Roadmap For Building Big Data Lake From ETL Perspective

Niha Mohanty, Sneha Videkar
Data and Knowledge Engineering
Otto Von Guericke University
Magdeburg, Germany
Email: niha.mohanty@st.ovgu.de, sneha.videkar@st.ovgu.de

***Abstract*—The Data lake has been the new buzzword in all the enterprises.The Data lake is considered an optimistic approach for managing large volumes of structured and unstructured data, known as Big Data. Companies are ready to dive in the world of Big Data and Data Lake. On the other hand, they fail to implement Data lake platform along with the correct ETL tools due to the lack of research strategy. This research paper will help gain knowledge of Data lake and will provide guidelines for building Data lake from an ETL perspective.**

***Index Terms*—Big data, Data lake, Data warehouse, ETL**

I. Introduction

Nowadays, we live in a world which constantly generates digital data. Financial and insurance organization has always been data-driven and oriented. Data also plays a vital role in the government sector, education, and manufacturing [1] [2]. For many of us, we use applications, websites to fulfill our daily needs. Every click and touch on the internet is viewed and analyzed to make our work easier by providing intelligent assistance. In other terms, applications such as financial, insurance applications were developed as per business users requirement. Latest applications such as Google map, for example, assist in the decision and improvise on the day to day activity. To summarize, the world is revolving around the data due to the potentials it contains [3]. Today new features like variety, volume, and velocity are essential aspects of the data. Strategy to hold such *Big Data* and extract real meaning from it has become the need of customer-oriented business.

One of the famous strategies was *Data warehouse* which could store, process and manage large volumes of structural data. With the new need for analyzing *Big data*, *Data lake* has been a promising way to handle both structured and unstructured data. Big data technologies are not only focused on the volume but also emphasize the real meaning of the data. In the customer-oriented business era, constant analysis of the behavior of customers plays an important role which supports the growth of the organization. Because of all these new needs of big data technologies, many organizations have either developed a Data lake or are in the process of building a Data lake.

However, Data lake architecture is complex and composed of different layers [4]. Various Big data technologies are integrated to process various forms of data, to extract the meaning of the data, to find patterns within the data and lastly also to find hidden values which could be necessary for the business. Though Data lake is capable of storing a large volume of data but it should not result in the storage of unnecessary pile of data. Additionally, this large amount of data should be properly cataloged so that it can be useful in the future. If no proper data cataloging is performed, it will be difficult to track and manage the data. Lastly, important data could be ignored or missed during an Extract, Transform and Load (ETL) process. Ignacio Terrizzano et al. have also mentioned that without finding proper meaning of the data it will be hard to interpret the statistical data [5]. Due to all these issues, many companies fail to setup the proper Data lake platform. Research performed by Ignacio Terrizzano et al. was inclined on the incoming data purification but have not focused on guidelines which will help Big data industry to build robust Data Lake.

Hence, lack of technological knowledge and guidelines to setup Data lake is impacting the goal of establishing efficient Big data lake platform. In our research, we have focused on formulating guidelines to build Big Data lake reasonable from ETL perspective. So basically our research will revolve around this question: *Does guidelines for building Big Data Lakes from ETL perspective will be beneficial from an implementation point of view?*

The remainder of this paper is structured as follows. In the next Section II, we give a brief idea about background topics such as Big data, Data warehouse, ETL (Extract, Transform and Load) process and Data lake. Additionally, we will also discuss Data warehouse guidelines which will be state-of-the-art for this research. A particular research methodology has been followed to formulate data lake guidelines. Section III covers details about this research methodology. Section IV is the core of this research. Here we propose the set of rules for building Data lake from ETL perspective. Finally, we summarize the research topic in conclusion Section V.

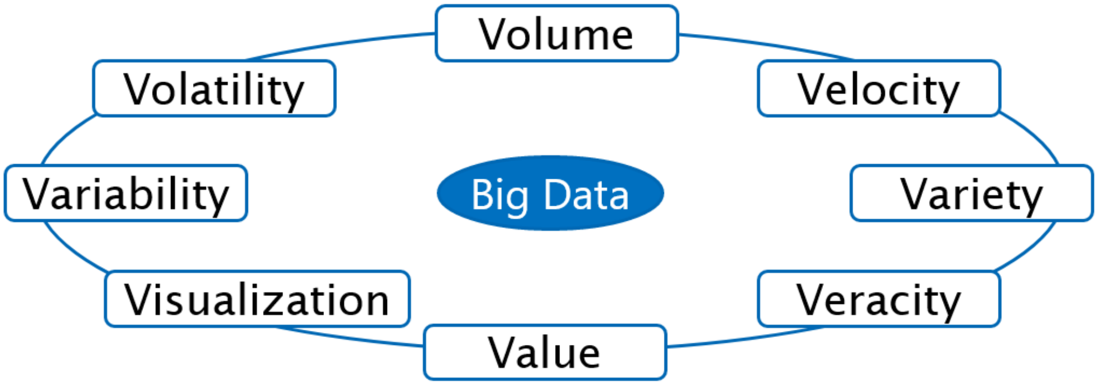


Fig. 1. Big Data Characteristics *

II. State Of The Art

In this section, we provide a brief background about key terms which forms the base and state-of-the-art of the research topic.

A. Background

The focus of this section is to give insights about the key terms like Big Data, Data Warehouse, and Data Lakes along with the ETL process.

1) Big Data: Big enterprises need an overview of the data used in the enterprise to make a better decision and achieve high growth in operations [6]. Nowadays, data is generated in large volumes by various sources at a tremendous rate [7]. Typically, this is called as *Big Data*.

This so-called Big Data is differentiated from traditional data by one or more of the *V*'s and its definition is strongly grounded to the *V*'s. These *V*'s are represented in Fig 1. To get a clear idea on Big Data [8] let's try to understand this *V*'s in-depth.

- *VOLUME: Volume* refers to the growing quantity of data that is generated and stored from various sources. These data are considered to be very big and their size ranges from few *Kilobytes* to *Petabytes* [9]. This extremely large volume of data is one of the major characteristics of Big Data and it must be understood to make decisions which are based on data base.

  * *For example:* The volume range varies because a text file is of a few *Kilobytes* whereas a sound file is of few *Megabytes* and a full-length movie is of few *Gigabyte* and so on.

- *VELOCITY: Velocity* measures the speed of data processing i.e. the speed at which new data is being generated or flows in from different sources and as well as how quickly they are being accessed for further processing and analysis. This flow of data is massive and uninterrupted. This extremely high velocity is another major characteristic of Big Data [10].

  * *For example:* Suppose, 80 hours of video is uploaded to Youtube every minute, indicates the velocity. Another major source of high velocity of data is social media.

- *VARIETY: Variety* refers to the range of different data sources and types. These data can be either structured, unstructured or semi-structured, media, audio, videos, etc. These different varieties of data sources are nowadays considered for the analysis applications [11].

  The *structured* data sources indicate the data that mostly have a definitive structure like, call detail records in a telecom company. The *unstructured* data sources indicate those that are sourced from websites like product reviews on twitter. The *semi-structured* data sources mainly indicate the graph data.

  Apart from the above-mentioned data sources, the inflow also occurs from machines, sensors, GPS signals from cell phones, and other sources, making it difficult and complex to manage and maintain. Hence, with regards to unstructured data there persist some issues related to storage, mining and analyzing. Hence, there is a necessity of integrating these data. So, *Variety* of data is the biggest challenge to use effectively from an analytic perspective.

- *VERACITY: Veracity* refers to the biases, noise, and abnormality in the data which are due to the inconsistency or incompleteness in the data. In data analysis, this is the biggest challenge. These data need to be cleaned before being integrated, i.e. the management of the reliability

* Source: https://www.vapulus.com/en/five-characteristics-of-big-data/

and predictability of intrinsic inaccurate data, like testing a different hypothesis, training vast samples, etc. [12]

- *VALUE: Value* refers to the data that are worthwhile and have importance for business [13].

  * *For example:* Cost effective and cheaper storage can be unreliable, and thus can cause a risk.

- *VISUALIZATION: Visualization* mostly is related to the visual representation and insights for decision making. Here, any type of data is represented in a graphical format which will make it easy to understand and interpret [14].

- *VARIABILITY: Variability* refers to the inconsistency in data that can be found at times which hampers the process of efficient handling and managing the data. Moreover, the data that are available sometimes may be untidy or not trustworthy. The quality and accuracy are also difficult to control because of the generation of a wide variety of big data types [15].

  * *For example:* A Twitter post has hashtags, typos and abbreviations.

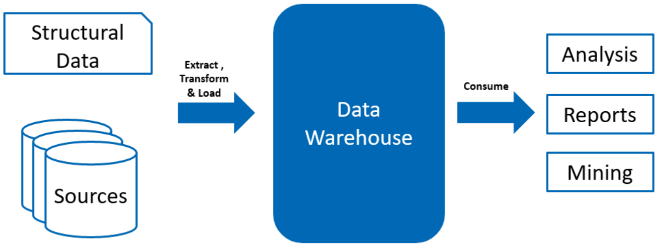
- *VOLATILITY: Volatility* represents the fact that how long is the data valid and how long it should be stored which indicates the point where the data becomes irrelevant for the current analysis [16].

  * *For example:* One year old customer purchase history may not be relevant for an online e-commerce company.

Big data needs should be processed efficiently, correctly and timely such that useful information is available for the business analysts.

2) Data Warehouse- DW: Business needs cumulative data for making the strategic decision [17]. Data warehouse holds integrated and frequently updated data for taking strategic decisions which leads to growth of the enterprise. Transnational processing data is normally stored in the normalized format this means data is scattered across different tables which makes operational transaction query processing faster. However, for analytical reports generation, a combined view of a few sets of tables is required. Such a report generation from normalized tables takes a longer time because fetching a huge amount of data from various tables needs complex join operation. Till now Data warehouse is the only leading storage repository to handle analytic data where data is stored in de-normalized way. Data warehouse mainly focus on 4 W's : who, what, when and where [18].

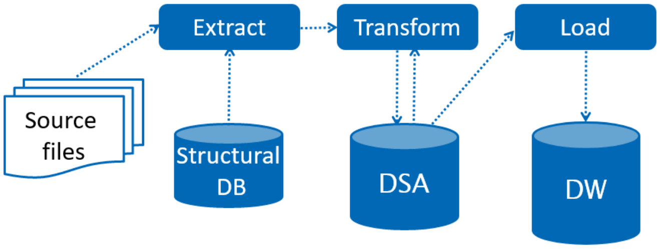
As shown in the Fig. 2, structural data is extracted, transformed to conform to data warehouse schema and loaded in to Data warehouse. This is performed with help of *Extract, Transform and Load - ETL* process.



Source: Referred from [18]

Fig. 2. ETL Process Of Data Warehouse

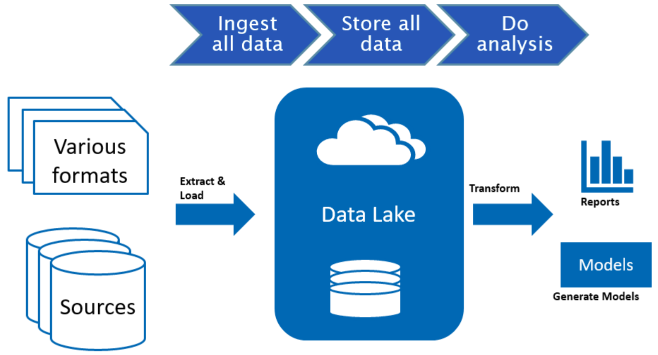
3) Extract Transform and Load: The main goal of *Extract, Transform and Load(ETL)* process is to extract data from various sources, to clean, and transform the input data, and finally load the data in data storage for report generation. General *ETL* process is shown in the Fig. 3. Data is extracted from different sources such as OLTP databases, text files or, spreadsheets. This data is transformed and cleansed in staging layer and then finally data is loaded in *Data warehouse*. ETL process is on going process which handles periodically updates of data and frequent requirement changes for full filling business needs [19]. In our research, we have analyzed ETL processes for *Data lake* in depth.



Source: Reprinted from [19]

Fig. 3. ETL Process

4) Data Lake: A massive repository is required in order to handle *Big Data*. *Data lake* serves the purpose of storage repository and handles raw data on a large scale in its original format until it is required by end-user/business analyst to improve profit of the Business [20]. Fig. 4 gives a brief idea about the data lake in general. Data in various formats like structured, semi-structured or unstructured will be read from different sources like a relational database, mainframe systems, and social media sites [11]. This data is loaded in the *Data lake* in its native format [21]. Moreover, *Data lake* has a processing engine which ingests the data along with its source data structure. Once the data is ingested in the *Data lake* it is available to be used by end-user. Transformation of the data in *Data lake* is performed once end user requests for a particular report or needs to generate a model for analysis purpose [18].



Source: Reprinted from [18]

Fig. 4. ETL Process Of Data Lake

B. Data Warehouse Guidelines

Kimball et al. have discussed various use cases and formalized guidelines for the data warehouse implementation [22]. For our research, we have considered these guidelines as state-of-the-art for our research. Data warehouse guidelines are as follows:

1) Understanding The Requirements: The first guidelines focus on the need for the Data warehouse from a business perspective. Before starting the implementation, user requirement should be gathered. All the key requirements are gathered and if there are any additional suggestions, this should be discussed with business. Additional suggestions should also be stated and noted if agreed by the customer. Data compliance is another important factor which should be maintained in the initial phase. In compliance, a legal agreement is made with the customer for input data and steps which will be carried to transform the data. The compliance can be used for verification of transformed data. Secondly, data quality should be agreed with the customer if there is compromised made for any of the data.

2) Extract The Data: Next, the extraction of the data should be performed. Input data sources should be identified. Data profiling is important activity under data extraction in which suitability of the identified input data is checked. It is always better to finalize the candidate input data sources in the early stage. Getting rid of unwanted data sources during implementation would cause a delay in the data warehouse implementation. At the end of data profiling, realistic development schedule can be set for the implementation phase, limitation on the data sources, and best data source capture practices can be identified upfront.

Furthermore, data in data sources would be evolving frequently due to transactions performed on it. Refreshing entire data warehouse tables are not feasible hence transferring only the relevant updated data is the way to handle refresh data. Hence, change data capture should be structured to handle future upcoming data.

3) Clean And Conform: The most important activity of the ETL process is cleaning and conforming data. This is the core step where transnational data will be transformed to make value for the organization. In a Data warehouse, data will be read from multiple input data sources which might have a different schema from each other. In cleaning and conforming phase, the uniform schema has to be prepared to handle the data from various sources. Invalid data or duplicate data should be handled before it is fed to the Data warehouse system. Now once the common structure is finalized, this can be then used in the delivery phase.

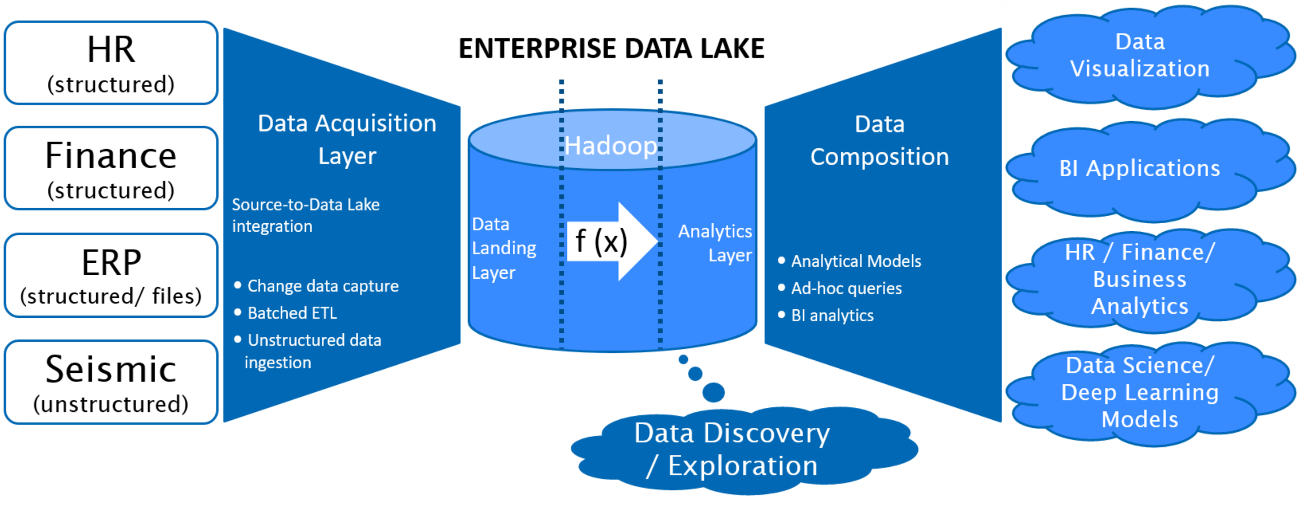
4) Delivering - Prepare For Presentation: The ultimate goal of the ETL process for data warehouse will provide the data from different sources with different schema in one combined schema format which will be used for report generation. Extracted, cleaned and conformed data should be stored in a de-normalized way in intermediary tables. In Data warehouse, such intermediary tables are called as dimension tables and facts tables. These tables will be populated with data from various sources. Data from these tables is then consumed during analytical report generation. These tables which are created should be scalable and maintainable in order to handle upcoming new data.

5) Manage The ETL Process: Further, the entire process should be managed continuously. There are various tasks which should be considered under management of ETL. First one is scheduling of some jobs which will take care of upcoming new data, logging the updates and issues for recovery purpose. In case if there is a failure in any load there should be a way to notify concerned teams. Moreover, Data warehouse stores a large volume of data and are subjected to face failures. To handle failures regular backups of the data warehouse should be taken. So even after failure, Data warehouse can be reconstructed to continue working as before.

These above mentioned five guidelines are discussed by Kimball et al. in detailed [22]. We refer above mentioned Data warehouse guidelines and background of Data Lake to construct guidelines required for building a successful Data lake in the next section.

III. Methodology

Main aim of this research to formulate guidelines for building Data lake. In order to achieve mentioned goal, detailed architecture of the Data lake is analyzed in depth. Sourabh Gupta and Venkata giri have given detailed architecture of data lake in their book [23]. This detailed architecture of data lake will help us understand input data, storage features, processing features, and output data. As per Fig. 5, the Data lake is partitioned into a data landing(or mirror) layer and an analytical layer. Mirror/landing layer



Source: Reprinted from [23]

Fig. 5. Data Lake Architecture

contains the raw data. Analytical layer makes sure the data is ready for consumption by a business analyst or data scientist. Major pillar for data lake: Storage is handled with the help of Hadoop storage and processing framework using hive.

As discussed in the state-of-the-art, predefined Data warehouse guidelines are currently in use by the industry. These guidelines help us in our research topic. Also, various questionnaires helped us to gather all the need for Data lake implementation. List of question is as follows:

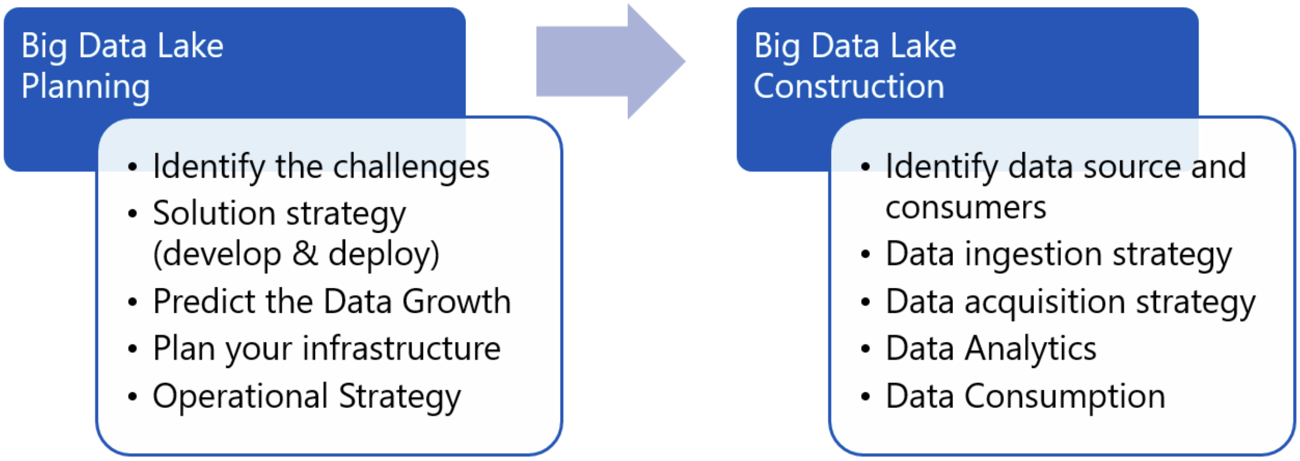
Question: What is the basic need?
A data lake can be built for various application scenarios. IBM research [5] has concluded that research scientist need readily available collection of large volume of contextual data for innovation purpose. Hence we need to understand if there is any specific business requirement before building Big data lake. Meticulous study should be conducted before starting the implementation. Discussions should be conducted with business domain experts and Data scientist to understand the problem statement and need of Data lake in depth.

Question: Which all data sources will be read and used?
A data lake is supposed to handle big data which can be structured, unstructured, and semi-structured data. In advance, all the data types which might be read should be noted. The complexity of the incoming data is another main factor. Approximate size of data which will be fed to Data lake system should be analyzed in advance. Data lake should be scalable and capable to handle frequently growing data. Lastly, the frequency of the data is also significantly important. Time intervals for recent data update should be examined.

Question: How cleaning and transforming of the big data can be performed?
Source data system which will feed in the data to Data lake system is expected to have redundant, and irrelevant data. In other words, cleaning is a vital part when multiple source systems are read during the ETL process. Additionally, incoming data could vary in terms of the internal schema. From ETL perspective, the transformation will help to sync the schema for various systems to find the meaning of the data contained in one format. One of the biggest advantages of Data lake is it can store the data in its native format. So handling of source data in its raw format should also be planned before implementation. Later on, this cleaned raw data would need transformation as per Data lake analytical layer.

Question: How data can be made available for the use by a business analyst or Data Scientist?
Business analyst and Data scientist will be at the consumer's end for the Data lake system. These are the main end-users of the Data Lake in which they will query the data frequently or periodically depending upon the need. Data extracted from the various system should be loaded into the Data lake in such a way that both business analyst and Data scientist can view the data in the appropriate format required for taking strategic decision. Also, a decision should be taken for which and what data will be useful for the end-users. In other words, data cataloging should also be focused.

Question: What to consider for management of big data inside Big data lake?
Finally, along with the basics of the ETL process, there is a need for maintenance of the entire process of the building Data lake. One has to imagine the amount of the data which could fit in the Data lake, but there must also be attention paid to archival of the used data and easy recovery strategy in case of system failure. Initial transformation and load


Big Data Lake Planning
• Identify the challenges
• Solution strategy (develop & deploy)
• Predict the Data Growth
• Plan your infrastructure
• Operational Strategy
Big Data Lake Construction
• Identify data source and consumers
• Data ingestion strategy
• Data acquisition strategy
• Data Analytics
• Data Consumption

Source: Reprinted from [23]

Fig. 6. Big Data Lake Guidelines

performed for the first feed in the Data lake must not be repeated in case of failure however efficient data recovery should be used to restore the data. Lastly, data security also plays an important role in the maintenance of the Data lake.

Methodology can be derived once we have found solutions to above mentioned questions. The next section is the core of our research in which define actual rules for Data lake implementation.

IV. Assessment

By taking into consideration the data warehouse guidelines from ETL perspective and the facts present in book [23], we tried to formulate the similar guidelines for Big Data Lakes. The guidelines are basically divided into two broad phases as represented in Fig. 6 and each phase has 5 different processes where each one of them has their own specialization towards building the guidelines. To get a detailed insight into these guidelines, let's try to understand how and for what reasons these phases work.

A. Big Data Lake Planning

As shown in Fig. 7, the first phase is the pre-processing step which is referred to as *planning phase*. In this step, 5 different processes are involved which ensures that a specific requirement of the organization is addressed in detail. The planning is required to make any process workflow work smoothly. Hence it is the foremost step in building Big Data Lakes.

In this phase, first the requirement of an organization is analyzed for its validity. If it is valid then the second rule, i.e. optimal solution strategy is created for the requirement. This is followed by data growth prediction, infrastructure planning and finally the operational strategies. The subsections mentioned below focus on these rules to give us a detailed insight.

1) Identify The Challenges: The first rule of the planning phase focuses on identifying or figuring out the business requirement, i.e. the issues that needs to be focused on is first determined. It is the part of the pre-implemention process, where all the requirements, necessity of an operation domain is first addressed to formulate the challenge that they facing.

2) Solution Strategy (Develop And Deploy): The second rule of this phase is based on the business problem statement where there is a need of strategies to know what to develop and deploy afterwards. The acknowledgement of problem statement will help to achieve the solution in a hassle-free way.

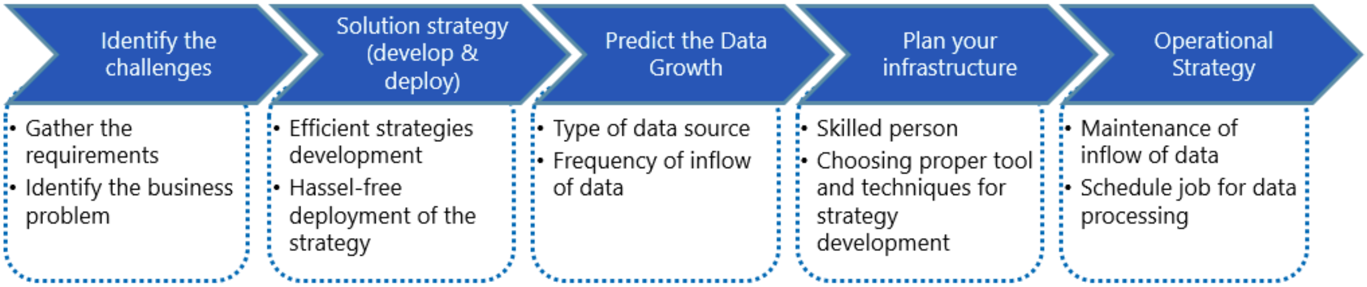
3) Predict The Data Growth: This is the third rule of this phase which is also based on the business problem statement but is about the data which will be required for the problem statement. Here there is a need to have an idea on the type of data, their growth rate and how to handle them.

4) Plan Your Infrastructure: The fourth rule of this phase demands the need to have proper infrastructure to make Big Data Lake successful. *For example*, have the correct person with adequate knowledge on Big Data Lake (possibly by giving proper training or hiring new ones).

5) Operational Strategy: This is the fifth or the final rule of the planning phase which includes how data will be maintained in due course of time.

B. Big Data Lake Construction

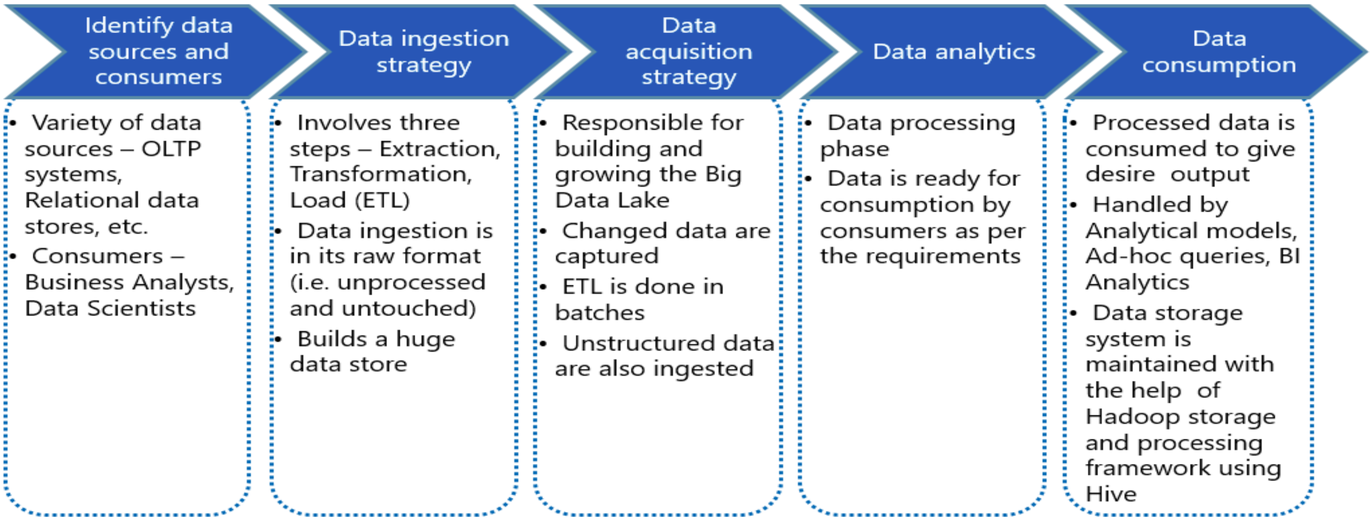
As shown in Fig. 8, the second phase is the execution step which is referred to as *construction phase* and which will be performed after the successful achievement of the first (planning) phase. In this step, 5 different processes



Identify the challenges
Solution strategy (develop & deploy)
Predict the Data Growth
Plan your infrastructure
Operational Strategy
• Gather the requirements
• Identify the business problem
• Efficient strategies development
• Hassel-free deployment of the strategy
• Type of data source
• Frequency of inflow of data
• Skilled person
• Choosing proper tool and techniques for strategy development
• Maintenance of inflow of data
• Schedule job for data processing

Source: Reprinted from [23]

Fig. 7. Big Data Lake Planning Phase



Identify data sources and consumers
Data ingestion strategy
Data acquisition strategy
Data analytics
Data consumption
• Variety of data sources – OLTP systems, Relational data stores, etc.
• Consumers – Business Analysts, Data Scientists
• Involves three steps – Extraction, Transformation, Load (ETL)
• Data ingestion is in its raw format (i.e. unprocessed and untouched)
• Builds a huge data store
• Responsible for building and growing the Big Data Lake
• Changed data are captured
• ETL is done in batches
• Unstructured data are also ingested
• Data processing phase
• Data is ready for consumption by consumers as per the requirements
• Processed data is consumed to give desire output
• Handled by Analytical models, Ad-hoc queries, BI Analytics
• Data storage system is maintained with the help of Hadoop storage and processing framework using Hive

Source: Reprinted from [23]

Fig. 8. Big Data Lake Construction Phase

are also involved which ensures that the requirement of the organization which addressed in detail in the planning phase is executed properly and efficiently without any failure. This is required to make sure that the above said business requirement is handled in a hassel-free way. Hence it is the last but vital step in building Big Data Lakes.

In this phase, first the data source and the type of consumers are analysed before proceeding further to build the data lake. Subsequently, remaining rules like data ingestion strategy, data acquisition strategy, Data analytics, and Data consumption are followed. The subsections below focuses on these rules to give us a detailed insight.

1) Identify Data Source And Consumers: The first rule of the construction phase is to identify data sources which can potentially ingest data into Big Data Lakes. These data sources could be any of these *OLTP systems and relational data stores, Data management systems, Legacy systems, Sensors and IoT devices, Web content & Geographical details* [24]. Moreover, this rule also need identify the consumers who can be either *Business Analyst* or *Data Scientists*.

2) Data Ingestion Strategy: The second rule of this phase is a strategy where a framework captures data from a variety of data sources and ingests it into Big Data Lake [23]. All the data ingestion is in its raw format (i.e., unprocessed and untouched) to build a huge data store. Data ingestion is majorly focused on data extraction .

3) Data Acquisition Strategy: This is the third vital rule of this phase where it is responsible for building and growing the data lake. The data extraction from source data systems and the organization of ingestion strategies into the data lake is clubbed in framework which is laid by data acquisition. Change data are captured, ETL is done in batches, unstructured data are also ingested.

4) Data Analytics: The fourth rule of this phase is about the data which is fetched from data ingestion step. Here mainly cleaning and conforming of the data is carried out as per business needs. This prepared data is then ready for consumption by business analyst or data scientists. This can also be called as *Data Processing phase*.

5) Data Consumption: The fifth or the final rule of the construction phase plays the vital part in building Big Data Lakes. Here, the data will be consumed to give the desired or required outputs. These are basically handled by Analytical models, Ad-hoc queries, BI Analytic. This data storage system is basically maintained with the help of Hadoop storage and processing framework using Hive. (includes deliveries and management of the state-of-the-art guidelines).

V. Conclusion

In this paper, we have addressed main challenges faced during the Data lake implementations. Complex architecture, a pile of unnecessary data and lack of data cataloging are main challenges among them. These issues lead to either delay in the implementation or failure of the Data lake platform. In our research, we have discussed Big Data, Data Lake, ETL ,and state-of-the-art Data warehouse guidelines in brief.

The study helped us to formulate guidelines in terms of two phases - planning and construction which will assist in Big Data lake implementation from an ETL perspective. The planning phase should be focused to avoid issues during the actual Big data lake implementation. Finally, the construction phase along with different sub-phases will guide organizations to capture, process and manage relevant data as per the business needs. These guidelines will help organizations to streamline and define a process framework of Data lake implementation. To conclude, set of guidelines mentioned in this research will make the Data Lake implementation process sustainable and robust.

References

[1] A. Gorelik, *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reilly Media, 2019.

[2] L. Haas, M. Cefkin, C. Kieliszewski, W. Plouffe, and M. Roth, "The ibm research accelerated discovery lab," *ACM SIGMOD Record*, vol. 43, no. 2, pp. 41–48, 2014.

[3] M. Onuralp Gökalp, K. Kayabay, M. Zaki, and A. Koçyiğit, "Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools Customer Experience Analytics: Dynamic-customer centric model View project Classification of Noisy Data: A Data mining challenge View project," 2017.

[4] P. Pasupuleti and B. S. Purra, *Data lake development with big data*. Packt Publishing Ltd, 2015.

[5] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data wrangling: The challenging yourney from the wild to the lake.," in *CIDR*, 2015.

[6] J. Pokorný, "Big Data storage and management: Challenges and opportunities," in *IFIP Advances in Information and Communication Technology*, vol. 507, pp. 28–38, Springer New York LLC, 2017.

[7] M. Kowalczyk and P. Buxmann, "Big data and information processing in organizational decision processes: A multiple case study," *Business and Information Systems Engineering*, vol. 6, pp. 267–278, 10 2014.

[8] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, pp. 2032–2033, 8 2012.

[9] D. Klein, P. Tran-Gia, and M. Hartmann, "big data," *Informatics Spectrum*, vol. 36, no. 3, pp. 319–323, 2013.

[10] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big data," *Business & Information Systems Engineering*, vol. 5, pp. 65–69, Apr 2013.

[11] N. Miloslavskaya and A. Tolstoy, "Big Data, Fast Data and Data Lake Concepts," in *Procedia Computer Science*, vol. 88, pp. 300–305, Elsevier B.V., 2016.

[12] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 48–55, May 2013.

[13] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, pp. 171–209, Apr 2014.

[14] D. Keim, H. Qu, and K. Ma, "Big-data visualization," *IEEE Computer Graphics and Applications*, vol. 33, pp. 20–21, July 2013.

[15] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and good practices," in *2013 Sixth International Conference on Contemporary Computing (IC3)*, pp. 404–409, Aug 2013.

[16] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar, "Computational health informatics in the big data age: a survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 12, 2016.

[17] B. Hüsemann, J. Lechtenbörger, and G. Vossen, "Conceptual Data Warehouse Design Computational Semiotics for Intelligent Decision-Making Support View project Reverse Engineering Database Queries via Intelligent Algorithms View project Conceptual Data Warehouse Design," tech. rep., Universität Münster, 2000.

[18] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," *ITM Web of Conferences*, vol. 17, p. 03025, 2018.

[19] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University - Computer and Information Sciences*, vol. 23, pp. 91–104, 7 2011.

[20] M. R. Llave, "Data lakes in business intelligence: Reporting from the trenches," in *Procedia Computer Science*, vol. 138, pp. 516–524, Elsevier B.V., 2018.

[21] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management*, vol. 36, no. 5, pp. 700–710, 2016.

[22] R. Kimball and Ross Margy, *The Data Warehouse Toolkit*. John Wiley & Sons, Inc., Indianapolis, Indiana, 0.

[23] S. Gupta and V. Giri, *Practical Enterprise Data Lake Insights*. Apress, 6 2018.

[24] W. Kim, O.-R. Jeong, and C. Kim, "A Holistic View of Big Data," *International Journal of Data Warehousing and Mining*, vol. 10, pp. 59–69, 10 2014.