A Study Review of Common Big Data Architecture for Small-medium Enterprise

Ridwan Fadjar Septian[1], Fajri Abdillah[2], Tajhul Faijin Aliyudin[3]
{ridwanbejo@gmail.com[1], clasense04@gmail.com[2], tajhulfaijin@gmail.com[3]}

Master of Information System, Faculty of Postgraduate, Universitas Komputer Indonesia, Indonesia [1], Senior Software Engineer, Horangi Cyber Security [2], Senior Software Engineer, Tado.live[3]

Abstract. This paper will be composed by conducting a document review and collect papers that related to big data solution and infrastructure. The survey is trying to cover the phases for building a common big data pipelines by following these structures: enterprise vision and mission that affect the big data architecture, defining the data sources, building the backend services to collect the datapoint from the data sources, utilize the data pipeline to prevent data loss, consuming the data pipeline message and deliver it to the data storage, build the raw data storage, define business-driven data visualization and analytics, build the ETL, build business-oriented data storage, applying statistical and machine learning, perform the business-driven visualization and implementing the monitoring for the big data pipelines. The survey will emphasize that many big data components could help the small-medium enterprise to tackle their big data operational issues.

Keywords: Big data Architecture, Small-medium enterprise, Open-source, Large-scale dataset, Data engineering

1 Introduction

An enterprise could leverage big data infrastructure whenever they put attention to large-scale datasets produced by their online services which are supported their business lines. Big data is a set of facilities that process those large-scale datasets in a complex way besides the traditional infrastructure that only applied with a small amount of dataset. Amount of users, mobile devices and the internet of thing become a factor for the enterprise which could gain more dataset from their user to be analyzed and processed to support their businesses with distributed processing and large-scale storage. On the other hand, big data should apply fast processing through large-scale dataset which is the traditional data processing could not achieve for [1].

Big data has six common characteristics that distinguish itself from traditional data processing. Volume: The quantity per each day could be more than gigabytes until petabytes depend on the scale of the business. Its also affected by enormous volume which is coming from sensors such as the Internet of Things (IoT) and mobile devices, streaming data such as video from CCTV and also human interactions through social media that yield a dataset such as text and photo for every minute. Variety: talk about the type of datasets such as structured data, semi-structured data, and unstructured data. Veracity: related to the quality of the dataset

MSCEIS 2019, October 12, Bandung, Indonesia

DOI 10.4108/eai.12-10-2019.2296535

even before and after data preprocessing. Velocity: Collecting the dataset might be done in an intensive process and request and bring a large-scale dataset from huge clients. Also, big data technology should not be allowed to loses the dataset on the production level. Value: unstructured and structured large-scale datasets could contain a piece of useful information and pattern to be converted as a knowledge for business purposes. The value from the big data should be handled through data processing techniques with the capability of distributed processing. Variability: the dataset could be sparse and must be cleaned up through the preprocessing phase before it is being used for further processing such as trained as a dataset for machine learning application. The source of truth of the data itself must be validated also before performing analyze and process those datasets [1].

Big data term has several kinds of methodology which are consist of the acquisition of the dataset, organizing the dataset, analyze the dataset and finally deciding the result [2]. In more detail, there are three more steps such as data acquisition, data storing, data management, data analysis and data visualization [4]. As a form of advanced data analysis, some researcher tries to apply machine learning for the processed dataset to gain more insight from the dataset such as applying an algorithm for regression learning, classification, natural language processing, association rule mining, clustering, forecasting and deep learning [2,3,5,7]. As an emergence topic both in research and industry, big data has brought a lot of issues as well that need to be handled from the industry or research perspective. There is some risk in big data such as data security and privacy protection, data ownership and transparency, data quality, cost of implementation, infrastructure maintenance, developer failure, security system and compliance with the regulation. [1,2,3,6,7].

Big data is adopted in several fields such as in education when big data is built to support institutional analytics, information technology analytics, academic analytics and learning analytics [6]. In environmental issue, in order to reduce the climate change, big data is performed to apply energy efficiency for the resource usage in manageable way, build a sustainable farming and agriculture that keep the positive impact for the ecosystem, designing smart city that could reduce the electricity usage, establish surveillance and monitoring system for disaster management and creating national strategy without deforestation and damaging the ecosystem [5]. In the healthcare industry, big data has taken part for involved in personalized patient care healthcare, improve pharmaceutical to produce an innovative medicine based on medical record and genomic information, shaping medical device design and manufacturing, fraud detection on medical claim, research on preventive action for certain diseases based on genomic analysis [4]. On the other hand, in the construction industry, big data has taken some role in performing resource and waste optimization, facility management with the internet of things, and energy management of buildings [3]. Some of the use cases also being applied to social network analysis and natural language processing [2].

In that case, the survey wants to pick a study review to find a common big data architecture that might utilize some open source software to be included in the architecture. Some infrastructure that might be useful to establish big data architecture is consists of message queue, data storage, ETL application, Hadoop-based analytics tools, data warehouse, and visualization. Not only the infrastructure, but The survey also wants to bring some process from a business perspective to make the big data architecture more aligned with the business goals. The survey wants to introduce the TOGAF and CRISP-DM together to see the potentiality for sharing big data architecture.

2 Methods

The qualitative methodology of this research was taken by the following phases which are applied during the research; 1) Research Planning, 2) Scoping the research, 3) Data Acquisition by reviewing the number of papers, 4) Conclusion and Recommendation. Research planning is set to perform a study review for the common architecture of big data. The study review aims to find the architecture that could be implemented by the small-medium enterprise with expected low cost. Scoping the research into big data architecture only is the primary goal for this study review. This study review didn't aim to cover a detail component of machine learning or technical matters of reviewed open-source software. Also, the study didn't cover the cost calculation for building the architecture. Data acquisition is conducted by searching expected keywords over journal repository then obtain some papers that match with this study review. Around 42 papers are obtained to become references for this study review then gain facts and innovations from those papers. Conclusion and recommendation will be stated after the result and discussion as a standpoint for extended research regarding the big data architecture.

3 Result and Discussion

3.1 Enterprise Architecture for Big Data Project

The Open Group offers a framework that emphasizes an architecture for an enterprise that helps them to align the enterprise roadmap and information technology. The framework is TOGAF v9. Big data is still part of the information technology project that has to be aligned with the business needs of the enterprise. By leveraging core phases of TOGAF v9, the enterprise might gain a good result to build robust big data architecture. TOGAF v9 has 9 phases that could be applied during architecting the big data project. That consist of Preliminary Phase, Phase A: Architecture Vision, Phase B: Business Architecture, Phase C: Information System Architectures, Phase D: Technology Architecture, Phase E: Opportunities and Solutions, Phase F: Migration Planning, Phase G: Implementation Governance and Phase H: Architecture Change Management [8].

Research showed that leveraging the enterprise architecture framework such as TOGAF v9 might rise the potential of stability for implementing the big data infrastructure. Occasionally, the enterprise has several issues when implementing big data infrastructures such as no business goal and no alignment with business requirements, poor planning and failure to recognize project scope, lack of access to data and communication between stakeholders, change management issues, focus on technology rather than business opportunities [9]. TOGAF also has a Target Capability Model that could drive the oil and gas enterprise to build big data to manage their big data infrastructure and migrate from the former data infrastructure [10].

3.2 Event Sources

Event sources for big data could be ingested from various sources such as mobile phone applications, internet of thing, web application or external existing cleansed data sources.

Those various event sources could be tracked by an enterprise to shape the business plan or change their strategy to reach their vision and mission.

The web application has also potentiality to gain insight from its users. Enterprise could track the user by their clickstream behavior, what they are searching for, what they are liking and sharing, what they are recommending to their friends, etc. Those events are valuable information to gain by the enterprise to help shape their businesses such as by collecting tweets from Twitter and perform analysis upon the collected tweets [10].

Sensors from the internet of things and IT infrastructures such as servers and networking devices could help the enterprise to monitor their performance and gain valuable information from the hardware. For instance, there is a project called SmartCampus that utilize sonar for detecting parking lot occupation and temperature sensor for heating regulation on the university area could be tracked by the university board to evaluate the policy for the parking and heating regulation. From that project, there could be useful for 390 Gb of database storage per year [11]. There is also a project in the industry area that collects 10.000 sensor values from the production site every 15 minutes and they also use Apache Nifi and MQTT protocol to collect the datasets [12]. In the manufacturing field, there is research that collects dataset about Return Air Temperature (RAT) and Set Point Temperature (SPT) from smart sensors then processed it through message queue to be cleansed for dashboard and reporting tools [13].

External data sources could be treated as an event source whenever the enterprise needs to load external data sources to their big data infrastructure. For instance, there is research that performing migration from a text file to PostgreSQL as a data warehouse by using S-Store as ETL tools that contain transaction data created by Transaction Processing Performance Council (TPC-DI) [14].

The mobile phone has several sensors that built-in from the manufacturer. On the other hand, an operating system vendor has their support to retrieve those sensor data through apps built by the enterprise then collect the data with or without consent from their users. For instance, geolocation data from mobile phone apps could help the enterprise in targeting its users with some campaign and advertisement. There is ta research that geotagged tweets could be collected to analyze Twitter's trending topics to be mapped to the map chart an visualize the relevant geo-tagged tweet in realtime [15]. It's not limited to the tagged content such as mentioned before, there is also research that enables the mobile apps to collect the geolocation of travelers and monitored them whenever they arrived in the city of their destination and spending time on the place visited by the tourists [16].

3.3 Message Queue

Message Queue is the terminology for software that retains some message for a certain period from producers and it will be processed by the consumers. Message queues are organized by certain of a topic and each topic has partition key. In that case, the consumer could be more than one instance and perform a reliable process to process the messages. The message queue is a conjunction for web application and data storage to prevent lost data during processing requests from the clients [17].

There are so many open source products that could be leveraged by the enterprises to work with such as Apache Kafka, RabbitMQ, Apache Storm, etc. Some of the products are

also supported for cluster set configuration to perform more resilient availability during receiving the requests from the clients. Below is the benchmark from some conducted researches for existing open-source message queue products during this study review as shown in table 1.

3.4 Data Lake

A data lake is a kind form of massive data storage that collects raw datasets before the dataset gets further processing on top of the Hadoop File System that initially introduced by James Dixon. Data lake could be utilized by using an open-source solution such as Hadoop File System from Apache Hadoop. Enterprise can build their data lake with mentioned open source products in their environment and more cost-effective than process it on the database [24]. The dataset could be stored on a data lake with various extensions such as CSV, Apache Avro, Apache Orc, Textfile, JSON, XML, etc.

Data lake have several key features such as large scale batch processing, schema on reading, able to store large scale data volume with low cost, could be accessed using SQL-like systems even they are not in SQL format dataset, complex processing that's even apply machine learning operation, store raw dataset instead compact format such as in SQL format, low cost for distributed processing [25]. The data lake has core components on the backend side such as catalog storage, batch job performance and scheduling, fault tolerance and garbage collection of metadata. On the other hand, for the frontend side, data lake has core components such as dataset profile pages, dataset search and team dashboards [26].

Data lake was introduced by The IBM Research Accelerated Discovery Lab in 2012. Their data lake implementation tended to support 500 researchers across multiple labs. They also divided the data lake into an external cloud and internal cloud data lake [27]. There are some approaches also to maintain the dataset inside the data lakes start from vetting data for licensing and legal usage, obtaining the dataset, describing the dataset, grooming dataset, provisioning dataset and preserving the dataset [28].

3.5 Extract Transform Load (ETL)

ETL (Extract, Transform, Load) is the part of big data that has a role to perform a conversion from raw dataset into a cleansed dataset. ETL is a set of tools that could help the enterprise to perform real-time analytics and decisions. ETL could be divided into three approaches that consist of micro-batch, near real-time and streaming. ETL must be a combination of scheduler and ETL script. The scheduler could use software such as Cron Jobs while Apache Kafka could be used for streaming approach. On the other hand, ETL script could be executed in distributed processes through a cluster of workers such as by using Apache Spark or executed in a single node such as using plain SQL or programming language approach. ETL could transform dataset from SQL format into other text formats or vice versa. Supported data format by ETL technology could be XML, CSV, JSON, Apache Avro, Apache Orc, etc. [14].

Data preprocessing is also part of the ETL phase that consists of imperfect data handling, dimensionality reduction, instance reduction, discretization, imbalanced data, incomplete data, and instance reduction. Data preprocessing could take apart in improving the result of the machine learning model [29].

3.6 Data Warehouse

The data warehouse is a kind of denormalized database that have generic information to cover management level question. A data warehouse is a kind of centralized data source that could supply the data to develop the strategic plan of the enterprise to make a better decision based on historical data that stored historical data. A data warehouse is built to perform online analytical processing (OLAP) over the data source to answer the business needs. The data warehouse receives the cleansed dataset that processed by the ETL part in the big data pipeline. On the other hand, data warehouses could be a source for data mining tasks to perform forecasting, classification, pattern recognition, clustering, etc [30].

Table 1. Open source message queue product comparison based on the referenced papers.

Products	Dataset	Hardware Specification	Process Throughput
Desk, Apache Kafka, Apache Spark [18]	Lightsource: raw the dataset in the APS format, an average encoded message size of 2MB	32 nodes, 1536 virtual cores of total nodes, 4 TB of RAM of total nodes	390 MB / sec
Apache Kafka, Apache Flink [19]	Illumina HiSeq 3000 at the CRS4: 12 human genomic samples, up to 47.8 Gb	12 nodes, 32 virtual cores per node, 244 Gb of RAM per node, disk x 1.9 Tb SSD per nodes, the network of 10 Gb Ethernet per node	450 MB / sec
Apache Kafka, Apache Spark, Apache Cassandra [10]	Twitter's tweets	3 nodes, 2 virtual cores, 4Gb of RAM	466.700 tweets / 10.7 minutes or 726 tweets / sec
Scribe, Stylus [20]	Facebook trending events	Not mentioned in the paper	135 MB / sec
RabbitMQ, Plain Consumer written in Python [21]	Not mentioned in the paper	RabbitMQ: 8x Intel Core i7 CPU 4GHz, 16 Gb of RAM, 1 TB of Harddisk. Publisher and Subscriber: Intel	50.000 message/second 1 – 3 of the subscriber.

Products	Dataset	Hardware Specification	Process Throughput
		Core i7 CPU 2.60 GHz, 8 Gb of RAM, 256 GB of Harddisk.	
Apache Kafka, Apache Storm [22]	Industrial IoT dataset	Microsoft Azure HDInsight	40.000 data / sec
Apache Kafka [23]	Not mentioned in the paper	5 nodes, 12 cores @2300.13Mhz per node, 16 GB of RAM per node, 16Gbps of network interface	100.000 ~ 200.000 message / sec
RabbitMQ [23]	Not mentioned in the paper	5 nodes, 12 cores @2300.13Mhz per node, 16 GB of RAM per node, 16Gbps of network interface	22.500 ~ 30.000 message / sec

In some cases, such as in the educational sector, the data warehouse has key roles to perform feasibility assessment and data analysis for an educational system from different viewpoints and make a quick decision to evaluate the educational system. Star schema could be also useful to build the data warehouse. Data warehouses could help the university to collect a huge amount of datasets from several kinds of existing databases and unify those data sources into a single data source. In that way, the university could perform a decision support system (DSS) technique over the data warehouse [31].

3.7 Data Mining Methodology

Data Mining is one of the methodologies over big data. There are some known methodologies to perform better data mining processes such as the KDD process, CRISP-DM, RAMSYS, DMIE, DMEE, ASUM-DM, AABA, etc. For instance, KDD has several steps that consist of selection, preprocessing, transformation, data mining, knowledge gain. On the other hand, a popular CRISP-DM has several steps that perform a complete operation that consist of Business understanding, data understanding, data preprocessing, modeling, evaluation, and deployment [32].

There is also a methodology that extended from CRISP-DM such as ASUM-DM which had developed by IBM. ASUM-DM has several steps that consist of analyze-design-configure & build, deploy and operate & optimize. A methodology that integrates an agile method such as AABA could become a robust method to perform the data mining process. AABA has several steps that consist of value discovery, design concepts catalog, architecture (BDD), implementation, testing and deployment [33]. DMME is also an extension of CRISP-DM by adding the step of technical understanding & conceptualization and technical realization & testing after business understanding step and before data understanding step. After the deployment step, DMME adds one additional step which is technical implementation [34].

Some research applied CRISP-DM in various sectors. In the retail sector, CRISP-DM is applied to perform a data mining process that uses association rule mining to predict the sales pattern [35]. Research on climatology has been conducted also by applying CRISP-DM and KDD as a combination to improve the data mining result [32]. Another research on social

media analysis has shown a result that CRISP-DM made the data mining processes gave a better result to perform the favorite TV series classification by applying the Decision Tree algorithm [36].

3.8 Data Visualization

Visualization could deliver more engagement to management level or other users so they can understand the insight that gains from the big data. Visualization could be formed as a web dashboard or document that contains an explanation and story. Data could be visualized with various graphs such as bar charts, line charts, pie charts, scatter charts, map charts, etc. Data visualization itself has several different types that consist of linear data visualization, planar data visualization, volumetric data visualization, temporal data visualization, multidimensional data visualization, and network data visualization [37].

Data visualization is not only established over the software alone, but it also needs a manageable way to work with. There is research that establishes a managed data visualization by using a collaborative framework. The data visualization is presented in the visualization room that everyone could join in. The main devices could share the interactivity with shared devices so everyone could mark their findings and decide together on the main device as a snapshot. Then the history for every activity is also saved to be evaluated for the next discussion [38].

Enterprise could also utilize an advance visualization such as graph visualization to help gain an understanding of security monitoring over network topology. The visualization could help the operator to identify which node who are being attacked faster [39]. On the other research on research paper clustering, advanced visualization is utilized to visualize the clustered dataset after being processed by the Hierarchical Clustering algorithm using a keyword map to see the correlation between topics among the clusters. A Strategic diagram is also utilized on that research to see the intersection between different clusters [40]. But besides generating the visualization manually, there is also a project that could generate visualizations over the dataset using a technology called DeepEye. That technology applies machine learning and expert rules to solve three main problems that consist of visualization recognition, visualization ranking, and visualization selection. Finally, DeepEye could generate an intuitive dashboard over the dataset as an input with minimum effort for the data analysts [41].

Some various open-source products can help the enterprise to visualize their insight from big data such as Pentaho, Apache Zeppelin, Apache Superset, Metabase, etc. Those products have capabilities to connect with Hadoop File System also with various database products that commonly used by the market.

3.9 Security and Compliance

With the vast amount size of the dataset and the potential threat to breach and attack, big data infrastructure should be secured and well-configured from the attacker to avoid misuse of the stored dataset. Some security approaches that might be applied for big data infrastructures such as encryption, security as a service, real-time monitoring, privacy by design, data

protection and authorization, log management, authentication, data anonymization and secure communication line [1].

For a more compacted manner, there is also a methodology issued by the NIST. NIST had a reference architecture proposal for big data that could help the enterprise establish a security system over its big data infrastructure. NIST for big data offers several components that consist of system orchestrator, data provider, big data application provider, big data framework provider, and the data consumer. NIST offers not only the management procedures within that proposal. Security and privacy have the same important position as well as management procedures [42].

There is a new standard also from ISO that focusing the compliance on big data. ISO is releasing ISO/IEC 20546:2019 which has been started to make compliance for big data infrastructure within the enterprise. On the other hand, the enterprise could also involve ISO 27001:2017 as a standard to keep their big data infrastructure secure from the attacker and unauthorized person [42].

3.10 Small Medium Enterprise in Big Data Era

Based on TOGAF standard, enterprise could be considered as a whole corporation or the division of that corporation, government agency or single government department, distributed organizations that linked together by common ownership and separated geographically, groups of countries or governments that working together to create common or shareable deliverables and partnerships or alliances of business that working together [8]. While small-medium enterprise is an enterprise that defined based on annual work unit, annual turnover, and annual balance sheet according to the European Commission [43, 44]. Small-medium enterprise is a differentiation of the enterprise itself but based on the size of the three indicators.

Based on prior researches, the small enterprise could have an annual work unit greater than equal 10 and less than 50 per year with annual turnover and annual balance sheet less than 10 million euros. Above 50 annual work unit until less than 250 annual work unit, the enterprise could be categorized into the medium enterprise if the company have an annual turnover and annual balance of less than 50 million euro. In that case, we could state that small-medium enterprise has an annual work unit around 10 until 250 annual work unit based on European Commission with European Union Standards [43, 44].

Some experts denied the necessity of big data for the small-medium enterprise. But there is also a statement who suggest small-medium enterprise learn their pattern from past transaction combine with external data to understand the market behavior to gain competitive advantage and growth from uncovering insight based on their data. Leveraging big data for small-medium enterprises could increase their product improvement and innovation against the competitor [45].

4 Conclusions

From the discussion above, this study review comes with a conclusion that could emphasize the common big data architecture for small-medium enterprises that could be categorized into three components that consist of design and architecture component, infrastructure component and operational component. First, on the design and architecture component, a small-medium enterprise could leverage enterprise architecture frameworks such

as TOGAF based on the study review. Second, on the infrastructure component, it could consist of event sources, message queue, data lake, extract-transform-load (ETL), data warehouse and data visualization. Third, on the operational component, the study found that data mining methodology is running on top of the infrastructure.

Furthermore, by using the methodology such as CRISP-DM or the other methodologies could produce a better data mining result. As an additional, security and compliance could be included in the operational component since the security and compliance is a life cycle for the sustainability of big data infrastructure and architecture against the threat, vulnerability, and risk. Based on this study review, the small-medium enterprise could found some open source products that could be established as a part of the big data infrastructure such as, Apache Nifi, Apache Hadoop, Apache Kafka, Apache Spark, Apache Storm, Scribe, RabbitMQ, Apache Zeppelin, Metabase, PostgreSQL, etc. Small-medium enterprises could start their big data projects with minimum cost over a software license for the early implementation of the big data project. There are so many proprietary products as well such as offered by Microsoft, AWS, Google, IBM and many more that could use by the small-medium enterprise.

Acknowledgment. We would like to say thank you to our companion (Zaky, Wildan, Rulsyah, Mirza, Nugraha, Firman, Doni) who help us to learn the big data and serverless technology. We would like to send our gratitude to our supervisor, Hubert Chan, who had to guide us to achieve big data knowledge during our employment.

References

[1] Bisht, P. and Singh, K. : Big Data Security: A Review of Big Data, Security Issues and Solutions. (2016)

[2] Emani, C.K., Cullot, N. and Nicolle, C. : Understandable big data: a survey. Computer science review, 17, pp.70-81 (2015)

[3] Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. and Pasha, M. : Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced engineering informatics, 30(3), pp.500-521 (2016)

[4] Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A. and Chandrakumarmangalam, S. : Big Data in healthcare management: a review of literature. American Journal of Theoretical and Applied Business, 4(2), pp.57-69 (2018)

[5] Hassani, H., Huang, X. and Silva, E. : Big Data and Climate Change. Big Data and Cognitive Computing, 3(1), p.12 (2019)

[6] Daniel, B.K. : Big Data and data science: A critical review of issues for educational research. British Journal of Educational Technology, 50(1), pp.101-113 (2019)

[7] Samuel, S.J., Rvp, K., Sashidhar, K. and Bharathi, C.R. : A survey on big data and its research challenges. ARPN J. Eng. Appl. Sci, 10(8), pp.3343-3347 (2015)

[8] The Open Group : TOGAF (The Open Group Architecture Framework) Version 9. ISBN: 978-90-8753-230-7. Dosscument Number:G091 (2009)

[9] Cameron, B.H. : The Need for Enterprise Architecture for Enterprise-Wide Big Data. ISJLP, 10, p.827 (2014)

[10] Nazeer, H., Iqbal, W., Bokhari, F., Bukhari, F. and Baig, S.U.R. : Real-time Text Analytics Pipeline Using Open-source Big Data Tools. arXiv preprint arXiv:1712.04344 (2017)

[11] Cecchinel, C., Jimenez, M., Mosser, S. and Riveill, M. : June. An architecture to support the collection of big data in the internet of things. In 2014 IEEE World Congress on Services (pp. 442-449). IEEE (2014)

[12] Sarnovsky, M., Bednar, P. and Smatana, M. : Big Data Processing and Analytics Platform Architecture for Process Industry Factories. Big Data and Cognitive Computing, 2(1), p.3 (2018)
[13] O'Donovan, P., Leahy, K., Bruton, K. and O'Sullivan, D.T. : An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. Journal of Big Data, 2(1), p.25 (2015)
[14] Meehan, J., Aslantas, C., Zdonik, S., Tatbul, N. and Du, J. : January. Data Ingestion for the Connected World. In CIDR (2017)
[15] Cha, S. and Wachowicz, M. : Towards Real-time Streaming Analytics based on Cloud Computing. International Journal of Big Data (ISSN 2326-442X), 2(1), pp.28-39 (2015)
[16] Klimek, R. : Towards Recognising Individual Behaviours from Pervasive Mobile Datasets in Urban Spaces. Sustainability, 11(6), p.1563 (2019)
[17] Blamey, B., Hellander, A. and Toor, S. : Apache Spark Streaming and HarmonicIO: A Performance and Architecture Comparison. arXiv preprint arXiv:1807.07724 (2018)
[18] Chantzialexiou, G., Luckow, A. and Jha, S. : December. Pilot-streaming: A stream processing framework for high-performance computing. In 2018 IEEE 14th International Conference on e-Science (e-Science) (pp. 177-188). IEEE (2018)
[19] Versaci, F., Pireddu, L. and Zanetti, G. : March. Kafka interfaces for composable streaming genomics pipelines. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 259-262). IEEE (2018)
[20] Chen, G.J., Wiener, J.L., Iyer, S., Jaiswal, A., Lei, R., Simha, N., Wang, W., Wilfong, K., Williamson, T. and Yilmaz, S. : June. Realtime data processing at Facebook. In Proceedings of the 2016 International Conference on Management of Data (pp. 1087-1098). ACM (2016)
[21] S. Sneha, J. Kiran, P. Kaustubh : Performance Analysis of RabbitMQ as a Message Bus. In International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) (pp. 229-234) (2018)
[22] Han, S., Gong, T., Nixon, M., Rotvold, E., Lam, K.Y. and Ramamritham, K. : October. RT-DAP: A real-time data analytics platform for large-scale industrial process monitoring and control. In 2018 IEEE International Conference on Industrial Internet (ICII) (pp. 59-68). IEEE (2018)
[23] John, V. and Liu, X. : A survey of distributed message broker queues. arXiv preprint arXiv:1704.00411 (2017)
[24] Miloslavskaya, N. and Tolstoy, A. : Big data, fast data and data lake concepts. Procedia Computer Science, 88, pp.300-305 (2016)
[25] Meena, S.D. and Meena, M.S.V. : Data lakes-A new data repository for big data analytics workloads. International Journal of Advanced Research in Computer Science, 7(5), pp.65-66 (2016)
[26] Halevy, A., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S. and Whang, S.E. : June. Goods: Organizing google's datasets. In Proceedings of the 2016 International Conference on Management of Data (pp. 795-806). ACM (2016)
[27] Haas, L., Cefkin, M., Kieliszewski, C., Plouffe, W. and Roth, M. : The IBM research accelerated discovery lab. ACM SIGMOD Record, 43(2), pp.41-48 (2014)
[28] Terrizzano, I.G., Schwarz, P.M., Roth, M. and Colino, J.E. : January. Data Wrangling: The Challenging Yourney from the Wild to the Lake. In CIDR. (2015)
[29] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F. : Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), p.9 (2016)
[30] Arif, M. and Mujtaba, G. : A Survey: Data Warehouse Architecture. International Journal of Hybrid Information Technology, 8(5), pp.349-356 (2015)
[31] Abdullah, Z.A. and Obaid, T.A. : Design and implementation of educational data warehouse using OLAP. International Journal of Computer Science and Network-IJCSN, 5(5) (2016)
[32] Gruedl, D., Wieland, T. : A Process Model for the Discovery of Knowledge in Sensor-Based Indoor Climate Data. The European Test and Telemetry Conference (ETTC). pp.293-299 (2018)
[33] Ponsard, C., Touzani, M. and Majchrowski, A. : Combining Process Guidance and Industrial Feedback for Successfully Deploying Big Data Projects. Open Journal of Big Data (OJBD), 3(1), pp.26-41 (2017)

[34] Wiemer, H., Drowatzky, L. and Ihlenfeldt, S. : Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model. Applied Sciences, 9(12), p.2407 (2019)
[35] Alfiah, F., Pandhito, B.W., Sunarni, A.T., Muharam, D. and Matusin, P.R. : Data Mining Systems to Determine Sales Trends and Quantity Forecast Using Association Rule and CRISP-DM Method. International Journal of Engineering and Techniques, 4(1), pp.186-192 (2018)
[36] Böhmová, L. and Chudán, D. : Analyzing social media data for recruiting purposes. Acta Informatica Pragensia, 7(1), pp.4-21 (2018)
[37] Jena, B. : A Review on data visualisation tools Used for Big Data. International Research Journal of Engineering and Technology (IRJET), 4(11), pp.492-495 (2017)
[38] Lukasczyk, J., Liang, X., Luo, W., Ragan, E.D., Middel, A., Bliss, N., White, D., Hagen, H. and Maciejewski, R. : January. A Collaborative Web-Based Environmental Data Visualization and Analysis Framework. In EnvirVis@ EuroVis (pp. 25-29) (2015)
[39] Kolomeec, M., Chechulin, A., Pronoza, A. and Kotenko, I.V. : Technique of Data Visualization: Example of Network Topology Display for Security Monitoring. JoWUA, 7(1), pp.58-78 (2016)
[40] Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J. and Möller, T. : Visualization as seen through its research paper keywords. IEEE Transactions on Visualization and Computer Graphics, 23(1), pp.771-780 (2016)
[41] Luo, Y., Qin, X., Tang, N. and Li, G. : April. DeepEye: Towards Automatic Data Visualization. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 101-112). IEEE (2018)
[42] Moreno, J., Serrano, M.A., Fernandez-Medina, E. and Fernandez, E.B. : Towards a Security Reference Architecture for Big Data. In DOLAP (2018)
[43] Holátová, D.D. and Monika, B. : Basic characteristics of small and medium-sized enterprises in terms of their goals. International Journal of Business and Social Science, 4(15) (2013)
[44] Berisha, G. and Pula, J.S. : Defining Small and Medium Enterprises: a critical review. Academic Journal of Business, Administration, Law and Social Sciences, 1(1), pp.17-28 (2015)
[45] Sen, D., Ozturk, M. and Vayvay, O. : An overview of big data for growth in SMEs. Procedia-Social and Behavioral Sciences, 235, pp.159-167 (2016)