School of Computing

Social Media Data Filtering and Analysis using MapReduce Programming Model

By

Umit Demirbaga

(160267235)

August 2017

Supervisor: Dr Rajiv Ranjan

MSc. Computer Science

Newcastle University

Declaration

This dissertation was carried out in accordance with the requirements for the degree of MSc Computing Science at Newcastle University.

I declare that this dissertation represents my own work except where otherwise stated.

Umit Demirbaga

August 2017

Abstract

Social media has been generating a large amount of data every minute because of its universal acceptance in recent years. Twitter, one of the most important social media sites, is a wealth of information for millions of users on the web. In this respect, the purpose of the studies on Twitter data analysis is to find valuable patterns and information from Twitter data.

Due to tweeting millions of tweets per minute, large data clusters are created. The analysis of these data requires efficient and advanced processing techniques. For this reason, the Twitter social media platform is suitable for implementing the MapReduce algorithm.

The aim of this project is to develop a concept to classify tweets about landslides. The focus will be on the Big Data architecture, which can manipulate Twitter data on clusters. For this reason, the MapReduce algorithm has been studied on Hortonworks' Hadoop platform for big data analysis. As a result of these studies, data clusters will be created based on the date of creation of tweets and the countries where the tweets are tweeted. In addition, the information will be visualised on Google Earth.

Keywords: Big Data, Hadoop, MapReduce, Data Filtering, Data Analysis, Social Media, Twitter.

Acknowledgements

After an intensive period of 11 months, while I am about to finish my Master's journey, I would like to express my gratitude to many people for their support and encouragement:

Foremost, I would like to express my sincere gratitude to my dissertation supervisor, Dr. Rajiv Ranjan, for his patience, motivation, enthusiasm, and immense knowledge throughout this study. He helped me with all the research and writing of this thesis. His feedback was very helpful, and studying with him was a pleasure. He is truly inspirational to his students. I am thankful to be a part of his team.

Besides my advisor, I would like to thank the doctoral students, Mr. Nipun Balan Thekkummal and Mr. Jedsada Phengsuwan, for the many stimulating conversations, the friendship they have offered me, and for helping me to improve my project.

I am highly indebted to the State of the Republic of Turkey and the Ministry of National Education for giving me this opportunity to complete my MSc, and thanks for their financial and emotional support. Allow me to express my sincerest gratitude for this opportunity they have given me. I value their trust in me and will work hard to maintain it.

A special thank you to my fiancée, Kübra Kırca. Words cannot describe how lucky I am to have her in my life. She has selflessly given more to me than I ever could have asked for. I love you and look forward to our lifelong journey.

I would also like to thank all my family for their wise counsel. You are always there for me. Thank you for your continuous support in many ways.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 About this project

Information technology takes place in almost every part of our lives, and every work and process made in the digital world leave a mark. Everything we do in daily life (using internet banking, taking a picture with my mobile phone and sharing it, sending a message to a friend) is a new data created by us. It is inevitable that the mass of data that is generated by the millions of people using the internet and changing the way of use increases day by day. In addition to personal use, storage, processing, and analysis of data generated in large institutions' systems is a big concept. Blogs, photos, videos, web server log files shared on social media are produced a very large data per day. According to the McKinsey Global Institute, data volume is growing 40% per year and will grow 44x between 2009 and 2020 (Dijcks, 2012). The evaluation of these data is very important. If there is no analysis of these data, they cannot go beyond being a growing data repository at any moment.

The devices that can be used to store data facilitate storing data, preferences, purchases, and habits of users, especially over the Internet. In this case, every movement of Internet users corresponds to a record in the database. All these developments have led to the rapid growth of data and the emergence of the Big Data concept. It is almost impossible to process Big Data with classical methods and derive meaning from these datasets, which has led to the emergence of Hadoop technology. With Hadoop, it is possible to process very large amounts of data quickly. An important part of Big Data is Social media. For this reason, the number of studies on social media analysis has been increasing daily.

This project studies Twitter data and aims to classify them using the MapReduce programming model on the Hadoop framework and then visualise them on Google Earth.

1.2 The reasons for this choice

Twitter generates millions of tweets daily because it is an environment where people can clearly express their feelings and thoughts about each topic. This data is very important and useful to conclude any subject. For example, information about any natural disaster event's

frequency, location, and date can be accessed via Twitter. By analysing millions of tweets, the desired information can be easily accessed.

For these reasons, this project aims to provide numerical and visual information about natural disasters. In light of this information, searching for information such as the location, time, and content of the tweets about the landslide occurring in various parts of the world is necessary. For this reason, the choice of this project has been made.

1.3 Expected achievements

The project is expected to first convert JSON files containing tweets under the landslip heading to CSV files, then group the tweet information in CSV files on the Hadoop framework using the MapReduce algorithm by the country code and creation date, and finally visualise- all tweets' information on Google Earth.

1.4 Aim

This project aims to use the MapReduce algorithm to make Twitter data analysable and to group the tweet information after filtering on Hadoop, and lastly after these operations, to visualise the obtained information on Google Map.

1.5 Objectives

To provide a practical solution for this project, the objectives consist of some clear and logical steps:

- Investigate and analyse whether the software and hardware of the computer that will work on the project meet the system requirements.
- Evaluate and test the accuracy of the analysed data and the visual system as a final stage.

1.6 Outline

The sections summarised below will be performed sequentially to achieve the objectives of the project:

- Chapter 2 contains detailed information on the background of the project.
- Chapter 3 presents information regarding the system requirements, design phases, the detailed implementation of the stage and the activities carried out.
- Chapter 4 presents the overall evaluation of the project and then makes a general assessment.
- Chapter 5 specifies whether the project's objectives are met or not, the positive and negative aspects of the project, and finally, how the project can be developed.

2 Background Research

2.1 Big Data

There is little doubt that the amount of data currently available is huge, but this new data is only one of the important features of the ecosystem. Big Data attracts attention not only because of its size but also because of its interest in other data. Due to data collection and collection efforts, Big Data is connected to the network. The value comes from a set of data that can be obtained between a piece of data about a person and groups of people or simply about the information structure (Boyd and Crawford, 2011).

Many purposes cause the concept of big data to become widespread. One of the most important of these aims is that companies want to take themselves further and reach more people. Besides increasing the profits and sales of the companies, there are also significant advantages provided by the big data in some areas, such as urbanism, policy, security, etc.

It is possible to encounter resources that constitute big data in every area of life. This massive data is produced even when driving, using the phone, shopping, travelling through stores and especially surfing the web. In addition, internet statistics, social media publications, blogs, microblogs, climate perceptions and similar perceptions can be given as examples.

When it is looked at the internet and social media usage statistics in the world (Chaffey, 2017);

• 2.5 billion people around the world use the internet. 1.8 billion of these users have accounts on social media networks.

• The number of active users of social media is increasing every year. Facebook maintains its leadership among social networks, with several active users of 1.860 billion.

• According to active user statistics, the top 10 most popular social media platforms are as follows:

1. Facebook (1.860 billion)
2. QQ (Tencent) (816 million)
3. Qzone (632 million)
4. WhatsApp (400 million)
5. Google+ (300 million)
6. WeChat (272 million)

7. LinkedIn (259 million)
8. Twitter (232 million)
9. Tumblr (230 million)
10. Tencent Weibo (220 million)

The Importance of Big Data

Big Data's main emphasis is on increasing the efficiency of using large volumes of data in different species. Organisations can get a better view when Big Data is properly identified and used accordingly. Thus, sales can gain efficiency in other areas, such as improving the manufactured product(Elena Geanina Ularu, Florina Camelia Puican, Anca Apostu, 2012).

Big data can be used efficiently in the following areas:

- To improve security and troubleshooting in information technology,

- To increase customer satisfaction in customer service using data from call centres,

- Developing services and products using social media content, knowing the preferences of potential customers

- To develop services and products using the content of social media of potential customers,

- Detecting fraud in online transactions.

2.1.1 Big Data Architecture

Big data architecture involves collecting, protecting, processing, and transforming data into file systems or database structures. There are many layers in big data architecture. Four of these logical layers are described below:

1- Big data sources layer: The data reach the organisations through this layer from company servers and sensors or any data providers, such as mobile devices, sensors, social media, and email. The company's sales records include all data such as customer database, social show channels, marketing lists, and email archives. Companies can use this data to reach everything they need and create new resources.

2- Data storage layer: This is where large amounts of data are gathered from sources after they are collected. If unstructured data exists, it stores them in a form that the tools can understand.

3- Analysis layer: The data must be analysed to get a desired result from the stored data. This layer works with stored data to extract data that would be useful for business intelligence. For this, MapReduce, a programming model, is used. MapReduce selects the data you want to analyse and insert a format that can be collected. Tools such as Apache PIG or HIVE are often preferred to analyse data.

4- Consumption layer: This layer receives and presents analysis results to the related output layer. In other words, the analysed data are transmitted to the relevant people for analysis.

Every layer consists of a few types of components, as shown below:

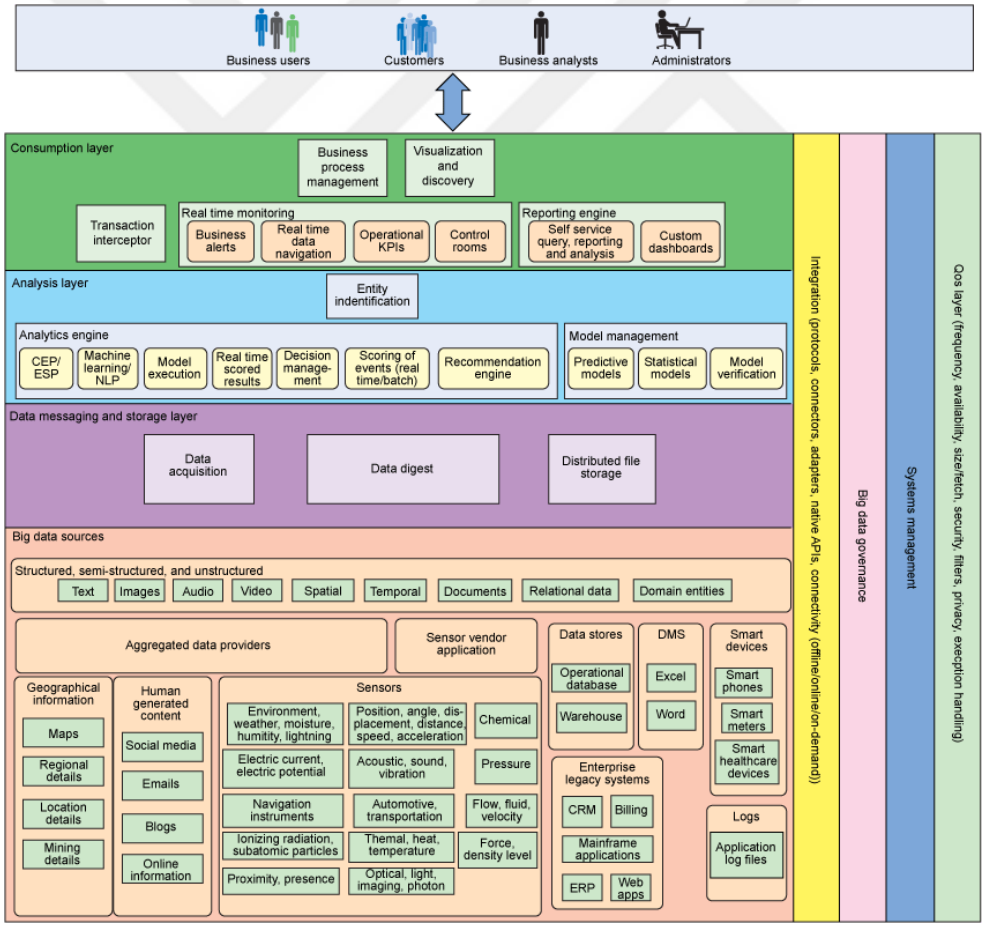


Figure 1 Logical layers of a big data solution (Mysore, Khupat and Jain, 2013)

As Marcos (2015) says, preparing data for analysis and processing is one of the most time-consuming and challenging tasks of analytics and investigative approach, and Big Data causes this problem. It requires influential methods to analyse big data records. Cloud analytics

solutions must consider enterprises' multiple Cloud deployment models (as cited in Guru and Suhil, 2015).

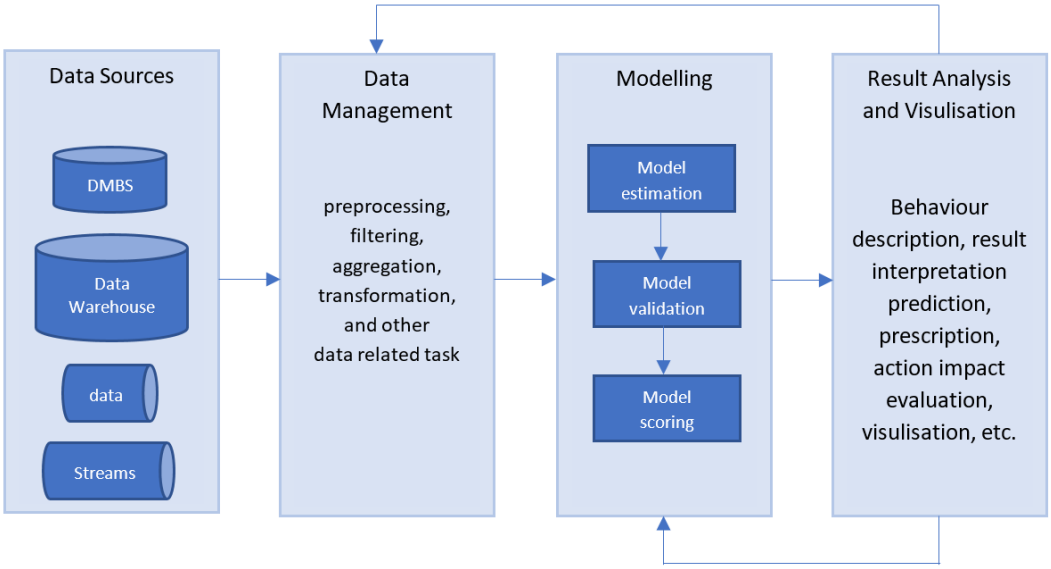


Figure 2 The Overview of the analytics workflow for Big Data (Assunção *et al.*, 2015)

2.1.2 The Structure of Big Data

Companies use various kinds of data. For this reason, data are categorised depending on the dimension of the data structure and the data source.

Depending on the dimension of the data structure, they are distinguished as structured (relational databases), unstructured (text from articles or email messages) and semi-structured data (XML, HTML-tagged text).

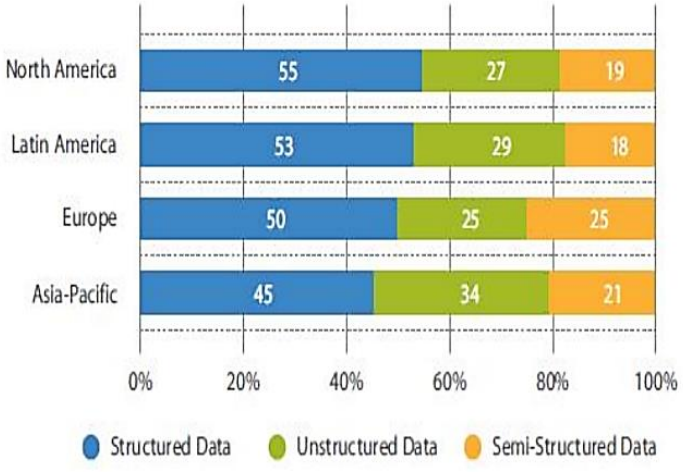


Figure 3 Data Structure Graphic (Nedelcu, 2013)

Additionally, depending on the dimension of the data source, they are distinguished as internal data (gathered from a company's sales) and external data (public social media sites).

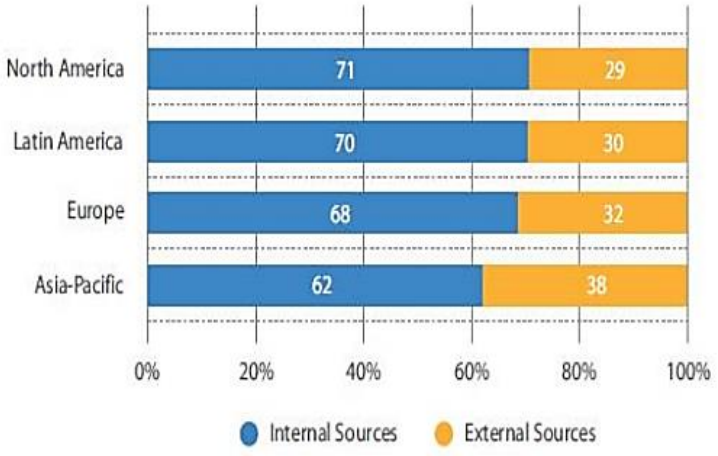


Figure 4 Data Sources Graphic (Nedelcu, 2013)

Results from studies in different parts of the world show that 51% of data is structured, 27% is unstructured, and 21% is semi-structured (Nedelcu, 2013).

2.1.3 Big Data Characteristics

Big data platforms enable data analysis in the virtual environment while simultaneously providing the ability to store more data at a low cost. At this point, Big Data characteristics, called 5V model, volume, variety, velocity, variability, and veracity, have become very important. So, there are five components in forming the Big Data platform. They are referred to as 5V using the initials, as shown in Figure 5.

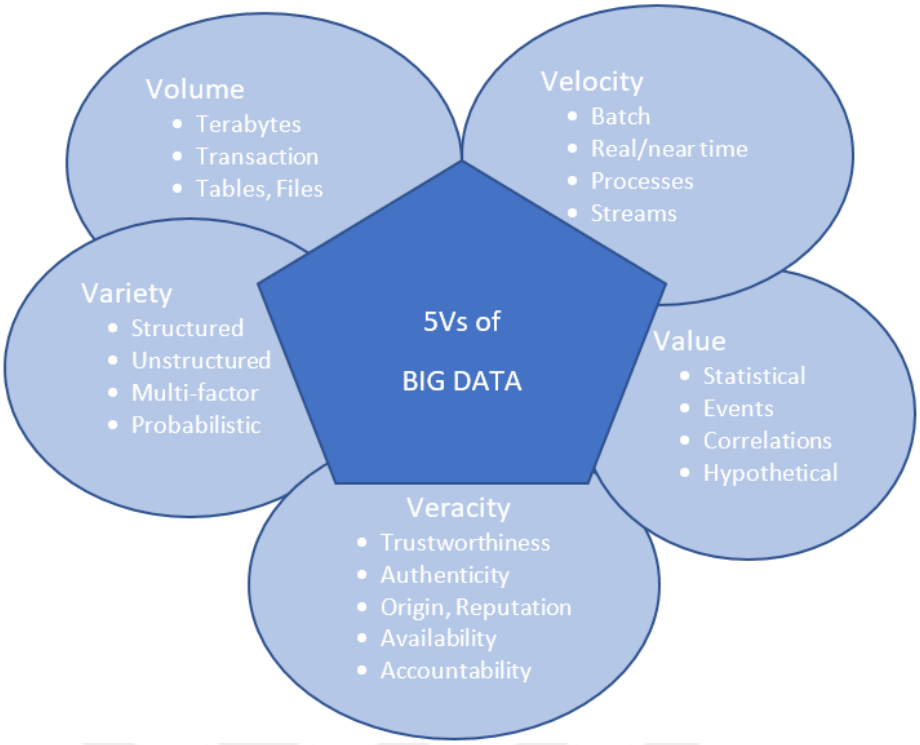


Figure 5 5Vs of Big Data

Volume

Traditionally, the data volume requirements for analytical and transactional applications were at terabyte levels. However, in recent years, more organisations in different industries have settled analytical data volume requirements for terabytes and petabytes. For example, on Facebook alone, more than 10 billion messages are sent daily (Dan Ariely, 2016). As Troester (2012) says, estimates produced by studies launched in 2005 show that the amount of data in the world has doubled every two years, and if this increase continues, it will be 50 times as much as it was in 2011 until 2020 (as cited in ISO / IEC, 2015). Other estimates show that 90% of the data generated was created in the last two years (NIST Big Data Public Working Group: Technology Roadmap Subgroup *et al.*, 2015).

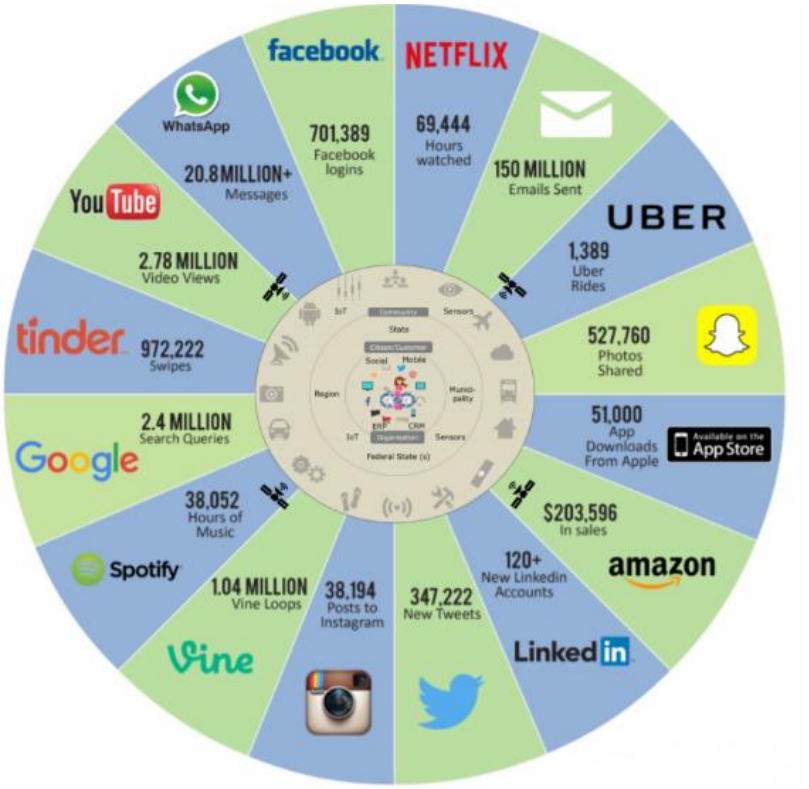


Figure 6 Volume of Big Data (Dan Ariely, 2016)

Big Data technology could be stored and used with the help of distributed systems, a collection of independent computers that are seen as a single coherent system for users. A clear and planned strategy for this is inevitable. Otherwise, the data cannot go beyond being a problem (Dan Ariely, 2016).

Variety

Data stacks are sometimes different from the same data type and structure. The diversity feature specifies data types and resources. Data sources can be very different. These are like mobile phones, peripherals or social media. These data, which come from other places as sources, also vary in form. Data forms are divided into structural, semi-structural, and non-structural data. We can give the example of structured data contained in the relational database or held in certain conditions. User feedback, email, or log files can be examples of semi-structured data. An example of unstructured data is image, audio, or text files. Today's research reveals that 95% of the data are unstructured, and the remaining 5% are structured (Cukier, 2010).

Figure 7 Variety of Big Data (Dan Ariely, 2016)

Velocity

Big data production is accelerating daily, and these data reach an incredible size now. Rapidly growing data reveals that the number of transactions that need that data and the multiplicity of the transactions increase simultaneously. Therefore, it is necessary to meet this density regarding software and hardware.

With the rapid development of technology, the speed of data generation has also increased. This increase in data rate and the use of persistent data sources such as detectors cause the data stacks to be very dynamic and variable. This makes it difficult to analyse and process the data stack. For example, let's consider a device that constantly controls the metabolism of a patient with a life-threatening condition and injects the drug into the patient's body accordingly. This tool should also decide on the next step while continuing to generate data. Perhaps because of this decision, the speed of human life is of great importance. Data processing speed has become very important due to using these and similar systems that facilitate human life in our daily lives. Amazon Web Services Kinesis is an example of a streaming application that handles the velocity of data (Williamson, 2017).

Variability

Variability refers to the state of the variability of the data. In this way, inconsistent data can be managed effectively. The data in different places of different systems are combined. Data coming from other sources cause a complex data size in different species. Big data provides rapid management of this complex data. In addition, variability indicates a change in data flow rates. Often, big data velocity is inconsistent and has periodic peaks that are constantly changing (Williamson, 2017).

Veracity

The veracity of the Big Data indicates the quality of the data. Sometimes, it is called validity or variability by addressing the data's lifetime. Because big data might be noisy, imprecise, erratic, and full of anomalies. If the data are not correct, it is not valuable. The quality of the results of large data analysis shows only the amount of data analysed. So, the evidence provided is useful and valid depending on the fact that the data have a qualitative level (Altintas and Gupta, 2017).

2.1.4 Big Data and Cloud Computing

What is Cloud Computing?

With the invention of information technology, storing and accessing big data on the internet has become possible. Thanks to this convenience, a big database based on Industry 4.0 has become applicable to the industry. The foundations of the cloud computing idea are based on the 1950s. Amazon, one of the information technology circuits, put its Amazon S3 into service in 2006 by improving its data centres (TechCrunch, 2017). Gartner, Consulting and Research Company has highlighted the importance of cloud computing, which can change the relationship between users and suppliers (Gartner, 2008). Since 2008, cloud computing has begun to be used extensively worldwide and continues to grow rapidly.

With the development of technology, data storage capacity has become a problem as users want to store more personal data. For this reason, users started to prefer faster and higher-capacity devices. This caused an increase in the prices. Cloud Computing has been shown as a solution to all these problems. Thanks to Cloud Computing, even with devices with low speed and

capacity, any information or personal data can be reached in a short time. For all these operations, it is necessary to establish a multi-server connection via a digital network.

Cloud Computing Services Models

Cloud computing is based on four basic models.

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)
- Communication as a Service (CaaS)

Software as a Service (SaaS): It is a software service that provides programs, such as CRM, ERP, finance and accounting software, that users need through the cloud (Rao, Leelarani and Kumar, 2013). For companies giving services in different locations, SaaS offers significant economic advantages without extra software costs. The best example of SaaS is Gmail. With this service provided by Google, sending mail, editing documents and backing up the files can be done easily without any software knowledge.

Platform as a Service (PaaS): A platform service that allows developers to develop their projects by providing hardware and software layers. This service offers platforms like system management, operating systems, programming language environments, and databases. Since the service provider performs system administration, users manage only their applications and data (Rittinghouse and Ransome, 2010). For example, if a software is developed with PHP, the software developer does not have to deal with the SQL and web server infrastructure of this software. PaaS only provides the platforms your software needs to work with.

Infrastructure as a Service (IaaS): It is an infrastructure service that is the basic service of Cloud Computing. IaaS creates a virtual server and provides a cloud server service to the users. With the cloud infrastructure, the virtual server resources are dedicated to the user, and the user has a flexible infrastructure with IaaS. For example, bus companies need more resource servers during the holiday season. The resources used can be increased or decreased when requested, thanks to the structure of Cloud Computing.

Communication as a Service (CaaS) is a solution for outsourced enterprise communication. CaaS manages the hardware and software needed to provide cloud-based solution providers,

voice-over IP (VoIP), instant messaging (IM), and video conferencing services. This model has begun to evolve from the telecommunication (Telco) industry and is no different from the emergence of the SaaS model in the software distribution services industry. All hardware and software management is consumed by the user base under the responsibility of CaaS vendors (Rittinghouse and Ransome, 2010).

Cloud Computing Deployment Models

Cloud hosting deployment models represent the entire category of the cloud environment and are often distinguished in three ways: ownership, size, and access. Cloud explains the purpose and nature of computing. Most organisations prefer implementing the cloud system because it reduces capital expenditures and controls operating costs. It is necessary to know the four distribution models to know which one meets the organisations' website requirements.

Public Cloud: General cloud applications are offered to the public by a service provider, like storage or other resources. It is deployed where available to the people and off-site over the internet. A public Cloud provides resources that are shared with high efficiency. These services are either free or charged with a pay-per-use model. Generally, cloud providers such as AWS, Microsoft and Google do not provide direct links; they only provide access via the internet.

Private Cloud: A private cloud is a cloud infrastructure that is run only for a single organisation, can be managed internally or by a third party, and can also be hosted internally or externally (National Institute of Science and Technology, 2013). When a private cloud project is undertaken, a certain level and contract level must be specified for the virtualisation of the business environment. In addition, the organisation must re-evaluate its existing resources. When done correctly, it may positively affect the business, but the steps in the project are in danger of being seriously damaged. The private cloud identifies the security issues that must be addressed to block these possibilities. Furthermore, it is a cloud technology preferred by companies with significant knowledge. All information can only be seen by the administrator. Access is secure, and privacy is high. Microsoft provides this with the help of Hyper-V and the System Center Product Family.

Hybrid Cloud: A hybrid cloud combines two or more clouds. These different clouds are independent but depend on each other, thus offering the possibility of multiple placement modes (National Institute of Science and Technology, 2013). Hybrid cloud architecture

provides companies and individuals with immediate local availability without needing an Internet connection. It also reduces the likelihood of making mistakes. Hybrid cloud architecture requires both in-house resources and external server-based cloud infrastructure. In hybrid clouds, in-house applications should be flexible, secure and specific. Hybrid cloud providers provide the flexibility of in-house applications with fault tolerance and scalability of cloud-based services. The hybrid cloud can be deployed with critical applications in the private cloud, while less cloudy applications can be deployed in the cloud.

Community Cloud:

It is cloud technology that hosts services shared with a few companies. In other words, community members have access to all practices and data. It is used by different organisations which have the same aims. It can be managed internally or by a third party and hosted internally or externally. As an example of this cloud technology, government agencies can access information about each person from a common cloud.

The Relationship Between Big Data and Cloud Computing

The large data term relates to the fact that data clusters are so large that they cannot be stored and analysed by typical database systems. The data is very big because it is no longer conventional structured data. These data are provided in many new and different sources like e-mail, social media, shopping on the internet, credit card usage and internet-accessible sensors. (McKinsey & Company, 2011). Large data also leads to challenges, such as data storage and data analysis for businesses.

Network-attached storage is a typical mode for internally storing big data (Sliwa, 2011). A network-attached storage (NAS) pod consisting of multiple computers connected to a computer used as a (NAS) device creates the first step of the configuration. Several NAS pods are interconnected using a computer used as a NAS device. Clustered NAS storage is an expensive option for small and medium-sized businesses. A cloud service provider can provide The required storage space at a lower cost (Purcell, 2013).

Cloud computing provides an environment to implement big data technology. Big data offers many benefits for businesses in decision-making support and innovation in business models, products, and services (McKinsey & Company, 2011). Small and medium-sized businesses prefer to use cloud computing for large data-processing applications because it provides a great

advantage in reducing hardware and processing costs and testing the value of large data sets. The biggest concerns about cloud computing are a security vulnerability and inspection deficiency (Géczy, Izumi and Hasida, 2012).

2.2 Apache Hadoop

Hadoop is inspired by the Google File System and MapReduce programming model, which is the work of Google's company. After the Google company published its work describing the Google File System and MapReduce programming model in 2003 and 2004, the people in the open-source community led by Doug Cutting used these two applications in the Nutch Search Engine. They have completed many applications. In 2006, Google File System and MapReduce programming models were named themselves Hadoop, and within the same year, Doug Cutting began working for Yahoo. Hadoop was an invention that would meet the needs of Yahoo and Google companies. These companies had to buy servers with more powerful hardware and larger data storage units constantly because of the ever-growing amount of data on the internet. With this new structure, instead of purchasing a new server, it was aimed that the work to be done would be separated and processed in parallel. A middle-level server could be added to the system if the resources were insufficient. Yahoo is the largest Hadoop cluster that has 4,500 nodes and more than 100,000 CPUs in over 40,000 servers. Yahoo stores 455 petabytes of data in Hadoop (Ashish Bakshi, 2016). Many popular companies like Amazon, eBay, Facebook, and LinkedIn use this technology to analyse large amounts of data (Hadoop Wiki, 2017).

Hadoop, developed by the Hadoop Open Source Software Team after Google launched it, has become available in many different areas. Unlike conventional database systems, Hadoop is designed to scan over large amounts of data and produce results quickly, thanks to the distributed architecture (Zikopoulos *et al.*, 2013).

2.2.1 Hadoop Architecture

Hadoop is designed to handle structured and non-structured data in petabyte size. Hadoop works with the cluster structure that ordinary servers come together. Hadoop can detect system breakdowns instantly and work without interruption by automatically applying the changes (Judith Hurwitz, Alan Nugent, Dr. Fern Halper, 2013).

The following five key features have been considered when constructing Hadoop's system architecture (Nemschoff, 2013):

1. Scalable: Although Hadoop has hundreds of servers running in parallel, it can add a new node point without changing the data, unlike traditional relational database systems (RDBMS).

2. Cost Effective: Handling large amounts of data with traditional relational database management systems is extremely costly. The CPU power required will also increase as the amount of data increases. With a cheaper server infrastructure, Hadoop provides the high processor power to operate high-capacity data.

3. Flexible: With Hadoop architecture, structured and unstructured data can be easily processed from any source. Data in different structures coming from other sources can be combined. And businesses can easily access new data sources. In addition, Hadoop can perform various transactions such as suggestion systems, data warehouses, market campaign analyses, and fraud detection.

4. Fast: Hadoop has a distributed file system that stores the data by maps technique regardless of where it is located on a cluster. Data processing tools and data are usually the same servers. For this reason, data processing is very fast. Hadoop can process terabytes of data efficiently in a matter of minutes.

5. Resilient to Failure: It is one of the most important advantages of Hadoop. While the data is being sent to a node, it is also copied to the other nodes in the cluster. In the event of a node failure, another copy is available.

Apache Hadoop provides the application infrastructure to process large amounts of data on the cluster structure created using ordinary servers without any limit on the number of lines. The Hadoop framework application works through an environment that provides distributed storage and computation among computer clusters (tutorialspoint, 2017a).

Hadoop consists of MapReduce, the Hadoop distributed file system (HDFS) and several related projects, such as Apache Hive HBase. MapReduce and Hadoop distributed file system (HDFS) are the main components of Hadoop.

2.2.2 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is designed to reliably store very large data clusters and stream these data clusters to user applications with a high bandwidth (Shvachko *et al.*, 2010). Thousands of servers in a large group have direct connectable storage space, and these servers execute user application tasks. Storage and computing are done by distributing too many servers, and demands can increase this number. Thanks to a distributed file system and a framework presented by Hadoop, very large data sets can be easily analysed and transformed using the MapReduce paradigm. Dividing data and calculations between thousands of computers and running application calculations in parallel with the data is the most important feature of Hadoop. Hadoop clusters at Yahoo! store 25 petabytes of application data spread across 25,000 servers. The largest is the 3500 server (Dagli and Mehta, 2014).

HDFS Objectives

HDFS has many remarkable features\ such as fault tolerance, data access, consistency, portability, scalability, economy, efficiency, and reliability. With these features, HDFS aims to the following titles:

Hardware failure: Each of the thousands of servers that make up HDFS stores a portion of the file system data. Some components of HDFS are not always functional because of the large number of members in HDFS and the impossibility that each piece may fail simultaneously. For this reason, it is a basic architectural goal of HDFS to detect errors and provide fault tolerance by applying a quick auto-recovery.

Streaming data access: Unlike general-purpose applications that work in general-purpose file systems, while applications in HDFS are running, they need streaming access to their data sets. HDFS is focused on batch processing rather than interactive use by users. The main purpose is to achieve high efficiency in data access.

Large data sets: With the feature of supporting very large files, HDFS allows very large data sets to run on it, and high data bandwidth is required for each cluster to help millions of files and must be scaled to many nodes.

Simple Coherency Model: A write-once-read-many access model for files is needed for HDFS applications. Because after creating or writing a file, there is no need to change it. Thus, problems related to data consistency are simplified and high throughput data access is ensured.

HDFS Architecture

HDFS consists of NameNode and DataNode sub-services.

NameNode: NameNode is the controller process responsible for the distribution of blocks on servers, creation, deletion, reconstruction when a problem occurs in a block, and any file access. In short, information about all files on HDFS (metadata) is stored and managed by NameNode. Only one NameNode can be in each cluster. The HDFS namespace is a factor of files and directories. Inodes on the NameNode represent files and directories. Some information, such as permissions, modification and access times, namespace is recorded by inodes. The file content is usually split into blocks of 128 megabytes, but the user can specify this capacity according to the file (Shvachko *et al.*, 2010).

DataNode: It is an agent process that keeps function blocks. Each DataNode is responsible for the data in its local disk. It also contains backups of data from other DataNodes. DataNodes can be more than one in a cluster (Patodi, 2011).

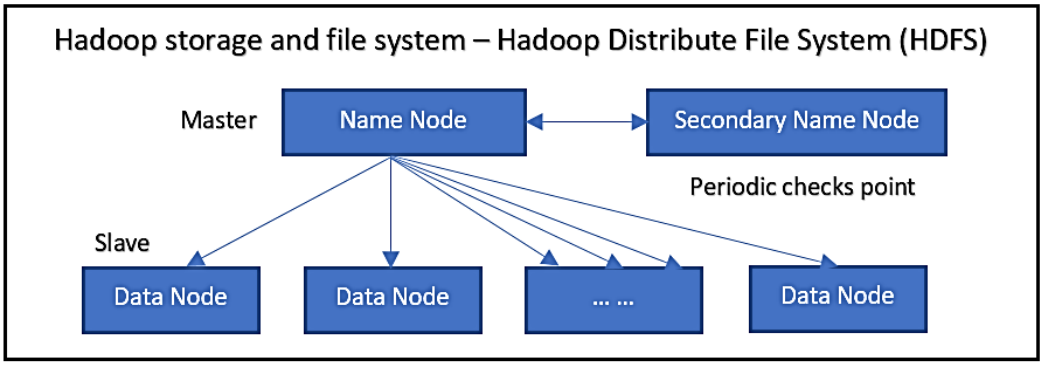


Figure 8 The components (daemons) of HDFS

Hadoop combines disks from ordinary servers to create a large and single disk space thanks to its file system HDFS. Thus, many large files can be safely stored and backed up in this file system. These files are backed up in a structure like RAID by being distributed on multiple and different servers in blocks. By this means, data loss is prevented.

2.2.3 MapReduce

MapReduce is a way to handle large files on HDFS. The program, which consists of the Map function used to filter the data you want and the Reduce functions that allow you to get results from these data, is run on Hadoop after it is written. Hadoop is responsible for distributing the Map and Reduce threads over the cluster, processing them simultaneously, and reassembling the resulting data (Dean and Ghemawat, 2004).

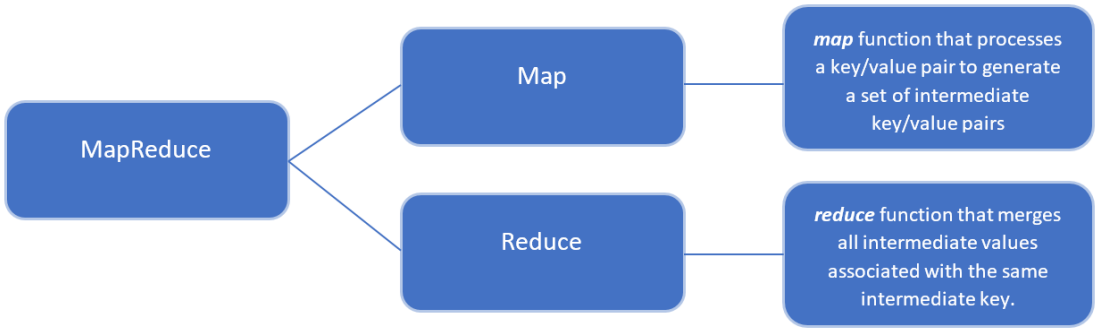


Figure 9

The power of Hadoop comes from the fact that the processed files are always linearly scaled by reading the corresponding node from the local disk, not engaging in network traffic, and handling multiple jobs simultaneously. As the number of nodes in the Hadoop cluster increases, the performance also increases linearly, as in the graph below.

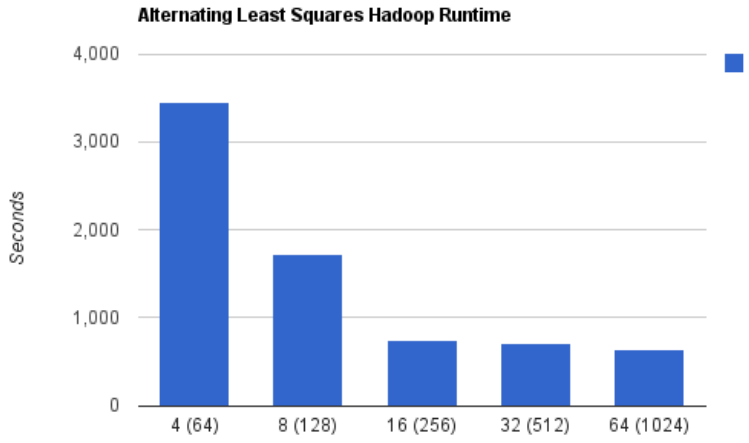


Figure 10 Number of HPC machines (x16 cores) (Patodi, 2011)

One of the advantages of Hadoop is that the traffic of the files distributed with Map does not go through the network. Another advantage is that it can process more than one job simultaneously and complete the result with a single Reduce operation.

MapReduce consists of JobTracker and TaskTracker processes. JobTracker is responsible for running the MapReduce program written on the cluster. It is the responsibility of JobTracker to terminate or restart that thread in any problem that may arise during the execution of distributed threads. TaskTracker runs on servers with DataNodes. JobTracker requests the workpiece to be completed. With the help of NameNode, JobTracker gives TaskTracker the most appropriate Map job based on the data on the local disk of the DataNode. The threads given in this way are completed. The result is written as a file on HDFS, and the program is terminated.

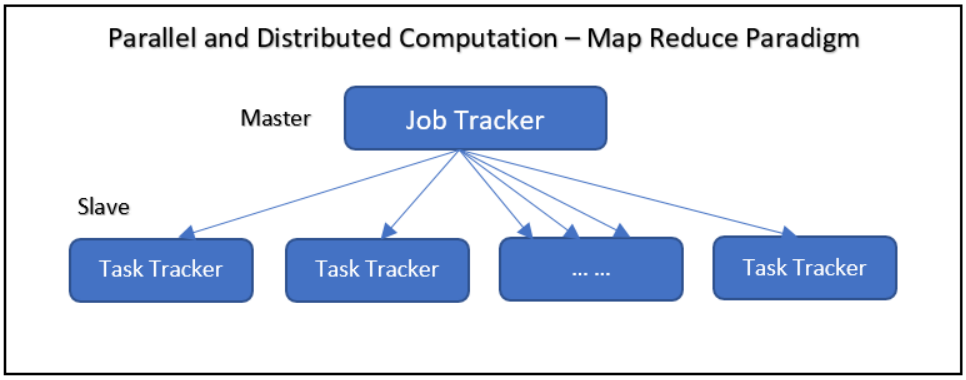


Figure 11 The components (daemons) of MapReduce

The processes of MapReduce

1. The client submits the job to JobTracker.
2. JobTracker asks NameNode for the location of data.
3. As per the reply from NameNode, the JobTracker asks respective TaskTrackers to execute the task on their data.
4. All the results are stored on some DataNode, and the NameNode is informed.
5. The TaskTrackers inform the job completion and progress to JobTracker.
6. The JobTracker informs the completion of the client.
7. The client contacts the NameNode and retrieves the results.

2.3 Hortonworks Data Platform (HDP)

Hortonworks Data Platform (HDP), an open-source Apache Hadoop distribution, includes built-in security, governance, and operations. HDP contains a variety of tools used in big data applications. It consists of data management, data access, security and handling of transactions.

HDP combines the power and cost-effectiveness of Apache Hadoop with several features, as indicated below. It makes it easier and more efficient for Hadoop to be successfully installed and managed in an enterprise environment (Hortonworks, 2016).

- Integrated and Tested Package
- Easy Installation
- Management & Monitoring Services
- Data Integration Services
- Centralized Metadata Services

Hortonworks Data Platform Components

The Hortonworks Data Platform includes components that make it easy to set up, configure and monitor Apache Hadoop components for big data projects. These components are shown in the figure below:

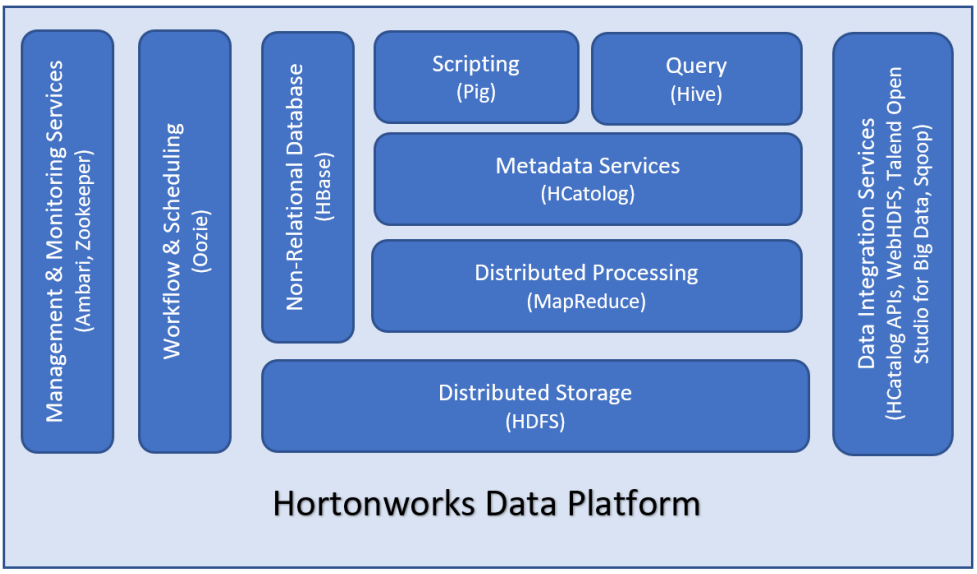


Figure 12 The members of Hortonworks Data Platform

2.3.1 Hortonworks Sandbox with VirtualBox

When the Big Data term first emerged, it was challenging to find an effective way to start working with Hadoop without trusting professional services or staff because of the complexity and rapidly changing nature of Hadoop and its ecosystem.

Hortonworks indicates that Sandbox is a personal and portable Hadoop environment that contains many interactive tutorials and the most exciting developments from the latest HDP distribution packaged in a virtual environment (Hortonworks, 2017).

Sandbox is provided as a self-contained virtual machine for the following three virtualisation environments (Hortonworks, 2015):

1. Oracle VirtualBox
2. VMware Fusion or Player
3. Microsoft Hyper-V.

After downloading the Sandbox VM image, it is needed to import the appliance into VirtualBox. To start a Sandbox session, it is required to open a web browser and enter a preconfigured IP address (e.g. http://127.0.0.0:8888/) after the Sandbox VM is created.

Hortonworks Sandbox is a very effective learning environment for beginners to Hadoop. A user can learn the detailed mechanics behind the scenes after a period of time, thanks to Sandbox's interactive and easy-to-use training environment. In addition to this, it offers a rich set of video, graphical, and text-based instructions when informative feedback and errors occur during exercises (Hsieh *et al.*, 2014).

Having a 64-bit system with hardware support and at least 4 GB of RAM is enough to run Sandbox.

2.4 Apache Mahout

Apache Mahout is an open-source project developed by the Apache Software Foundation (ASF) to create scalable machine-learning algorithms. Mahout includes implementations for clustering, categorisation, CF, and evolutionary programming. In addition, using the Apache Hadoop library enables Mahout to scale effectively in the cloud (Ingersoll, 2009).

Mahout is an open-source project to implement all kinds of machine learning techniques and focuses on three key areas of machine learning: recommender engines, clustering, and classification.

Recommender engines

The Recommender engine is the easiest recognisable machine learning technique. Many sites try to recommend products, movies or music related to visitors' past actions, such as Amazon, Netflix, Facebook, Instagram, etc.

These recommendation lists are produced with the help of recommender engines. Mahout provides recommender engines of several types such as (Ingersoll, 2009):

- **User-based recommenders:** Recommends items by finding similar users. Scaling is generally harder due to the dynamic nature of users.
- **Item-based recommenders:** Calculate the similarity between items and then make recommendations. This could be computed offline because items generally do not change much.
- **Slope-One:** A quick and simple item-based recommendation approach that can be implemented when users rate.
- **Model-based:** Provides recommendations on improving a user's model and ratings.

Clustering

It is often useful to group or cluster large items of data automatically, whether text or numeric. For example, daily sports news can be automatically grouped according to the specified categories. So, the reader can read the information without passing on non-interest sports.

Clustering's job is to group together similar items by calculating the similarity between items in the collection. In most clustering implementations, items in the collection are generally represented as vectors in an n-dimensional space. When vectors are considered, the distance between two elements can be calculated using measures such as Manhattan Distance, Euclidean distance, or cosine similarity. Enterprises might use this technique to discover hidden groupings among users or create a large collection of documents (Ingersoll, 2009).

Classification

Classification techniques determine whether something is in a type or category. The purpose of classification is to label invisible documents so that they can be put together. Like clustering, classification learns by examining many category elements to derive classification rules. Classification helps to decide whether a new entry should fit a previously observed pattern. It is often used to classify unknown behaviours or patterns. Classification features include words, weights of words based on frequency, and parts of speech. With these features, a document can be associated with a label. It can also be used to detect fraud on the Internet.

A few of the real-life examples are mentioned below.

- Classification is used by the iTunes application to prepare a new playlist.
- Gmail uses this technique to decide whether a new post is spam. First, the categorisation algorithm trains itself to mark certain posts as spam, depending on user habits. Based on this training, it then decides whether a new mail should be left in the inbox or spam groups.

As mentioned earlier, one of Mahout's purposes is to implement these techniques that can be scaled up to huge input (Ingersoll, 2009).

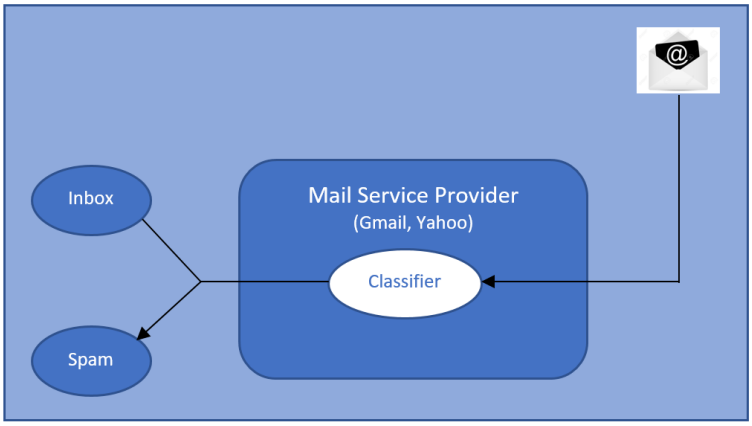


Figure 13 Classification for spam mail

How Classification Works

While the classifier system classifies a particular data set, it performs the following operations in order (tutorials point, 2017b):

1. Firstly, a new data model is prepared by the selected learning algorithm.
2. The accuracy of the prepared data model is tested.
3. Based on this data model, new data are evaluated, and their classes are determined.

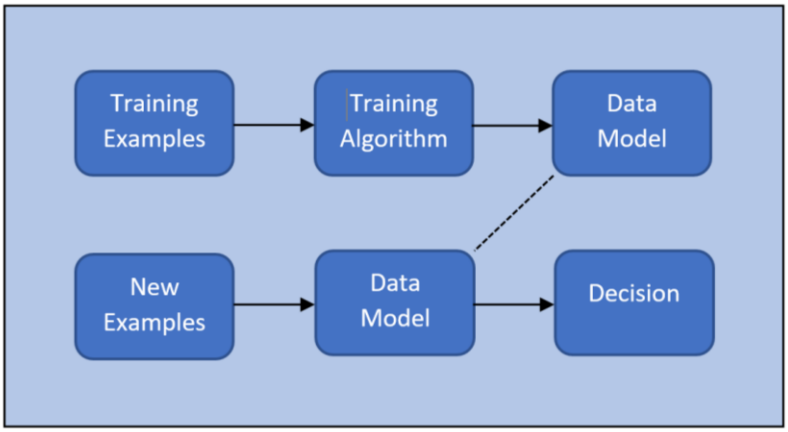


Figure 14 Working Principle of Classification

Algorithms Supported in Mahout

In Mahout, the algorithms are implemented in two different ways: Sequential Algorithms are algorithms which are executed sequentially and do not use Hadoop scalable processing, such as user-based collaborative filtering, logistic regression, Hidden Markov Model, multi-layer perceptron, singular value decomposition. Parallel algorithms reduce parallel processing by using the Hadoop map. For this reason, they easily support high-capacity data at the petabyte level (Press, Anderson and Jacobson, 2015). Random Forest, Naïve Bayes, canopy clustering, k-means clustering and spectral clustering are examples of this algorithm.

2.5 Natural Language Processing (NLP)

The NLP is a model and methodology developed in the United States about the functioning of the mind due to research on how we perceive, think, and behave in a way that we have automatically realised without thinking about it. Natural language processing (NLP) processes human language as automatic or semi-automatic (Copestake, 2002).

Thanks to software developed by Natural Language Processing (NLP) that generates and understands the natural language, the user can speak naturally through the computer, not through any programming or artificial languages. Diller is divided into two: Machine language

and natural language used by humans. Computers need to use natural language processing knowledge to communicate with people and understand them (Seker, 2015a).

Natural language processing is divided into two according to inputs. These are based on text processing and speech processing. The first is to give a voice response to the person in case of speaking, and the other is to analyse a written text. There needs to be more work on the voice. Generally, studies are carried out on writing voice narratives and, later on, processing written texts. Natural language processing focuses on four main subjects: morphological-lexical, syntactic, semantic, and pragmatic discourse. Morphology is the study of how words are formed and how they relate to other words in the same language. It analyses the structure of words and stems, root words, prefixes, and suffixes. Syntactic: the arrangement of the words in a sentence. It deals with how a word is spoken. Semantic examines the meanings of the used sentences. In order for the natural language to be processed correctly by the computer, it must first be understood correctly. The pragmatics discourse deals with the words and meanings used in a conversation. First, speech is understood by the computer and then answered using correct words (Seker, 2015b).

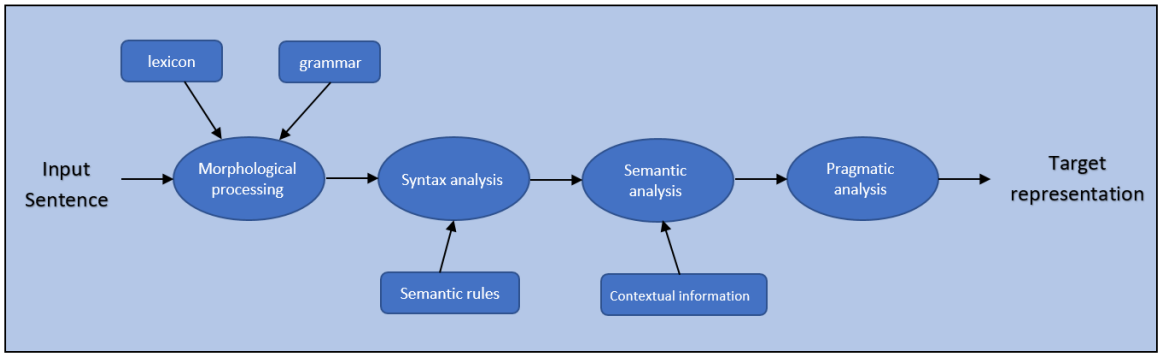


Figure 15 The logical steps in NLP

Morphological Processing

Morphological processing is the preliminary step performed before the syntax analysis. In this stage, the strings of language input are broken into sets of tokens corresponding to discrete words, sub-words and punctuation forms.

For instance, a word like "incorrectly" can be broken into three sub-word tokens:

in-correct-ly

Morphology is recognising how basic words are derived from words with similar semantics but different syntactic structures. The modification is generally performed by adding prefixes and/or suffixes to the word, but other textual changes take place. Generally, there are three basic cases of word change (Teesside University, 2017).

1. **Inflection:** Textual representations of words change because of the syntax. For example, in English, plural nouns are suffixed with -s suffix. In addition, regular adjectives receive -es -est suffixes when comparisons are made.
2. **Derivation:** To derive new words from existing words takes place by morphological rules. For example, to produce a name using some adjectives, the adjectives take -ness as a suffix.
3. **Compounding:** At least two words form a new word. Examples include football, blackboard, breakwater, underworld, and highlight words.

Syntax and Semantics

Many studies on natural language processing are based on assumptions, and when the assumption is made, the key point is the sentence. A sentence expresses ideas, suggestions, thoughts, or emotions and indicates something about them. Thus, it is a very important issue to extract the meaning from a sentence. However, the sentences are not simply linear arrays and are therefore widely known to require an analysis of each sentence to perform this task. This involves an analysis of the clue. The NLP approaches based on productive linguistics are based on determining the syntactic or grammatical structure of each sentence (Indurkhya and Damerau, 2010).

Many operations must be performed in language processing based on syntax and semantic analysis. Syntax analysis has two purposes. The first is to check that a culmination is correctly formed and to divide the sentence according to a structure that shows the syntactic relations of the different words. This is done by the syntactic analyser using a set of word definitions and syntax rules (grammar). A simple lexicon contains the syntactic category of every word. A simple grammar explains rules that show how syntactic categorisations are combined to form different types of patterns (Teesside University, 2017).

A simple lexicon and grammar examples can be as follows:

Lexicon	
cat	noun
chased	verb
large	adjective
rat	noun
the	article

Table 1 An example of the lexicon

Grammar
Sentence → NounPhrase, VerbPhrase
VerbPhrase → Verb, NounPhrase
NounPhrase → Article, Noun
NounPhrase → Article, Adjective, Noun

Table 2 An example of grammar

The sentence "The large cat chased the rat" could be destructured by this grammar-lexicon combination as follows:

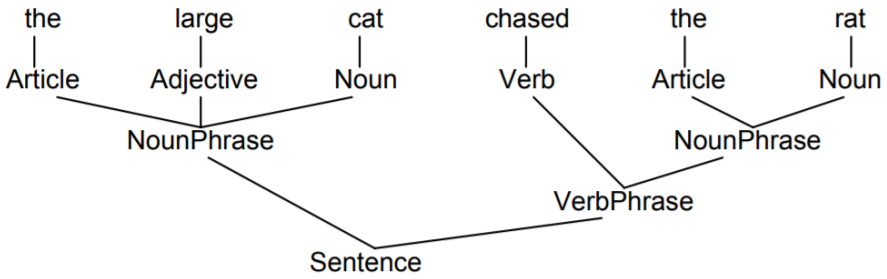


Figure 16 An example of a grammar-lexicon combination

In general, the task of a language processor is to analyse a sentence in a language such as English and produce an expression in some formal notation that briefly expresses the semantics of a sentence in terms of the computer system. An interface to a database, for example, a language processor, may need to convert English or German sentences into SQL queries. The production of this formalised semantic representation is called semantic analysis. To carry out the semantic analysis, the grammar must be extended to identify the semantics of each word it

contains and to determine how the semantics of any semantics of the semantics of sentence is formed. To capture semantic information, grammar and vocabulary can be extended using a simplified logic format (Teesside University, 2017).

Semantics and Pragmatics

After the semantic analysis phase, the pragmatic stage is dealt with. However, there is yet to be a universally accepted distinction between semantics and pragmatics. Pragmatics are not the main concern of many NLP systems. However, once the ambiguities in syntactic or semantic levels have emerged, the content and purpose of the expression are found for analysis. Consider a problem that uses pragmatism in such "support" capacity: ambiguous noun phrases.

One of the most common problems when translating natural language entries into a computer system is the resolution of ambiguous references in noun phrases. For example, suppose an NLP system encounters an exact respect to "X". In that case, there may not be any indication of a discussion of the current cumulated X. This could be a reference to a previously mentioned part of the discourse, or this sentence might contain the first word of X. Without this knowledge, attributes that are not relevant at all can be associated with such an object. A similar problem may arise with ambiguous noun clauses. Certain indefinite references, such as "X", do not define a particular X, and the dialogue may not have had a previous one (Cherpas, 1992).

2.6 Social Network Analysis

What is a Social Network?

Thanks to people being able to relocate easily and especially the development of telecommunication services, people's interaction and communication with each other has increased at an important level. Sound and video conversations can be provided in real-time, even with people located miles away.

A network is a graphical representation of a group of nodes connected by edges (Kim *et al.*, 2011). A social network is a way of sharing and managing people's relationships with each other in a virtual environment. In short, the social network is the interaction of multiple entities. Personal information can be shared through computer-aided communication platforms, such as

online social networking sites, and inter-individual relationships can be defined. The rapid development of internet technologies and their worldwide expansion have led people to spend more time in the virtual world, where they can express themselves more easily. Social actions such as chatting, organising activities or sharing anything are commonly done in the virtual environment. So, people can communicate with each other easily and quickly in the social arena whenever they want.

What is Social Network Analysis?

Social networks include many kinds of relationships, such as personal, official, family, and geographical relationships. Many computer software are used to analyse these networks, and the widespread use of the software has led to an area called social network analysis. Social network analysis is the use of this data in the analysis of any event after quantifying the relationships between people. Social network analysis aims to map associations that connect people as a network to understand society and then reveal relationships between key persons, groups, and components within the network. With SNA, it is easy to make comparisons by analysing various networks. SNA uses social networking to generate sociograms (Meese and McMahon, 2012). A sociogram is a graphical representation prepared to show the relationships among the various members, demonstrated as nodes (Figure 17), of a group (Didactic Encyclopedia, 2015). Demonstration of relationships is made through links between individuals.

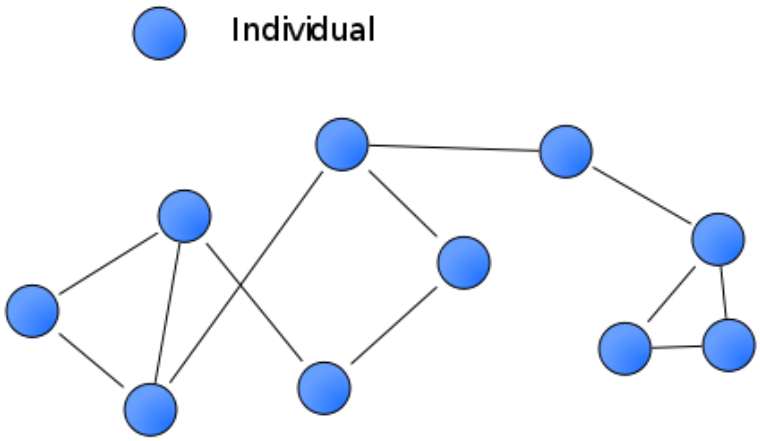


Figure 17 A social network diagram (Wikipedia, 2016)

2.6.1 Social Media Use in Natural Disasters

With the ever-increasing popularity of the internet and the proliferation of smartphones, many users have begun to spend their time in social media environments. While social media is an important part of our lives, these environments also change how people interact and do business. Social networks like Facebook, Twitter, Instagram, Vine, and LinkedIn have created a new industry that includes social entrepreneurs, application creators, social media executives, and consumers. Social media tools have become inseparable from users' lives with the numerous virtual interaction options available for leisure time. At the same time, they have also begun to affect how users communicate. In the same way, people have started to use social media resources more than traditional media resources, such as newspapers, television, etc., about what is happening in the world. In addition, users have started to prefer this media more frequently because of the speed of social media and the power to reach mass media in case of disaster communication. An example is the flood disaster in Calgari in Canada in 2013. When the evacuation was taking place in the affected areas of the city, Twitter was the most commonly used social media tool, and authorities could identify the greatest need (Antoniou and Ciaramicoli, 2013). Likewise, in 2011, the news about the flood disaster in Queensland, located in Australia's south-eastern province, was updated and shared via Twitter by both the authorities and citizens, which showed that Twitter is an effective media tool (Bruns *et al.*, 2012).

Social Media Analysis for Disaster Management

Social Media Analysis is critical in extracting useful information from social media. Social media includes both very important information and a lot of personal and useless information for the community. For this reason, information on social media should be analysed. Many companies and organisations are doing sentiment analysis for their products and services. For example, investors might save millions of dollars from investing in bad investments thanks to the valuable information. Sentiment analysis can also create early warning systems, such as the detection and prediction of natural disasters and hazardous situations. For instance, by predicting bad weather conditions, aircraft can be prevented from taking off, and hundreds of lives can be saved (Jain, 2015).

2.6.2 Twitter

Twitter is an online social network used by millions of people worldwide. People stay connected with family, friends, and colleagues through Twitter. With the development of smartphones, access to Twitter has become easier and faster. Thus, people can connect anywhere that internet connection is available anytime. Twitter allows users to send a message called a tweet with a maximum of 140 characters, typically at most 30 words. In addition, users can share and change messages sent by other users. Tweets can be shown only to their followers or to all users, according to the wishes of the sharing user. A user may follow, remove from their own follow-up list or block other users. Currently, the total number of monthly active Twitter Users is 317 million, and an average of 500 million tweets are tweeted daily (Aslam, 2017).

The most common Twitter terms

Trending Topic: It is abbreviated as TT. Twitter shows a list of the ten most spoken topics. This section is on the left side of Twitter's main page. TT list varies according to country.

Friday Follow: It is abbreviated as FF. This name was given only on Fridays. FF is used to mean that followers follow you.

Retweet: One of the most used features of Twitter, Retweet is abbreviated as RT. It is about sharing other users' Tweets, like quotes, for their followers to see.

Direct Message: It is abbreviated as DM. It is the name of the private messages sent between the users of Twitter. The message feature is also the e-mail function.

Mention: If you put a @ sign at the beginning of the Tweet, you will have committed. It is a feature that allows you to show the person you want to see as the target most shortly and accurately. You can see the mentioned posts from what is mentioned on Twitter.

Hashtag: It can also be called the topic title. If you put a '#' mark at the beginning of the words, you would have done Hashtag.

The Importance of Twitter

Twitter has become a significant resource in the field of data mining thanks to its many features. It has various users from various parts of the world. Thanks to information technology, it spread rapidly among millions of people. People can send messages in real time. As mentioned above, nearly 500 million tweets a day are being sent. For this reason, it becomes almost impossible to follow the information. At this point, the hashtag feature plays a significant role. It is very easy to search for or filter messages by a subject of a user using hashtags. Many people create hashtags to communicate with their users, allow them to ask questions, and answer those questions. For example, BBC makes a hashtag like #AskBbcLive while broadcasting a live channel and allows its users to do the things mentioned above. Twitter also can secretly post tweets for the safety of its users. People often share information that is not personal but valuable for data mining. This information can be subject to such issues as politics, traffic congestion, and natural disasters (landslides, earthquakes, floods, fires, storms). For this reason, people can use Twitter to be aware of the latest news. Tweets can also be used in early warning systems. (Jain, 2015). When the tsunami disaster occurred in Japan in 2011, Twitter was used as a communication tool (Taylor, 2011).

Communication in emergencies like natural disasters comes to the forefront with its real-time communication feature. Social media platforms like Facebook, Twitter, Flickr and Foursquare are important for crisis communication and communication researchers. Depending on the need for increased communication during natural disasters, the importance of social media tools has begun to be addressed in a different context. Some organised groups are investigating issues related to social media tools, especially Twitter, on emergency awareness, communication and information sharing on natural disasters. In many natural disasters and emergencies that have taken place in recent years, Twitter has taken on an important communication role thanks to its multiprotocol structure. Emergency services also use case studies to inform the public about the current situation through the feedback of relevant hashtags. The work of social media on natural disasters and emergencies is generally based on two main topics. The first focuses on the real-time follow-up of feedback from the social media tool and the creation of a reliable research infrastructure (Bruns *et al.*, 2012). The other focuses on the communicative dimension of why social media is important in such situations (Lindsay, 2011).

3 Implementation

3.1 System Requirements

- Windows 10
- Apache Hadoop 2.6.0
- Virtual Machine (Preferred Oracle)
- Putty
- WinSCP
- Web Browser (any)

3.2 Tweet Extraction

Twitter is an online social network that produces about 500 million tweets a day. This enormous data collection can be accessed via Twitter APIs.

The application program interface (API) consists of some protocols and tools for building software applications that identify how software components should interact (Beal, 2017).

Twitter has an Application Programming Interface (API) that allows access to all its users' data through a programmable query method (Go, Bhayani and Huang, 2009). The Twitter platform provides access to the server's data through the APIs that belong to this data. Software specialists can develop different applications through these APIs. Figure 17 shows the general framework for tweet extraction.

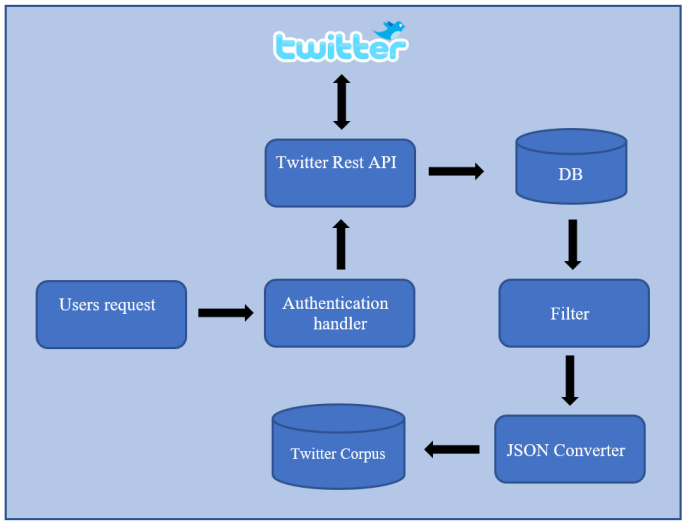


Figure 18 Tweet Extraction Framework

Authentication Handler

OAuth is an open standard authorisation protocol that internet users use to access third-party websites without revealing their passwords to their Google, Microsoft, Facebook, Twitter, or One Network accounts (Gordon, 2012). Twitter supports some authentication methods. It is an obligation to obtain an OAuth access token on behalf of a Twitter user to make authorised calls to Twitter's APIs. Below are the steps required for this process.

1. The user clicks on https://apps.twitter.com/ after the Twitter account is opened.

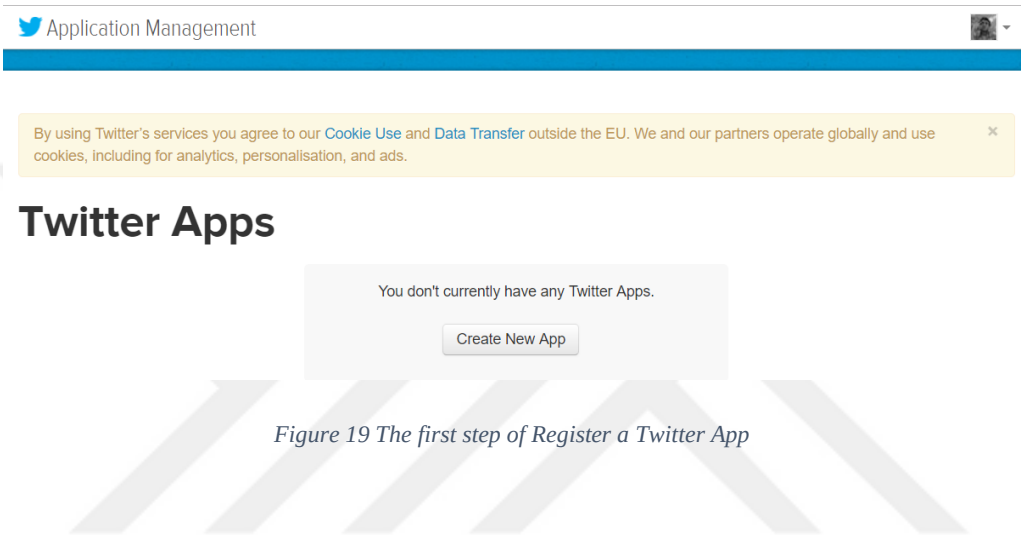


Figure 19 The first step of Register a Twitter App

2. Click the Create New App button and fill in the blanks.

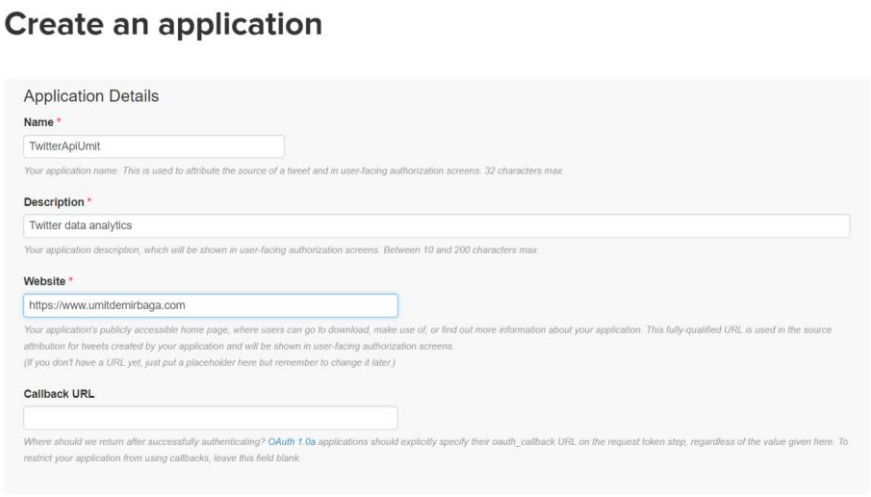


Figure 20 The second step of Register a Twitter App

3. The default access type is read-only When a Twitter app is created. If the user wants the app to delete or write data and access direct messages, the option should be selected as in the photo.

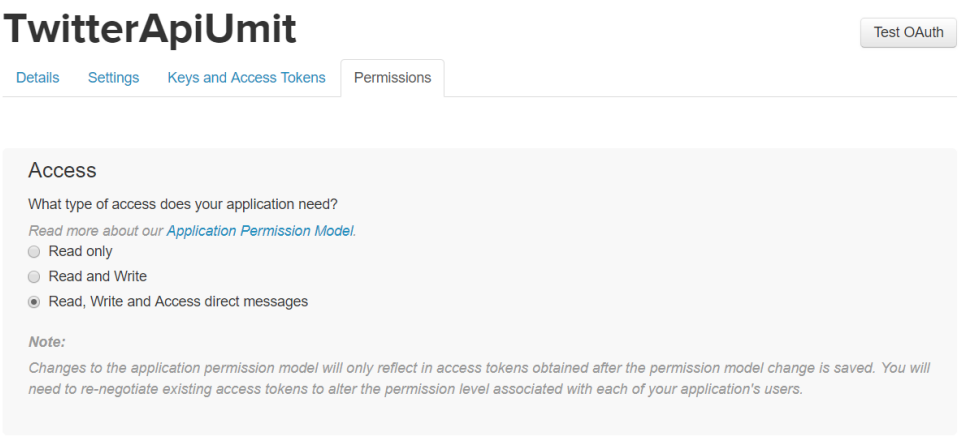


Figure 21 The third step of Register a Twitter App

Then click on the Test OAuth button in the upper right corner.

4. The following figure will show the four long character strings required for the user's Twitter application. These are:
    - Consumer Key,
    - Consumer Secret,
    - OAuth Access Token,
    - OAuth Access Token Secret.

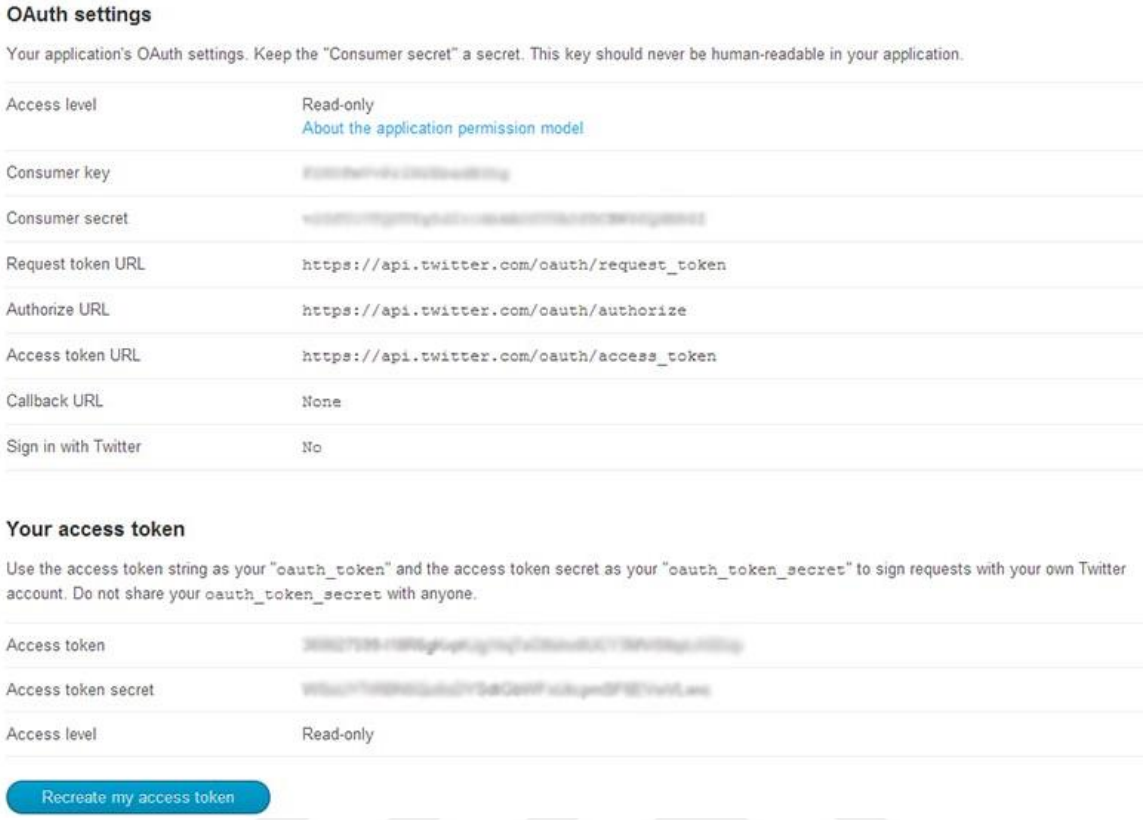


Figure 22 The fourth step of Register a Twitter App

3.2.1 Filtering

According to a study by some academic staff (Zhou, Chen and He, 2015), two methods have been proposed to filter tweets. The first is the method of mapping the dictionary. First, a glossary of keywords is created. Then, only the tweets that contain the words in this dictionary are stored. The second is another approach, which casts tweet filtering as a binary classification problem. When a set of tweets M = (m1, ..., mk) is given, the classifier generates a class C ∈ {event, non-event}. To create a good classifier, an accurate feature set must be designed.

3.2.2 JSON Conversion

JSON is a form of data exchange in which many systems agree on data communications (Lindsay, 2015). JSON is a data definition format independent of programming languages but similar in writing to C-derived languages.

JSON is built on two data structures:

- A collection of name/value pairs. In various programming languages, this is defined as an "object, record, struct, dictionary, hash table, keyed list or associative array".
- Sequential value list. In most programming languages, this is defined as "array, vector, list, or sequence".

Once the tweet objects have been extracted, they could be symbolised as name/value pairs.

Figure 22 shows the snapshot of a JSON file.

Figure 23 JSON format of Twitter data

3.2.3 Twitter API

Twitter has three different APIs: Rest, Search and Streaming. Representational State Transfer (REST) API uses a structure that includes some network design rules that reach data by showing the data address. It provides functionality to Twitter through the simple interfaces that it has. REST architecture means Web syndication, the process by which an application collects information from a source and distributes it to various sources (Strickland and Chandler, 2017). The Twitter Search API, a part of the REST API, allows only popular tweets or tweets published in the last seven days to be queried (Twitter, 2017b). With this API model, only tweets that are pre-tweeted and are limited by Twitter's rate limits could be queried. The maximum number of tweet queries for a user is 3200. For a given keyword, a maximum of 5,000 tweets can be received. In addition, the number of requests can be made within 15

minutes is 180. A streaming API is a persistent connection that leaves the HTTP connection open for as long as possible. To receive new data from the server, the client does not have to request again. The main advantage of this API is that it reduces network latency by producing a continuous stream of data on Twitter.

How to Select Twitter Data

The first step to analysing Twitter data is to collect the data. About 500 million tweets are tweeted on Twitter per day. This large data contains a wide variety of information. Therefore, there are three separate ways to collect the data to be analysed: geo-tagged tweets, tweets using a keyword (hashtag), and tweets from a user.

A keyword (hashtag) is used in tweets.

Using a hashtag is the best way to separate a tweet's topic into categories and search for other tweets about these topics. Putting the pound sign (#) before the word or phrase without using spaces or punctuation to create a hashtag is necessary. When a hashtag is tweeted, it automatically becomes a clickable link. If somebody clicks on that, he is directed to a page with all the most recent tweets containing that specific hashtag. For example, suppose #newcastleUniversity or #NewcastleUniversity is typed in the search box at the top of the Twitter page. In that case, all tweets under the newcastleuniversity title are displayed because there is no case sensitivity. A tweet about using the keyword (hashtag) is shown in Figure 24.

Figure 24 An example tweet about Newcastle University

Geo-tagged tweets

Thanks to Twitter's Tweet with Your Location feature, users can add general location information, such as Istanbul, Newcastle Upon Tyne, to their tweets. In addition, users tweet at the exact geographic location coordinates of their tweeted location using some applications. A tweet tweeted at Sultanahmet Square in Istanbul, Turkey, is shown in Figure 17. Also, the information about the site of some cities is as follows (GPS-coordinates, 2017)

- Latitude and longitude coordinates for Ankara, Turkey's capital city:

  39.933363, 32.859741.

- Latitude and longitude coordinates for London is the capital of England and the United Kingdom:

  51.507351, -0.127758.

A tweet tweeted at Sultanahmet Square in Istanbul, Turkey, is shown in Figure 25.

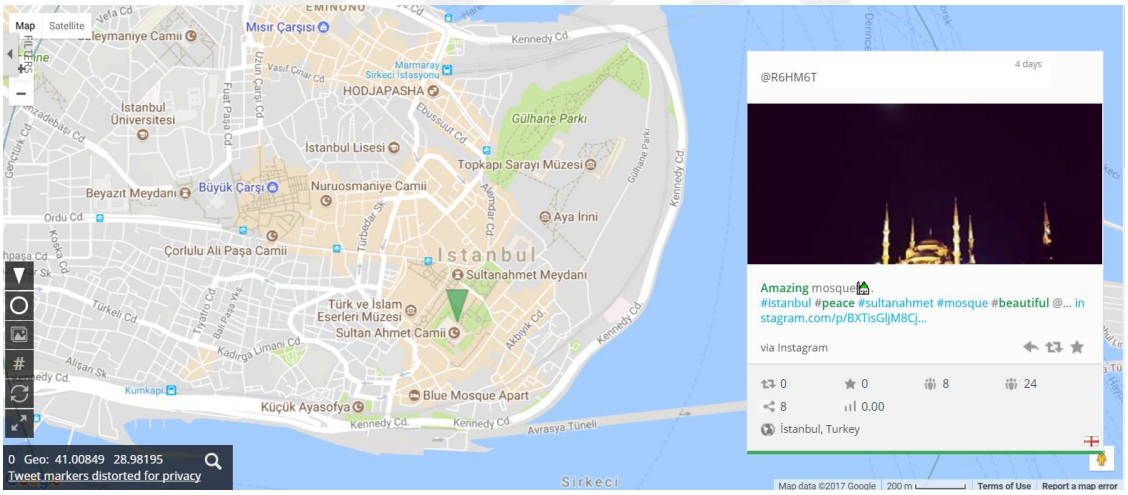


Figure 25 An example of a geo-tagged tweet

After collecting the tweets obtained using the locations of cities, some data about the natural and cultural richness of an area where the data belongs can be reached. Especially when any natural disaster strikes, it can be used as a communication tool.

Tweets from a user

Thanks to the features provided by Twitter, it is possible to receive tweets from the desired user. For this purpose, the Rest API should be used using screen_name or user_id parameters. However, if the tweets belong to a protected user, those tweets may only be accessed by an approved follower of the protected user. However, Twitter has limited the number of tweets obtained with this method to 3200 (Twitter, 2017a).

3.3 Text Normalization

Any text, especially tweets, must be normalised for any NLP task, unlike normal documents or texts. After the tweets are collected, they are in an irregular form. Because users often use language flexibility when expressing their emotions or views. For example, slang, URLs, hashtags, abbreviations, and emoticons are used in tweets, making pre-processing difficult. It is required to systematically pre-process tweets to enhance the sentiment analyser's accuracy. The steps of preprocessing are shown in Figure 25.

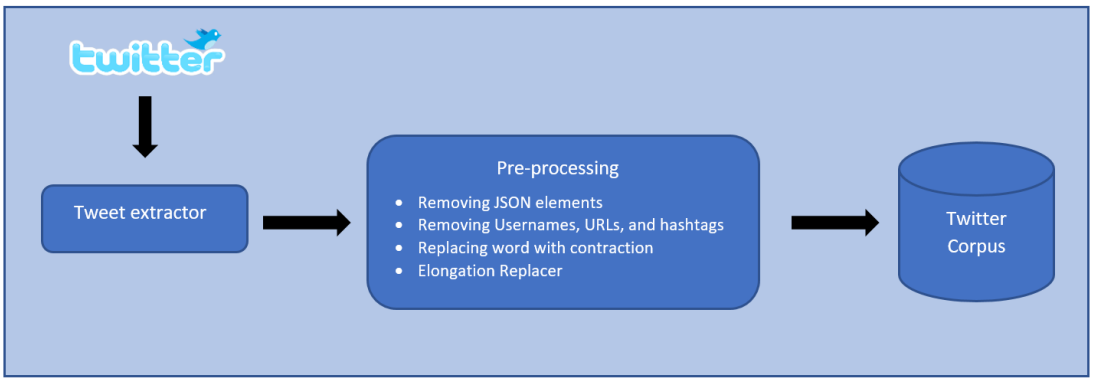


Figure 26 Pre-processing steps

3.3.1 Removing JSON elements

As seen in Figure 27, the JSON file contains much different information, such as index, type, id, score, source, id_str, text, truncated, name, screen_name, location, URL, description, protected, etc. All the information in a tweet can be handled differently depending on the project's needs. In this project, only country_code, text, user_id, created_at and location (latitude and longitude) will be used from all these data.

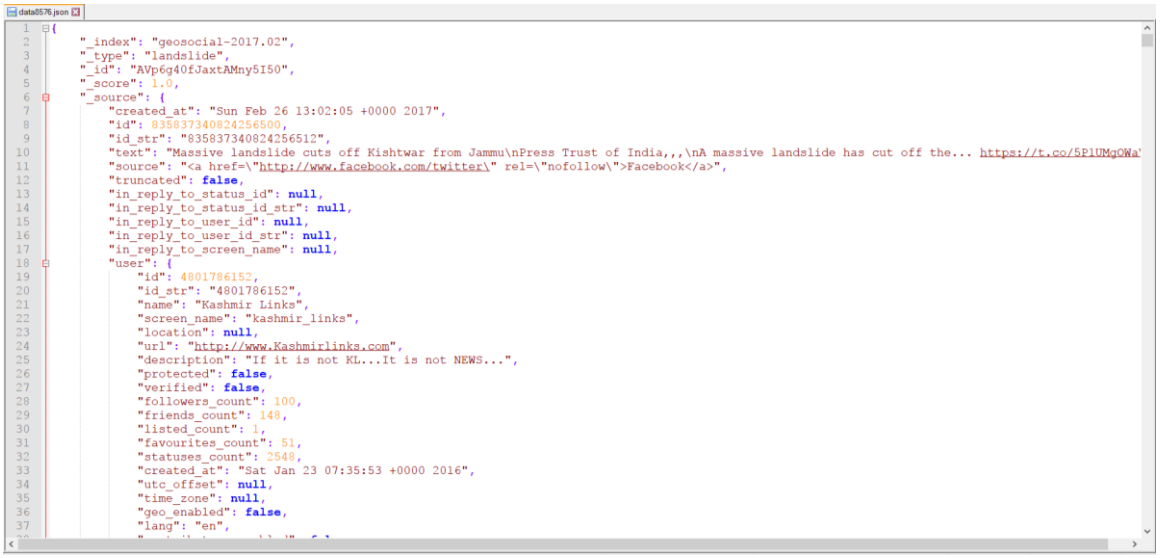


Figure 27 The part of the Tweet used in the project.

3.3.2 Removing usernames, links, and hashtags

Using a link within the tweets, things like a source for the detailed content description, news, or photo can be included. Similarly, users use user mentions to refer to or mention another user. The @ symbol is used for this, followed by a username. Another thing used in tweets is the hashtag, a word that does not contain a space character, starts with the # symbol, and is usually meaningless. URLs, user mentions, and hashtags do not need to be analysed because they can differ from tweet content.

3.3.3 Replacing Word with Contraction

People usually use contractions like didn't, haven't, can't, and might've except official correspondence when writing. For Natural Language Processing (NLP), these expressions should be replaced by equivalent extended language expressions. The definition of which words to replace the contractions must be prepared beforehand. Following this process, each contraction phrase is replaced by its expanded language expressions. The following table shows some examples.

Contracted Forms	Extended Language Expressions
I didn’t take it.	I did not take it.
They still haven’t found a good home.	They still have not found a good home.
She can't change me.	She cannot change me.
He might’ve forgotten your birthday.	He might have forgotten your birthday.

Table 3 Some examples of contraction

3.3.4 Elongation Replacer

Twitter users commonly use elongation or repeated characters in their tweets. People often write words they wish to emphasise, like “goooooood”. The following table shows frequently used elongations. The following table indicates frequently used elongations. Some researchers have suggested limiting the number of repeating characters to two as a solution to extension. But this method causes some words to lose their meaning. For example, if the expression “perfetttt” is normalised according to this suggestion, the result would be “perfectt”, which would be wrong.

Elongated word	Appropriate word
Goooooood	Good
Wellllll	Well
Muchhh	Much
Niceee	Nice

Table 4 Some examples of elongation

3.4 Twitter Data Analysis

3.4.1 Parsing the Twitter Data

Analysing the collected raw big Twitter data with classical methods causes a great waste of time. At this point, Hadoop has significant importance. From this point of view, to analyse the JSON files containing Twitter data on Hadoop, one of the project's main aims, firstly, a console application consisting of 17 classes, was developed in Java programming language. By this application, only the information needed in the project, like country_code, text, user_id, created_at and location (latitude and longitude), were parsed from Twitter data, including lots of information.

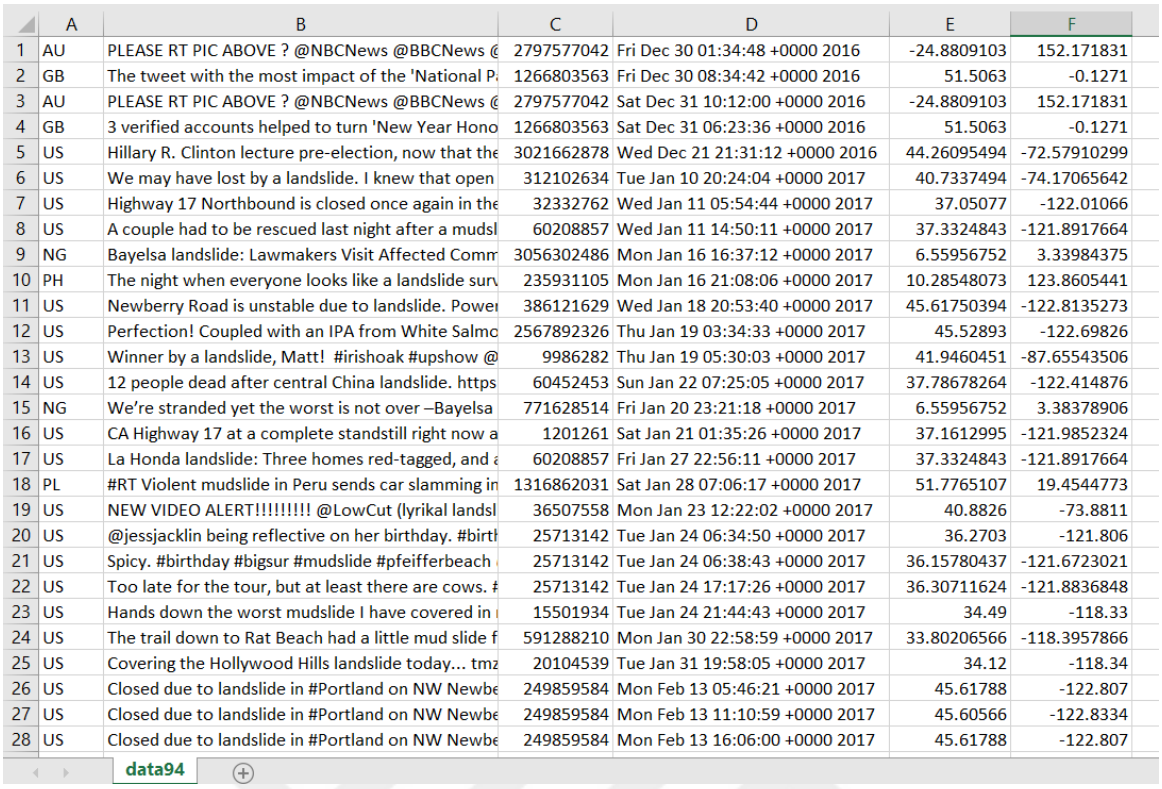
Figure 28 Twitter data structure in JSON file

3.4.2 Saving Parsed Data in CSV Format

One way to process the analysed data in Hadoop is to convert the JSON files in which the Twitter data is located into CSV format. With the application developed, it was saved to a CSV file after parsing the nested Twitter data.

After parsing the nested Twitter data, this data was saved in a CSV file with the same application. Figure 29 shows the converted CSV file after parsing the six-specific tweet information in the JSON file.

A	Country code
B	Text
C	User id
D	Creation date and time
E	Location (latitude)
F	Location (longitude)



	A	B	C	D	E	F
1	AU	PLEASE RT PIC ABOVE ? @NBCNews @BBCNews @	2797577042	Fri Dec 30 01:34:48 +0000 2016	-24.8809103	152.171831
2	GB	The tweet with the most impact of the 'National P	1266803563	Fri Dec 30 08:34:42 +0000 2016	51.5063	-0.1271
3	AU	PLEASE RT PIC ABOVE ? @NBCNews @BBCNews @	2797577042	Sat Dec 31 10:12:00 +0000 2016	-24.8809103	152.171831
4	GB	3 verified accounts helped to turn 'New Year Hono	1266803563	Sat Dec 31 06:23:36 +0000 2016	51.5063	-0.1271
5	US	Hillary R. Clinton lecture pre-election, now that the	3021662878	Wed Dec 21 21:31:12 +0000 2016	44.26095494	-72.57910299
6	US	We may have lost by a landslide. I knew that open	312102634	Tue Jan 10 20:24:04 +0000 2017	40.7337494	-74.17065642
7	US	Highway 17 Northbound is closed once again in the	32332762	Wed Jan 11 05:54:44 +0000 2017	37.05077	-122.01066
8	US	A couple had to be rescued last night after a mudsl	60208857	Wed Jan 11 14:50:11 +0000 2017	37.3324843	-121.8917664
9	NG	Bayelsa landslide: Lawmakers Visit Affected Comm	3056302486	Mon Jan 16 16:37:12 +0000 2017	6.55956752	3.33984375
10	PH	The night when everyone looks like a landslide surv	235931105	Mon Jan 16 21:08:06 +0000 2017	10.28548073	123.8605441
11	US	Newberry Road is unstable due to landslide. Power	386121629	Wed Jan 18 20:53:40 +0000 2017	45.61750394	-122.8135273
12	US	Perfection! Coupled with an IPA from White Salmo	2567892326	Thu Jan 19 03:34:33 +0000 2017	45.52893	-122.69826
13	US	Winner by a landslide, Matt! #irishoak #upshow @	9986282	Thu Jan 19 05:30:03 +0000 2017	41.9460451	-87.65543506
14	US	12 people dead after central China landslide. https	60452453	Sun Jan 22 07:25:05 +0000 2017	37.78678264	-122.414876
15	NG	We're stranded yet the worst is not over –Bayelsa	771628514	Fri Jan 20 23:21:18 +0000 2017	6.55956752	3.38378906
16	US	CA Highway 17 at a complete standstill right now a	1201261	Sat Jan 21 01:35:26 +0000 2017	37.1612995	-121.9852324
17	US	La Honda landslide: Three homes red-tagged, and a	60208857	Fri Jan 27 22:56:11 +0000 2017	37.3324843	-121.8917664
18	PL	#RT Violent mudslide in Peru sends car slamming in	1316862031	Sat Jan 28 07:06:17 +0000 2017	51.7765107	19.4544773
19	US	NEW VIDEO ALERT!!!!!!!!! @LowCut (lyrikal landsl	36507558	Mon Jan 23 12:22:02 +0000 2017	40.8826	-73.8811
20	US	@jessjacklin being reflective on her birthday. #birth	25713142	Tue Jan 24 06:34:50 +0000 2017	36.2703	-121.806
21	US	Spicy. #birthday #bigsur #mudslide #pfeifferbeach	25713142	Tue Jan 24 06:38:43 +0000 2017	36.15780437	-121.6723021
22	US	Too late for the tour, but at least there are cows. #	25713142	Tue Jan 24 17:17:26 +0000 2017	36.30711624	-121.8836848
23	US	Hands down the worst mudslide I have covered in	15501934	Tue Jan 24 21:44:43 +0000 2017	34.49	-118.33
24	US	The trail down to Rat Beach had a little mud slide f	591288210	Mon Jan 30 22:58:59 +0000 2017	33.80206566	-118.3957866
25	US	Covering the Hollywood Hills landslide today... tmz	20104539	Tue Jan 31 19:58:05 +0000 2017	34.12	-118.34
26	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 05:46:21 +0000 2017	45.61788	-122.807
27	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 11:10:59 +0000 2017	45.60566	-122.8334
28	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 16:06:00 +0000 2017	45.61788	-122.807

data94

Figure 29 A view from CSV file content

3.4.3 Grouping the tweets using the MapReduce algorithm on Hadoop

A Map-Reduce console application was created in the Java programming language to handle the CSV file in Hadoop. With the Map method, tweets are grouped by country codes. Later, these tweets grouped according to the country codes by the Reduce method are recorded as separate files in HDFS.

3.4.3.1 Grouping the Tweets by Country Code

Processing CSV file in Hadoop:

1- Firstly, the WinSCP program transfers the CSV and jar files to run in Hadoop to the virtual machine (VM) environment. WinSCP connects to the Linux server or computer via the desired protocol and port through a laptop with Windows installed and allows the users to manage the files listed in the folder, like the structure of Windows.

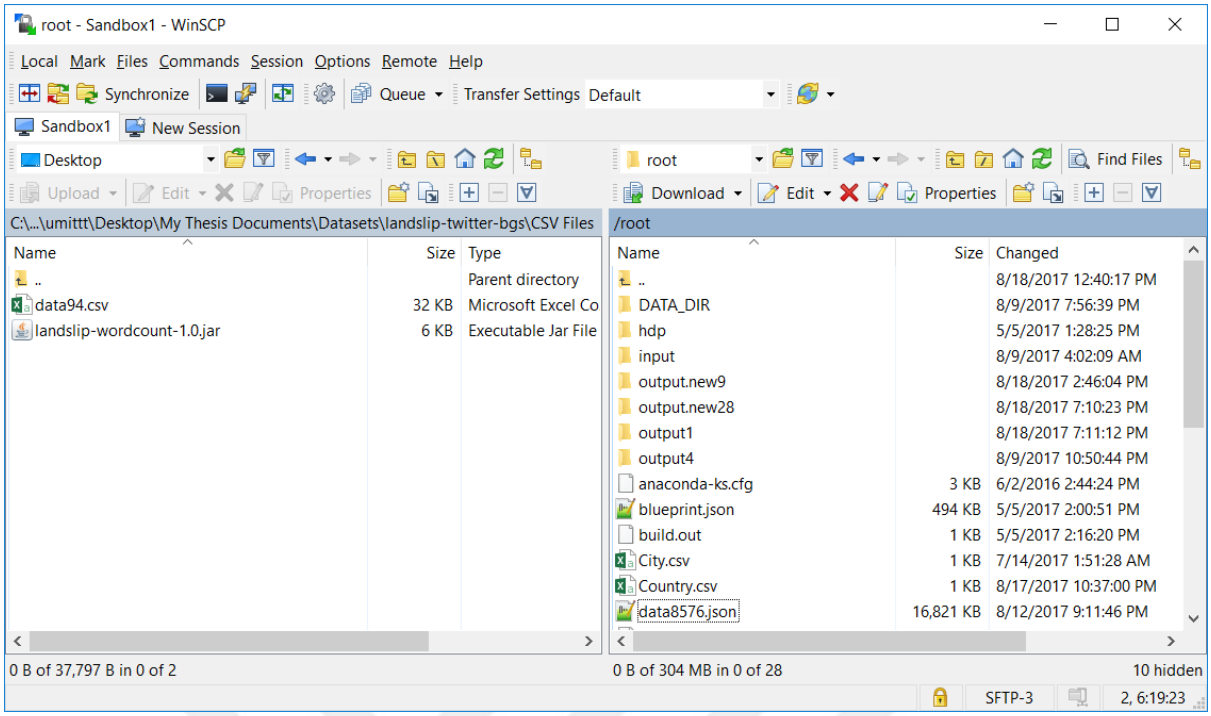
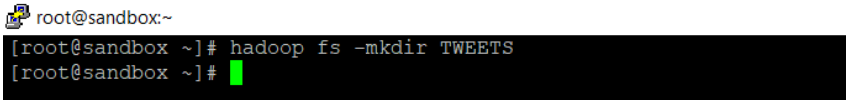


Figure 30 Transfer CSV and jar files to VM

2- To transfer the CSV file to HDFS and process it on Hadoop, the following codes are executed in the Putty program.

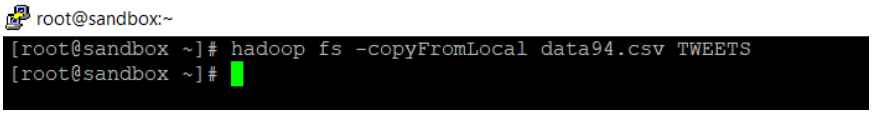
- *hadoop fs -mkdir TWEETS*

With this command, a folder named TWEETS will be created to copy the data94.csv file to be processed.



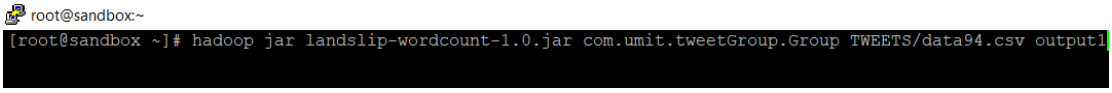
- *hadoop fs -copyFromLocal data94.csv TWEETS*

The data94.csv file is copied to the TWEETS folder.

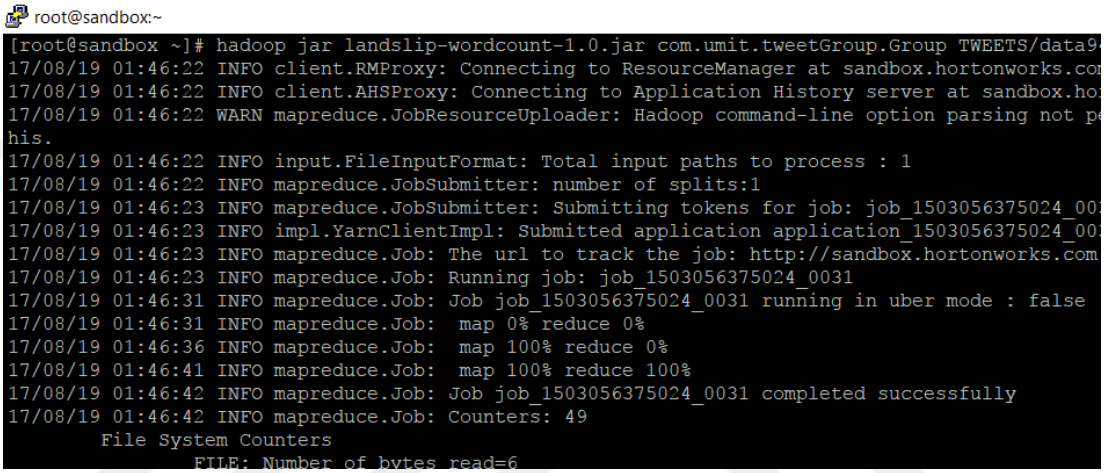


- *hadoop jar landslip-wordcount-1.0.jar com.umit.tweetGroup.Group TWEETS/data94.csv output1*

Then, the jar file is executed by the code to group the tweets in the data94.csv file according to the country code and then to save the tweets belonging to each country code to a separate file. The output1 folder is where the processed results will be saved.



- Map and Reduce processes start after the enter is pressed.



- *hadoop fs -ls output1*

Outputs can be viewed with the following code after the process is finished.

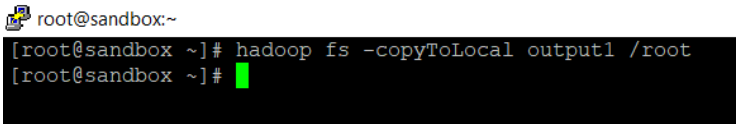
```
Found 36 items
-rw-r--r--   1 root hdfs       8116 2017-08-22 00:47 output1/AU-m-00000
-rw-r--r--   1 root hdfs        716 2017-08-22 00:47 output1/BR-m-00000
-rw-r--r--   1 root hdfs       1001 2017-08-22 00:47 output1/CA-m-00000
-rw-r--r--   1 root hdfs        194 2017-08-22 00:47 output1/CD-m-00000
-rw-r--r--   1 root hdfs        196 2017-08-22 00:47 output1/CH-m-00000
-rw-r--r--   1 root hdfs        193 2017-08-22 00:47 output1/CR-m-00000
-rw-r--r--   1 root hdfs        409 2017-08-22 00:47 output1/DE-m-00000
-rw-r--r--   1 root hdfs        382 2017-08-22 00:47 output1/DO-m-00000
-rw-r--r--   1 root hdfs       1320 2017-08-22 00:47 output1/ES-m-00000
-rw-r--r--   1 root hdfs        506 2017-08-22 00:47 output1/FR-m-00000
-rw-r--r--   1 root hdfs      30303 2017-08-22 00:47 output1/GB-m-00000
-rw-r--r--   1 root hdfs        358 2017-08-22 00:47 output1/GH-m-00000
-rw-r--r--   1 root hdfs        195 2017-08-22 00:47 output1/ID-m-00000
-rw-r--r--   1 root hdfs        501 2017-08-22 00:47 output1/IE-m-00000
-rw-r--r--   1 root hdfs        758 2017-08-22 00:47 output1/IN-m-00000
-rw-r--r--   1 root hdfs        867 2017-08-22 00:47 output1/IT-m-00000
-rw-r--r--   1 root hdfs        703 2017-08-22 00:47 output1/KE-m-00000
-rw-r--r--   1 root hdfs        190 2017-08-22 00:47 output1/MO-m-00000
-rw-r--r--   1 root hdfs       1319 2017-08-22 00:47 output1/MX-m-00000
-rw-r--r--   1 root hdfs        186 2017-08-22 00:47 output1/MY-m-00000
-rw-r--r--   1 root hdfs      13233 2017-08-22 00:47 output1/NG-m-00000
-rw-r--r--   1 root hdfs        191 2017-08-22 00:47 output1/NP-m-00000
-rw-r--r--   1 root hdfs        761 2017-08-22 00:47 output1/NZ-m-00000
-rw-r--r--   1 root hdfs       1252 2017-08-22 00:47 output1/PH-m-00000
-rw-r--r--   1 root hdfs        189 2017-08-22 00:47 output1/PK-m-00000
-rw-r--r--   1 root hdfs        505 2017-08-22 00:47 output1/PL-m-00000
-rw-r--r--   1 root hdfs        371 2017-08-22 00:47 output1/PT-m-00000
-rw-r--r--   1 root hdfs        214 2017-08-22 00:47 output1/SE-m-00000
-rw-r--r--   1 root hdfs        379 2017-08-22 00:47 output1/TH-m-00000
-rw-r--r--   1 root hdfs        123 2017-08-22 00:47 output1/TR-m-00000
-rw-r--r--   1 root hdfs        165 2017-08-22 00:47 output1/TT-m-00000
-rw-r--r--   1 root hdfs      89973 2017-08-22 00:47 output1/US-m-00000
-rw-r--r--   1 root hdfs        657 2017-08-22 00:47 output1/VE-m-00000
-rw-r--r--   1 root hdfs        472 2017-08-22 00:47 output1/ZA-m-00000
-rw-r--r--   1 root hdfs          0 2017-08-22 00:47 output1/_SUCCESS
-rw-r--r--   1 root hdfs          0 2017-08-22 00:47 output1/part-r-00000
[root@sandbox ~]#
```

- *hadoop fs -cat output1/GB-m-00000*

This code can be used to look at the contents of any file, such as *GB-m-00000*.

- *hadoop fs -copyToLocal output1 /root*

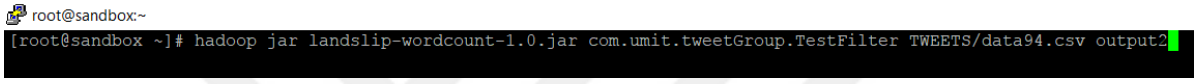
To get these files from HDFS to the user's computer via the WinSCP program, the output1 folder is copied to the VM with the code above.



3.4.3.2 Grouping the Tweets by the Date

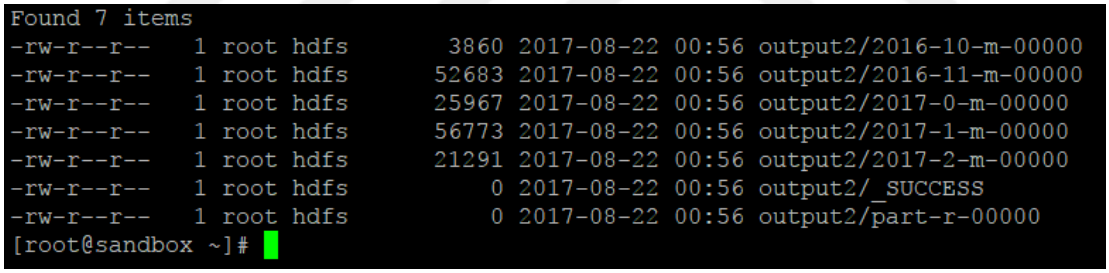
For this process, it is necessary to change only the name of the class.

- *hadoop jar landslip-wordcount-1.0.jar com.umit.tweetGroup.TestFilter TWEETS/data94.csv output2*



- *hadoop fs -ls output2*

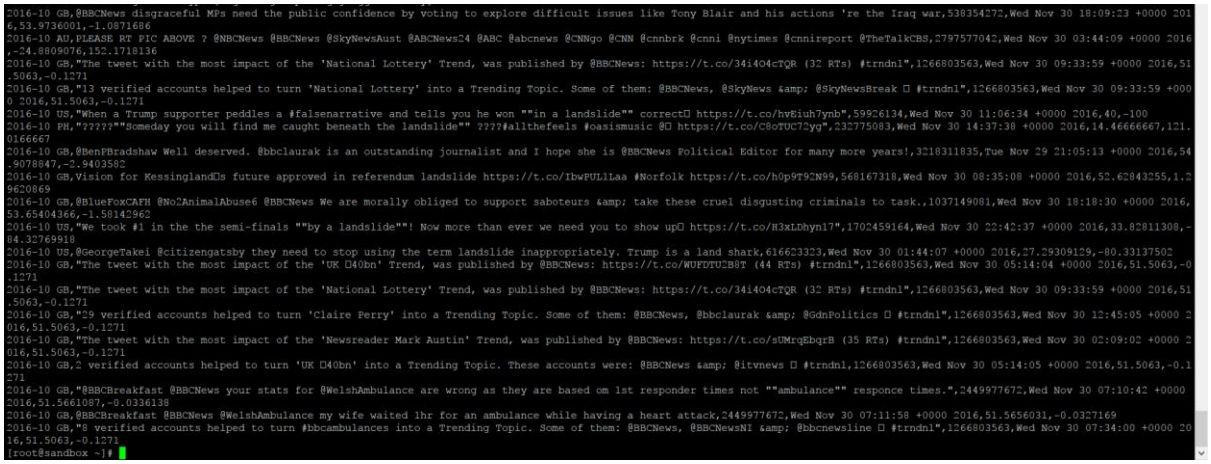
Outputs can be viewed with the following code after the process is finished.



- *hadoop fs -cat output2/2016-10-m-00000*

This code can be used to look at the contents of any file, such as *2016-10-m-00000*.

Note: 2016-10 specifies the October of 2016.



- *hadoop fs -copyToLocal output2 /root*

To get these files from HDFS to the user's computer via the WinSCP program, the output1 folder is copied to the VM with the code above.

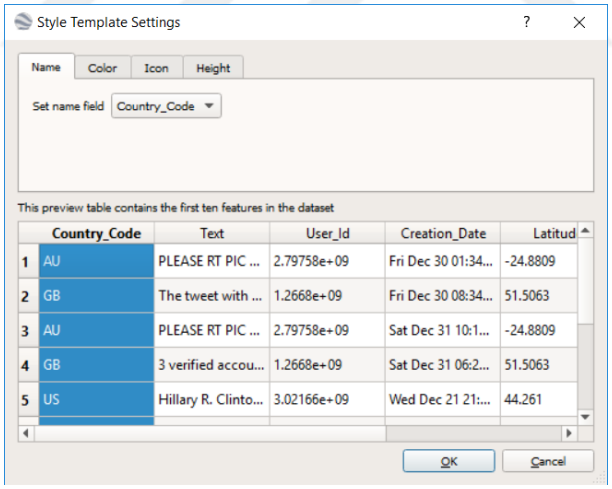
3.4.4 Visualising Twitter Geo Data

Google Earth, which uses Earth's imagery to create Google Earth, benefits many people, not just individuals but corporations. Throughout the area scanned on Google Earth, people can see where they are or visit regions they are curious about.

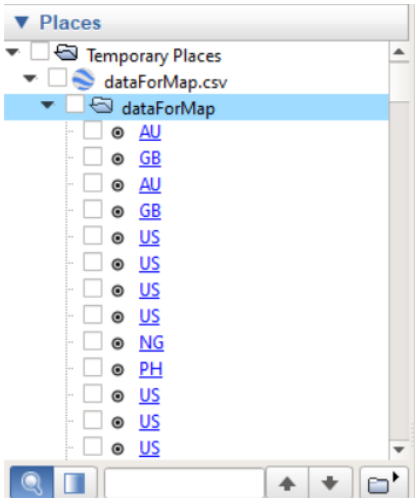
The CSV file generated by analysing the JSON file with the developed Java program and retrieving the desired data contains the country_code, text, user_id, created_at and location (latitude and longitude) information. This information will be displayed on Google Earth using play and longitude information. Therefore, the Google Earth application is needed. It can be downloaded at https://www.google.com/earth/download/gep/agree.html. After opening the application, the following screen will be displayed.

Figure 31 Google Earth app splash screen

Then, the CSV file is imported by selecting File > Import; the following menu is displayed. The Set name field determines which title to use to choose the Tweet.



After the process is completed, tweets are displayed on the left side of the window.



Once you click on any of the tweets listed here, they will be viewed on Google Earth. Some examples are shown below.

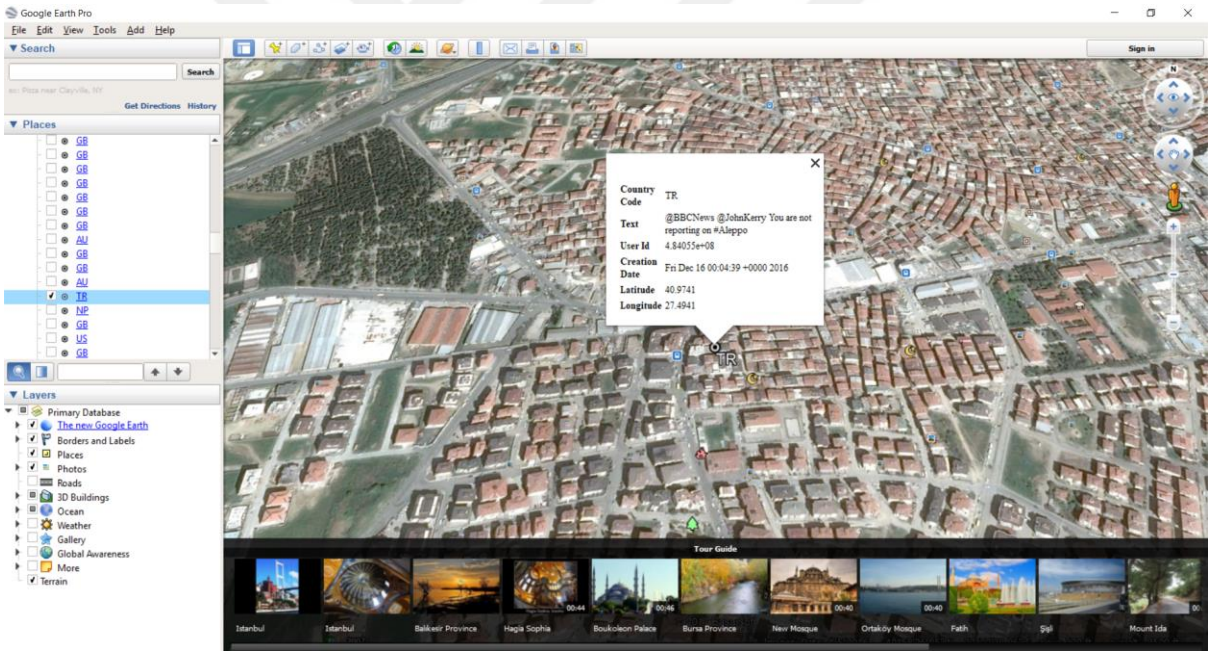


Figure 32 A tweet from Tekirdag, Turkey

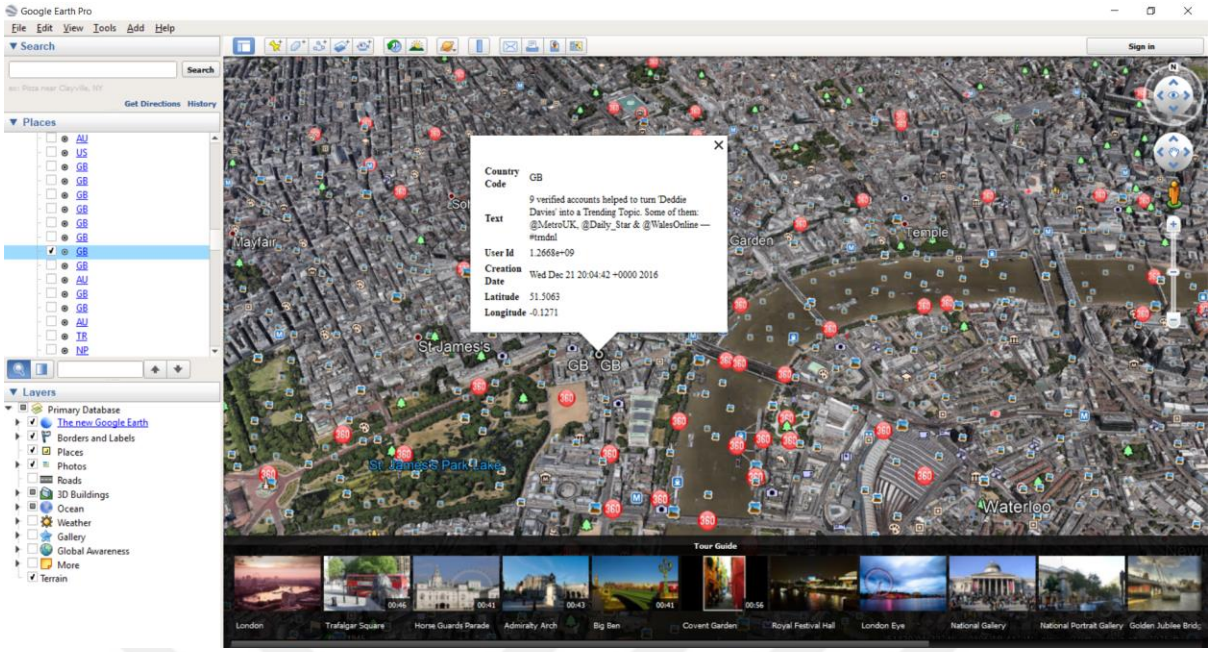


Figure 33 A tweet from London, United Kingdom (Great Britain)

4 Result and Evaluation

4.1 Presenting the result

Tweets must be filtered and analysed to get meaningful results from Big Twitter data in a nested structure. For this reason, the necessary Java programs have been created to enable data processing in the Hadoop environment, allowing analysis of Big data. One of these programs is software that works in a Windows environment and converts the nested JSON files into CSV format after the necessary fields are determined. The other is a jar file created in the Java programming language that groups data and saves them as separate files in the Hadoop environment. Test results are shown below.

The output of the converted CSV file looks like this:

	A	B	C	D	E	F
1	AU	PLEASE RT PIC ABOVE ? @NBCNews @BBCNews @	2797577042	Fri Dec 30 01:34:48 +0000 2016	-24.8809103	152.171831
2	GB	The tweet with the most impact of the 'National P	1266803563	Fri Dec 30 08:34:42 +0000 2016	51.5063	-0.1271
3	AU	PLEASE RT PIC ABOVE ? @NBCNews @BBCNews @	2797577042	Sat Dec 31 10:12:00 +0000 2016	-24.8809103	152.171831
4	GB	3 verified accounts helped to turn 'New Year Hono	1266803563	Sat Dec 31 06:23:36 +0000 2016	51.5063	-0.1271
5	US	Hillary R. Clinton lecture pre-election, now that the	3021662878	Wed Dec 21 21:31:12 +0000 2016	44.26095494	-72.57910299
6	US	We may have lost by a landslide. I knew that open	312102634	Tue Jan 10 20:24:04 +0000 2017	40.7337494	-74.17065642
7	US	Highway 17 Northbound is closed once again in the	32332762	Wed Jan 11 05:54:44 +0000 2017	37.05077	-122.01066
8	US	A couple had to be rescued last night after a mudsl	60208857	Wed Jan 11 14:50:11 +0000 2017	37.3324843	-121.8917664
9	NG	Bayelsa landslide: Lawmakers Visit Affected Comm	3056302486	Mon Jan 16 16:37:12 +0000 2017	6.55956752	3.33984375
10	PH	The night when everyone looks like a landslide surv	235931105	Mon Jan 16 21:08:06 +0000 2017	10.28548073	123.8605441
11	US	Newberry Road is unstable due to landslide. Power	386121629	Wed Jan 18 20:53:40 +0000 2017	45.61750394	-122.8135273
12	US	Perfection! Coupled with an IPA from White Salmo	2567892326	Thu Jan 19 03:34:33 +0000 2017	45.52893	-122.69826
13	US	Winner by a landslide, Matt! #irishoak #upshow @	9986282	Thu Jan 19 05:30:03 +0000 2017	41.9460451	-87.65543506
14	US	12 people dead after central China landslide. https	60452453	Sun Jan 22 07:25:05 +0000 2017	37.78678264	-122.414876
15	NG	We're stranded yet the worst is not over –Bayelsa	771628514	Fri Jan 20 23:21:18 +0000 2017	6.55956752	3.38378906
16	US	CA Highway 17 at a complete standstill right now a	1201261	Sat Jan 21 01:35:26 +0000 2017	37.1612995	-121.9852324
17	US	La Honda landslide: Three homes red-tagged, and a	60208857	Fri Jan 27 22:56:11 +0000 2017	37.3324843	-121.8917664
18	PL	#RT Violent mudslide in Peru sends car slamming in	1316862031	Sat Jan 28 07:06:17 +0000 2017	51.7765107	19.4544773
19	US	NEW VIDEO ALERT!!!!!!!!! @LowCut (lyrikal landsl	36507558	Mon Jan 23 12:22:02 +0000 2017	40.8826	-73.8811
20	US	@jessjacklin being reflective on her birthday. #birth	25713142	Tue Jan 24 06:34:50 +0000 2017	36.2703	-121.806
21	US	Spicy. #birthday #bigsur #mudslide #pfeifferbeach	25713142	Tue Jan 24 06:38:43 +0000 2017	36.15780437	-121.6723021
22	US	Too late for the tour, but at least there are cows. #	25713142	Tue Jan 24 17:17:26 +0000 2017	36.30711624	-121.8836848
23	US	Hands down the worst mudslide I have covered in	15501934	Tue Jan 24 21:44:43 +0000 2017	34.49	-118.33
24	US	The trail down to Rat Beach had a little mud slide f	591288210	Mon Jan 30 22:58:59 +0000 2017	33.80206566	-118.3957866
25	US	Covering the Hollywood Hills landslide today... tmz	20104539	Tue Jan 31 19:58:05 +0000 2017	34.12	-118.34
26	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 05:46:21 +0000 2017	45.61788	-122.807
27	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 11:10:59 +0000 2017	45.60566	-122.8334
28	US	Closed due to landslide in #Portland on NW Newbe	249859584	Mon Feb 13 16:06:00 +0000 2017	45.61788	-122.807

data94

Figure 34 CSV file created after running the developed Java Application

The following figure shows the appearance of the output after the execution of other Java applications to classify the data in the Hadoop environment.

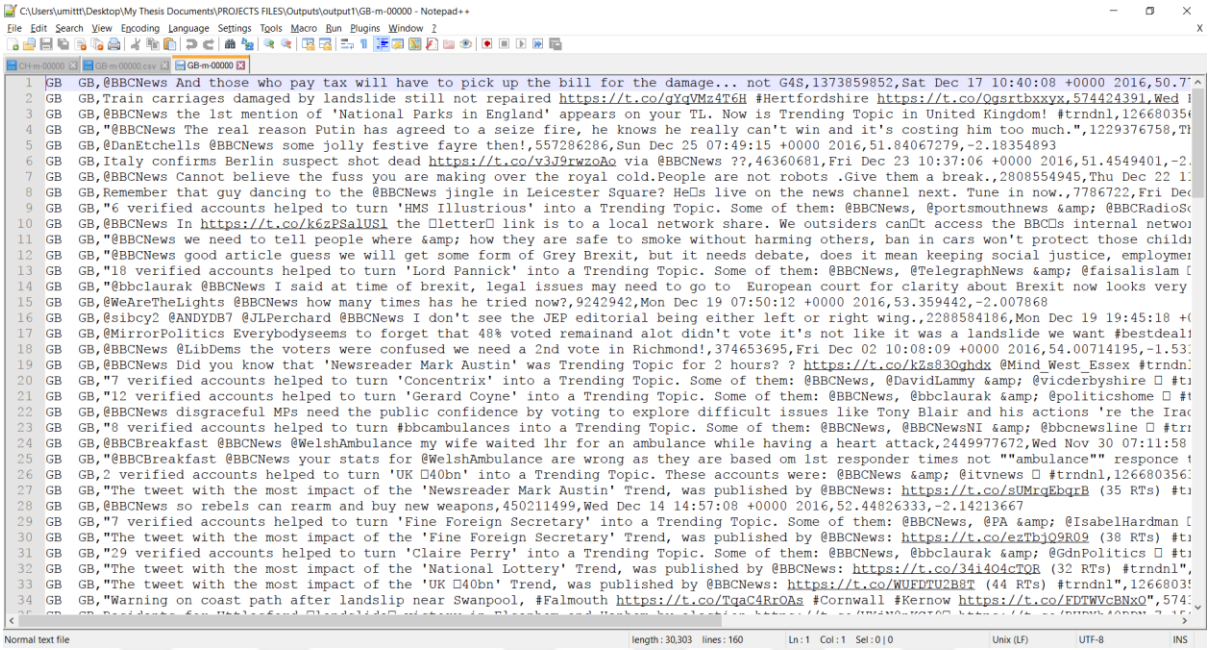


Figure 35 A view from one of the outputs after grouping by country code

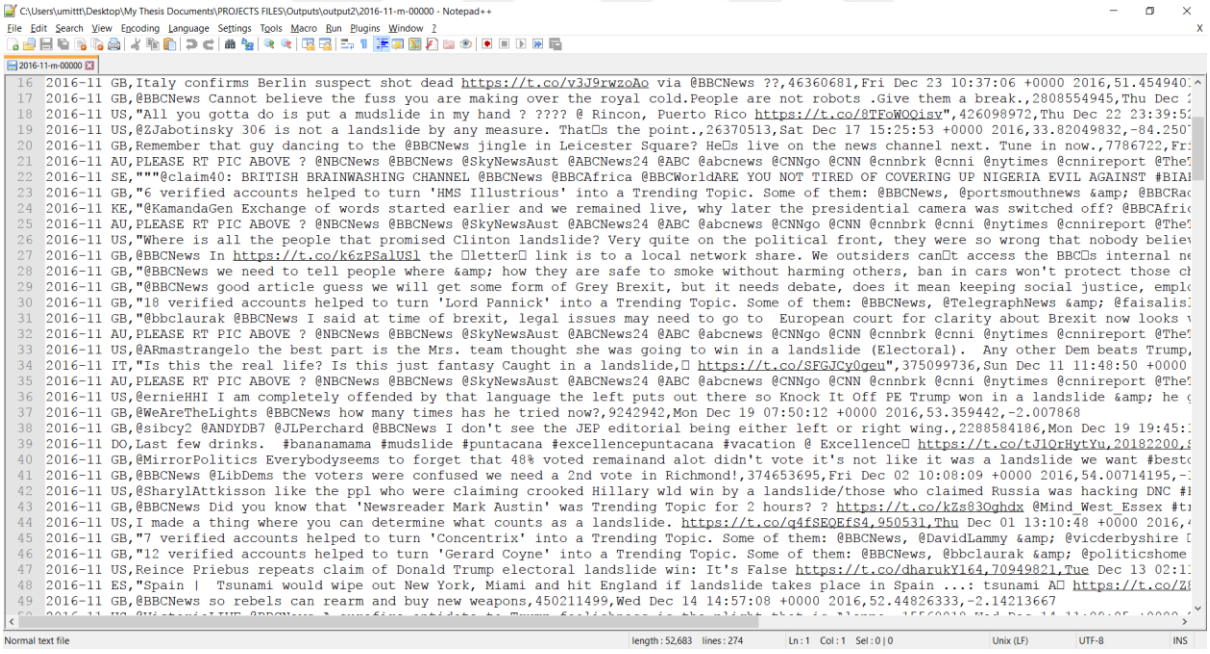


Figure 36 A view from one of the outputs after grouping by the creation date (month)

4.2 Evaluation of the system

A tweet can be classified and analysed in many aspects because it contains many different information. However, due to this project's Hadoop architecture and time limitation, there was

no opportunity to classify the data by other information. Nevertheless, this work has become major in classifying and analysing big data. In addition to this, this study contains useful and guiding information on data classification.

5 Conclusion

The system has achieved the desired data classification and analysis results in the Hadoop environment. The big nested Twitter data is successfully categorised and saved as separate files as desired.

5.1 Meeting the original objectives

The following table shows how much the expected results are met in the project, aiming to process and analyse Big Twitter data.

By reviewing this chapter, the expected and obtained results from the system are seen easily. It has been proven that all the requirements mentioned in the "Expected achievements" section are met. The system meets expectations and saves a great deal of time using a Hadoop system.

Test Description	Expected Outcome	Actual Outcome	Decision (Successful / Unsuccessful)
CSV Conversion	The Java application should correctly convert the JSON file to CSV after filtering.	After running the application, the CSV file with the selected fields was created.	Successful
Classification by Country_code on Hadoop	Country codes should classify the CSV file in Hadoop.	After executing the jar file on Hadoop, the country code categorised the tweets via the Map function.	Successful
Saving separate files by country code on Hadoop	Tweets classified by country codes should be saved as individual files.	The Reduce function saved tweets classified by country codes as files in HDFS.	Successful
Classification by Country_code on Hadoop	The CSV file in Hadoop should be classified by creation date (month).	After executing the jar file on Hadoop, the tweets were organised by the creation date (month) via the Map function.	Successful
Saving separate files by creation date on Hadoop	Tweets classified by creation date should be saved as individual files.	The Reduce function saved tweets organised by creation date as files in HDFS.	Successful
Visualisation on Google Earth	Analysed tweets should be visualised on Google Earth with all the information they contain.	Analysed tweets were visualised on Google Earth with all the information they contained.	Successful

Table 5 Meeting the original objectives

5.2 Positive and negative aspects

Looking at the results from the project, the project's positive aspects have outweighed the negative ones. To see this, it would be more descriptive to examine some of the features of the developed system. The tweet contains many different pieces of information, separated by a comma (,) in the JSON file. Therefore, many CSV readers separate Twitter data based on commas. However, when expressing their feelings or opinions, users use many special characters, including commas. For this reason, mistakes come to the fore when parsing the data. This problem was encountered in the first stages of the project. But this was overcome by using another CSV library. In addition, if the user uses an enter character while writing a text for a tweet, it is considered to have two different tweets. The improved algorithm preceded this problem. All possibilities were considered while the algorithm was being developed, making the system much more powerful.

As a negative aspect of the project, the classification was made in a small number. As mentioned in the "Evaluation of the system" section, because of the time limitation for this project, only two titles could be classified, even though a tweet has many tags.

5.3 Lessons learned

This project has played a significant role in acquiring the following important information about the subject:

- Understanding the structure of a tweet.

- Collecting, analysing, classifying and transforming Twitter data.

- Understanding the structure of Hadoop, the processing, and storage of data.

- Learning and implementing the MapReduce programming model.

- Being able to manage the project properly in a limited time.

- Being able to plan and implement a real project.

- Taking advantage of software engineering aspects of the software development process.

5.4 Future Work

Despite all efforts, many different algorithms, classification, and processing methods have been left for the future due to time limitations. There might be more questions that this study presented in the thesis has answered. This section presents a few proposed topics to improve the project.

- Tweets can be grouped by location information. In this way, people close to each other can be classified.
- Natural Language Processing can be applied to tweets. Thus, an emergency assessment can be made by analysing the tweets about landslides.

6 References

Altintas, I. and Gupta, A. (2017) *Characteristics of Big Data - Veracity*. Available at: https://www.coursera.org/learn/big-data-introduction/lecture/Ln0lI/characteristics-of-big-data-veracity (Accessed: 26 July 2017).

Antoniou, N. and Ciaramicoli, M. (2013) 'Social media in the disaster cycle - Useful tools or mass distraction?', *Secure World Foundation*, 14, pp. 10780–10789. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84904621136&partnerID=tZOtx3y1.

Ashish Bakshi (2016) *Why the world's largest Hadoop installation may soon become the norm*. Available at: https://www.edureka.co/blog/hdfs-tutorial (Accessed: 2 August 2017).

Aslam, S. (2017) *Twitter by the Numbers: Stats, Demographics & Fun Facts*. Available at: https://www.omnicoreagency.com/twitter-statistics/ (Accessed: 29 July 2017).

Assunção, M. D. *et al.* (2015) 'Big Data computing and clouds: Trends and future directions', *Journal of Parallel and Distributed Computing*. Elsevier Inc., 79–80, pp. 3–15. doi: 10.1016/j.jpdc.2014.08.003.

Beal, V. (2017) *API - application program interface*. Available at: http://www.webopedia.com/TERM/A/API.html (Accessed: 6 August 2017).

Boyd, D. and Crawford, K. (2011) 'Six Provocations for Big Data', *Computer*, 123(1), pp. 1–17. doi: 10.2139/ssrn.1926431.

Bruns, A. *et al.* (2012) 'Crisis Communication on Twitter in the 2011 South East Queensland Floods', *Methodology*, (Cci), pp. 1–57. doi: 10.1007/978-3-642-39527-7_16.

Chaffey, D. (2017) *Global social media research summary 2017*. Available at: http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (Accessed: 25 July 2017).

Cherpas, C. (1992) 'Natural language processing, pragmatics, and verbal behavior', (September). doi: 10.1007/BF03392880.

Copestake, A. (2002) 'Natural Language Processing', *Language*, pp. 1–34.

Cukier, K. (2010) 'Data, data everywhere: a special report on managing information', *The Economist*, pp. 1–10. doi: 10.1136/bmj.f725.

Dagli, M. K. and Mehta, B. B. (2014) ‘Big Data and Hadoop: A Review’, *Online) IJARES*, 2(2), pp. 2347–9337. doi: 10.1080/15389588.2012.716880.

Dan Ariely (2016) ‘Big Data’. doi: 10.1007/978-3-658-11589-0.

Dean, J. and Ghemawat, S. (2004) ‘MapReduce: Simplified Data Processing on Large Clusters’, *Proceedings of 6th Symposium on Operating Systems Design and Implementation*, pp. 137–149. doi: 10.1145/1327452.1327492.

Didactic Encyclopedia (2015) ‘Concept and Definition of Sociogram’, pp. 3305–3308. Available at: https://edukalife.blogspot.co.uk/2015/09/what-is-meaning-of-sociogram-concept.html%0A.

Dijcks, J. (2012) ‘Oracle: Big data for the enterprise’, *Oracle White Paper*, (June), p. 16. Available at: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

Elena Geanina Ularu, Florina Camelia Puican, Anca Apostu, M. V. (2012) ‘Perspectives on Big Data and Big Data Analytics’, *Database Systems Journal*, III(4), pp. 3–14. doi: 10.5406/jaesteduc.46.4.iii.

Gartner (2008) *Gartner Says Cloud Computing Will Be As Influential As E-business*. Available at: http://www.gartner.com/newsroom/id/707508 (Accessed: 31 July 2017).

Géczy, P., Izumi, N. and Hasida, K. (2012) ‘Cloudsourcing: Managing Cloud Adoption.’, *Global Journal of Business Research (GJBR)*, 6(2), pp. 57–70. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=66423829&site=eds-live.

Go, A., Bhayani, R. and Huang, L. (2009) ‘Twitter Sentiment Classification using Distant Supervision’, *Processing*, 150(12), pp. 1–6. doi: 10.1016/j.sedgeo.2006.07.004.

Gordon, W. (2012) *Understanding OAuth*. Available at: http://lifehacker.com/5918086/understanding-oauth-what-happens-when-you-log-into-a-site-with-google-twitter-or-facebook (Accessed: 8 August 2017).

gps-coordinates (2017) *Get Latitude and Longitude*. Available at: https://www.gps-coordinates.net/ (Accessed: 6 August 2017).

Guru, D. S. and Suhil, M. (2015) ‘A novel Term-Class relevance measure for text categorization’, *Procedia Computer Science*. Elsevier Masson SAS, 45(C), pp. 13–22. doi:

10.1016/j.procs.2015.03.074.

Hadoop Wiki (2017) *Powered by Apache Hadoop*. Available at: https://wiki.apache.org/hadoop/PoweredBy (Accessed: 2 August 2017).

Hortonworks (2015) 'Hortonworks Sandbox with VirtualBox INSTALL', *Hortonworks*, pp. 1–12.

Hortonworks (2016) 'Hortonworks Data Platform'. Available at: http://hortonworks.com/products/data-center/hdp/.

Hortonworks (2017) *Hortonworks Sandbox*. Available at: https://hortonworks.com/tutorial/getting-started-with-hdf-sandbox/ (Accessed: 28 July 2017).

Hsieh, G. *et al.* (2014) 'Lessons Learned : Building a Big Data Research and Education Infrastructure'.

Indurkhya, N. and Damerau, F. J. (2010) *NATURAL LANGUAGE PROCESSING*.

Ingersoll, G. (2009) 'Introducing Apache Mahout', *Machine Learning*, pp. 1–19.

ISO / IEC (2015) 'Big Data', *Iso / Iec*, pp. 1–36. doi: 10.1016/B978-0-7020-2920-2.50020-0.

Jain, S. (2015) 'Real-Time Social Network Data Mining For Predicting The Path For A Disaster', *Computer Science Theses*. Available at: http://scholarworks.gsu.edu/cs_theses/79%5Cnhttp://scholarworks.gsu.edu/cs_theses/79/.

Judith Hurwitz, Alan Nugent, Dr. Fern Halper, M. K. (2013) *Big Data For Dummies*.

Kim, Y. *et al.* (2011) 'Structural investigation of supply networks: A social network analysis approach', *Journal of Operations Management*. Elsevier B.V., 29(3), pp. 194–211. doi: 10.1016/j.jom.2010.11.001.

Lindsay, B. (2015) *Introduction to JavaScript Object Notation*.

Lindsay, B. R. (2011) 'Social Media and Disasters: Current Uses, Future Options and Policy Considerations.', *Congressional Research Service Reports*, p. 13. Available at: http://fas.org/sgp/crs/homesec/R41987.pdf.

McKinsey & Company (2011) 'Big data: The next frontier for innovation, competition, and productivity', *McKinsey Global Institute*, (June), p. 156. doi: 10.1080/01443610903114527.

Meese, N. and McMahon, C. (2012) 'Analysing sustainable development social structures in

an international civil engineering consultancy', *Journal of Cleaner Production*. Elsevier Ltd, 23(1), pp. 175–185. doi: 10.1016/j.jclepro.2011.10.018.

Mysore, D., Khupat, S. and Jain, S. (2013) *Understanding the architectural layers of a big data solution*. Available at: https://www.ibm.com/developerworks/library/bd-archpatterns3/index.html (Accessed: 3 August 2017).

National Institute of Science and Technology (2013) 'NIST Cloud Computing Standards Roadmap', *NIST Cloud Computing Standards*, pp. 1–3. doi: 10.6028/NIST.SP.500-291r2.

Nedelcu, B. (2013) 'About Big Data and its Challenges and Benefits in Manufacturing', *Database Systems Journal*, IV(3), pp. 10–19.

Nemschoff, M. (2013) *Big data: 5 major advantages of Hadoop*. Available at: http://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/ (Accessed: 2 August 2017).

NIST Big Data Public Working Group: Technology Roadmap Subgroup *et al.* (2015) 'NIST Volume 7 - Roadmap', *NIST Special Publication*, 4(Ref of [DLM+14]), p. 46. doi: 10.6028/NIST.SP.1500-5.

Patodi, R. (2011) *Hadoop: A Soft Introduction*. Available at: https://www.javacodegeeks.com/2011/05/hadoop-soft-introduction.html (Accessed: 27 July 2017).

Press, I., Anderson, J. and Jacobson, D. (2015) *Learning Apache Mahout Classification*.

Purcell, B. (2013) 'Big data using cloud computing', *Journal of Technology Research*, 5, pp. 1–8. doi: 10.1109/ICACCE.2015.112.

Rao, C. C., Leelarani, M. and Kumar, Y. R. (2013) 'Cloud: Computing Services And Deployment Models', *International Journal Of Engineering And Computer Science*, 2(12), pp. 3389–3392. doi: 10.1109/CloudCom.2013.179.

Rittinghouse, J. W. and Ransome, J. F. (2010) *Cloud Computing: Implementation, Management, and Security*, *Biography An Interdisciplinary Quarterly*. doi: 10.1201/9781439806814.ch5.

Seker, S. E. (2015a) 'Computerized argument delphi technique', *IEEE Access*, 3, pp. 368–380. doi: 10.1109/ACCESS.2015.2424703.

Seker, S. E. (2015b) ‘Natural Language Processing’, pp. 1–16.

Shvachko, K. *et al.* (2010) ‘The Hadoop distributed file system’, *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010*, pp. 1–10. doi: 10.1109/MSST.2010.5496972.

Sliwa, C. (2011) *Scale-out NAS, object storage, cloud gateways replacing traditional NAS*. Available at: http://searchstorage.techtarget.com/feature/Scale-out-NAS-object-storage-cloud-gateways-replacing-file-storage (Accessed: 1 August 2017).

Strickland, J. and Chandler, N. (2017) *How Twitter Works*. Available at: http://computer.howstuffworks.com/internet/social-networking/networks/twitter.htm (Accessed: 6 August 2017).

Taylor, C. (2011) *Twitter Users React To Massive Quake, Tsunami In Japan*. Available at: http://mashable.com/2011/03/10/japan-tsunami/#tUY6jv7FqkqL (Accessed: 30 July 2017).

TechCrunch (2017) *Amazon Web Services*. Available at: https://techcrunch.com/topic/product/amazon-web-services/ (Accessed: 31 July 2017).

Teesside University (2017) *Natural Language Processing (NLP), Teesside University*. Available at: https://www.scm.tees.ac.uk/isg/aia/nlp/NLP-overview.pdf (Accessed: 28 July 2017).

tutorialspoint (2017a) *Hadoop Tutorial*, *tutorialspoint*. Available at: http://www.tutorialspoint.com/hadoop/ (Accessed: 26 July 2017).

tutorialspoint (2017b) *Mahout Tutorial*. Available at: https://www.tutorialspoint.com/mahout (Accessed: 28 July 2017).

Twitter (2017a) *GET statuses/user_timeline*. Available at: https://dev.twitter.com/rest/reference/get/statuses/user_timeline (Accessed: 7 August 2017).

Twitter (2017b) *The Search API*. Available at: https://dev.twitter.com/rest/public/search (Accessed: 6 August 2017).

wikipedia (2016) *Sociogram*. Available at: https://en.wikipedia.org/wiki/File:Social-network.svg (Accessed: 5 August 2017).

Williamson, J. (2017) *The 4 V’S oF Big Data*. Available at: http://www.dummies.com/careers/find-a-job/the-4-vs-of-big-data/ (Accessed: 26 July 2017).

Zhou, D., Chen, L. and He, Y. (2015) ‘An Unsupervised Framework of Exploring Events on Twitter : Filtering , Extraction and Categorization’, pp. 2468–2474.

Zikopoulos, P. *et al.* (2013) *Harness the power of big data: the IBM big data platform*.