**CHAPTER 1**

**INTRODUCTION**

We live in a society where the textual data on the Internet is growing at a rapid pace and many companies are trying to use this deluge of data to extract people’s views towards their products. Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional “pulse of the nation”, and indeed the global community. A great source of unstructured text information is included in social networks, where it is unfeasible to manually analyze such amounts of data. There is a large number of social networks websites that enable users to contribute, modify and grade the content, as well as to express their personal opinions about specific topics. Some examples include blogs, forums, product reviews sites, and social networks, like Twitter (http://twitter.com/). Twitter (San Francisco, CA, USA) is a micro blogging site that offers the opportunity for the analysis of expressed mood, and previous studies have shown that geographical, diurnal, weekly, and seasonal patterns of positive and negative affect can be observed.

Micro blogging and more particularly Twitter is used for the following reasons:

• Micro blogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people’s opinions.

• Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.

• Twitter’s audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.

• Twitter’s audience is represented by users from many countries

As the audience of micro blogging platforms and services grows every day, data from these sources can be used in opinion mining and sentiment analysis tasks.

**CHAPTER 2**

**Literature Review**

In the past years, many works has been released in sentiment analysis. Implementation of sentiment analysis has been carried out for a variety of applications over a wide range of classification algorithms and for varying data size. There exist many possible variants; some of them are discussed in following section

**2.1 Lin, Jimmy, and Alek Kolcz. "Large-Scale Machine Learning at Twitter." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804. ACM, 2012.**

This paper presents a case study of Twitter’s integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. The core of this work lies in recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specifically on supervised classification. In particular, the authors have identified stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data.

**2.2 Bian, Jiang, Umit Topaloglu, and Fan Yu. "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events" In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 25-32. ACM, 2012.**

In this paper, the authors describe an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing Natural Language Processing (NLP) and to build Support Vector Machine (SVM) classifiers. Due to the size nature of the dataset (i.e., 2 billion Tweets), the experiments were conducted on a High Performance Computing (HPC) platform using Map Reduce, which exhibits the trend of big data analytics. The results suggest that daily-life social networking data could help early detection of important patient safety issues.

**2.3 Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013.**

Machine learning technologies are widely used in sentiment classification because of their ability to “learn” from the training dataset to predict or support decision making with relatively high accuracy. However, when the dataset is large, some algorithms might not scale up well. In this paper, the authors evaluate the scalability of Naive Bayes classifier (NBC) in large-scale datasets. They have presented a simple and complete system for sentiment mining on large datasets using a Naive Bayes classifier with the Hadoop framework. Instead of using Mahout Library, they implemented NBC to achieve finegrain control of the analysis procedure for a Hadoop implementation.

**CHAPTER 3**

**OBJECTIVE**

The main objective of this project is to focus on how data generated from Social Media can be stored and utilized by different companies to make targeted, real time and informed decisions about their product that can increase their market share. This can be done by using Hadoop concepts. The given project will focus on how data generated from Social Media Websites can be analyzed and utilized. There are multiple applications of this project.

Companies can use this project to understand how effective and penetrative their marketing programs are. In addition to the view counts, subscribers and shares, audience retention count, companies can also evaluate views according to date range. This can tell the companies when the slow period or spike in viewership is and attribute the same to certain marketing campaign. Applications for Social Media data can be endless. For example, Companies can analyze how much a product people like.

This project can also help in analyzing new emerging trends and knowing about people's changing behavior with time. In addition, people in different countries have different preferences. By analyzing the comments/feedbacks/likes/view counts etc. of the videos, images uploaded, companies can understand what are the likes/dislikes of people around the world and work on their preferences accordingly.

**CHAPTER 4**

**SCOPE**

Social media is one of the popular media right now to share opinions or variety of topics and twitter is very popular social site to share everything related to opinions on variety of topics and discussions on current issues. These tweets, Status, Stories, Video and Image Uploaded, tags, like/dislike etc. generates the huge information related to different area like government, election, etc. millions of tweets, Status, Stories, Video and Image Uploaded, tags, like/dislike etc. are generated every day and which is very useful in decision making because everyone is share their view and opinions on issues or variety of topics. Twitter, Facebook, Instagram, LinkedIn and many more websites receives petabytes of data every day and these data is nothing but a collection of tweets so these data is very important in real life to analyze different scenario through which its helps us in decision making. The analysis of Social Media data gives real view or different user opinions regarding what they think and to analysis, these data provide a better way for making any decision.

"Twitter, Facebook, Instagram, LinkedIn and many more social media website has over a billion users and everyday people tweets, uploads hundreds of status and tag people and spends millions of hours on these websites and generate billions of views”. "Every day, people across the world are uploading 1.2 million videos, images and gif on these sites, or over 100 hours per minute and this number is ever increasing.

To analyze and understand the activity occurring on such a massive scale, a relational SQL database is not enough. Such kind of real time data is well suited to a massively parallel and distributed system like Hadoop.

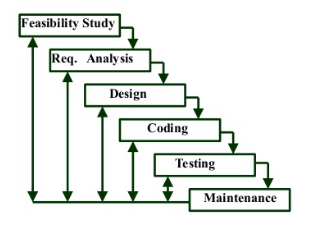
**CHAPTER 5**

**PROBLEM STATEMENT**

In today’s world lot of focus is on the study habits and study processes which improve the knowledge of the students, which are really good. But there are no approaches which mainly concentrate on the mental health of students. Having a specialized consultant at a college level decreases the mental stress by a certain level. With the advent of social media applications like facebook, twitter, etc have lot of emotions placed by various people at different age levels. This huge amount of data can be very useful for various conclusions like who is most liked actor, who is the most favorite politician etc., In this work student’s learning experiences are taken from twitter, facebook i.e., the emotions of the students and then analyze them to make decisions about the problems faced by students. So that steps can be taken to resolve those problems by using ‘*Hadoop and Pig*’. The proposed work is used to demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data analyzing techniques to explore engineering students’ informal conversations on Twitter, Facebook and other Social Media in order to understand issues and problems students encounter in their learning experiences.

**CHAPTER 6**

**SOFTWARE DEVELOPMENT METHODOLOGY**



**Figure: 1**

**WATERFALL ITERATIVE MODEL** contains 6 phases:

* Feasibility study: The feasibility study activity involves the analysis of the problem and collection of the relevant information relating to the product. The main aim of the feasibility study is to determine whether it would be financially and technically feasible to develop the product.
* Requirement analysis and specification: The goal of this phase is to understand the exact requirements of the customer and to document them properly (SRS)
* Design: The goal of this phase is to transform the requirement specification into a structure that is suitable for implementation in some programming language.
* Implementation and unit testing: During this phase the design is implemented. Initially small modules are tested in isolation from rest of the software product. Integration and system testing: In this all the modules are integrated and then tested altogether.
* Operation and maintenance: Release of software inaugurates the operation and life cycle phase of the operation.

**CHAPTER 7**

**SYSTEM REQUIREMENT SPECIFICATIONS**

**HARDWARE SPECIFICATIONS:**

* **PROCESSOR: Intel Core i3**
* **RAM: 4 GB**
* **HARD DISK: 500 GB**
* **MONITOR: LCD Monitor**
* **KEYBOARD: 101 keys or more**
* **MOUSE: Simple Mouse**
* **NETWORK: Internet**

**SOFTWARE SPECIFICATIONS:**

* **PLATFORM: Windows 10**
* **SOFTWARE: VMWARE Work Station**

**IBM Infosphere**

**Hadoop (HDFS and Map Reduce)**

**Pig**

**SOFTWARE SPECIFICATIONS**

1. **VMWARE Work Station:**

**V**Mware Workstation is a hosted [hypervisor](https://en.wikipedia.org/wiki/Hypervisor) that runs on [x64](https://en.wikipedia.org/wiki/X64) versions of Windows and Linux operating systems (an [x86](https://en.wikipedia.org/wiki/X86) version of earlier releases was available); it enables users to set up [virtual machines](https://en.wikipedia.org/wiki/Virtual_machine) (VMs) on a single physical machine, and use them simultaneously along with the actual machine. Each virtual machine can execute its own [operating system](https://en.wikipedia.org/wiki/Operating_system), including versions of [Microsoft Windows](https://en.wikipedia.org/wiki/Microsoft_Windows), [Linux](https://en.wikipedia.org/wiki/Linux), [BSD](https://en.wikipedia.org/wiki/BSD), and [MS-DOS](https://en.wikipedia.org/wiki/MS-DOS). VMware Workstation is developed and sold by [VMware, Inc.](https://en.wikipedia.org/wiki/VMware,_Inc.), a division of [Dell Technologies](https://en.wikipedia.org/wiki/Dell_Technologies). There is a free-of-charge version, VMware Workstation Player, for non-commercial use. An operating systems license is needed to use proprietary ones such as Windows. Ready-made Linux VMs set up for different purposes are available from several sources. VMware Workstation includes the ability to group multiple virtual machines in an inventory folder. The machines in such a folder can then be powered on and powered off as a single object, useful for testing complex client-server environments.

1. **IBM Infosphere:**

IBM Infosphere DataStage is an [ETL](https://en.wikipedia.org/wiki/Extract,_transform,_load) tool and part of the IBM Information Platforms Solutions suite and IBM Info Sphere. It uses a graphical notation to construct data integration solutions and is available in various versions such as the Server Edition, the Enterprise Edition, and the MVS Edition.

1. **Hadoop:**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop **MapReduce** is a software framework for easily writing applications, which process big amounts of data in parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The **Hadoop Distributed File System (HDFS)** is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data.

1. **PIG:**

**Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007. Pig consists of a language and an execution environment. Pig’s language, called as PigLatin, is a data flow language - this is the kind of language in which you program by connecting things together. Pig can operate on complex data structures, even those that can have levels of nesting. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs. PigLatin is not Turing complete on itself, but it can be Turing complete when extended with User-Defined Functions.**

**Execution environment: There are two choices of execution environment: a local environment and distributed environment. A local environment is good for testing when you do not have a full distributed Hadoop environment deployed. You tell Pig to run in the local environment when you start Pig’s command line interpreter by passing it the -x local option. You tell Pig to run in a distributed environment by passing -x mapreduce instead. Alternatively, you can start the Pig command line interpreter without any arguments and it will start it in the distributed environment. There are three different ways to run Pig. You can run your PigLatin code as a script, just by passing the name of your script file to the pig command. You can run it interactively through the grunt command line launched using Pig with no script argument. Finally, you can call into Pig from within Java using Pig’s embedded form.**

**CHAPTER 8**

**DATA FLOW DIAGRAMS**

A graphical tool used to describe and analyze the moment of data through a system manual or automated including the process, stores of data, and delays in the system. Data Flow Diagrams are the central tool and the basis from which other components are developed. The transformation of data from input to output, through processes, may be described logically and independently of the physical components associated with the system. The DFD is also know as a data flow graph or a bubble chart.

DFDs are the model of the proposed system. They clearly should show the requirements on which the new system should be built. Later during design activity this is taken as the basis for drawing the system’s structure charts. The Basic Notation used to create a DFD’s are as follows:

**1. Dataflow:** Data move in a specific direction from an origin to a destination.

**2. Process:** People, procedures, or devices that use or produce (Transform) Data. The physical component is not identified.

**3. Source:** External sources or destination of data, which may be People, programs, organizations or other entities.

**4. Data Store:** Here data are stored or referenced by a process in the System.

* **DATA FLOW DIAGRAM Level 0**

DFD level 0 data flow diagram shows the entire system as a single process, and gives no clues as to its internal organization.

Hadoop

Framework

Classification Result

Social Media Data

**FIGURE 2**

Data Collection from Social Media acts as an input where we read the data on the topics from twitter, Facebook, Instagram using hadoop. Hadoop used in designing the system are Open Authentication 1 and Open Authentication 2 along with secret key. Open Authentication 1 is used for registration to access the twitter account, Open Authentication 2 is used to access the account and the secret key is used to support encryption and decryption. The system is responsible for finding the best classification of tweets among five categories, namely Lack of Social Engagement, Sleep Problems, Diversity issues, Heavy Study Load Problems and Negative Emotions. It produces cleaned data, probability, contingency, and enhanced contingency as intermediate results and classification of tweets as final output.

* **DATA FLOW DIAGRAM Level 1**

Social Media Data

Probability Computation

Data

Collection

Social Media

Probability

*DATA STORE*

Classification Result

Analyzed

data

Classification Result

**FIGURE 3**

The DFD describes the various modules present in the system and also how the data flows from one module to another to complete the action of understanding tweets description before classification. Tweets for various cases acts as input to the data collection module. These tweets are collected from the twitter using Hadoop. For each of the tweets the stop words are removed and clean tweets are obtained. The Cleaned tweets are processed by probability computation module, here each tweet is compared against the words belonging to different categories. The Probability against each category is computed by using Naive Bayes algorithm. Ranking the problem module categorizes tweets into specific category based on maximum probability, and this is done to remove uncertainty if exists after cleaning the data.

* **DATA FLOW DIAGRAM Level 2**

Data Collected from Social Media Data

Comparison

& Compute the Data

Social Media Data

Social Media Data

Compute Data

*Data Store*

Result

Analyzed data

Classification of data into catagories

Compute the probability

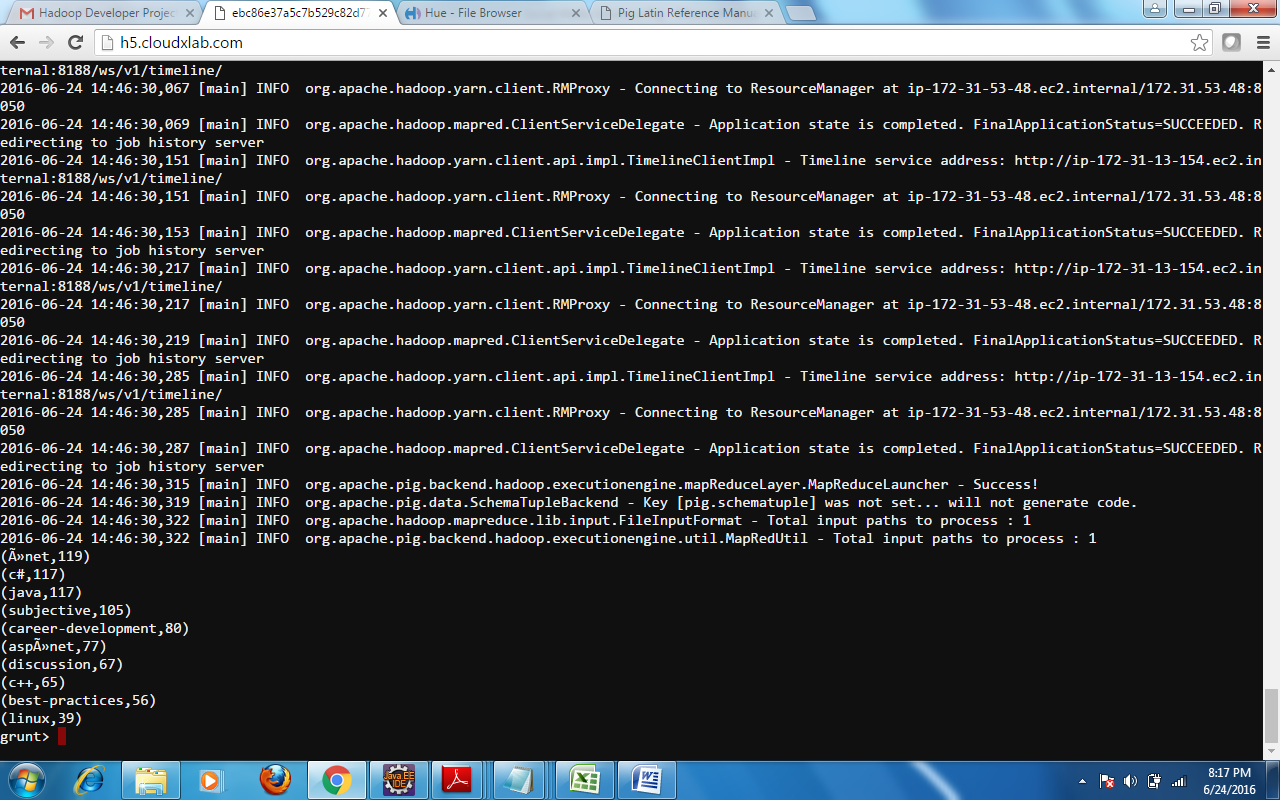
**FIGURE 4**

Data collection and cleaning module of DFD level 1 of student learning experience system is sub divided into 2 sub-systems namely data collection from Twitter and removal of stop words from Tweets. Removal of stop words from tweets is performed by using chi-square algorithm. Probability computing module is further divided into 4 sub modules, Comparison and Compute the Weight, Compute the Probability, Classification into categories and Compute the total weight for each category. Using the category weight ranking the problem module performs the categorization.

**CHAPTER 9**

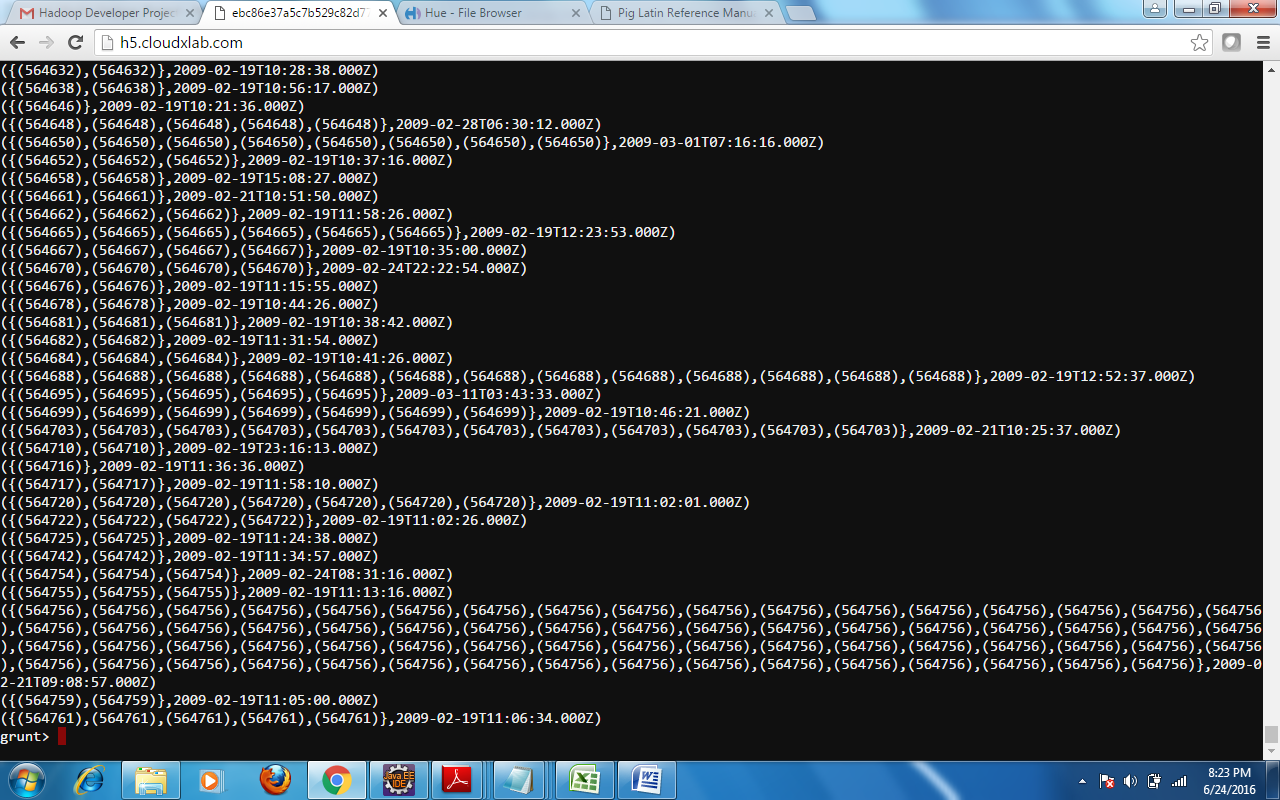
**OUTPUT**

**Question 1: Top 10 most commonly used tags in this data set**



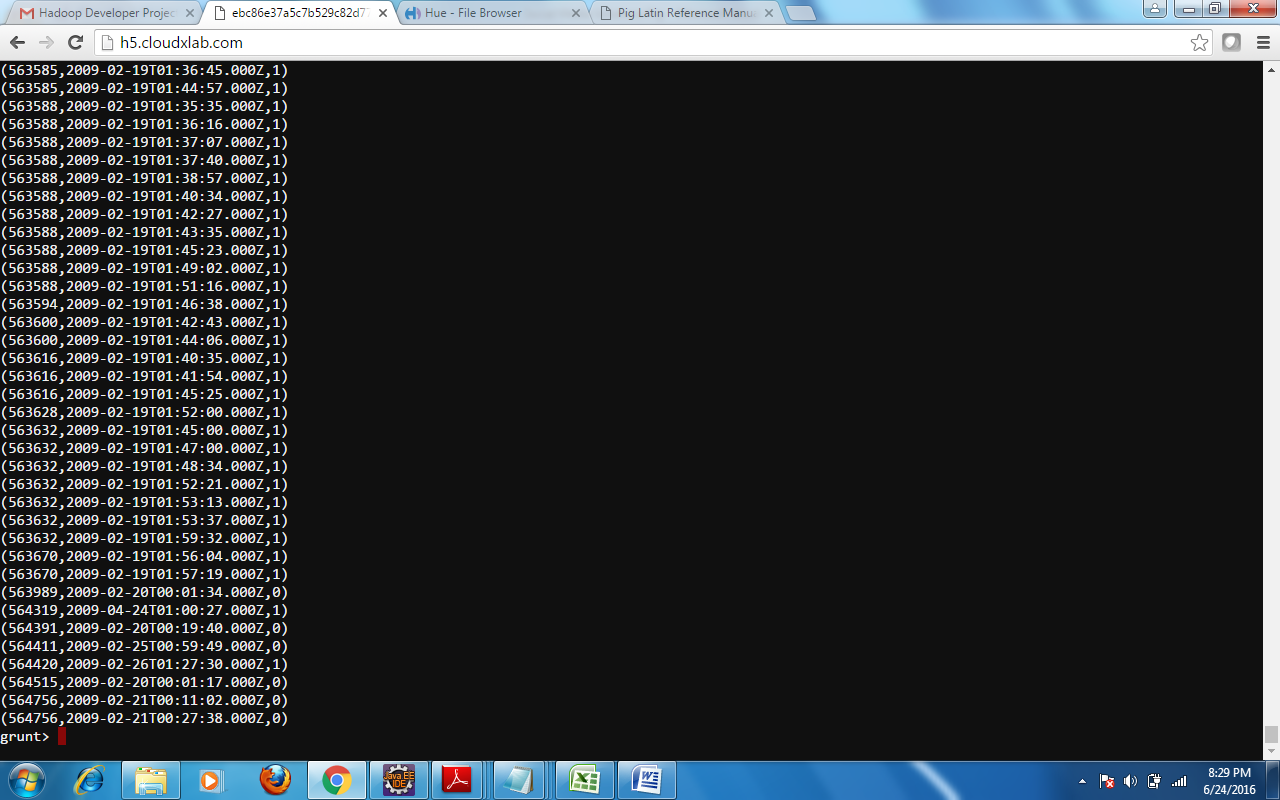
**FIGURE 5**

**Question 2: Average time to answer questions**



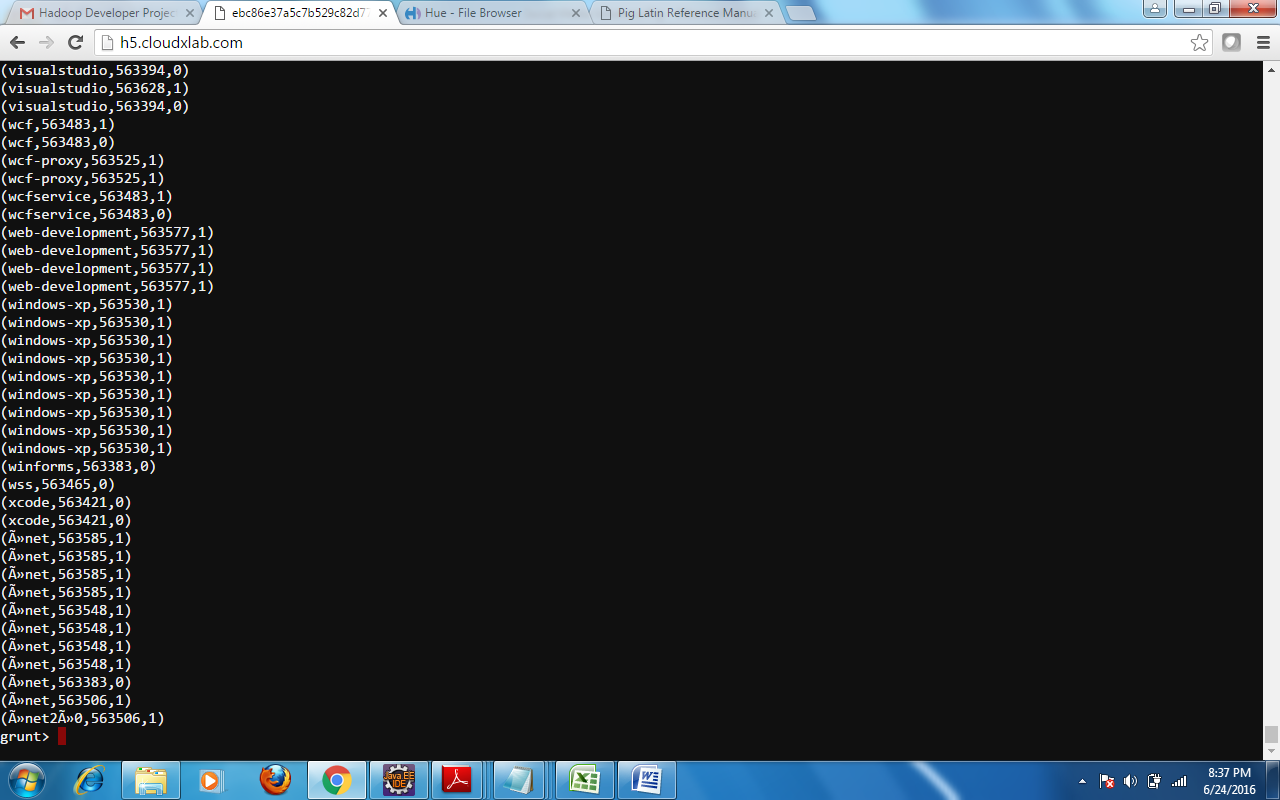
**FIGURE 6**

**Question 3: Number of questions which got answered within 1 hour**



**FIGURE 7**

**Question 4: Tags of questions which got answered within 1 hour**



**FIGURE 8**

**CHAPTER 10**

**References**

1. T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of HLT and EMNLP. ACL, (2005), pp. 347–354
2. A. Gates, “Programming Pig”, O'Reilley, (2011).
3. T. White, “Hadoop: The Definitive Guide”, Third Edition, O'Reilley, (2012).
4. S. Kumar, F. Morstatter and H. Liu, “Twitter Data Analytics”, Springer Science & Business Media, (2013).