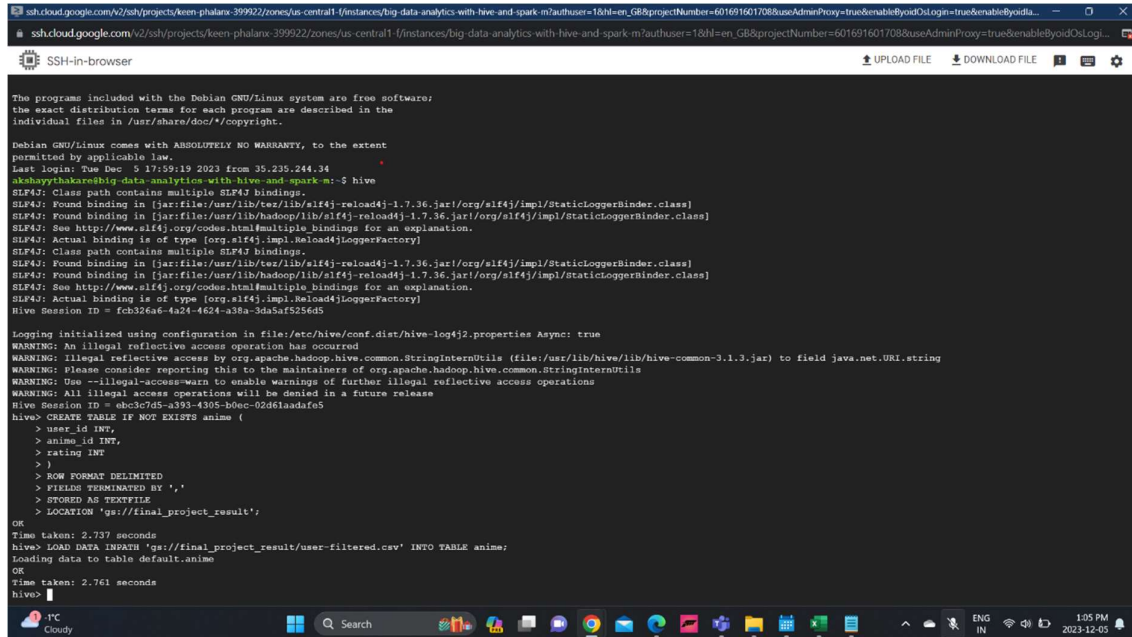


Final Project- Hive Data Loading and Cleaning

This are my SSH code for Hive

Data Loading



```
ssh.cloud.google.com/v2/ssh/projects/keen-phalanx-399922/zones/us-central1-f/instances/big-data-analytics-with-hive-and-spark-m?authuser=1&hl=en_GB&projectNumber=601691601708&useAdminProxy=true&enableByodOsLogin=true&enableByodOsLogi...
SSH-in-browser
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Tue Dec 5 17:59:19 2023 from 35.235.244.34
akashpythakare@big-data-analytics-with-hive-and-spark-m:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.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.Reload4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.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.Reload4jLoggerFactory]
Hive Session ID = fcb326a6-4a24-4624-a38a-3da5af5256d5
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hive.common.StringInternUtils (file:/usr/lib/hive/lib/hive-common-3.1.3.jar) to field java.net.URI.string
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hive.common.StringInternUtils
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Hive Session ID = ebc3c7d5-a393-4305-b0ec-02d61adafe5
hive> CREATE TABLE IF NOT EXISTS anime (
  > user_id INT,
  > anime_id INT,
  > rating INT
  > )
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE
  > LOCATION 'gs:///final_project_result';
OK
Time taken: 2.737 seconds
hive> LOAD DATA INPATH 'gs:///final_project_result/user-filtered.csv' INTO TABLE anime;
Loading data to table default.anime
OK
Time taken: 2.761 seconds
hive>
```

Displaying Data

```
hive> SELECT * FROM anime LIMIT 10;
OK
NULL      NULL      NULL
0          67        9
0          6702     7
0          242     10
0          4898    0
0          21      10
0          24      9
0          2104    0
0          4722    8
0          6098    6
Time taken: 2.574 seconds, Fetched: 10 row(s)
hive> set hive.cli.print.header
> ;
hive.cli.print.header=false
hive> set hive.cli.print.header=true;
hive> SELECT * FROM anime LIMIT 10;
OK
anime.user_id  anime.anime_id  anime.rating
NULL          NULL          NULL
0             67           9
0             6702          7
0             242          10
0             4898           0
0             21           10
0             24           9
0             2104          0
0             4722          8
0             6098          6
Time taken: 0.329 seconds, Fetched: 10 row(s)
hive>
```

Data Cleaning

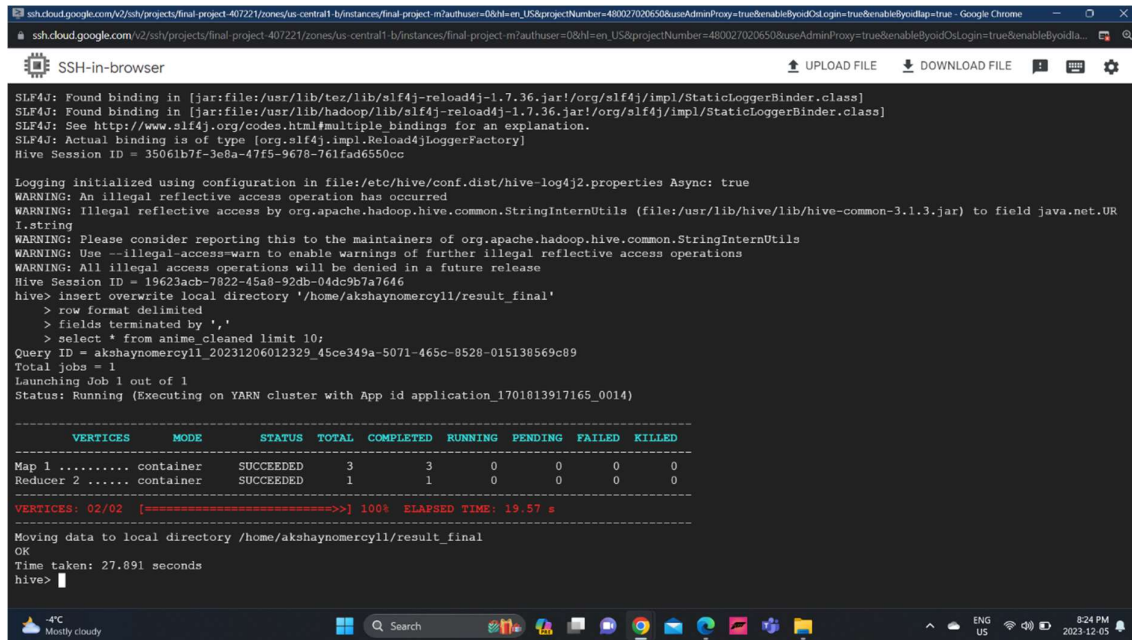
```

> LOCATION 'gs://final_project_result';
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.q1.exec.DDLTask. java.io.IOException: Error accessing Bucket final_project_result
hive> CREATE TABLE anime_cleaned (
>   user_id INT,
>   anime_id INT,
>   rating INT
> )
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> STORED AS TEXTFILE
> LOCATION 'gs://final_project_result';
OK
Time taken: 0.43 seconds
hive> INSERT OVERWRITE TABLE anime_cleaned
> SELECT
>   COALESCE(anime.user_id, 0) as anime.user_id,
>   COALESCE(anime.anime_id, 0) as anime.anime_id,
>   COALESCE(anime.rating, 0) as anime.rating
> FROM anime;
FAILED: ParseException line 3:37 missing EOF at '.' near 'anime'
hive> INSERT OVERWRITE TABLE anime_cleaned
> SELECT
>   COALESCE(anime.user_id, 0) as user_id,
>   COALESCE(anime.anime_id, 0) as anime_id,
>   COALESCE(anime.rating, 0) as rating
> FROM anime;
Query ID = akshaynomeroy11_20231206005200_37c8b0bd-5b3e-4547-a945-b00acf9b064a
Total jobs = 1
Launching Job 1 out of 1
Task session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1701813917165_0009)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1         container  RUNNING    3        0          1          2        0        0
Reducer 2     container  INITED     1        0          0          1        0        0
-----
VERTICES: 00/02 [>-----] 0% ELAPSED TIME: 63.76 s
-----
akshaynomeroy11@final-project-m1:~$

```

Moving Data to gs bucket

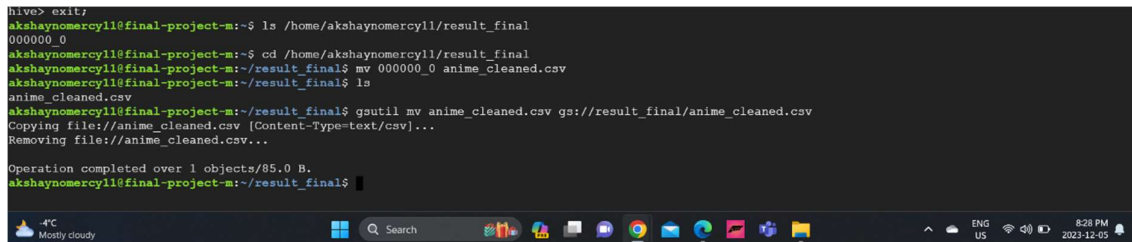


```
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Hive Session ID = 35061b7f-3e8a-47f5-9678-761fad6550cc

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hive.common.StringInternUtils (file:/usr/lib/hive/lib/hive-common-3.1.3.jar) to field java.net.UR
I.string
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hive.common.StringInternUtils
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Hive Session ID = 19623acb-7822-45a8-92db-04dc9b7a7646
hive> insert overwrite local directory '/home/akshaynmercyl1/result_final'
> row format delimited
> fields terminated by ','
> select * from anime_cleaned limit 10;
Query ID = akshaynmercyl1_20231206012329_45ce349a-5071-465c-8528-015138569c89
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1701813917165_0014)

-----
VERTICES    MODE          STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  3      3            0         0         0         0
Reducer 2 ..... container  SUCCEEDED  1      1            0         0         0         0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 19.57 s
-----
Moving data to local directory /home/akshaynmercyl1/result_final
OK
Time taken: 27.891 seconds
hive>
```

Saving data as csv file



```
hive> exit;
akshaynmercyl1@final-project-m1:~$ ls /home/akshaynmercyl1/result_final
0000000_0
akshaynmercyl1@final-project-m1:~$ cd /home/akshaynmercyl1/result_final
akshaynmercyl1@final-project-m1:~/result_final$ mv 0000000_0 anime_cleaned.csv
akshaynmercyl1@final-project-m1:~/result_final$ ls
anime_cleaned.csv
akshaynmercyl1@final-project-m1:~/result_final$ gutil mv anime_cleaned.csv gs://result_final/anime_cleaned.csv
Copying file://anime_cleaned.csv [Content-Type=text/csv]...
Removing file://anime_cleaned.csv...

Operation completed over 1 objects/85.0 B.
akshaynmercyl1@final-project-m1:~/result_final$
```