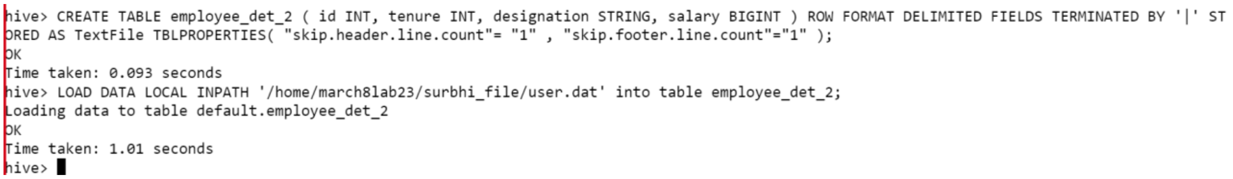
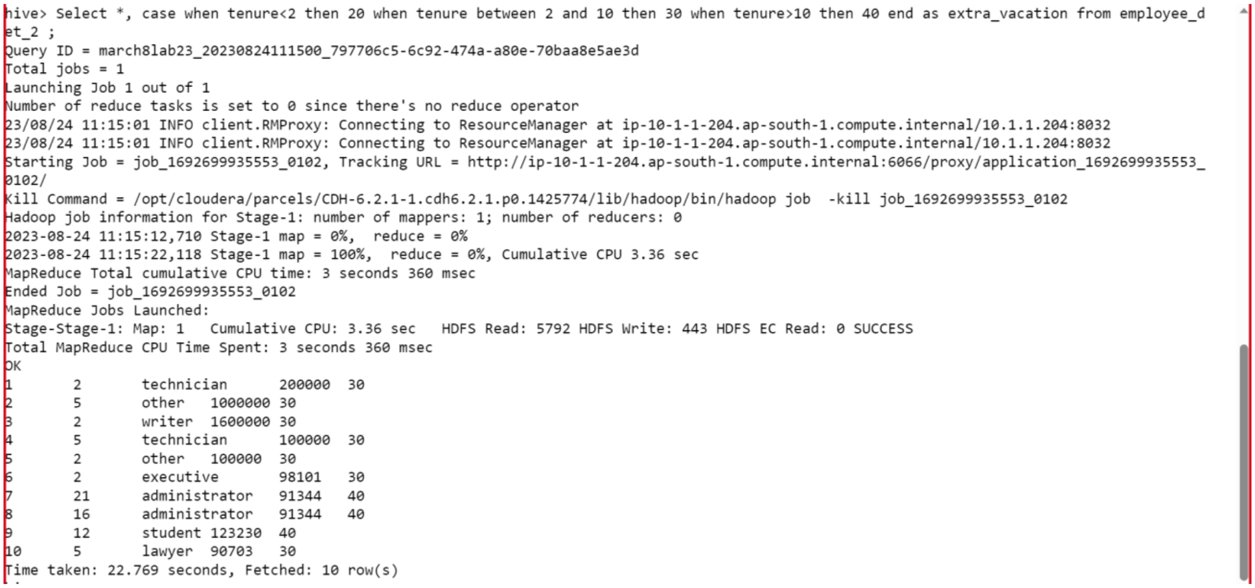
1. Create a table with the schema as specified below and load the data.
Write a query to derive a new column extra_vacation based on the tenure served, the logic is as given below.
1. If tenure < 2, Then 20
2. If tenure is 2-10 then 30 days
3. If tenure > 10 then 40 days

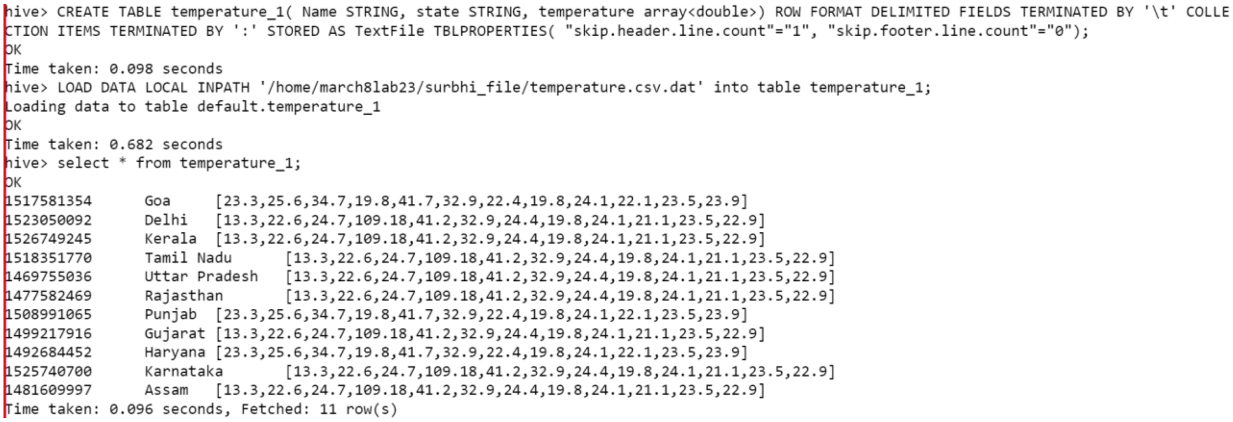


```
hive> CREATE TABLE employee_det_2 ( id INT, tenure INT, designation STRING, salary BIGINT ) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' ST
ORED AS TextFile TBLPROPERTIES( "skip.header.line.count"= "1" , "skip.footer.line.count"="1" );
OK
Time taken: 0.093 seconds
hive> LOAD DATA LOCAL INPATH '/home/march8lab23/surbhi_file/user.dat' into table employee_det_2;
Loading data to table default.employee_det_2
OK
Time taken: 1.01 seconds
hive> █
```

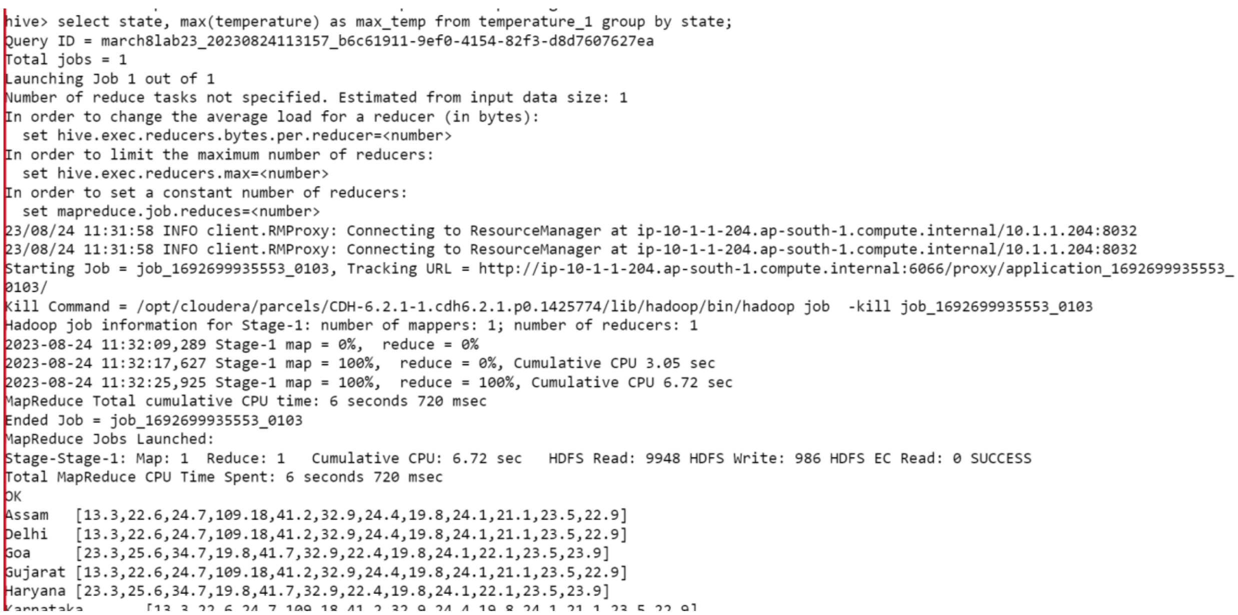


```
hive> Select *, case when tenure<2 then 20 when tenure between 2 and 10 then 30 when tenure>10 then 40 end as extra_vacation from employee_d
et_2 ;
Query ID = march8lab23_20230824111500_797706c5-6c92-474a-a80e-70baa8e5ae3d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/08/24 11:15:01 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 11:15:01 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0102, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0102/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0102
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-24 11:15:12,710 Stage-1 map = 0%,  reduce = 0%
2023-08-24 11:15:22,118 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.36 sec
MapReduce Total cumulative CPU time: 3 seconds 360 msec
Ended Job = job_1692699935553_0102
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.36 sec   HDFS Read: 5792 HDFS Write: 443 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 360 msec
OK
1       2       technician      200000  30
2       5       other   1000000 30
3       2       writer  1600000 30
4       5       technician      100000  30
5       2       other   100000  30
6       2       executive       98101   30
7       21      administrator   91344   40
8       16      administrator   91344   40
9       12      student 123230  40
10      5       lawyer  90703   30
Time taken: 22.769 seconds, Fetched: 10 row(s)
```

2. Create a table “temperature” to store the dataset as mentioned in the schema and load the data
Write a query to calculate the maximum temperature of each state.

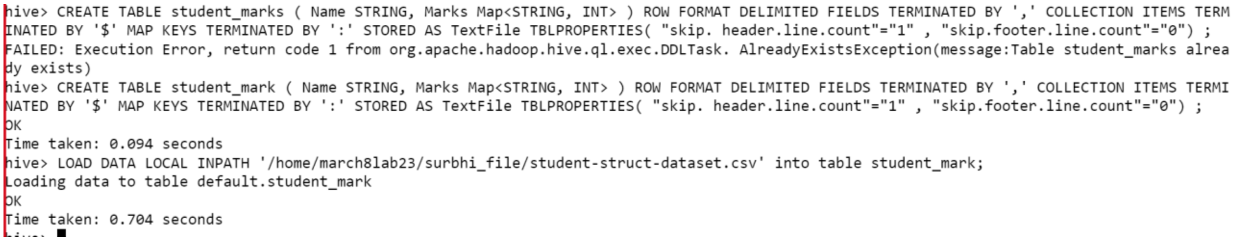


```
hive> CREATE TABLE temperature_1( Name STRING, state STRING, temperature array<double>) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' COLLE
CTION ITEMS TERMINATED BY ':' STORED AS TextFile TBLPROPERTIES( "skip.header.line.count"="1", "skip.footer.line.count"="0");
OK
Time taken: 0.098 seconds
hive> LOAD DATA LOCAL INPATH '/home/march8lab23/surbhi_file/temperature.csv.dat' into table temperature_1;
Loading data to table default.temperature_1
OK
Time taken: 0.682 seconds
hive> select * from temperature_1;
OK
1517581354      Goa     [23.3,25.6,34.7,19.8,41.7,32.9,22.4,19.8,24.1,22.1,23.5,23.9]
1523050092      Delhi   [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1526749245      Kerala  [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1518351770      Tamil Nadu      [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1469755036      Uttar Pradesh   [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1477582469      Rajasthan       [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1508991065      Punjab  [23.3,25.6,34.7,19.8,41.7,32.9,22.4,19.8,24.1,22.1,23.5,23.9]
1499217916      Gujarat [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1492684452      Haryana [23.3,25.6,34.7,19.8,41.7,32.9,22.4,19.8,24.1,22.1,23.5,23.9]
1525740700      Karnataka       [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
1481609997      Assam   [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
Time taken: 0.096 seconds, Fetched: 11 row(s)
```



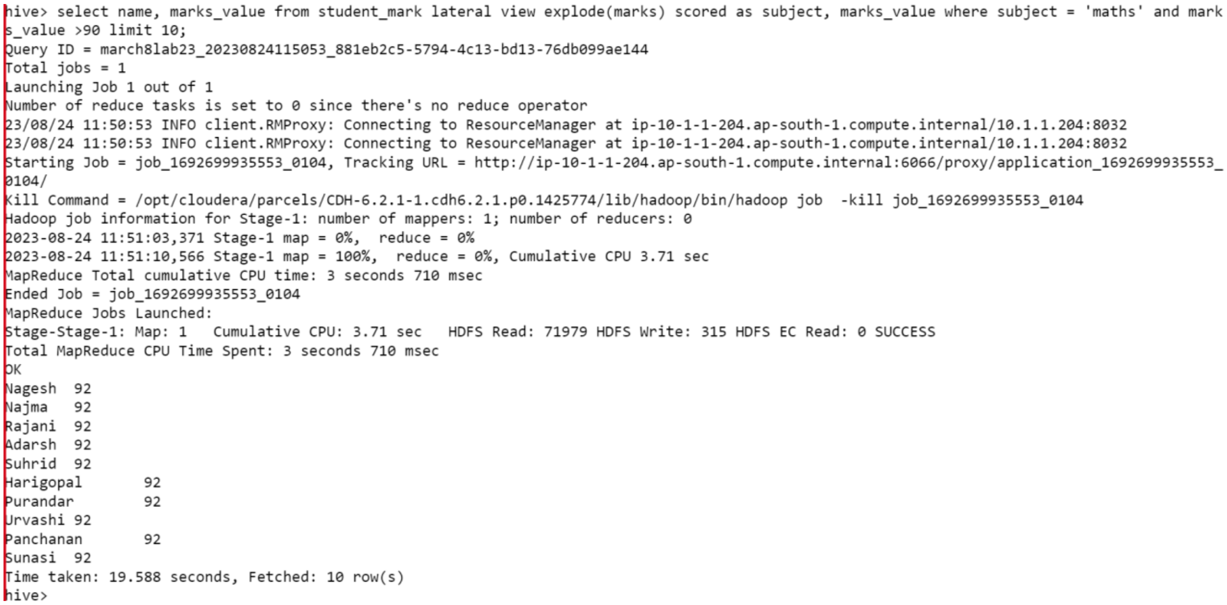
```
hive> select state, max(temperature) as max_temp from temperature_1 group by state;
Query ID = march8lab23_20230824113157_b6c61911-9ef0-4154-82f3-d8d7607627ea
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average 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>
23/08/24 11:31:58 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 11:31:58 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0103, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0103/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0103
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-08-24 11:32:09,289 Stage-1 map = 0%,  reduce = 0%
2023-08-24 11:32:17,627 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2023-08-24 11:32:25,925 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.72 sec
MapReduce Total cumulative CPU time: 6 seconds 720 msec
Ended Job = job_1692699935553_0103
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.72 sec   HDFS Read: 9948 HDFS Write: 986 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 720 msec
OK
Assam   [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
Delhi   [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
Goa     [23.3,25.6,34.7,19.8,41.7,32.9,22.4,19.8,24.1,22.1,23.5,23.9]
Gujarat [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
Haryana [23.3,25.6,34.7,19.8,41.7,32.9,22.4,19.8,24.1,22.1,23.5,23.9]
Karnataka       [13.3,22.6,24.7,109.18,41.2,32.9,24.4,19.8,24.1,21.1,23.5,22.9]
```

3) Create a table 'student_marks' with schema as shown above and load the data into the 'student_marks' table.



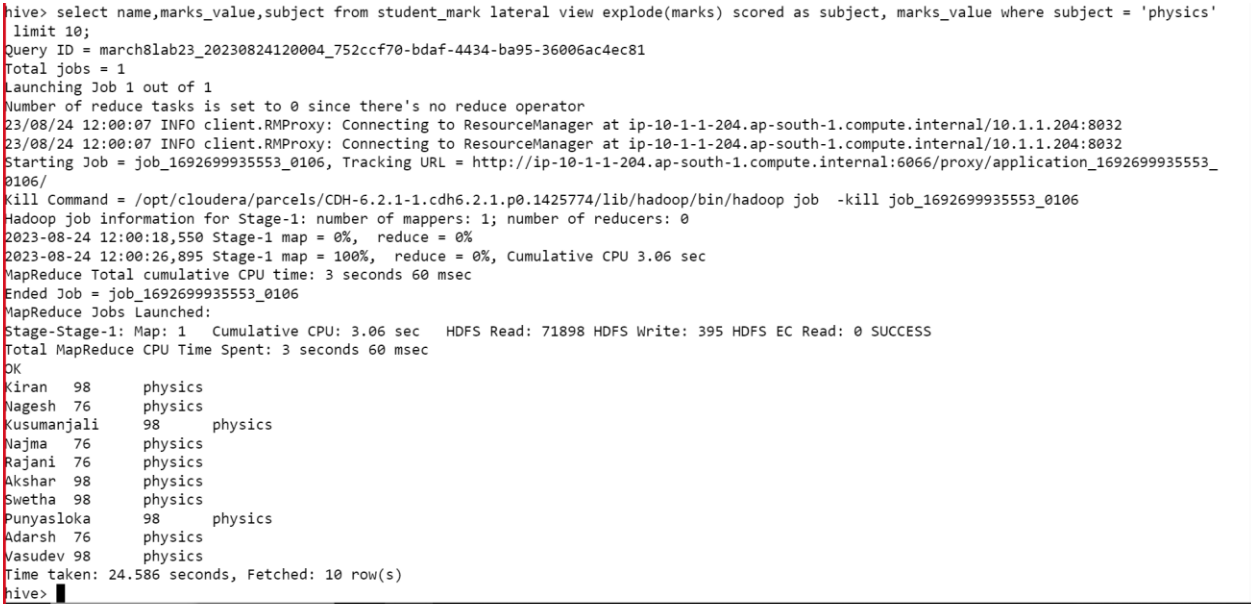
```
hive> CREATE TABLE student_marks ( Name STRING, Marks Map<STRING, INT> ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' COLLECTION ITEMS TERM
INATED BY '$' MAP KEYS TERMINATED BY ':' STORED AS TextFile TBLPROPERTIES( "skip. header.line.count"="1" , "skip.footer.line.count"="0") ;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table student_marks alrea
dy exists)
hive> CREATE TABLE student_mark ( Name STRING, Marks Map<STRING, INT> ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' COLLECTION ITEMS TERMI
NATED BY '$' MAP KEYS TERMINATED BY ':' STORED AS TextFile TBLPROPERTIES( "skip. header.line.count"="1" , "skip.footer.line.count"="0") ;
OK
Time taken: 0.094 seconds
hive> LOAD DATA LOCAL INPATH '/home/march8lab23/surbhi_file/student-struct-dataset.csv' into table student_mark;
Loading data to table default.student_mark
OK
Time taken: 0.704 seconds
hive> ■
```

a)Write a query to perform below mentioned tasks: 1. Display NAME who have scored more than 90 in subject Maths subject



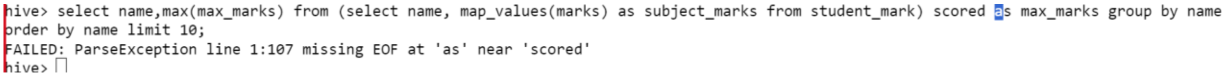
```
hive> select name, marks_value from student_mark lateral view explode(marks) scored as subject, marks_value where subject = 'maths' and mark
s_value >90 limit 10;
Query ID = march8lab23_20230824115053_881eb2c5-5794-4c13-bd13-76db099ae144
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/08/24 11:50:53 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 11:50:53 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0104, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0104/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0104
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-24 11:51:03,371 Stage-1 map = 0%,  reduce = 0%
2023-08-24 11:51:10,566 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.71 sec
MapReduce Total cumulative CPU time: 3 seconds 710 msec
Ended Job = job_1692699935553_0104
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.71 sec   HDFS Read: 71979 HDFS Write: 315 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 710 msec
OK
Nagesh  92
Najma   92
Rajani  92
Adarsh  92
Suhrid  92
Harigopal       92
Purandar        92
Urvashi 92
Panchanan       92
Sunasi  92
Time taken: 19.588 seconds, Fetched: 10 row(s)
hive>
```

b) Display NAME and marks scored in physics subject.



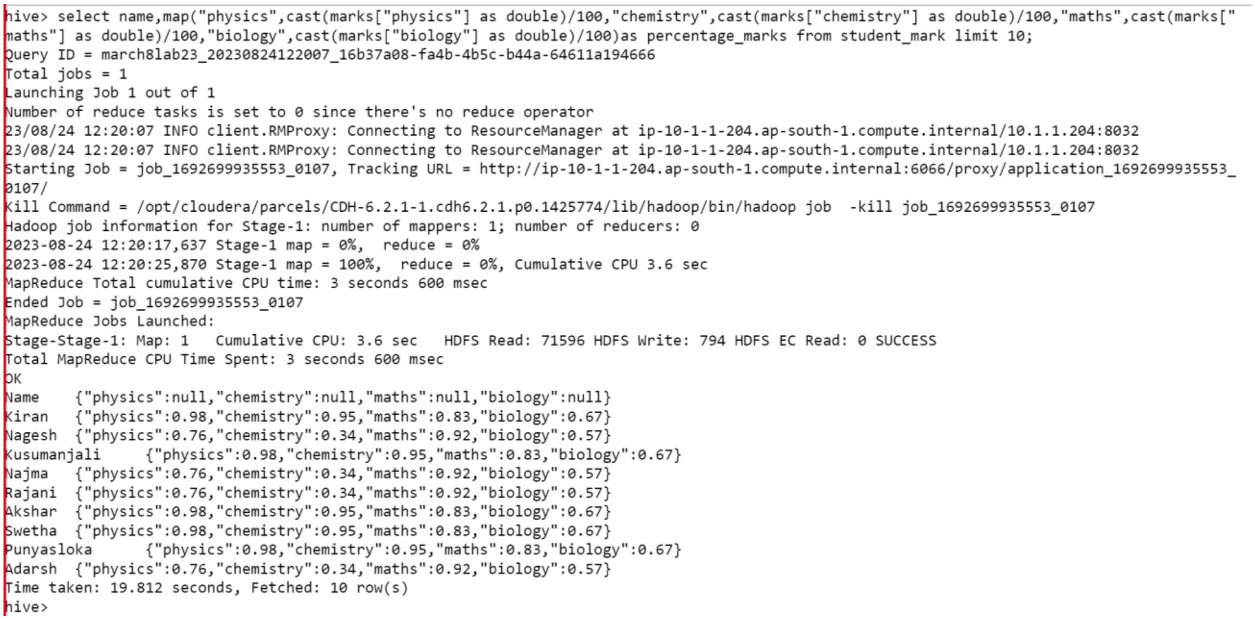
```
hive> select name,marks_value,subject from student_mark lateral view explode(marks) scored as subject, marks_value where subject = 'physics'
 limit 10;
Query ID = march8lab23_20230824120004_752ccf70-bdaf-4434-ba95-36006ac4ec81
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/08/24 12:00:07 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 12:00:07 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0106, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0106/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0106
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-24 12:00:18,550 Stage-1 map = 0%,  reduce = 0%
2023-08-24 12:00:26,895 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.06 sec
MapReduce Total cumulative CPU time: 3 seconds 60 msec
Ended Job = job_1692699935553_0106
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.06 sec   HDFS Read: 71898 HDFS Write: 395 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 60 msec
OK
Kiran   98      physics
Nagesh  76      physics
Kusumanjali     98      physics
Najma   76      physics
Rajani  76      physics
Akshar  98      physics
Swetha  98      physics
Punyasloka      98      physics
Adarsh  76      physics
Vasudev 98      physics
Time taken: 24.586 seconds, Fetched: 10 row(s)
hive> █
```

C) Display NAME, and <maximum-subject-marks>



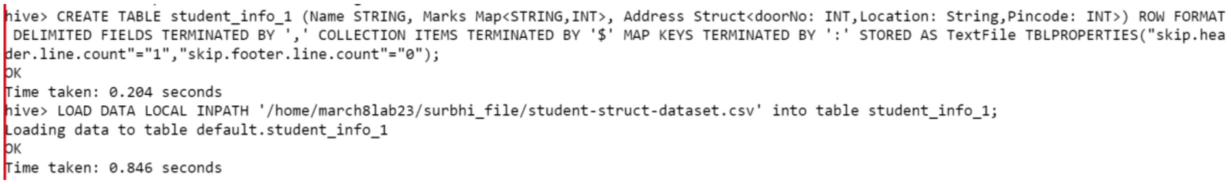
```
hive> select name,max(max_marks) from (select name, map_values(marks) as subject_marks from student_mark) scored as max_marks group by name
order by name limit 10;
FAILED: ParseException line 1:107 missing EOF at 'as' near 'scored'
hive> □
```

d) Display NAME and <percentage of marks>



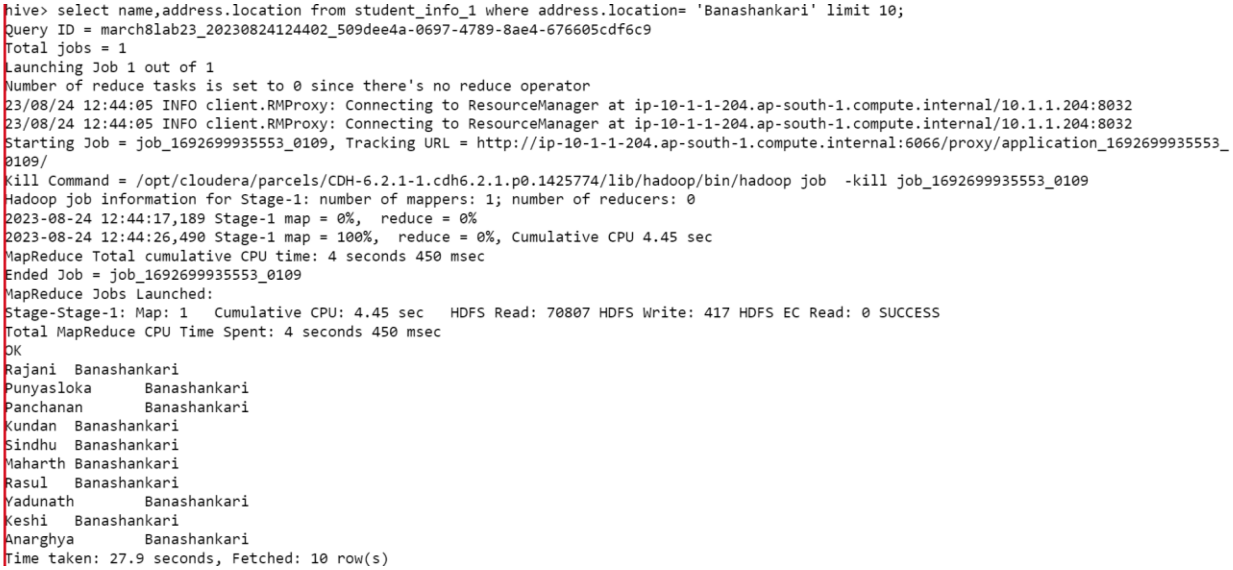
```
hive> select name,map("physics",cast(marks["physics"] as double)/100,"chemistry",cast(marks["chemistry"] as double)/100,"maths",cast(marks["
maths"] as double)/100,"biology",cast(marks["biology"] as double)/100)as percentage_marks from student_mark limit 10;
Query ID = march8lab23_20230824122007_16b37a08-fa4b-4b5c-b44a-64611a194666
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/08/24 12:20:07 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 12:20:07 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0107, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0107/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0107
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-24 12:20:17,637 Stage-1 map = 0%,  reduce = 0%
2023-08-24 12:20:25,870 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.6 sec
MapReduce Total cumulative CPU time: 3 seconds 600 msec
Ended Job = job_1692699935553_0107
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.6 sec   HDFS Read: 71596 HDFS Write: 794 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 600 msec
OK
Name    {"physics":null,"chemistry":null,"maths":null,"biology":null}
Kiran   {"physics":0.98,"chemistry":0.95,"maths":0.83,"biology":0.67}
Nagesh  {"physics":0.76,"chemistry":0.34,"maths":0.92,"biology":0.57}
Kusumanjali     {"physics":0.98,"chemistry":0.95,"maths":0.83,"biology":0.67}
Najma   {"physics":0.76,"chemistry":0.34,"maths":0.92,"biology":0.57}
Rajani  {"physics":0.76,"chemistry":0.34,"maths":0.92,"biology":0.57}
Akshar  {"physics":0.98,"chemistry":0.95,"maths":0.83,"biology":0.67}
Swetha  {"physics":0.98,"chemistry":0.95,"maths":0.83,"biology":0.67}
Punyasloka      {"physics":0.98,"chemistry":0.95,"maths":0.83,"biology":0.67}
Adarsh  {"physics":0.76,"chemistry":0.34,"maths":0.92,"biology":0.57}
Time taken: 19.812 seconds, Fetched: 10 row(s)
hive>
```

4)Create a table "student_info" with schema as show below and load the data



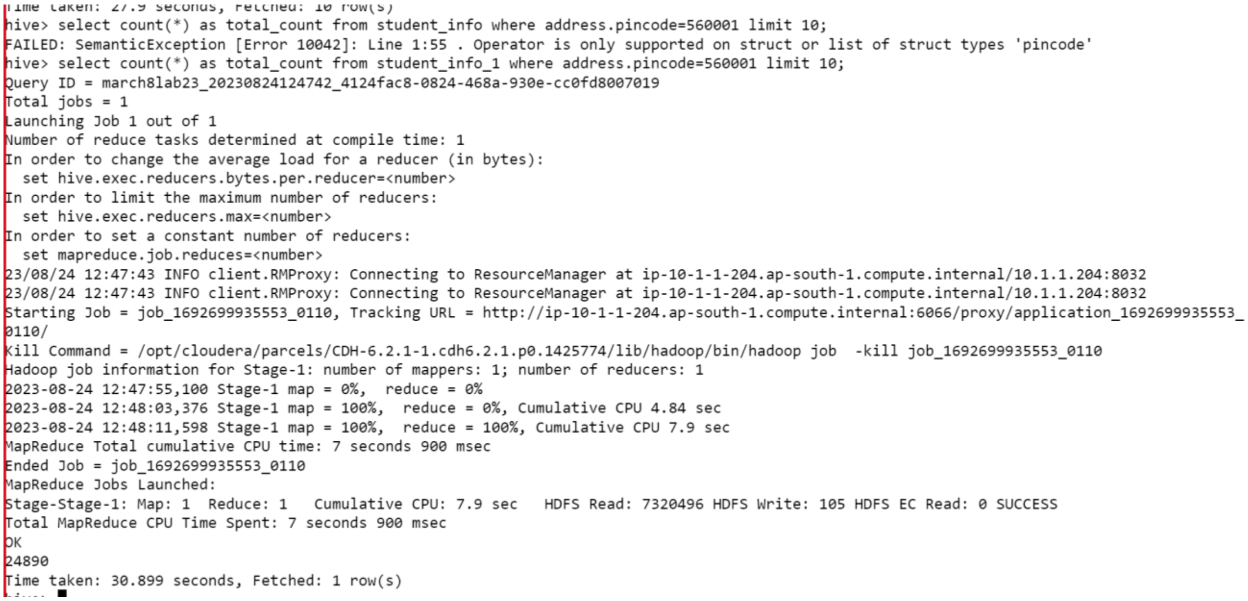
```
hive> CREATE TABLE student_info_1 (Name STRING, Marks Map<STRING,INT>, Address Struct<doorNo: INT,Location: String,Pincode: INT>) ROW FORMAT
 DELIMITED FIELDS TERMINATED BY ',' COLLECTION ITEMS TERMINATED BY '$' MAP KEYS TERMINATED BY ':' STORED AS TextFile TBLPROPERTIES("skip.hea
der.line.count"="1","skip.footer.line.count"="0");
OK
Time taken: 0.204 seconds
hive> LOAD DATA LOCAL INPATH '/home/march8lab23/surbhi_file/student-struct-dataset.csv' into table student_info_1;
Loading data to table default.student_info_1
OK
Time taken: 0.846 seconds
```

a) Display all "NAME" who is located in Banashankari



```
hive> select name,address.location from student_info_1 where address.location= 'Banashankari' limit 10;
Query ID = march8lab23_20230824124402_509dee4a-0697-4789-8ae4-676605cdf6c9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/08/24 12:44:05 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 12:44:05 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0109, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0109/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0109
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-08-24 12:44:17,189 Stage-1 map = 0%,  reduce = 0%
2023-08-24 12:44:26,490 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.45 sec
MapReduce Total cumulative CPU time: 4 seconds 450 msec
Ended Job = job_1692699935553_0109
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.45 sec   HDFS Read: 70807 HDFS Write: 417 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 450 msec
OK
Rajani  Banashankari
Punyasloka      Banashankari
Panchanan       Banashankari
Kundan  Banashankari
Sindhu  Banashankari
Maharth Banashankari
Rasul   Banashankari
Yadunath        Banashankari
Keshi   Banashankari
Anarghya        Banashankari
Time taken: 27.9 seconds, Fetched: 10 row(s)
```

b) Calculate the total count who is staying in pin code 560001



```
Time taken: 27.9 seconds, Fetched: 10 row(s)
hive> select count(*) as total_count from student_info where address.pincode=560001 limit 10;
FAILED: SemanticException [Error 10042]: Line 1:55 . Operator is only supported on struct or list of struct types 'pincode'
hive> select count(*) as total_count from student_info_1 where address.pincode=560001 limit 10;
Query ID = march8lab23_20230824124742_4124fac8-0824-468a-930e-cc0fd8007019
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>
23/08/24 12:47:43 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/08/24 12:47:43 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1692699935553_0110, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1692699935553_
0110/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1692699935553_0110
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-08-24 12:47:55,100 Stage-1 map = 0%,  reduce = 0%
2023-08-24 12:48:03,376 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.84 sec
2023-08-24 12:48:11,598 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.9 sec
MapReduce Total cumulative CPU time: 7 seconds 900 msec
Ended Job = job_1692699935553_0110
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.9 sec   HDFS Read: 7320496 HDFS Write: 105 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 900 msec
OK
24890
Time taken: 30.899 seconds, Fetched: 1 row(s)
hive> 
```