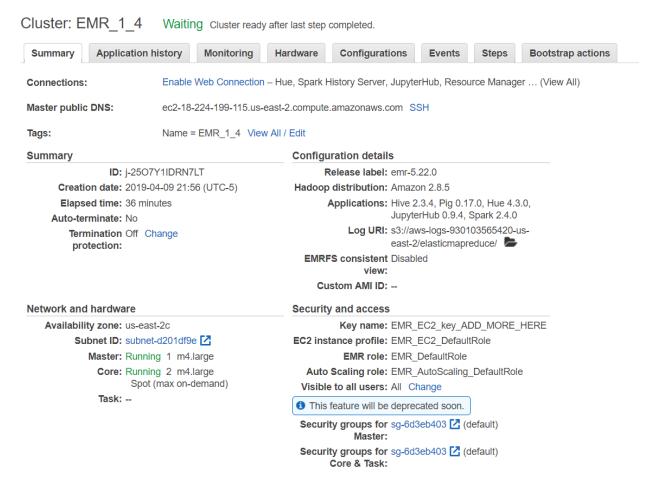
Project Lab - Load Full Data-Set to Hive

Deliverables

Screen shot of provisioned EMR cluster from AWS EMR dashboard:



Screen shot/shots from the EMR Master terminal, clearly showing all executed commands:

1) Download full dataset and save in Windows:

2) Pre-process data set to remove special characters and reformat to tab delimited format

The .csv file that was downloaded in the above step pre-processed to format it to tab delimited format using python3 as given below:

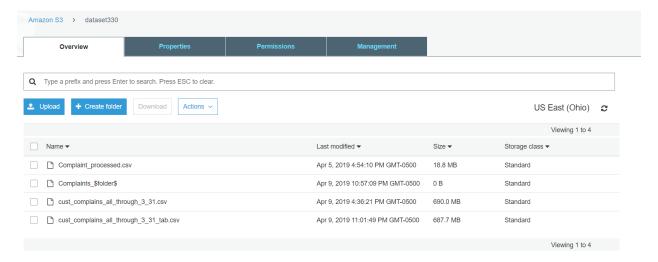
```
In [1]: import csv
In [2]: input file name = "cust complains all through 3 31.csv"
            output file name = "cust complains all through 3 31 tab.csv"
In [3]: def remove_n_r(column):
                 column = column.replace('\n','')
column = column.replace('\r','')
                 column = column.replace('\t',
                 return column
In [4]: in_txt = csv.reader(open(input_file_name, encoding="utf8"), delimiter = ',')
In [5]: out_csv = open(output_file_name, 'w', newline='\n', encoding="utf8")
csv_writer = csv.writer(out_csv, delimiter = '\t', quotechar='|', quoting=csv.QUOTE_MINIMAL)
            line count = 0
            for row in in_txt:
                 if line_count == 0:
                      print('Column names are:', row)
                       line_count += 1
                      new_row = []
                      for col in row:
                           new row.append(remove n r(col))
                       csv_writer.writerow(new_row)
                      line count += 1
            out_csv.close()
            Column names are: ['Date received', 'Product', 'Sub-product', 'Issue', 'Sub-issue', 'Consumer complaint narrative', 'Company pu blic response', 'Company', 'State', 'ZIP code', 'Tags', 'Consumer consent provided?', 'Submitted via', 'Date sent to company', 'Company response to consumer', 'Timely response?', 'Consumer disputed?', 'Complaint ID']
In [6]: print('Processed',line_count,' lines.')
            Processed 1256553 lines.
```

From the above screenshot it can it said that 1256553 records were processed. Thus, the output file "cust_complaints_all_through_3_31_tab.csv" consists of 1256552 records and a header. This output file is then uploaded in aws S3 in location: S3://dataset330/

3) Create database customer_complaints and External table cust_complaints_all_raw as below:

```
hive> create database customer_complaints;
Time taken: 0.074 seconds
hive> CREATE EXTERNAL TABLE customer complaints.cust complaints all raw
   > (date_received
                           string,
    > product
                           string,
    > sub_product
                           string,
    > issue
                           string,
    > sub_issue
                           string,
    > consumer_complaint_narrative string,
    > company_public_response string,
                           string,
    > company
    > state
                           string,
                           string,
    > zip_code
    > tags
                           string,
    > consumer_consent_provided string,
    > submitted_via
                           strina.
    > date_sent_to_company string,
    > company_response
                           string,
    > timely_response
                           string,
    > consumer_disputed
> complaint_id
                           string,
                           string)
    > ROW FORMAT DELIMITED
     FIELDS TERMINATED BY '\t'
      STORED AS
    > INPUTFORMAT
        "com.amazonaws.emr.s3 select.hive.S3 Selectable TextInputFormat" \\
        'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
      LOCATION 's3://dataset330/Complaints'
    > TBLPROPERTIES (
        "s3select.format" = "csv",
        "s3select.headerInfo" = "ignore"
Time taken: 0.631 seconds
```

The data that will be loaded in this table must be tab delimited. Also, the data that will be loaded will be saved in location "S3://dataset330/Complaints" as shown in the screenshot below:



4) Load data in the External Table:

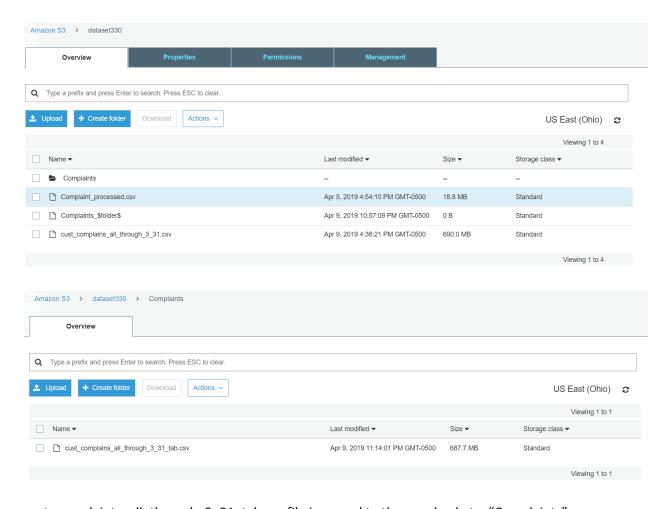
The below code was executed to load data in the External Table.

```
hive> LOAD DATA inpath 's3://dataset330/cust_complains_all_through_3_31_tab.csv'
> INTO TABLE customer_complaints.cust_complaints_all_raw;
Loading data to table customer_complaints.cust_complaints_all_raw
OK
Time taken: 5.115 seconds
```

This is a one-time process and this step is not required to be executed every time this table is created as location is mentioned while creating this external table.

As cluster is terminated, all the tables and databases created get lost. So, whenever a new cluster is created, step 2 is required to be executed every time. As data is now already stored in the location mentioned while creating the external table, data is automatically loaded in the table.

The status of AWS S3://dataset330/ looks like below:



cust_complaints_all_through_3_31_tab.csv file is moved to the new bucket – "Complaints".

5) Create table to store data in PARQUET format

```
hive> CREATE TABLE customer complaints.cust complaints all
    > (date received
                         string,
   > product
                          string,
   > sub_product
                        string,
   > issue
                          string,
   > sub issue
                          string,
   > consumer complaint narrative string,
   > company public response string,
                          string,
   > company
   > state
                          string,
   > zip code
                          string,
   > tags
                          string,
   > consumer consent provided string,
   > submitted via
                         string,
   > date sent to company string,
   > company response
                        string,
   > timely response
                         string,
   > consumer disputed
                         string,
   > complaint id
                          string)
   > ROW FORMAT DELIMITED
   > FIELDS TERMINATED BY ','
   > STORED AS PARQUET
    > TBLPROPERTIES ("parquet.compression"="SNAPPY");
0K
Time taken: 0.111 seconds
```

Create a table "cust_complaints_all" and insert data in it from table cust_complaints_all_raw as shown below:

Run basic queries:

Data has been loaded into table cust_complaints_all. Let's run some basic queries:

(i) Display first 5 records:

```
from customer complaints.cust complaints all limit 5;
hive> select *
                                                                          Loan servicing, payments, escrow account 0 02/13/2015 Closed with explanation Ye
02/13/2015 Mortgage
CWEN LOAN SERVICING LLC VA
                                     Conventional fixed mortgage
                                     239XX
                                                       N/A
                                                                 Web
        No
                  1239425
                                                                                             Attempted to collect wrong
02/13/2015
                  Debt collection Medical False statements or representation
                           Global Financial Services Group SC
amount
                                                                                                               02/21/2015
                                                                          29073
                                                                                             N/A
                                                                                                     Web
Closed with monetary relief
                                    Yes
                                            No
                                                        1239104
02/13/2015 Debt collection I do
egal Prevention Services, LLC. NY
in colief Yes Yes 12400
                  Debt collection I do not know
                                                       Cont'd attempts collect debt not owed
                                                                                                      Debt is not mine L
                                                                                                      Closed with non-mo
                                              109XX
                                                                N/A
                                                                          Web
                                                                                   02/13/2015
                                    1240053
                  Debt collection Other (i.e. phone, health club, etc.) Cont'd atte
not mine AllianceOne Recievables Management
02/13/2015
                                                                                   Cont'd attempts collect debt not o
         Debt is not mine
                                                                                                               98021
wed
                                                                                                      WΔ
                  02/24/2015
         Web
                                     Closed with explanation Yes
                                                                         No
                                                                                   1240181
02/13/2015
                  Debt collection Credit card
                                                       Cont'd attempts collect debt not owed
                                                                                                      Debt resulted from
                                     ENCORE CAPITAL GROUP INC.
 identity theft
                                                                                                               Web
                                                                                                                        02
                                                                          TI
                                                                                   608XX
                                                                                                      N/A
                                                                          1239003
                  Closed with non-monetary relief Yes
/13/2015
                                                                 No
Time taken: 0.15 seconds, Fetched: 5 row(s)
hive>
```

(ii) Display total no of records in the table:

```
hive> select count(*) as Total_Complaints from customer_complaints.cust_complaints_all;
OK
1256552
Time taken: 0.436 seconds, Fetched: 1 row(s)
hive> ■
```

(iii) Display top 10 companies against which most complaints have been filed along with count:

```
count(*) as Count from customer_complaints.cust_complaints_all group by company orde
r by Count desc limit 10;
Query ID = hadoop_20190410043331_f4047638-1c24-497f-8941-2b5ea915e08b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application 1554865673450 0004)
        VERTICES
                      MODE
                                   STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container
                                SUCCEEDED
                                                                    0
                                                                              0
                                                                                      0
                                                                                               0
Reducer 2 ..... container
                                SUCCEEDED
                                                                    0
                                                                              0
                                                                                               0
Reducer 3 ..... container
                                SUCCEEDED
                                                                    0
                                                                                      Θ
                                                                                               0
VERTICES: 03/03 [===
                                             =>>] 100% ELAPSED TIME: 17.65 s
0K
EQUIFAX, INC. 112203
Experian Information Solutions Inc.
                                          100584
TRANSUNION INTERMEDIATE HOLDINGS, INC.
                                         93334
BANK OF AMERICA, NATIONAL ASSOCIATION
                                         81346
WELLS FARGO & COMPANY
JPMORGAN CHASE & CO.
CITIBANK, N.A. 48324
                         59332
CAPITAL ONE FINANCIAL CORPORATION
Navient Solutions, LLC. 28923
OCWEN LOAN SERVICING LLC
                                 27642
Time taken: 18.25 seconds, Fetched: 10 row(s)
hive>
```

(iv) Display top 10 product and sub product that have been complained about along with count:

```
hive> select product, sub_product, count(*) as Count from customer_complaints.cust_complaints_all group by product, sub_product order by Count desc limit 10;
Query ID = hadoop_20190410043753_48493c06-395e-4b0b-9ed2-07e6e7665ba2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1554865673450_0004)
       VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
                                                                                    0
Map 1 ..... container SUCCEEDED
                                                             0
Reducer 2 ..... container
Reducer 3 ..... container
                            SUCCEEDED
                                                             0
                                                                     0
                                                                             0
                                                                                    0
                         SUCCEEDED
                                           1
                                                             0
                                                                     0
                                                                             0
                                                                                    0
.....
Credit reporting, credit repair services, or other personal consumer reports
                                                                          Credit reporting
Credit reporting
Credit card
                      89190
Mortgage
              Other mortgage 86636
              Conventional fixed mortgage
Mortgage
Bank account or service Checking account
Debt collection I do not know 49134
                                            59045
Debt collection Other (i.e. phone, health club, etc.)
                                                  44555
Credit card or prepaid card
                             General-purpose credit card or charge card
                                                                          35071
Mortgage
              FHA mortgage
                             31334
Time taken: 17.623 seconds, Fetched: 10 row(s)
hive>
```

(v) Display top 10 issue and sub issue that have been reported along with count: