International Journal of Advanced Engineering and Management Research
Vol. 2 Issue 5, 2017

http://ijaemr.com/ ISSN: 2456-3676

IMPLEMENTATION OF BIG DATA FRAMEWORK IN WEB ACCESS LOG ANALYSIS

Imam Fahrur Rozi[1], Ridwan Rismanto[2], Siti Romlah[2]
State Polytechnic of Malang, Malang, East Java, Indonesia

ABSTRACT

Big data has now grown into one of the innovations in the field of data management. Its ability to handle complex data in large sizes in a short period of time becomes an advantages. Especially at this time the development of data in the world of industry and on the internet quite rapidly. The need for big data has encouraged the birth of Hadoop platform which is currently a big data platform that is widely used. As a preliminary study of the implementation of big data, this research took the case of server log analysis. Characteristics of unstructured and rapidly growing log data make it relevant to be a case study of big data implementation. From this research, the Hadoop-based architecture involves HDFS, Yarn and MapReduce, and involves Zeppelin with Pig interpreters to visualize the results of the analysis. Log data is stored in HDFS, and analyzed by MapReduce-based programs. The analysis results are then translated by Pig and then displayed by Zeppelin. The results of the analysis resulting from this study include web access analysis based on IP address, type of browser used, bot's access, daily access, and also weekly access.

Key Words: log analysis, web server access log, big data, hadoop

1. INTRODUCTION

Every organization, whether government or private, is currently living in a powerful business world driven by data. How an organization defines its strategy and approach to data (including the selection of its supporting technology) will be a critical factor determining future competitiveness [1]. So it has been a long time, the process of management and data analysis become one of the special attention of an organization.

Starting from the 1960s (when the field of computing began to enter the commercial market), data is stored in a flat files. And in the 1970s, there emerged an innovation with a significant impact that is the discovery of Relational Data model and Relational Database Management System (RDBMS). The RDBMS is able to fulfill the organization's need for

effective and integrated transaction data management. As the growth of increasingly large data and business process developments that want to analyze data from various transactional data, then comes the Data Warehouse and Data Mart (started to enter the commercial market in the era of the 1990s). Data Warehouse Technology has actually been able to work effectively on data that is structured with large volumes. But nowadays, the type of data that needs to be processed is not only structured, but also unstructured, with very large volumes. This is what prompted the emergence of a new innovation that is Big Data [2].

The idea of Big Data began to appear in the early 2000s, and began to grow from the 2005s, marked by the emergence of Web 2.0 technology and Hadoop (Big Data plalform currently widely used) [3]. As one of the new innovations in the field of data management, the Big Data is currently very relevant to be examined its application in various cases. Moreover in the next few years is predicted to be the era of data explosion. This is due to the growth rate of data in the coming years will be faster than the growth of data in previous years. Even by 2020 it is estimated that there will be about 1.7 megabytes of new data appearing in every second of every human being in the world [4]. These data can have different shapes and structures, ranging from textual and multimedia data.

In this research will be examined the application of Big Data framework on the process of data analysis of web server access log. Web server access log is a file in which is recorded a list of accessing activities of a web server. This file is generated and maintained automatically by the server. Web server access logs are one example of unstructured data, and on enterprise-scale servers, server log files can be very large. Analysis of web server access logs needs to be done with respect to the recognition of the pattern of recorded activity. This makes the case analysis of web server access log data is very relevant to applied framework big data in its processing. This research is intended as a preliminary study of application of Big Data framework. From this research will get model application of Big Data technology in processing textual data in the form of web server access log. So that can be used as the basis for the implementation of big data on data processing with higher complexity.

Related Works

Analysis of various important server log data is performed, in order to know the information contained in it. The importance of the process of analysis of data logs, has encouraged researchers and practitioners to research about the process of analysis of it. In 2011, Grace, et.al. doing research on the analysis of web logs and web users on web mining. The analysis is intended to determine the pattern of web access made by visitors or users. There are three stages of web mining process that has been done, namely preprocessing, pattern discovery (association rule, sequential patterns, cluster, classification) and pattern analysis (site filter, mWAP, BC-WAPT) [5]. Similar research has also been conducted by Pamutha, et.al. in 2012, focusing on data preprocessing on web server log files. From the preprocessing results, statistical

information about the session user, including: total unique IP, total unique page, total session, session length and frequency of pages visited [6].

In addition to the process of web mining, log analysis is also widely applied to network security, as has been done by Cheon, et.al. who did research on distributed processing on Snort alert logs using Hadoop. Researchers use distributed systems because the number of messages related to network performance is growing fast enough and large. So it will be less effective when analyzed using one computer only. In the study, used Hadoop, HDFS with 8 nodes. From the test results, obtained an increase in speed of 426% when compared with a process performed by a single computer [7]. The growing volume of data and increasing demands to use the data in real time, encourage the existence of a distributed system to process the data. This causes Hadoop as a big data framework increasingly popular to use. Almeer, M.H. apply Hadoop Map Reduce for analysis of remote sensing results. After testing, it was concluded that Hadoop could work efficiently and scalable enough to multiply large images, as in the case of remote sensing image analysis [8]. Hadoop can also be used to build distributed systems for video transcoding in cloud environments, as has been done by Kim, et.al. in 2013. Researchers used a cluster with 28 nodes to transcode video. The system works well, with great speed and good video quality [9].

Big Data Architecture

To understand the high level architecture aspect of Big Data, it must first understand the logical information architecture for structured data. Figure 1 shows two data sources using integration techniques (ETL / Change Data Capture) to transfer data into a DBMS data warehouse or operational data store, and then provide a variety of variations of analytics capability to display data. Some of these analytical skills include dashboards, reports, BI applications, summary and query statistics, semantic interpretations for textual data, and visualization tools for heavy data.

Figure 1. Ability of Traditional Information Architecture [10]

Unlike the traditional information system platform, on the big data architecture there are some things that need special attention such as volume, acceleration, variation, and values that become demands. So in big data technology usually involves real-time processing and batch processing. NoSQL is an example of technology that can be involved to perform data processing in real-time. For batch processing, a technique known as Map Reduce, filtering data based on

data specific to the discovery strategy. Once the filtered data is found, it will be analyzed directly, inserted into another unstructured database, delivered into the mobile device or merged into the traditional data warehouse environment and correlated to the structured data.

Figure 2. Big Data Information Architecture Capabilities [10]

Hadoop Framework

Hadoop is an open-source framework developed by Apache that is intended to perform distributed processing of large datasets, involving multiple computer clusters with a simple programming model. To perform processing of large datasets, Hadoop offers distributed data processing solutions on multiple computers (each computer has its own data storage and processing) instead of upgrading a single server specification to process the data. Services provided by Hadoop include data storage, data processing, data access, data governance, security and operations. The main components in the Haddop such as [11]:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS ™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

2. RESEARCH METHOD

Data

The data to be processed in this research is the access log data from a web server. As a case study, the access log will be obtained from the web server machine in the UPT. Puskom, State Polytechnic of Malang. The server access log data will then be processed by the bigdata platform, so it will generate an access pattern to the server. The example of log server format to be processed is shown in Figure 3.

```
172.16.80.53 - - [06/Feb/2017:17:46:17 +0700] "GET /wp-content/plugins/
jti-alumni/public/js/jti-alumni-public.js?ver=1.0.0 HTTP/1.1" 200 828
"http://jti.polinema.ac.id/" "Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.79
Safari/537.36 Edge/14.14393"
```

Figure 3. Example of Access Log Format

Data processing

In this research, the steps to be implemented are planned as diagrams in Figure 4. Log data is available on some server machines, in the first step, the data is retrieved and saved to the Hadoop cluster with the HDFS file system. HDFS is a file system provided by Hadoop to store data in large volumes.

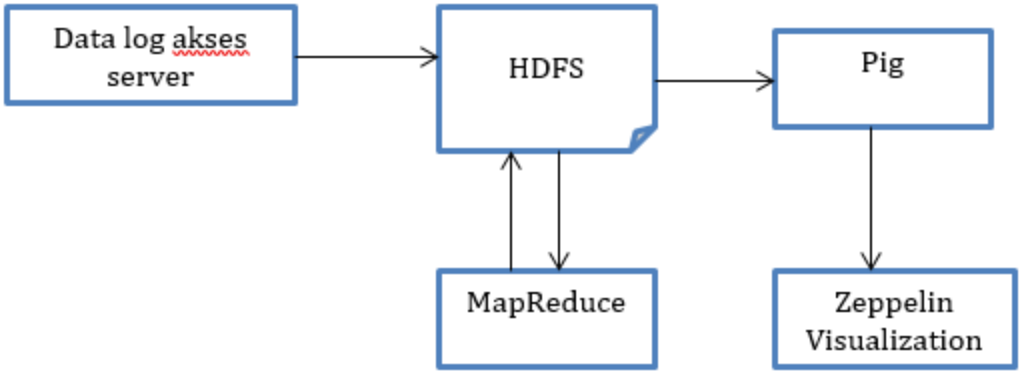

Figure 4. Flow of Data Processing

The next step, log data that has been stored in HDFS will be processed and analyzed to retrieve the desired information. Data processing is made with MapReduce-based, in order to run in parallel in the cluster. The data generated by MapReduce is a key-value pair that indicates a certain information. The result data is then saved again to HDFS. Furthermore, by using Zeppelin the data will be visualized into the display of various forms of charts. Before visualizing the MapReduce results data is loaded by Pig and subsequently formatted into a more structured form according to the desired data format.

The process of analysis performed on the web server access log data is intended to obtain information in the form of:

1. The number of server access based on the IP address of the accessor
2. The number of server access, viewed based on the type of browser used
3. Number of server access performed by bots
4. Daily server access
5. Weekly server access
6. Monthly server access

3. RESULT

In this research, Hadoop is installed and set in single cluster mode. Hadoop version 2.7.3 is installed on a computer with the following hardware specifications:

- Processor: Intel Core i7, 2.7 GHz

- Memory: 4 GB 1333 MHz DDR3

Other software specifications required include:

- Operating System: OS X Yosemite, version 10.10.5
- Java: Java version 1.8.0_60
- Hadoop: Hadoop version 2.7.3
- Pig
- Zeppelin
- IDE Eclipse Version: Mars.1 Release (4.5.1)

After Hadoop is installed, the first thing to do is to format the name node with the command **bin / hdfs namenode -format**. The HDFS service is then run with the **sbin / start-dfs.sh** command and YARN is executed with the **sbin / start-yarn.sh** command. Once the HDFS and YARN service is running, then Hadoop is ready to save the file and manage the process in the cluster.

The next log file to be analyzed is entered into the created HDFS folder with the **bin / hdfs dfs -put etc / hadoop / user / macbook / input** command. In this case the HDFS folder created to save the file is / user / macbook / input.

In the next process made MapReduce program that will process log data that has been stored in HDFS. In this research MapReduce program is written in Java programming language. In the Mapper program will extract the log file into the key-value pair. Suppose that in the process of analyzing the number of access based on the visitor's IP address, the key used is the IP address and its value is the time of visit, such as:

103.9.22.88 2017/03 / 25-10: 23: 20
103.9.22.88 2017/03 / 25-10: 30: 20
103.9.22.88 2017/03 / 25-10: 40: 20
191.252.63.86 2017/03 / 25-11: 20: 20

Then the key-value pair will be reduced by Reducer into the desired key-value pair format. In the case of an access analysis based on its IP key it is an IP address and its value is the number of accesses, such as:

103.9.22.88 3
191.252.63.86 1

The results of the Reducer are stored in HDFS and will then be translated by Pig, and visualized using Zeppelin. The following syntax written in the Zeppelin notebook will read the MapReduce results file showing how many times access is made by an IP address, and display it in the bentu chart.

%pig

```
rawIPAccessCount = load '/user/macbook/input/logbyipcount';

loggedIPAccessCount = foreach rawIPAccessCount generate $0 as ip, $1 as cnt;

loggedIPAccessCount = foreach loggedIPAccessCount generate ip as ip, (int)cnt;

loggedIPAccessCount = filter loggedIPAccessCount by cnt > 1000;


%pig.query

loggedIPAccessCount = filter loggedIPAccessCount by cnt > ${minCount=2000};

groupedIPAccessCount = group loggedIPAccessCount by ip;

foreach groupedIPAccessCount generate group as ip, SUM(loggedIPAccessCount.cnt) as
    access_count;
```

The pie chart view obtained is shown in Figure 5. In the figure is seen percentage comparison of the number of access from each IP address that has accessed the web server.

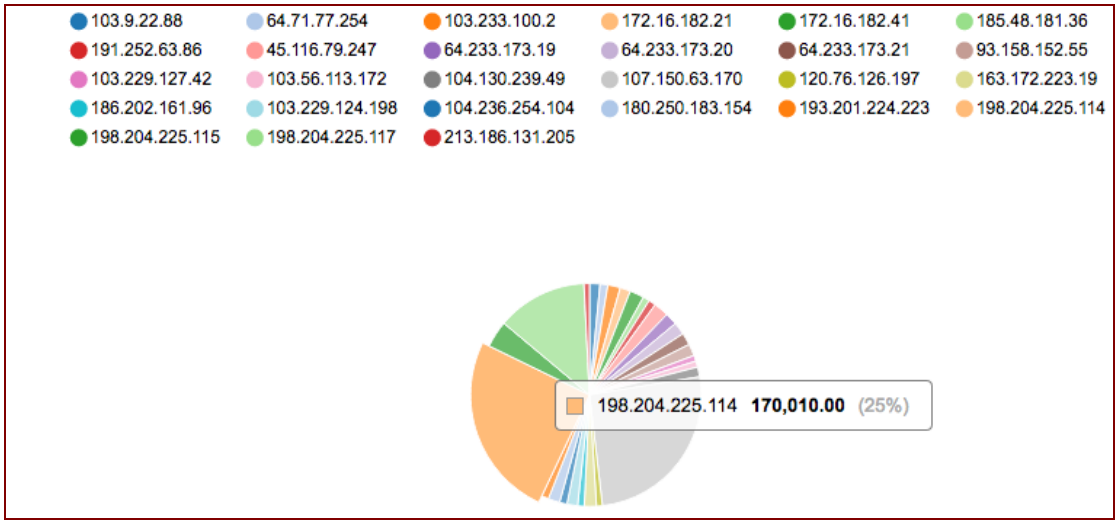

Figure 5. Pie Chart for Data Number of Visits Based on IP Address

In a similar way to the process of analyzing access numbers based on IP address, MapReduce was created to process log data into the desired key-value pairs and created a notebook in Zeppelin with a Pig interpreter to visualize the results of the analysis. Figure 6 shows the results of log analysis based on the number of accesses performed by bots. It appears that Googlebot and Yandexbot are the most active bots to access the system.

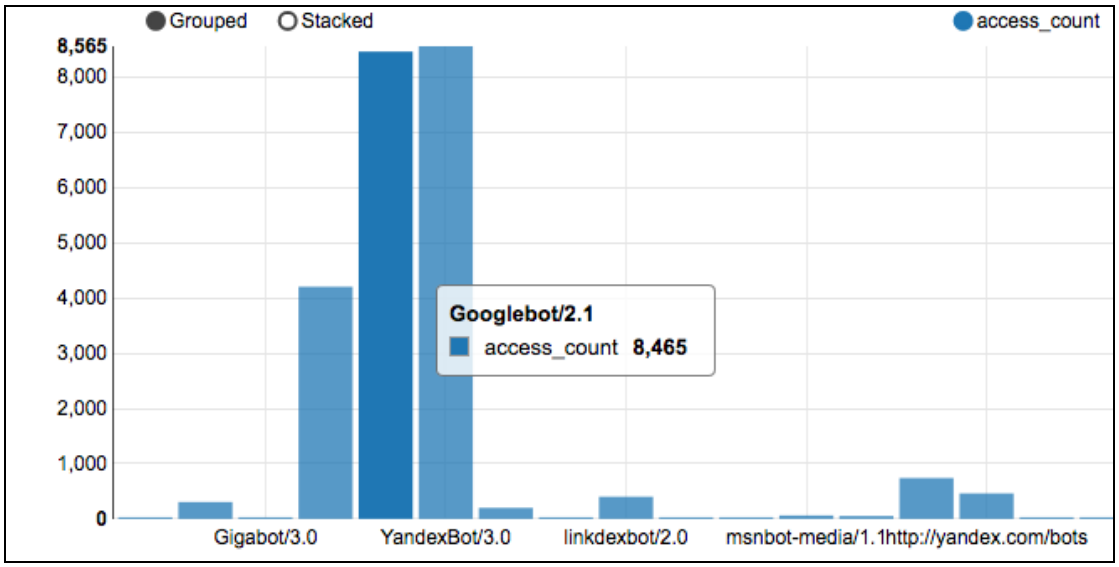

Figure 6. Bar Chart for Data Access Number by Bot

As for access from users, the most widely used web browser is Mozilla Firefox, as shown in Figure 7.

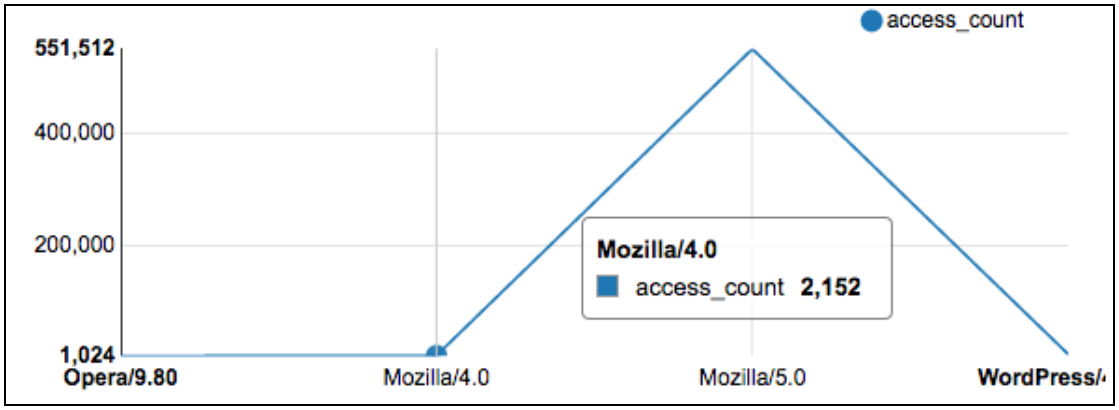

Figure 7. Line Chart for Data Access Based by Browser

From Figures 8, 9 and 10 each show access trends over web servers in daily, weekly and monthly periods. Seen initially the number of requests against the web server is relatively small which then increasingly increases, although fluctuating up and down the number of access.

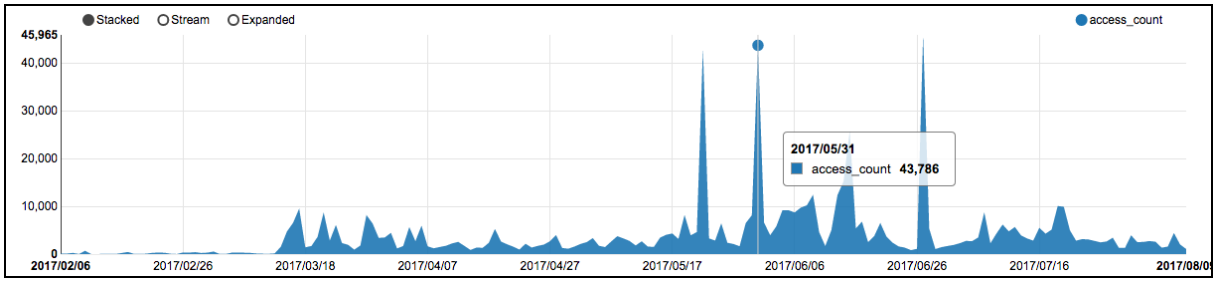

Figure 8. Chart Area for Daily Access Data

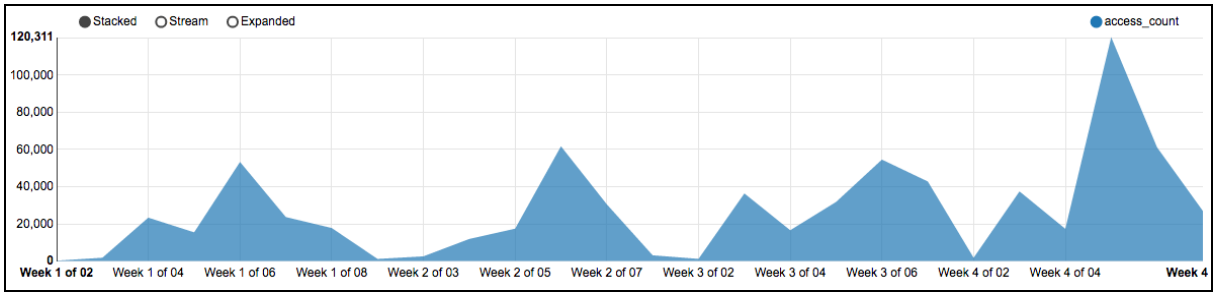

Figure 9. Chart Area for Weekly Number of Data Visits

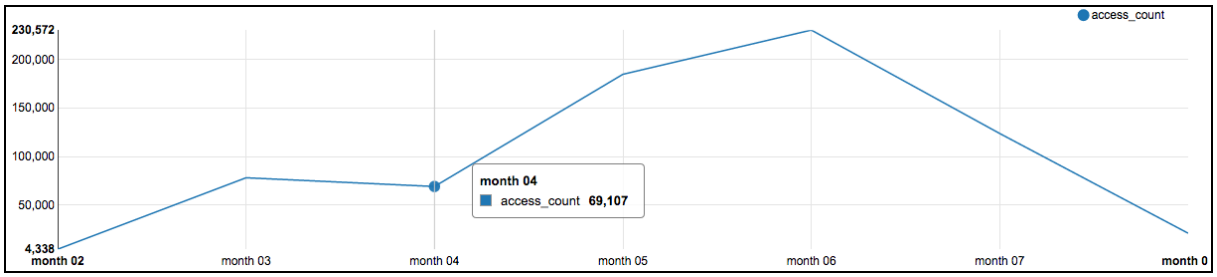

Figure 10. Chart Area for Monthly Number of Data Visits

4. CONCLUSION

From the research that has been done by taking the data object server log and perform the analysis process using big data technology platform, the conclusion that can be taken such as:

1. The architecture used in this research is based on Hadoop technology with one cluster involved. The log data to be processed is input into HDFS and then processed in parallel using Map Reduce. The processing results are also stored in HDFS.
2. Log analysis process used include access analysis based on IP address, web browser type, access made by bot, daily access number, weekly access number, and monthly access number. The analysis process is implemented in the Map Reduce program.
3. The results of the next analysis are translated by Zeppelin by using the Pig interpreter and displayed into the chart view. From observations made during the research, Zeppelin with Pig interpreter used when processing the data was also running on Map Reduce technology.

REFERENCES

1 Oracle Big Data. 2017. *The Foundation for Data Innovation*. Diakses dari https://www.oracle.com/big-data/index.html, tanggal akses 5 Januari 2017

2 Hurwitz, Judith, et.al. 2013. *Big Data for Dummies*. Canada. John Willey and Sons, Inc.

3 Marr, Bernard. 2015. *A Brief History of Big Data Everyone Should Read*. Diakses dari https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr. Tanggal akses 5 Januari 2017

4 Marr, Bernard. 2015. *Big Data: 20 Mind-Boggling Facts Everyone Must Read*. Diakses dari https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read. Tanggal akases 3 Agustus 2017

5 Grace, L.K. Joshila, et.al. 2011. *Analysis of Web Logs dan Web User in Web Mining*. International Journal of Network Security & Its Application (IJNSA) vol.3. January, 2011

6 Pamutha, T. et.al. 2012. *Data Preprocessing on Web Server Log Files for Mining Users Access Patterns*. International Journal of Research and Reviews in Wireless Communication, Vol. 2, No. 2, June 2012.

7 Cheon, Jeongjin, et.al. 2013. *Distributed Processing of Snort Alert Log using Hadoop*. International Journal of Engineering and Technology (IJET), Vol. 5, No. 3, Jun-Jul 2013

8 Almeer, Mohamed.H. 2012. *Cloud Hadoop Map Reduce for Remote Sensing Image Analysis*. Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 4, April 2012.

9 Kim, Myoungjin,et.al. 2013. *Towards Efficient Desing and Implementation of a Hadoop-based Distributed Video Transcoding System in Cloud Computing Environment.* International Journal of Multimedia and Ubiquituous Engineering, Vol. 8, No. 2, March, 2013

10 Sun, H., & Heller, P. (2012). Oracle Information Architecture. *Oracle Information Architecture*.

11 Haddop, Apache. 2017. What is Apache Hadoop?. Diakses dari http://hadoop.apache.org/. Tanggal akses: 20 Juli 2017