

TWITTER DATA ANALYSIS USING HADOOP



MANTHAN SHAH

**Abstract**

Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a reasonable amount of time. The challenge is not only storing the data, but also accessing and analyzing the required data in specified amount of time. One of the ways to solve the above challenges of big data is using Hadoop. Hadoop is well-known open-source implementation of the MapReduce programming model for processing big data in parallel of data-intensive jobs on clusters of commodity servers. It is a highly scalable computing platform. Hadoop enables users to store and process large amount of data which is not possible while using less scalable techniques. Twitter, one of the largest social media sites, receives tweets in millions every day in the range of zettabyte per year. This huge amount of raw data can be used for industrial or business purposes by organizing data according to our requirement and processing. This report provides a way of analyzing big data such as Twitter data using Apache Hadoop which will process and analyze the tweets on Hadoop clusters.

**Introduction**

Big data generally exceeds the capacity of normal conventional traditional databases. The value of big data to an organization falls into two categories: analytical use and enabling new products based on existing ones. Big data can reveal the issues hidden by data that are too costly to process and performs analytics such as user transactions, social and geographical data issues faced by the industry.

In the distributed environment, the data should be available to users and capable of performing different analysis from databases in a specified amount of time. Hadoop is the open source flexible infrastructure for large scale computation and data processing on a network commodity of hardware systems. MapReduce allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The term MapReduce refers to two tasks in Hadoop i.e. Map and Reduce. In the first step, it takes the data and converts it into another set of data. Each word is referred as key and the number of occurrences is treated as value. So, MapReduce tuple consists of key value pairs. Then, the second step consists of reduce operation where it takes the output from a map as input and combines those data tuples into a smaller set of tuples.

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Apache Hive (HiveQL) with Hadoop Distributed File System is used for Analysis of data. Hive provides a SQL-like interface to process data stored in HDP.

**Proposed System**

The main scope of the project is to analyze and fetch the Twitter IDs of those users whose statuses have been tweeted the most by the user whose tweets are being analyzed. Analysis is done to also find out the words used most frequently and the most active user.

First, the system involves collecting the tweets from the social network using the Twitter API. Then, it consists of a standard platform such as Hadoop to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data for ease of handling. Finally, the collected tweets are analyzed to find out the most frequent words and bi-grams, as well as the most active user.

Twitter data is collected using API streaming of tokens using R by mentioning keywords and hashtags related to various types of Cryptocurrency. Then, the tweets are uploaded into Hadoop file system using HDFS commands. Finally, MapReduce is applied to find out the most frequently used words and the most active user.

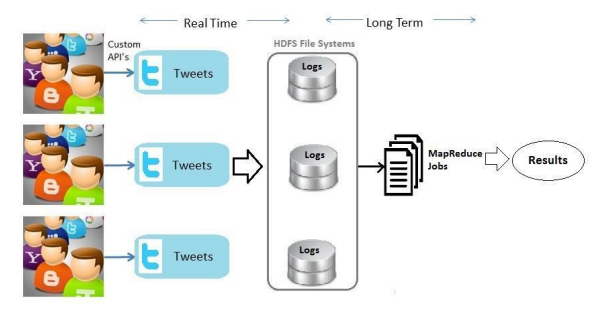


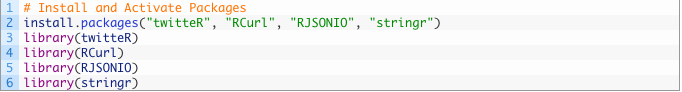
Fig: System Architecture

**Extraction of Twitter data**

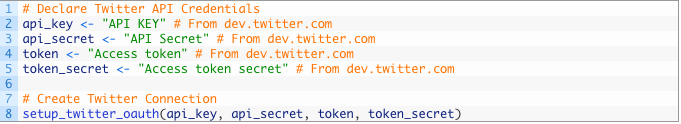
We extracted the data using the Twitter Search API. The Twitter search API, one of three such APIs (search, streaming, “firehose”), allows access to a subset of popular or recent tweets (in the last 4-6 days). That is, it allows querying past tweets. So, to be able to query the twitter search API and import the data into R, it is required to accomplish the following tasks:

1. Sign up for Twitter & create an application.
2. Install R and required R packages
3. Understand the Twitter Search API query structure
4. Run our first query, and save to database
5. **Sign up for Twitter & create an application:** We created a Twitter account and then we moved to dev.twitter.com. After logging in, we clicked on twitter thumbnail (upper right corner) and click on “My Applications”. Then we click on “Create New App” and added the required name, description and website. Once created, click on “modify app permissions” and allow the application to read, write and access direct messages. Then, we clicked on API Keys tab and scroll to the bottom of the page and clicked on “Create my access token”. These access tokens will be used when we fire up R.
6. **Install ‘R’ and required packages:** We installed R using **r-project.org** as per the platform we had. After installation, we installed the following packages:

* [TwitteR](http://cran.r-project.org/web/packages/twitteR/index.html)
* [ROAuth](http://cran.r-project.org/web/packages/ROAuth/index.html)
* [RCurl](http://cran.r-project.org/web/packages/RCurl/index.html)
* [Rjsonio](http://cran.r-project.org/web/packages/RJSONIO/index.html)
* [stringr](http://cran.r-project.org/web/packages/stringr/)



After the installation of the above packages, we created the connection with the Twitter API using the following code and entered the required tokens:



The API keys needs to be in quotation marks. The setup\_twitter\_oauth function is used to create the connection with Twitter.

**Twitter Search API Structure**

With above packages and code, we have already build the connection with Twitter. Per Twitter, the best way to build a query and test if it’s valid and will return matched tweets is to first try it at **twitter.com/search**. This in essence uses the same API that we are calling. We have selected the below terms and then, run the query:

* Bitcoin
* Dash
* Ethereum
* Litecoin
* Ripple
* Cash

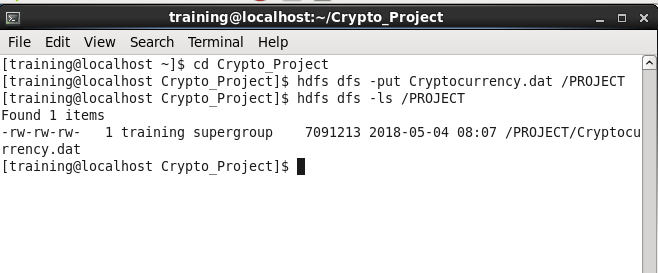
We changed the n value and extracted 1000 tweets at a time. After that, we transformed it into a data frame.

../Desktop/Screen%20Shot%202018-04-30%20at%205.29.06%20PM.pn

So, we have 24890 datasets in total by combining all the data in CSV format. The data was cleaned and filtered to remove redundancy and null values.

**Importing the Data into HDFS**

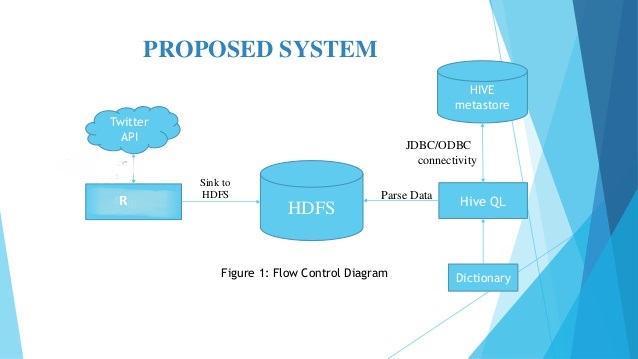
Querying Twitter data in a traditional RDBMS is inconvenient, since the Twitter Streaming API outputs tweets in a JSON format which can be arbitrarily complex. In the Hadoop ecosystem, the Hive project provides a query interface which can be used to query data that resides in HDFS. The query language looks very similar to SQL, but allows us to easily model complex types, so we can easily query the type of data we have. The first step is to get Twitter data into HDFS, and then to tell Hive where the data resides and how to read it.

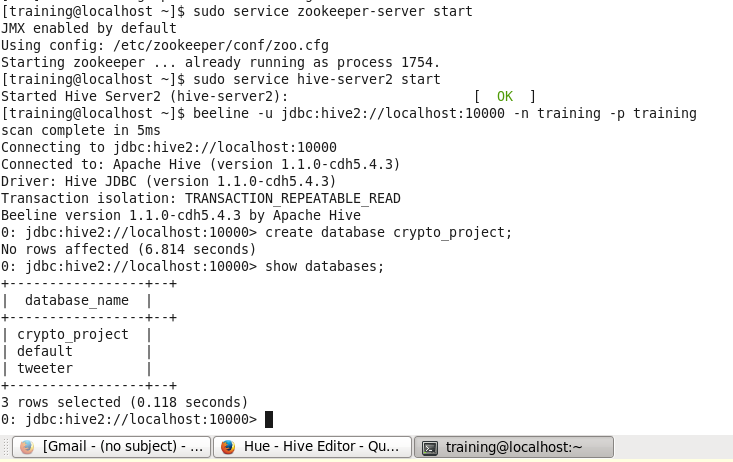


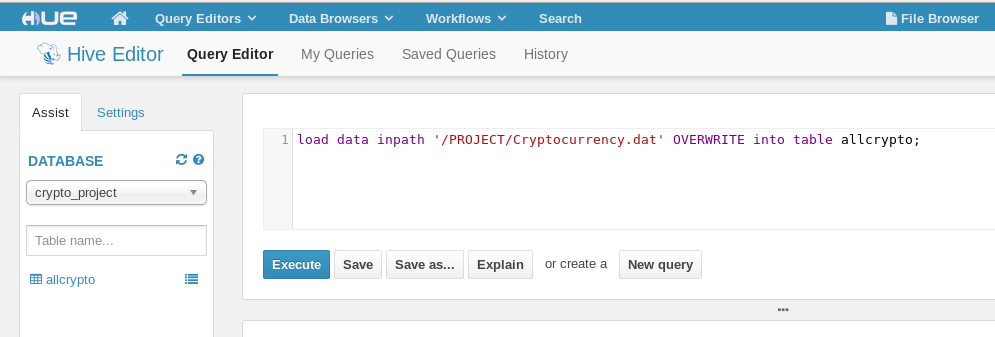
Once we have the Twitter data loaded into HDFS, we can stage it for querying by creating an external table in Hive. Using an external table will allow us to query the table without moving the data from the location where it ends up in HDFS. To ensure scalability, as we add more and more data, we’ll need to also partition the table.

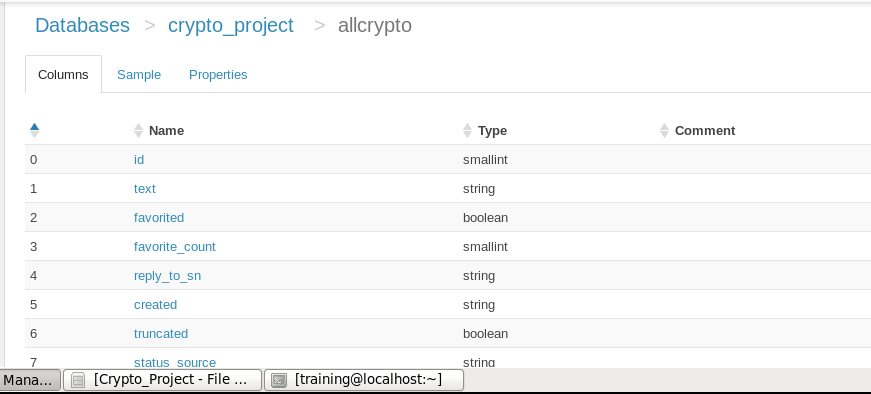
**Querying Complex Data with Hive**

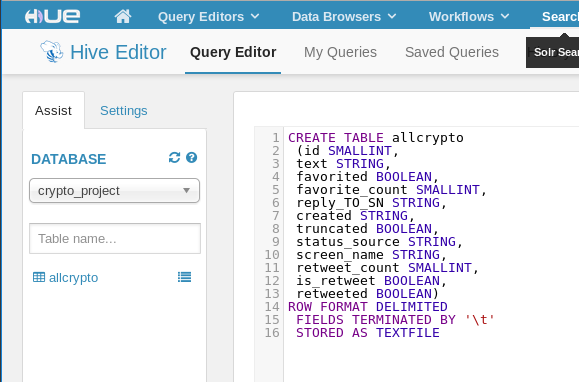
Before we can query the data, we need to ensure that the Hive table can properly interpret the JSON data. By default, Hive expects that input files use a delimited row format, but our Twitter data is in a JSON format, which will not work with the defaults. Hive allows us to flexibly define, and redefine, how the data is represented on disk.

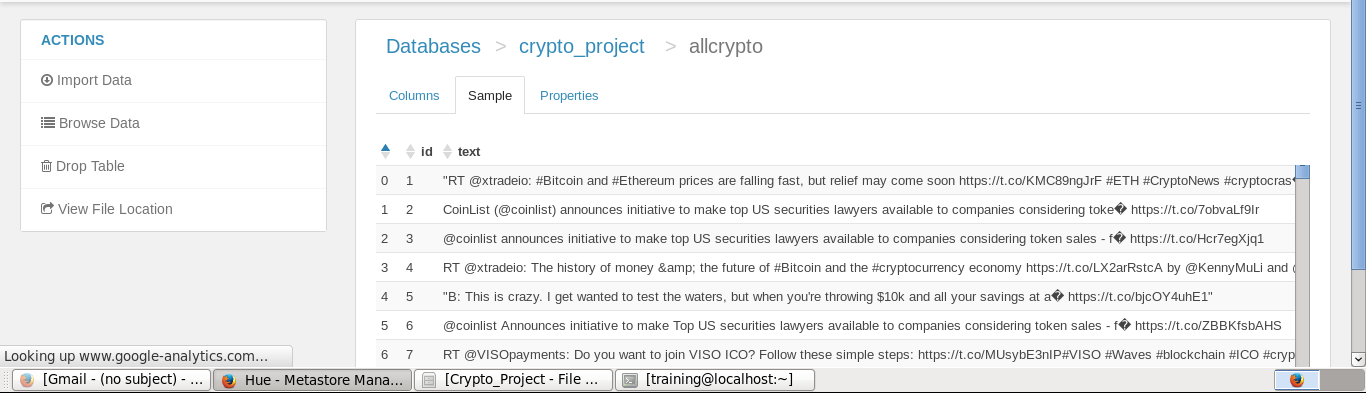


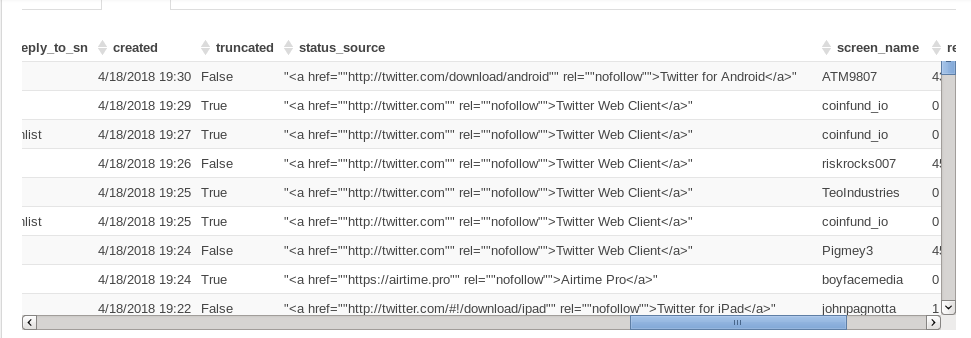










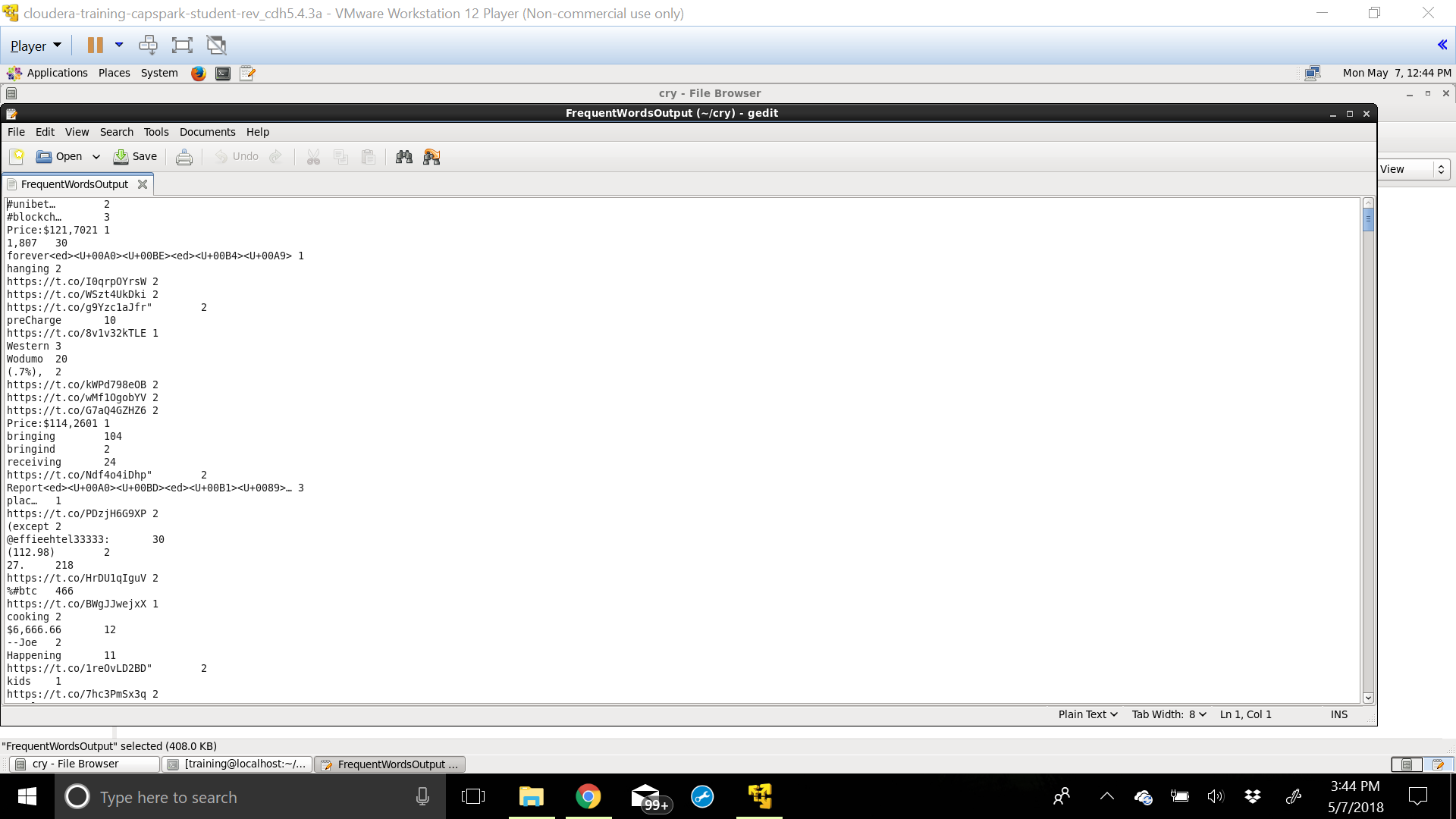


**Results**

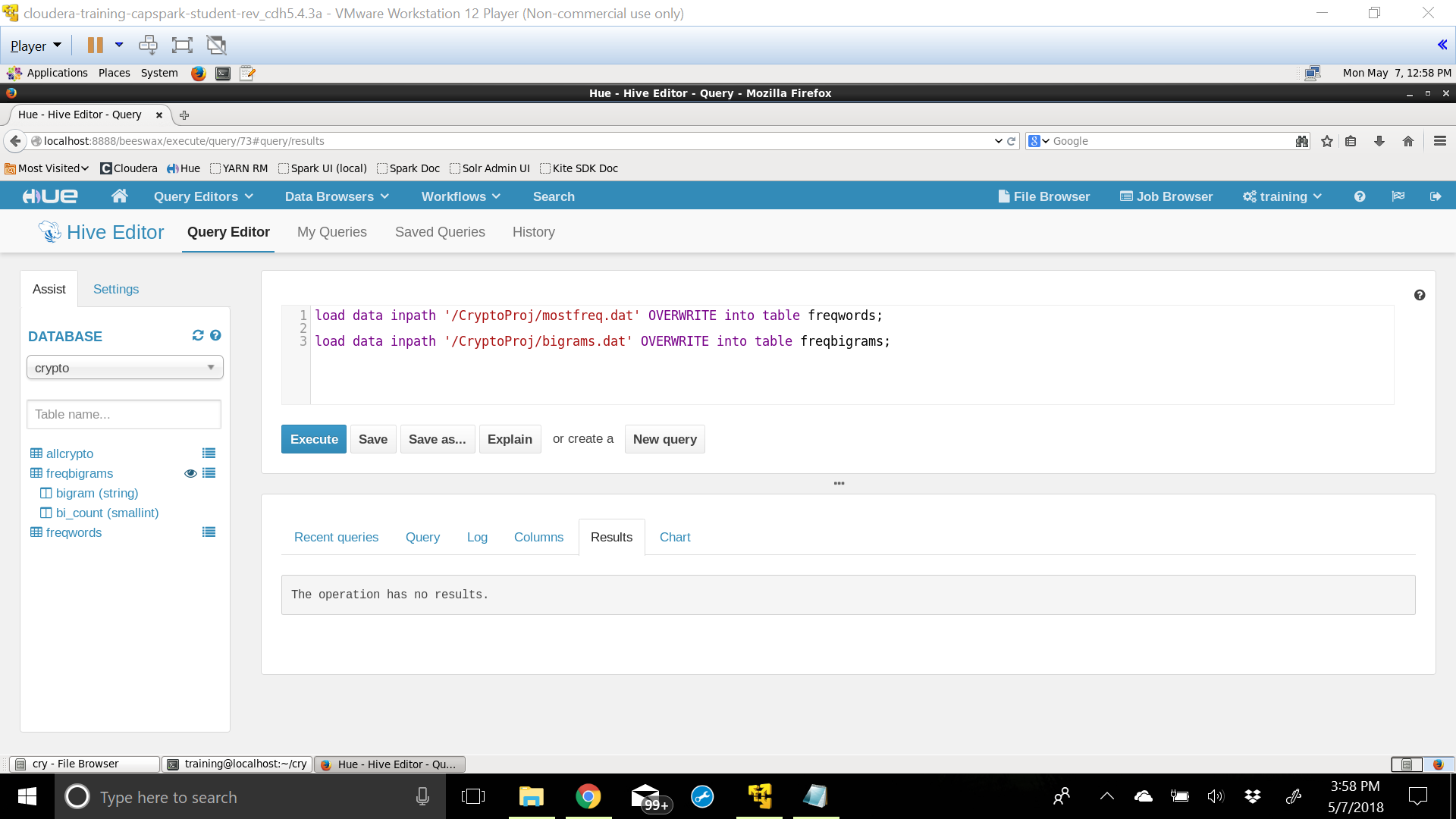
We collected data over 2 weeks filtering on a set of keywords relating to different forms of Cryptocurrency such as Bitcoin, Dash, Ethereum, Litecoin, Ripple and Zcash. The data was collected using R and combined into a single CSV file. The semi-structured quality of the data makes the data very difficult to query in a traditional RDBMS. Hive can handle this data much more gracefully.

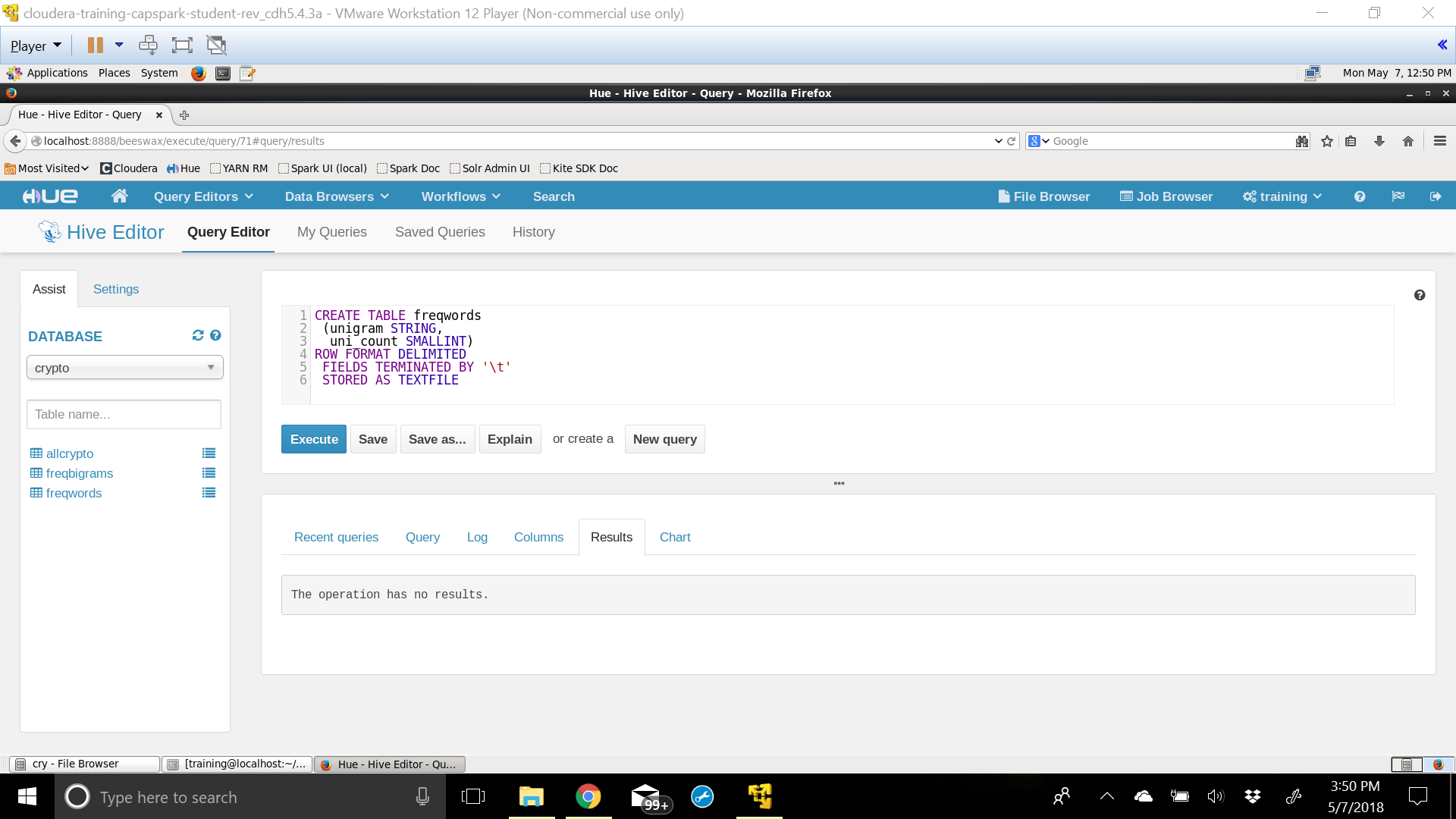
1. **Finding the most frequently used words**

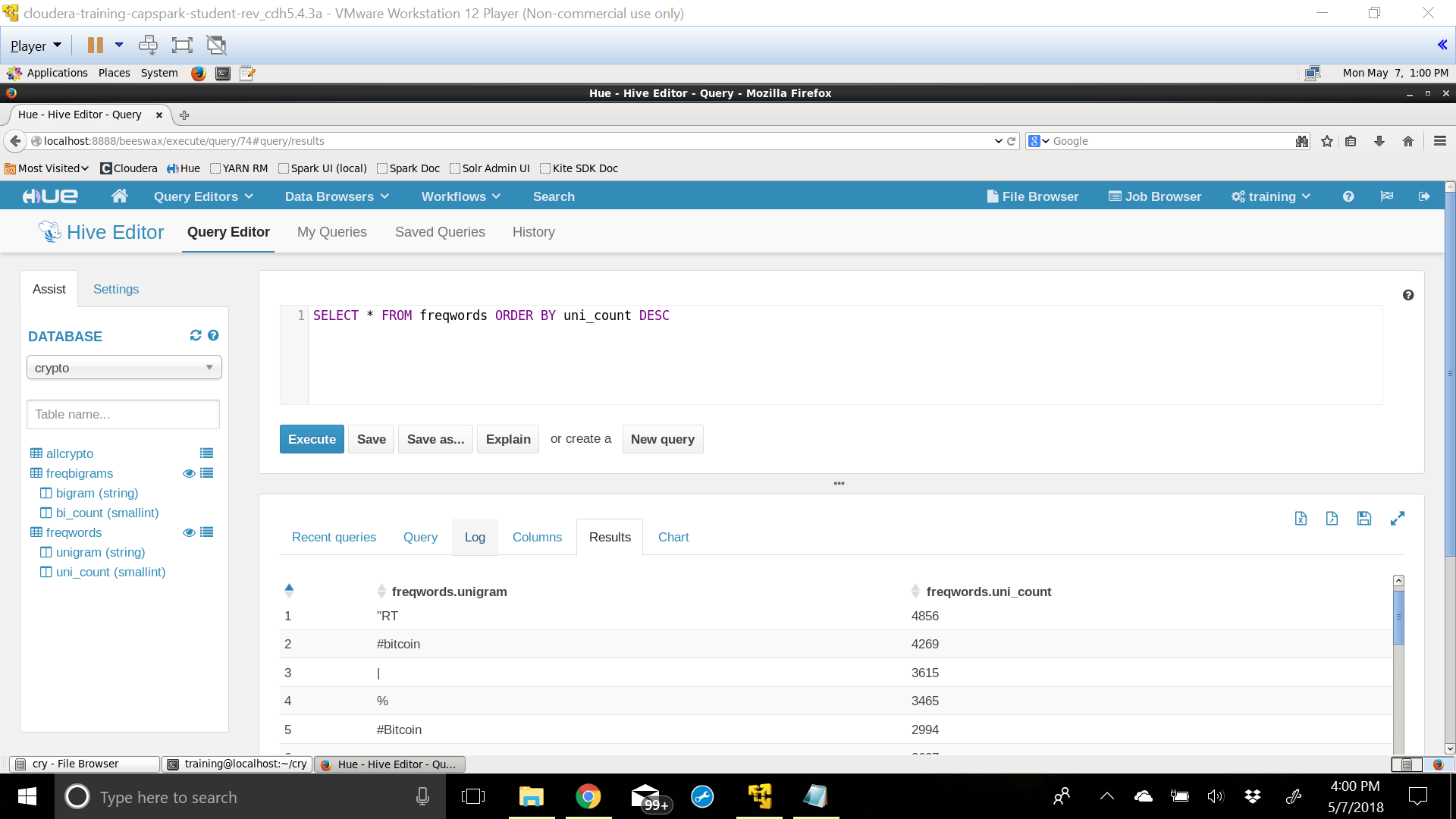
MapReduce jobs were created in Python to find the most frequently used words from the data and the below output was received.



We created tables in Hive to store the most frequent words and run the below query.



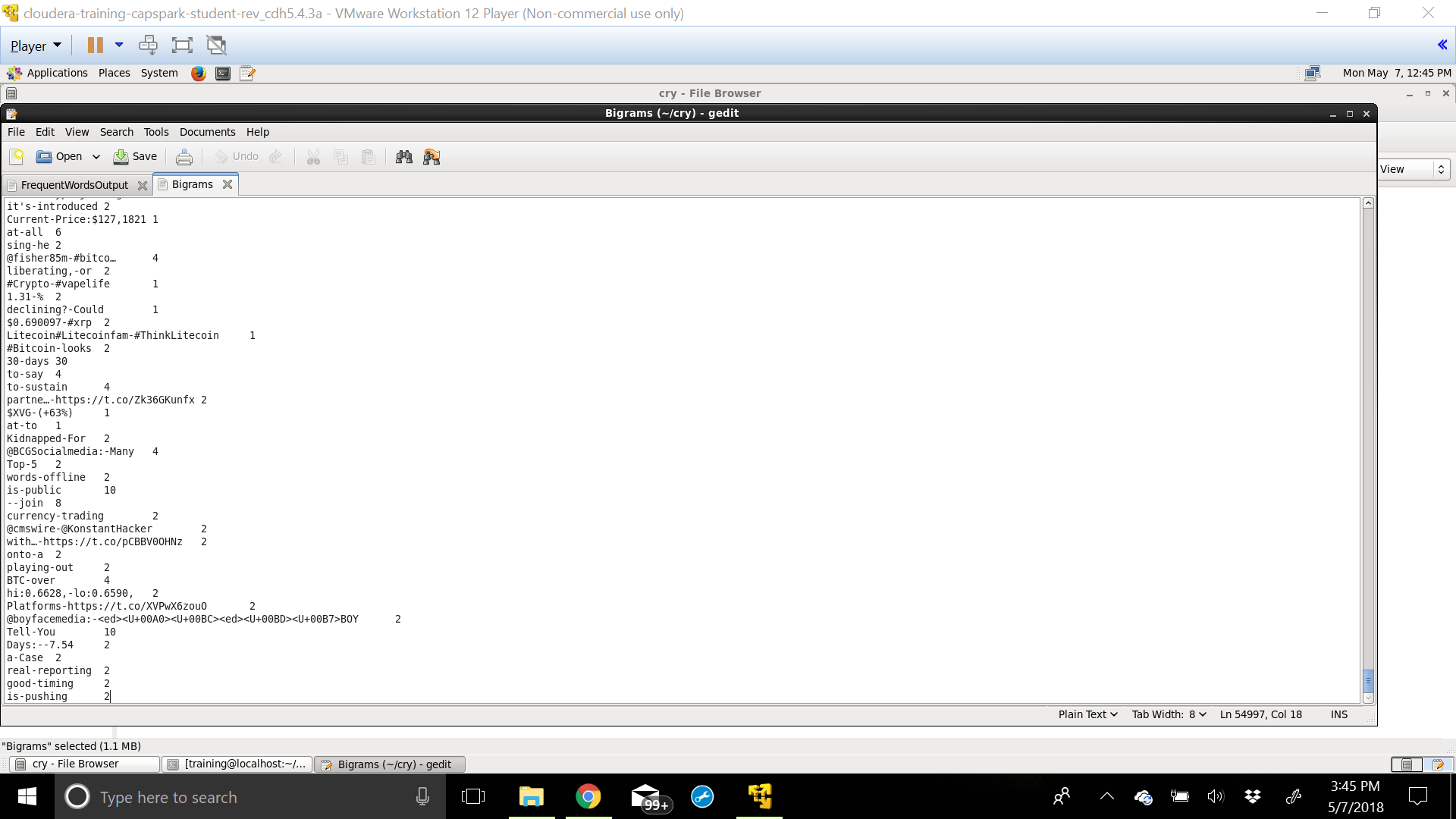


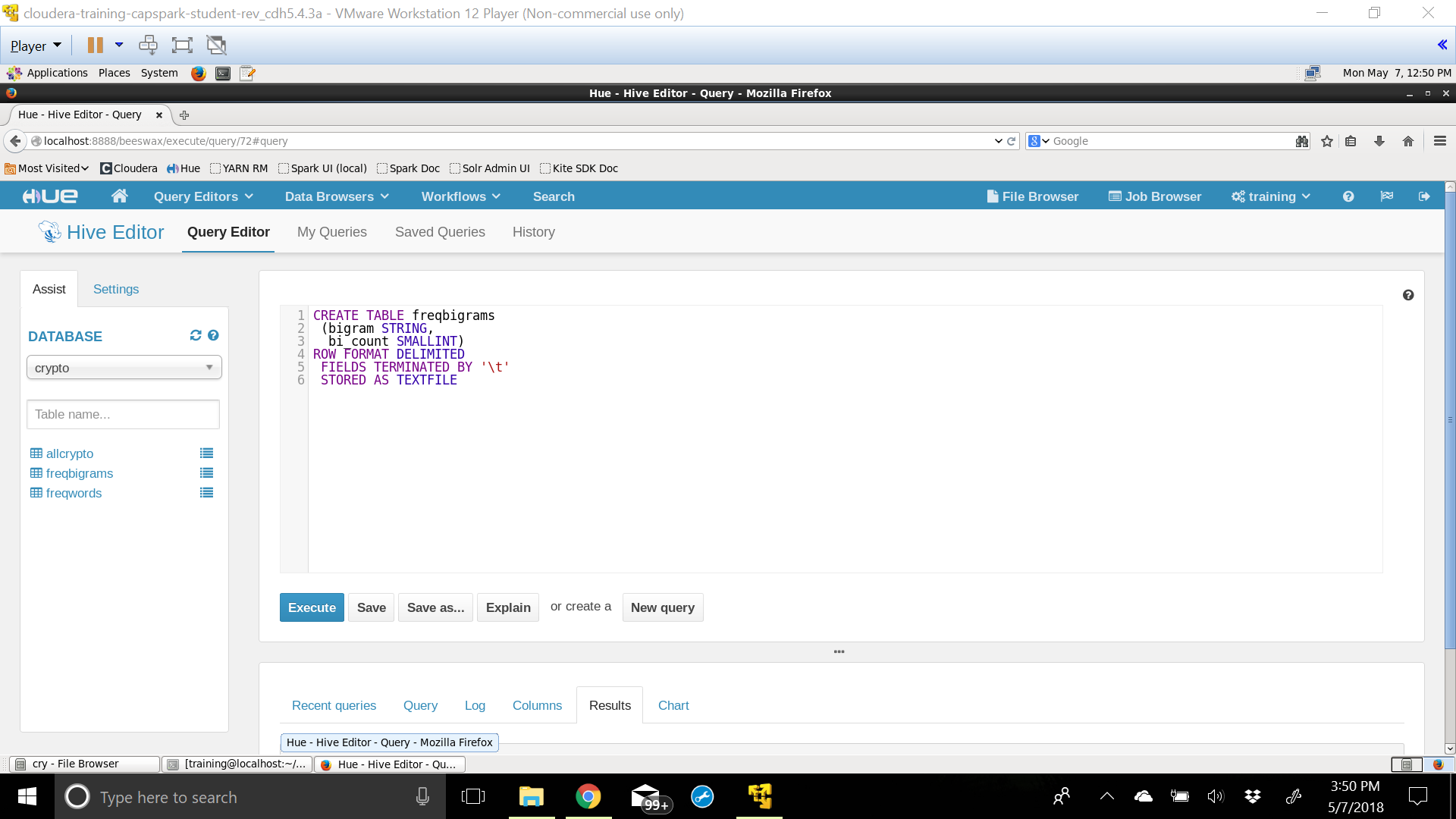


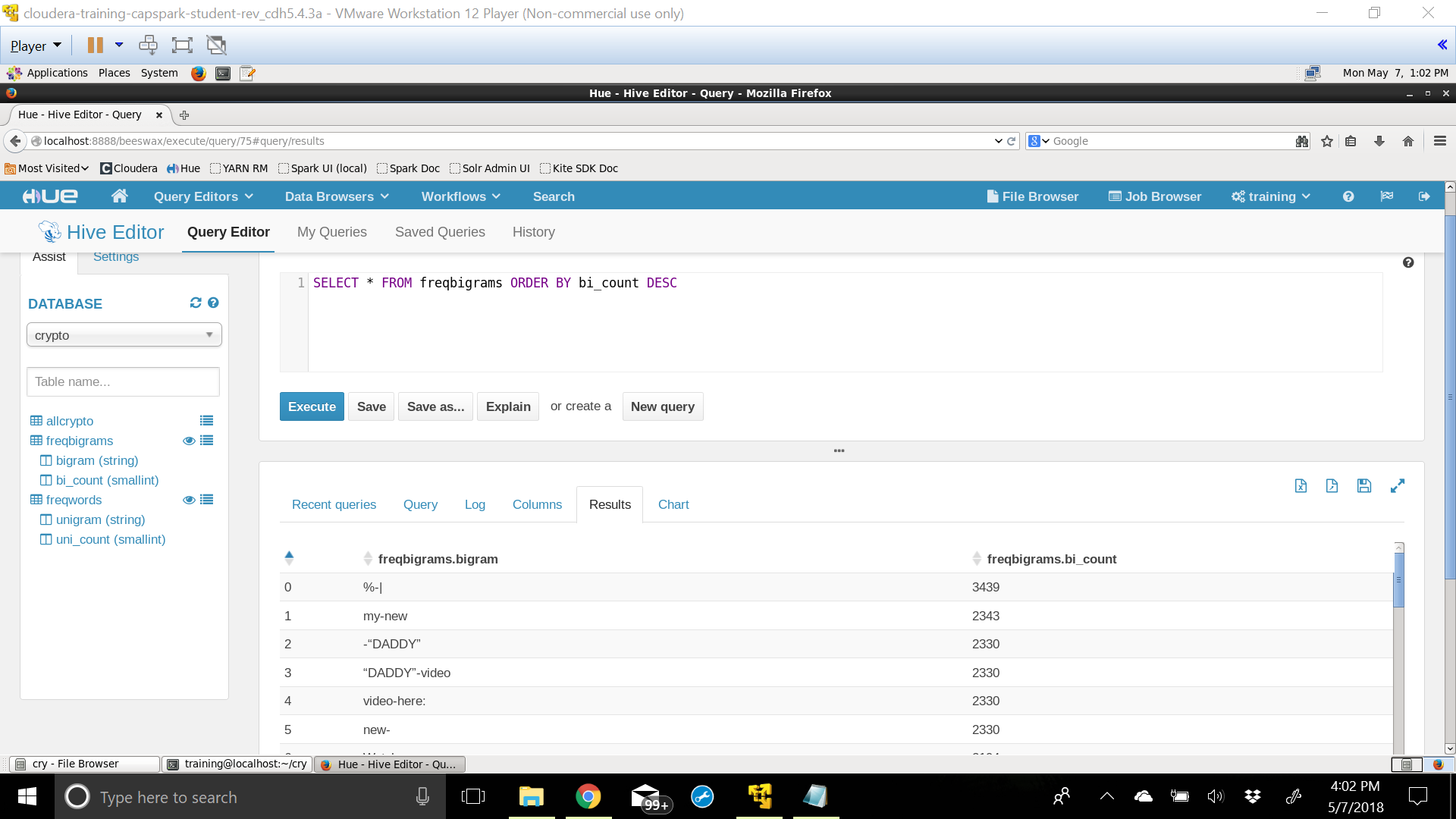
It can be seen that the most frequent word is ‘RT’ followed by ‘#bitcoin’.

1. **Finding the most frequent bi-grams**

MapReduce jobs were created in Python to find the most frequently used words from the data and tables were created in Hive.

****

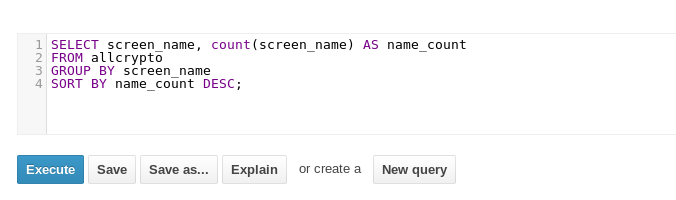
****

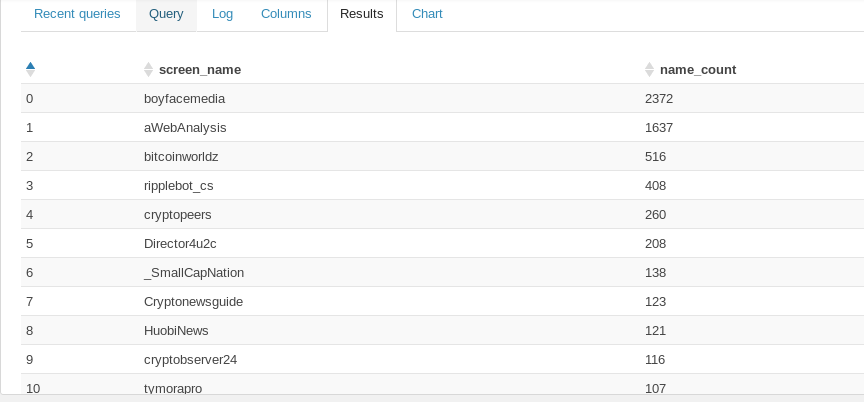
****

The bi-grams are separated by a hyphen. It can be seen that the most frequently used bi-grams ‘%-|’ and ‘my-new’. There is a room for some improvement by conducting Natural Language Processing to get better keywords.

1. **Finding the most active user**

The query below will find usernames and the number of tweets they have generated across all the data that was extracted from Twitter.





Hence, it can be seen that the user with the screen name ‘boyfacemedia’ is the most active user and has tweeted 2372 times.

**Applications**

The proposed system can have the following applications:

* Finding out the Twitter ID’s of those persons whose tweets are retweeted maximum number of times.
* Finding the most number of follows in the social networking sites.
* This system can be useful to track the business analysis of organizations.
* Allowing researchers to retrieve and analyze data easily from large datasets.

**Conclusion**

Traditional Enterprise Data Warehouses do not have the ability to keep up with rapidly increasing social media data. With this system, one can build a dashboard to monitor the sentiment of Twitter traffic around any given topic in near real-time allowing users to take advantage of near real-time Twitter sentiment for business insights.

In this report we have explained how some of the components of Cloudera’s Distribution of Apache Hadoop can be combined to create an end-to-end data management system. This architecture could be used for a variety of applications designed to look at Twitter data, such as identifying spam accounts, or identifying clusters of keywords. Taking the system even further, the general architecture can be used across numerous applications, such as analyzing web logs. Hence, we can conclude that processing time and retrieving capabilities are made very easy when compared to other processing and analyzing techniques for large amounts of data.

**Future Scope of Implementation**

Nowadays, the need for analyzing and processing of information has increased to a great extent. This report implemented the analyses of big data (tweets) only for text. Further analysis can be done for images and all types of multimedia files based on index support. The result of text mining and data analysis would help in suggesting related pages based on different types of data so that industries make the data easily available to people who are using and trying to access such type of data. Also, opinion mining can also be done on that data for finding polarity (positive, negative, neutral) of tweets collected.

**References**

1. <http://bogdanrau.com/blog/collecting-tweets-using-r-and-the-twitter-search-api/>
2. Manoj Kumar Danthala, “*Tweet Analysis: Twitter Data Processing Using Apache Hadoop*”, International Journal of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015, ISSN: 2348 9510
3. <http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/>
4. Sangeeta Yogi, “*Twitter Data Analysis Using FLUME & HIVE on Hadoop FrameWork*”
5. Ms. Pooja S. Patil1, Ms. Pranali B. sable, Ms. Reshma J. Fasale, Mr. P. A. Chougule, “*Sentiment Analysis on Twitter Data Using Apache Flume and Hive*”, International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 02 | Feb-2016, e-ISSN: 2395 -0056