Identification of Big Data Opportunities and Challenges in Statistics Indonesia

Siti Mariyah
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
sitimariyah@stis.ac.id

***Abstract*— Statistics Indonesia is in charge of doing government duties in the field of statistics. Censuses, surveys and compilation of statistical data have been being done. In some cases, survey environment is getting worse and statistical activities work less efficient. In other hand, a phenomenon called big data arises and many researches learn how to extract big data value. In this paper, researcher identified that big data can be combined in statistical methodology as a part of data source. Researcher identified in three surveys consists of Consumer Price Survey, Business Tendency Survey and Data Satisfaction Survey. In these surveys, researcher found that big data could give opportunities in generating statistical data. Making online submission system, information extraction system, sentiment analysis, and establishing legal cooperation with private sector are some ways to get opportunities delivered by big data. Nevertheless, big data also offer many challenges when combined in statistical methodology. Researcher identified big data challenges come from each stage of statistical methodology. Challenges also come from legislation, security and privacy, storage, processing and data access, skill requirement, and financial.**

Keywords—statistical data; big data; Statistics Indonesia

I. INTRODUCTION

Statistics Indonesia is a non-ministerial government institution who is carrying out government duties in the field of statistics. Statistics Indonesia has some roles, as follows: provides data to government and public; assists statistics division of government departments and other institutions in developing statistical system; develops and promotes standards to be incorporated in the implementation of statistical techniques and methods; and establishes corporation with international institutions and other countries for the benefit of Indonesia's statistical development [16].

For carrying out all roles, Statistics Indonesia does statistical activities, which divided into two big groups; censuses & surveys and compilation of statistical data. In 2012, Statistics Indonesia has conducted 182 censuses and surveys, 67 compilations of statistical data, 605 sectoral statistics and 359 specific statistical activities [17]. In 2013, Statistics Indonesia has conducted 83 censuses and surveys, 24 compilations of statistical data, 112 sectoral statistics and 24 specific statistical activities [18]. Censuses, surveys, compilations of statistical data, sectoral statistics, and specific statistical activities are routinely conducted each year. All of these statistical activities generate large amounts of data. All data is collected from enumeration area and computerized stored in each provincial statistical office. Then all data is computerized sent to central office for processing and analyzing. Statistics Indonesia has been using computers for statistical data processing since 1963. Many kinds of information delivered to public, which grouped into information of social and demographic, information of economic and trade, and information of agriculture and mining.

All statistical data is manually collected, enumerators come to respondents, conduct interview, record the respondents' answers in the questionnaire, entry those in computer application and stored in computer application or database. During conducting statistical activities, Statistics Indonesia faces some issues, including technically or non-technically obstacles. Some issues faced by Statistics Indonesia are survey environment is getting worse and environmental condition. These issues make some statistical activities work less efficient and trigger researcher to identify new alternative in collecting statistical data.

In information and communication technology (ICT) era, there is a phenomenon famous the term big data. A phenomenon caused of the use of information and communication technology such as internet, web and mobile applications, cloud computing, etc. Internet technology has change life style of most people. Many aspects of life such as business, education, scientific research, public administration, etc. have been connected to internet. Recent conservative studies concluded that 9.57 zetabytes of information were processed by servers in companies worldwide in 2008 [1]. Over 6 billion hours of video are watched each month on YouTube and 100 hours of video uploaded to YouTube every minute [2]. In every second, 9100 tweets happen [3]. In 2008 Facebook had 100 million users and as of March 2013 has 1.11 Billion users [4]. The amount of data and information is estimated to increase double every two years [5]. These

conditions has produced huge amount of data and information that now is familiar with term “big data”.

Big data and issues faced by Statistics Indonesia have triggered researcher to identify opportunities and challenges of big data in Statistics Indonesia. According to other researches in the field of big data, big data gives values to organization if organization makes sense of data. Each census and survey need much time and workers, nevertheless number of human resource and time is limited. This research aims to identify opportunities and challenges that will be delivered by big data when we try to extract values from big data to enhance the way or methodology in collecting and generating statistical data.

II. Related Work

Some researchers have become related work to this paper. In [6], Katal, et al. described the issues, challenges, tools and best practices of big data. In [12], authors Joseph and Johnson explained how can big data and associated analytics help e-government evolve into transformational government. In [7], Zheng, et al. described an overview of service-generated big data and big data-as-a-service. Authors described three types of service-generated big data are exploited to enhance system performance. Then, Big Data-as-a-Service, including Big Data Infrastructure-as-a-Service, Big Data Platform-as-a-Service, and Big Data Analytics Software-as-a-Service, is employed to provide common big data related services. From the specific aspects, in [8] authors explained challenges and opportunities of big data analytics in healthcare. In [13], Moon and Cho proposed big data and policy design for data sovereignty.

III. Research Methodology

This research has been done by doing literature review and interview. The researcher made a literature review on research papers in the field of big data to know the state-of-the-art of big data. Later, researcher identified three surveys that have the possibility to enhance the statistical techniques and methods by combining big data as a part of way to generate statistical data. Objects research in this paper are three surveys consists of Consumer Price Survey, Business Tendency Survey, and Data Satisfaction Survey. Researcher learned source of data, statistical techniques used, and identified source of data from big data. Researcher also interviewed subject matter of survey who knows the detail of survey and chief information officer who know information technology used in Statistics Indonesia.

IV. Statistical Data VS Big Data

Big data are collection of very huge data sets generated from whole electronic signal during computing process with a great diversity of types that difficult to be handled using the existing traditional system. Big data has some characteristics that can explain big data well. Big data has 3Vs characteristics; volume, variety, and velocity. Volume means that the amount of data is too big. The volume of data are in petabytes, zetabytes, even more. Variety means that data are on not only structured format, but also semi structured and unstructured format. Social medias have contributed on producing semi and unstructured format. Velocity deals with the speed of data. Data are coming from various sources, speed of incoming data and speed at data flows depicts velocity [6]. Some researchs also identified other characteristics of big data; variability, complexity, value. There are much values produced by big data. Big data can be a service that consists of big data analytics software as a service, big data platform as a service, and big data infrastructure as a service [7]. Because of these characteristics, big data creates opportunities (roles) and challenges when we try to extract values from big data.

Statistical data are different with big data. Statistical data produced by Statistics Indonesia have been conducted under law and government regulation. Statistical data need technique and methodology that have been well designed before collecting data. Statistical data also have different sources of data depend on object research. Sources of data come from individuals, households, industries, and even from government institutions. It needs sampling technique to determine the right respondents.

V. Result of Research

Statistics Korea has developed a pilot project for using big data directly in statistical business process. According to the standard business process, when producing The Industrial Production Index, every month enumerators visit a sample of establishments. Data on industrial classification, items, sales, etc. are edited, and then the Index is published. In the pilot project, the editing process was redesigned to use media data and the big data processing model was inserted. Statistics Netherlands have studied several big data case. Data sources studied as potential input for statistics were a) traffic loop detection, b) mobile phone data, c) social media messages. Researcher has identified the possibility of combining big data in statistical methods when producing statistical data by Statistics Indonesia. Researcher classified into three cases, as follows:

First case- Consumer Price Survey aims to get consumer price data completely. Observation units of this survey are retailers from traditional and modern market, households, and institutions. Main collected variables are prices from 774 goods and services. Data are collected monthly, weekly, and daily. Researcher has identified and found that big data can be combined as a part of data source. Big data can give contribution in providing consumer prices from modern market. Indonesia has various types of modern market consists of 85 Carrefour outlets, 100 Hypermart outlets, and 199 Superindo outlets located in most provinces in Indonesia. Each these modern market always publish around 20.000-40.000 goods price in each official website and routinely updated. For getting and extracting all data, researcher proposes two

alternatives as follows: first, building information extraction application to extract relevant data from each official website. Information extraction application should can extract not only text but also image or graphic; Second, establishing a legal cooperation with modern market for getting goods price. Statistics Indonesia can build online reporting system which supports the direct submission of raw data (goods price) collected from modern market to Statistics Indonesia.

Second case- Business Tendency Survey (BTS) aims to calculate and analyze index of business tendency that quarterly describe early information about business condition in Indonesia. Respondents of surveys are 2500 companies consist of medium and large company. Each company fulfill questionnaire then enumerator will collect questionnaires from respondents and record the answers to computer application. Researcher has identified and found that big data can be combined as a part of data source.

Table 1. Sources of documents that contain data related to BTS.

No	Sources	Number of documents
1	Website of Indonesia Stock Exchange (IDX) (www.idx.co.id) or http://www.idx.co.id/id-id/beranda/perusahaantercatat/profilperusahaantercatat.aspx	Provide documents of company profile from 495 companies.
2	Website of Indonesia Stock Exchange (IDX) (www.idx.co.id) or http://www.idx.co.id/id-id/beranda/perusahaantercatat/laporankeuangandantahunan.aspx	Provide documents of company finance report from the first quarter 2005 until now. On the third quarter 2013, IDX have published document of company finance report from 500 companies in Indonesia.
3	Website of Indonesia Stock Exchange (IDX) (www.idx.co.id) or http://www.idx.co.id/id-id/beranda/publikasi/ringkasankinerjaperusahaantercatat.aspx	Provide document of company performance report. There are 464 companies who have published the documents with PDF format. These companies describe trading activities and finance of company every month since 2010 until now
4	Website of news media (http://financeroll.co.id) and (http://pasarmodal.inilah.com)	Financeroll provides finance condition from 25 companies and Pasarmodal provides finance condition from 44 companies in HTML format.

All sources provide some variables and data related and needed by BTS in internet freely and routinely. From these sources, researcher can extract data about company profile, operating revenues, number of employees, lending rates, financial and credit. Researcher is building an information extraction application that can extract needed data from various defined sources. Extracted data are stored in structured format in database.

Third case- Data Satisfaction Survey aims to know consumer demand for statistical data as well as the level and period of data and to know performance of Statistics Indonesia according to data consumer. Data are collected by interviewing the data consumer who comes to central/ province/district offices using questionnaire. Besides interviewing data consumer, researcher proposes new alternative, i.e. sentiment analysis. Sentiment analysis aims to determine the attitude, respond, criticism of a speaker or a writer with respect to statistical data topic. Corpus for sentiment analysis can be collected from various social media such as Twitter, Facebook, and news website or news blog. Wijayanto, et al. have explored corpus from Facebook and Twitter and made sentiment analysis applications which provide recommendation for Statistics Indonesia in making decisions and policy related to statistical data publication.

VI. Big Data Opportunities for Statistics Indonesia

In three cases that have been described in the previous section, researcher identified and learned that big data give some opportunities for Statistics Indonesia. Big data give opportunities as a control, a substitution, and a new product.

Big data as a control, it means that statistical data generated from big data can be a control for statistical data generated from censuses and surveys. In case of Business Tendency Survey (BTS), by applying information extraction system which can extract related data in defined documents, researcher can get some data that related and needed by BTS without doing survey. Extracted data are company profile, operating revenues, number of employees, lending rates, financial and credit from around 400 companies in Indonesia. Extracted data can be compared to data that collected by survey. Comparison conducted on same companies, same time range, and same variables. If results of comparison are different, Statistics Indonesia should evaluate to know the cause of difference.

Big data as a substitution, it means that statistical data generated from big data can substitute statistical data generated from censuses and surveys. In case of Consumer Price Survey, collecting consumer prices daily, weekly, and monthly exert many employees, costly, and require long time. Establishing legal cooperation with modern market and building online submission system that allow modern market to share the consumer prices with Statistics Indonesia so that can save labors and time.

Big data as a new product, it means that analyzing big data can provide new statistical data that never produced by Statistics Indonesia. In case of Data Satisfaction Survey,

Statistics Indonesia can get corpus from social media then extract them to analyze sentiment from users. Sentiment analysis has not yet been done. By doing sentiment analysis, Statistics Indonesia can know perception of data consumer about quality of statistical data, quality of public services, and performance of Statistics Indonesia.

VII. Big Data Challenges in Statistics Indonesia

Big data has public participation [20], data are collected from social and applications searching that are main component of big data. It means 80% big data are generated spontaneously (non-administratively) from information subject and it reflects the real requirement and preference of public [20]. Researcher identified big data challenges in each stage of statistical methodology for producing statistical data.

Program Designing Stage, statistical designer should figure out how the existing data come from. Big data are quite an opposite flow compared with traditional statistical designing. If big data become a part of data source, Statistics Indonesia should consider, evaluate, and modify sampling frame.

Data Collecting Stage, for statistical methodology, the quality of raw data depends on the feedback of respondent as well as high cost. However, if parts of raw data are extracted from records collected technology, it will weaken the data fraud and lower cost. Statistics Indonesia must build the proper information extraction system and intelligent management system to join the statistical data.

Data Analyzing Stage, big data consists of not only structured data but also semi-structured and unstructured data. It means that the traditional database management system is not quite to handle such data. Statistics Indonesia must prepare other storage system such as NoSQL database that can stored any kind types of data.

Data Releasing Stage, big data require transparency and pertinence of statistical data [20]. Statistical data without detailed of explanations and fairly statistical methods lead to worthless will be queried by public and replaced by others. On the other hand, digging on the potential users and providing them with the most concerned data will make statistical data more competitive and authority.

Legislation, many potential big data sources are collected by non-governmental organizations or are freely available on the web. Everyone or organization can catch up the data; situations may not be covered by the existing legislation. It needs law or government regulation that can describe rights, obligation, and responsibility for both of Statistics Indonesia and private sector.

Security and Privacy, the personal information of a person when combined with external large data sets leads to the interface of new facts about that person and it is possible that these kinds of facts about the person is secretive and the person might not want the data owner to know or any person to know about them [11].

Storage, processing, and data access, because data are collected from many sources such as surveys, censuses, websites, sensors devices, social media, and log website, the volume of data become bigger and the storage available is not enough for storing the large amount of data. The existing traditional data processing system also must be evaluated, is still to process and maintain the data. Network and access mechanism should be considered to maintain data access stabile.

Skill Requirement, to achieve the value delivered by big data and handle big data's challenges, Statistics Indonesia needs to have various type of skill. Skill to maintain and data in very huge scale, skill to be database administrator, data analyst, network specialist, business intelligence skill such as neural networks, statistical skills such as hands-on experience with Bayesian models, etc. Statistics Indonesia should try to understand about Hadoop, Map Reduce, and related frameworks such as Hbase, Pig, and Hive.

Financial, Statistics Indonesia should analyze potential costs of sourcing data versus benefits. Statistics Indonesia will need costs to acquire big data, especially big data held by the private sectors and especially if legislation is silent on financial modalities.

VIII. Conclusion

In three cases described in previous section, big data can be a part of statistical data as a new data source. Big data offer many opportunities when Statistics Indonesia tries to extract the value of big data. Besides that, big data also offer many challenges that must be learned and analyzed well. Researcher suggest that Statistics Indonesia need to study comprehensively and continuously to understand how much effort must be done and how many benefits will be got when trying to extract the value of big data.

IX. Future Work

In the future research, it is good to continue this research that focus on specific big data that can be applied by Statistics Indonesia and research all supported technology. Big data has five phases: acquisition / extraction, transformation, storage, processing, analysis and visualization. NSO must consider all phases to get benefits and value from big data. In the next time, it is possibility to research five phases of big data.

References

[1] James E. Short, Roger E. Bohn, & Chaitanya Baru. *How Much Information? 2010* Report on Enterprise Server Information.

[2] www.youtube.com/yt/press/statistics.html. Accessed on 3 April 2014.

[3] www.statisticbrain.com/twitter-statistics. Accessed on 3 April 2014.

[4] www.statisticbrain.com/facebook-statistics. Accessed on 3 April 2014.

[5] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *J. Parallel Distrib. Comput.*, Feb. 2014.

[6] A. Katal, "Big Data : Issues , Challenges , Tools and Good Practices", *The Sixth International Conference on Contemporary Computing,* pp. 404–409, 2013.

[7] Z. Zheng, J. Zhu, and M. R. Lyu, "Service-generated Big Data and Big Data-as-a-Service : An Overview", *2013 IEEE International Congress on Big Data,* 2013.

[8] Nambiar, R and Adhiraaj Sethi, "A Look at Challenges and Opportunities of Big Data Analytics in Healthcare", *2013 IEEE International Conference on Big Data*, pp. 17–22, 2013.

[9] Ludena, D.A and Alireza Ahrary, "A Big Data approach for a new ICT Agriculture Application Development", *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2013.

[10] I. Paik, T. Tanaka, H. Ohashi, and W. Chen, "Big Data Infrastructure for Active Situation Awareness on Social Network Services", *2013 IEEE International Conference on Big Data*, pp. 411–412, 2013.

[11] http://www.whitehouse.gov/blog/2014/01/23/big-data-and-future-privacy. Accessed on 3 April 2014.

[12] R. C. Joseph, P. State, and N. A. Johnson, "Big Data and Transformational Government", *IEEE Computer Society*, pp. 43–48, 2013.

[13] H. Moon and H. S. Cho, "Big Data and Policy Design for Data Sovereignty: A Case Study on Copyright and CCL in South Korea," *2013 International. Conference.on Soial. Computing.*, pp. 1026–1029, Sep. 2013.

[14] James Manyika, Michael Chui, brad Brown, jacques bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute

[15] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data," *2012 Joint. Working IEEE/IFIP Conference.on Software Archiecturet.and European. Conference Software Architecture.*, pp. 215–218, Aug. 2012.

[16] www.bps.go.id. Accessed on 1 June 2014.

[17] Statistics Indonesia, "Ringkasan Kegiatan Metadata Statistik Indonesia 2012", 2012.

[18] Statistics Indonesia, "Ringkasan Kegiatan Metadata Statistik Indonesia 2013", 2013.

[19] Statistics Indonesia, "Indeks Tendensi Bisnis dan Indeks Tendensi Konsumen 2013". *Publication*. 2013.

[20] Damin, Liang and Jinjing, Cheng, "Big Data and Official Statistics in China", 2014. Working Paper. Meeting on The Management of Statistical Information System.

[21] Itou, Takao, "Big Data and Official Statistics: Analysis of Recent Discussion in Statistical Communities", 2014. Working Paper. Meeting on The Management of Statistical Information System.

[22] Albert, Jose Ramon G, "Challenges, Opportunities and Issues on Using Big Data for Meeting Current and Emerging Demands on Measuring Progress and Development", 2014. Working Paper. Meeting on The Management of Statistical Information System.

[23] Ahn, Jeong-Im and Young-Ja Hwang, "Production of Official Statistics by Using Big Data", 2013. Working Paper. Meeting on The Management of Statistical Information System.

[24] Daas, Piet and Loo, Mark van der, "Big Data (and Official Statistics)", 2013. Working Paper. Meeting on The Management of Statistical Information System.

[25] http://www.carrefour.co.id/en/shop/carrefour/. Accessed on 1 August 2014.

[26] http://wartaekonomi.co.id/berita20996/trans-retail-indonesia-remodeling-gerai-di-karawang-jawa-barat.html. Accessed on 28 Auguts 2014.

[27] http://www.superindo.co.id/tentang_kami/sejarah. Accessed on 28 August 2014.

[28] http://www.hypermart.co.id/en/about- hypermart/store-location/43-store-location#panel1. Accessed on 28 August 2014.

[29] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," pp. 1320–1326

[30] Wijayanto, Arie Wahyu, Indri Djoko T., A. Muchlis A., "Integrated Mobile Social Media Sentiment Analysis Based on Internet of Thing untuk Pendukung Kebijakan Pimpinan Badan Pusat Statistik", Institut Teknologi Bandung, 2014.