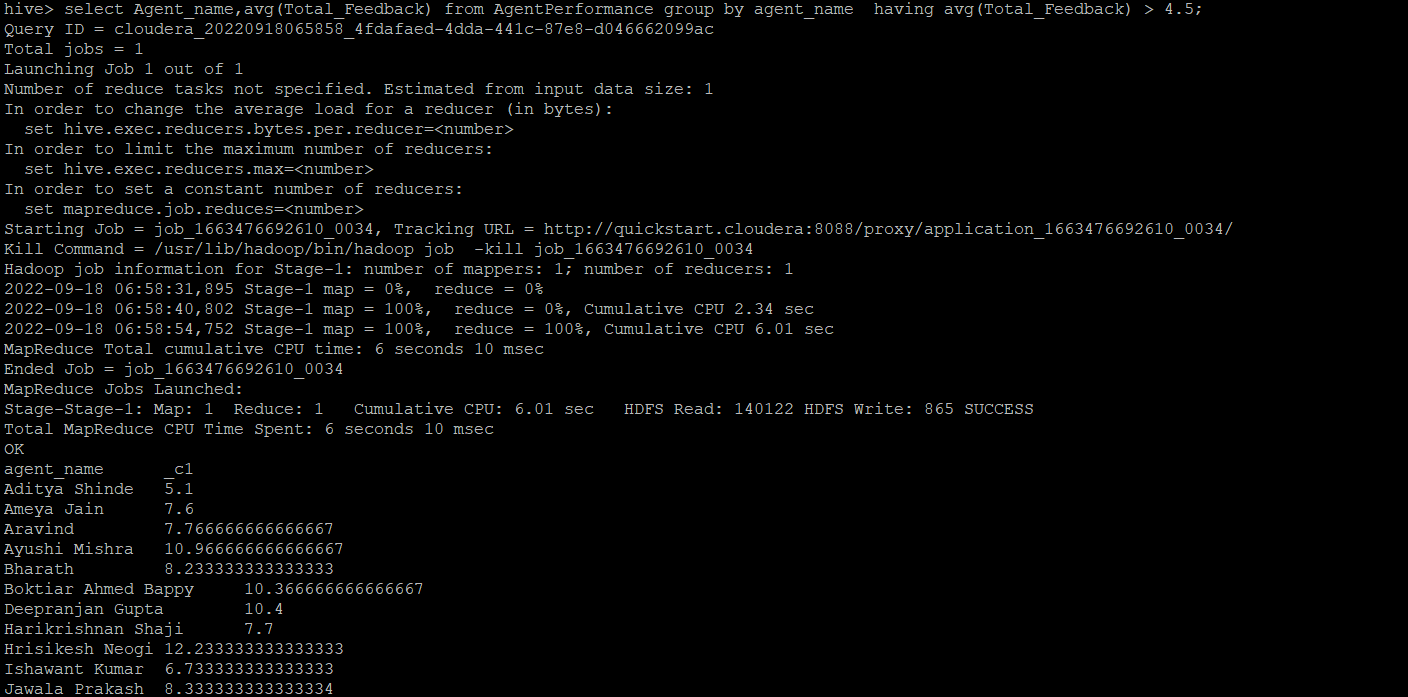
10. Agent name who have rating more than 4.5

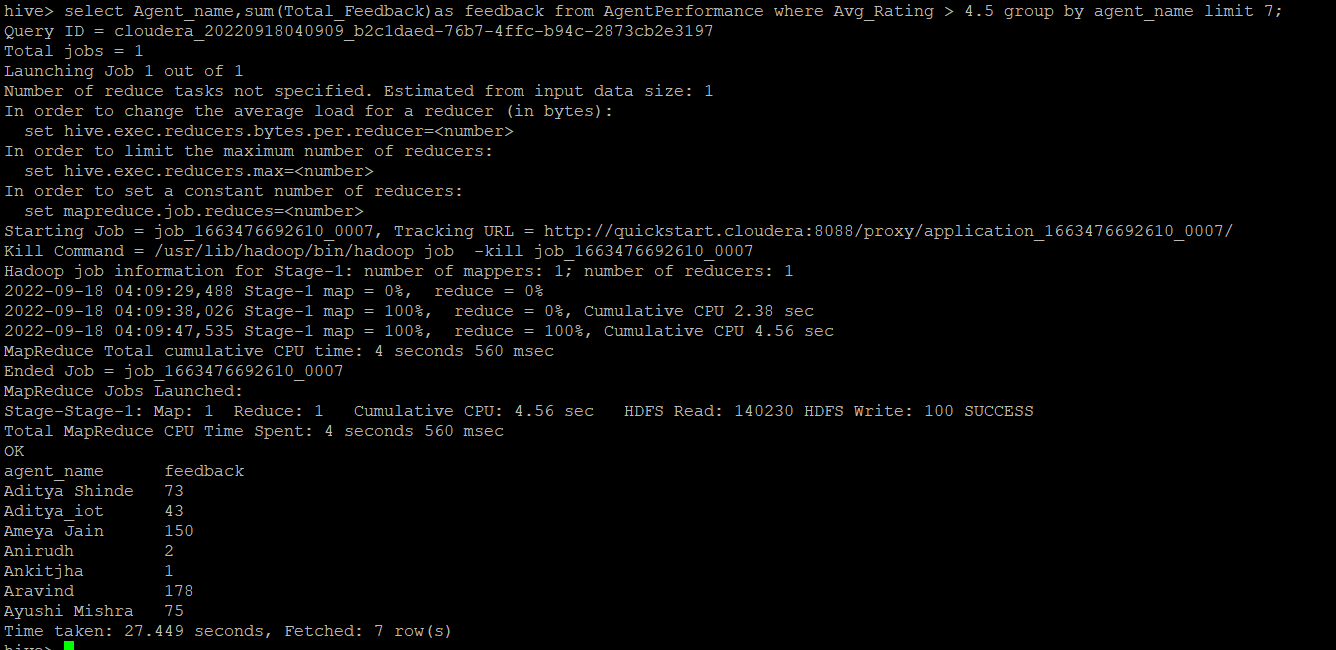
**Hive> select Agent\_name,Avg\_Rating from AgentPerformance where Avg\_Rating > 4.5;**



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

**Hive> select Agent\_name,avg(Total\_Feedback) from AgentPerformance having avg(Total\_Feedback) > 4.5;**

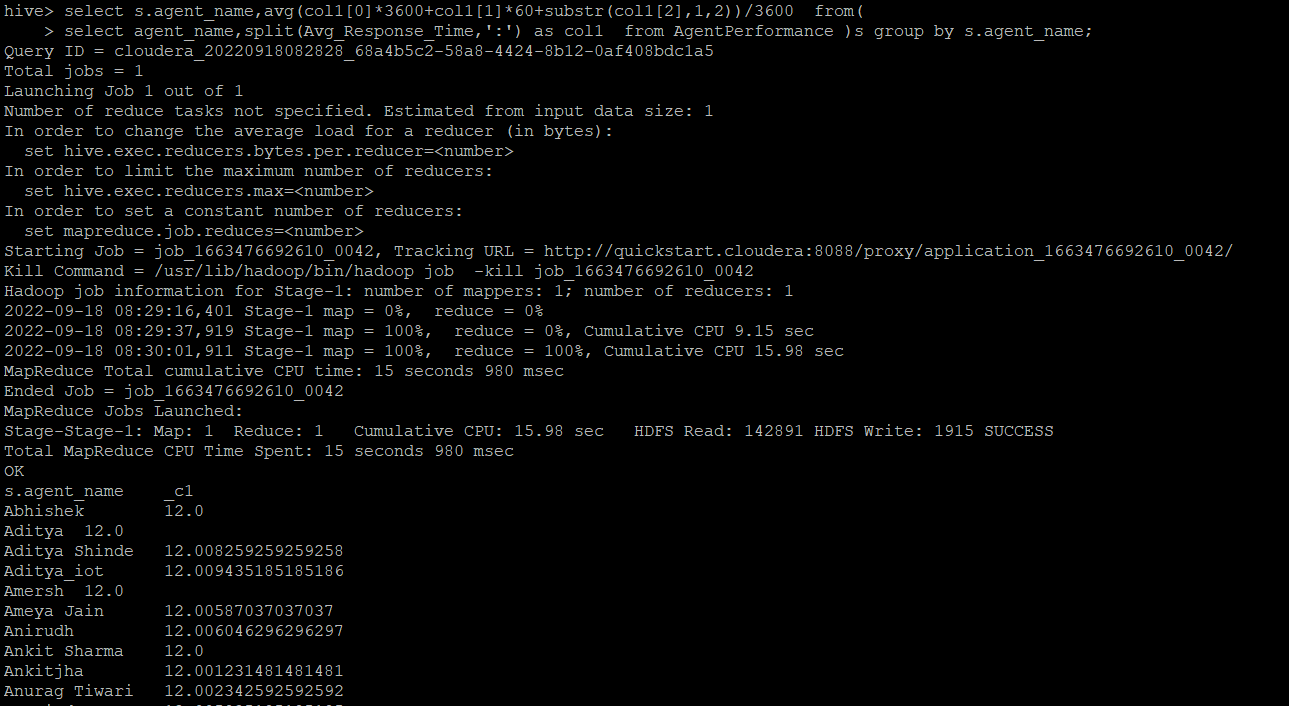




12. average weekly response time for each agent.

**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;**

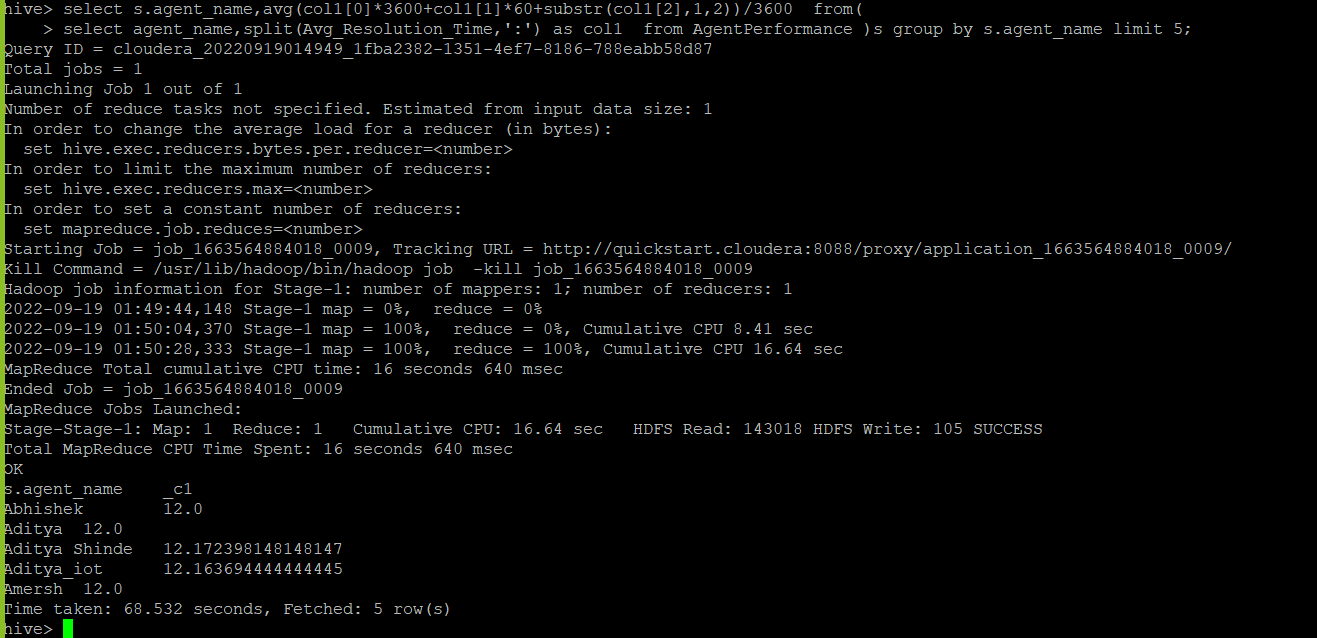


**Hive> select agent\_name,avg(Avg\_Response\_Time)as Avg\_Response\_Time,weekofyear(Date) as weekly from AgentPerformance group by agent\_name,weekofyear(Date);**

13. average weekly resolution time for each agents

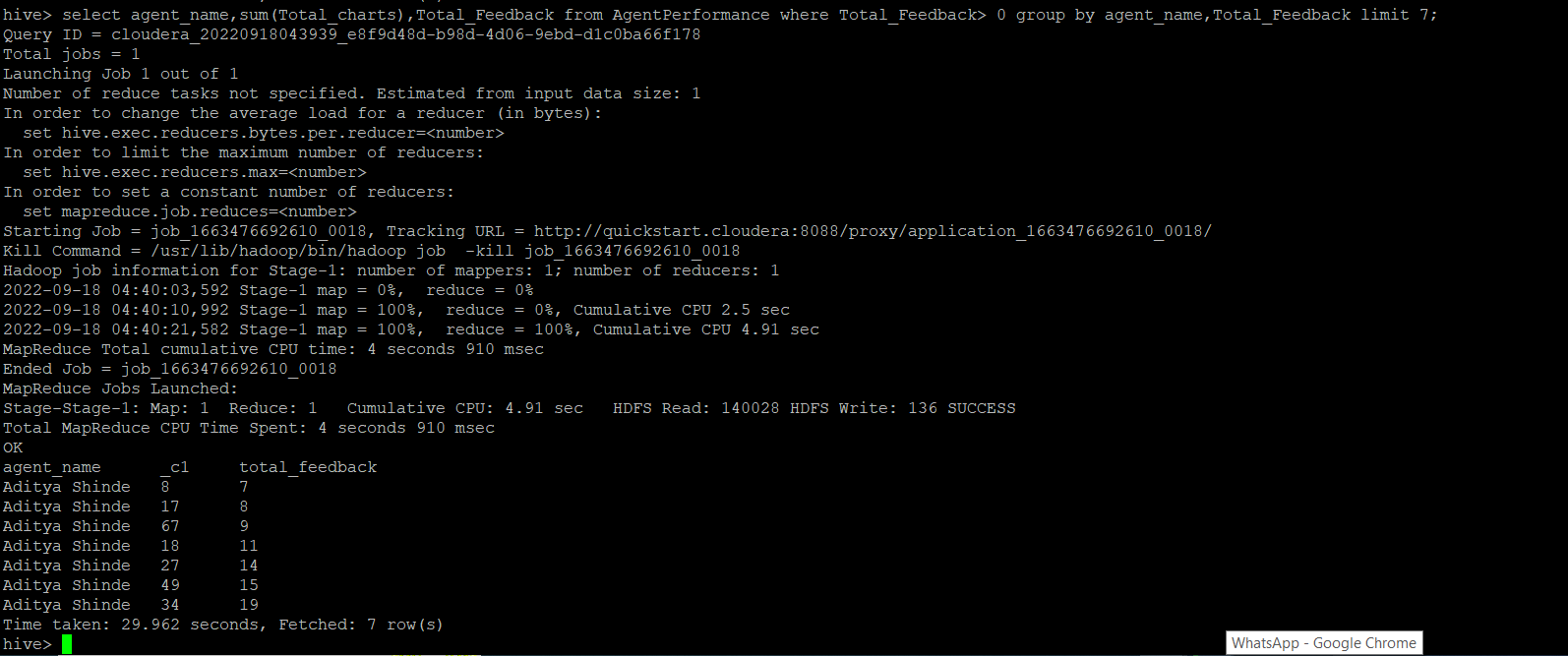
**Hive> select s.agent\_name,avg(col1[0]\*3600+col1[1]\*60+substr(col1[2],1,2))/3600 from(**

**select agent\_name,split(Avg\_Resolution\_Time,':') as col1 from AgentPerformance )s group by s.agent\_name;**



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

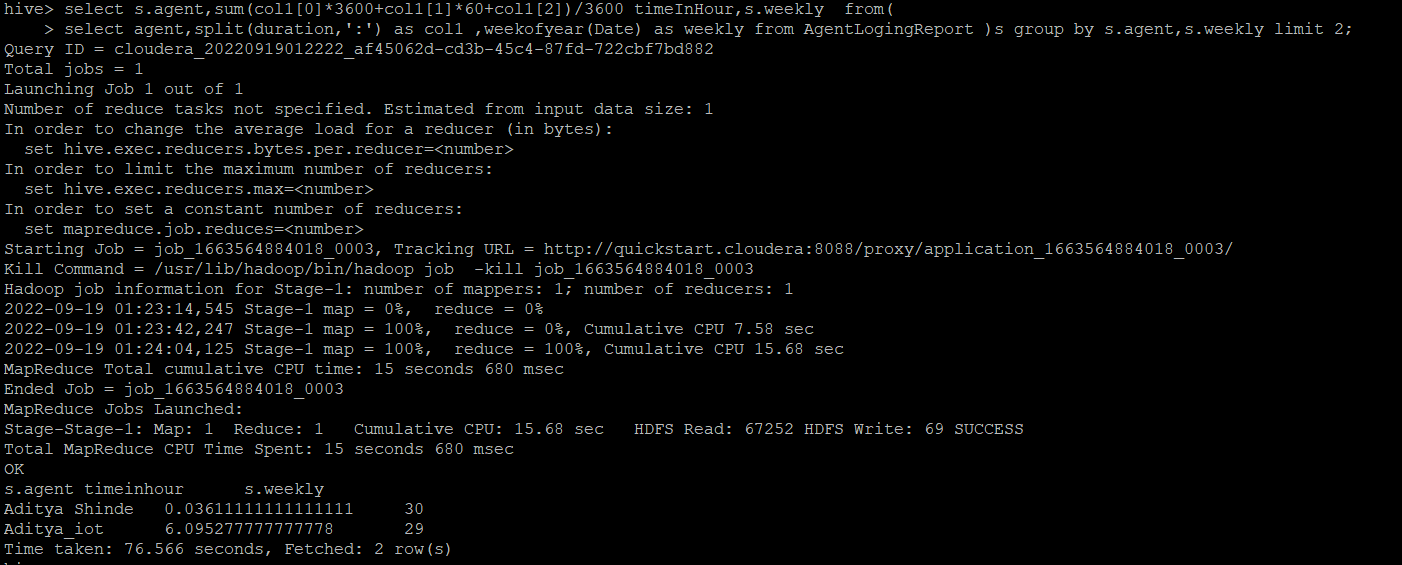
**Hive> select agent\_name,sum(Total\_chats),Total\_Feedback from AgentPerformance where Total\_Feedback> 0 group by agent\_name,Total\_Feedback;**



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(Date) as weekly from AgentLogingReport )s group by s.agent,s.weekly limit 2;**



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:**

**hive -e 'select a.agent,a.date,a.Duration,b.Total\_chats,b.Total\_Feedback from challenge.AgentLogingReport a join challenge.AgentPerformance b on a.agent = b.agent\_name' > /home/cloudera/sidd/Challenge/mini\_project\_1/inner\_join.csv;**

**left join:**

**hive -e 'select a.agent,a.date,a.Duration,b.Total\_chats,b.Total\_Feedback from challenge.AgentLogingReport a left join challenge.AgentPerformance b on a.agent = b.agent\_name' > /home/cloudera/sidd/Challenge/mini\_project\_1/left\_join.csv;**

**left join with performance improved due to /\*+ streamtable(a) \*/ hint:**

**hive -e 'select /\*+ streamtable(a) \*/a.agent,a.date,a.Duration,b.Total\_charts,b.Total\_Feedback from challenge.AgentLogingReport a left join challenge.AgentPerformance b on a.agent = b.agent\_name' > /home/cloudera/sidd/Challenge/mini\_project\_1/left\_join.csv;**

**Right join:**

**hive -e 'select 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/right\_join.csv;**

**Right join with performance improved due to /\*+ streamtable(a) \*/ hint:**

**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,**

**Date date,**

**Login string,**

**Logout string,**

**Duration string**

**)partitioned by (Agent string)**

**CLUSTERED BY (Date) sorted by (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,Date,Login,Logout,Duration,Agent from AgentLogingReport;**

**Hive> Create table AgentPerformance\_partitioned**

**(**

**sr\_no int,**

**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 (Date) sorted by (Date) INTO 8 BUCKETS**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ',';**

**Hive> insert into table AgentPerformance\_partitioned partition(Agent\_name) select sr\_no,Date,Total\_charts,Avg\_Response\_Time,Avg\_Resolution\_Time,Avg\_Rating,Total\_Feedback,Agent\_name from AgentPerformance;**