OPEN ACCESS
53
https://humicpress.org/journals/index.php/ijohce
Intl. Journal of Human Centric Engineering
Vol. 1, No. 1, Mar 2024. pp.53-63

Social Media-based Polling Data Processing using MapReduce on the Hadoop Framework

Yusuf Yunadian[1] , Hilal H. Nuha[2], Sidik Prabowo[3]

[1,2,3]*Faculty of Informatics,* [1,2,3]*Telkom University*
Bandung

[1]yyunadian@student.telkomuniversity.ac.id, [2]hilalnuha@gmail.com,
[3]sidikprabowo@telkomuniversity.ac.id

Abstract

The processing of data obtained from a polling system holds significant importance as the results can serve as a reference for addressing community issues. With the escalating use of social media, Indonesia ranks fifth globally in Twitter usage. When dealing with substantial data sets, processing speed becomes a concern, prompting the author to develop a more time-efficient system for handling polling data gathered through social media. Hadoop emerges as a prime candidate for this purpose due to its two main modules: Hadoop Distributed File System (HDFS) for distributed storage and MapReduce for algorithmic computation. Through experimentation using both the wordcount program with MapReduce on Hadoop and the version without MapReduce, it was observed that MapReduce outperforms the latter in terms of data processing speed. On average, employing MapReduce on Hadoop resulted in a 1.3 times faster data processing rate compared to the method without MapReduce.

Keywords:Polling, Hadoop, MapReduce, wordcount, processing speed.

Abstrak

Pengolahan data yang diperoleh dari sistem jajak pendapat memiliki arti penting karena hasilnya dapat menjadi referensi untuk mengatasi masalah-masalah yang ada di masyarakat. Dengan meningkatnya penggunaan media sosial, Indonesia berada di peringkat kelima di dunia dalam hal penggunaan Twitter. Ketika berhadapan dengan kumpulan data yang besar, kecepatan pemrosesan menjadi perhatian, sehingga mendorong penulis untuk mengembangkan sistem yang lebih efisien dalam menangani data jajak pendapat yang dikumpulkan melalui media sosial. Hadoop muncul sebagai kandidat utama untuk tujuan ini karena dua modul utamanya: Hadoop Distributed File System (HDFS) untuk penyimpanan terdistribusi dan MapReduce untuk komputasi algoritmik. Melalui eksperimen menggunakan program wordcount dengan MapReduce di Hadoop dan versi tanpa MapReduce, terlihat bahwa MapReduce mengungguli versi tanpa MapReduce dalam hal kecepatan pemrosesan data. Secara rata-rata, penggunaan MapReduce di Hadoop menghasilkan kecepatan pemrosesan data 1,3 kali lebih cepat dibandingkan dengan metode tanpa MapReduce

Kata Kunci: polling, Hadoop, MapReduce, wordcount, kecepatan proses.

I. INTRODUCTION

A. *Background*

Social media users in Indonesia are increasing along with the number of smartphone (smart phone) users. The function of social media today is not just to communicate, share information and exist. The public can also conduct opinion polls on social media. The polling method is the most effective method for obtaining quick information about problems or issues developing in society. Polls are used to obtain information about a phenomenon, in this case what we want to obtain from polls are people's attitudes, views and beliefs towards developing issues [1]. Therefore, it can also be said that polling is the practical application of survey methods, the use of survey methods to measure public opinion such as political issues. Social Media is a communication platform that has the ability to quickly find out public opinion on an issue [1].

Usually, to conduct a poll we need at least large enough data, to get maximum results. This large data is often referred to as Big Data. Big Data is a collection of data that is large, very varied, and may be unstructured [2]. Big Data is so large that it is difficult for conventional procedures to analyze big data. With large data, the analysis of a phenomenon can be more perfect, and if successful in analyzing the data it will help in making better decisions [2]. It is also hoped that the data from the analysis can efficiently represent the votes obtained, cheaper boarding, and speed in processing. So a special algorithm is needed so that in-depth information is easily obtained and helps in making better decisions.

Hadoop is a distributed system intended for processing large data [3]. Hadoop is a framework for large-scale data storage and processing comprises various modules, including Hadoop Common and Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce [4].

This final project implements the MapReduce method in the Hadoop Framework which is used to process polling data to make it more efficient in terms of data processing time. The data to be processed comes from the web (happypoll.id), where respondents who will carry out the poll are required to log in first. This is to obtain information, including address, username, email, and other respondent Twitter account information. Data processing utilized MapReduce computing with the wordcount program, which was carried out several times. Namely processing polling data with a wordcount program with MapReduce on Hadoop and a wordcount program without MapReduce or a regular wordcount program in Java language. This was done to find out how effective it would be if Hadoop was used to process social media-based polling data in this final assignment.

B. *Formulation of the problem*

Based on the background described above, there are several problems discussed in this final assignment, namely as follows:

1. How to design a system for processing social media-based polling data by utilizing MapReduce on Hadoop?
2. How to implement the wordcount program with MapReduce on Hadoop to process polling data?
3. How fast is wordcount data processing with MapReduce and wordcount without MapReduce?
4. Can Blocks affect data processing in Hadoop that is set up with a Multi Node Cluster?

C. *Scope of problem*

The limitations of the problems in this final assignment are:

1. The data used is data from poll results on the happypoll.id website, where voters must log in first via Twitter.

2. Poll data is in the form of a customized .text file.
3. Data processing utilizes MapReduce computing on the Hadoop Framework with the wordcount program.
4. Tests were carried out to compare the word count program utilizing MapReduce on Hadoop compared to the word count program without Hadoop.
5. Testing was carried out on Hadoop with an increased block size.

D. *Objective*

The objectives of this final assignment are as follows:

1. Create a social media-based polling data processing system, using MapReduce in the Hadoop framework.
2. Running a wordcount application that runs with MapReduce computing on the Hadoop Framework configured in a MultiNode Cluster.
3. Testing the wordcount program with MapReduce Hadoop and without MapReduce Hadoop, to find out which method is superior in terms of data processing speed.
4. Testing the wordcount program with MapReduce Hadoop, by increasing the block size to determine the effect of the block.

E. *Writing Organization*

This TA research is structured with the following structure: The first part is an introduction which contains the background, problem formulation, problem limitations, objectives and writing organization. The second part contains related studies, references related to the author's TA. The third section contains an explanation of the system being built. The fourth part concerns the analysis and evaluation of the system being built. The fifth section contains conclusions and finally the bibliography.

II. RELATED STUDIES

A. *Polling Using the Internet*

In 2015 Hays, Liu, and Kapteyn in their journal stated that using the internet to collect survey data is cheaper and takes less time compared to traditional methods. The landscape of survey research has evolved, showing reduced reliance on interviews and a growing adoption of innovative technologies to collect data [5]. However, there are still many things that need to be learned about the advantages and disadvantages of using the internet to collect survey data, such as using the web. And in the future there is an opportunity for mobile devices and social media platforms to do the same.

Increased ownership and use of hardware among students benefit from enhanced learning opportunities, fostering faculty innovation in creating engaging educational settings. With widespread Wi-Fi availability across campuses, students can utilize their mobile devices effectively (cellphones, tablets, or laptops) to engage in learning, especially about leadership concepts. Noel, Stover, McNutt in 2015 proposed mobile-based polling as an Audience Response System (ARS) [6]. And it turns out that the results of the mobile-based polling experience show that students become very involved on three levels (behavioral, emotional, and cognitive). Additionally, responses from surveys suggest mobile-based polling is feasible for use outside the classroom[6].

From several of the studies above, the author also used poll data obtained from poll results using the internet in this research. Where the poll is carried out on a website that requires voters to log in with Twitter first. By utilizing the Twitter API to obtain personal data from voters conducting polls

B. Big Data

Big data is simply data that requires processing capacity beyond processing in conventional database systems. In general, data that falls into the big data category is data with a volume exceeding one tera-byte. The characteristics of big data include: volume (size), velocity (speed), variety (variety), and veracity (data uncertainty) [7].

C. Apache Hadoop

In 2017, Merla and Liang in their paper conducted research on the use of Hadoop in processing data from YouTube social media. The system built starts from extracting data from the YouTube API, saving data to the Hadoop Distributed File System (HDFS), Mapper and Reducer to visualizing results using pie charts regarding trending videos based on categories. From the results of YouTube data processing carried out using Hadoop, a conclusion was obtained in the form of information on trending YouTube videos based on categories [8].

Hadoop is the most popular tool in Big Data, which is widely used in social networks such as Google [9]. Hadoop makes it possible to process large data in a distributed manner involving clusters of computers [4]. Hadoop has several advantages such as efficiency in execution time, minimizing computing costs, and increasing performance [10]. Hadoop, an open-source framework, is employed for handling extensive datasets. in parallel and has several modules, including:

1. Hadoop Distributed File System (HDFS)

   HDFS is a distributed storage system that can store files distributed on HDFS nodes.

2. Hadoop Common

   Hadoop Common are libraries and tools required by other Hadoop modules.

3. Hadoop YARN

   Hadoop YARN is used to manage the resources that will be used.

4. MapReduce

   MapReduce is a program model for data processing techniques based on distributed computing [3].

D. MapReduce

The MapReduce programming model is used to efficiently process large data sets in parallel. MapReduce consists of three stages, namely, the map stage, shuffle, and finally the reduce stage.

1. The Map stage processes input data in the form of files stored in HDFS, these files are then converted into tuples, namely pairs of keys and values.
2. The Reduce stage processes the input data from the results of the map process, which then results in the new data set being stored in HDFS again.

For more details in understanding the stages of the MapReduce process, see Figure 3.

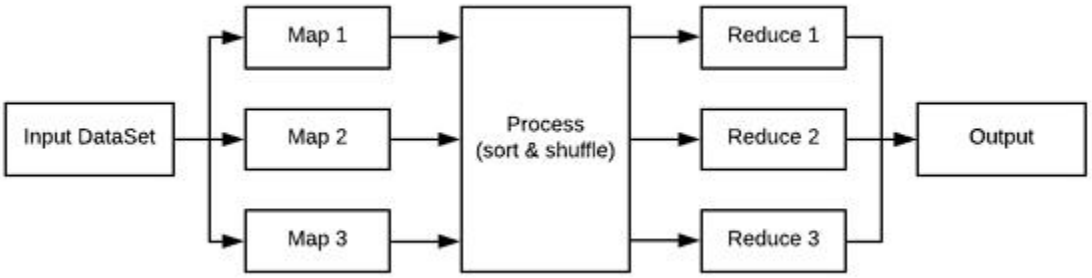

Fig. 1. MapReduce Programming Process

E. YARN (Resource Manager)

For any very large cluster with around 4000 nodes or more, existing MapReduce systems had scalability problems, In 2010, a team at Yahoo! began conceptualizing the next iteration of MapReduce, leading to the development of YARN., which is short for Yet Another Resource Negotiator [11].

YARN overcomes the lack of scalability of classic MapReduce by dividing the job tracker's responsibilities into several parts. Jobtracker is responsible for job scheduling (matching tasks with task trackers) and monitoring the progress of tasks (keeping track of tasks and restarting tasks that are slow or failed, as well as doing bookkeeping such as recording the total tasks so far) [12].

F. Amazon EC2

Amazon EC2, known as Amazon Elastic Compute Cloud, offers a web-based service ensuring secure computing resources, flexibly sized computing capacity in the cloud. Amazon EC2 is design aims to simplify web-scale cloud computing for developers, with Amazon EC2 offering a user-friendly web services interface for acquiring and configuring capacity with ease. Amazon EC2 provides full control of computing resources and allows working in Amazon's proven computing environment [13].

III. SYSTEM DESIGN AND IMPLEMENTATION

In this research, the design of the system to be built is preceded by the installation and configuration process of Hadoop-2.7.1 on the Linux Ubuntu 14.04.6 LTS operating system in a MultiNode Cluster on Amazon EC2. After the installation and configuration process is complete, the input data is in the form of a .text file which is input from local data to HDFS. This data was obtained from poll results via the happypoll.id web application, where voters must log in first via the social media Twitter. Data that has been input to HDFS will be processed by the MapReduce (wordcount) program. In the testing process, several experiments were carried out by comparing the wordcount program with MapReduce on Hadoop and the usual wordcount program (without MapReduce on Hadoop). For more details, data processing using MapReduce on Hadoop can be seen in Figure 2.

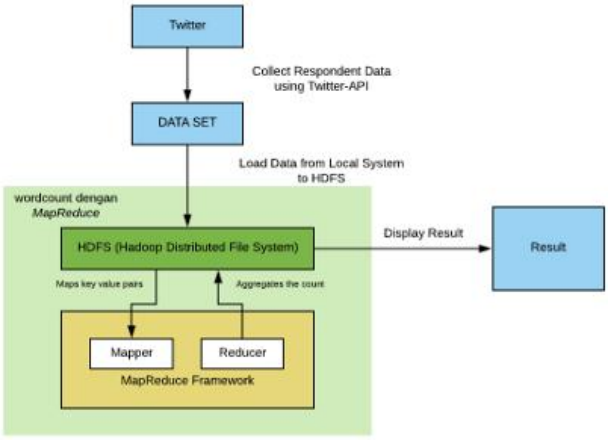

Fig. 2. Process Data with MapReduce

Testing was also carried out with the wordcount program without MapReduce on Hadoop. Namely, with ordinary Java programming the process can be seen in Figure 3.

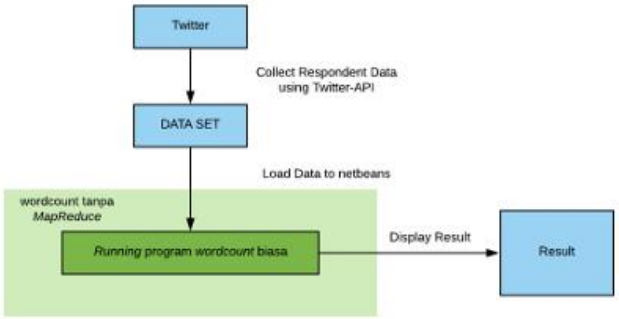

Fig. 3. Process Data without MapReduce

After that, testing was also carried out on Hadoop which processes data with the default block size of 128 MB and with the block size increased to 512 MB. This was done to determine the effect of blocks in HDFS on data processing speed.

A. Data Used for Processing

The data used in this final assignment is data originating from the HappyPoll.id website application, where the data is data from the results of a poll conducted, with the issue of Mr. Basuki Tjahaja Purnama (Ahok) who was appointed as President Commissioner at Pertamina. In this poll, respondents were faced with the choice of agreeing or disagreeing with Mr. Basuki Tjahja Purnama who was elected as President Commissioner at Pertamina. The data from the poll results is exported in the form of a text file (.text) to then be loaded into HDFS and processed using MapReduce programming in the Hadoop framework and a regular wordcount program with Java programming. The author in this final assignment duplicates the data obtained into several sizes, namely data1.text with a size of 55 MB, data2.text with a size of 107.3 MB, data3.text with a size of 214.7 MB, data4.text with a size of 536, 8 MB and data5.text with a size of 1.07 GB. This was done to determine the speed of data processing in the wordcount program with Hadoop MapReduce and wordcount without MapReduce.

B. Data Processing with MapReduce

From the data collection process, it continues with the data being loaded into HDFS for further processing using distributed computing by MapReduce in the Hadoop framework, with a multi node cluster configuration. To run commands from the MapReduce process, there are several steps to take, including:

1. Format the filesystem:

$$\$\ \text{bin/hdfs namenode} - \text{format}$$

2. Run namenode and datanode:

$$\$\ \text{sbin/start} - \text{dfs.sh}$$

3. Create a directory in HDFS for the MapReduce jobs execution process:

$$\$\ \text{bin/hdfs dfs} - \text{mkdir /user}$$

$$\$\ \text{bin/hdfs dfs} - \text{mkdir /user/} < \text{username} >$$

4. Copy files from local system to HDFS:

$$\$\ \text{bin/hdfs dfs} - \text{mkdir input}$$

$$\$\ \text{bin/hdfs dfs} - \text{put etc/hadoop/} * \text{.xml input}$$

5. Run MapReduce:

$$\$\ \text{bin/hadoop jar share/hadoop/mapreduce/hadoop} - \text{mapreduce} - \text{examples} - 3.0.3.\text{jar grep input output 'dfs[az.] + '}$$

As for the MapReduce process itself, there are several process stages, including:

1. The Map stage processes input data in the form of text files stored in HDFS, these files are then converted into tuples, namely pairs of keys and values.
2. The Reduce stage processes the input data from the results of the map process, which then stores the new data set in HDFS again for the next process.

C. Wordcount Application Testing

Two wordcount application tests were carried out, namely the wordcount program with MapRedue and without MapReduce. This was done to find out which one is more effective in processing poll data obtained from conducting polls on the social media-based happypoll.id website.

This test is carried out by looking at the speed of data processing in each program, in each data processed. There are 5 data processed to determine the speed of each program.

D. Software Components Used to Build the System

The software used in implementing this system is as follows:

1. Operating system
   The operating system used on the computer is Linux Ubuntu 14.04.6 LTS
2. Java
   Java is used to support the use of the Hadoop Distributed File System. A Hadoop cluster requires a Java Run Time Environment (JRE) and Java Development Kit (JDK) with Java version jdk-1.8.0.
3. Hadoop
   The Hadoop used is Hadoop-2.7.1 with a multi node cluster setting.
4. Open SSH
   Hadoop runs on top of an SSH server.
5. Web Application
   To get opinion poll data, voters who carry out an opinion poll are required to log in first with Twitter.
6. Google Chrome
   Used to access HDFS and resourcemanager.
7. Amazon EC2
   Used for multi-node Hadoop cluster installations, and for processing large amounts of data that cannot be done by just a writing computer.

IV. EVALUATION

A. *Comparison of Data Processing Speed with MapReduce and Without MapReduce*

In this final assignment, several data processes were also carried out, by testing 5 data of different sizes. Against 2 programs, namely the wordcount application utilizing MapReduce compared to the word count application without MapReduce. This is to find out how fast the two programs are in processing social media-based polling data obtained by the author. The results obtained can be seen from Figure 4

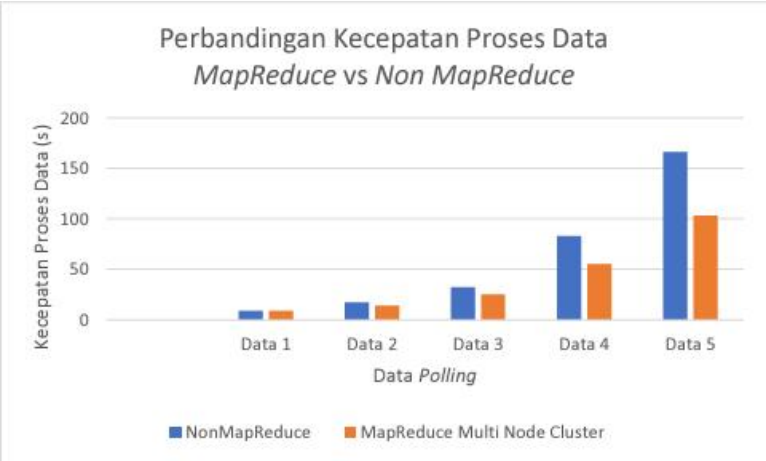

Fig. 4. Comparison of MapReduce vs Non MapReduce Data Processing Speed

From the graph above, for 1 data the wordcount program with MapReduce takes 9 seconds while without MapReduce it also takes 1 second to process the data. For data 2, the wordcount program with MapReduce takes 14 seconds while without MapReduce it takes 17 seconds. For data 3, the wordcount program with MapReduce takes 25 seconds while without MapReduce it takes 32 seconds. For data 4, the wordcount program with MapReduce takes 55 seconds while without MapReduce it takes 83 seconds. And finally for data 5 in the wordcount program with MapReduce it only takes 103 seconds while without MapReduce it takes 166 seconds. For more details, see the following table:

TABLE I
DATA PROCESSING SPEED COMPARISON

Data	NonMapReduce	MapReduce Multi Node Cluster
Data 1 (55 MB)	9 seconds	9 seconds
Data 2 (107.3 MB)	17 seconds	14 seconds
Data 3 (214.7 MB)	32 seconds	25 seconds
Data 4 (536.8 MB)	83 seconds	55 seconds
Data 5 (1.07 GB)	166 seconds	103 seconds

From the 5 data tested, it is known that the Wordount program with MapReduce Hadoop is effectively used to process large data. Even though both programs need the same amount of time for data 1, which is smaller in size, after several tests with larger data, the wordcount program with MapReduce is superior. The larger the data processed, the more superior the wordcount program with MapReduce is by widening the comparison distance of data processing based on time.

B. *Comparison of Data Processing Speed of 128MB Block Size with 512MB Block Size*

After testing Hadoop with an increased block size, it was discovered that the enlarged block was no faster than the default block size (128MB). For more details, see Figure 5.

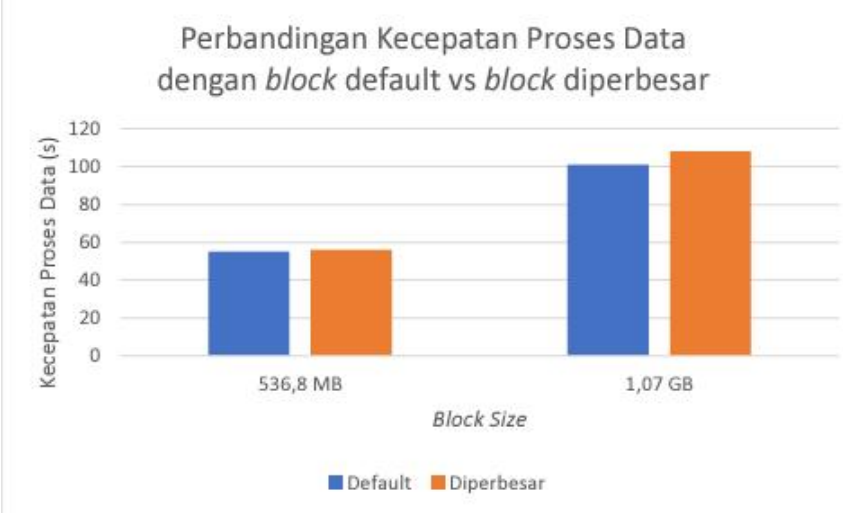

Fig. 5. Comparison of Data Processing Speed with Increased Block Size

From this graph, a block with the default size to process 536.8 MB data takes 55 seconds, while an enlarged block takes 56 seconds. And for data measuring 1.07 GB, a block with the default size takes 101 seconds and for an enlarged block it takes 108 seconds.

C. *Analysis Results from Testing*

From the tests carried out, polling data processing using MapReduce in the Hadoop Framework for the data obtained by the author is optimal for use. This is known from the polling data processing that has been carried

out, from several data that were tested. Data processing with the wordcount program with MapReduce on Hadoop requires a relatively short time for large data. This is because the characteristics of Hadoop itself are intended to process data in a distributed manner, so data processing will be faster using Hadoop.

And from the results of the comparison carried out by comparing the wordcount program with and without MapReduce. The result is that the wordcount program for processing polling data is superior to the program with MapReduce Hadoop. From testing, the larger the data for the method without Hadoop, the performance decreases, with data processing taking quite a long time. Meanwhile, for the wordcount program with MapReduce Hadoop, even though the data processed is greater, the performance in processing the data is stable. This is based on the greater the data obtained, the greater the time gap required between the wordcount program with Hadoop and wordcount without Hadoop. Using Hadoop can process data 1.3 times faster than without Hadoop.

For testing the block size, after testing data measuring 53.6 MB and 1.07 GB, the default block size was slightly superior to the increased block size.

V. CONCLUSIONS AND RECOMMENDATIONS

A. *Conclusion*

Based on the analysis of the test results carried out on the system built in this final project, there are several conclusions, namely:

1. MapReduce in Hadoop is optimally used to process polling data because processing the data requires relatively little time.

2. Wordcount, which is a simple program from MapReduce, is suitable for use in polling data cases.

3. Of the 5 data processed with different sizes, wordcount with MapReduce Hadoop is superior to the regular wordcount program. By averaging 5 test data, MapReduce Hadoop can be 1.3 times faster than without MapReduce Hadoop.

4. The increased HDFS block size affects data processing speed, but the results take longer than the default block size.

B. *Suggestion*

There are several suggestions for development of this final project, namely by utilizing data streaming in Hadoop to obtain larger data. And compare Hadoop with other methods like apache spark, mongdb etc.

REFERENCES

[1] HL Li, VTY Ng, and SCK Shiu, “Predicting short interval tracking polls with online social media,” Proc. 2013 IEEE 17th Int. Conf. Comput. Support. Coop. Work Dec. CSCWD 2013, no. Idc, pp. 587–592, 2013.

[2] PM Bante and K. Rajeswari, "Big Data Analytics Using Hadoop Map Reduce Framework and Data Migration Process," 2017 Int. Conf. Comput. Commun. Auto Control. ICCUBEA 2017, pp. 1–5, 2018.

[3] T. Advancements, G. Jagdev, B. Singh, and M. Mann, “Subcontinent,” Int. J. Sci. Tech. Adv., vol. 1, no. 3, 2015.

[4] C. Verma and R. Pandey, “Big Data representation for grade analysis through Hadoop framework,” Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016, pp. 312–315, 2016.

[5] R.D. Hays, H. Liu, and A. Kapteyn, “Use of Internet panels to conduct surveys,” Behav. Res. Methods, vol. 47, no. 3, pp. 685–690, 2015.

[6] D. Noel, S. Stover, and M. McNutt, “Student perceptions of engagement using mobile-based polling as an audience response system : Implications for leadership studies,” J. Leadersh. Educ., no. Summer, pp. 53–70, 2015.

[7] G. Kapil, A. Agrawal, and RA Khan, “A study of big data characteristics,” Proc. Int. Conf. Commun. Electrons. Syst. ICCES 2016, 2016.

[8] P. R. Merla and Y. Liang, “Data analysis using hadoop MapReduce environment,” Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, pp. 4783–4785, 2017.

[9] K. Rattanaopas and S. Kaewkeeree, “Improving Hadoop MapReduce performance with data compression: A study using wordcount jobs,” ECTI-CON 2017 - 2017 14th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol., pp. 564–567, 2017.

[10] SR Suthar, VK Dabhi, and HB Prajapati, “Machine learning techniques in Hadoop environment: A survey,” 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. 2017-January, pp. 1–8, 2017.

[11] K. Basuki, HN Palit, and LP Dewi, "Implementation of Hadoop: Case Study of Petra Christian University Library Loan Data Processing," 2015.

[12] T. C. Bressoud and Q. Tang, “Results of a model for Hadoop YARN MapReduce tasks,” Proc. - IEEE Int. Conf. Clust. Comput. ICCC, pp. 443–446, 2016.

[13] N. Ekwe-Ekwe and A. Barker, “Location, location, location: Exploring amazon EC2 spot instance pricing across geographical regions,” Proc. - 18th IEEE/ACM Int. Symp. Clust. Cloud Grid Comput. CCGRID 2018, pp. 370–373, 2018.