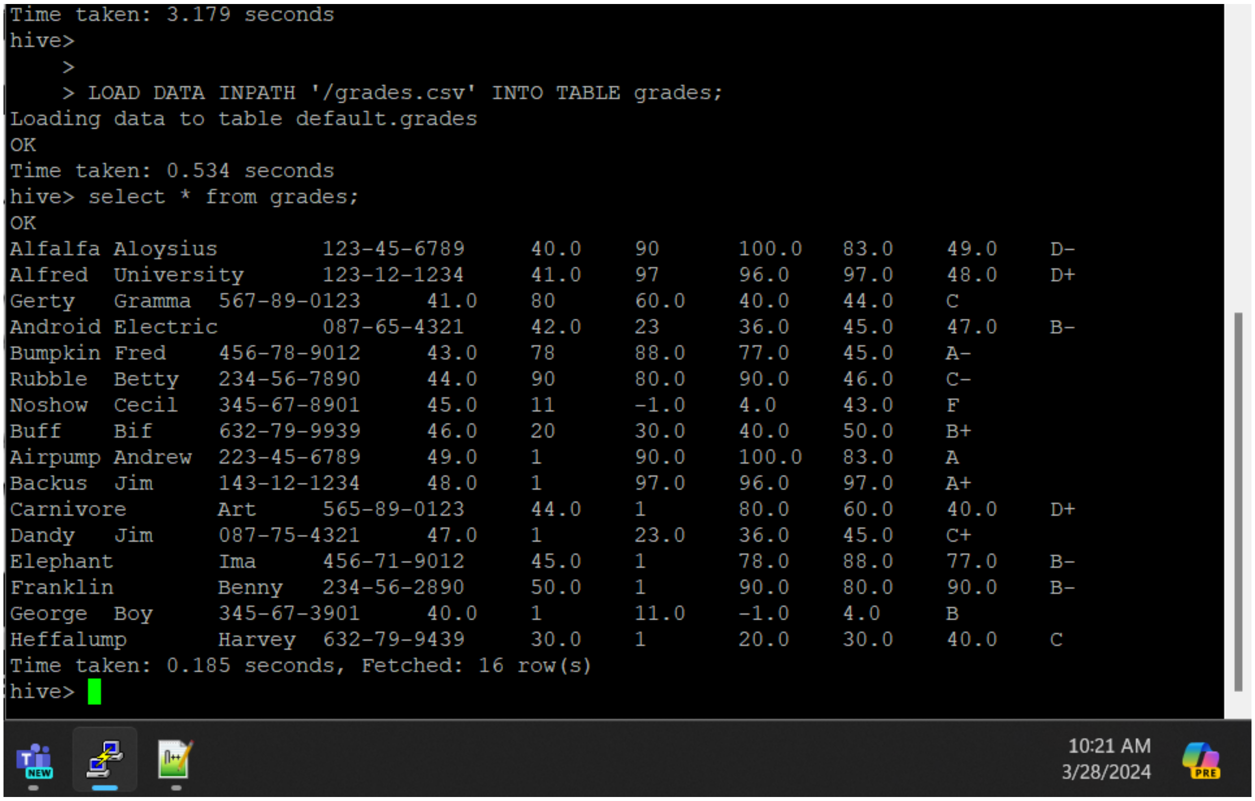
DSC 650

Week 3 Assignment

Eyram Kueviakoe

March 28, 2024

Screenshot 1: SELECT * FROM grades;



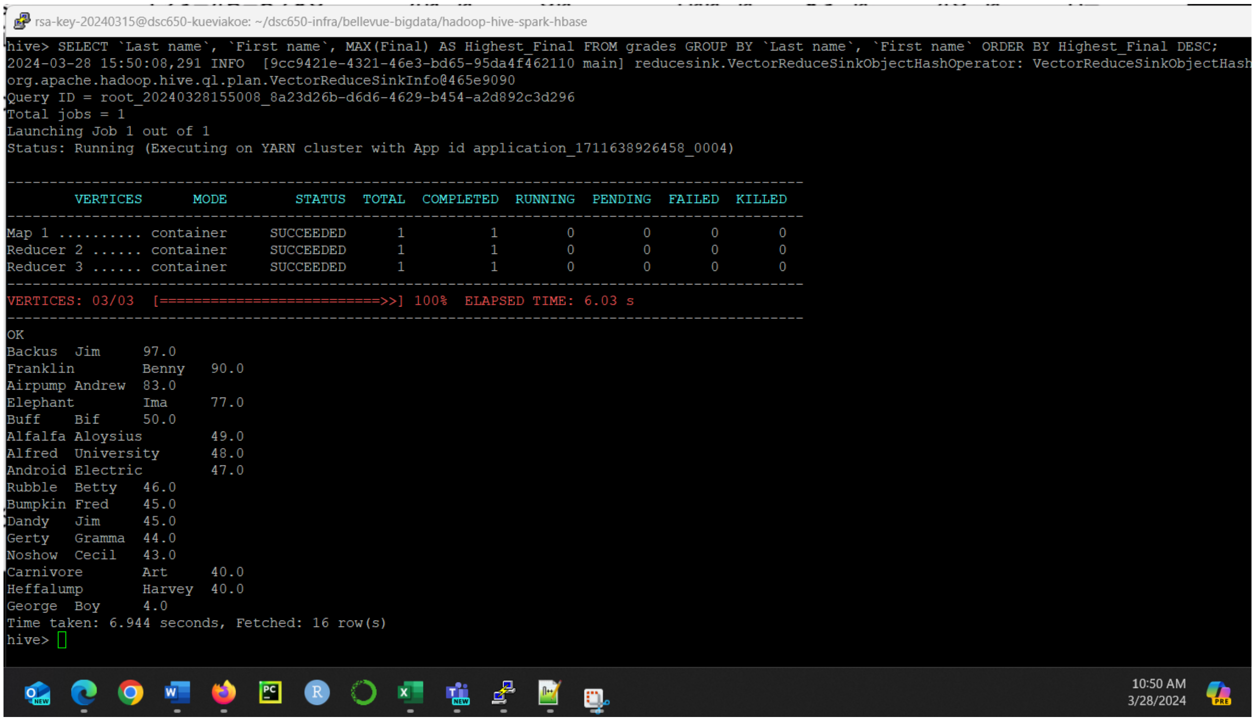
```
Time taken: 3.179 seconds
hive>
    >
    > LOAD DATA INPATH '/grades.csv' INTO TABLE grades;
Loading data to table default.grades
OK
Time taken: 0.534 seconds
hive> select * from grades;
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      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: 0.185 seconds, Fetched: 16 row(s)
hive>
```

Run 3 different SQL commands on the grades data

Query 1:

Screenshot 2: Students with highest Final exam score

SELECT ` Last name` , ` First name` , MAX(Final) AS Highest_Final FROM grades GROUP BY ` Last name` , ` First name` ORDER BY Highest_Final DESC;

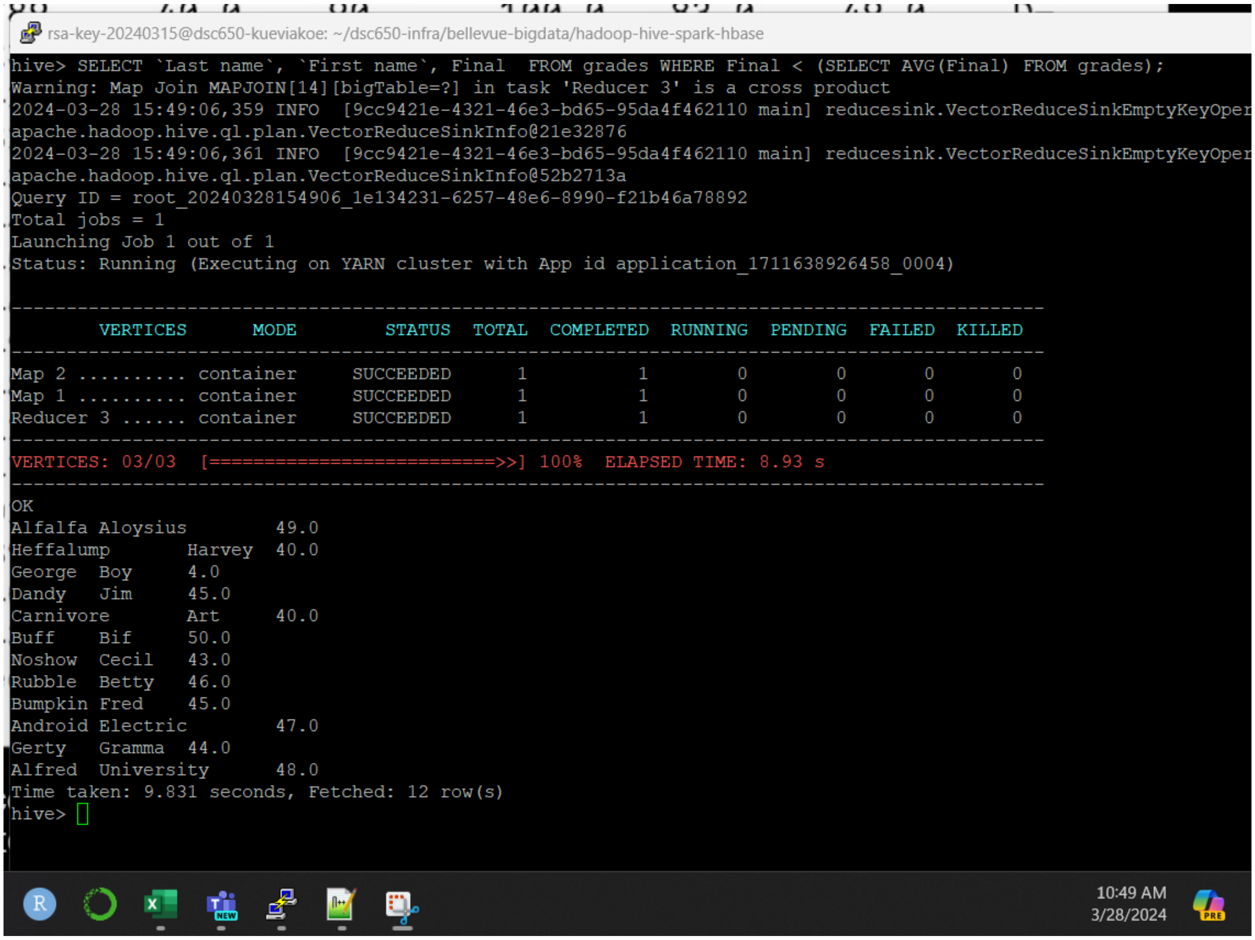

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT `Last name`, `First name`, MAX(Final) AS Highest_Final FROM grades GROUP BY `Last name`, `First name` ORDER BY Highest_Final DESC;
2024-03-28 15:50:08,291 INFO  [9cc9421e-4321-46e3-bd65-95da4f462110 main] reducesink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObjectHash
org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@465e9090
Query ID = root_20240328155008_8a23d26b-d6d6-4629-b454-a2d892c3d296
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711638926458_0004)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.03 s
----------------------------------------------------------------------------------------------
OK
Backus  Jim     97.0
Franklin        Benny   90.0
Airpump Andrew  83.0
Elephant        Ima     77.0
Buff    Bif     50.0
Alfalfa Aloysius        49.0
Alfred  University      48.0
Android Electric        47.0
Rubble  Betty   46.0
Bumpkin Fred    45.0
Dandy   Jim     45.0
Gerty   Gramma  44.0
Noshow  Cecil   43.0
Carnivore       Art     40.0
Heffalump       Harvey  40.0
George  Boy     4.0
Time taken: 6.944 seconds, Fetched: 16 row(s)
hive> 
```

Query 2: List of students who scored less than the average final exam score

SELECT ` Last name` , ` First name` , Final FROM grades WHERE Final < (SELECT AVG(Final) FROM grades);

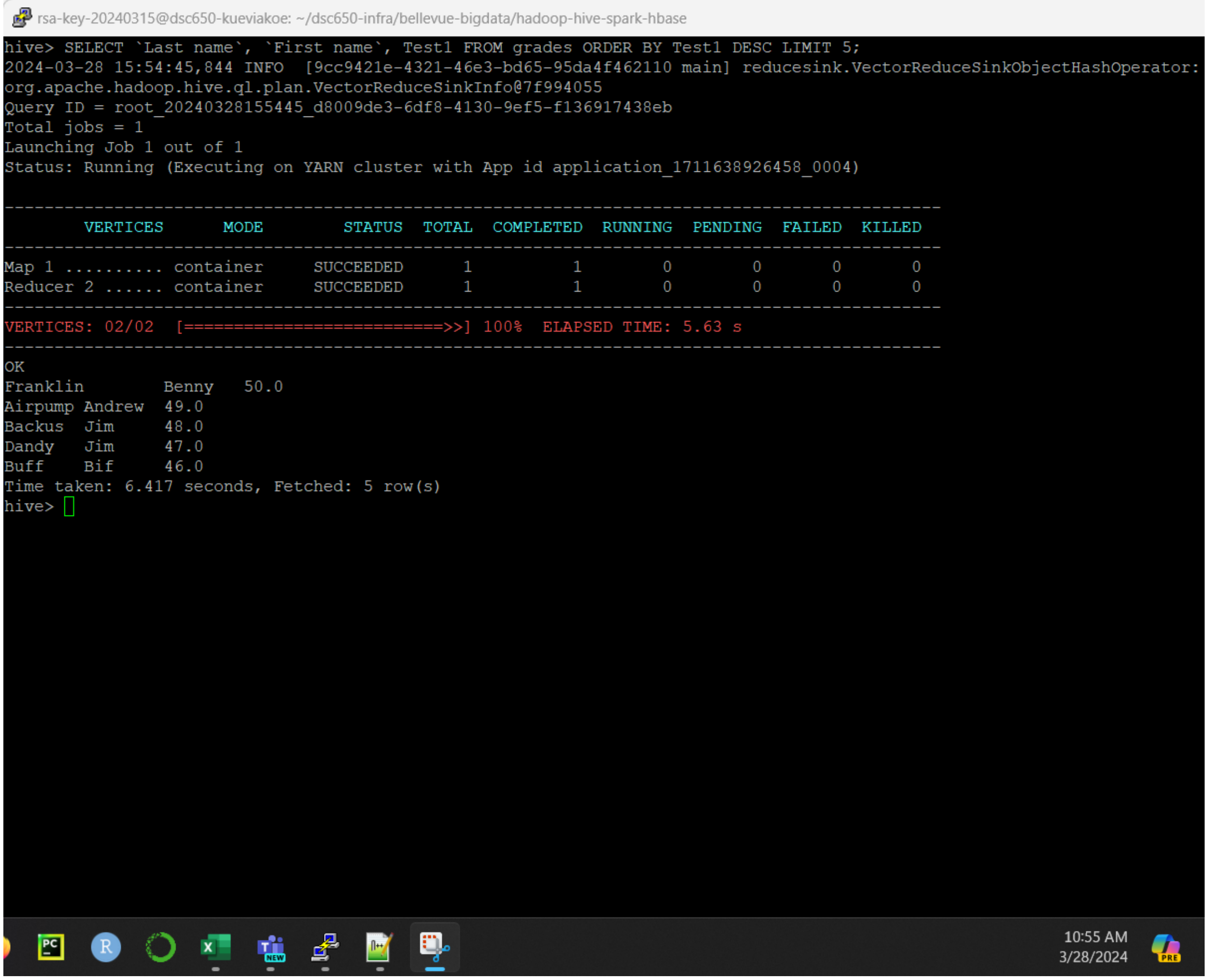

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
10:49 AM
3/28/2024

```
hive> SELECT `Last name`, `First name`, Final  FROM grades WHERE Final < (SELECT AVG(Final) FROM grades);
Warning: Map Join MAPJOIN[14][bigTable=?] in task 'Reducer 3' is a cross product
2024-03-28 15:49:06,359 INFO  [9cc9421e-4321-46e3-bd65-95da4f462110 main] reducesink.VectorReduceSinkEmptyKeyOper
apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@21e32876
2024-03-28 15:49:06,361 INFO  [9cc9421e-4321-46e3-bd65-95da4f462110 main] reducesink.VectorReduceSinkEmptyKeyOper
apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@52b2713a
Query ID = root_20240328154906_1e134231-6257-48e6-8990-f21b46a78892
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711638926458_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.93 s
----------------------------------------------------------------------------------------------
OK
Alfalfa Aloysius        49.0
Heffalump       Harvey  40.0
George  Boy     4.0
Dandy   Jim     45.0
Carnivore       Art     40.0
Buff    Bif     50.0
Noshow  Cecil   43.0
Rubble  Betty   46.0
Bumpkin Fred    45.0
Android Electric        47.0
Gerty   Gramma  44.0
Alfred  University      48.0
Time taken: 9.831 seconds, Fetched: 12 row(s)
hive> 
```

Query 3: Top 5 students with the highest scores in Test 1

SELECT ` Last name` , ` First name` , Test1 FROM grades ORDER BY Test1 DESC LIMIT 5;



rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT `Last name`, `First name`, Test1 FROM grades ORDER BY Test1 DESC LIMIT 5;
2024-03-28 15:54:45,844 INFO  [9cc9421e-4321-46e3-bd65-95da4f462110 main] reducesink.VectorReduceSinkObjectHashOperator:
org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@7f994055
Query ID = root_20240328155445_d8009de3-6df8-4130-9ef5-f136917438eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711638926458_0004)

----------------------------------------------------------------------------------------------
        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: 5.63 s
----------------------------------------------------------------------------------------------
OK
Franklin        Benny   50.0
Airpump Andrew  49.0
Backus  Jim     48.0
Dandy   Jim     47.0
Buff    Bif     46.0
Time taken: 6.417 seconds, Fetched: 5 row(s)
hive> 
```

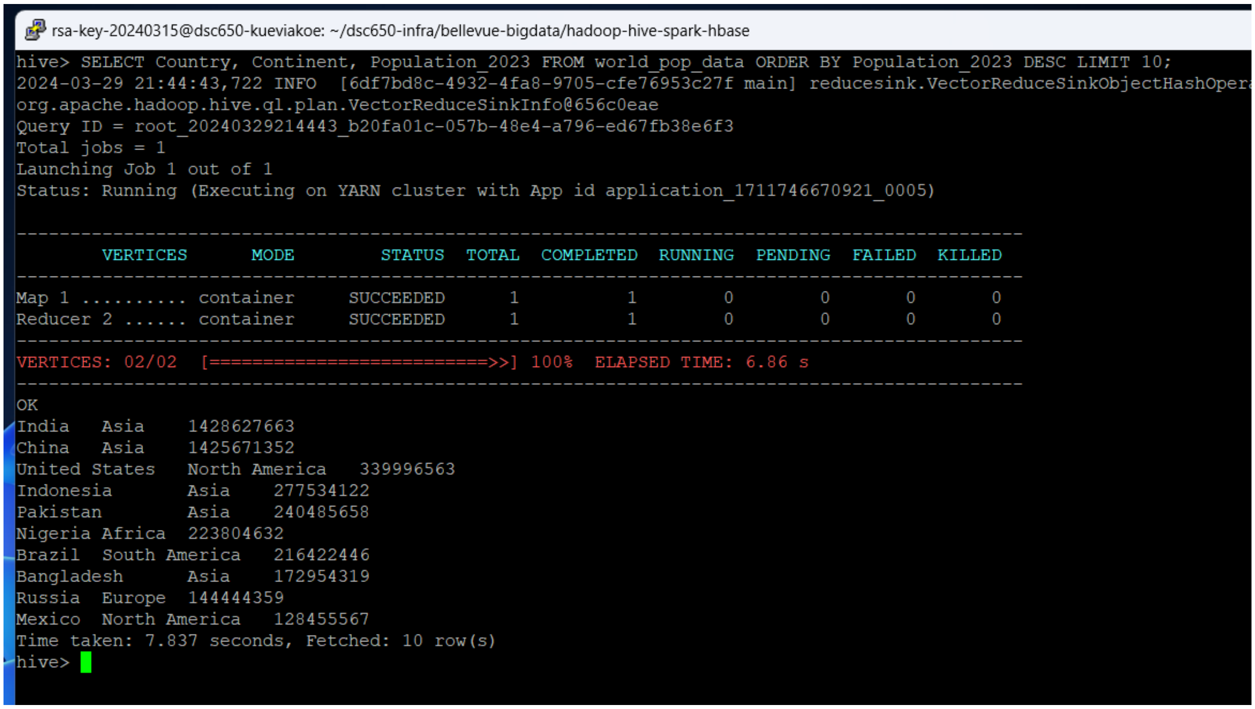
For the assignment, I am choosing the world population data. The data can be found on Kaggle, at https://www.kaggle.com/datasets/sazidthe1/world-population-data

I chose the world population dataset because it has information about population dynamics over time, including data from 1970 and 2023, along with details about country's areas and continents.

This dataset will allow us to compare demographics, analyze population density, and track the evolution of populations across different continents.

Query 1: 10 most populated countries in 2023

SELECT Country, Continent, Population_2023 FROM world_pop_data ORDER BY Population_2023 DESC LIMIT 10;


rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT Country, Continent, Population_2023 FROM world_pop_data ORDER BY Population_2023 DESC LIMIT 10;
2024-03-29 21:44:43,722 INFO  [6df7bd8c-4932-4fa8-9705-cfe76953c27f main] reducesink.VectorReduceSinkObjectHashOpera
org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@656c0eae
Query ID = root_20240329214443_b20fa01c-057b-48e4-a796-ed67fb38e6f3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711746670921_0005)

----------------------------------------------------------------------------------------------
        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: 6.86 s
----------------------------------------------------------------------------------------------
OK
India   Asia    1428627663
China   Asia    1425671352
United States   North America   339996563
Indonesia       Asia    277534122
Pakistan        Asia    240485658
Nigeria Africa  223804632
Brazil  South America   216422446
Bangladesh      Asia    172954319
Russia  Europe  144444359
Mexico  North America   128455567
Time taken: 7.837 seconds, Fetched: 10 row(s)
hive>
```

Query 2: Percentage of population growth between 1970 and 2023

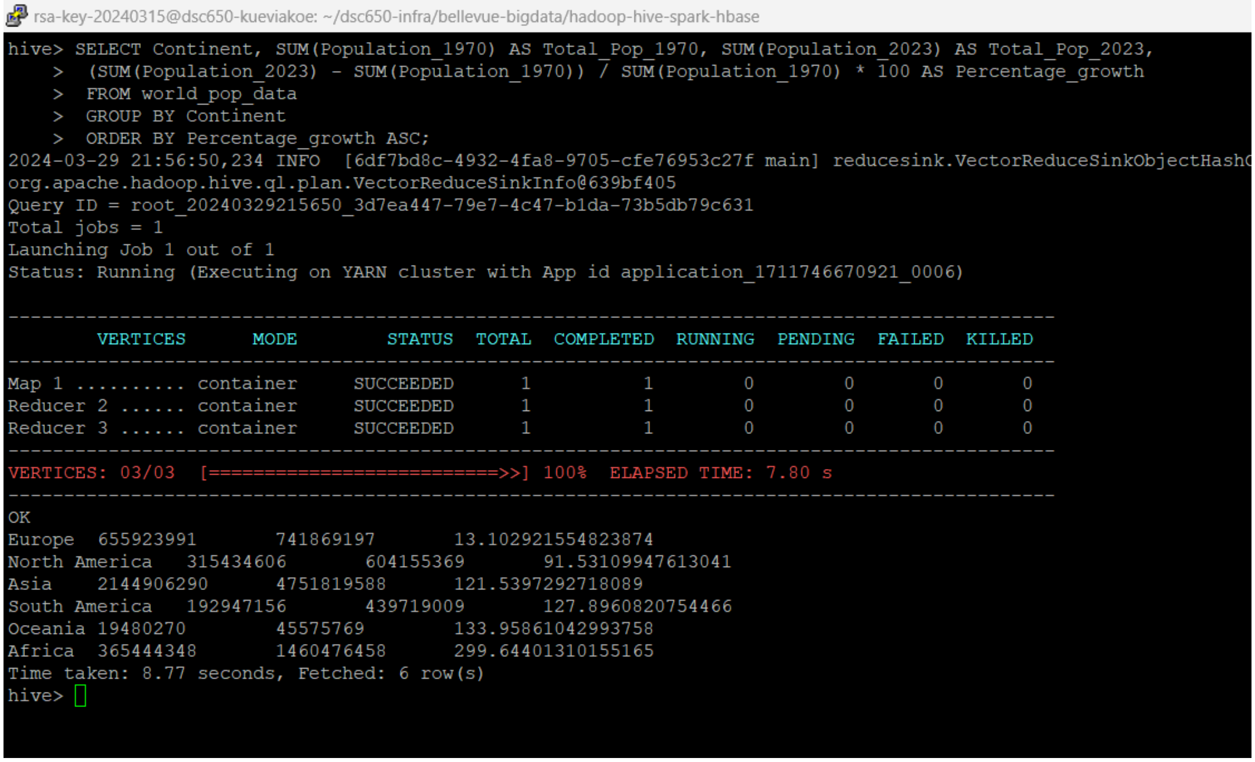
SELECT Continent, SUM(Population_1970) AS Total_Pop_1970, SUM(Population_2023) AS Total_Pop_2023,

*(SUM(Population_2023) - SUM(Population_1970)) / SUM(Population_1970) * 100 AS Percentage_growth*

FROM world_pop_data

GROUP BY Continent

ORDER BY Percentage_growth ASC;



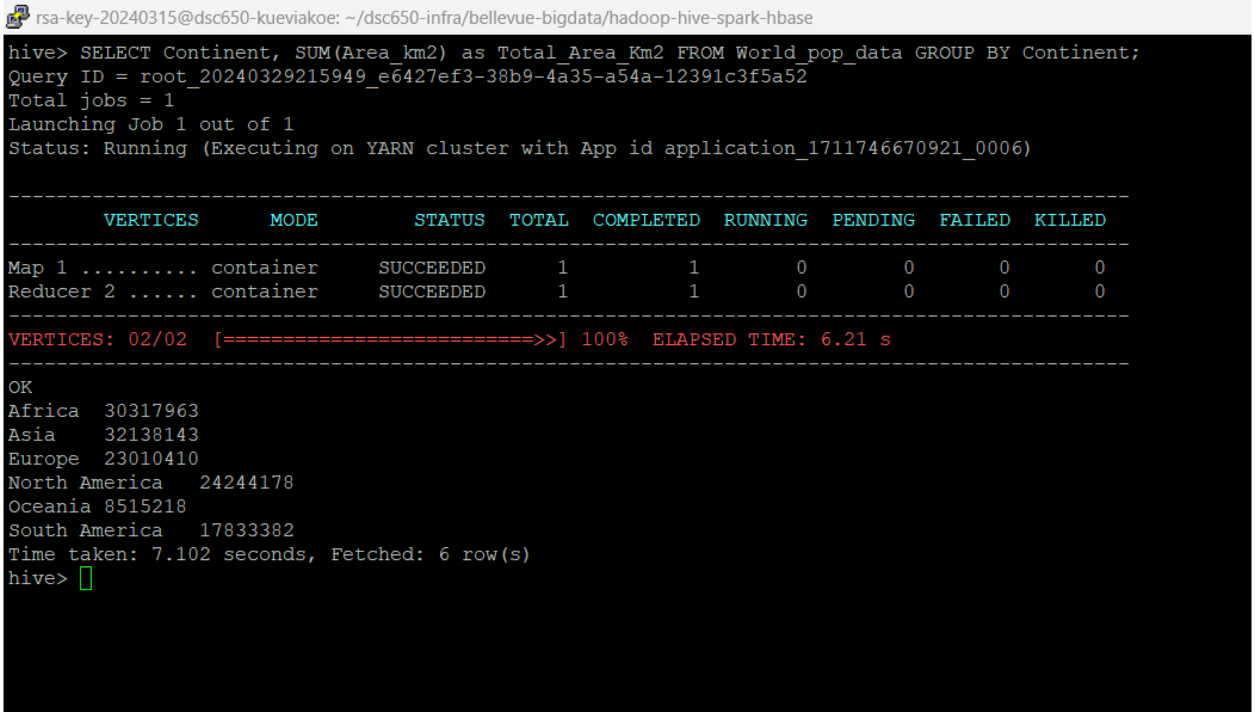
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT Continent, SUM(Population_1970) AS Total_Pop_1970, SUM(Population_2023) AS Total_Pop_2023,
    >  (SUM(Population_2023) - SUM(Population_1970)) / SUM(Population_1970) * 100 AS Percentage_growth
    >  FROM world_pop_data
    >  GROUP BY Continent
    >  ORDER BY Percentage_growth ASC;
2024-03-29 21:56:50,234 INFO  [6df7bd8c-4932-4fa8-9705-cfe76953c27f main] reducesink.VectorReduceSinkObjectHashC
org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@639bf405
Query ID = root_20240329215650_3d7ea447-79e7-4c47-b1da-73b5db79c631
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711746670921_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.80 s
----------------------------------------------------------------------------------------------
OK
Europe  655923991       741869197       13.102921554823874
North America   315434606       604155369       91.53109947613041
Asia    2144906290      4751819588      121.5397292718089
South America   192947156       439719009       127.8960820754466
Oceania 19480270        45575769        133.95861042993758
Africa  365444348       1460476458      299.64401310155165
Time taken: 8.77 seconds, Fetched: 6 row(s)
hive> 
```

Query 3: Size of each continent

SELECT Continent, SUM(Area_km2) as Total_Area_Km2 FROM World_pop_data GROUP BY Continent;



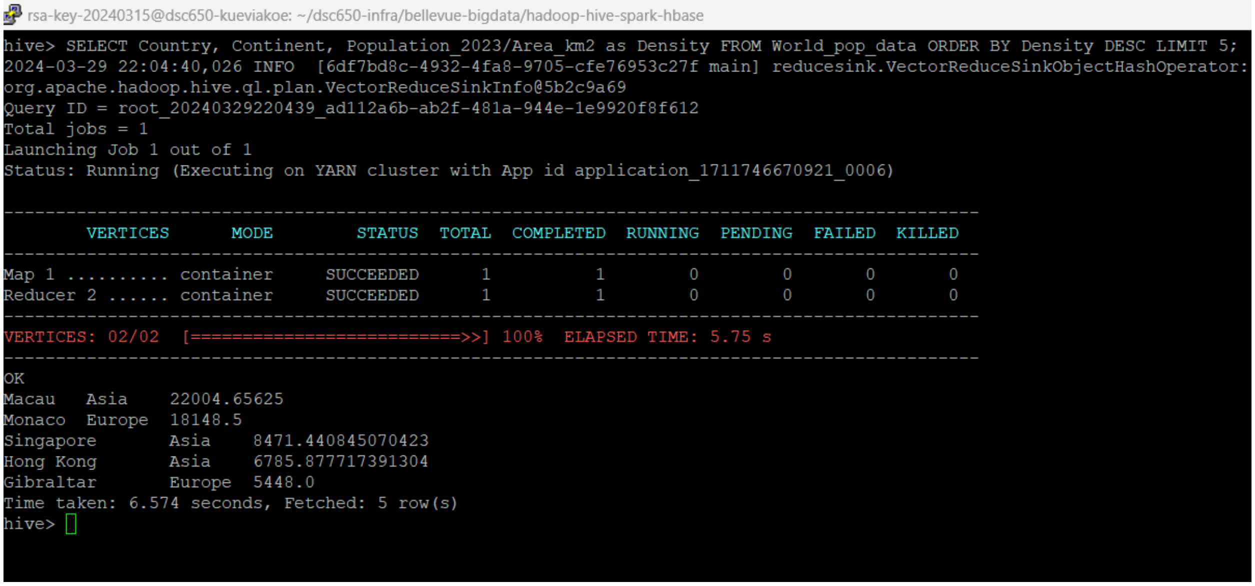
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT Continent, SUM(Area_km2) as Total_Area_Km2 FROM World_pop_data GROUP BY Continent;
Query ID = root_20240329215949_e6427ef3-38b9-4a35-a54a-12391c3f5a52
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711746670921_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: 6.21 s
----------------------------------------------------------------------------------------------
OK
Africa  30317963
Asia    32138143
Europe  23010410
North America   24244178
Oceania 8515218
South America   17833382
Time taken: 7.102 seconds, Fetched: 6 row(s)
hive>
```

Query 4: What are the top 5 countries with highest density in 2023

SELECT Country, Continent, Population_2023/Area_km2 as Density FROM World_pop_data ORDER BY Density DESC LIMIT 5;


rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> SELECT Country, Continent, Population_2023/Area_km2 as Density FROM World_pop_data ORDER BY Density DESC LIMIT 5;
2024-03-29 22:04:40,026 INFO  [6df7bd8c-4932-4fa8-9705-cfe76953c27f main] reducesink.VectorReduceSinkObjectHashOperator:
org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@5b2c9a69
Query ID = root_20240329220439_ad112a6b-ab2f-481a-944e-1e9920f8f612
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711746670921_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: 5.75 s
----------------------------------------------------------------------------------------------
OK
Macau   Asia    22004.65625
Monaco  Europe  18148.5
Singapore       Asia    8471.440845070423
Hong Kong       Asia    6785.877717391304
Gibraltar       Europe  5448.0
Time taken: 6.574 seconds, Fetched: 5 row(s)
hive> 
```

Query 5: Continents' density in 1970 and 2023

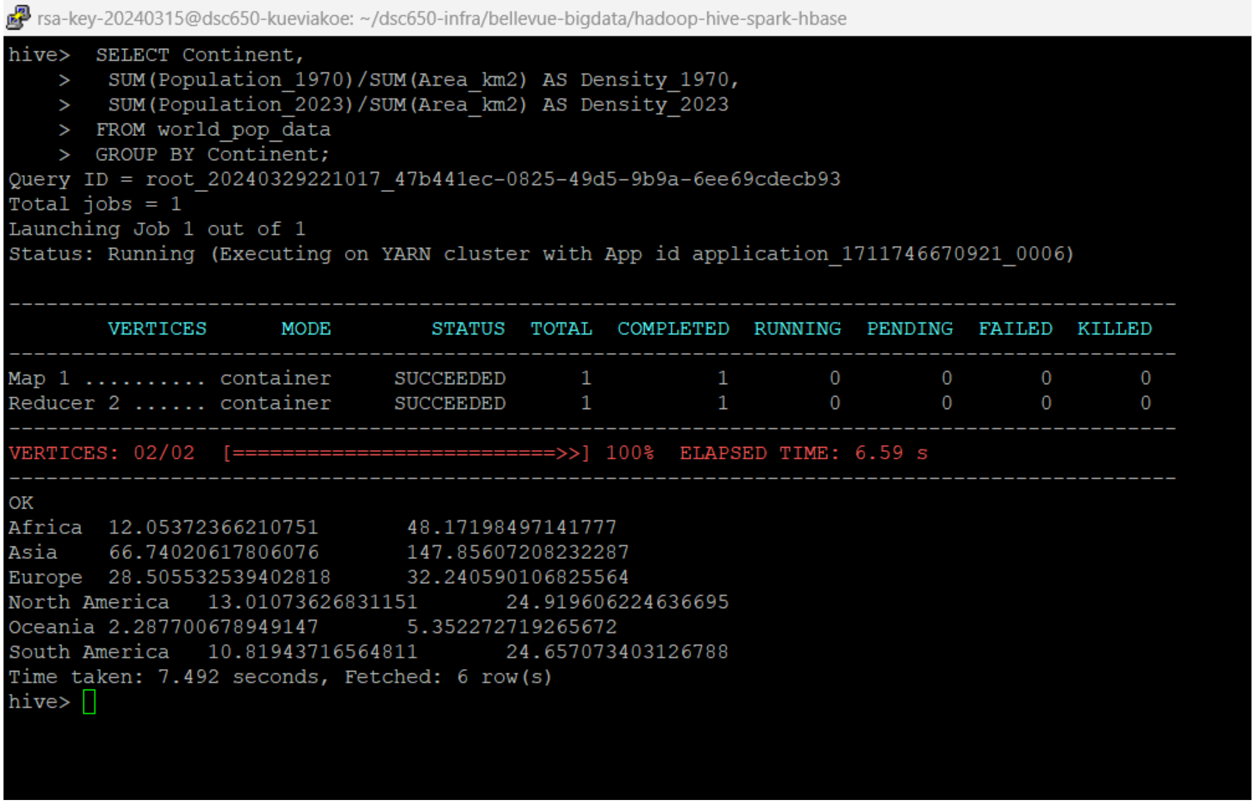
SELECT Continent,

SUM(Population_1970)/SUM(Area_km2) AS Density_1970,

SUM(Population_2023)/SUM(Area_km2) AS Density_2023

FROM world_pop_data

GROUP BY Continent;


rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive>  SELECT Continent,
    >   SUM(Population_1970)/SUM(Area_km2) AS Density_1970,
    >   SUM(Population_2023)/SUM(Area_km2) AS Density_2023
    >  FROM world_pop_data
    >  GROUP BY Continent;
Query ID = root_20240329221017_47b441ec-0825-49d5-9b9a-6ee69cdecb93
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711746670921_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: 6.59 s
----------------------------------------------------------------------------------------------
OK
Africa  12.05372366210751       48.17198497141777
Asia    66.74020617806076       147.85607208232287
Europe  28.505532539402818      32.240590106825564
North America   13.01073626831151       24.919606224636695
Oceania 2.287700678949147       5.352272719265672
South America   10.81943716564811       24.657073403126788
Time taken: 7.492 seconds, Fetched: 6 row(s)
hive> 
```