PalArch's Journal of Archaeology of Egypt / Egyptology

A REVIEW ON OIL AND GAS ORGANIZATION'S DATA LAKES IMPLEMENTATION BEST PRACTICES

Mohd Hilmi Hasan[1], Azlinda Abdul Malik[2], Ariana Yunita[3], Ade Irawan[4], Meredita Susanty[5]

[1]Center for Research in Data Science, Computer & Information Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar, Perak, Malaysia.

[2]Petroleum Engineering Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia

[3,4,5]Computer Science Department, Universitas Pertamina, Jl. Teuku Nyak Arief, Kebayoran Lama, South Jakarta, Indonesia

E-mail: [1]mhilmi_hasan@utp.edu.my

Mohd Hilmi Hasan, Azlinda Abdul Malik, Ariana Yunita, Ade Irawan, Meredita Susanty. A Review On Oil And Gas Organization's Data Lakes Implementation Best Practices-- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(10), 967-977. ISSN 1567-214x

Key Words: About Four Key Words Or Phrases In Alphabetical Order, Separated By Commas. Best Practices, Data Lakes, Data Lakes Best Practices, Oil And Gas Data Lakes

ABSTRACT

Business organizations have leveraged on data analytics since 1950s, but in recent years, the trend has been increasing mainly due to the availability of fast and efficient computation nowadays. This has encouraged the emergence of big data analytics. One of the main components in big data analytics is the ability of the framework/architecture to manage scalable data and different types of data i.e. whether it is structured, semi-structured or unstructured date. This has become the reason for the introduction of data lakes; to support massive scalable data storage, which can hold the three types of data mentioned above. However, despites their significant benefits, many have found that data lakes eventually turned into data swamps. This is because data lakes are normally implemented without consideration on the necessary fundamentals to operationalize the generated insights. Therefore, we initiated a research to investigate the data lakes implementation best practices. The case study for this research is oil and gas organization. In this paper, we present a review on data lakes implementation best practices for oil and gas organizations. The review will support our research in developing a guideline containing best practices of data lake implementation to support oil and gas big data initiative.

INTRODUCTION

Decades ago, even in 1950s, businesses had used data analytics to uncover insights and trends. However, that was done essentially based on manual examination upon data in spreadsheets. A few years ago, this kind of data analytics became mainstream again but with the new benefits of speed and efficiency. The main technology that contributes to this new era of big data analytics is the powerful processing tools that help businesses identify insights for immediate decisions. This ability makes organizations work faster as well as gives them a competitive edge that they did not have before. More and more organizations now understand the importance of capturing data, streaming them into businesses, applying analytics and gaining significant values from it [1].

Big data management involves processes like data collection, storage, and processing of datasets of size and structure that are beyond the capabilities of traditional IT tools such as database, software etc. [2]. Moreover, those data can be structured, semi-structured and unstructured which cannot effectively be managed and processed using traditional tools. Furthermore, in the current world, organizations' data grow rapidly, which require scalable repository that can be used to store those data until there is a reason to mine them [3].

These new characteristics of data and its management require a new technology so that the advantages of big data analytics can be benefited. A few years ago, data lakes were introduced that refers to a massively scalable storage space. It can hold vast amount of data, in its native format, until it is needed for processing [4, 5]. Data lakes can also store structured, semi-structured and unstructured data.

However, despite their noteworthy benefits in supporting the emerging big data analytics-driven innovations, data lakes are generally perceived as ineffective [6 – 9]. The main reason is that companies adopting big data analytics initiatives generally begin data lakes implementation without the fundamentals necessary to operationalize the generated insights. Data lakes implementation frequently lack forethought on what they are supposed to achieve. Gartner listed three causes for the failure of data lakes; when they lack governance, self-disciplined users, and a rational data flow. These culprits will turn data lakes into data swamps. An overview of the whole processes of data lakes and their descriptions/purposes (match with organization's goals) must be in place prior to implementation to ensure data lakes employment on track. The processes are data acquisition/ingestion, insight discovery and development, optimization and governance, and analytics consumption [10]. These processes are shown in Figure 1.

Hence, our research will answer the following research questions:

- What are the tasks required for data acquisition/ingestion process and their best practices?
- What are the best practices of insight discovery and development for data lakes?

- What are the tasks required for optimization and governance of insights/analytics process and their best practices?
- Who are the audience and their profiling information that will consume the insight/analytics?

All these four research questions are critical to be investigated and answered so that a guideline for data lakes implementation best practices can be produced.

Oil and gas industry cannot be exempted from this big data analytics initiative. Based on our initial engagement with some organizations, data lakes have become one of the components contained in their information management architecture. Therefore, our research aims at investigating the best practices for data lakes implementation and will produce a guideline that contains the overall required tasks/processes that must be planned prior to data lakes implementation. The produced guideline will be mapped with oil and gas big data management.

The objective of this paper is to present a review on data lakes implementation best practices for oil and gas organizations. The paper is structured as follows; next section will present related works, section 3 will contain proposed methodology to carry out this research, and finally the paper will end with conclusion section that summarizes the outcomes from this review as well as outlines future works.

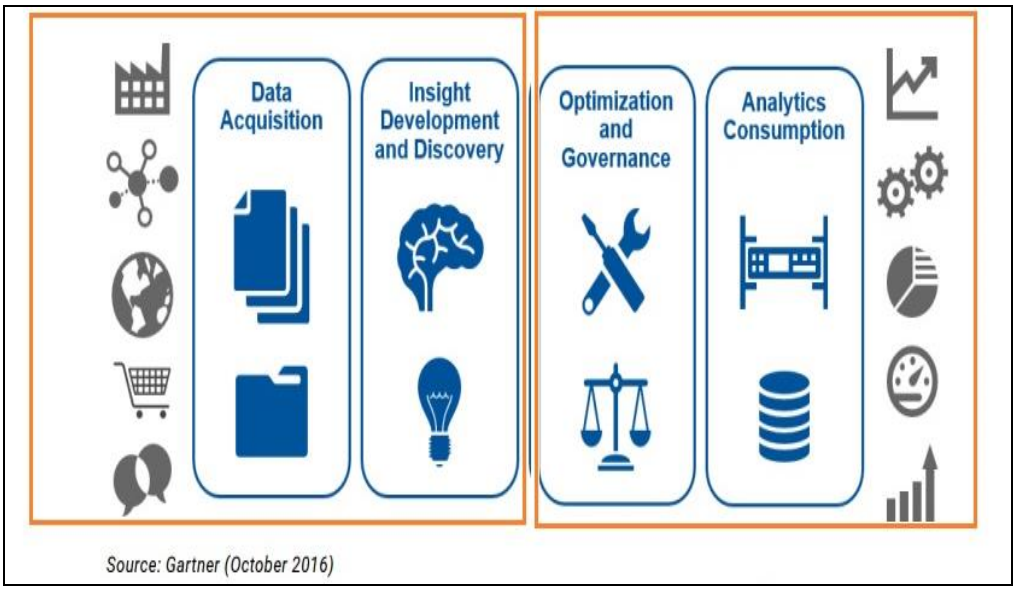

Figure 1: Processes involved in data lakes

LITERATURE REVIEW

Ref. [11] claimed that data lakes are the emerging technology for the development of the next generation of systems for handling big data. Organizations are gearing towards data lakes so that they have the capabilities to manage and use data with increased volume, variety and velocity. Ref. [12] stated that data lakes apply a flat architecture to store data in raw format. Each

data is assigned with a unique ID and a set of extended metadata, while users may propose their schemas to query relevant data. This allows analysis of smaller set of data to help answering consumers' questions. Another work also proposed the use of data lakes to support the massive growth of data in big data analytics era. By having data lakes, this gigantic amount of data may be put in repositories first in its raw format, which will then be ingested later for insights [13].

In term of data type, data warehouse is; structured, processed; while data lake can be structured, semi-structured, unstructured or raw. Then, in data warehouse, only selected data will be stored; mainly due to expensive storage space and scalability; in Data Lake, all data will be stored. Furthermore, data warehouse has slower speed while Data Lake has faster. In term of storage, data warehouse is costly because it needs large data volume while Data Lake is designed for low cost. Regarding agility, data warehouse is less agile, and it has fixed configuration; Data Lake is highly agile, configured and reconfigured as needed. Lastly, for processing, data warehouse has schema-on-write; Data Lake has schema-on-read.

The above works showed that data lakes are significant components that support the big data analytics initiative. Data lakes provide a place for all data to be stored and dormant until someone or something needs it. This gives benefits in terms of flexibility for users to use specific data types/sources they need, how much they need, when they need those data, and what type of analytics that the need to derive. Their characteristics can close the gaps that exist due to deficiencies of traditional data warehouse technology. Table 1 summarizes the key differences between data lakes and data warehouse [14 – 17].

Data acquisition

Today, data can be found everywhere, inside the organization, in cloud-based applications, public sources, and open data sources that anyone can access, in various format, structured, unstructured, and semi-structured. Most of this data never makes it into the data warehouse, nor should it, because all data must go through some processes before it can be made available. Thus, a key reason to have a data lake: making data available without doing unnecessary work. The acquisition layer is expected to be flexible to accommodate a variety of schema specifications and at the same time, it must have a fast connect mechanism to seamlessly push all the translated data messages into the data lake [18]. In data lake, acquiring data should be independent from data management to avoid premature data standardization and modeling [19]. The separation allows data collection; data processing takes place in different time, as well as permitting different processes to use the data. Immutable data store for raw data is suggested. Using immutable layer data can be collected at any latency, from real-time streams to bulk file loads, original data can be retained. Data processing such as reprocessing derived data, and extra work a relational database does can be performed after writing the data. The investigation on data acquisition matters are covered under research question (1).

Insights discovery and development for Data Lake

Insights discovery and development for Data Lake is a fundamental process from data lake implementation which can be done by data scientists, often working in a team. After data has been acquired, data scientists will apply various techniques to analyze the data which may include simulation, text analysis, machine learning approaches and graph analysis [20]. The discovery and development of new analytical models includes two items: data interrogation, model discovery and analytic development. The study on these matters is covered under research question (2).

Insights' optimization, governance and audience profiling

After the discovery and development of insights produce significant new analytical questions, all of the inputs and outputs involved in those insights must be optimized and governed. Without optimization and governance, data lakes may lead to underutilization, if not failure. Among the considerations that need to be tackled in this process are how data are integrated and visualized/executed in an optimized way. Furthermore, this process also optimizes data integrity and quality, hence promotes continuous improvement of the whole big data analytics initiative. Therefore, the understanding on the whole governance and optimization plus their planning are vital due to the ever-increasing volumes of data nowadays [21 – 23]. The investigations on these matters are covered under the research question (3).

Then, after the analytical models and data are optimized, the workflow should be clear enough to define the most relevant audience for those models and data. Different audience expects different kinds of capabilities and presentations. Among the considerations that must be tackled are the number of concurrent users that will utilize the insights and the business insight platforms/tools that they will use. Moreover, good user profiling will help customers in the navigation and search for the data according to their business needs [24, 25]. This kind of data profiling and cataloging will also enhance customer experience. The studies on these matters are covered under research question (4).

Table 1: Comparison between data warehouse and data lakes

Criteria	Data Warehouse	Data Lakes
Data	Structured, processed	Structured, semi-structured, unstructured, raw
Data	Only selected data will be stored; mainly due to expensive storage space and scalability	Store all data
Speed for insights	Slower	Faster
Processing	Schema-on-write	Schema-on-read
Storage	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, configure and reconfigure as needed

METHODOLOGY

Figure 2 shows the methodology that will be carried out in our research. The first phase will be beginning with structured literature review. The objective of this phase is to investigate and determine the best practices of data lakes implementation in terms of data acquisition/ingestion, insights discovery and development, optimization and governance, and analytics consumption/user profiling. The detail activities in this phase are shown in Figure 3, and explained as the following:

Framing questions for a review

The main objective of this phase is to firm up the problems to be addressed in this research. The activity involved is defining problems into set of questions in the form of clear, unambiguous, and structured.

Identifying relevant works

This phase will involve extensive work on searching for relevant works. The search criteria are relaxed, for example the relevant works are searched among multiple resources (digital or printed) and no language restrictions. The search will be based on the questions identified in previous phase. Any reasons for inclusion/exclusion of works will be recorded.

Assessing the quality of studies

The output of the second phase will be a huge amount of information/data/references of relevant works. In this phase, those works will be further refined and filtered based on stricter search criteria. This phase will begin with defining these new levels of filtering criteria so that it can be used to determine whether or not any works will be included.

Summarizing the evidence

The collected data/information/references from the selected works will be reviewed, synthesized and analyzed. These activities will be carried out based on those works' characteristics, quality and effects. Statistical method may be used to explore relationship among findings.

Interpreting the findings

In this phase all issues highlighted in the previous four phases will be consolidated and solved. The results of synthesis and analysis from previous phase will be converted into a guideline for effective implementation of data lakes.

As mentioned above, the outcome of this structured literature review phase is a guideline for data lakes implementation best practices.

The next phase involves case study preparation activities. The objective of this phase is to prepare the case and interview-related items to verify the guideline identified through structured literature review. This phase will begin with selecting the case study. Since this research is a collaboration work between two universities, we plan to have one case study each for an oil and gas organization in Malaysia and Indonesia. The specific organization, as well as the department for case study will be identified in this phase. Then we will prepare interview instruments and tools. This activity will involve among others, generating interview questions as well as verifying the questions. Lastly, the respondents for each case will be identified.

The next phase will be data collection. This phase will be divided into four cases namely data acquisition/ingestion, insights discovery and development, optimization and governance, and analytics consumption/user profiling.

Then, it will be continued with the data analysis phase whereby all outcomes from the data collection cases will be analyzed. The analysis will involve within each of the cases as well as cross cases analysis. The outcomes from this phase will be compared against the data lakes implementation best practices identified earlier. The final outcome will be a verified guideline for data lakes implementation best practices for oil and gas organizations.

CONCLUSION

Despites having benefits to organizations, data lakes face a problem of ineffective implementation, which mainly because of improper planning and descriptions/purposes prior to its implementation. This has caused many of data lakes implementation become data swamps. Hence, in this research we will investigate the best practices for data lakes implementation. Our case study will be oil and gas organizations.

Since this is an initial part of our research, this paper presents a preliminary review on the best practices' implementation of data lakes in oil and gas

organizations. We identified four components of data lakes implementation that will become our case study. They are data acquisition/ingestion, insights discovery and development, optimization and governance, and analytics consumption/user profiling. This paper also presents the methodology of research that will be carried out in our research.

For future work, we will begin conducting our research by carrying out all the stated phases and activities in the research methodology.

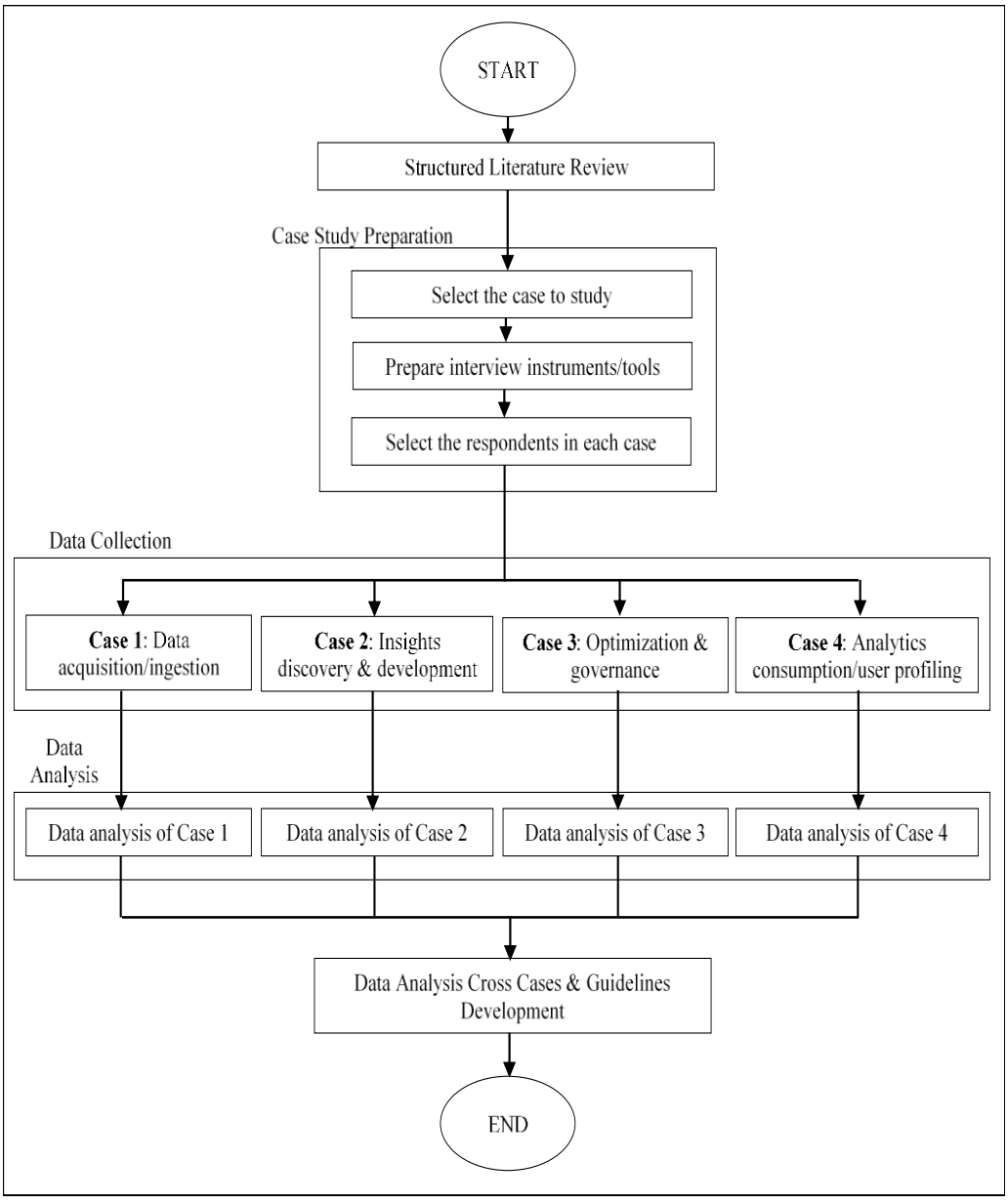

Figure 2: Methodology of research

ACTIVITY	OBJECTIVE	DELIVERABLE
Framing the question	Firm up the problems to be addressed Gather information from experts	A clear, unambiguous and structured questions to be answered in this research
Identifying relevant works	Extensive search for previous studies Record reasons for inclusion/exclusion of studies	List of relevant works
Assessing the quality of studies	Set the minimum acceptable level for inclusion of previous studies Filter the collected previous studies using the minimum acceptable level	Refined list of relevant works
Summarizing the evidence	Review, synthesize and analyze the collected data	Raw findings of guideline for effective data lake implementation
Interpreting the findings	Consolidate and convert findings into guideline for effective data lake implementation	A guideline for effective data lake implementation

Figure 3: Structured literature review activities, objectives, and deliverables

AKNOWLEDGMENT

This research is an ongoing research supported by Yayasan UTP (L0029); a grant funded by Universiti Teknologi PETRONAS, Malaysia.

REFERENCES

A. Dhankhar, K. Solanki. A comprehensive review of tools and techniques for big data analytics, International Journal of emerging Trends in Engineering Research, 7(11), pp. 556-562, 2019.

N. Miloslavskaya, A. Tolstoy. Big Data, Fast Data and Data Lake Concepts, 7th Annual International Conference on Biologically Inspired Cognitive Architecture. 2016.

N. Shalom. The next big thing in big data: fast data, https://venturebeat.com/2014/06/25/the-next-big-disruption-in-big-data/ . 2014.

Laskowski. Data lake governance: A big data do or die, http://searchcio. techtarget.com/feature/ Data-lake-governance-A-big-data-do-or-die. 2016.

J. Kachaoui, A. Belangour. From single architectural design to a reference conceptual meta-model: an intelligent data lake for new data insights, International Journal of emerging Trends in Engineering Research, 8(4), pp. 1460-1465, 2020.

S. Early. Data virtualization and digital agility, IT Professional, 18 (5). 2016.

C. Stamford. Gartner says beware of the data lake fallacy, http://www.gartner.com/newsroom/id/2809117. 2014.

S. Brobst. Is your data lake destined to be useless?, https://www.forbes.com/sites/teradata/2016/07/01/is-your-data-lake-destined-to-be-useless/#cf4f4f6f40c7. 2016.

Teradata. How to stop a data lake turning into a data swamp, http://www.technologydecisions.com.au/content/it-management/article/how-to-stop-a-data-lake-turning-into-a-data-swamp-1087785566#axzz4tjt5iwTe. 2014.

Gartner. https://www.gartner.com/doc/3483017/best-practices-designing-data-lake. 2016.

H. Fang. Managing data lakes in big data era – What's a data lake and why has it became popular in management ecosystem, 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems. 2015.

H. Alrehamy, C. Walker C. Personal data lake with data gravity pull, IEEE 5th International Conference on Big Data and Cloud Computing, 2015.

A. Alserafi, A. Abello, O. Romero, T. Calders. Towards information profiling: Data lake metadata management, IEEE 16th International Conference on Data Mining Workshops, 2016.

M. Knight. Data warehouse vs. data lake technology: different approaches to managing data, http://www.dataversity.net/data-warehouse-vs-data-lake-technology-different-approaches-managing-data/. 2017.

C. Campbell. Top five differences between data lakes and data warehouses, https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses. 2015.

P. Simon. Data lake and data warehouse – know the difference, https://www.sas.com/en_us/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html. Access 2017.

KDnuggets. Data lakes vs data warehouse: key differences, http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html. 2015.

M. Madsen. How to Build an Enterprise Data Lake: Important Considerations Before Jumping In, Available at: http://resources.idgenterprise.com/original/AST-0163853_how-to-build-an-enterprise-data-lake.pdf.

T. John, P. Misra. Data Lake for enterprises. Birmingham: Packt Publishing Ltd. 2017.

N. Heudecker. Best Practices For Desiginign Your Data Lake, https://www.gartner.com/doc/3483017/best-practices-designing-data-lake. 2017.

A. Erraissi, A. Belangour. Meta-modeling of big data management layer, International Journal of emerging Trends in Engineering Research, 7(7), pp. 36-43, 2019.

T. King. The Growing Importance of Data Governance, https://solutionsreview.com/data-integration/the-growing-importance-of-data-governance/. 2016.

R. Marvin. Big Data Basics: How to Build a Data Governance Plan, https://www.pcmag.com/article/ 347785/big-data-basics-how-to-build-a-data-governance-plan. 2016.

Knowledgent White Paper. How to design a successful data lake, https://knowledgent.com/ whitepaper/ design-successful-data-lake/. 2014.

F.Z. Bouseba, H. Zeghdoudi, G. Amine. On variance and volatility swaps in oil markets, journal of computer science & computational mathematics, 7(2), 2017.