HBase

Overview:

Download amazon video game review data from the link https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz (Links to an external site.) . After extracting you will get a ".tsv" file where the columns are 'tab' separated. The description of the columns can be found in the link https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt (Links to an external site.) . For this assignment, you must run different queries on **HBase** using this dataset.

Data File: amazon_reviews_us_Video_Games_v1_00.tsv

- Removed the first line of the given tsv file (in the question) which contains the column names
- Added only those columns which were required in the question

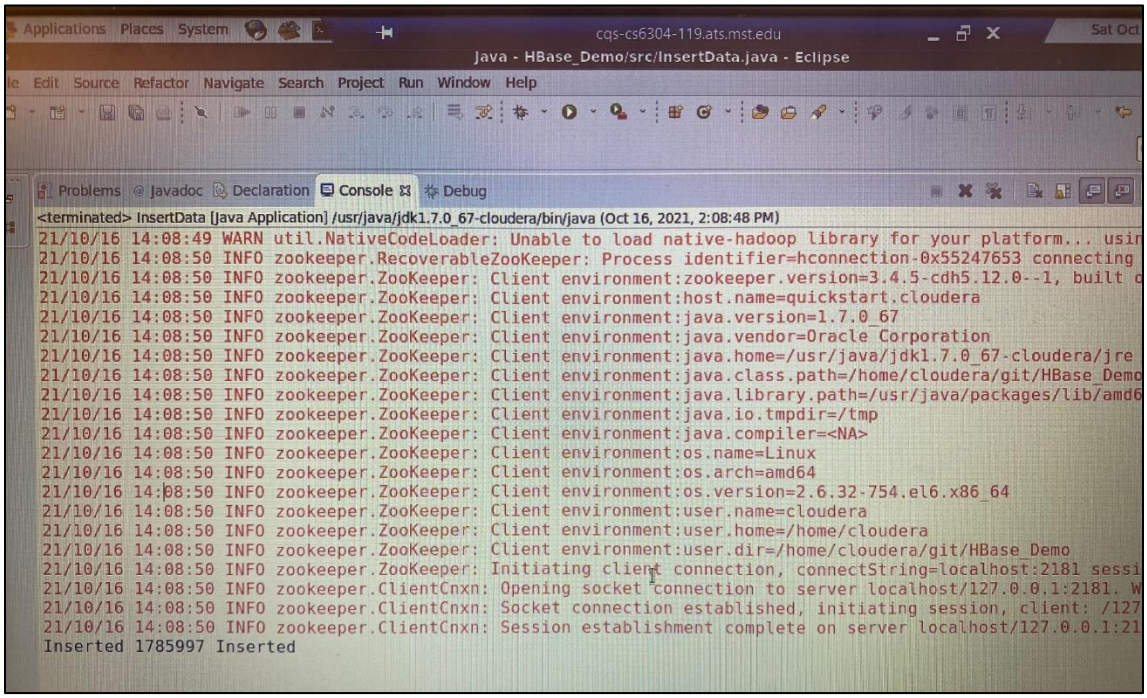
Part 1:

Download and Insert this dataset to a HBase table named "GameTable":

- Run this file to insert the data from .tsv file into the table

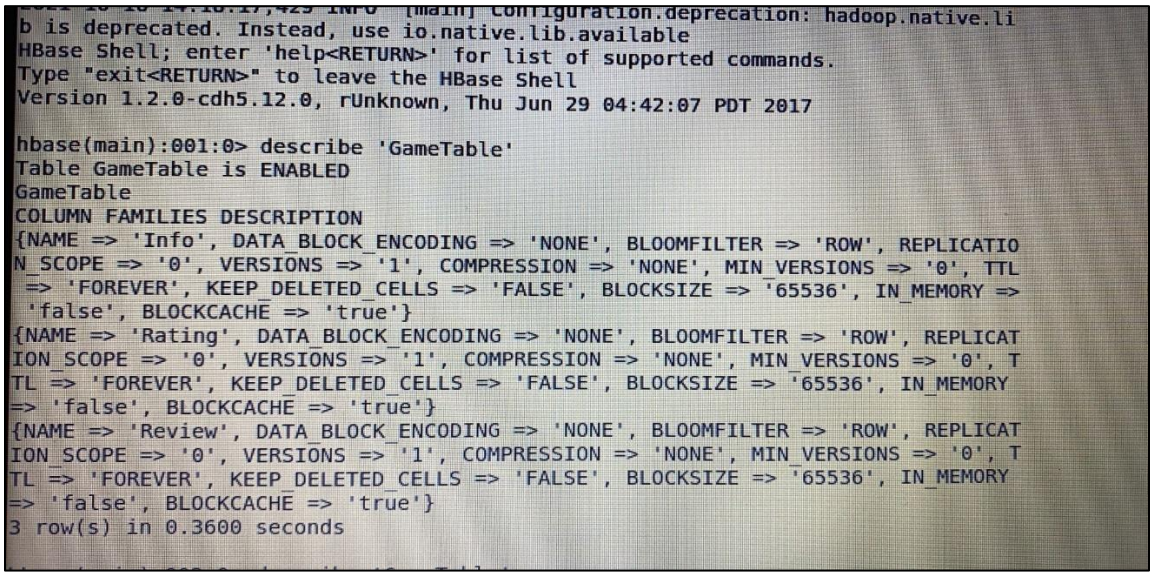
InsertData.java

Data Inserted: 1785997 rows



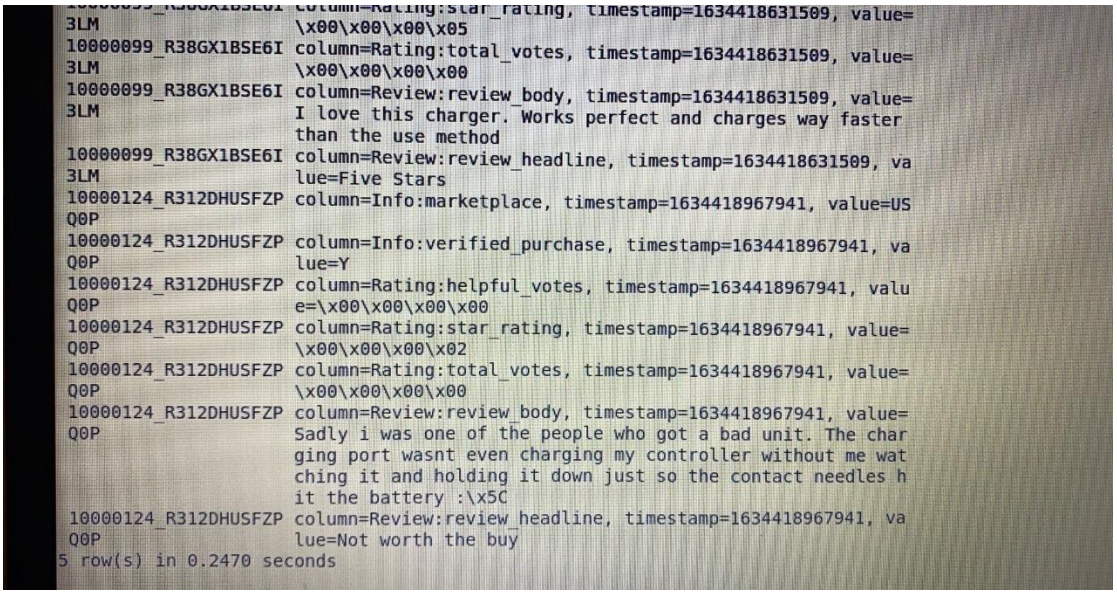
Command: Describe

hbase(main):001:0> describe 'GameTable'



Command: Scan, Limit

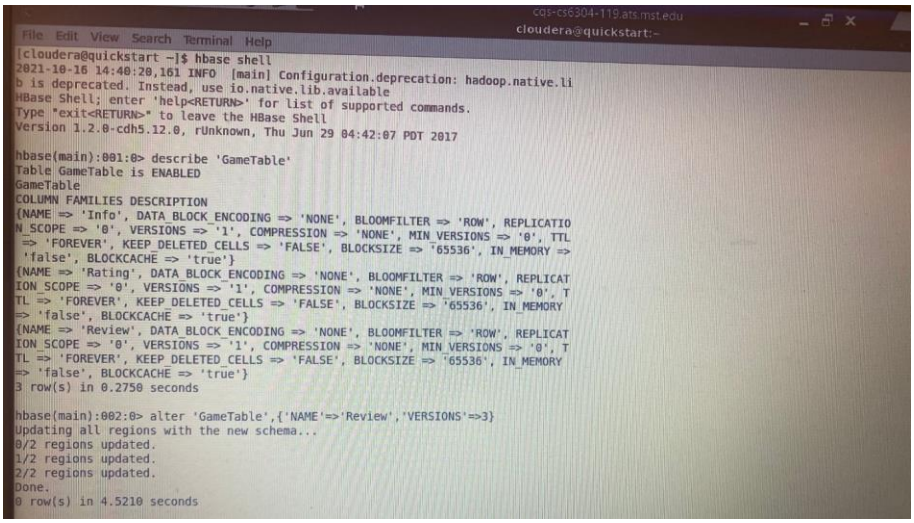
hbase(main):004:0> scan 'GameTable', {'LIMIT'=>5}



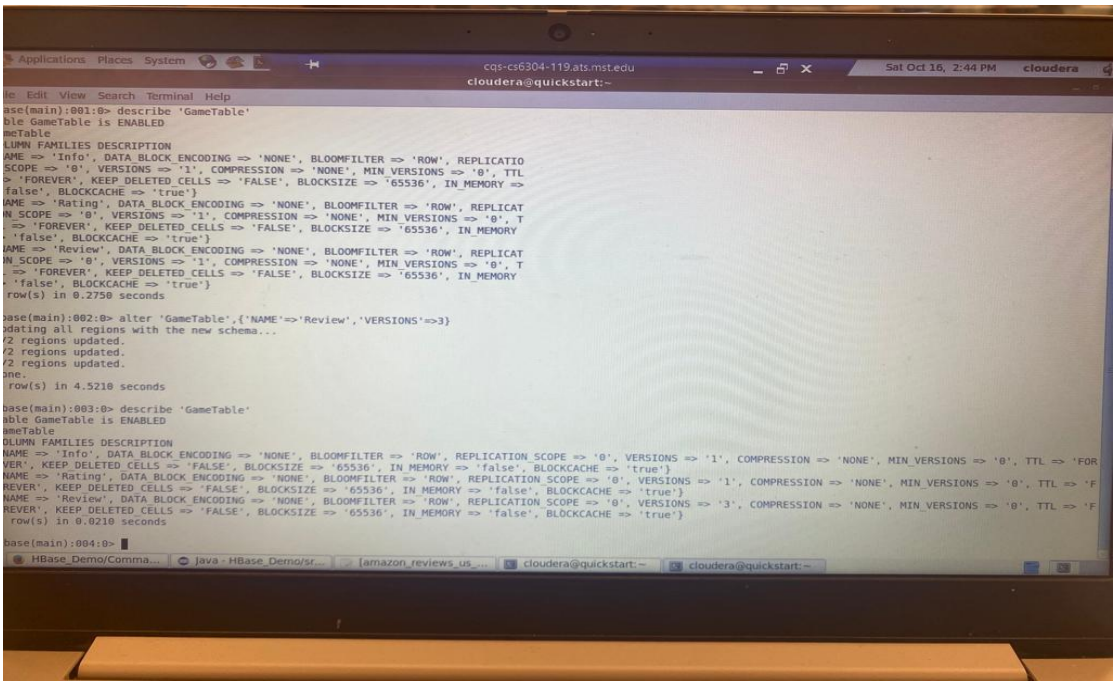
Part 2:

Command: Alter to change versions

hbase(main):017:0> alter 'GameTable', {'NAME'=>'Review', 'VERSIONS'=>3}



hbase(main):001:0> describe 'GameTable' => to show the change in versions



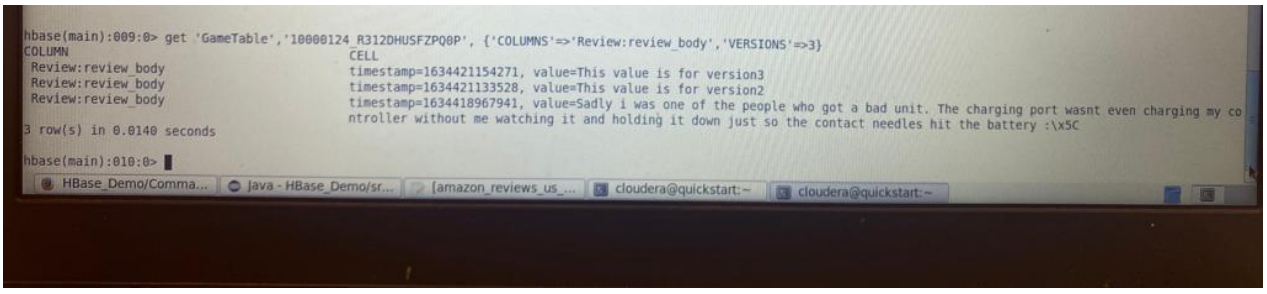
Random row to put additional 2 different 'review_body':

hbase(main):017:0> put 'GameTable','10000124_R312DHUSFZPQ0P','Review:review_body', 'This value is for version2'

hbase(main):017:0> put 'GameTable','10000124_R312DHUSFZPQ0P','Review:review_body', 'This value is for version3'

hbase(main):017:0>

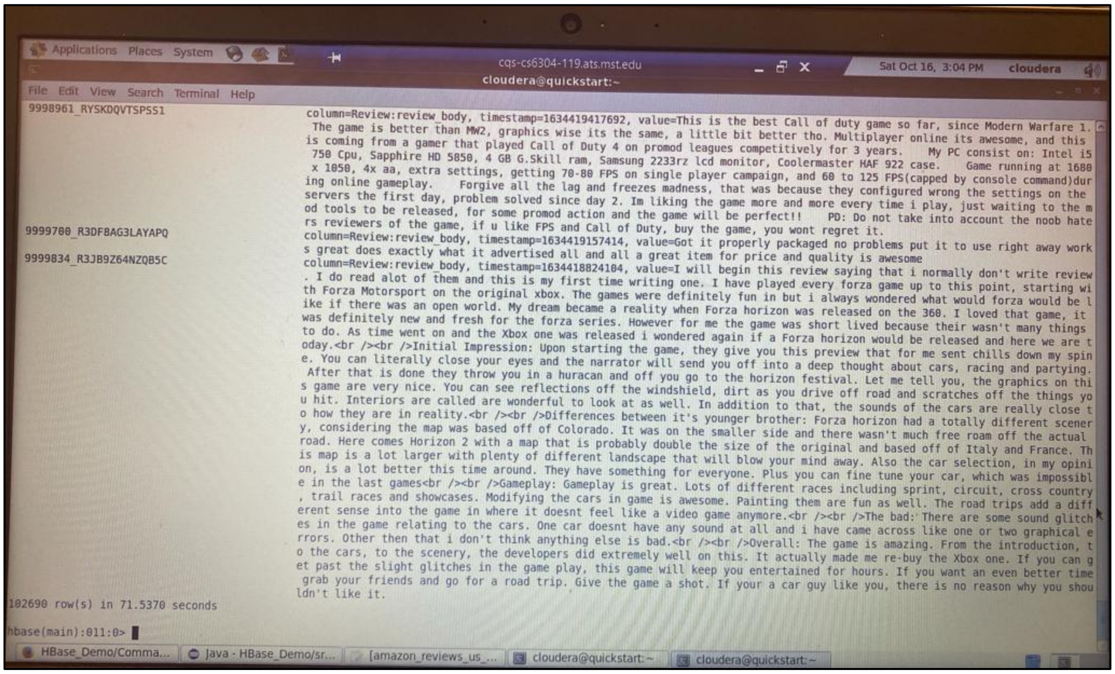
get 'GameTable','10000124_R312DHUSFZPQ0P', {'COLUMNS'=>'Review:review_body','VERSIONS'=>3}



Find the 'review_bodys' that have the word 'awesome'. => 102690 rows

hbase(main):017:0>

scan 'GameTable', {'COLUMNS'=>'Review:review_body', 'FILTER'=>"ValueFilter(=, 'substring:awesome')"}

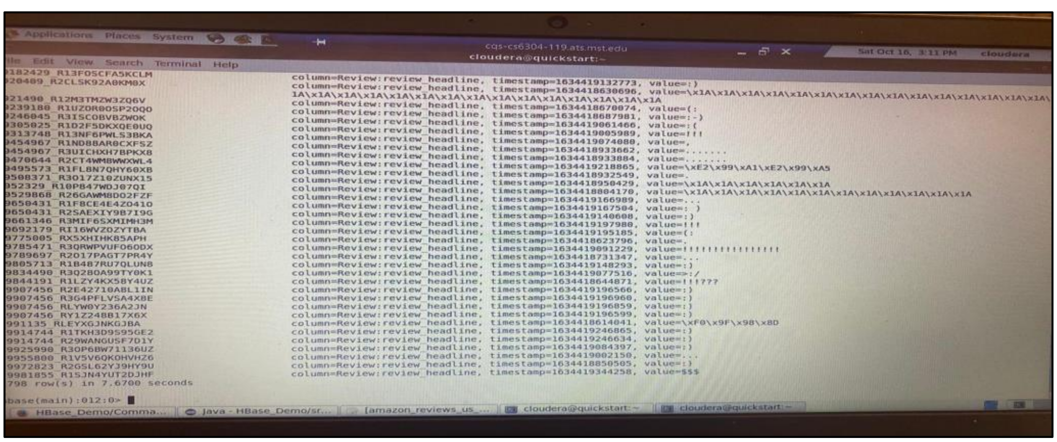


Find the 'review_headlines' that have any characters apart from alphanumerical characters (use regex)

=> 1798 rows

hbase(main):017:0>

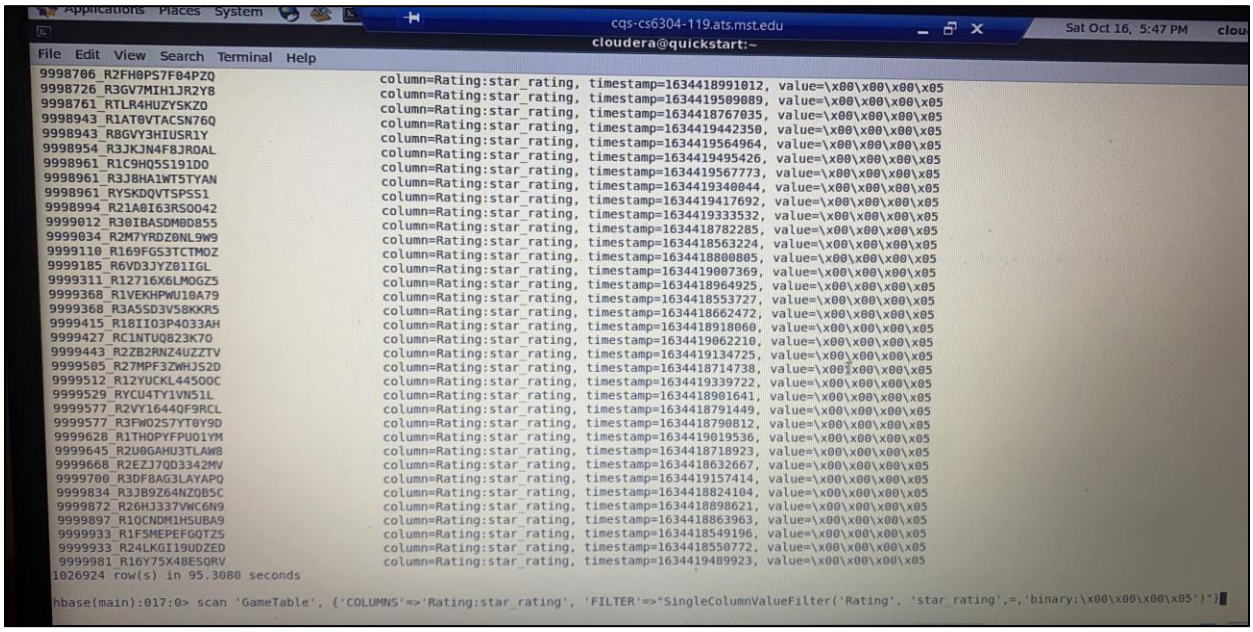
scan 'GameTable', {'COLUMNS'=>'Review:review_headline','FILTER'=>"ValueFilter(=, 'regexstring:^[^a-zA-Z0-9]*$')"}



Part 3:

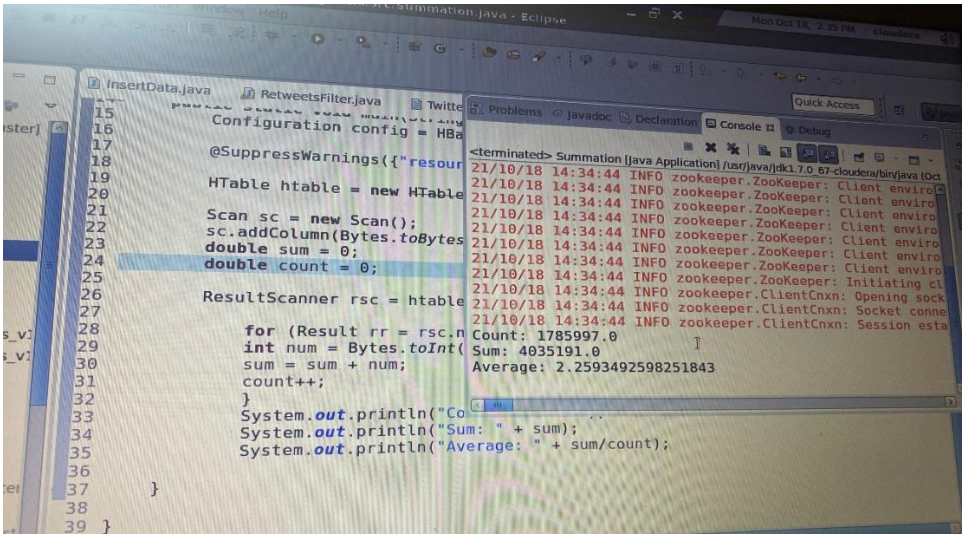
i) **Find reviews have 'star_rating' equal to 5:** 1026924 rows

hbase(main):017:0> scan 'GameTable', {'COLUMNS'=>'Rating:star_rating', 'FILTER'=>"SingleColumnValueFilter('Rating', 'star_rating',=,'binary:\x00\x00\x00\x05')"}



ii) Find the average 'helpful_vote' in the dataset.

Summation.java



iii) Show the 'review_headlines' that got 1 'star_rating'

hbase(main):004:0> scan 'GameTable',
{'COLUMNS'=>['Review:review_headline','Rating:star_rating'], 'LIMIT'=>15,
'FILTER'=>"SingleColumnValueFilter('Rating', 'star_rating', =, 'binary:\x00\x00\x00\x01')"}