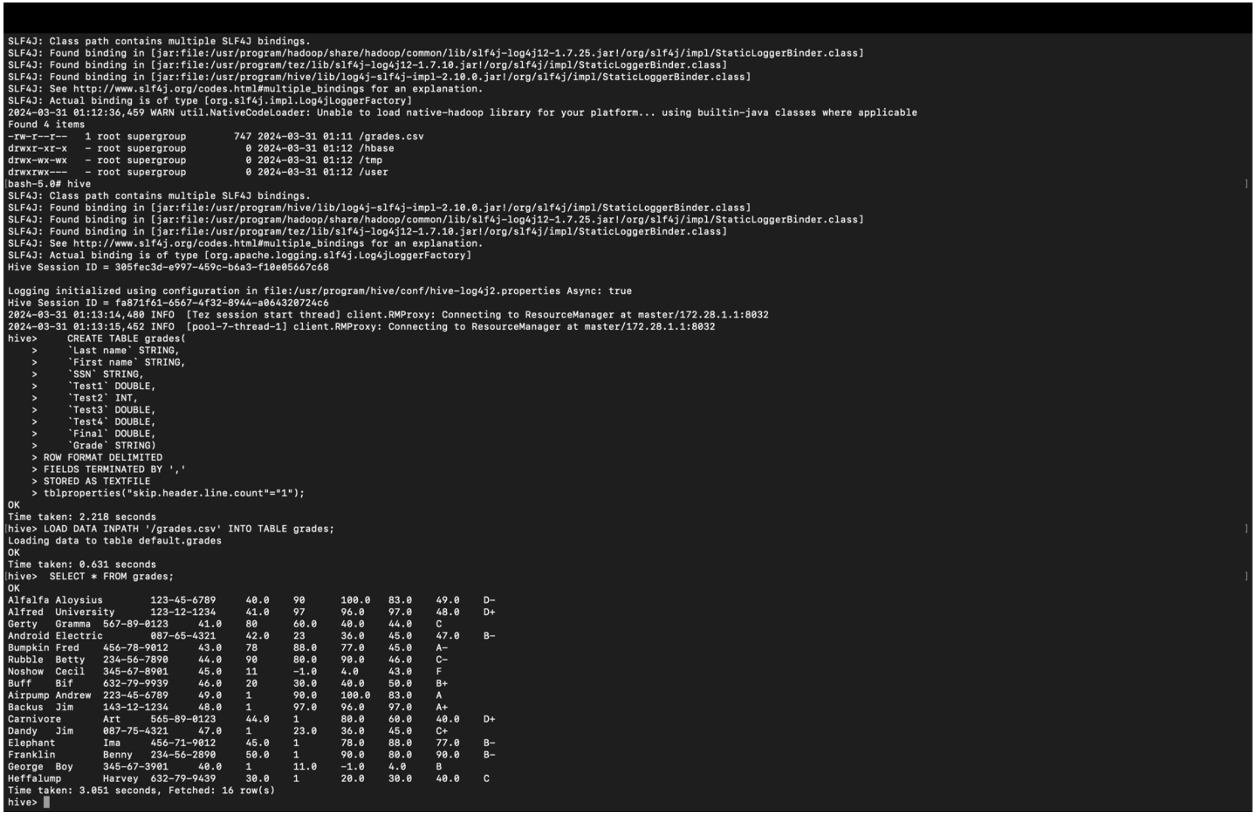
DSC 650
Inman, Gracie
Week 3
03/31/2024

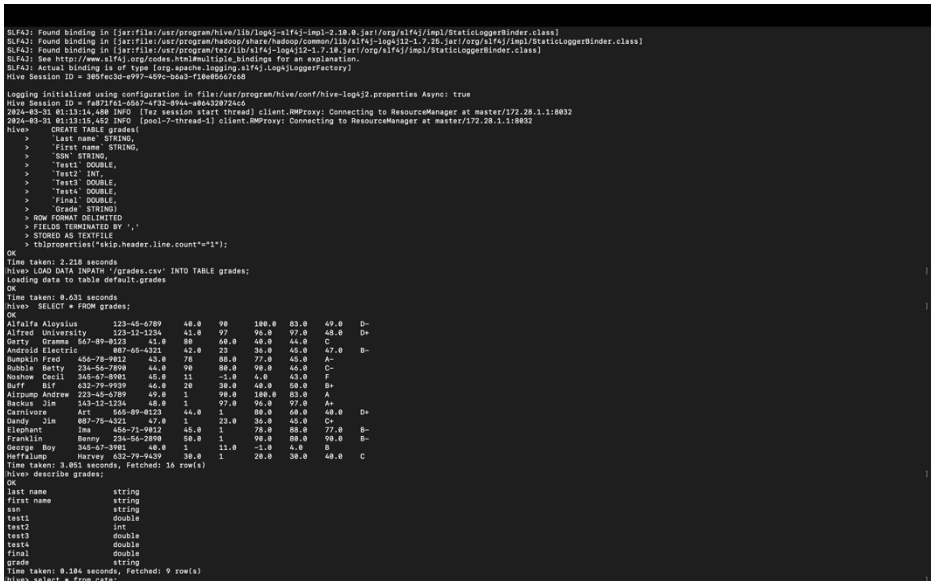
Query Results:



```
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-31 01:12:36,459 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 root supergroup        747 2024-03-31 01:11 /grades.csv
drwxr-xr-x   - root supergroup          0 2024-03-31 01:12 /hbase
drwx-wx-wx   - root supergroup          0 2024-03-31 01:12 /tmp
drwxrwx---   - root supergroup          0 2024-03-31 01:12 /user
bash-5.0# hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 305fec3d-e997-459c-b6a3-f10e05667c68

Logging initialized using configuration in file:/usr/program/hive/conf/hive-log4j2.properties Async: true
Hive Session ID = fa871f61-6567-4f32-8944-a064320724c6
2024-03-31 01:13:14,480 INFO  [Tez session start thread] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
2024-03-31 01:13:15,452 INFO  [pool-7-thread-1] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
hive>     CREATE TABLE grades(
    >     `Last name` STRING,
    >     `First name` STRING,
    >     `SSN` STRING,
    >     `Test1` DOUBLE,
    >     `Test2` INT,
    >     `Test3` DOUBLE,
    >     `Test4` DOUBLE,
    >     `Final` DOUBLE,
    >     `Grade` STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.218 seconds
hive> LOAD DATA INPATH '/grades.csv' INTO TABLE grades;
Loading data to table default.grades
OK
Time taken: 0.631 seconds
hive>  SELECT * FROM grades;
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      123-12-1234     41.0    97      96.0    97.0    48.0    D+
Gerty   Gramma  567-89-0123     41.0    80      60.0    40.0    44.0    C
Android Electric        087-65-4321     42.0    23      36.0    45.0    47.0    B-
Bumpkin Fred    456-78-9012     43.0    78      88.0    77.0    45.0    A-
Rubble  Betty   234-56-7890     44.0    90      80.0    90.0    46.0    C-
Noshow  Cecil   345-67-8901     45.0    11      -1.0    4.0     43.0    F
Buff    Bif     632-79-9939     46.0    20      30.0    40.0    50.0    B+
Airpump Andrew  223-45-6789     49.0    1       90.0    100.0   83.0    A
Backus  Jim     143-12-1234     48.0    1       97.0    96.0    97.0    A+
Carnivore       Art     565-89-0123     44.0    1       80.0    60.0    40.0    D+
Dandy   Jim     087-75-4321     47.0    1       23.0    36.0    45.0    C+
Elephant        Ima     456-71-9012     45.0    1       78.0    88.0    77.0    B-
Franklin        Benny   234-56-2890     50.0    1       90.0    80.0    90.0    B-
George  Boy     345-67-3901     40.0    1       11.0    -1.0    4.0     B
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Time taken: 3.051 seconds, Fetched: 16 row(s)
hive> 
```

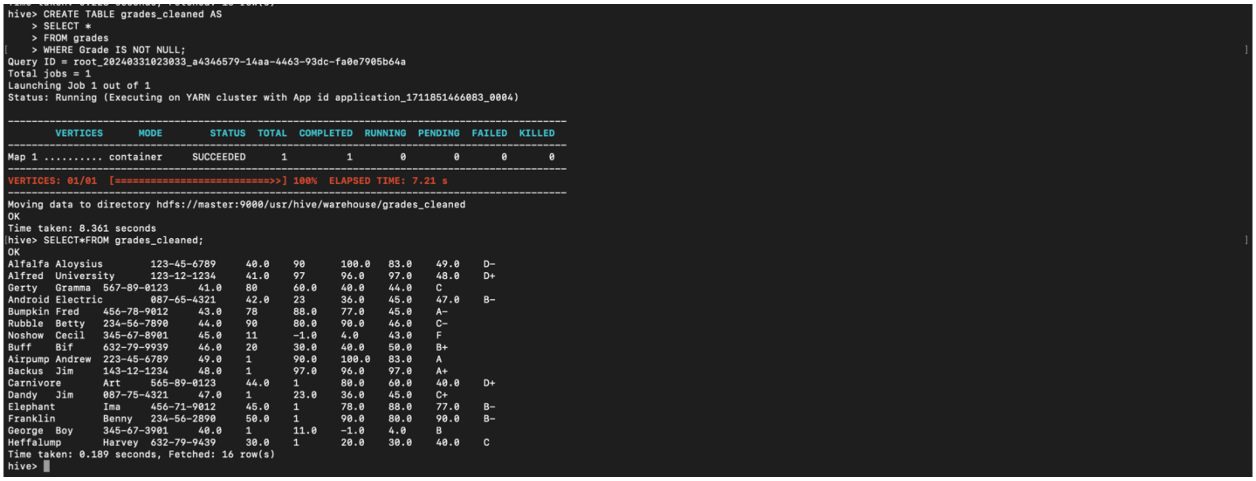
Three SQL Commands:

- Describe table



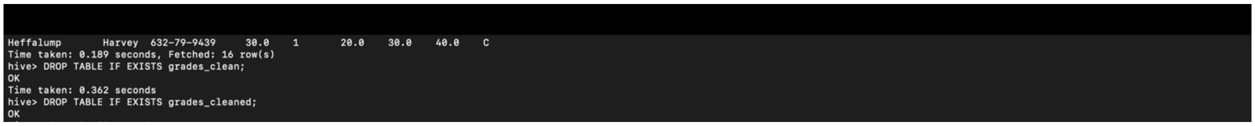
```
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 305fec3d-e997-459c-b6a3-f10e05667c68

Logging initialized using configuration in file:/usr/program/hive/conf/hive-log4j2.properties Async: true
Hive Session ID = fa871f61-6567-4f32-8944-a064320724c6
2024-03-31 01:13:14,480 INFO  [Tez session start thread] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
2024-03-31 01:13:15,452 INFO  [pool-7-thread-1] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
hive>     CREATE TABLE grades(
    >     `Last name` STRING,
    >     `First name` STRING,
    >     `SSN` STRING,
    >     `Test1` DOUBLE,
    >     `Test2` INT,
    >     `Test3` DOUBLE,
    >     `Test4` DOUBLE,
    >     `Final` DOUBLE,
    >     `Grade` STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.218 seconds
hive> LOAD DATA INPATH '/grades.csv' INTO TABLE grades;
Loading data to table default.grades
OK
Time taken: 0.631 seconds
hive>  SELECT * FROM grades;
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      123-12-1234     41.0    97      96.0    97.0    48.0    D+
Gerty   Gramma  567-89-0123     41.0    80      60.0    40.0    44.0    C
Android Electric        087-65-4321     42.0    23      36.0    45.0    47.0    B-
Bumpkin Fred    456-78-9012     43.0    78      88.0    77.0    45.0    A-
Rubble  Betty   234-56-7890     44.0    90      80.0    90.0    46.0    C-
Noshow  Cecil   345-67-8901     45.0    11      -1.0    4.0     43.0    F
Buff    Bif     632-79-9939     46.0    20      30.0    40.0    50.0    B+
Airpump Andrew  223-45-6789     49.0    1       90.0    100.0   83.0    A
Backus  Jim     143-12-1234     48.0    1       97.0    96.0    97.0    A+
Carnivore       Art     565-89-0123     44.0    1       80.0    60.0    40.0    D+
Dandy   Jim     087-75-4321     47.0    1       23.0    36.0    45.0    C+
Elephant        Ima     456-71-9012     45.0    1       78.0    88.0    77.0    B-
Franklin        Benny   234-56-2890     50.0    1       90.0    80.0    90.0    B-
George  Boy     345-67-3901     40.0    1       11.0    -1.0    4.0     B
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Time taken: 3.051 seconds, Fetched: 16 row(s)
hive> describe grades;
OK
last name               string
first name              string
ssn                     string
test1                   double
test2                   int
test3                   double
test4                   double
final                   double
grade                   string
Time taken: 0.104 seconds, Fetched: 9 row(s)
hive> select * from cate;
```

- Filter out NA values from the Grade column



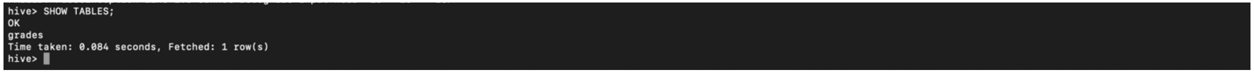
```
hive> CREATE TABLE grades_cleaned AS
    > SELECT *
    > FROM grades
    > WHERE Grade IS NOT NULL;
Query ID = root_20240331023033_a4346579-14aa-4463-93dc-fa0e7905b64a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711851466083_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.21 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://master:9000/usr/hive/warehouse/grades_cleaned
OK
Time taken: 8.361 seconds
hive> SELECT*FROM grades_cleaned;
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      123-12-1234     41.0    97      96.0    97.0    48.0    D+
Gerty   Gramma  567-89-0123     41.0    80      60.0    40.0    44.0    C
Android Electric        087-65-4321     42.0    23      36.0    45.0    47.0    B-
Bumpkin Fred    456-78-9012     43.0    78      88.0    77.0    45.0    A-
Rubble  Betty   234-56-7890     44.0    90      80.0    90.0    46.0    C-
Noshow  Cecil   345-67-8901     45.0    11      -1.0    4.0     43.0    F
Buff    Bif     632-79-9939     46.0    20      30.0    40.0    50.0    B+
Airpump Andrew  223-45-6789     49.0    1       90.0    100.0   83.0    A
Backus  Jim     143-12-1234     48.0    1       97.0    96.0    97.0    A+
Carnivore       Art     565-89-0123     44.0    1       80.0    60.0    40.0    D+
Dandy   Jim     087-75-4321     47.0    1       23.0    36.0    45.0    C+
Elephant        Ima     456-71-9012     45.0    1       78.0    88.0    77.0    B-
Franklin        Benny   234-56-2890     50.0    1       90.0    80.0    90.0    B-
George  Boy     345-67-3901     40.0    1       11.0    -1.0    4.0     B
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Time taken: 0.189 seconds, Fetched: 16 row(s)
hive> 
```

- Delete Tables



```
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Time taken: 0.189 seconds, Fetched: 16 row(s)
hive> DROP TABLE IF EXISTS grades_clean;
OK
Time taken: 0.362 seconds
hive> DROP TABLE IF EXISTS grades_cleaned;
OK
```

  - Show a list of Tables to show dropped tables were dropped.

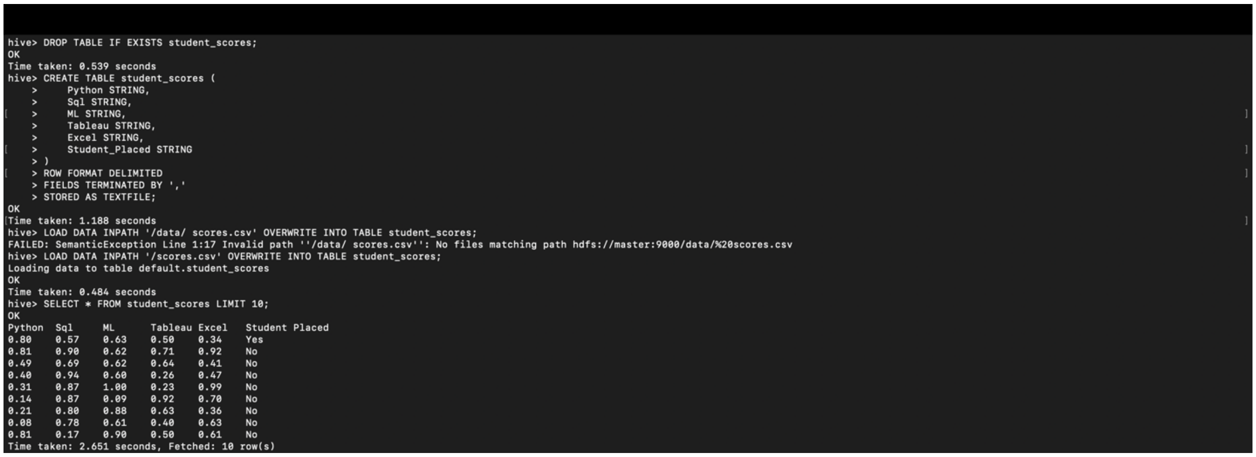


```
hive> SHOW TABLES;
OK
grades
Time taken: 0.084 seconds, Fetched: 1 row(s)
hive> 
```

Three SQL Commands on New Dataset:

I chose the student scores dataset because I wanted to practice doing some of the commands I would usually do when looking at a dataset. I am hoping to gain insight into what scores result in a student being placed.

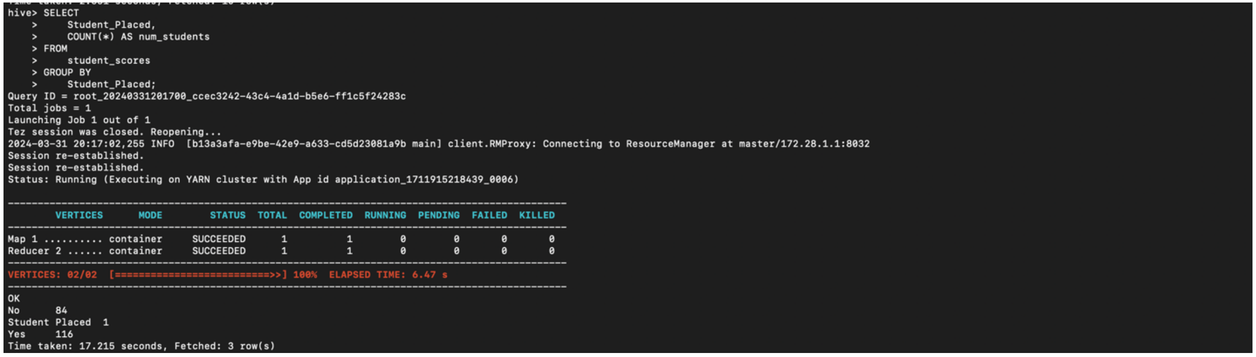
- Loaded Data +



[redacted]

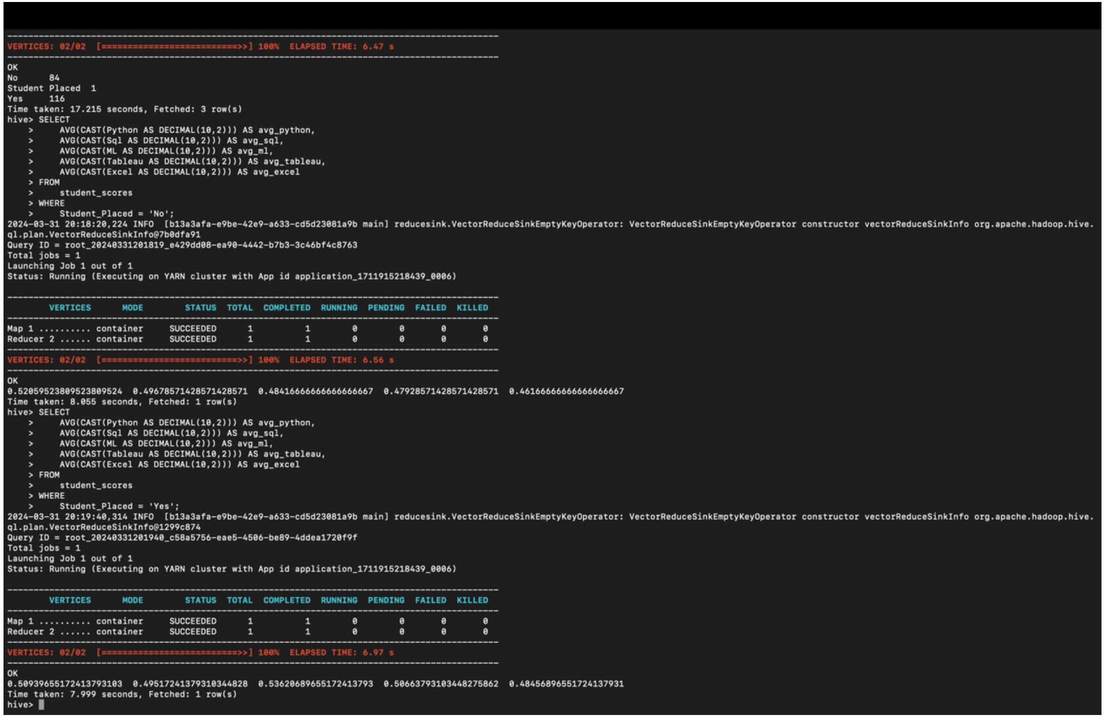
```
hive> DROP TABLE IF EXISTS student_scores;
OK
Time taken: 0.539 seconds
hive> CREATE TABLE student_scores (
    >     Python STRING,
    >     Sql STRING,
    >     ML STRING,
    >     Tableau STRING,
    >     Excel STRING,
    >     Student_Placed STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 1.188 seconds
hive> LOAD DATA INPATH '/data/ scores.csv' OVERWRITE INTO TABLE student_scores;
FAILED: SemanticException Line 1:17 Invalid path ''/data/ scores.csv'': No files matching path hdfs://master:9000/data/%20scores.csv
hive> LOAD DATA INPATH '/scores.csv' OVERWRITE INTO TABLE student_scores;
Loading data to table default.student_scores
OK
Time taken: 0.484 seconds
hive> SELECT * FROM student_scores LIMIT 10;
OK
Python  Sql     ML      Tableau Excel   Student Placed
0.80    0.57    0.63    0.50    0.34    Yes
0.81    0.90    0.62    0.71    0.92    No
0.49    0.69    0.62    0.64    0.41    No
0.40    0.94    0.60    0.26    0.47    No
0.31    0.87    1.00    0.23    0.99    No
0.14    0.87    0.09    0.92    0.70    No
0.21    0.80    0.88    0.63    0.36    No
0.08    0.78    0.61    0.40    0.63    No
0.81    0.17    0.90    0.50    0.61    No
Time taken: 2.651 seconds, Fetched: 10 row(s)
```

- Discover how many students are in each placement category.



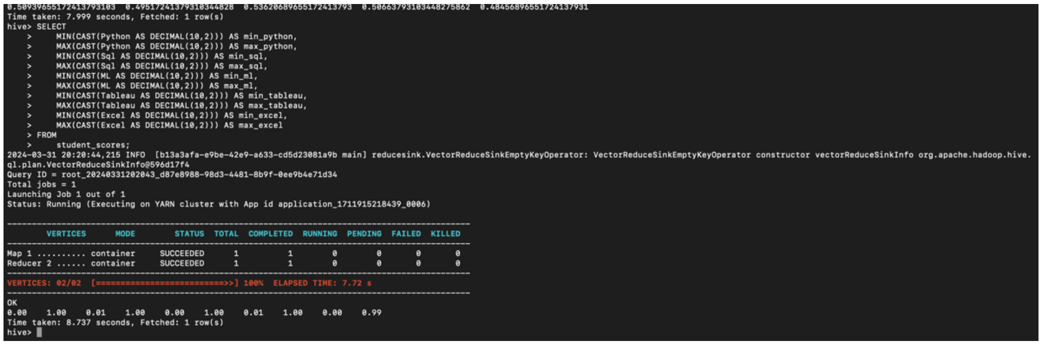
```
hive> SELECT
    >     Student_Placed,
    >     COUNT(*) AS num_students
    > FROM
    >     student_scores
    > GROUP BY
    >     Student_Placed;
Query ID = root_20240331201700_ccec3242-43c4-4a1d-b5e6-ff1c5f24283c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-03-31 20:17:02,255 INFO  [b13a3afa-e9be-42e9-a633-cd5d23081a9b main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1711915218439_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.47 s
----------------------------------------------------------------------------------------------
OK
No      84
Student Placed  1
Yes     116
Time taken: 17.215 seconds, Fetched: 3 row(s)
```

- Average scores for each column for both groups



```
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.47 s
----------------------------------------------------------------------------------------------
OK
No      84
Student Placed  1
Yes     116
Time taken: 17.215 seconds, Fetched: 3 row(s)
hive> SELECT
    >     AVG(CAST(Python AS DECIMAL(10,2))) AS avg_python,
    >     AVG(CAST(Sql AS DECIMAL(10,2))) AS avg_sql,
    >     AVG(CAST(ML AS DECIMAL(10,2))) AS avg_ml,
    >     AVG(CAST(Tableau AS DECIMAL(10,2))) AS avg_tableau,
    >     AVG(CAST(Excel AS DECIMAL(10,2))) AS avg_excel
    > FROM
    >     student_scores
    > WHERE
    >     Student_Placed = 'No';
2024-03-31 20:18:20,224 INFO  [b13a3afa-e9be-42e9-a633-cd5d23081a9b main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.
ql.plan.VectorReduceSinkInfo@7b0dfa91
Query ID = root_20240331201819_e429dd08-ea90-4442-b7b3-3c46bf4c8763
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711915218439_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.56 s
----------------------------------------------------------------------------------------------
OK
0.52059523809523809524  0.49678571428571428571  0.48416666666666666667  0.47928571428571428571  0.46166666666666666667
Time taken: 8.055 seconds, Fetched: 1 row(s)
hive> SELECT
    >     AVG(CAST(Python AS DECIMAL(10,2))) AS avg_python,
    >     AVG(CAST(Sql AS DECIMAL(10,2))) AS avg_sql,
    >     AVG(CAST(ML AS DECIMAL(10,2))) AS avg_ml,
    >     AVG(CAST(Tableau AS DECIMAL(10,2))) AS avg_tableau,
    >     AVG(CAST(Excel AS DECIMAL(10,2))) AS avg_excel
    > FROM
    >     student_scores
    > WHERE
    >     Student_Placed = 'Yes';
2024-03-31 20:19:40,314 INFO  [b13a3afa-e9be-42e9-a633-cd5d23081a9b main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.
ql.plan.VectorReduceSinkInfo@1299c874
Query ID = root_20240331201940_c58a5756-eae5-4506-be89-4ddea1720f9f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711915218439_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.97 s
----------------------------------------------------------------------------------------------
OK
0.50939655172413793103  0.49517241379310344828  0.53620689655172413793  0.50663793103448275862  0.48456896551724137931
Time taken: 7.999 seconds, Fetched: 1 row(s)
hive> 
```

- Minimum and Max for each column



```
0.50939655172413793103  0.49517241379310344828  0.53620689655172413793  0.50663793103448275862  0.48456896551724137931
Time taken: 7.999 seconds, Fetched: 1 row(s)
hive> SELECT
    >     MIN(CAST(Python AS DECIMAL(10,2))) AS min_python,
    >     MAX(CAST(Python AS DECIMAL(10,2))) AS max_python,
    >     MIN(CAST(Sql AS DECIMAL(10,2))) AS min_sql,
    >     MAX(CAST(Sql AS DECIMAL(10,2))) AS max_sql,
    >     MIN(CAST(ML AS DECIMAL(10,2))) AS min_ml,
    >     MAX(CAST(ML AS DECIMAL(10,2))) AS max_ml,
    >     MIN(CAST(Tableau AS DECIMAL(10,2))) AS min_tableau,
    >     MAX(CAST(Tableau AS DECIMAL(10,2))) AS max_tableau,
    >     MIN(CAST(Excel AS DECIMAL(10,2))) AS min_excel,
    >     MAX(CAST(Excel AS DECIMAL(10,2))) AS max_excel
    > FROM
    >     student_scores;
2024-03-31 20:20:44,215 INFO  [b13a3afa-e9be-42e9-a633-cd5d23081a9b main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.
ql.plan.VectorReduceSinkInfo@596d17f4
Query ID = root_20240331202043_d87e8988-98d3-4481-8b9f-0ee9b4e71d34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711915218439_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.72 s
----------------------------------------------------------------------------------------------
OK
0.00    1.00    0.01    1.00    0.00    1.00    0.01    1.00    0.00    0.99
Time taken: 8.737 seconds, Fetched: 1 row(s)
hive> 
```

It was discovered that more students were accepted into the program. After performing the analysis, it appears as though although the students placed had a lower average Python and SQL score, the average scores for machine learning, Tableau, and Excel were higher. There was no large discrepancy in the minimum and maximum of each column.

Reference:

Khan, S. S. (2022, July 17). *Student's scores*. Kaggle.
    https://www.kaggle.com/datasets/samarsaeedkhan/scores