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.

Part 1

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

* Keep three column familys:
(i) Info (contains columns 'marketplace' and 'verified_purchase')
(ii) Rating (contains columns 'star_rating', 'helpful_votes', 'total_votes',)
(iii) Review (contains columns 'review_headline', 'review_body')

The columns under 'Rating' should be integer. The columns under 'Info' and 'Review' should be string. Concat 'customer_id' and 'review_id' by an underscore to use it as the Row Key. (for example, some random Row Key for a certain row can be 11_22, where 11 is the customer_id and 22 is the review_id. After creating the table and inserting data, run 'describe' command to show that the table has been created perfectly and then run 'scan' & 'limit' command to show 5 rows.

Part 2

* Alter 'Review' column family to support 3 versions. Now, take a random row to put additional 2 different 'review_body'. Then show the all 3 'review_body' for this row.

* Find the 'review_bodys' that have the word 'awesome'.

* Find the 'review_headlines' that have any characters apart form alphanumerical characters (use regex).

Part 3:

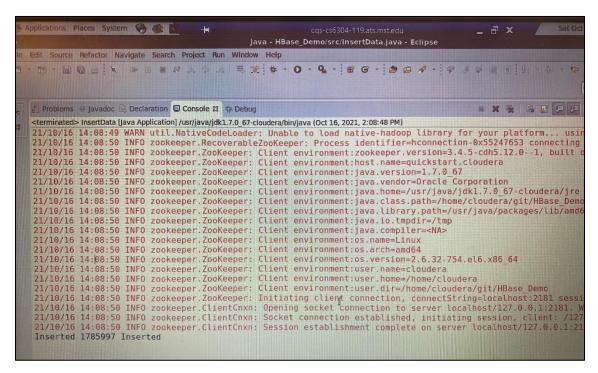* Find how many reviews have 'star_rating' equal to 5.

* Find the average 'helpful_vote' in the dataset.

* Show the 'review_headlines' that got 1 'star_rating'.

Part 1:

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

InsertData.java

Data Inserted: 1785997 rows


```
 Applications  Places  System                                    cqs-cs6304-119.ats.mst.edu                              _ ⬜ X            Sat Oct
                                             Java - HBase_Demo/src/InsertData.java - Eclipse
 le  Edit  Source  Refactor  Navigate  Search  Project  Run  Window  Help

   Problems  @ Javadoc  Declaration  Console ⊠  Debug                                             ⬛ ✖ ✖  🔲 🔲 🔲 🔲
 <terminated> InsertData [Java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Oct 16, 2021, 2:08:48 PM)
 21/10/16 14:08:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
 21/10/16 14:08:50 INFO zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x55247653 connecting
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.5-cdh5.12.0--1, built c
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:host.name=quickstart.cloudera
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.version=1.7.0_67
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.home=/usr/java/jdk1.7.0_67-cloudera/jre
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.class.path=/home/cloudera/git/HBase_Demo
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd6
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:os.name=Linux
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:os.arch=amd64
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:os.version=2.6.32-754.el6.x86_64
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:user.name=cloudera
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:user.home=/home/cloudera
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Client environment:user.dir=/home/cloudera/git/HBase_Demo
 21/10/16 14:08:50 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessi
 21/10/16 14:08:50 INFO zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. W
 21/10/16 14:08:50 INFO zookeeper.ClientCnxn: Socket connection established, initiating session, client: /127
 21/10/16 14:08:50 INFO zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:21
 Inserted 1785997 Inserted
```

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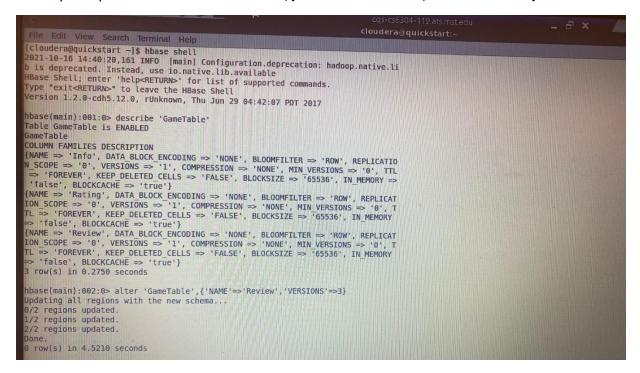


```
2021-10-16 14:10:17,425 INFO  [main] configuration.deprecation: hadoop.native.li
b is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.12.0, rUnknown, Thu Jun 29 04:42:07 PDT 2017

hbase(main):001:0> describe 'GameTable'
Table GameTable is ENABLED
GameTable
COLUMN FAMILIES DESCRIPTION
{NAME => 'Info', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATIO
N_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL
 => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY =>
 'false', BLOCKCACHE => 'true'}
{NAME => 'Rating', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICAT
ION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', T
TL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY
=> 'false', BLOCKCACHE => 'true'}
{NAME => 'Review', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICAT
ION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', T
TL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY
=> 'false', BLOCKCACHE => 'true'}
3 row(s) in 0.3600 seconds
```

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

```
10000099_R38GX1BSE6I column=Rating:star_rating, timestamp=1634418631509, value=
3LM                   \x00\x00\x00\x05
10000099_R38GX1BSE6I column=Rating:total_votes, timestamp=1634418631509, value=
3LM                   \x00\x00\x00\x00
10000099_R38GX1BSE6I column=Review:review_body, timestamp=1634418631509, value=
3LM                   I love this charger. Works perfect and charges way faster
                      than the use method
10000099_R38GX1BSE6I column=Review:review_headline, timestamp=1634418631509, va
3LM                   lue=Five Stars
10000124_R312DHUSFZP column=Info:marketplace, timestamp=1634418967941, value=US
Q0P
10000124_R312DHUSFZP column=Info:verified_purchase, timestamp=1634418967941, va
Q0P                   lue=Y
10000124_R312DHUSFZP column=Rating:helpful_votes, timestamp=1634418967941, valu
Q0P                   e=\x00\x00\x00\x00
10000124_R312DHUSFZP column=Rating:star_rating, timestamp=1634418967941, value=
Q0P                   \x00\x00\x00\x02
10000124_R312DHUSFZP column=Rating:total_votes, timestamp=1634418967941, value=
Q0P                   \x00\x00\x00\x00
10000124_R312DHUSFZP column=Review:review_body, timestamp=1634418967941, value=
Q0P                   Sadly i was one of the people who got a bad unit. The char
                      ging port wasnt even charging my controller without me wat
                      ching it and holding it down just so the contact needles h
                      it the battery :\x5C
10000124_R312DHUSFZP column=Review:review_headline, timestamp=1634418967941, va
Q0P                   lue=Not worth the buy
5 row(s) in 0.2470 seconds
```
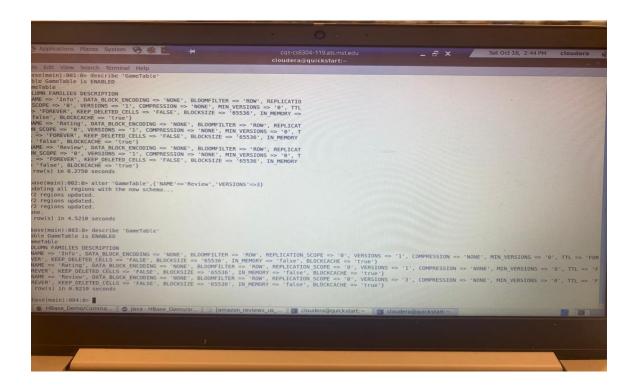
Part 2:

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

```
                                                    cqs-cs6304-119.ats.mst.edu              _ 🗗 ✕
                                                    cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ hbase shell
2021-10-16 14:40:20,161 INFO  [main] Configuration.deprecation: hadoop.native.li
b is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.12.0, rUnknown, Thu Jun 29 04:42:07 PDT 2017

hbase(main):001:0> describe 'GameTable'
Table GameTable is ENABLED
GameTable
COLUMN FAMILIES DESCRIPTION
{NAME => 'Info', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATIO
N_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL
 => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY =>
'false', BLOCKCACHE => 'true'}
{NAME => 'Rating', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICAT
ION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', T
TL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY
=> 'false', BLOCKCACHE => 'true'}
{NAME => 'Review', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICAT
ION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', T
TL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY
=> 'false', BLOCKCACHE => 'true'}
3 row(s) in 0.2750 seconds

hbase(main):002:0> alter 'GameTable',{'NAME'=>'Review','VERSIONS'=>3}
Updating all regions with the new schema...
0/2 regions updated.
1/2 regions updated.
2/2 regions updated.
Done.
0 row(s) in 4.5210 seconds
```
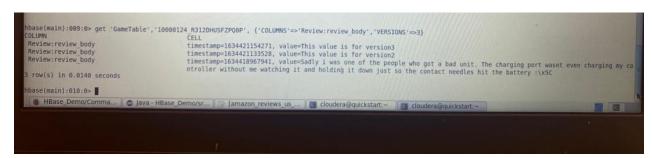
hbase(main):001:0> describe 'GameTable'

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'
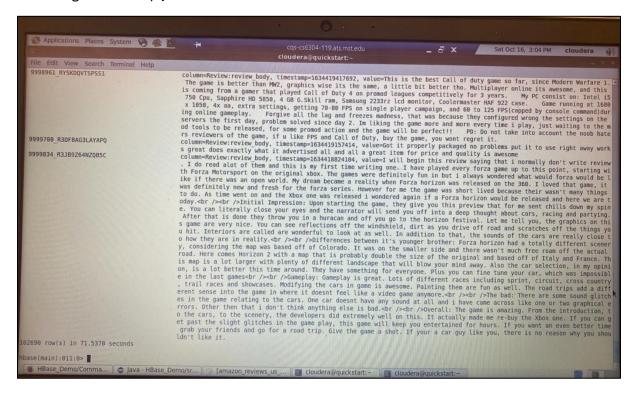
hbase(main):017:0>

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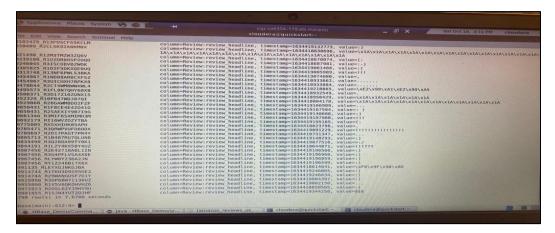

Substring: "awesome" 102690 rows

hbase(main):017:0>

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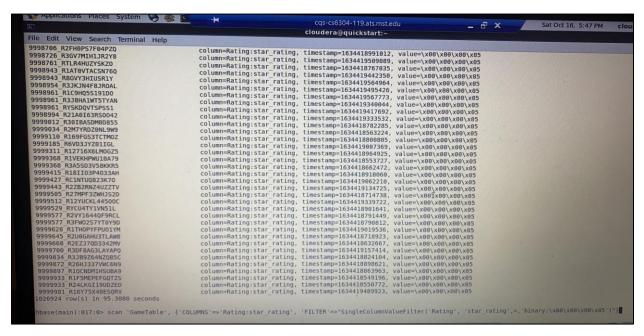

Regex string : 1798 rows

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')"}