The Utilization of Big Data Technology in Handling Fake News (Hoax) Content

Ahmad Budi Setiawan[1], Bambang Mudjiyanto[2]
The Center of Research and Development of Informatics Application, Information and Public Communication, The Agency of Human Resources Research and Developement
Minstry of Communication and Information Technology, Indonesia[1,2]

{ahma003@kominfo.go.id[1], bamb037@kominfo.go.id[2]}

Abstract. Big Data technology is a phenomenon and many organizations try to integrate it into their business processes to get added value and also support business processes. Big Data can be utilized in various ways, one of which is the handling of negatively charged Internet content. Along with the increasing dissemination of negatively charged internet content, especially hoaxes, the Ministry of Communication and Information facilitated various kinds of stakeholders in handling the hoax using the utilization of Big Data technology. Internet content monitoring carried out so far is through manual crawling, controlling negative content named Trust + based on Domain Name Server (DNS System) technology where the blocking mechanism can only be done by using a domain name and / or server name . By utilizing AIS machines that adopt Big Data technology, it can speed up the process of crawling negative content that has been done manually. The machine was not provided with a "killer weapon" site, social media accounts, news portals, and others. This study was conducted through a qualitative approach to describe the use of Big Data technology on AIS machines in monitoring negative internet content. The results of this study outline the use of Big Data technology in terms of handling negative internet content, especially dissemination of hoaxes.

Keywords: Hoax; Big Data; utilization; negative internet content; fake news

1 Introduction

A scientific article entitled "The Spread of True and False News Online" published in Science magazine in March 2018 revealed a surprising finding: "Fictitious news spreads much faster than factual news." In this study, Vosoughi, Roy, and Aral as researchers analyzed a set of 126,000 rumors (hoax) spread by around 3 million Twitter account owners from 2006 to 2017 [1]. In addition to finding that fictitious news spreads much faster than factual news, studies it also found that fictitious news was able to reach more audiences than factual news. They succeeded in observing that from about 1 (one) percent of the most favorite fictitious news was able to reach 1,000 to 100,000 people, while for factual news almost no one was able to reach more than 1,000 people. Hoax distribution generally has three patterns. First, the information disseminated utilizes the confusion of public opinion so that it can easily get the

ICA 2019, October 16-17, Bali, Indonesia

DOI 10.4108/eai.16-10-2019.2304345

attention of the public; Second, hoaks generally use references to people who are known to the public even though the information is often twisted, cut, or even fabricated; Third, the spread of hoaks moves in syndication by disseminating information through various social media.

The rapid development of technology supports the emergence of new data whose development is also rapid in number. The data is supported by the existence of personal data that can be generated by individual data sharing activities (volunteered data), personal activities (obsevered data), as well as data from the relationship between the two (inferred data). The effects of the growing data are felt by social media companies like Facebook, LinkeIn, Google, Microsoft, or Yahoo search engines. The data is irregular, varied, structured or unstructured and has different formats. These data sets are commonly known as Big Data [2], [3].

Indonesian Internet Network Providers Association (APJII) believes that internet penetration to users is the biggest contributor to hoaxes [4]. Internet users in Indonesia are increasing every year. Until 2017 the number of internet users in Indonesia was 143.27 million (more than 50% of Indonesia's population). The ease of internet access coupled with the presence of social media has increased the spread of hoaxes in society. The 2017 Mastel survey identified that the highest media for spreading hoax was social media (92.4%). The behavior of social media users is not supported by adequate digital literacy capabilities (post-truth era). This causes the spread of hoax to increase.

On the other hand, along with the massive presence of hoax in the community, Big Data technology is currently present along with the development of web 2.0 technology. Through Big Data technology, a large amount of data and large volumes of data are increasingly being explored. This is not limited to a collection of data that is processed, processed and used in processing a business process. Starting from the massive data, the information generated is able to make a radical contribution to the mindset and way of looking at a problem [5].

Regarding information, Big Data is able to present more complete data, not only in one or two-way relationships. For example on Facebook, the relationships that occur between users have a growing possibility, ranging from friendship, work, hobbies or school friends. The relationship is ultimately needed in solving certain problems [6]. One of the characteristics of Big Data is that it contains unexpected, hidden, and mass information that is not necessarily obtained from an ordinary data mining process. That information is the reason Big Data becomes a very important thing for an organization or enterprise.

Enterprise goals can be achieved if the IT function can collaborate with management. Therefore we need a governance and management of IT functions so that they are able to make the most of IT resources [7][8]. Big Data is one of the new phenomena in the IT world and is able to provide benefits to organizations. For the application of Big Data to have benefits in the organization, all involved in the Big Data integration process must be carried out appropriately and thoroughly and may not be partial [9]. The analogy of this theory is illustrated as blind people and elephants, where each part has a non-holistic interpretation. In other words, each part of the organization only understands its part so that it has the opportunity to produce biased information. Big Data in this case is seen as an integrated system of an organization in order to find its benefits.

Managing a variety of data and with very large amounts, will require an effective way to process it, especially if the information generated from the data is needed to support decisions for policy makers, it takes time to be able to process the data into information. In addition to online and print media, social networking media is also a concern to be monitored. Because the use of social media is very large makes it an area that must be considered, especially to see public opinion. Therefore, the Big Data principle is very suitable to be applied, where the Big

Data principle is to be able to manage a lot of data and various, and process it into the desired information in a very fast time.

Duties and Functions of Media Monitoring become a very important part of the Ministry of Communication and Information. This is in accordance with Presidential Instruction No. 9 of 2015 concerning Management of Public Communications, where the Ministry of Communication and Information has the task of conducting a study of information that is developing as well as monitoring and analyzing media content. Based on the problems that have been raised, this study discusses the implementation of Big Data in terms of handling content or negative news.

Big Data is defined as tools, processes and procedures that enable an organization to create, manipulate and manage very large, varied, and high-speed datasets. Big Data is an organizational asset that is information, whose use and processing is useful for improving knowledge and decision making processes [10][11]. Gartner describes the dimension of Big Data as 3V, namely Volume, Velocity, Variety [12]. Along with its development, Big Data not only covers 3V but expands to 5V namely Volume, Velocity, Value, Veracity, and Variety. Technically, Big Data is a large set of data both structured, semi, and unstructured so that it cannot be processed using ordinary relational database tools [13][14][15][16][17][18]. Based on this understanding, Big Data is a terminology for massive data with a very large size, fast, varied, with a complex structure and able to provide information that is very important and crucial. Big Data is a prediction consisting of artificial intelligence, machine learning, and other characteristics that support the prediction process [19][20]. The data that appears has the opportunity to be able to provide a policy guide without being realized before [21][22].

In addition to offering a big leap in organizational decision management and decision making, Big Data integration has several challenges and implementation risks. Big Data has a capacity beyond ordinary database processing capabilities due to the characteristics of Big Data that includes data volume, velocity, diversity / variety type and data structure, variability / data variability, and value [23]. Therefore, implementing Big Data requires a framework. To get a suitable framework, first determine the characteristics of Big Data to be applied as follows: data volume and data processing and capabilities; variety of data / data sources; the speed and timeliness required; targeted services, products, solutions and applications; presentation of data, uses and interpretations; and privacy, error handling and security.

Big Data emerges from all things that are related in this world with a very large size, complex, with a high speed. Big Data is a technological trend for new approaches to understanding the world and making business decisions [24]. These decisions are made based on data in very large volumes of structured, unstructured and complex (for example, tweets, videos, commercial transactions). According to Bill Schmarzo, the process of integrating Big Data in an enterprise has a business maturity index consisting of several phases as follows [11]: business monitoring, business insights, business optimization, data monetization, and business metamorphosis.

Big Data is one of the new phenomena in the IT world and is able to provide benefits to organizations. For the application of Big Data to have benefits in the organization, all involved in the Big Data integration process must be carried out appropriately and thoroughly and may not be partial [25]. The analogy of this theory is illustrated as blind people and elephants, where each part has a non-holistic interpretation. In other words, each part of the organization only understands its part so that it has the opportunity to produce biased information.

Big Data in this case is seen as an integrated system of an organization in order to find its benefits. A data is not categorized as "Big Data" just because of the large amount of data, but there are several characteristics that distinguish Big Data from other systems. Some

characteristics of Big Data, namely, Big Data Systems have a VOLUME of very large data, which usually exceeds ordinary servers in general and this data will continue to grow every day. The amount of data can reach more than 100 TB and is usually stored in external infrastructure (not maintained alone). In addition, not only is the amount of data that many, Big Data also has a variety of data (VARIETY), with the format and type of data that is very diverse so that it requires a special process to be able to process it. In addition to the large and varied amount of data, Big Data must also be able to process the data in a very fast time (VELOCITY) so that the data can be useful not only because of the information generated but also because of the speed needed to process it. The fourth characteristic of Big Data is the truth of the data itself (VERACITY). Information that is processed from these data in order to become useful and trustworthy information, we also have to look at the source of the data used. Therefore, the Big Data, the truth of the data is one thing that must be considered as well.

2 Method

This study uses a qualitative approach to try to explore the existing needs. To find alternative implementation options, this study uses the Modified Waterfall method. Although this method was originally used in the field of software engineering, this method is suitable for exploring various alternative options for implementing Big Data. But in accordance with the limitations of the study, in this study not to the creation or implementation of the system. Therefore, if it refers to the Modified Waterfall method, for this study only at the System Design stage. This research will go through four stages: (1) Formulation of System Requirements: At the beginning of the research will formulate the requirements needed by the Big Data System to monitor various public media; (2) Alternative System Design Options: After formulating the needs of the Big Data System, the next step is designing appropriate alternative choices to be applied to the Big Data System; (3) Data Collection: Stages Data collection is carried out through the FGD process to obtain Expert Judgment from relevant experts; (4) Strengthening alternative choices: The results of the FGD and survey will be used to strengthen alternative implementation options that have been previously designed; (5) Formulation of Recommendations: After that, a recommendation will be made to implement the Big Data system based on alternative choices that have been generated from this research.

3 Result And Discussion

3.1 Formulation of System Requirements

Initially in 1970-2000 the data that was built was a structured data model and was a relational database such as MySQL, Oracle, and others. Then in 1995 the next business intelligence was built to use a structured and relational database with systems such as cognos, pentaho and others. In 2010 until now a system that has 3V goals (volume, velocity, varity) or 4V (added value) was built, and with various technologies such as map reduce, high performance computers cluster and others. This means that Big Data is part of business intelligence, Big Data can be used to form a business that has intelligence to support decision making. But in this case there are several things that are different in terms of volume that is not only a large amount of data, but the growth of data is very fast so that in a short span of

time the data can grow very quickly and large (velocity), and existing data have variations very much (variety), of course, in the big data itself, especially in the formation of data warehouses there has been a lot of transform load extraction done to handle varieties of the data so that the data can be a good standard cleared of various noise also transformed so that the data is much more in line with existing business processes or those currently running for certain organizations. Business intelligence in which there is a use of big data also requires a technology that can support business processes that are in business intelligence itself, so that it can run as expected. So it needs to be built an appropriate infrastructure and can overcome the needs of big data, one of which is the data processing is very fast even though at the same time the data is large and growing quickly.

Problems and challenges in this process are data acquisition, data recording, extraction, cleaning, annotation, integration, aggregation, representation, analysis, modeling, interpretation, and visualization. Big data itself has applications and benefits for various fields as mentioned above before. There are two technologies in infrastructure in Big Data, namely: (1) High Performance Computing Cluster (HPCC) or can be referred to as Data Analytics Supercomputer (DAS); (2) Hadoop Platform (Map Reduced-Based Platform).

High Performance Compputing Clusters themselves basically build a super computer that consists of more than one computer with certain specifications (usually the same) to help support each other, or divide tasks with each other so that together they can process data, especially in terms of data search. Large processes that usually run by themselves are like, Extract, Transform, and Load, then after that an analysis is carried out to obtain information that is more in line with the organization's business needs.

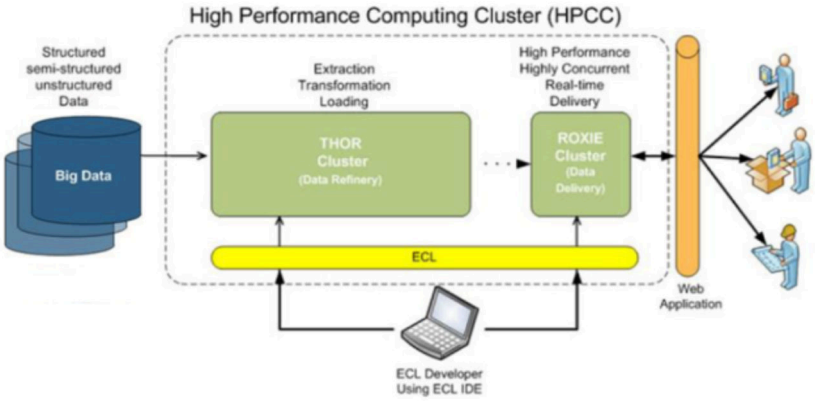

Fig 1. High Performance Computing Clustering

Meanwhile, the Hadoop Platform itself is a technology project developed by Apache in managing large data so that it is far more effective and efficient. In Hadoop itself consists of various components, even to Hadoop itself has its own distributed file system called (HDFS). The advantages of HDFS itself are: (1) Fault tolerance, and is deployed for low cost hardware; (2) Write Onece, Read many, is a simple coherence, and moreover the framework that is built in Hadoop when we will use Hadoop, using Java technology; (3) Moving computing / processes faster than moving data; (4) Similar to Google File System, but HDFS divides files into blocks in distributed cluster nodes; (5) Core components: master vs slave, name node vs. data node, job tracker vs. task tracker.

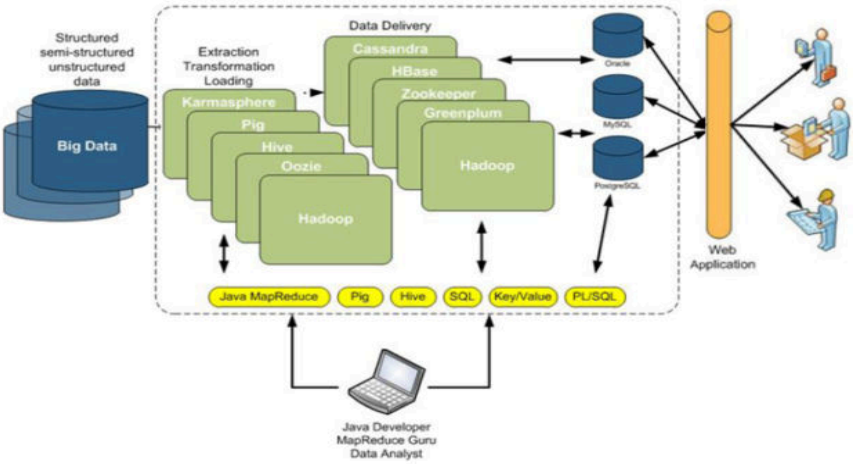

Fig 2. Hadoop Platform

And here is one integration architecture between HPCC and Hadoop platforms:

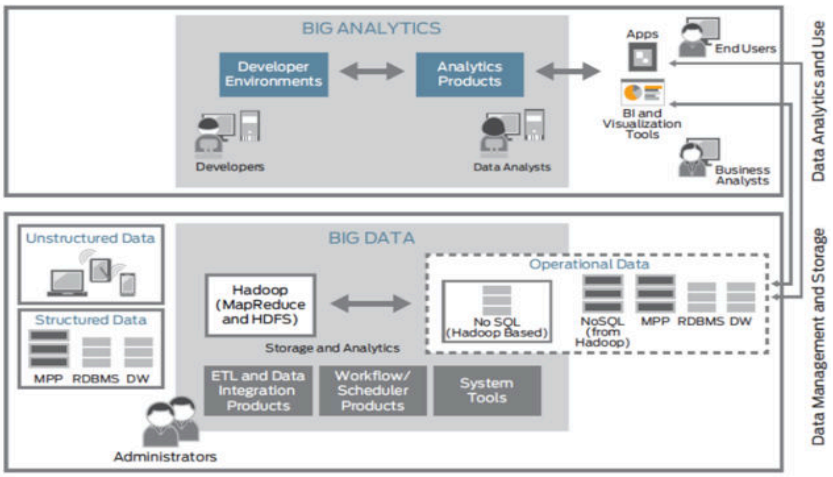

Fig 3. Integrated Architecture of HPCC and Hadoop

At this time there are differences in database management systems, which basically have a correlation database between data that we usually sebt with data that is structured or organized, and database management system tools which are software that can be used to manage databases such as MySQL, Oracle, PostgreSQL and others. At the moment there is another need for database management in the presence of big data or becoming a Big Data Management System.

3.2 System Planing

Before you can determine the Big Data model as to what would be suitable for use in handling negative content, you must first be able to define what the needs of the Big Data system are handling negative content. The formulation of this need will be set forth in a

Business Process so that it will be clear what the process of activities of the Big Data system is handling negative content, especially in Media Monitoring. Of the two tasks, the government simply wants to see if there is a "gap" between issues circulating in the media and the public when viewed with government policies and programs, so that if there is, the government can anticipate it to fill the "gap", for example by providing information relating to the issue, reviewing the program of activities carried out, etc. Based on the data that has been collected through FGDs that have been done, broadly Media Monitoring conducted using the Big Data System in terms of media monitoring, produces two routine results, namely Public Issue Monitoring (MIP) and Monitoring Content Analysis (MCA).

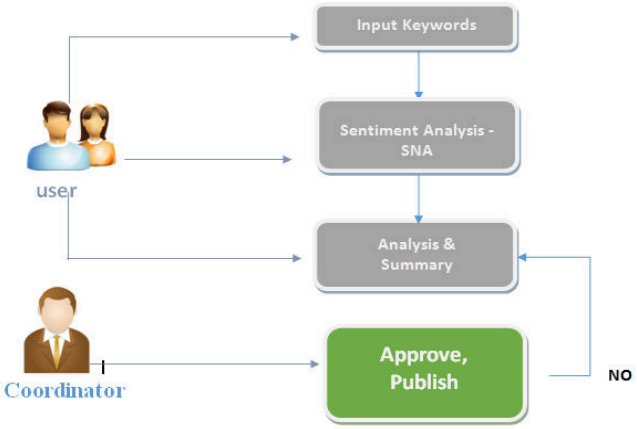

Fig 4. Business Process of Media Monotoring

Public Issue Monitoring (MIP) is a media monitoring activity based on news headlines. The headline that was seen was an issue of the news which was being used a lot in other news. In making MIP reports, using Sentiment Analysis can accelerate the process of news analysis coupled with Social Network Analysis (SNA) enriching the results of the analysis by adding information such as the linkages of one issue with other media. Meanwhile, Monitoring Content Analysis (MIP) is a media monitoring activity by viewing the contents of the news. The news content that will be viewed are the news with the most headlines. In making MIP reports, using Sentiment Analysis can accelerate the process of news analysis coupled with Social Network Analysis (SNA) enriching the results of the analysis by adding information such as the linkages of one issue with other media. The difference with MIP which only sees news headlines, at this stage the analysis is more generated from the content of the news.

4 Conclusion

Sentiment analysis methods can be combined with other methods such as Social Network Analysis to enrich the analysis results of media monitoring. So that using this method can cover the shortcomings that exist in the analysis sentiment and strengthen it. Sentiment analysis methods should be combined with Social Network Analysis so as not to stop with media monitoring, but also to find key persons or actors who are related to the issue. This is very important in order to maximize the function of Big Data to be implemented. Because it would be very unfortunate if the large resources that have been allocated to implement Big

Data are only used for monitoring, but it would be better if you could do predictive analysis to estimate how big an issue could be. So that in the future Kominfo together with other government agencies can react more quickly to issues predicted through the Big Data system, which has the potential to become big, before the issue becomes an uncontrollable problem. In simple terms the processes that occur in the Big Data System that will be implemented can be divided into three categories, namely Input, Process, and Output. The recommended combination of methods is at the process stage. So, it is hoped that with this recommendation, in addition to being able to maximize the function of Big Data to be implemented, it will not greatly change the function of Media Monitoring.

For Big Data implementation it is better to use the option to build a partial system, so that important system parts such as Corpus can be owned by themselves and the system that is built can be designed and monitored as desired. This option can reduce the government's dependence on other parties and also reduce the expenditure of large funds to pay vendors each year.

References

[1] S. Vosoughi, D. Roy, S. Aral, "The Spread Of True And False News Online", MIT Initiative On The Digital Economy Research Brief, 2017.

[2] K. Krishnan. *Data Warehousing in the Age of Big Data*. USA: MK Publications, 2013.

[3] S. Sagiroglu, D. Sinanc. Big Data: A Review. *IEEE International Congress on Big Data* : Gazi University, Department of Computer Engineering, Faculty of Engineering, Ankara, Turkey, 42-47, 2013.

[4] Internet Service Providers Association, Penetration & Behavior Profile of Indonesian Internet Users, Survey Reports, 2019.

[5] V. Mayer-Schonberger, K. Cukier. *Big Data: A Revolution that Will Transform How we Live, Work, And Think*. New York: Houngthon Mifflin, Harcourt Publishing, 2013.

[6] S.K. Associates. Introduction to Big Data : Challenges, Opportunities and Realities. *IEEE 47th Hawaii International Conference on System Science*, American University, USA, 728, 2014.

[7] ISACA. COBIT 5 - *A Business Framework for the Governance and Management of Enterprise IT*. USA: ISACA, 2012.

[8] ISACA. COBIT 5 *Enabling Information*. USA: ISACA, 2012.

[9] J. Zhang, M.L. Huang. 5Ws Model for Big Data Analysis and Visualization. *IEEE 16th International Conference on Computational Science and Engineering, School of Computer Software*, University of Technology, Sydney Australia, Tiajin University Tiajin China, 129-134, 2013.

[10] M.A. Milton. *A Brain-Friendly Guide : Head First Data Analysis*. USA: O'Reilly, 2019.

[11] S. Bill. *Understanding How Data Powers Big Business*. USA: John Wiley & Sons, Inc, 2013.

[12] Gartner. *Big Data*. Gartner Group Press Release, 2009.

[13] F. Tekiner, J.A. Keane. Big Data Framework. *IEEE International Conference on Systems, Man, and Cybernetics*, School of Computer Science, The University of Manchester, Manchester, UK, , 1494-1499, 2013.

[14] K. Krishnan. *Data Warehousing in the Age of Big Data*. USA: MK Publications, 2013.

[15] N. Sawant, H. Shah. *Big Data Application Architecture Q & A*. New York: Springer Science+Business Media, 2013.
[16] P.C. Zikopoulos. *The Power of Big Data: The IBM Big Data Platform*. USA: Mc Graw Hill, 2013.
[17] S. Sagiroglu, D. Sinanc. Big Data: A Review. *IEEE International Congress on Big Data* : Gazi University, Department of Computer Engineering, Faculty of Engineering, Ankara, Turkey, 42-47, 2013.
[18] R. Schell. Security – A Big Question for Big Data. *IEEE International Conference on Big Data* University of Southern California, USA, 2013.
[19] V. Prajapati. *Big Data Analytics with R and Hadoop*. Birmingham, UK: Packt Publishing, 2013.
[20] Z. Liu, P. Yang, L. Zhang. A Sketch of Big Data Technologies. *Seventh International Conference on Internet Computing for Engineering and Science*, School of Information Science and Technology, Shanghai Sanda University Shanghai, China, 26-29, 2013.
[21] M. Milton. *A Brain-Friendly Guide : Head First Data Analysis*. USA: O'Reilly, 2019.
[22] V. Mayer-Schonberger, K. Cukier. *Big Data: A Revolution that Will Transform How we Live, Work, And Think*. New York: Houngthon Mifflin, Harcourt Publishing, 2013.
[23] P.C. Zikopoulos. *The Power of Big Data: The IBM Big Data Platform*. USA: Mc Graw Hill, 2013.
[24] F.O. John. *Big Data Analytics : Turning Big Data Into Big Money*. USA: Wiley & Sons, 2013.
[25] J. Zhang, M.L. Huang. 5Ws Model for Big Data Analysis and Visualization. *IEEE 16th International Conference on Computational Science and Engineering, School of Computer Software*, University of Technology, Sydney Australia, Tiajin University Tiajin China, 129-134, 2013.
[26] U. Papen. "Hymns, Prayers, and Bible Stories: The Role of Religious Literacy Practices in Children's Literacy Learning". *Ethnography and Education*, Vol. 13, No. 1, pp. 119-134, 2018.
[27] R. Ismail, R. Ibrahim. "Teachers' Perception on Digital Game: A Preliminary Investigation towards Educational Game Application for Islamic Religious Primary Schools", *Proceeding 2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, pp. 36-41, 2018.