Instructions for CIS graduate class Term Paper Manuscript

Group 2
(Hut Tangchitnob, Swarnim Jambhule, Maria Espinoza, Julio Rojas)
Department of Information Systems, California State University
Los Angeles
E-mails : stangch2@caltstatela.edu
sjambhu@calstatela.edu
mespino6@calstate.edu
jrojas2@calstatela.edu

Abstract: This manuscript describes the findings from an Amazon Book Product Review data set hosted by Amazon. Using HiveQL and Power BI, specific product sentiment and overall review sentiment can be analyzed.

1. Introduction

Big Data is an increasingly important field that is growing in use with business analytics. In this instance, we use Oracle Big Data Compute Edition with the following platform specs: CPU Speed, 2.2 GHz, 10 OCPU, 150 GB Memory, 678 GB Storage, and 147 GB HDFS Capacity, to download and extract data for analysis with HiveQL and Power BI. Using HiveQL, we can create tables and query data seamlessly with HDFS and Oracle Cloud. Using Power BI, we can visualize the data using a wide range of tools. Tempo-spatial analysis cannot be

2. Related Works

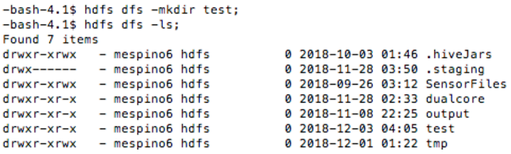
The paper [1] presents the issues and challenges of data integration in Big Data environment and techniques for big data integration. The paper [1] also presents a new ETL framework to handle future research in the big data environment. On the other hand, this paper presents a methodical approach to big data analysis with HiveQL and Power BI.

In paper [2], a scientific workflow framework is proposed for big geoscience data analytics by utilizing cloud computing, MapReduce, and Service Oriented Architecture. The proof-of-concept prototype tests the performance of this framework by showing the efficiency of big geodata science analytics in data processing time reduction. On the other hand, this paper presents a methodical approach to big data analysis with HiveQL and Power BI.

2. General Instructions

Using Oracle Cloud, first download and unzip the data, then upload it to HDFS, and finally, query it using HiveQL. Use dictionary table to assign polarity values to each word in the review body. Use HiveQL's average function to create a table of average polarity relative to average star rating. Download this data onto disk and upload into Power BI to being visualization. Place data onto bar graph and assign appropriate X, Y values to get final visualization presented in Figure 5.

Step 1: Remotely connect to Oracle Cloud[1].

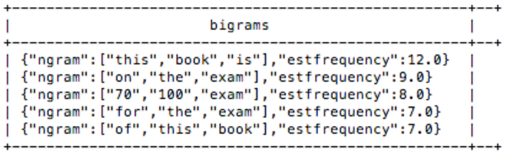


```
-bash-4.1$ hdfs dfs -mkdir test;
-bash-4.1$ hdfs dfs -ls;
Found 7 items
drwxr-xrwx   - mespino6 hdfs          0 2018-10-03 01:46 .hiveJars
drwx------   - mespino6 hdfs          0 2018-11-28 03:50 .staging
drwxr-xrwx   - mespino6 hdfs          0 2018-09-26 03:12 SensorFiles
drwxr-xr-x   - mespino6 hdfs          0 2018-11-28 02:33 dualcore
drwxr-xr-x   - mespino6 hdfs          0 2018-11-08 22:25 output
drwxr-xr-x   - mespino6 hdfs          0 2018-12-03 04:05 test
drwxr-xrwx   - mespino6 hdfs          0 2018-12-01 01:22 tmp
```

Figure 1: Folders present in Oracle Cloud.

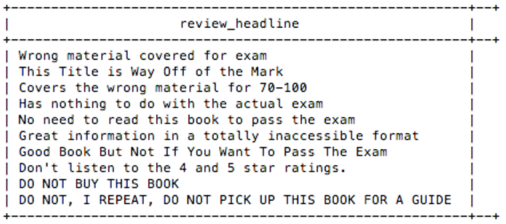
Step 2: Download data into Oracle Big Data Compute Edition.

Step 3: Create HiveQL tables and Queries to Analyze Data



```
+-----------------------------------------------------------+--+
|                          bigrams                          |
+-----------------------------------------------------------+--+
| {"ngram":["this","book","is"],"estfrequency":12.0}        |
| {"ngram":["on","the","exam"],"estfrequency":9.0}          |
| {"ngram":["70","100","exam"],"estfrequency":8.0}          |
| {"ngram":["for","the","exam"],"estfrequency":7.0}         |
| {"ngram":["of","this","book"],"estfrequency":7.0}         |
+-----------------------------------------------------------+--+
```

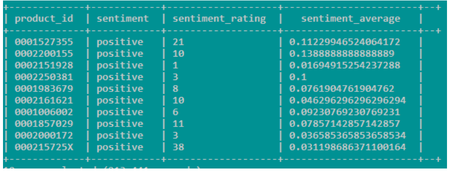
Table 1: Trigrams analysis results.



```
+-----------------------------------------------------------+--+
|                      review_headline                      |
+-----------------------------------------------------------+--+
| Wrong material covered for exam                           |
| This Title is Way Off of the Mark                         |
| Covers the wrong material for 70-100                      |
| Has nothing to do with the actual exam                    |
| No need to read this book to pass the exam                |
| Great information in a totally inaccessible format        |
| Good Book But Not If You Want To Pass The Exam            |
| Don't listen to the 4 and 5 star ratings.                 |
| DO NOT BUY THIS BOOK                                      |
| DO NOT, I REPEAT, DO NOT PICK UP THIS BOOK FOR A GUIDE    |
+-----------------------------------------------------------+--+
```

Table 2: Headline query results for lowest rated product.

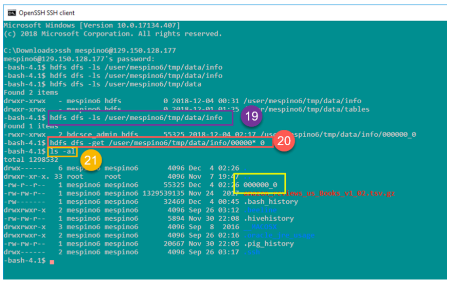
Step 4: Create Hive Queries to analyze the sentiment of data using dictionary and download data into disk.



```
+-------------+------------+-------------------+-----------------------+--+
| product_id  | sentiment  | sentiment_rating  |   sentiment_average   |
+-------------+------------+-------------------+-----------------------+--+
| 0001527355  | positive   | 21                | 0.11229946524064172   |
| 0002200155  | positive   | 10                | 0.1388888888888889    |
| 0002151928  | positive   | 1                 | 0.01694915254237288   |
| 0002250381  | positive   | 3                 | 0.1                   |
| 0001983679  | positive   | 8                 | 0.0761904761904762    |
| 0002161621  | positive   | 10                | 0.046296296296296294  |
| 0001006002  | positive   | 6                 | 0.09230769230769231   |
| 0001857029  | positive   | 11                | 0.07857142857142857   |
| 0002000172  | positive   | 3                 | 0.036585365853658534  |
| 000215725X  | positive   | 38                | 0.031198686371100164  |
+-------------+------------+-------------------+-----------------------+--+
```

Table 3: Table with sentiment averages.

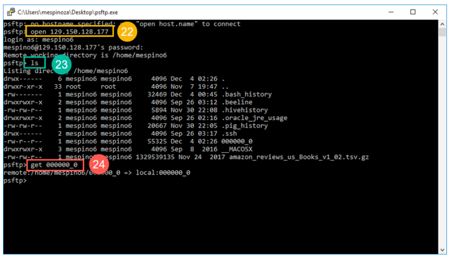
[1] You must have an IP address to login to Oracle Cloud



```
Microsoft Windows [Version 10.0.17134.407]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Downloads>ssh mespino6@129.150.128.177
mespino6@129.150.128.177's password:
-bash-4.1$ hdfs dfs -ls /user/mespino6/tmp/data/info
-bash-4.1$ hdfs dfs -ls /user/mespino6/tmp/data/info
-bash-4.1$ hdfs dfs -ls /user/mespino6/tmp/data
Found 2 items
drwxr-xrwx   - mespino6 hdfs          0 2018-12-04 00:31 /user/mespino6/tmp/data/info
drwxr-xrwx   - mespino6 hdfs          0 2018-12-01 01:25 /user/mespino6/tmp/data/tables
-bash-4.1$ hdfs dfs -ls /user/mespino6/tmp/data/info
Found 1 items
-rwxr-xrwx   2 bdcsce_admin hdfs      55325 2018-12-04 02:17 /user/mespino6/tmp/data/info/000000_0
-bash-4.1$ hdfs dfs -get /user/mespino6/tmp/data/info/00000* 0
-bash-4.1$ ls -al
total 1298532
drwx------    6 mespino6 mespino6       4096 Dec  4 02:26 .
drwxr-xr-x. 33 root     root           4096 Nov  7 19:47 ..
-rw-r--r--    1 mespino6 mespino6      55325 Dec  4 02:26 000000_0
-rw-rw-r--    1 mespino6 mespino6 1329539135 Nov 24  2017 amazon_reviews_us_Books_v1_02.tsv.gz
-rw-------    1 mespino6 mespino6      32469 Dec  4 00:45 .bash_history
drwxrwxr-x    2 mespino6 mespino6       4096 Sep 26 03:12 .beeline
-rw-rw-r--    1 mespino6 mespino6       5894 Nov 30 22:08 .hivehistory
drwxrwxr-x    3 mespino6 mespino6       4096 Sep  8  2016 __MACOSX
drwxrwxr-x    2 mespino6 mespino6       4096 Sep 26 02:16 .oracle_jre_usage
-rw-rw-r--    1 mespino6 mespino6      20667 Nov 30 22:05 .pig_history
drwx------    2 mespino6 mespino6       4096 Sep 26 03:17 .ssh
-bash-4.1$
```

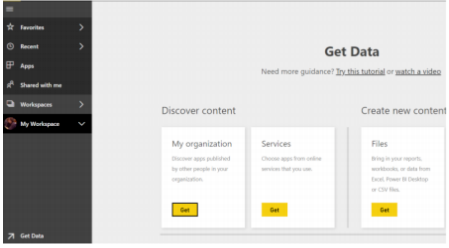
Figure 2: Place file in Oracle Cloud.



```
psftp: no hostname specified; use "open host.name" to connect
psftp> open 129.150.128.177
login as: mespino6
mespino6@129.150.128.177's password:
Remote working directory is /home/mespino6
psftp> ls
Listing directory /home/mespino6
drwx------    6 mespino6 mespino6     4096 Dec  4 02:26 .
drwxr-xr-x   33 root     root         4096 Nov  7 19:47 ..
-rw-------    1 mespino6 mespino6    32469 Dec  4 00:45 .bash_history
drwxrwxr-x    2 mespino6 mespino6     4096 Sep 26 03:12 .beeline
-rw-rw-r--    1 mespino6 mespino6     5894 Nov 30 22:08 .hivehistory
drwxrwxr-x    2 mespino6 mespino6     4096 Sep 26 02:16 .oracle_jre_usage
-rw-rw-r--    1 mespino6 mespino6    20667 Nov 30 22:05 .pig_history
drwx------    2 mespino6 mespino6     4096 Sep 26 03:17 .ssh
-rw-r--r--    1 mespino6 mespino6    55325 Dec  4 02:26 000000_0
drwxrwxr-x    3 mespino6 mespino6     4096 Sep  8  2016 __MACOSX
-rw-rw-r--    1 mespino6 mespino6 1329539135 Nov 24  2017 amazon_reviews_us_Books_v1_02.tsv.gz
psftp> get 000000_0
remote:/home/mespino6/000000_0 => local:000000_0
psftp>
```

Figure 3: psftp GET command.

Step 5: Load data into Power BI



Get Data

Need more guidance? Try this tutorial or watch a video

Discover apps published by other people in your organization.

Choose apps from online services that you use.

Bring in your reports, workbooks, or data from Excel, Power BI Desktop or CSV files.

Figure 4: Power BI dashboard,

Step 6: Visualize data in Power BI

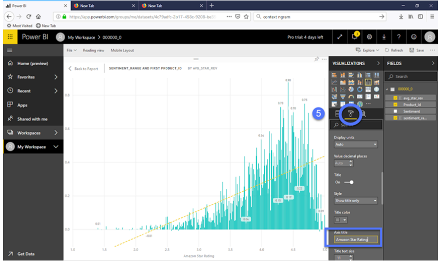


Figure 5: Visualized data in Power BI.

References

[1] L. Arockiam, "A Review on Big Data Integration" *International Journal of Computer Applications*, Vol.4, no. 2, pp.21-26, 2014.

[2] L. Zhenlong, "Enabling Big Geoscience Data Analytics with a Cloud-Based, MapReduce-Enabled and Service-Oriented Workflow Framework", PLos One, 2015.