DATA CLEANING USING HIVE

This document covers all the steps taken to import the CSV to hive, clean, and export the CSV.

Step 1: Using DataProc create an Apache Hadoop cluster on Google Cloud Platform and Google Cloud Bucket to add the CSV dataset

Step 2: Launch the SSH session and import your CSV file using the GSUTIL URL to your local instance

SSH-in-browser

```
Linux strikereport-m 5.10.0-26-cloud-amd64 #1 SMP Debian 5.10.197-1 (2023-09-29) x86_64

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.
imgauravmehta@strikereport-m:~$ pwd
/home/imgauravmehta
imgauravmehta@strikereport-m:~$ mkdir strike_report
imgauravmehta@strikereport-m:~$ ls -l
total 4
drwxr-xr-x 2 imgauravmehta imgauravmehta 4096 Nov 30 19:38 strike_report
imgauravmehta@strikereport-m:~$ cd strike_report
imgauravmehta@strikereport-m:~/strike_report$ gsutil cp gs://strike_report/STRIKE_REPORTS.csv /^Cth/to/destination
imgauravmehta@strikereport-m:~/strike_report$ ^C
imgauravmehta@strikereport-m:~/strike_report$ gsutil cp gs://strike_report/STRIKE_REPORTS.csv /home/imgauravmehta/strike_report/.
Copying gs://strike_report/STRIKE_REPORTS.csv...
| [1 files][185.9 MiB/185.9 MiB]
Operation completed over 1 objects/185.9 MiB.
imgauravmehta@strikereport-m:~/strike_report$ ls -l
total 190400
-rw-r--r-- 1 imgauravmehta imgauravmehta 194965534 Nov 30 19:49 STRIKE_REPORTS.csv
imgauravmehta@strikereport-m:~/strike_report$
```

Step 3: Once the file is copied to the local instance copy the same file to the HADOOP file system

```
imgauravmehta@strikereport-m:~/strike_report$ hadoop fs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-11-30 19:33 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-11-30 19:33 /user
drwxrwxrwt   - hdfs hadoop          0 2023-11-30 19:33 /var
imgauravmehta@strikereport-m:~/strike_report$ hadoop fs -mkdir /user/strikereport
imgauravmehta@strikereport-m:~/strike_report$ hadoop fs -ls /user
Found 12 items
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/dataproc
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/hbase
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/hdfs
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/hive
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/mapred
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/pig
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/solr
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/spark
drwxr-xr-x   - imgauravmehta hadoop          0 2023-11-30 19:51 /user/strikereport
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/yarn
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/zeppelin
drwxrwxrwt   - hdfs          hadoop          0 2023-11-30 19:33 /user/zookeeper
imgauravmehta@strikereport-m:~/strike_report$
```

```
imgauravmehta@strikereport-m:~/strike_report$ hadoop fs -copyFromLocal /home/imgauravmehta/strike_report/STRIKE_REPORTS.csv /user/strikereport/.
imgauravmehta@strikereport-m:~/strike_report$ hadoop fs -ls /user/strikereport
Found 1 items
-rw-r--r--   1 imgauravmehta hadoop  194965534 2023-11-30 19:54 /user/strikereport/STRIKE_REPORTS.csv
imgauravmehta@strikereport-m:~/strike_report$
```

Step 4: Once the file is saved to our HADOOP file system, Initiate HIVE

```
imgauravmehta@strikereport-m:~/strike_report$ 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 = 1ec1081c-f602-4ae9-8ba7-2726cdf63cdd

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 = e1754683-f018-4aa4-93c4-7077e57bc9f3
hive>
```

Step 5: Create a new database in HIVE as air_accidents and use the same database to create the table

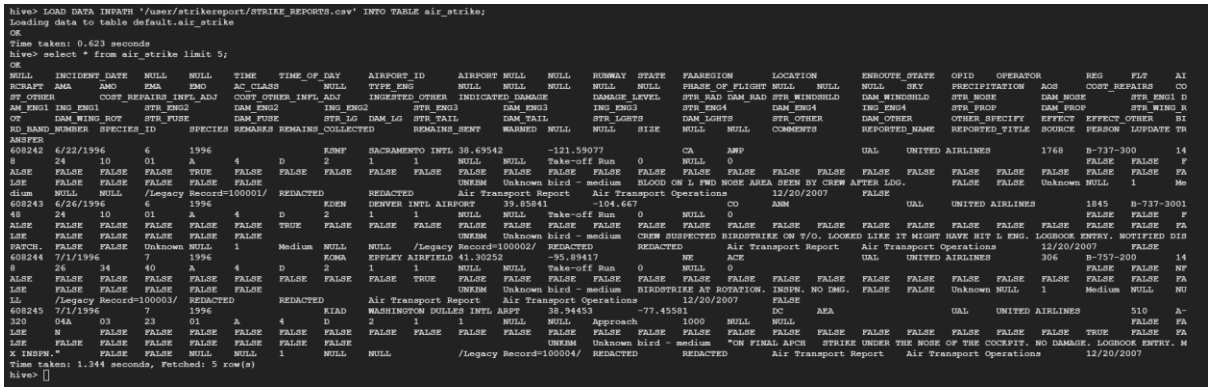
```
hive> create database air_accidents;
OK
Time taken: 0.534 seconds
```

```
hive> use air_accidents;
OK
Time taken: 0.06 seconds
hive>
```

Step 6: Create an EXTERNAL TABLE in HIVE with the columns and their data types along with the delimiter passed as {,} to ensure the CSV file standards are maintained.

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS air_strike_data (
    >   INDEX_NR INT,
    >   INCIDENT_DATE STRING,
    >   INCIDENT_MONTH INT,
    >   INCIDENT_YEAR INT, `TIME` STRING,
    >   TIME_OF_DAY STRING,
    >   AIRPORT_ID STRING,
    >   AIRPORT STRING,
    >   LATITUDE DOUBLE,
    >   LONGITUDE DOUBLE,
    >   RUNWAY STRING,
    >   STATE STRING,
    >   FAAREGION STRING,
    >   LOCATION STRING,
    >   ENROUTE_STATE STRING,
    >   OPID STRING,
    >   OPERATOR STRING,
    >   REG STRING,
    >   FLT STRING,
    >   AIRCRAFT STRING,
    >   AMA STRING,
    >   AMO STRING,
    >   EMA STRING,
    >   EMO STRING,
    >   AC_CLASS STRING,
    >   AC_MASS INT,
    >   TYPE_ENG STRING,
    >   NUM_ENGS INT,
    >   ENG_1_POS INT,
    >   ENG_2_POS INT,
    >   ENG_3_POS INT,
    >   ENG_4_POS INT,
    >   PHASE_OF_FLIGHT STRING,
    >   HEIGHT INT,
    >   SPEED INT,
    >   DISTANCE INT,
    >   SKY STRING,
    >   PRECIPITATION STRING,
    >   AOS STRING,
    >   COST_REPAIRS INT,
    >   COST_OTHER INT,
    >   COST_REPAIRS_INFL_ADJ INT,
    >   COST_OTHER_INFL_ADJ INT,
    >   INGESTED_OTHER BOOLEAN,
    >   INDICATED_DAMAGE BOOLEAN,
    >   DAMAGE_LEVEL STRING,
    >   STR_RAD BOOLEAN,
    >   STR_ENG4 BOOLEAN,
    >   DAM_ENG4 BOOLEAN,
    >   ING_ENG4 BOOLEAN,
    >   STR_PROP BOOLEAN,
    >   DAM_PROP BOOLEAN,
    >   STR_WING_ROT BOOLEAN,
    >   DAM_WING_ROT BOOLEAN,
    >   STR_FUSE BOOLEAN,
    >   DAM_FUSE BOOLEAN,
    >   STR_LG BOOLEAN,
    >   DAM_LG BOOLEAN,
    >   STR_TAIL BOOLEAN,
    >   DAM_TAIL BOOLEAN,
    >   STR_LGHTS BOOLEAN,
    >   DAM_LGHTS BOOLEAN,
    >   STR_OTHER BOOLEAN,
    >   DAM_OTHER BOOLEAN,
    >   OTHER_SPECIFY STRING,
    >   EFFECT STRING,
    >   EFFECT_OTHER STRING,
    >   BIRD_BAND_NUMBER STRING,
    >   SPECIES_ID STRING,
    >   SPECIES STRING,
    >   REMARKS STRING,
    >   REMAINS_COLLECTED BOOLEAN,
    >   REMAINS_SENT BOOLEAN,
    >   WARNED BOOLEAN,
    >   NUM_SEEN INT,
    >   NUM_STRUCK INT,
    >   SIZE STRING,
    >   NR_INJURIES INT,
    >   NR_FATALITIES INT,
    >   COMMENTS STRING,
    >   REPORTED_NAME STRING,
    >   REPORTED_TITLE STRING,
    >   SOURCE STRING,
    >   PERSON STRING,
    >   LUPDATE STRING,
    >   TRANSFER STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/strikereport/STRIKE_REPORTS';
OK
Time taken: 0.564 seconds
hive>
```

Step 7: Verify the table (air_strike) created



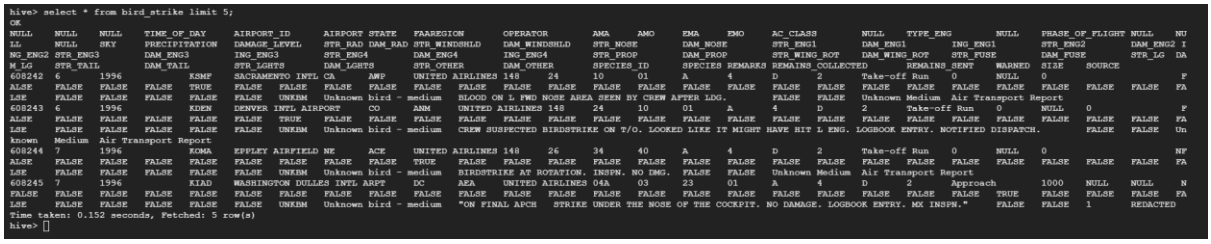
```
hive> LOAD DATA INPATH '/user/strikereport/STRIKE_REPORTS.csv' INTO TABLE air_strike;
Loading data to table default.air_strike
OK
Time taken: 0.623 seconds
hive> select * from air_strike limit 5;
OK
NULL    INCIDENT_DATE   NULL    NULL    TIME    TIME_OF_DAY     AIRPORT_ID      AIRPORT NULL    NULL    RUNWAY  STATE   FAAREGION       LOCATION        ENROUTE_STATE   OPID    OPERATOR        REG     FLT     AI
RCRAFT  AMA     AMO     EMA     EMO     AC_CLASS        NULL    TYPE_ENG        NULL    NULL    NULL    NULL    NULL    PHASE_OF_FLIGHT NULL    NULL    NULL    SKY     PRECIPITATION   AOS     COST_REPAIRS    CO
ST_OTHER        COST_REPAIRS_INFL_ADJ   COST_OTHER_INFL_ADJ     INGESTED_OTHER  INDICATED_DAMAGE        DAMAGE_LEVEL    STR_RAD DAM_RAD STR_WINDSHLD    DAM_WINDSHLD    STR_NOSE        DAM_NOSE        STR_ENG1 D
AM_ENG1 ING_ENG1        STR_ENG2        DAM_ENG2        ING_ENG2        STR_ENG3        DAM_ENG3        ING_ENG3        STR_ENG4        DAM_ENG4        ING_ENG4        STR_PROP        DAM_PROP        STR_WING_R
OT      DAM_WING_ROT    STR_FUSE        DAM_FUSE        STR_LG  DAM_LG  STR_TAIL        DAM_TAIL        STR_LGHTS       DAM_LGHTS       STR_OTHER       DAM_OTHER       OTHER_SPECIFY   EFFECT  EFFECT_OTHER    BI
RD_BAND_NUMBER  SPECIES_ID      SPECIES_REMARKS REMAINS_COLLECTED       REMAINS_SENT    WARNED  NULL    SIZE    NULL    NULL    COMMENTS        REPORTED_NAME   REPORTED_TITLE  SOURCE  PERSON  LUPDATE TR
ANSFER
608242  6/22/1996       6       1996            KSMF    SACRAMENTO INTL 38.69542        -121.59077              CA      AWP                     UAL     UNITED AIRLINES         1768    B-737-300       14
8       24      10      01      A               4       D       2       1       1       NULL    NULL    Take-off Run    0       NULL    0                                               FALSE   FALSE   F
ALSE    FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE                                   UNKBM   Unknown bird - medium   BLOOD ON L FWD NOSE AREA SEEN BY CREW AFTER LDG.        FALSE   FALSE   Unknown NULL    1       Me
dium    NULL    NULL    /Legacy Record=100001/  REDACTED                REDACTED                Air Transport Report    Air Transport Operations        12/20/2007      FALSE
608243  6/26/1996       6       1996            KDEN    DENVER INTL AIRPORT     39.85841        -104.667                CO      ANM                     UAL     UNITED AIRLINES         1845    B-737-3001
48      24      10      01      A               4       D       2       1       1       NULL    NULL    Take-off Run    0       NULL    0                                               FALSE   FALSE   F
ALSE    FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE                                   UNKBM   Unknown bird - medium   CREW SUSPECTED BIRDSTRIKE ON T/O. LOOKED LIKE IT MIGHT HAVE HIT L ENG. LOGBOOK ENTRY. NOTIFIED DIS
PATCH.  FALSE   FALSE   Unknown NULL    1       Medium  NULL    NULL    /Legacy Record=100002/  REDACTED                REDACTED                Air Transport Report    Air Transport Operations        12/20/2007      FALSE
608244  7/1/1996        7       1996            KOMA    EPPLEY AIRFIELD 41.30252        -95.89417               NE      ACE                     UAL     UNITED AIRLINES         306     B-757-200       14
8       26      34      40      A               4       D       2       1       1       NULL    NULL    Take-off Run    0       NULL    0                                               FALSE   FALSE   NF
ALSE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE                                   UNKBM   Unknown bird - medium   BIRDSTRIKE AT ROTATION. INSPN. NO DMG.  FALSE   FALSE   Unknown NULL    1       Medium  NULL    NU
LL      /Legacy Record=100003/  REDACTED                REDACTED        Air Transport Report    Air Transport Operations        12/20/2007      FALSE
608245  7/1/1996        7       1996            KIAD    WASHINGTON DULLES INTL ARPT     38.94453        -77.45581               DC      AEA                     UAL     UNITED AIRLINES         510     A-
320     04A     03      23      01      A       4       D       2       1       1       NULL    NULL    Approach        1000    NULL    NULL                                            FALSE   FALSE   FA
LSE     N       FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE                   UNKBM   Unknown bird - medium   "ON FINAL APCH  STRIKE UNDER THE NOSE OF THE COCKPIT. NO DAMAGE. LOGBOOK ENTRY. M
X INSPN."       FALSE   FALSE   NULL    NULL    1       NULL    NULL            /Legacy Record=100004/  REDACTED        REDACTED        Air Transport Report    Air Transport Operations        12/20/2007
Time taken: 1.344 seconds, Fetched: 5 row(s)
hive>
```

Step 8: Data Cleaning

1. DROPPING COLUMNS

- Dropping the columns that will not be required at all for analysis purposes especially with maximum NULL values.
  Columns Dropped: Time, Runway, Location, Enroute_State, OPID (Operator Column already present), AOS, COST_REPAIRS, COST_OTHER, COST_REPAIRS_INFL_ADJ, COST_OTHER_INFL_ADJ, OTHER_SPECIFY, EFFECT_OTHER, BIRD_BAND_NUMBER, NUM_SEEN, NUM_STRUCK, NR_INJURIES, NR_FATALITIES, COMMENTS, REPORTED_NAME, REPORTED_TITLE, SOURCE, PERSON, LUPDATE, REASON
- Dropping LATITUDE and LONGITUDES as we have AIRPORT name and AIRPORT ID.
- Removing ENGINE_POSITIONS as well as this variable does not help in analysis purposes.
- INCIDENT_DATE is yet another column that can be dropped as the values are inconsistent and INCIDENT_MONTH and INCIDENT_YEAR can help us for analysis purposes.

Once the columns to be dropped are verified, create a new table as bird_strike from air_strike with the new table wanted for further analysis

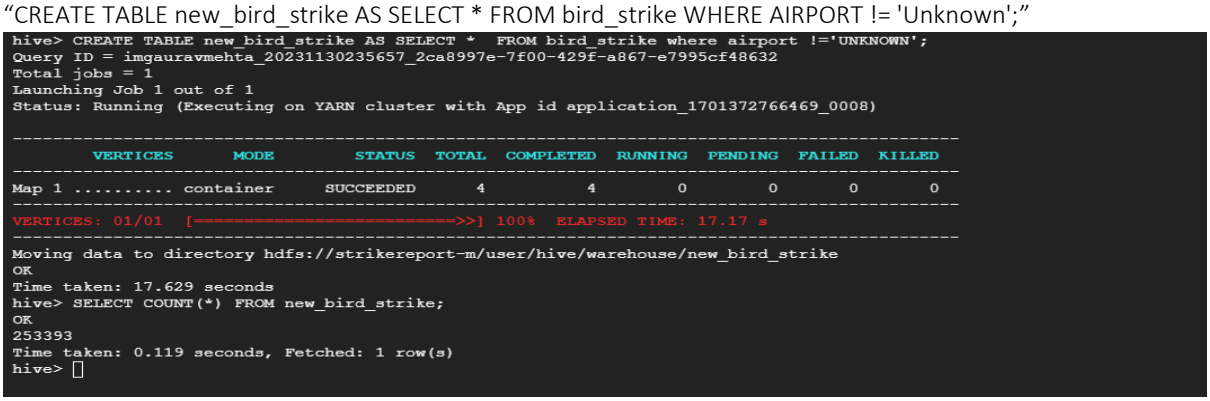


```
hive> select * from bird_strike limit 5;
OK
NULL    NULL    NULL    TIME_OF_DAY     AIRPORT_ID      AIRPORT STATE   FAAREGION       OPERATOR        AMA     AMO     EMA     EMO     AC_CLASS        NULL    TYPE_ENG        NULL    PHASE_OF_FLIGHT NULL    NU
LL      NULL    SKY     PRECIPITATION   DAMAGE_LEVEL    STR_RAD DAM_RAD STR_WINDSHLD    DAM_WINDSHLD    STR_NOSE        DAM_NOSE        STR_ENG1        DAM_ENG1        ING_ENG1        STR_ENG2        DAM_ENG2 I
NG_ENG2 STR_ENG3        DAM_ENG3        ING_ENG3        STR_ENG4        DAM_ENG4        ING_ENG4        STR_PROP        DAM_PROP        STR_WING_ROT    DAM_WING_ROT    STR_FUSE        DAM_FUSE        STR_LG  DA
M_LG    STR_TAIL        DAM_TAIL        STR_LGHTS       DAM_LGHTS       STR_OTHER       DAM_OTHER       SPECIES_ID      SPECIES_REMARKS REMAINS_COLLECTED       REMAINS_SENT    WARNED  SIZE    SOURCE
608242  6       1996            KSMF    SACRAMENTO INTL CA      AWP     UNITED AIRLINES 148     24      10      01      A       4       D       2       Take-off Run    0       NULL    0               F
ALSE    FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   UNKBM   Unknown bird - medium   BLOOD ON L FWD NOSE AREA SEEN BY CREW AFTER LDG.        FALSE   FALSE   Unknown Medium  Air Transport Report
608243  6       1996            KDEN    DENVER INTL AIRPORT     CO      ANM     UNITED AIRLINES 148     24      10      01      A       4       D       2       Take-off Run    0       NULL    0               F
ALSE    FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   UNKBM   Unknown bird - medium   CREW SUSPECTED BIRDSTRIKE ON T/O. LOOKED LIKE IT MIGHT HAVE HIT L ENG. LOGBOOK ENTRY. NOTIFIED DISPATCH.        FALSE   FALSE   Un
known   Medium  Air Transport Report
608244  7       1996            KOMA    EPPLEY AIRFIELD NE      ACE     UNITED AIRLINES 148     26      34      40      A       4       D       2       Take-off Run    0       NULL    0               NF
ALSE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE   UNKBM   Unknown bird - medium   BIRDSTRIKE AT ROTATION. INSPN. NO DMG.  FALSE   FALSE   Unknown Medium  Air Transport Report
608245  7       1996            KIAD    WASHINGTON DULLES INTL ARPT     DC      AEA     UNITED AIRLINES 04A     03      23      01      A       4       D       2       Approach        1000    NULL    NULL    N
FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   TRUE    FALSE   FALSE   FALSE   FALSE   FA
LSE     FALSE   FALSE   FALSE   FALSE   FALSE   UNKBM   Unknown bird - medium   "ON FINAL APCH  STRIKE UNDER THE NOSE OF THE COCKPIT. NO DAMAGE. LOGBOOK ENTRY. MX INSPN."      FALSE   FALSE   1       REDACTED
Time taken: 0.152 seconds, Fetched: 5 row(s)
hive>
```

Data now looks much more sorted and cleaned.

2. Dealing with NULL values and making values consistent

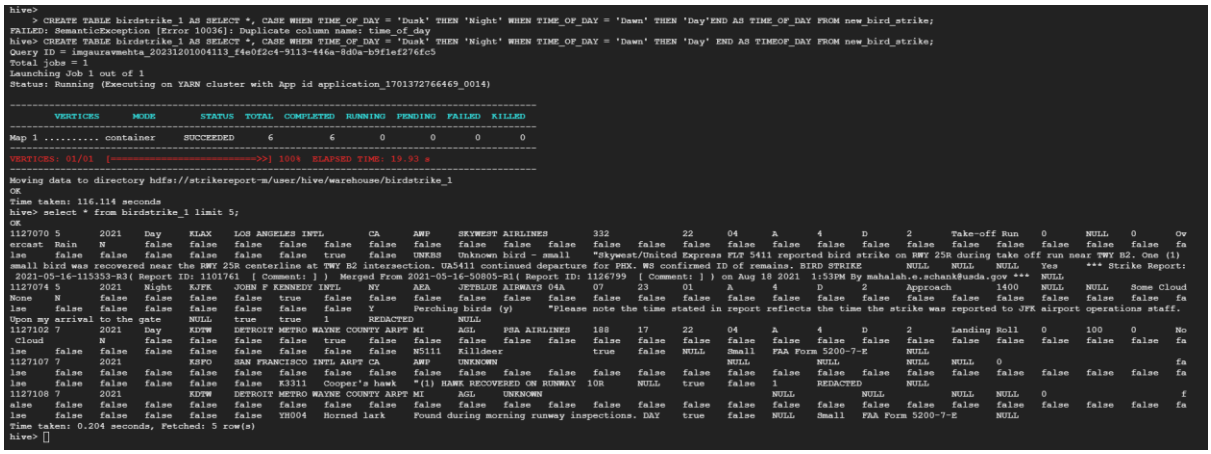
- Drop rows with airport as unknown or missing and create a new table as new_bird_strike using the query "CREATE TABLE new_bird_strike AS SELECT * FROM bird_strike WHERE AIRPORT != 'Unknown';"



```
hive> CREATE TABLE new_bird_strike AS SELECT *  FROM bird_strike where airport !='UNKNOWN';
Query ID = imgauravmehta_20231130235657_2ca8997e-7f00-429f-a867-e7995cf48632
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1701372766469_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 17.17 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://strikereport-m/user/hive/warehouse/new_bird_strike
OK
Time taken: 17.629 seconds
hive> SELECT COUNT(*) FROM new_bird_strike;
OK
253393
Time taken: 0.119 seconds, Fetched: 1 row(s)
hive>
```

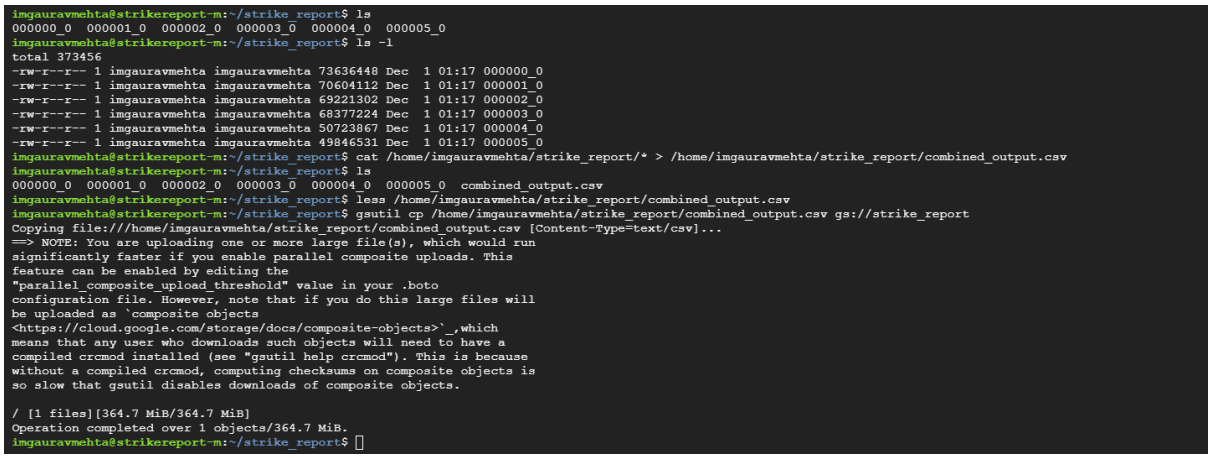
- TIME_OF_DAY values with DUSK will be replaced with NIGHT and DAWN with DAW to remove data inconsistencies (Table name: birdstrike_1)



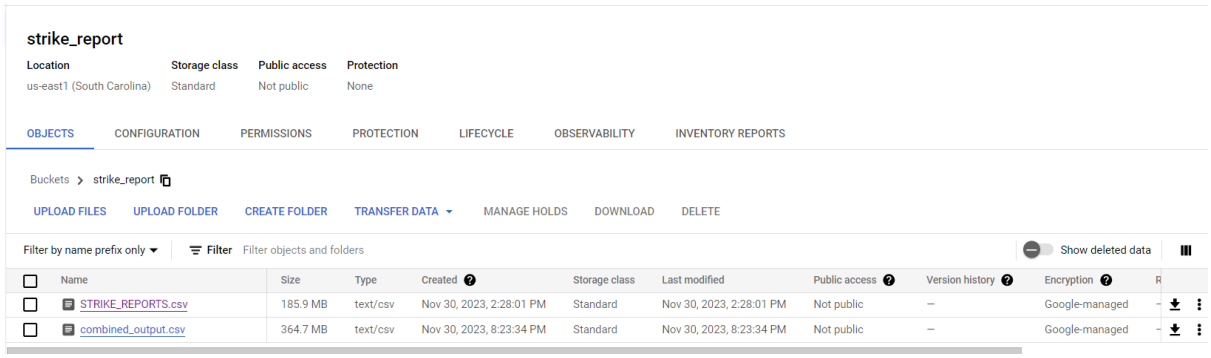
Step 9: The table birdstrike_1 can now be exported to google bucket for further analysis. While exporting the file from HIVE to local instance, the file has multiple outputs.

The multiple output files can be combined into 1 single file using the –cat command, the file can be now named combined_output.csv.

Using the GSUTIL COPY command, export combined_output.csv to Cloud Storage Bucket.



Step 10: Verify the file in Google Bucket (strike_report)



The output file (combined_output.csv) can now be downloaded and loaded into Jupyter Notebook with Spark to perform further cleaning and analysis.