Cloud-Based Processing for Data Science Visualization

Ahmad Ashari
Department of Computer Science
and Electronics
Faculty of Mathematic and Natural
Sciences, Universitas Gadjah Mada
Yogyakarta, Indonesia

A Min Tjoa
Institute of Software Technology
Vienna University of Technology
Vienna, Austria

Mardhani Riasetiawan
Department of Computer Science
and Electronics
Faculty of Mathematic and Natural
Sciences,
Universitas Gadjah Mada
Yogyakarta, Indonesia

***Abstract*—Data scientists need to process and visualize data science for scientific and decision purposes. The data have different size, type, real-time or batch forms, and validity. Data science visualization has a challenge in processing, management, and technique. The research works to investigate, design, and develop the cloud-based processing for data science visualization. The research uses Google Drive as file storage, Google App Engine as the processing tool, and Google Fusion for the visualization. Financial and banking data from Indonesia are used in the research to provide geolocation data, transaction flows, and bank networks information. Cloud-based processing consists of a data mapping process, data tagging, data manipulation, and data visualization. The research focus is on the data source manipulation, data preparation, storage management, data processing, and visualization. This research contributes to delivering cloud-based approach to handle data science visualization of financial-banking data networks in Indonesia.**

Keywords—component; cloud-based processing; display; data science; big data

I. Introduction

The increase in data size, type of data, data stream or batch, and the data structure is one of the issues in big data processing [1]. Computer processing has a different method and approach based on the data characteristics. It will become complicated for the data scientist to deliver the processing plan. There is also a need to understand the business process, information architecture, information system design, data structures, and delivery system designs [2]. In the term data science, we need to define the business process that should be used to deliver the information. The data science needs the word of knowledge to define business process [3]. The data that come from different sources is managed together in the store and arranged in structured or unstructured formats.

The information architecture specifies the detail of data and information [4]. The structure is used to define the data feature in the first process and the results. The information system design needs to know the information structure to describe the process and related information [5]. The interaction between information architecture and information system design requires establishing the process. The data architecture manages the data science collection by identifying the data details, in this case metadata and content [6]. The data processing method can deliver in several ways, such as integration, offline by using tools, online by using web application, and hybrid by using them in combination [7]. The processing technology approach uses real-time, batch, and stream. The method and technical approach are combined based on the purposes.

The research investigates cloud-based processing in data process for data science visualization. The research designs the cloud-based processing steps for managing the data. The technology approach used in this study are cloud-based applications such as Google Drive, Google App Engines, and Google Fusion. The study uses the financial-banking data in Indonesia provided by Open Data Indonesia [8]. The research goal is to deliver the data science visualization of intercity-network bank in Indonesia. This research has a contribution to the methods of cloud-based processing for data visualization as a best practice to deliver the data knowledge on particular issues. The paper has the following sections: Section II presents the current approach and method for data visualization. Section III delivers the step-by-step method on cloud-based processing. Section IV shows the result and discussion. The last section shows the conclusion and future direction.

II. Data Science Process and Visualization

The primary issues in data processing and display are big data and data science research, such as machine learning, data mining, semantic web, social networks, and information fusion [9]. The research is based on an investigation and discovers a new technique in data processing, data representation, pattern mining, data storage, and visualization. The combination of the algorithm and the process approach is the primary concern to the resulting information. The big data and little (small) data management can combine to support many purposes. The use of little (small) data as a sample and generated to answer a question has been used for many reasons. The little (small) data can be used for defining the sample of the big data. It will improve the quality of data and the process itself. The big data will enable in spreading data and enhance the quality of the sample and results [10].

Data management for long-term use and access, especially for big data, is an important issue in managing the data value and usage. Data processing has the capability to address the problem of long-term access and use, not only in the present but also in the future [11]. Data processing for big data can be done by using distributed data mechanism at the storage and and work management levels. The technique of distributed data storage can increase the efficiency when provided through an Internet-enabled environment [12]. The mechanism supports the system architecture for cloud-based processing. Data science needs an enormous volume of resources. In several cases, the processing needs to share with other resources to enhance capacity. The shared resources become the big data services that need protection. An authentication scheme is implemented to protect user privacy on the research conducted by Jeong and Shin [13]. Big data processing focuses on end-to-end processing of data science integration, model, and evidence [14]. The approach delivers by process mining and bridges the gap between data science and process science. The process mining use big data technologies, service based and cloud services.

The big data system architecture consists of several components, that is, data visualization, processing (include real-time, structured database, interactive analytics, and batch processing), data structure, and infrastructure. The data visualization in big data science delivers the intelligence visualization [15]. The intelligence visualization displays information and knowledge. The real-time process, analytics, and batch processing need to address speed, reliabilities, and data spread especially in processing purposes [16]. The data is classified into structured and unstructured data [17]. The infrastructure needs to address the high-performance infrastructure to support the processing needs [18]. Figure 1 presents the interaction between the components.

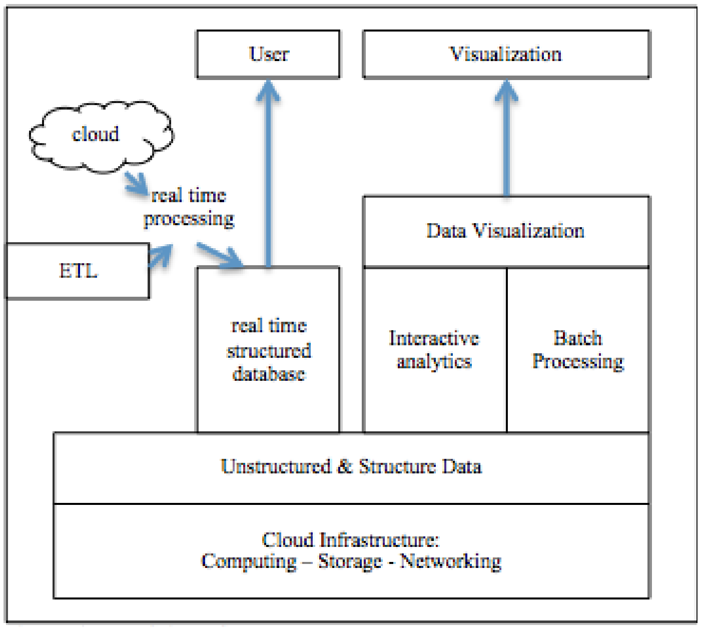


Fig. 1. Big Data Science System

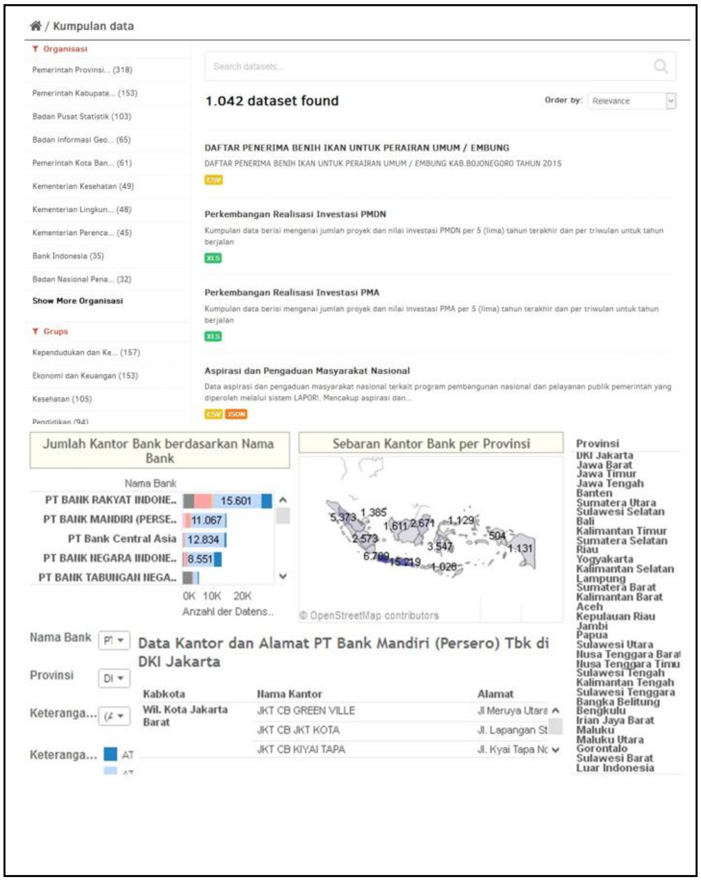
Emergency management is used in the case study and helps in overcoming the trending issue in emergency management. The visualization has been used to describe the difference between the type of record and history based on the provenance [19, 20]. The research is an organizational framework to specify the origin and design knowledge on it. Reactive Vega has presented a system architecture for graphic visualization and interaction [21]. The research constructs the data flow graph, scene graph, and interaction with streaming data. The display has been built with the help of time scale, relational, and hierarchical data.

III. Cloud -Based Processing

This section talks about the research design and works. The research was divided into several steps such as data preparation, storage management, data processing and manipulation, data integration, and data visualization.

A. Data Sources

The research uses the data from Indonesia Open Data portal. Open Data Portal (data.go.id) is a data portal built by the Indonesian Government to establish the open data movement and free data service. The open data portal itself has 1042 datasets, 31 institutions, and 18 groups of data. The research uses economic and financial data, provided by the Bank of Indonesia. There are 153 datasets consisting of economic and financial information from a broad range of regions in Indonesia. Figure 2 shows the set of the collections.


/ Kumpulan data

Organisasi

Pemerintah Provinsi... (318)
Pemerintah Kabupate... (153)
Badan Pusat Statistik (103)
Badan informasi Geo... (65)
Pemerintah Kota Ban... (61)
Kementerian Kesehatan (49)
Kementerian Lingkun... (48)
Kementerian Perenca... (45)
Bank Indonesia (35)
Badan Nasional Pena... (32)
Show More Organisasi

Grups

Kependudukan dan Ke... (157)
Ekonomi dan Keuangan (153)
Kesehatan (105)
Pendidikan (94)

Search datasets...

1.042 dataset found

Order by: Relevance

DAFTAR PENERIMA BENIH IKAN UNTUK PERAIRAN UMUM / EMBUNG
DAFTAR PENERIMA BENIH IKAN UNTUK PERAIRAN UMUM / EMBUNG KAB.BOJONEGORO TAHUN 2015
CSV

Perkembangan Realisasi Investasi PMDN
Kumpulan data berisi mengenai jumlah proyek dan nilai investasi PMDN per 5 (lima) tahun terakhir dan per triwulan untuk tahun berjalan
XLS

Perkembangan Realisasi Investasi PMA
Kumpulan data berisi mengenai jumlah proyek dan nilai investasi PMA per 5 (lima) tahun terakhir dan per triwulan untuk tahun berjalan
XLS

Aspirasi dan Pengaduan Masyarakat Nasional
Data aspirasi dan pengaduan masyarakat nasional terkait program pembangunan nasional dan pelayanan publik pemerintah yang diperoleh melalui sistem LAPOR!. Mencakup aspirasi dan...
CSV JSON

Nama Bank PT
Provinsi DI
Keteranga... (A
Keteranga... AT AT

Data Kantor dan Alamat PT Bank Mandiri (Persero) Tbk di DKI Jakarta

Kabkota	Nama Kantor	Alamat
Wil. Kota Jakarta Barat	JKT CB GREEN VILLE	Jl Meruya Utara
	JKT CB JKT KOTA	Jl. Lapangan St
	JKT CB KIYAI TAPA	Jl. Kyai Tapa Nc

Fig. 2. Open Data Portal

The research uses the dataset from the portal that was involved in the process, that is.:

- Bank Location

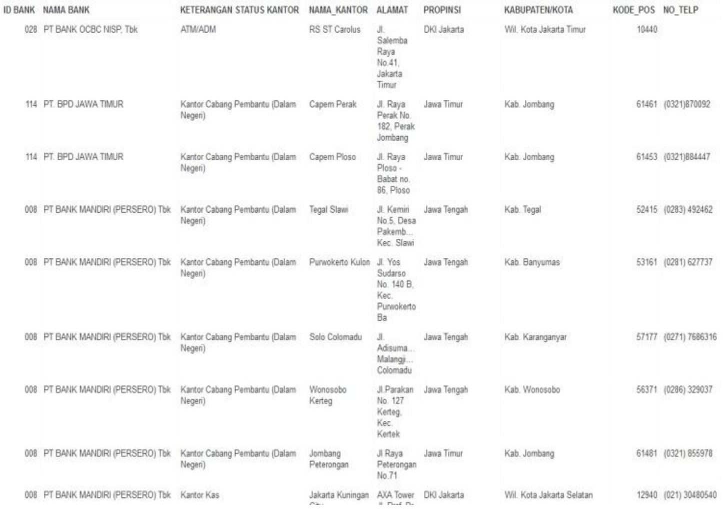
- Indonesia Bank Operation
- Transaction Volume
- Regional economic indicators

B. Data Preparation

The data preparation uses the data bank locator, operation, transaction volume, and economic indicators. The data preparation has several steps; there is data normalization, data cleansing, and data tagging. Data normalization standardizes the data. The normalization identifies the region name, the bank office, the name of the bank, and the region classification.

Data cleansing is done to minimize the data error in geotagging and relation. The data cleansing process consists of taking a data sample of at least 30 items of data. The data is transformed into the visualization prototype. The process is to figure out whether there are data items that cannot be processed based on the current data.

Data tagging has two options. Geotagging is used to give location information to the data object such as location, bank office, and transaction data. The second option or geolocation uses Google Map API facilities to attach to it. The result of this process is presented in Figure 3.



ID BANK	NAMA BANK	KETERANGAN STATUS KANTOR	NAMA_KANTOR	ALAMAT	PROPINSI	KABUPATEN/KOTA	KODE_POS	NO_TELP
028	PT BANK OCBC NISP, Tbk	ATM/ADM	RS ST Carolus	Jl. Salemba Raya No.41, Jakarta Timur	DKI Jakarta	Wil. Kota Jakarta Timur	10440	
114	PT. BPD JAWA TIMUR	Kantor Cabang Pembantu (Dalam Negeri)	Capem Perak	Jl. Raya Perak No. 182, Perak Jombang	Jawa Timur	Kab. Jombang	61461	(0321)870092
114	PT. BPD JAWA TIMUR	Kantor Cabang Pembantu (Dalam Negeri)	Capem Ploso	Jl. Raya Ploso - Babat no. 86, Ploso	Jawa Timur	Kab. Jombang	61453	(0321)884447
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Tegal Slawi	Jl. Kemiri No.5, Desa Pakemb... Kec. Slawi	Jawa Tengah	Kab. Tegal	52415	(0283) 492462
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Purwokerto Kulon	Jl. Yos Sudarso No. 140 B, Kec. Purwokerto Ba	Jawa Tengah	Kab. Banyumas	53161	(0281) 627737
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Solo Colomadu	Jl. Adisuma... Malangji... Colomadu	Jawa Tengah	Kab. Karanganyar	57177	(0271) 7686316
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Wonosobo Kerteg	Jl.Parakan No. 127 Kerteg, Kec. Kertek	Jawa Tengah	Kab. Wonosobo	56371	(0286) 329037
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Jombang Peterongan	Jl Raya Peterongan No.71	Jawa Timur	Kab. Jombang	61481	(0321) 855978
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Kas	Jakarta Kuningan	AXA Tower	DKI Jakarta	Wil. Kota Jakarta Selatan	12940	(021) 30480540

Fig. 3. Data Preparation

C. Storage Management

The cloud-based processing is stored the data and the process in the Internet facilities. The research uses Google Drive to place the data, Google App Engine to access the data stored in the intermediate storage and database engine, and Google Fusion to process the data and visualize it as presented in Figure 4.

D. Data Processing and Manipulation

The data processing and manipulation have several steps: card process, mapping, chart, and summary. The data proceed first into the card. In this process, the data are collected into the record. The data become an individual item that will be used to continue the relation and data network. The card process result is presented in Figure 5.

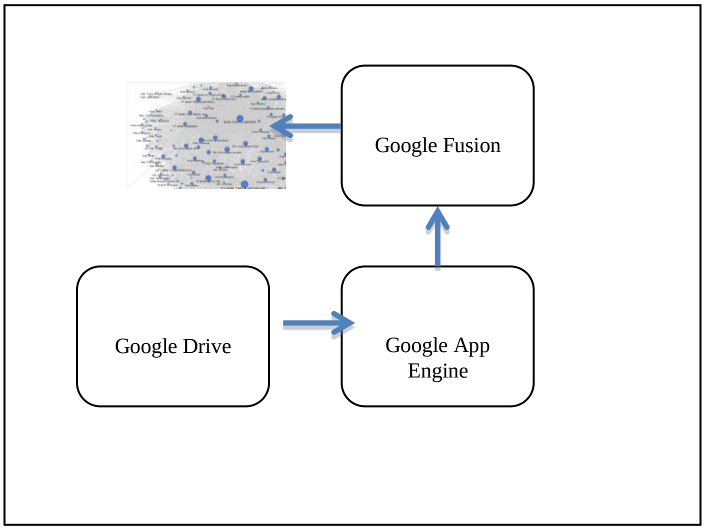
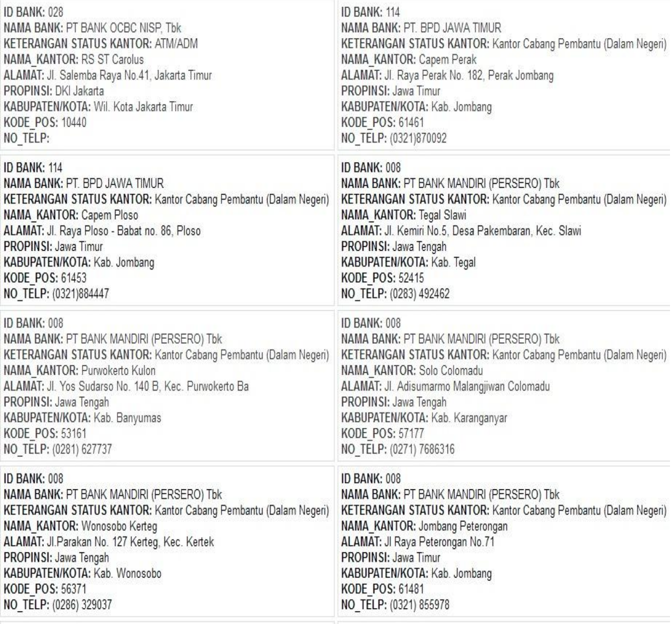


Fig. 4. Storage Management



ID BANK: 028
NAMA BANK: PT BANK OCBC NISP, Tbk
KETERANGAN STATUS KANTOR: ATM/ADM
NAMA_KANTOR: RS ST Carolus
ALAMAT: Jl. Salemba Raya No.41, Jakarta Timur
PROPINSI: DKI Jakarta
KABUPATEN/KOTA: Wil. Kota Jakarta Timur
KODE_POS: 10440
NO_TELP:

ID BANK: 114
NAMA BANK: PT. BPD JAWA TIMUR
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Capem Perak
ALAMAT: Jl. Raya Perak No. 182, Perak Jombang
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61461
NO_TELP: (0321)870092

ID BANK: 114
NAMA BANK: PT. BPD JAWA TIMUR
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Capem Ploso
ALAMAT: Jl. Raya Ploso - Babat no. 86, Ploso
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61453
NO_TELP: (0321)884447

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Tegal Slawi
ALAMAT: Jl. Kemiri No.5, Desa Pakembaran, Kec. Slawi
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Tegal
KODE_POS: 52415
NO_TELP: (0283) 492462

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Purwokerto Kulon
ALAMAT: Jl. Yos Sudarso No. 140 B, Kec. Purwokerto Ba
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Banyumas
KODE_POS: 53161
NO_TELP: (0281) 627737

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Solo Colomadu
ALAMAT: Jl. Adisumarmo Malangjiwan Colomadu
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Karanganyar
KODE_POS: 57177
NO_TELP: (0271) 7686316

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Wonosobo Kerteg
ALAMAT: Jl.Parakan No. 127 Kerteg, Kec. Kertek
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Wonosobo
KODE_POS: 56371
NO_TELP: (0286) 329037

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Jombang Peterongan
ALAMAT: Jl Raya Peterongan No.71
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61481
NO_TELP: (0321) 855978

Fig. 5. Card Process

The data is also used in the mapping process. The data uses the location parameter by rendering and process to have geolocation based on Google Map. The mapping process resulted in a card that had the information location. It also processes the transaction data. The next process is a chart and summary. The process is used to create a relation between datasets to map the network process. The summarizing will give weight to every data and the bank location to visualize in the representation burden.

E. Visualization

The visualization process works to display all the information that resulted from previous steps appropriately.. The visualization process itself has a particular format. The process identified the bank institution and location (city) as primary nodes, and transaction and other data as weight indicator for primary nodes. It needs not only for visualization and give the value of nodes. Table I shows the visualization process. The visualization process is rendered from the dataset and displayed in the HTML format.

TABLE I. VISUALIZATION PARAMETER

Primary	***Weight***
Location Bank Name	Transaction Volume Indicator

IV. RESULT

The research has resulted in a working visualization prototype for displaying the bank, location, and transaction weight based on the cloud-based processing. The display shows the network maps chart as shown in Figure 6.

The visualization result demonstrates the bank, location, and the transaction. The nodes have a different size based on the transaction weight on it. The visualization can be dynamic and comes out with the other data.

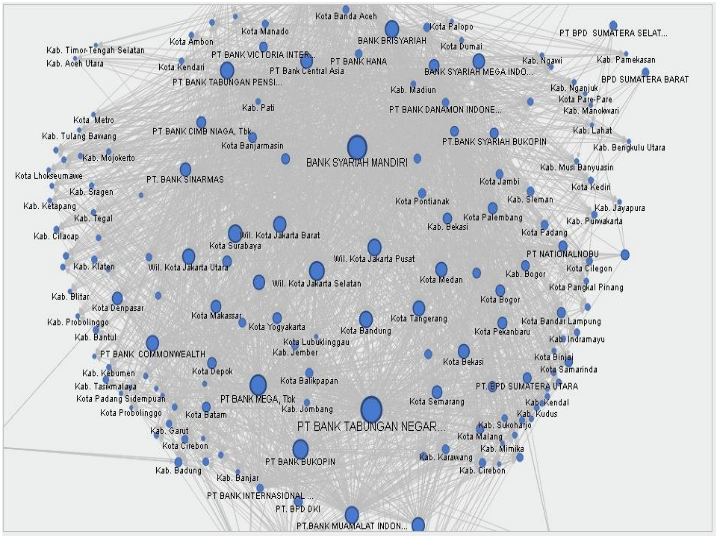


Fig. 6. Visualization Result

The visualization can display the network between the banks that operate in several cities, as shown in Figure 7. The relation between cities and banks is illustrated in Figure 8.

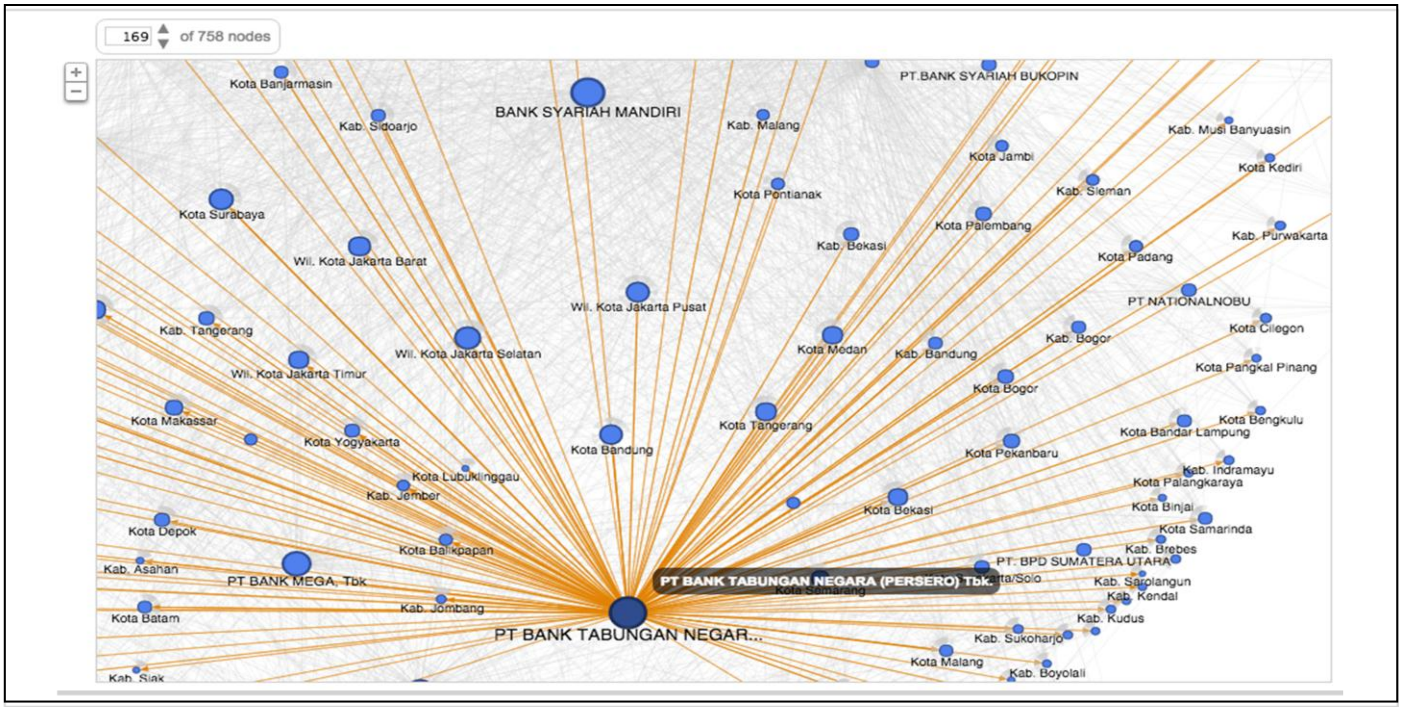


Fig. 7. Dynamic Visualization based on The Bank

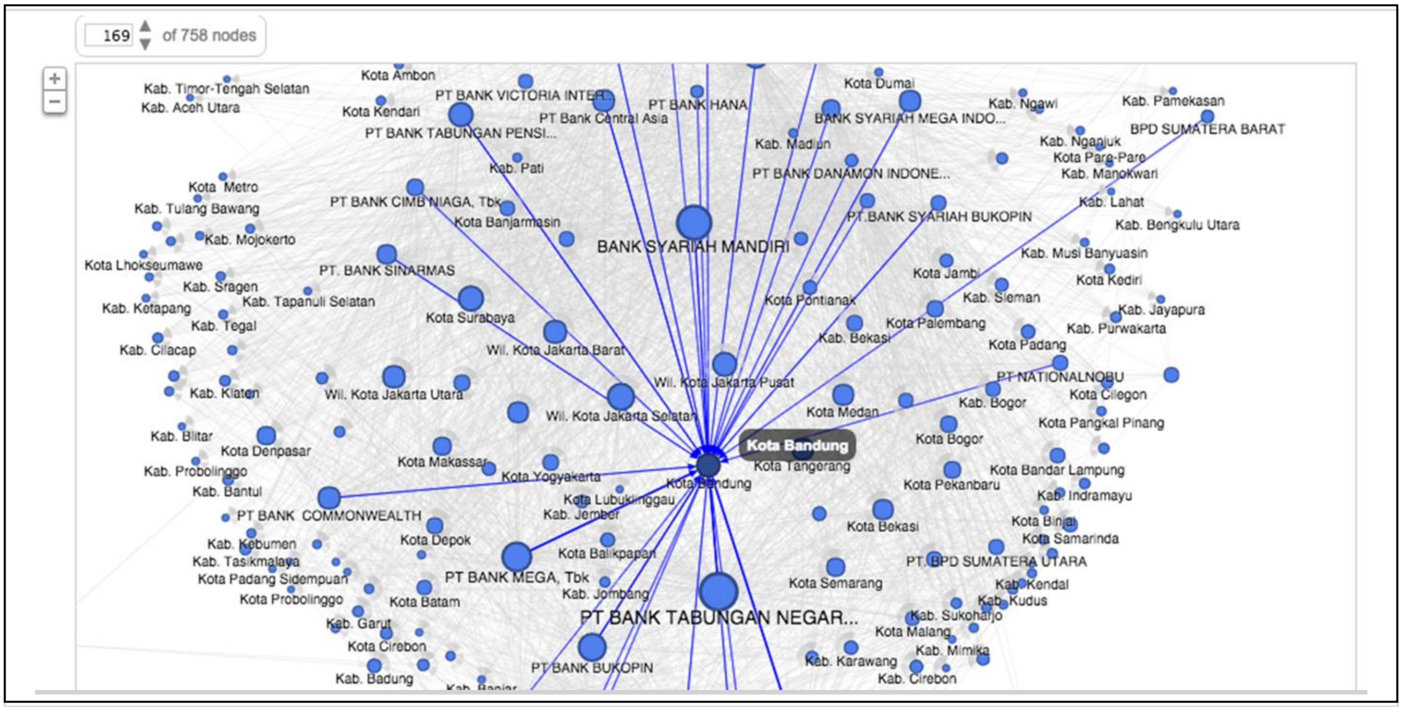


Fig. 8. Dynamic Visualization based on the city

V. Conclusion and Discussion

The research has shown the best practice of using cloud-based approach to process the data science, which is big data, with several steps. The conclusion of this research is that the cloud-based approach utilized for data science purposes, in this case, uses Google application. The research works with an open data sample, and the visualization is presented in http://makeiswork.com/2015/12/23/show-case/ as a working prototype. The research work has a contribution to the process of data science into the visualization that uses cloud-based approach. The process consists of data preparation, storage management, data processing and manipulation, and display. In this, every process needs a unique approach to ensure the quality of data. The data process is unique and depends on the data characteristic itself. The process involving more datasets will need more processing. The work on visualization depends on the process.

The research on data process and display has a challenge in the multiple datasets involved. The process even uses a framework tool but still needs to have a well-designed approach and methods. The cloud-based approach addresses the process on the Internet. The approach needs to address the multiple sources handled in the cloud-based process. The work on the approach will be the key for many organizations in business decision-making, business analysis, and intelligence, or scientific analysis.

Acknowledgment

The research was supported by ASEA UNINET, Vienna University, and Vienna University of Technology. The research uses sample data from Open Data Indonesia (data.go.id), and places the visualization in www.makeiswork.com which in turn is supported by DALA-intelligent technology initiatives.

References

[1] A.A. Al-Habsi, D.K. Kang, M.J. kim, “Enhancing Dataset Processing in Hadoop Yarn Performance for Big Data Applications,” Lecture Notes in Electrical Engineering, vol. 354, pp. 9-15, 2016.

[2] G. Quicmayr, “The Boeing Information Services 2010 Study,” Lecturing Materials, Institut fur Distributed and Multimedia Systems, Universitat Wien, 2015.

[3] F. Rahimi, C. Moller, L. Hvam, “Business Process Management and IT Management: The missing Integration, “ International Journal of Information Management, vol. 36, pp. 142-154, 2016.

[4] C. Maican, R. Kixandroiu, “A System Architecture Based on Open Source Enterprise Content Management System for Supporting Educational Institutions,” International Journal of Information Management, vol. 36, pp. 2017-214, 2016.

[5] R. Dijkman, I. Vanderfeesten, H.A Reijers, “Business Process Architecture: Overview, Comparison, and Framework,” Enterprise Information System, vol. 10, pp. 129-158, 2016.

[6] A.A. Neznanov, A.A. Parinov, “Distributed Architecture of Data Analysis System Based on Formal Concept Analysis Approach,” Studies in Computational Intelligence, vol. 616, pp. 265-271.

[7] C. Hu, X. Cheng, Z. Liu, “A Virtual Dataspaces Model for Large-scale Materials Scientific Data Access,” Future Generation Computer Systems, vol. 54, pp. 456-468, 2016.

[8] Open Government Indonesia, available at www.data.go.id

[9] G. Bello-Orgaz, J.J Jung, D. Camacho, “Social Big Data: Recent Achievements and New Challenges,” Information Fusion, vol. 28, pp. 45-59, 2016.

[10] R. Kitchin, T.P. Lauriault, “Small Data in the era of big data,” Geojournal, vol. 80, pp. 463-475, 2015.

[11] M. Riasetiawan, AK. Mahmood, “Managing and Preserving Large Data Volume in Data Grid Environment,” 2010 International Conference on Information Retrieval and Knowledge Management, Shah Alam Malaysia, pp. 91-96, 2010.

[12] S.T. Park, Y.R. Kim, S.P. Jeong, C.I. Hong, T.G. Kang, “A Case Study on Effective Technique of Distributed Data Storage for Big Data

Processing in The Wireless Internet Environment," Wireless Pers Communication, vol. 86, pp. 239-253, 2016.

[13] Y.S. Jeong, S.S. Shin, "An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services," Wireless Pers Communication, vol. 86, pp. 7-19, 2016.

[14] W.v.d. Aalst, E. Damiani, "Processes Meet Big Data: Connecting Data Science with Process Science," IEEE Transaction on Service Computing, vol. 8, pp. 810-819, 2015.

[15] E.A Mohammed, C. Naugler, B.H. Far, "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics," Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools, pp. 577-602, 2015.

[16] Cloudera, available at www.cloudera.com.

[17] B. Funk, P. Niemeyer, J.M. Gomez, "Information Technology in Environmental Engineering: Selected Contributions to the Sixth International Conference on Information Technologies in Environmental Engineering (ITEE 2013)," 2013.

[18] F. Teng. Management Des Donnees Et Ordonnancement Des Taches Sur Architectures Distributes. Thesis, Ecole Centrale Paris, 2011.

[19] F. Dusse, P:S. Junior, A.T Alves, R. Novais, V. Vieira, M. Mendoca, "Information Visualization for Emergency Management: A Systematic Mapping Study," Expert Systems with Application, vol. 45, pp.424-437, March 2016.

[20] E.D Ragan, A. Endert, J. Sanyal, J. Chen, "Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework for Provenance Types and Purposes," IEEE Transaction on Visualization and Computer Graphic, vol. 22, pp. 31-40, January 2016.

[21] A. Satyanarayan, R. Russel, J. Hoffswell, J. Heer, "Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization," IEEE Transaction on Visualization and Computer Graphic, vol. 22, pp. 659-668, January 2016.