UNIVERSITI PUTRA MALAYSIA

UNSTRUCTURED BIG DATA PROCESSING IN CLOUD COMPUTING ENVIRONMENT BY USING AMAZON ELASTIC MAP REDUCE

NORZAHARAWANI BUSU

FSKTM 2017 24

UNSTRUCTURED BIG DATA PROCESSING IN CLOUD COMPUTING ENVIRONMENT BY USING AMAZON ELASTIC MAP REDUCE

NORZAHARAWANI BUSU

MASTER OF COMPUTER SCIENCE

2017

UNSTRUCTURED BIG DATA PROCESSING IN CLOUD COMPUTING ENVIRONMENT BY USING AMAZON ELASTIC MAP REDUCE

By

NORZAHARAWANI BUSU

Thesis submitted to the Faculty of Computer Science and Information Technology, University Putra Malaysia, in fulfillment of the requirements for the Master of Computer Science

January 2017

Copyright Page

ABSTRACT

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for Master of Computer Science

UNSTRUCTURED BIG DATA PROCESSING IN CLOUD COMPUTING ENVIRONMENT BY USING AMAZON ELASTIC MAP REDUCE

By

NORZAHARAWANI BUSU

January 2017

Supervisor: Mrs. Sazlinah binti Hasan

Faculty: Computer Science and Information Technology

Nowadays, growing expansion of data content on the web delivers a huge amount of collective resources. Twitter, one of the biggest social media site collects tweets in millions every day in the range of Petabyte per year. Societies share their experiences, thoughts or simply talk just about whatever concerns

them online. Unstructured big data in social media plays vital roles in sentiment analysis or also known as opinion mining.

Continuous structured and unstructured data are being generated in a large scale every day. These data are meaningless if they are not being captured and analyzed accordingly. Traditional RDBMS technology becomes less reliable when dealing with huge amount of structured data and the processing speed of data becomes sluggish if the infrastructure is not being upgraded to match the big amount of data. Furthermore, RDBMS is not capable to deal with unstructured data.

Due to petabytes of records are generated every year on the net, capturing and analyzing big data can be challenging and cloud computing technologies are able to provide an on-demand infrastructures and services based on user requirements. Therefore, this thesis aims to use cloud based infrastructure which is Amazon Web Service to capture unstructured of big data, and afterward analyzing, visualizing and extracting useful information from large, diverse, distributed and mixed of data gathered from public data sets and Twitter's Application Programming Interface (API).

The results and explanation on the experiments mentioned in the chapter four; show the test bed result on collecting twitter data, test bed result on processing twitter input data and test bed result on output data. The analysis emphasizes on the elapsed time when collecting twitter data and also the performance of Amazon Elastic MapReduce (EMR). The infrastructures provided by Amazon Web Service are proficient enough to captured and manipulated large volume of unstructured big data on twitter. Afterward, this study have tested the capability of Amazon Elastic MapReduce (EMR) to process the input twitter data that had collected earlier, and transform them into a meaningful output that can be used for any decision making.

ABSTRAK

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan ijazah Master Sains Komputer (Pengkomputeran Teragih)

PEMPROSESAN DATA RAYA TIDAK BERSTRUKTUR DIDALAM PERSEKITARAN PENGKOMPUTERAN CLOUD MENGGUNAKAN AMAZON ELASTIC MAP REDUCE

Oleh

NORZAHARAWANI BUSU

Januari 2017

Penyelia: Puan Sazlinah binti Hasan

Fakulti: Fakulti Sains Komputer dan Teknologi Maklumat

Pembangunan data dalam talian yang meningkat dengan pesat pada waktu kini membawa kepada pengumpulan sumber yang mendadak. Twitter merupakan salah satu saluran pengumpulan data yang besar di laman media sosial yang melibatkan lebih sejuta data sehari dan merangkumi saiz petabait setiap tahun.

Masyarakat berkongsi pengalaman, luahan rasa dan bercakap mengenai apa sahaja yang diingini dilaman sosial twitter. Data tidak berstruktur di media sosial memainkan peranan penting didalam analisis sentimen atau dikenali juga sebagai luahan pendapat.

Data dihasilkan pada skala besar setiap hari. Data-data ini tidak bererti sekiranya tidak dikumpul dan dianalisa sebaiknya. Teknologi system pengkalan data (RDBMS) yang sedia ada menjadi tidak sesuai digunakan apabila berhadapan dengan isu data yang berskala sangat besar dan pemprosesan data menjadi lambat sekiranya infrastrukturnya tidak ditambah baik.

Data dalam talian yang begitu besar dihasilkan saban hari menyebabkan proses analisis data menjadi isu yang mencabar dan didapati teknologi cloud dapat menangani isu ini kerana ia menyediakan infrastruktur dan kemudahan perkhidmatan berdasarkan keperluan pengguna Justeru itu, tesis ini memfokuskan untuk menggunakan cloud berasakan infrastruktur iaitu Amazon Web Service untuk mengumpul data-data tidak berstruktur dan seterusnya menganalisis data tersebut untuk menghasilkan maklumat yang berguna daripada data yang telah dikumpul melalui set data luar daripada program aplikasi Twitter.

Hasil dan penerangan terhadap eksperimen dinyatakan dalam bab empat; yang menunjukkan ujian dibuat semasa pengumpulan data Twitter, ujian ke atas pemprosesan data input Twitter dan ujian ke atas data output. Analisa telah dijalankan ke atas masa yang diambil semasa pengumpulan data Twitter dan juga perlaksanaan Amazon Elastic MapReduce (EMR). Infrastruktur yang disediakan oleh EMR sangat mencukupi untuk mengumpul dan memanipulasi data tidak berstruktur di Twitter dalam skala yang besar. Seterusnya, kajian ini telah menguji keupayaan Amazon Elastic MapReduce (EMR) untuk memproses input data Twitter dan mengubahnya menjadi output yang bermakna dan boleh digunakan untuk membuat keputusan.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Allah s.w.t. for giving me strength and determination to complete this thesis. I would like to extend my sincere gratitude and appreciation to my Supervisor, Mrs. Sazlinah Hasan for her great supervision, suggestions and full support in all aspects throughout my research.

Sincere thanks to my beloved husband, Azizul Rizal Zulkapli for his support and valuable words of wisdom that helped me stay grooved and finish my study. Without his great loving, supporting and patience with me throughout this tough time, I may not able to make it this success.

I would like to extend my love and thanks to my lovely daughters, Mashitah Marcella Azizul Rizal and Sumayyah Sufiea Azizul Rizal for being understanding and cooperative during times when I was not able to spend quality time with the family, especially in the final stages of my thesis work.

My special thanks to my former lecturer Dr Noor Azliza Che Mat who has support me from the beginning and be my referral to further my study and give me some motivation to complete my master level. Not to forget, my classmates during my study for their great support and help. All parties contributions and accomplishments in this research study will be always remembered.

APPROVAL

Thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for Master of Computer Science.

_____

Project Supervisor,
Mrs. Sazlinah binti Hasan
Department of Communication Technology and Network Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

Date :

DECLARATION

I hereby confirm that:

- This thesis is my original work;
- Quotations, illustrations and citations have been duly referenced;
- This thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- Intellectual property from the thesis and copyright of thesis are fully-owned by University Putra Malaysia (UPM), as according to the Universiti Putra Malaysia (Research) Rules 2012;
- Written permission must be obtained from supervisor and the office of the Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- There is no plagiarism or data falsification/ fabrication in the thesis, and scholarly integrity was upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:_____ Date: January 20, 2017

Name and Matric No.: Norzaharawani Busu GS39640

Table of Contents

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF ABBREVIATIONS

AWS	Amazon Web Service
EMR	Elastic MapReduce
S3	Simple Storage Service
EC2	Elastic Compute Cloud
API	Application Program Interface
JSON	JavaScript Object Notation
AMI	Amazon Machine Image
DNS	Domain Name System
SSH	Secure Socket Shell
FTP	File Transfer Protocol

CHAPTER 1

INTRODUCTION

1.1 Research Background

The significant growth of data being generated and collected is astonishing as we evolve into digital culture. By using traditional analytical tools like relational database management system (RDBMS) or data warehouses technologies, the analysis of this persistently growing data becomes more challenging. Traditional database may not be able to handle billions lines of data or when it deals with semi-structured and unstructured data.

1.1.1 Big Data

Big data, as defined by Bigelow, S. (2014) is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. When we talk about big data, we may refer to Laney, D. (2001) on his 3V's concept in data management where the following three V's stood out as the characteristics of big data, which are Variety, Velocity and Volume as we can see in Figure 1.1. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing. According to the 3V's model, the challenges of big data management result from the expansion of all three properties, rather than just the volume alone which is the absolute amount of data to be managed.

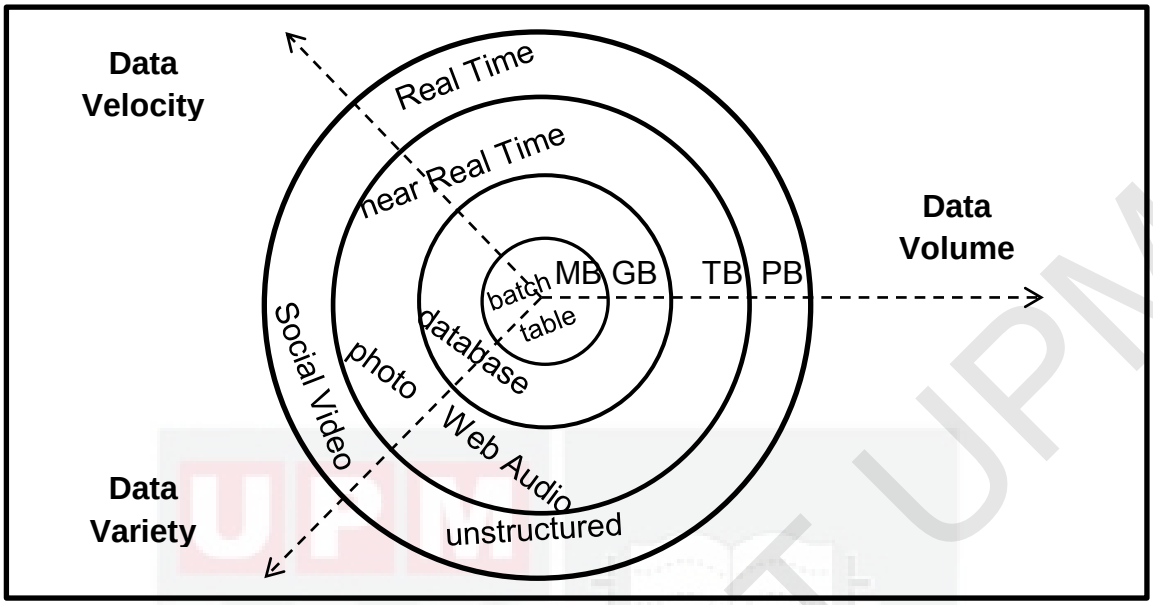

Figure 1.1: Characteristics of Big Data

Recently, Rijmenam, M. (2013) proposed to expand the concept of Big Data by introducing four more V's which are Veracity, Value, Visualisation and Vision to further understand the incredibly complex nature of Big Data as we can see in Figure 1.2 and Table 1. 1.

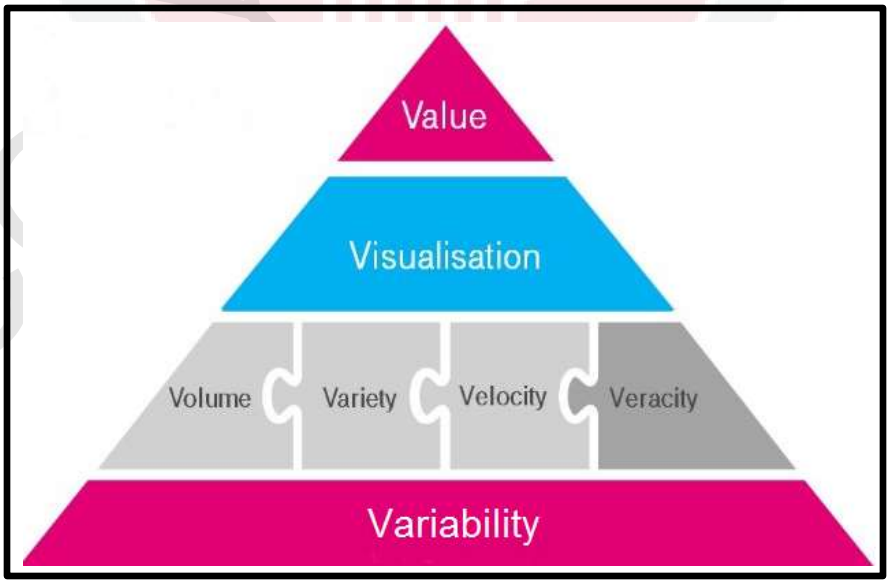

Figure 1.2: 7V's of Big Data

Table 1.1: Additional 4V's definition

Veracity	The accuracy of the data
Variability	Data whose meaning is constantly changing
Visualisation	Data comprehensible in a manner that is easy to understand and read
Value	The usefulness of the data for the company

Below are some interesting facts about Big Data.

- Each day, the world produces 2.5 quintillion bytes of data. That is 2.3 trillion gigabytes. (Optimus Information, 2015)
- By 2020, we will have created 40 zettabytes of data, which is 43 trillion gigabytes. (Optimus Information, 2015)
- The New York Stock Exchange captures 1 Terabytes of Trade Information everyday. (BlazeClan, 2014)
- Every minute of every day, users upload 100 hours of video on Youtube, send over 200 million emails and send 300,000 tweets. (Mcnulty, E. 2014)
- By 2020, at least a third of all data will be passed through the cloud. (Marr, B. 2015)
- The Obama 2012 political campaign used big data analytics and the sentiment analysis method to assemble a winning coalition vote by vote. (Issenberg, S. 2012)

1.1.2 Structure and Unstructured data

Big data is made up from structured and unstructured data. Figure 1.3 shows the illustration on both data. Unstructured data represents any data that does not have a recognizable structure. It is unorganized and raw and can be non-textual or textual. According to Nemshoff, M. (2014) unstructured data refers to information that either does not have a pre-defined data model and/or is not organized in a predefined manner. In a simpler term, it refers to any data that cannot be managed in Database Management System (DBMS). On the other hand, all data that can be stored and manipulated in DBMS are structured.

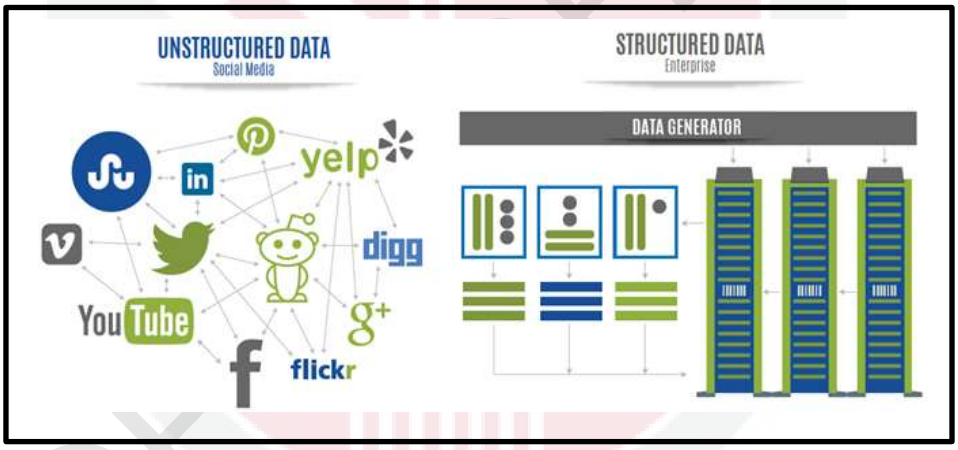

Figure 1.3: Unstructured vs. structured data

1.1.3 Sentiment Analysis

Unstructured big data in social media plays vital roles in sentiment analysis or also known as opinion mining. Thomas, J (2011) defines Sentiment Analysis as studying peoples' opinions, attitudes and feelings towards an event, product or an organization computationally. Twitter, one of the largest social media site receives tweets in millions every day in the range of Petabyte per year. This vast

amount of raw data can be used for marketing strategy, business purpose or even political campaign by organizing them according to the requirement and processing. Big Data firm like Datameer, analyzed unstructured big data in twitter for a week after the presidential debate between Donald Trump and Hillary Clinton. As a result of the sentiment analysis on big data, they found out Clinton is slightly ahead of Trump on a positive sentiment in Twitter posts. Sentiment analysis becomes a hot area in decision making as explain by Taurent, C. (2015). The main goal of analyzing sentiment is to analyze the reviews and examine the scores of sentiments.

1.1.4 Cloud Computing Technologies

Capturing and analyzing unstructured data like twitter posts can be challenging considering Petabytes of data are being generated a year on Twitter alone. Therefore, cutting-edge technologies are required in order to close the gap between the data being generated and the data that can be captured and analyze effectively. This is where cloud computing technologies like Amazon Web Service (AWS) provides the computing flexibility on a global infrastructure with access to many different geographic regions to build sophisticated big data applications.

Cloud computing is Internet based technology where the users can subscribe high quality of services from data and software that resides solely in the remote

servers. This provides many benefits for the users to create and store data in the remote servers thereby utilizing fewer resources in client system. Shawky, D.M. (2012) defines cloud computing as the most recent step in on-demand information technology services and products. It is a computing technology that provides abilities to access software and hardware resources from a virtual space. Data is neither stored on the local hard drive of your computer, nor on servers that are down in the basement of your company. Instead it is out in the cloud.

Analyzing large data sets requires significant compute capacity that can vary in size based on the amount of input data and the type of analysis. This characteristic of big data workloads is ideally suited to the pay-as-you-go cloud computing model, where applications can easily scale up and down based on demand. As requirements change, we can easily resize our environment on AWS to meet our needs, without having to wait for additional hardware or being required to over invest to provision enough capacity.

1.1.5 Amazon Elastic MapReduce (Amazon EMR)

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze enormous amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, data can be

processed for analytics purposes and business intelligence workloads. Additionally, Amazon EMR can be used to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

Amazon EMR's flexible framework reduces large processing problems and data sets into smaller jobs and distributes them across many compute nodes in a Hadoop cluster. This capability lends itself to many usage patterns with big data analytics. Figure 1.4 shows how Amazon EMR interacts with other AWS services.

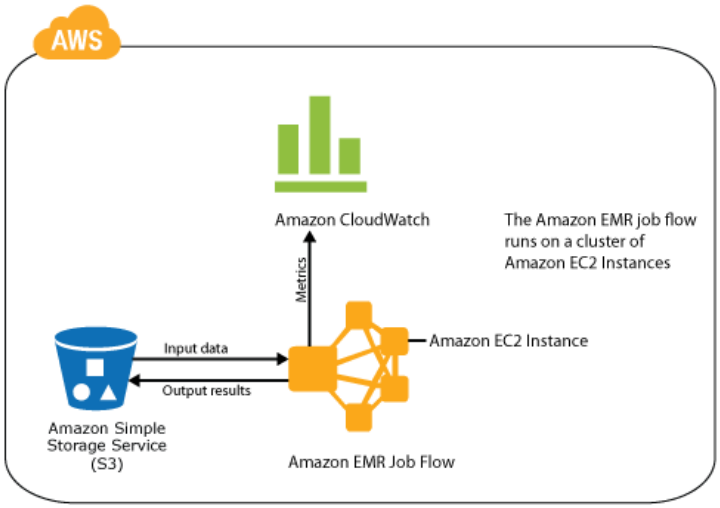

Figure 1.4: Amazon EMR Job Flow

1.2 Problem Statement

Continuous structured and unstructured data are being generated in a large scale every day. These data are very important for most organizations in the

world in order for them to make their decision making. Unstructured data specifically, can help organizations gain a better understanding of their customers, products, services and business in general. For example, data from Twitter streams, social media networks and web logs can help an organization estimates customer sentiment toward a product or service, or help identifies and addresses a potential service or quality issue before it becomes a full-fledged problem.

However, these data are meaningless if they are not being captured and analyzed accordingly. Marr, B. (2015) quoted that less than 0.5% of all data generated digitally is ever been analysed and used. Traditional RDBMS technology becomes less reliable when dealing with huge amount of structured data and the processing speed of data becomes sluggish if the infrastructure is not being upgraded to match the big amount of data. Furthermore, RDBMS is not capable to deal with unstructured data. Mcnulty, E. (2014) estimates that unstructured data contributes at stunning 90 percent of all digital data, much of which is locked away across a variety of different data stores. Therefore, this thesis aims to use cloud based infrastructure which is Amazon Web Service to capture unstructured of big data, and afterward analyzing, visualizing and extracting useful information from large, diverse, distributed and mixed of data gathered from public data sets and Twitter's Application Programming Interface (API).

1.3 Research Objectives

Based on the identified problems above, the objectives throughout this study are:

1. To collect unstructured big data from Twitter Application Program Interface (API) by using Amazon Cloud Service
2. To capture and analyze unstructured big data by using sentiment analysis.
3. To measure the elapsed time when collecting twitter data.
4. To measure the elapsed time and CPU usage when performing MapReduce jobs in Amazon Elastic MapReduce (EMR).

1.4 Research Questions

The study will address the following research challenges in order to solve the identified problems:

1. How to provide cloud based infrastructure to captured and manipulate large volume of unstructured big data to solve the data processing?
2. How the cloud technologies are proficient to capture and gives meaningful output to unstructured big data to solve the meaningless data?

1.5 Research Scope

The scopes of this study are:

1. This study is using Amazon Web Service such as Amazon Elastic MapReduce (Amazon EMR), Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3) with the limitation of free tier services, as the tools of Cloud Technologies.
2. Unstructured data are driven from Twitter Application Programming Interface (API) only.
3. This study is using Amazon EMR to perform sentiment analysis for unstructured big data.

1.6 Thesis Structure

Chapter 1 describes about introduction of research background, problem statement, objectives and scopes of the research. In this chapter, the study mentioned about Big Data and its characteristics, also explained about general information of cloud technologies environment such as Amazon Web Service and their approach in handling Big Data processing.

Chapter 2 starts with a highly focused review of literature provides a detail discussion on the related work in collecting and processing large unstructured data sets in cloud computing environment.

Chapter 3 explained about research methodology and framework design including algorithms, sample codes and few techniques that have chosen to run the experiments on big data.

Chapter 4 shows explanation about results and steps of the experiments done in the previous chapter. The analysis highlights on the elapsed time when collecting twitter data and also the performance of Amazon Elastic MapReduce (EMR) starts from creating the cloud environment and cluster, processing the twitter input data and produce a meaningful output.

Chapter 5 is show a conclusion and future research recommendation. It also briefs some summary of the research presented, limitation and offer future work recommendations for related research in the field of Big Data environment using cloud infrastructure.

1.7 Summary

Unstructured Big Data processing is requiring a technology of cloud when it's dealing with a huge amount of data. Amazon web service is one of the Cloud providers to capturing and analyzing the unstructured data by using sentiment analysis to become meaningful output.

REFERENCES

Bakshi, K. (2012). Considerations for big data: Architecture and approach. *In Aerospace Conference, 2012* (pp. 1-7). IEEE.

Bian, J. T. (2012). Towards large-scale twitter mining for drug-related adverse events. *In Proceedings of the 2012 international workshop on Smart health and wellbeing* (pp. 25-32). ACM.

Das, T. K. (2013). Big Data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, 153.

Gharehchopogh, F. S. (2011). Analysis and evaluation of unstructured data: text . *In Application of Information and Communication Technologies (AICT)* (pp. 1-4). IEEE.

Jadeja, Y. &. (2012). Cloud computing-concepts, architecture and challenges. *In Computing, Electronics and Electrical Technologies (ICCEET)* (pp. 877-880). IEEE.

Lin, J. &. (2012). Large-scale machine learning at twitter. *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 793-804). ACM.

Mohammed, E. A. (2014). Applications of the MapReduce programming framework to . *BioData mining*, 1.

Narendra, B. S. (2016). Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies. *International Journal of Intelligent Systems and Applications (IJISA)*, 8(8), 66.

Shawky, D. M. (2012). Defining a measure of cloud computing elasticity. (pp. 1-5). IEEE.

Smedt, T. D. (2012). Pattern for python. *Journal of Machine Learning Research*, (pp. 2063-).

Troussas, C. V. (2013). Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. *In Information, Intelligence, Systems and Applications (IISA), IEEE*, 1-6.