AN EFFECTIVE DATA ANALYSIS SYSTEM ON BIG DATA ANALYTICS

BY

Badhan Saha Setu
ID: 151-25-462

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By
Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
DECEMBER 2018

APPROVAL

This Thesis titled **"An Effective Data Analysis System On Big Data Analytics",** submitted by Badhan Saha Setu (ID: 151-25-462) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of MIS/M.Sc. in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

`

_____

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Internal Examiner

Dr. Md. Ismail Jabiullah
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Internal Examiner

Md Tarek Habib
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Muhammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this thesis has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of CSE** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:

_____

Badhan Saha Setu
ID: 151-25-462
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete this thesis successfully.

I really grateful and wish our profound our indebtedness to **Dr. Sheak Rashed Haider Noori**, **Associate Professor and Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "Big Data Analytics" influenced me to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to **Dr. Syed Akhter Hossain**, Professor and Head**,** Department of CSE, for his kind help to finish my thesis and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Big Data Analytics is a very important topic which have recently gained interest among researches in the Information and technology areas, which provide a vast amount of user-generated annotations and reflect the interests of millions of people. In this thesis, I proposed and discuss about a system which is very effective and organization benefited model for analyze the unstructured data using Apache Hadoop framework. The initial findings obtained from analyzing social media data like reddit.com, facebook.com. Apart from investigating bookmarking and tagging patterns in that data, we discuss evidence that how the unstructured data is to push this system to analyze and how the system process that types of data. Social bookmarking systems data are an example of this system input. Here I present a method, how to analyze data in an efficient way in this system. This system is not only for analyze social media data it analyzes any kind of unstructured data in data analysis and provide conclusions and directions for future research. All of researcher who are select their research field on data analysis they are very easily implement their concept in this model. This is a conceptual model of a data analysis using big data technology. This is not a research work it is a system where other researchers can research and analyze their idea in this system. I think it will be the appropriate system in this modern Information Technology world.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

In the Information era, everything are going to digitalized and automation process. Since information systems generate enormous amounts of records every day, every second, it seems the world is reaching the level of data overload. It is obvious now, that in order to process such volumes of data an enormous capacity is required in terms of storage and computing resources. Whereas the growth of capacity is limited by evolution of hardware and technologies, the growth of the data volume is in fact unlimited. Especially unstructured data is growing like the flow of sea water compared to semi structured and structured data. Massive uses of chat, emails, high concurrency application such as Facebook, search engines, Amazon, official documents, video/audio/ image etc. which use diverse type of non-relational databases (commonly referred as NoSQL; Not only structured query language). [1]

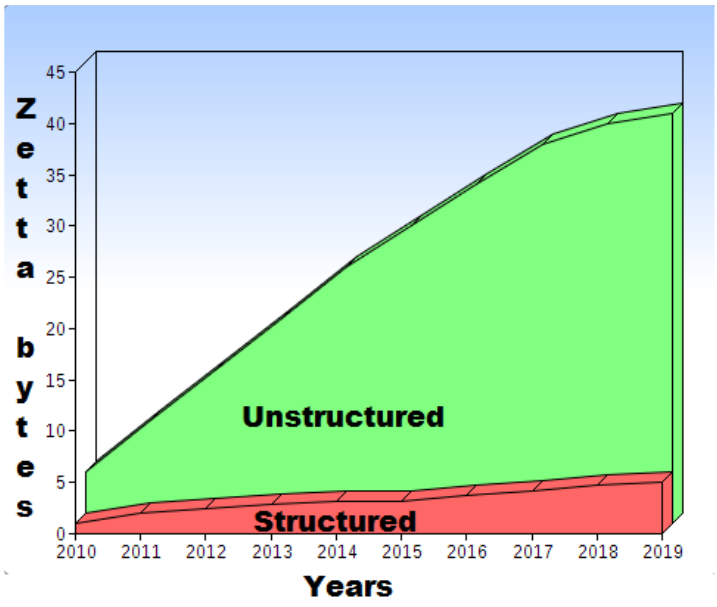

Figure: 1.1.1: Data growth per year [15]

The percentages of different unstructured data growth are shown in below figure 1.1.2

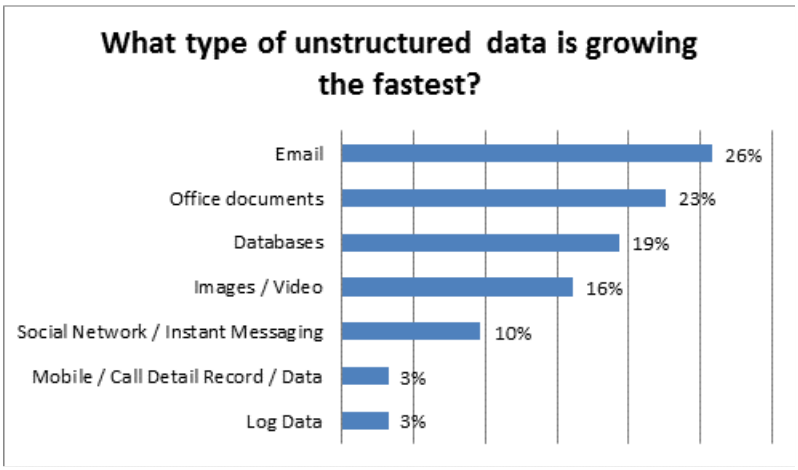

Figure 1.1.2: The percentages of different unstructured data growth. [3]

Beside the growth of this data, people also handle and process this huge amount of data by constructing huge datacenter and using high-end powerful computing systems, which consumes more space, more electricity, more man powered and more money. Again it is difficult to handle structured and instructed data at a time in same structure and framework.

1.2 THESIS OBJECTIVES

This thesis will review the Big-Data Analytics and proposed a system which is a data analysis platform, which analyze very big range unstructured data and generate result in a short time using Apache Hadoop framework of Big Data technology. The vital objective is listed below:

- Analyze and understand how the unstructured data is processing
- Understand Hadoop API
- Understand HDFS
- Understand Pig and Hive query
- Understand Hadoop MapReduce operation
- How Hadoop process unstructured data to semi structure format.

1.3 THESIS CASE STUDY

In this digitalized era, there are lots of data analysis technique are used to analyze the data in need basis. Some of them are very good and some of them are below average. But why we need a new system to analyze data in our life. One point is there are lots of good quality data processing and analyzing software but all of them are not process the unstructured data. Unstructured data means the data which is not structured format or text format, it could be audio, video, sound, picture, code, embedded software's data, and sensor and Internet of Things (IOT) data. So traditional database are not processing this kind of data. This is a very big problem in our working and research life. So Big Data database like NOSQL database process this type of data. But other reason is the time, cost and accuracy of data processing. There are lots of framework in Big Data technology. There is some problem in quality of service because all of them are not give the quality of service. So, in this thesis, there are proposed a very low cost system which provides a very good output in a short time.

CHAPTER 2

BACKGROUND OF BIG DATA

2.1 LITERATURE REVIEW OF BIG DATA ANALYTICS

A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value, it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value. Big data and analytics are intertwined, but analytics is not new. Many analytic techniques, such as regression analysis, simulation, and machine learning, have been available for many years. Even the value in analyzing unstructured data such as e-mail, video, audio and documents has been well understood. So, for the importance it can be used in different sectors in our society. Some of them are customer support and handling, optimizing business processes, performance optimization, improving healthcare and public health, sports area, improving science and research, improving security and law enforcement, financial field and so on. The below use case diagram are describe usage of Big Data in health sector.

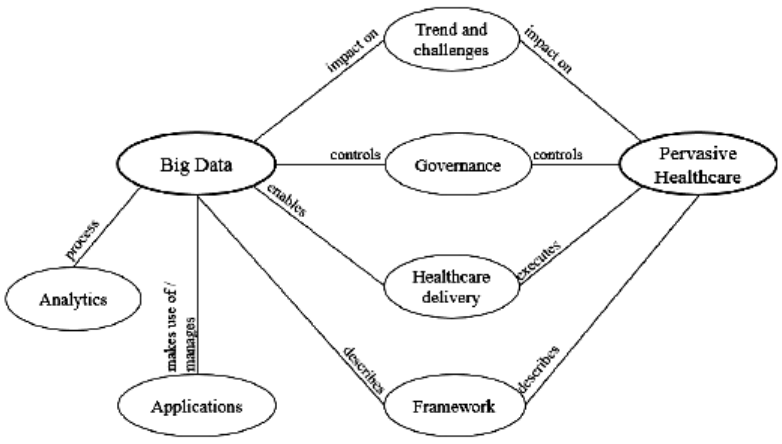

Figure 2.1: Use case diagram of Big-Data usage in health sector

2.2 RELATED WORKS

There are few theses on data analysis using Big Data tools because most of research in Big Data are related to the framework upgradation, deployment and performance analysis. In related of this thesis I select a field name analysis on social book marking data from social media. Because, most of generated data in everyday from social media like reddit.com, Facebook, YouTube, twitter, Skype, Instagram etc. Social bookmarking systems have recently gained interest among researches in the areas of data mining and web intelligence, as they provide a vast amount of user-generated annotations and reflect the interests of millions of people. Social bookmarking systems also provide a promising source for the detection of trends.

The authors of [2] provide an overview about the structure of collaborative tagging systems. Based on a small subset of the del.icio.us corpus, they investigate what motivates tagging and how tagging habits change over time. The authors of [9] present a taxonomy for the classification of tagging systems based on the design choices such as the tagging rights (who can tag what?) or the type of underlying resources (what can be tagged?). The authors also suggest to classify a tagging system according to the incentives of its users. The authors of [10] argue that social tagging systems can be described as tripartite graphs, involving users, tags and resources, extending the traditional bipartite ontology model by the user dimension.

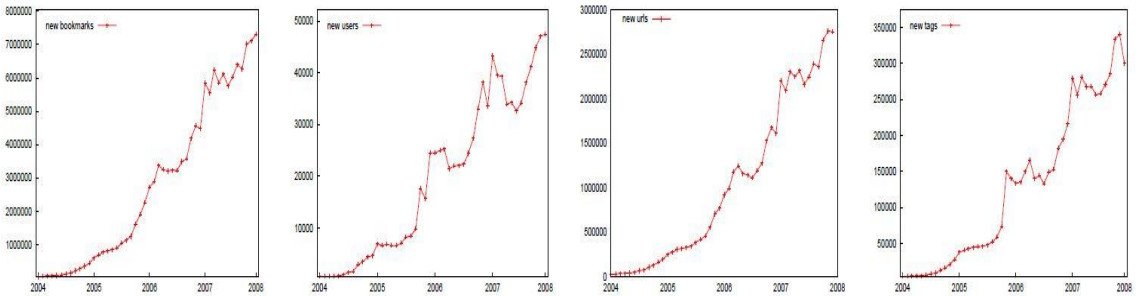

Figure 2.2: The monthly growth of del.icio.us between 2004 to 2008 by posted bookmarks, new users, new URLs and new tags

Based on the tripartite model, the authors show that semantically related tags can be clustered in order to discover emerging ontologies. The integration of collaborative tagging systems with the semantic web concept is also the goal of [11]. The authors combine filter and cluster techniques to extract the semantics emerging from the tag space. Both, [10] and [11], base parts of their analysis on del.icio.us data but only consider data sets with less than 1,00,000 bookmarks.

The authors of [5] apply the tripartite community model also found in [10] and a diffusion technique similar to Google's PageRank algorithm in order to detect trends in social resource sharing communities. Their algorithm enables the authors to rank items (users, tags, urls) with respect to a given topic preference vector. The authors can thus detect trends by comparing the popularity (rank) of items at different points in time.

This type of system is data analysis system. Current situation most of researcher are completing this types of research in some third party build up software or some low quality Big Data Frameworks.

CHAPTER 3

DATA ANALYSIS SYSTEM IN BIG DATA

3.1 EXISTING SYSTEM

Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration. Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as business, entertainment, science, crisis management and politics. For developing the internet and connectivity of computer networks now all of us badly depends on social media. All of us spend your lots of time in a day spent on social media. Currently we post some news, pictures, videos, job vacancy news, research work, messages, communicate with others by social media for example, Facebook, Twitter, Google+, LinkedIn, Wikipedia, Reddit, Pinterest and lots of social sites. We post personal and official news in this social media. Some are select social media as their income source and job in like digital marketing, social media marketing, email marketing and etc. types. Researcher and scientist are upload and share their innovation on social media for expanding knowledge of others in their innovation or model. Now-a-days social media are very popular for income source and communication purpose. So everyday Brillion of posts are appearing in this kinds of social sites. But think a minute that how many data are generate and increase of data in social media day by day. If we think like one bit of data is a small part of a paper, then the whole world will fill up by the small pieces of paper. [15]

All of we have account on any social site and we use that account. On using time, we found lots of advertisement from different company or when we seeing some video we see suddenly advertisement will appear and after completing the advertisement we

see again the video. For reference the authors of [5], [10], and [11] are shows that how many data are generated by user of social media.

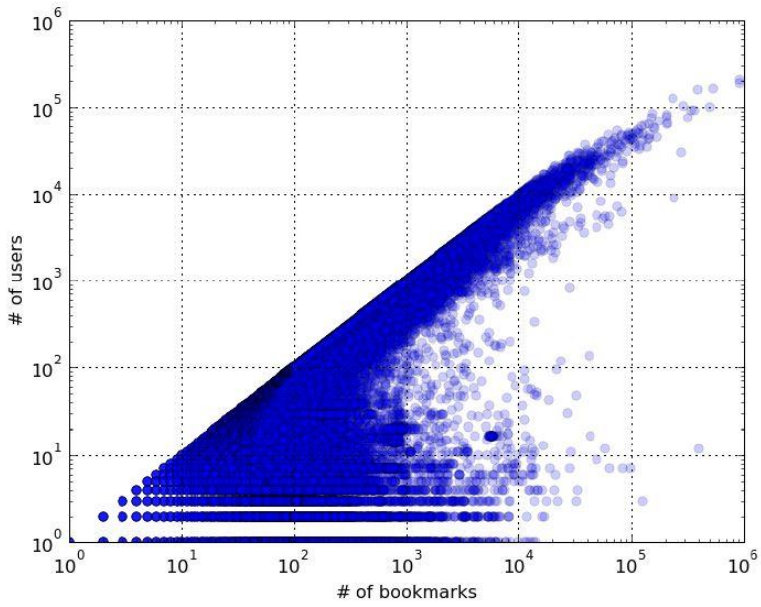

Figure 3.1.1: The number of bookmarks compared to the number of users Linking in a domain.

It will happen in most of social media site for example Facebook, and YouTube etc.

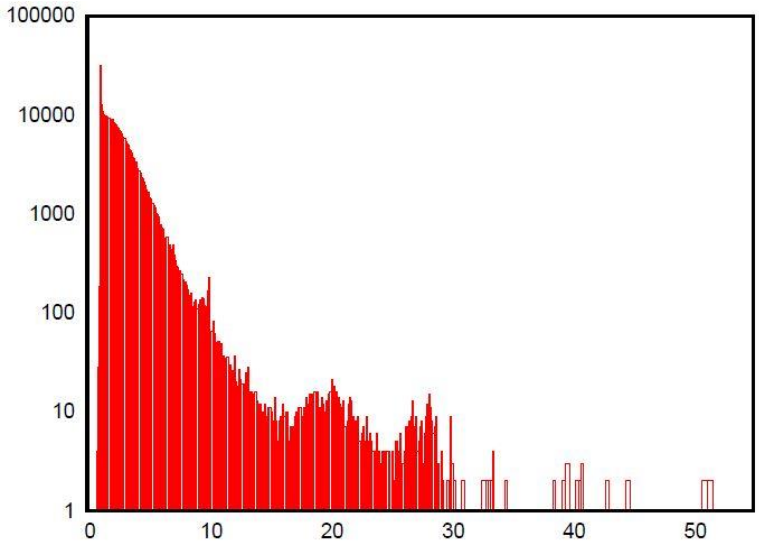

Figure 3.1.2: Tag posted on average by a user.

But thinking topics is how we get this type of advertisement and the advertisements are very closely to our habits or what we want and like. How advertisement will come

or notifying me their update and offer, why this comes, why advertisement is related to my want in social media, who working on behind the screen, who analyze my resource, is this secure. Our absent mind found this types of question and we cannot find the appropriate solution. The answer is every user's data on social media are analysis by some company. Now the point is how they analyze this huge data, are they use more powerful tools and technology to analyze the data sets? The answer is they use a Big Data technology to analyze the data and generate an effective result so that they can read the mind of user. In this reason they can found the related advertisement for specific user and we can see the advertisement in social media. [16] All the status updates, pictures, and videos posted by people on their social media contains information and that information about their demographics, their likes, their dislikes, etc. This information can be analyzed, and its applications are numerous. For example, social media data can be analyzed to reveal the proportion of social media users that enjoy a particular flavor of ice cream at a particular ice cream parlor at any given time of the day. The ice cream parlor can then identify which of his ice creams are best-sellers, and at which times of the day. This technique is known as data analytics, and it is one of the biggest trends of 2018

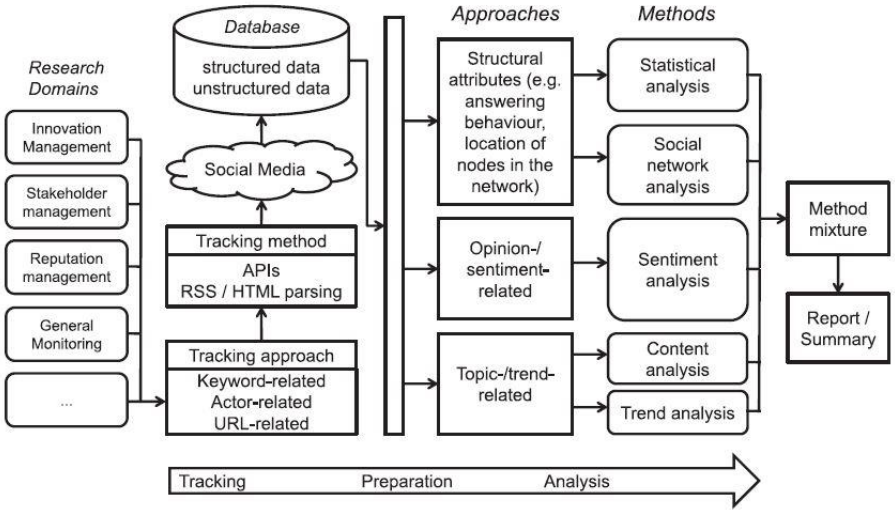

Figure 3.1.3: The total system of Social media data analysis

There are lots of system and tools for data analysis in social media: Sprout Social, Snaplytics, Iconosquare, Buzzsum Tailwind, Google Analytics, ShortStack, TapInfluence and etc.

3.2 PROBLEM OF EXISTING SYSTEM

We understand the tools and One thing it all of above tools are working for data processing their main target is process the data and that's it. But now a day's technology is incising day by day and user want their result in a very short time with accuracy. The main problem is this tools process only structured and little-bit semi-structure data. But when we need to process the data of Facebook, YouTube, Skype, LinkedIn etc. then this kind of tools cannot process the whole data and not serve our purpose. But this kind of third party tools are not reliable because there have security vulnerability and threats and malware issues. In this type of software are not reliable because if user data is leaked then it is very risk for business and promotional purpose. This kinds of software are costly and the authority of software charge more than Open Source Big Data frameworks. Big data approaches to analyzing social media data can increase understanding of how people think and act. Organizations can use this information to inform their activities, improve decision-making, target products and services more effectively, and to try to influence users' behaviors in the future. The rate of unstructured data production on social media makes it difficult to analyze using traditional methods that rely on human analysts. Social media analytics is a new field of study that is developing automated or semi-automated methods for analyzing data. One prominent technique is called sentiment analysis. This can be a useful tool to assess public reaction to a particular event, such as a protest or TV show. However, outside of specific contexts the insights that can be drawn from this technique are currently limited. Research is underway to improve the technology and apply it to wider settings. One example is the WeGov project, funded by the European Commission, which is building tools to analyze responses to government policies on social media. On that time the developer team can understand how Big Data can help them to analyze the data set.

3.3 PROPOSED SYSTEM

The unstructured data are increasing day by day rather than the structure data cannot increase like unstructured data. So it is very important and essential to process and analyze data in a framework which is effectively and accurately generate the result of data analysis. In a picture it will clear how the ratio of structure and unstructured data. Relational models and SQL provide an abstraction layer between the database's physical layer and the application layer. This feature lets users specify a query in a language dependent and declarative manner, while a query engine schedules and optimizes its execution. No similar solution exists for big data analysis.

Instead, NoSQL data stores offer various forms of data structures — such as document, graph, row-column, and key-value pair — that are directly exposed to users. So, users must understand data's physical organization and employ vendor-specific APIs to manipulate these data. Current state of the art attempts to devise a SQL layer on top of NoSQL, but without an abstract data model, this effort is ad hoc and limited to the underlying technology. Transactional data is processed initially on an online transaction processing (OLTP) system before flowing through an extract, transform, and load (ETL) process in a batch mode. Eventually, data are loaded into an online analytical processing (OLAP) data warehouse, where they're analyzed to provide strategic insights. This OLTPETL-OLAP approach trades timeliness for accuracy, given that a long delay occurs between when data becomes available and insight generation.

Aggregating data from multiple social networks enables data analytics that correlate the dataset's various networks. Given that social networking vocabulary varies from one network to another, we anticipate the need for cross-domain vocabulary mapping as a data preprocessing step. For example, the Twitter glossary defines terms such as "followers" and "tweet." Facebook defines terms such as "friends" and "status." Google Plus uses "circles" and "hangout." To perform cross-domain data analytics, we must develop and maintain a common ontology that will capture the differences

and similarities in terminologies and define relationships between terms within and across the network.

Finally, in this thesis I proposed a system that collect the form user or social media data analysis and bookmarking site for example, reddit.com data and analyze the data set in below proposed model:

- Use MapReduce code to convert XML data file to flat file like comma separated value (.csv) file.
- MapReduce Mymapper function which API is developing by Java program. So, Mymapper map the input data which is flat file formatted.
- On the mapping time XML driver use to import raw dataset and support the Hadoop framework all library file.
- CSV Input format will help to import all data.
- Use a PIG Script to analysis the output or result which from the Apache Hadoop MapReduce framework.
- Use an Open Source Linux based shell script to create some scripts which will allows user to bookmark, rate, review, and track the record on time & date on specific topics or various link on any topic.

3.4 THE WORK FLOW ALGORITHM

Step 01: Get the data set from cloud or user input panel.

Step 02: Convert web based input data to readable format.

Step 03: Use Mymapper for convert data to flat file format

Step 04: XML driver of Hadoop MapReduce import raw dataset.

Step 05: In data mapping time, XML driver upload the Hadoop Library File.

Step 06: MapReduce framework shuffle and sort the data.

Step 07: Reduce the sorting Data and prepare an output of short range data set.

Step 08: PIG script analyze the output data set and prepare a semi structure data as a primary output.

Step 09: Open Source Shell Script language Hive query to reanalyze the primary output data and insert the output into NOSQL Database.

Step 10: The relational Database will process the user level query and shows the output in user readable format.

CHAPTER 4

DESIGN AND IMPLEMENTATION

4.1 SYSTEM REQUIREMENT

To deploy this infrastructure in a system we will need system requirement as below.

- Operating System: Red Hat Enterprise Linux 7.0
- Processor: Intel core i7 processor-8250U @ 3.2 GHz
- System Memory : 16GB to 32GB
- Storage: 256GB SSD (For OS) and 4TB HDD
- VMware Player :VMware-player-4.0.6-1035888
- Sandbox and Hortonworks : Hortonworks+Sandbox+1.3+VMware+RC6
- Sample Database : NYSE-2000-2001.tsv
- Framework: Apache Hadoop framework version 2.9.2 (Hadoop 2.9.2.tar.gz)
- PIG Framework download link: https://archive.apache.org/dist/pig/
- Hive query Language: version 0.17.0 ; artifact : /dist/PROJ/foo.tar.gz [10]

4.2 SYSTEM DESIGN AND IMPLEMENTATION

The system is a conceptual system which is used Apache Hadoop Framework and its some feature.

Apache Hadoop MapReduce is needed with PIG and Hive query language to generate a result based report. First collect data from social media or social bookmark rating site, most of time data are in XML format. MapReduce convert the data to user-friendly format as like comma separated value (.csv).

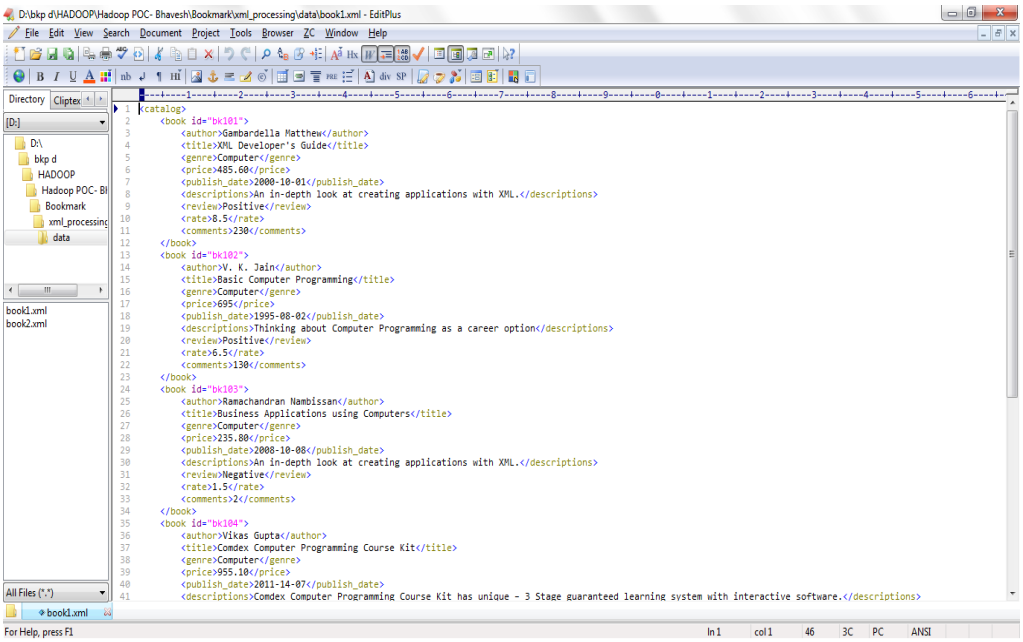

Figure 4.2.1: MapReduce Implementation

This output is comes from MapReduce technology but file format conversion is done by Hadoop Distributed File System (HDFS). We called the output as HDFS output then push this output and feed into PIG platform for creating programs which splits the result into some categories. This categories data called semi structure data.

```
import java.io.ByteArrayInputStream;
import java.io.IOException;
import java.io.InputStream;
import javax.xml.parsers.DocumentBuilder;
import javax.xml.parsers.DocumentBuilderFactory;
import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.w3c.dom.Document;
import org.w3c.dom.Element;
import org.w3c.dom.Node;
import org.w3c.dom.NodeList;
```

Figure 4.2.2: Partial code for HDFS conversion system

This semi structure data need to analyze in future in Hive project to process in relational database management system (RDBMS). Now, Hive query to make the semi-structure data to a structure format and then Hive query generate a result which is very structure format.

```
hadoop fs -rmr /BookMarkOutput/Type_Computer
hadoop fs -rmr /BookMarkOutput/Type_Database
hadoop fs -rmr /BookMarkOutput/Rating5+
hadoop fs -rmr /BookMarkOutput/Rating5-
```

Figure 4.2.3: Partial code for PIG query of system

```
hive -e 'drop table if exists ComputerBooks';
hive -e 'drop table if exists DatabaseBooks';
hive -e 'drop table if exists Highest_Rating';
hive -e 'drop table if exists Lowest_Rating';

hive -e "create external table ComputerBooks
(Bookid string,
author string,
title string,
genre string,
price float,
publish_date string,
```

Figure 4.2.4: Partial code for Hive query of system

Now, the output of Hive query are ready to process and analysis. In this proposed system sqoop technology push the Hive query output into relational database management system (RDBMS).

4.3 WORKING FLOW OF SYSTEM:

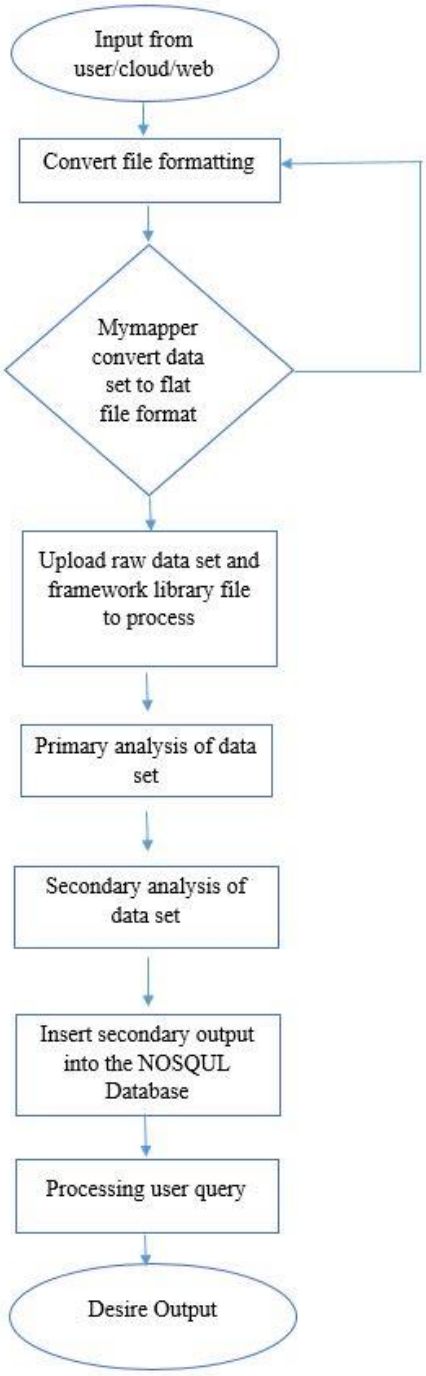

Figure 4.2.5: Work Flow process of System

Overall of the Infrastructure of this system are use very strong security protocol name MD5 and Kerberos password for ensuring data security in the processing time. In Linux based all operating system is used this two security protocol in different storage allocation and sharing time. Here using two security protocol and Hadoop framework for ensuring high data security and well organize in cost cutting methodology. The backend of this system will run sqoop which is command line interface application for transferring data between relational database and Hadoop. So here Sqoop output is input for relational database. So, in this system every component are set very organized way.

CHAPTER 05

IMPACT ANALYSIS OF PROPOSED SYSTEM

5.1 SYSTEM DEPLOYMENT

This proposed system is a system where one can analysis not only social media data but also all kinds of Big Data analytics research based data. It is a platform where one can have established in his research work step by step and very sequential way. In this system deploy in corporate and personal research work field. Now a day most of Inter-Government and International research center start their operational field survey and data analysis of different types of pattern. In Integrated Rural Development area, we can use this system to analyze their data. The main criteria of IRD is-

- Poverty Alleviation
- Environment & Climate change
- Income generation
- Agriculture
- Sustainable Development Goals
- Agriculture
- Agro-Processing
- Blue Economy
- Good Governance
- Health
- Information Management
- Renewable Energy
- Water & Sanitation

On the above area it is already used and there have some scope to start analyze data in this system. This proposed system will used not only rural development field but also in science and ICT sector. Like, when a person want to established a business in online

based then owner must have known about client want, client habits and clients satisfaction with related product when the business is startup. The successful businessman must have analyze the data of his product related. This is the best practice when he start his business on a big deal.

NASA already start their data analysis on Big Data using Apache Hadoop platform in Earth Observing System Data and Information System project.

5.2 EVALUATION OF PROPOSED SYSTEM

The evaluation part is very important for a system and it depends on some criteria. In the below figure shows the indicator of evaluate a system.

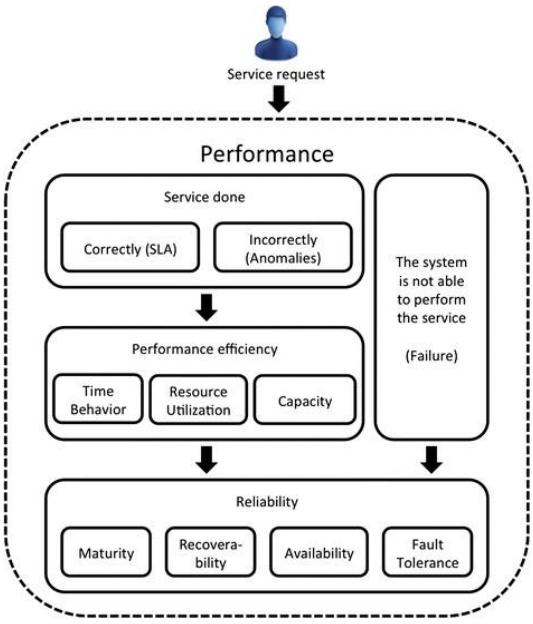

Figure 5.2: The Indicator of evaluate a system

i. Service: In Apache Hadoop frame work the service of a system better than other Big Data framework. It is very effective framework where any kind and any level of unstructured data able to process. In Big Data concepts, Hadoop technology is very powerful because this framework is based on

some powerful and efficient technology named MapReduce, HIVE, PIG, YARN, and HDFS.

ii. Data Processing: In this system used Apache Hadoop framework because the Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is developed for Very Big data set and its data processing rate so much high. So, the data processing rate is too much higher than other framework.

iii. Accuracy: The accuracy is a vital question when evaluation team are evaluate a system. This is generate the data set very accurately using Hive and Pig query.

iv. Security of Data: In this system used two security protocol which is used in Open Source operating system to protected data in a specific storage named MD5 and Kerberos password security policy.

v. Power Consumption: Less power consumption is the one of the biggest feature of Big Data Apache Hadoop framework. In Spark frame work need more system memory and processor to run its activity but Hadoop needs very less system memory, processor and power to up a system and execute every operation in a system.

vi. Cost: This system is developing on Open Source framework and open source operating system so that there is no need money to purchase system software and framework. Just physical infrastructure purchases and power cost is the total cost. So there are huge save the money if used this system.

vii. Plugin and version Update: This system is totally deploying in end user level and there is no need to purchase the enterprise plugin. So in the open source there are no complexity to plugin and version update.

5.3 SCOPE OF PROPOSED SYSTEM.

This system we can use in every sector in our real life where we need to analyze lots of data set. International and national level research institute, Hospital, Bank and Insurance Company, University and educational organization, NGO and INGO, Training Institute, Statistical Organization, Government Body and Ministry, Big Corporate Business, Agricultural institute and other institute where have data analysis is very important.

5.4 ADVANTAGE AND DISADVANTAGE OF PROPOSED SYSTEM

Advantage:

I. The performance is higher than other existing system.
II. This system is more secure because there are no chance to leak the organization data in web.
III. This is very design for Big data set so that lots of data can be process parallel way.
IV. This system is easy to use that anyone can process the data when he know the system password.
V. The cost of this system is cheaper than other system.
VI. This system operate with lots of supporting technology so that the data processing rate is higher.
VII. This system is limited plugin and also deploy on open source operating system so that there are no problem to update the plugin.
VIII. The processing data (output) are highly accurate because this system framework are used very reputed and big organization to analyze their data. If result will not accurate they cannot use it.
IX. This system is need less power supply.
X. This system is very suitable, if organization can switch their activity in spark or other framework it can be possible in this system.

Disadvantage: There few disadvantage in this system. The main disadvantage is this system framework named Apache Hadoop is designed for big data set and big volume. So in small data set the performance is not higher than other framework. Other side it can be need more developer when a big change in Infrastructure.

CHAPTER 6

CONCLUSION

6.1 FUTURE WORK

There are most of the world's total data is generated in last two years and with the increase in the use of digital and electronic devices and technologies like machine learning and IOT, there cannot be a degradation in the Big Data and its technologies for 100's of coming decade. In this system not only a dedicated platform of data analysis. This system will provide a research based solution for data scientist. In future plan, it will be integrated with cloud and IOT for analysis data from different system and different environment. There are scope to working with this system to business automation and digital marketing with taking result from existing technology.

Currently in this proposed system data can inserted in this system from local storage by user. But it is possible to adapt this platform with cloud server to collect data automatically and generate output and give some decision. After getting that output there will be some scope to add some IOT device take some initiative to doing the right work.

The system could be expanded further to allow the possibility for real time simultaneous analysis of multiple data set. This would require the restructuring of the monitoring system to be able to store large amounts of data. Furthermore, the future analytics system would need to process this data more efficiently strategies. Another possibility would be the integration of more social media platforms with different analysis strategies and static government data, to contribute to the complete analysis of Sustainable Development Goals.

6.2 CONCLUSION

Apache Hadoop is a framework where large and very big data set can be process in a very sequential way for unstructured data within a short time. This system is proposed for data analysis where the data are unstructured and size is very huge. Social media and social bookmarking system data is a single part of this system where the data set can be process in a very short time but output is very effective and accurate. The user can process different types of query in this proposed system. This system used Hadoop MapReduce technique in Apache Hadoop which operates exclusively on <Key, value> pairs that is the framework view the input to the job as a set of <Key, Value> pairs and produces a set of <Key, Value> pair as the output of the facilitate sorting by the framework. There have a MapReduce User Interface which provides a reasonable amount of detailed on every user facing aspect of the MapReduce framework. This should help users implement, configure and tune their jobs in a fine-grained manner. The Mapper option maps input key/value pairs to a set of intermediate key/value pairs. On the other section two powerful key components of Hadoop ecosystem named Hive and PIG used in this system. The PIG is used here for programming and Hive is used for creating reports. They helps unstructured data to build up a structured format of data set. Then Sqoop will help to process that data to send into the relational database. All of the above technology makes this system very effective and time saving data processing model. If any organization used this system they will more benefited from getting real time data processing feature with big volume, big velocity and variety of data processing feature.

REFERENCE

[1] Learn about Big Data Definition, available at http://www.wikipedia.org/, last accessed on 11-09-2018 at 7:00pm

[2] Scott A. Golder and Bernardo A. Huberman, 'Usage patterns of collaborative tagging systems', Journal of Information Science, 32(2), (2006).

[3] Learn about Big Data, available at https://www.techopedia.com/definition/27745/big-data, last accessed on 12-09-2018 at 10:00pm

[4] Learn about Importance of Big Data, available at https://insidebigdata.com/2017/09/09/big-data-important-business/, last accessed on 12-09-2018 at 10:00pm

[5] Andreas Hotho, Robert J¨aschke, Christoph Schmitz, and Gerd Stumme, 'Trend detection in folksonomies', in SAMT, volume 4306 of Lecture Notes in Computer Science. Springer, (2006).

[6] Wissem Inoubli, Sabeur Aridhi, "An Experimental Survey on Big Data Frameworks" Semantic Scholar vol. 10.106, pp. 2-3, July-2018.

[7] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis, 'Ht06, tagging paper, taxonomy, flickr, academic article, to read', in HYPERTEXT '06, New York, NY, USA, (2006). ACM.

[8] Peter Mika, 'Ontologies are us: A unified model of social networks and semantics', J. Web Sem., 5(1), (2007).

[9] Lucia Specia and Enrico Motta, 'Integrating folksonomies with the semantic web', 624–639, (2007).

[10] Learn about HDFS, available at http://www.wikipedia.org/, last accessed on 11-11-2018 at 10:00pm

[11] The Apache Software Foundation. Hadoop. Hadoop, available at http://hadoop.apache.org/ last accessed on 22-10-2018 at 10:47pm

[12] Learn about Spark, available at http://spark.apache.org/ last accessed on 23-10-2018 at 12:02 am.

[13] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina, 'Can social bookmarking improve web search?' in WSDM, July, 2018.

[14] Beate Krause, Andreas Hotho, and Gerd Stumme, 'The anti-social tagger-detecting spam in social bookmarking systems', in Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web.

[15] Learn about Big Data, available at https://www.geospatialworld.net/ last accessed on 29-11-2018 at 10:00 am.

[16] Learn about Big Data in Public Health, available at https://www.apple.com/ last accessed on 1-11-2018 at 9:00 am.

[17] Learn about Big Data in Science & Research, available at https://www.nytimes.com/ last accessed on 09-11-2018 at 11:00 am.

[18] Learn about Big Data in Security and Law Enforcement, available at http://www.dataversity.net/big-data-at-the-nsa/ last accessed on 23-11-2018 at 8:00 am.

[19] Learn about Big Data frameworks, available at https://veribilimleri.wordpress.com/ last accessed on 28-11-2018 at 10:00 am.

[20] Learn about Process of MapReduce, available at https://www.slideshare.net/ last accessed on 29-11-2018 at 10:00 am.