# Week 3 Assignment

Guruprasad Velikadu Krishnamoorthy

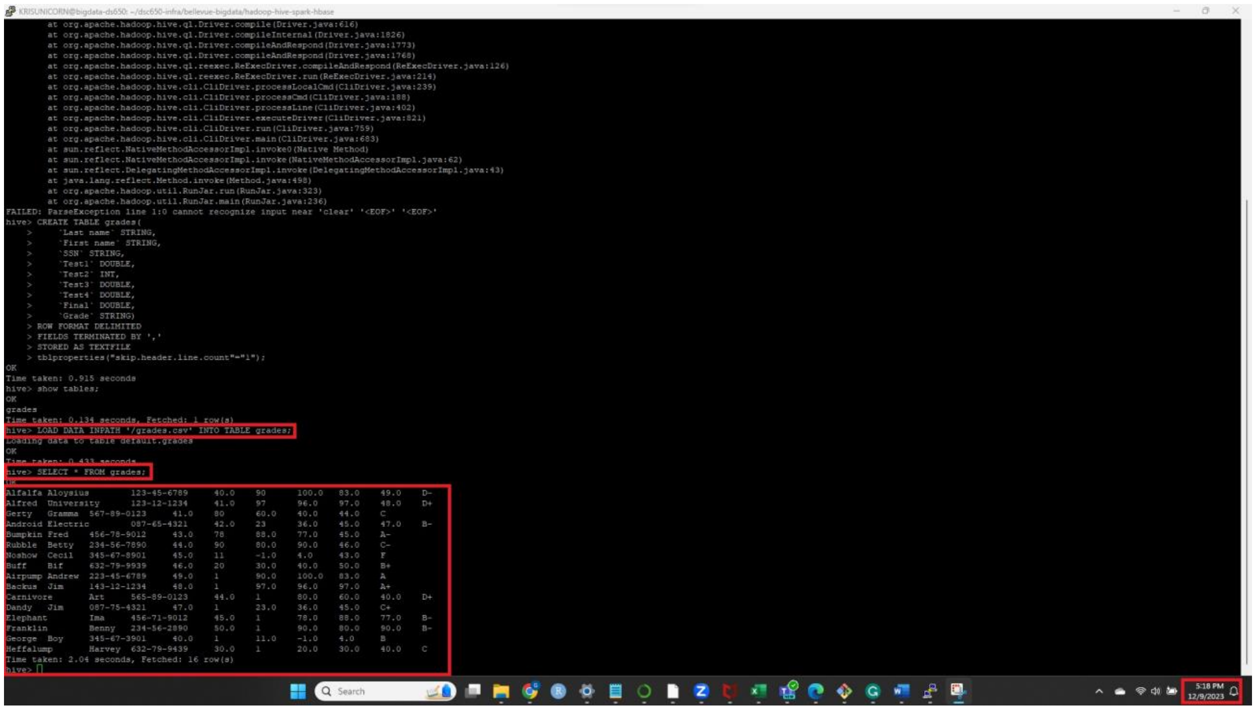
College of Science and Technology, Bellevue University

DSC650-T301: Big Data

Professor. Nasheb Ismaily

December 10, 2023

1. **Loading the grades.csv file into Hive:**
   The Screenshot below shows grades.csv is loaded into the Hive table and a select query is executed:



KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:0 cannot recognize input near 'clear' '<EOF>' '<EOF>'
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: 0.915 seconds
hive> show tables;
OK
grades
Time taken: 0.134 seconds, Fetched: 1 row(s)
hive> LOAD DATA INPATH '/grades.csv' INTO TABLE grades;
Loading data to table default.grades
OK
Time taken: 0.433 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: 2.04 seconds, Fetched: 16 row(s)
hive> 
```

5:18 PM
12/9/2023

2. **Screenshot of Commands executed on grades table:**

   Commands used are:

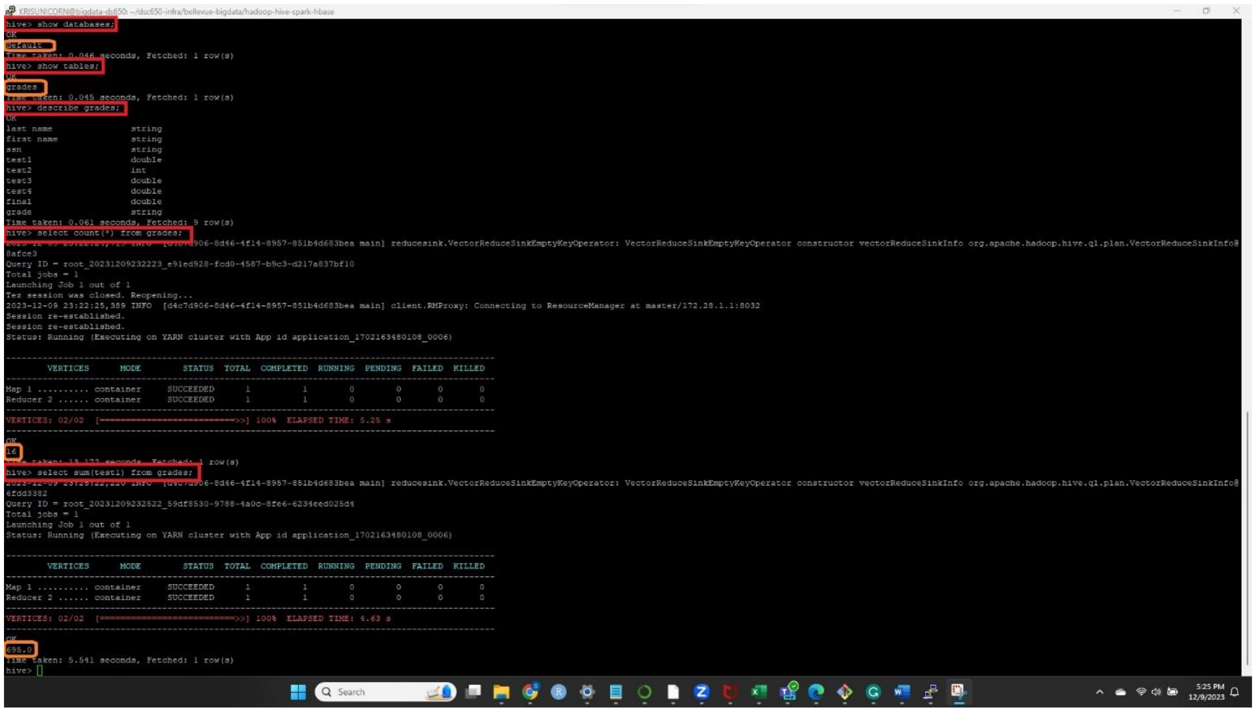
   1. Select count(*) from grades;
      ***Significance:*** *To check the count of rows in the table*

   2. Show databases;
      ***Significance:*** *Returns the list of databases.*

   3. Show tables;
      ***Significance:*** *Returns the list of tables.*

   4. Describe grades;
      ***Significance:*** *Returns the table structure with the name of the column and its datatype.*

   5. Select sum(test1) from grades;
      ***Significance:*** *returns the sum of all values of test1 variable in the grades table.*
   6. Select grade,count(*) from grades group by grade;
      ***Significance:*** *returns the total count of each grade scored by the students.*

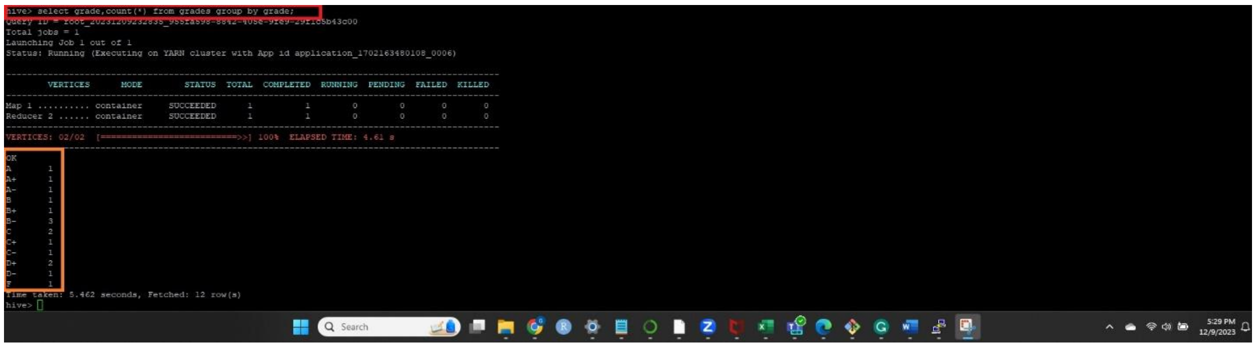


KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
hive> show databases;
OK
default
Time taken: 0.046 seconds, Fetched: 1 row(s)
hive> show tables;
OK
grades
Time taken: 0.045 seconds, Fetched: 1 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: 0.061 seconds, Fetched: 9 row(s)
hive> select count(*) from grades;
2023-12-09 23:22:21,719 INFO  [d4c7d906-8d46-4f14-8957-851b4d683bea main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@
8afce3
Query ID = root_20231209232223_e91ed928-fcd0-4587-b9c3-d217a837bf10
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2023-12-09 23:22:25,389 INFO  [d4c7d906-8d46-4f14-8957-851b4d683bea 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_1702163480108_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: 5.25 s
----------------------------------------------------------------------------------------------
OK
16
Time taken: 13.172 seconds, Fetched: 1 row(s)
hive> select sum(test1) from grades;
2023-12-09 23:25:22,220 INFO  [d4c7d906-8d46-4f14-8957-851b4d683bea main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@
6fdd3382
Query ID = root_20231209232522_59df8530-9788-4a0c-8fe6-6234eed025d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_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: 4.63 s
----------------------------------------------------------------------------------------------
OK
695.0
Time taken: 5.541 seconds, Fetched: 1 row(s)
hive>
```



```
hive> select grade,count(*) from grades group by grade;
Query ID = root_20231209232835_955fa598-8842-405e-9fe9-29f1c5b43c00
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_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: 4.61 s
----------------------------------------------------------------------------------------------
OK
A       1
A+      1
A-      1
B       1
B+      1
B-      3
C       2
C+      1
C-      1
D+      2
D-      1
F       1
Time taken: 5.462 seconds, Fetched: 12 row(s)
hive>
```

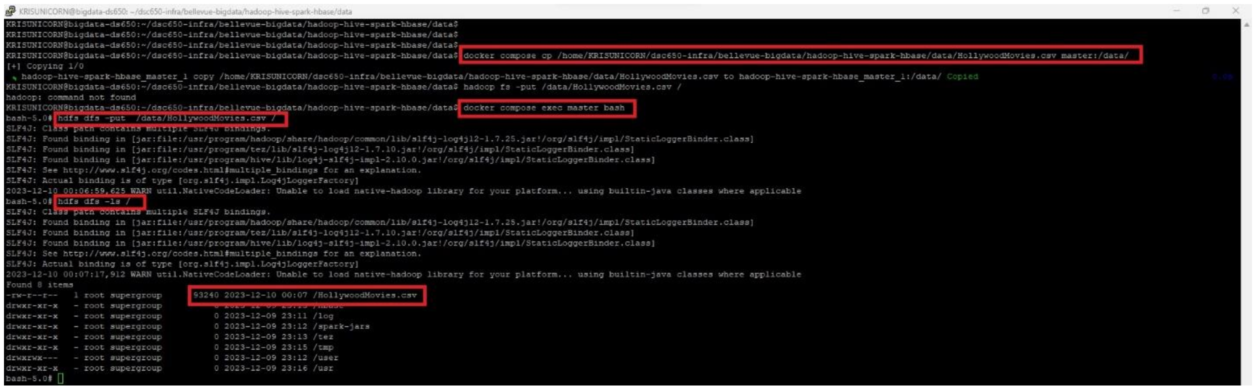
3. **Loading a new Dataset into Hive:**
   For this exercise, the Hollywood movies dataset from GitHub was downloaded from https://raw.githubusercontent.com/reisanar/datasets/master/HollywoodMovies.csv.

   Commands used are listed below:

   - *# To download the file from GitHub into Google Cloud VM.*
     wget https://raw.githubusercontent.com/reisanar/datasets/master/HollywoodMovies.csv

   - *# After starting the docker, copying the file into master container from the VM.*
     docker compose cp /home/KRISUNICORN/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data/HollywoodMovies.csv master:/data/

   - *# Copying the file from Master Container into HDFS filesystem in the / directory*
     hdfs dfs -put /data/HollywoodMovies.csv /



```
KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$ docker compose cp /home/KRISUNICORN/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data/HollywoodMovies.csv master:/data/
[+] Copying 1/0
 ✔ hadoop-hive-spark-hbase_master_1 copy /home/KRISUNICORN/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data/HollywoodMovies.csv to hadoop-hive-spark-hbase_master_1:/data/ Copied
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$ hadoop fs -put /data/HollywoodMovies.csv /
hadoop: command not found
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data$ docker compose exec master bash
bash-5.0# hdfs dfs -put  /data/HollywoodMovies.csv /
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]
2023-12-10 00:06:59,625 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bash-5.0# hdfs dfs -ls /
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]
2023-12-10 00:07:17,912 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
-rw-r--r--   1 root supergroup      93240 2023-12-10 00:07 /HollywoodMovies.csv
drwxr-xr-x   - root supergroup          0 2023-12-09 23:1[illegible] /hbase
drwxr-xr-x   - root supergroup          0 2023-12-09 23:11 /log
drwxr-xr-x   - root supergroup          0 2023-12-09 23:12 /spark-jars
drwxr-xr-x   - root supergroup          0 2023-12-09 23:13 /tez
drwxr-xr-x   - root supergroup          0 2023-12-09 23:15 /tmp
drwxrwx---   - root supergroup          0 2023-12-09 23:12 /user
drwxr-xr-x   - root supergroup          0 2023-12-09 23:16 /usr
bash-5.0# 
```

4. **The Reason for choosing the dataset?**

   The movie dataset contains 970 Rows and 16 columns that contain the details of the profitable Hollywood movies released from 2007 to 2013. As the number of records wasn't too small, it seemed perfect for this assignment project. Also, I am a movie enthusiast, so I thought it may be a little fun to do a Data Science project on Movie related datasets.

5. **The description of some of the commonly used columns is explained below:**
   - ***Movie:*** Name of the movie
   - ***LeadStudio:*** The Studio that produced the movie.
   - ***RottenTomatoes:*** Represents the score rated by Rotten Tomatoes critics.
   - ***AudienceScore***: Represents the score rated by Audience critics.
   - ***Genre***: Represents the Genre of the movie
   - ***TheatersOpenWeek***: Number of theatres where the movie was released in the opening week in the US and Canada. This is an integer value.
   - ***OpeningWeekend*****:** Box Office collection in the opening week in the US and Canada in USD. This is a floating number value.
   - ***DomesticGross:*** *Total domestic collection in the US and Canada represented in USD.*
   - ***ForeignGross:*** *Total collection from countries other than the US and Canada represented in USD.*
   - ***WorldGross:*** Sum of Domestic and Foreign Gross collection represented in USD.
   - ***Budget:*** Total movie budget represented in USD.
   - ***Profitability:*** Total profits made from the movie represented in USD
   - ***Year***: The year in which the movie was released.

6. **What is expected to be achieved by analyzing the dataset?**

   The questions that we are trying to answer by analyzing the dataset are listed below:

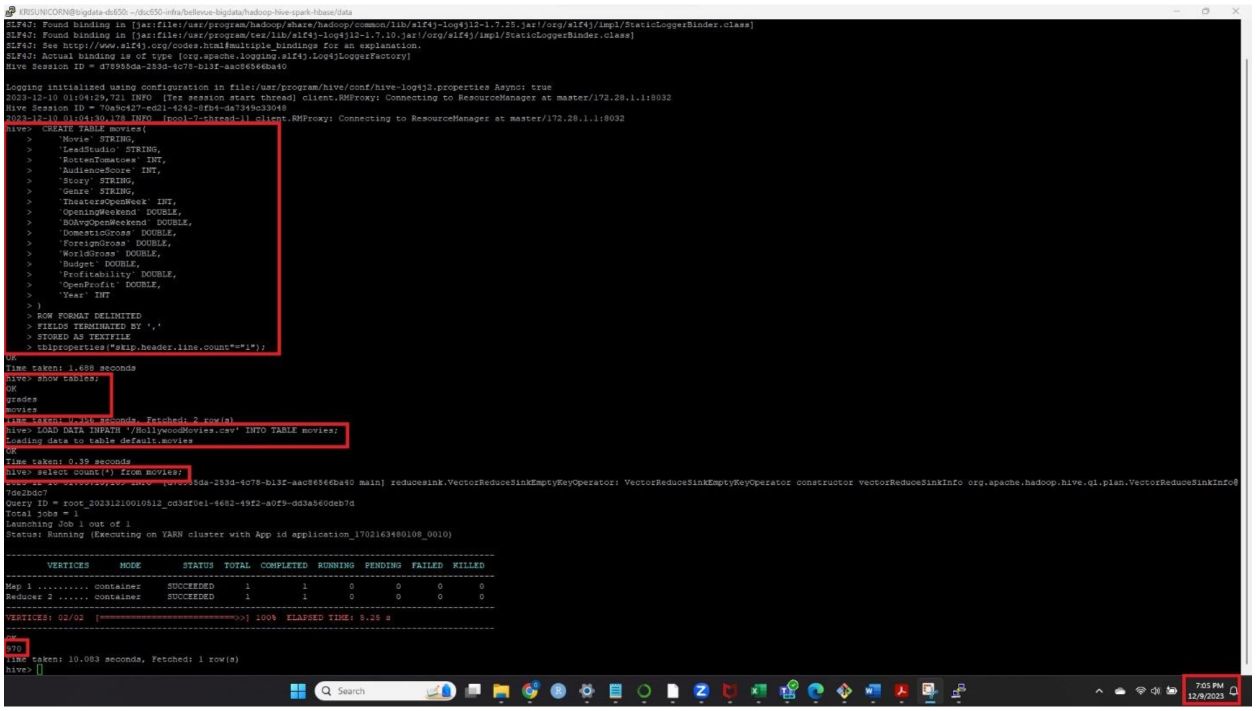
   1. How many movies had a perfect Rotten Tomatoes score of 95% and above in the dataset?
   2. How many Profitable movies were made each year?
   3. What was the average profit made each year in each of the Genres?

4. What are the variance, Mean, and Standard Deviation values of Domestic Gross each year?
5. What was the most profitable movie made each year and by how much?

**7. Creating tables and loading data in Hive:**

```
# Creating Hive table and defining the datatype of each column
CREATE TABLE movies(
   `Movie` STRING,
   `LeadStudio` STRING,
   `RottenTomatoes` INT,
   `AudienceScore` INT,
   `Story` STRING,
   `Genre` STRING,
   `TheatersOpenWeek` INT,
   `OpeningWeekend` DOUBLE,
   `BOAvgOpenWeekend` DOUBLE,
   `DomesticGross` DOUBLE,
   `ForeignGross` DOUBLE,
   `WorldGross` DOUBLE,
   `Budget` DOUBLE,
   `Profitability` DOUBLE,
   `OpenProfit` DOUBLE,
   `Year` INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
tblproperties("skip.header.line.count"="1");

# Loading the movies table with the csv file data.
LOAD DATA INPATH '/HollywoodMovies.csv' INTO TABLE movies;
```



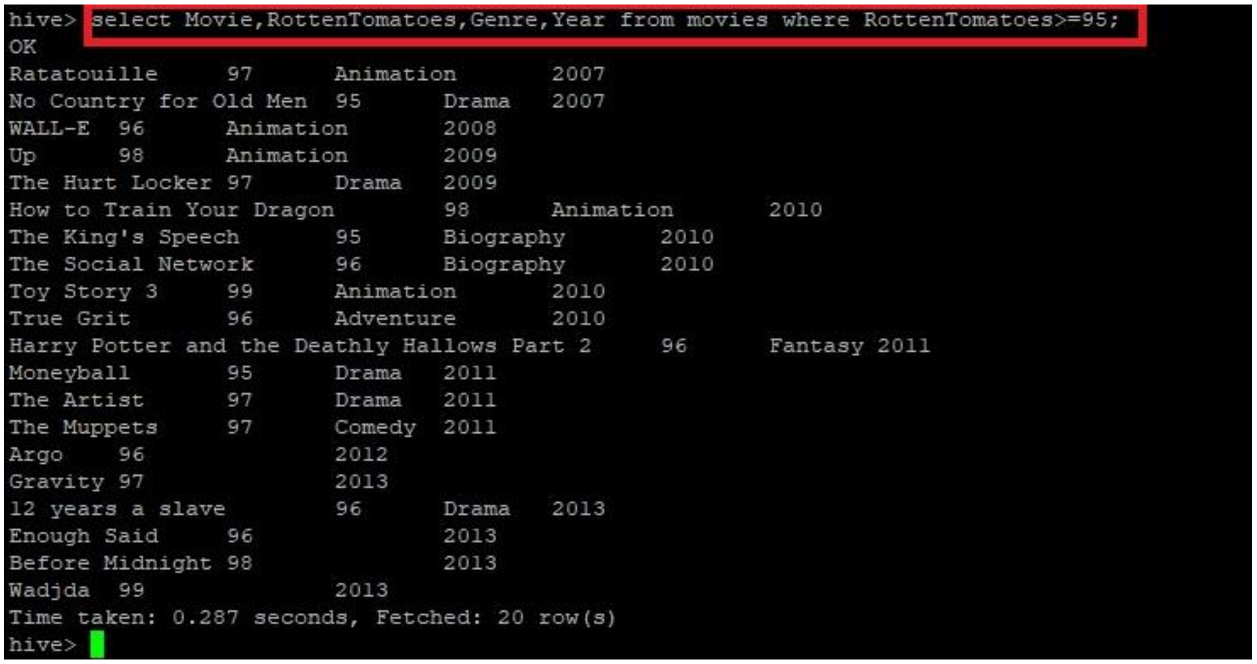
```
KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
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 = d78955da-253d-4c78-b13f-aac86566ba40

Logging initialized using configuration in file:/usr/program/hive/conf/hive-log4j2.properties Async: true
2023-12-10 01:04:29,721 INFO  [Tez session start thread] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Hive Session ID = 70a9c427-ed21-4242-8fb4-da7349c33048
2023-12-10 01:04:30,178 INFO  [pool-7-thread-1] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
hive>  CREATE TABLE movies(
    >      `Movie` STRING,
    >      `LeadStudio` STRING,
    >      `RottenTomatoes` INT,
    >      `AudienceScore` INT,
    >      `Story` STRING,
    >      `Genre` STRING,
    >      `TheatersOpenWeek` INT,
    >      `OpeningWeekend` DOUBLE,
    >      `BOAvgOpenWeekend` DOUBLE,
    >      `DomesticGross` DOUBLE,
    >      `ForeignGross` DOUBLE,
    >      `WorldGross` DOUBLE,
    >      `Budget` DOUBLE,
    >      `Profitability` DOUBLE,
    >      `OpenProfit` DOUBLE,
    >      `Year` INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.688 seconds
hive> show tables;
OK
grades
movies
Time taken: 0.356 seconds, Fetched: 2 row(s)
hive> LOAD DATA INPATH '/HollywoodMovies.csv' INTO TABLE movies;
Loading data to table default.movies
OK
Time taken: 0.39 seconds
hive> select count(*) from movies;
2023-12-10 01:05:10,169 INFO  [d78955da-253d-4c78-b13f-aac86566ba40 main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.ql.plan.VectorReduceSinkInfo@
7de2bdc7
Query ID = root_20231210010512_cd3df0e1-4682-49f2-a0f9-dd3a560deb7d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_0010)

----------------------------------------------------------------------------------------------
        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: 5.25 s
----------------------------------------------------------------------------------------------
OK
970
Time taken: 10.083 seconds, Fetched: 1 row(s)
hive>
```

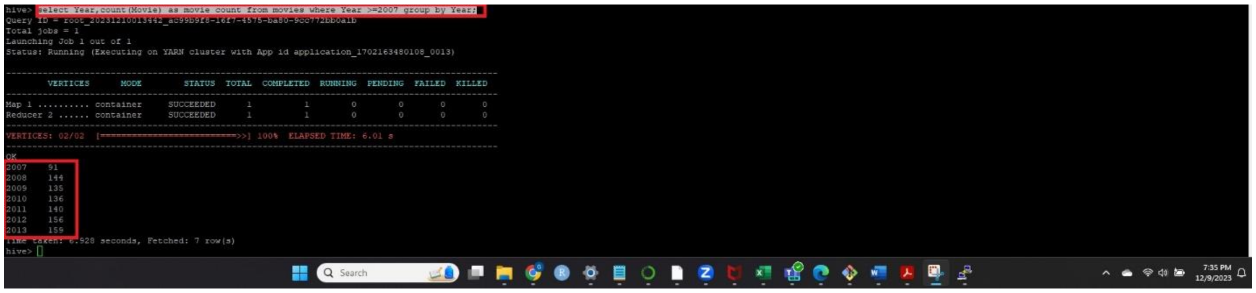
8. **Running Hive Queries:**

   - **Command 1:** *select Movie,RottenTomatoes,Genre,Year from movies where RottenTomatoes >=95;*
   - **Significance:** The query returns the movie name, Genre name and Year along with Rotten Tomates where the score is more than or equal to 95.



```
hive> select Movie,RottenTomatoes,Genre,Year from movies where RottenTomatoes>=95;
OK
Ratatouille     97      Animation       2007
No Country for Old Men  95      Drama   2007
WALL-E  96      Animation       2008
Up      98      Animation       2009
The Hurt Locker 97      Drama   2009
How to Train Your Dragon        98      Animation       2010
The King's Speech       95      Biography       2010
The Social Network      96      Biography       2010
Toy Story 3     99      Animation       2010
True Grit       96      Adventure       2010
Harry Potter and the Deathly Hallows Part 2     96      Fantasy 2011
Moneyball       95      Drama   2011
The Artist      97      Drama   2011
The Muppets     97      Comedy  2011
Argo    96              2012
Gravity 97              2013
12 years a slave        96      Drama   2013
Enough Said     96              2013
Before Midnight 98              2013
Wadjda  99              2013
Time taken: 0.287 seconds, Fetched: 20 row(s)
hive>
```

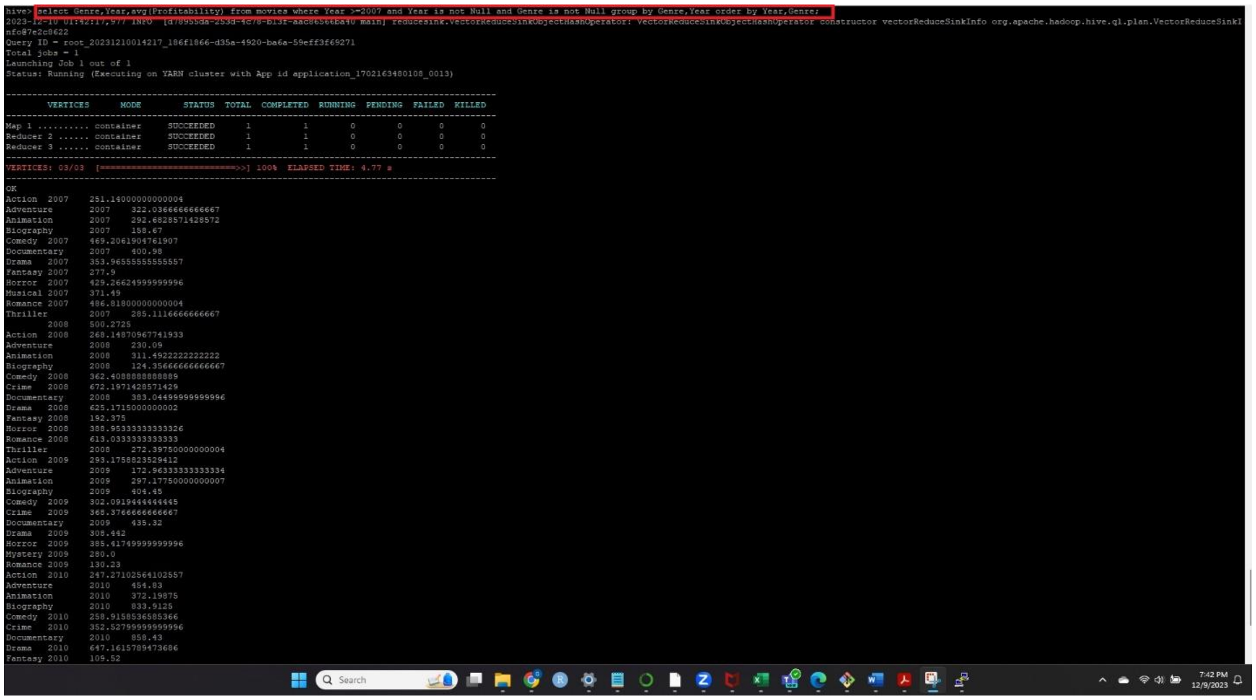
- **Command 2:** *select Year,count(Movie) from movies where Year >=2007 group by Year;*
- **Significance:** The query returns the count of number of movies from each year in the dataset.



```
hive> select Year,count(Movie) as movie count from movies where Year >=2007 group by Year;
Query ID = root_20231210013442_ac99b9f8-16f7-4575-ba80-9cc772bb0a1b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_0013)

----------------------------------------------------------------------------------------------
        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: 6.01 s
----------------------------------------------------------------------------------------------
OK
2007    91
2008    144
2009    135
2010    136
2011    140
2012    156
2013    159
Time taken: 6.928 seconds, Fetched: 7 row(s)
hive>
```

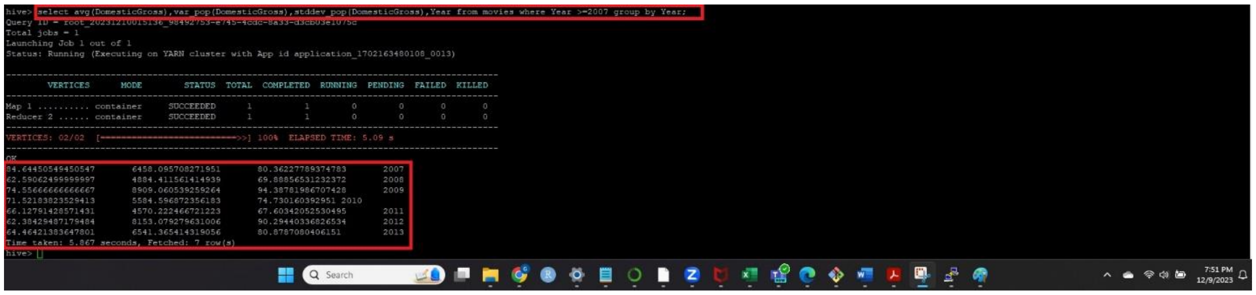
- **Command 3:** *select Genre,Year,avg(Profitability) from movies where Year >=2007 and Year is not Null and Genre is not Null group by Genre,Year order by Year,Genre;*
- **Significance:** The Query returns the average Profitability of the movies from the years 2007-2013 per Genre.



```
hive> select Genre,Year,avg(Profitability) from movies where Year >=2007 and Year is not Null and Genre is not Null group by Genre,Year order by Year,Genre;
2023-12-10 01:42:17,977 INFO  [d78955da-253d-4c78-b13f-aac86566ba40 main] reducesink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObjectHashOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.ql.plan.VectorReduceSinkI
nfo@7e2c8622
Query ID = root_20231210014217_186f1866-d35a-4920-ba6a-59eff3f69271
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_0013)

----------------------------------------------------------------------------------------------
        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: 4.77 s
----------------------------------------------------------------------------------------------
OK
Action  2007    251.14000000000004
Adventure       2007    322.0366666666667
Animation       2007    292.6828571428572
Biography       2007    158.67
Comedy  2007    469.2061904761907
Documentary     2007    400.98
Drama   2007    353.96555555555557
Fantasy 2007    277.9
Horror  2007    429.26624999999996
Musical 2007    371.49
Romance 2007    486.81800000000004
Thriller        2007    285.1116666666667
        2008    500.2725
Action  2008    268.14870967741933
Adventure       2008    230.09
Animation       2008    311.4922222222222
Biography       2008    124.35666666666667
Comedy  2008    362.4088888888889
Crime   2008    672.1971428571429
Documentary     2008    383.04499999999996
Drama   2008    625.1715000000002
Fantasy 2008    192.375
Horror  2008    388.95333333333326
Romance 2008    613.0333333333333
Thriller        2008    272.39750000000004
Action  2009    293.1758823529412
Adventure       2009    172.96333333333334
Animation       2009    297.17750000000007
Biography       2009    404.45
Comedy  2009    302.0919444444445
Crime   2009    368.3766666666667
Documentary     2009    435.32
Drama   2009    308.442
Horror  2009    385.41749999999996
Mystery 2009    280.0
Romance 2009    130.23
Action  2010    247.27102564102557
Adventure       2010    454.83
Animation       2010    372.19875
Biography       2010    833.9125
Comedy  2010    258.9158536585366
Crime   2010    352.52799999999996
Documentary     2010    858.43
Drama   2010    647.1615789473686
Fantasy 2010    109.52
```

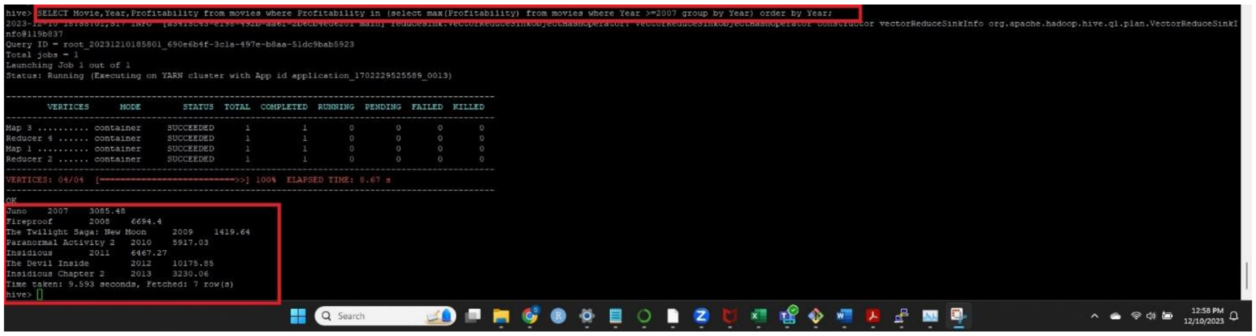
- **Command 4:** *select avg(DomesticGross),var_pop(DomesticGross),stddev_pop(DomesticGross),Year from movies where Year >=2007 group by Year;*
- **Significance:** The Query returns statistical measures of Domestic Gross such as Variance, Standard deviation, and mean.



```
hive> select avg(DomesticGross),var_pop(DomesticGross),stddev_pop(DomesticGross),Year from movies where Year >=2007 group by Year;
Query ID = root_20231210015136_98492753-e745-4cdc-8a33-d3cb03e1075c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702163480108_0013)

----------------------------------------------------------------------------------------------
        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: 5.09 s
----------------------------------------------------------------------------------------------
OK
84.64450549450547       6458.095708271951       80.36227789374783       2007
62.59062499999997       4884.411561414939       69.88856531232372       2008
74.55666666666667       8909.060539259264       94.38781986707428       2009
71.52183823529413       5584.596872356183       74.730160392951 2010
66.12791428571431       4570.222466721223       67.60342052530495       2011
62.38429487179484       8153.079279631006       90.29440336826534       2012
64.46421383647801       6541.365414319056       80.8787080406151        2013
Time taken: 5.867 seconds, Fetched: 7 row(s)
hive>
```

**Command 5**: *SELECT Movie,Year,Profitability from movies where Profitability in (select max(Profitability) from movies where Year >=2007 group by Year) order by Year;*
**Significance**: The query returns the most profitable movie of each year from 2007 to 2013.



```
hive> SELECT Movie,Year,Profitability from movies where Profitability in (select max(Profitability) from movies where Year >=2007 group by Year) order by Year;
2023-12-10 18:58:01,517 INFO  [a3413c43-e198-492b-aa61-2b6cb4ede201 main] reducesink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObjectHashOperator constructor vectorReduceSinkInfo org.apache.hadoop.hive.ql.plan.VectorReduceSinkI
nfo@119b837
Query ID = root_20231210185801_690e6b4f-3c1a-497e-b8aa-51dc9bab5923
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702229525589_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 8.67 s
----------------------------------------------------------------------------------------------
OK
Juno    2007    3085.48
Fireproof       2008    6694.4
The Twilight Saga: New Moon     2009    1419.64
Paranormal Activity 2   2010    5917.03
Insidious       2011    6467.27
The Devil Inside        2012    10175.85
Insidious Chapter 2     2013    3230.06
Time taken: 9.593 seconds, Fetched: 7 row(s)
hive>
```