Data lake: a new ideology in big data era

Pwint Phyu Khine[1,a,b,] and Zhao Shun Wang[1,c]*

[1]Department of Computer Science and Technology, School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), 10083, Beijing, China 13051513579, [a]pwintphyukhinecs@gmail.com; [b]pwintphyukhinecs@outlook.com; [c]zhswang@sohu.com

Abstract. Data Lake is one of the arguable concepts appeared in the era of big data. Data Lake original idea is originated from business field instead of academic field. As Data Lake is a newly conceived idea with revolutionized concepts, it brings many challenges for its adoption. However, the potential to change the data landscape makes the research of Data Lake worthwhile.

1 Introduction

In the era of big data, a new term called "Data Lake" came into view of the digital universe. The simplest intention of data lake is to munge every data produced by an organization to give more valuable insight in finer granularity. Big Data technologies are sometimes considered as destructive technologies as they revolutionized the traditional ways of doing things in this data intensive era. Concepts from distributed and parallel system are reapplied as the foundation of big data such as MapReduce paradigms for handling the big Vs characteristics – volume, velocity, variety, value and value. The incumbent SQL databases with ACID characteristics are challenged (and sometimes even replaced) by NoSQL databases with BASE characteristics. Now, Data Lake concept is trying to challenge the reliable, traditional data warehouses for storing heterogeneous complex data.

The idea of Data Lake was first initiated by Pentaho CEO Jame Dixon [9]. If a data warehouse or data mart is seem as a bottle of water cleaned and ready for consumption, then "Data Lake" is whole lake of data which is cleaned for ready use. Ref [16] added more in-depth definition for Data Lake as "a Data Lake stores disparate information while ignoring almost everything". Some believe that new data architecture is required [5] in the age of big data as this computational intensive era requests for new ideas and techniques for storing and processing voluminous, diverse, changing and evolving data. All data generated by an organization regardless of types, structures, or formats will be stored in Hadoop clusters or other similar framework in their original forms. When parts of the organizations need to use the data, that stored data will be loaded and transformed as required by that organization parts. Due to these, concepts in "Data Lake" seem to be challenging to the traditional ways of storing data i.e. data warehouse and data marts.

This paper is organized as follow. Introduction about data lake in big data era is described Section 1. Section 2 introduce about Data Lake concepts, describe the current data landscape and data lake architectural implementation. Comparison of data lake and data warehouse is explained in section 3. Section 4 describe the primary concerns and challenges imposed by Data Lake. Section 5 discuss the overview about data lake and conclude the paper.

* Pwint Phyu Khine: pwintphyukhinecs@gmail.com; pwintphyukhinecs@outlook.com;

2 Data lake concepts

As Data Lake is a relatively new concept, there are only a few academic literatures which are targeted to Data Lake. Ref [8] defines Data Lake as "*a methodology enabled by a massive data repository based on low cost technologies that improves the capture, refinement, archival, and exploration of raw data within an enterprise.*" A data lake may contain raw, unstructured or multi-structured data where most part of these data may have unrecognized value for the organization.

The basic idea of Data Lake is simple, all data emitted by the organization will be stored in a single data structure called Data Lake. Data will be stored in the lake in their original format. Complex preprocessing and transformation of loading data into data warehouses will be eliminated. The upfront costs of data ingestion can also be reduced. Once data are placed in the lake, it's available for analysis by everyone in the organization [6].
Ref [3] suggested more specifications for Data Lake especially from the viewpoint of business domain instead of research community.

- All data are loaded from source systems.
- No data is turned away.
- Data are stored at the leaf level in an untransformed or nearly untransformed state.

The distinct characteristic of Data Lake is that it attracts more attention from business fields instead of academic research fields. Data Lake is a relatively new concept even for big data domain. Therefore, its definitions and characteristics, architecture, creation (implementation) and usage are far more prevalent in web articles and practitioner blogs than academic papers.

Different data lake concepts from the opinion about data lake is reviewed in [12] from "*Yesterday's unified storage is today's enterprise data lake*" to "*a massively scalable storage data repository that holds vast amount of raw data in its native format which is ingested by processing system (engine) without compromising the data structure.*"

2.1 Today data landscape

When comes to core, there are only two operations in data processing – Transactional and Analytical. Daily operations such as Online Transactional Processing (OLTP) are mostly work with CRUD-Create, Replicate, Update, Delete operations of the data for daily routines. Data will be structured and stored in SQL databases. In big data era, not only structured data but also semi-structured and unstructured data will be stored in NoSQL databases. Nonetheless, data in these databases will also be selected, cleaned, integrated, summarized, and transformed according the structure of the data warehouse schema for the analytical purpose. Currently, data warehouses are the dominant approach for providing analytical data. Only transformed data will be stored in the data warehouse.

Data warehouse is created based on fact table with simple W questions - "who, what, when, where" [11]. Then, dimension tables are supplemented based on the fields from the databases. Data are extracted, transformed to conform to data warehouse schema, and loaded into the Data Warehouse (ETL operations). Enterprise gather data from multiple operational databases into a single data warehouse storage to run ad-hoc queries. The query run on the consolidated data can help retrieve business intelligence conveniently.

This cause the separation of concerns for the two important concepts. Daily transactions will be performed by systems such as OLTP (Online Transaction processing). Analytical process such as data analysis, reviewing historical data and correlate data will be performed by analytical system such as OLAP (online analytical processing). Transactions data remain in operational databases whereas complex ad-hoc queries will run on data warehouses

which serve for analytical purpose (thus they will not degrade the performance e.g. response time of transactions query). As DW are intended to build for analytical purpose, data are loaded into batch mode with the defined regular intervals. Data analytical process is performed on the data stored in the data warehouse to support the decision making of the enterprise and to obtain the valuable business insight.

However, as Data warehouses are very large and take time to create, "Data Marts" can be created. "Data Marts" are smaller than data warehouses, and intended to store the data of a part of the organization (i.e. a department in the enterprise). Data warehouse will store data of the whole enterprise. These data marts can be built separately. Or a part of the data warehouse intended for specific functionality or department can be extracted to create a data mart.

There are two famous definitions of Data Warehouse which come from their physical implementation. Inmon defined data warehouse is a subject-oriented (represent real-time object), integrated (data from different databases must be consolidated for showing a unified view), time-variant (data are loaded in time-interval and stored with appropriate timestamps for comparison and analysis), and non-volatile (i.e. instead of currently transition data, data will be collected and loaded into data warehouse in a point of time for analysis), thus no-update collection of data in support of management decision making. Data are stable and are not changed or deleted once are loaded [11]. Inmon approach, a top-down approach for implementation, makes the creation of data warehouse as a first step. In his approach, "data mart", targeted for departmental use, in a way a smaller portion of data warehouse, is created after the establishment of the main data warehouse. Kimball approach is different as it encourages the creation of data mart first. Then these departmental data marts are combined to create an organizational data warehouse. Therefore, Kimball approach is called bottom-up approach.

Data warehouses work better with bit-map indexes, applied materialized view for better optimization [11]. From bird's eye view, DWs are intentionally built for answering ad-hoc queries including historical data (and fall under the category of Analytical systems) which cannot be handled by transactional systems for daily operations. Data warehouses need to use operations like aggregate, join, etc. in one word – need to compute intensive query load. Therefore, most of the data warehouses are built as a single consolidated storage structure, highly summarized, ETLed data various transactional databases.

2.2 Data lake (architectural implementation)

Ref [7] give the explicit explanation of Data Lake from architectural view. "*A data lake uses a flat architecture to store data in their raw format. Each data entity in the lake is associated with a unique identifier and a set of extended metadata, and consumers can use purpose-built schemas to query relevant data, which will result in a smaller set of data that can be analyzed to help answer a consumer's question*". In essence, Data Lake is a data respiratory where all data in an enterprise i.e. structured, semi-structured, unstructured data + binary data are stored altogether regardless of types, format, or structure. The understanding of the data nature is delegated to the data consumer at the time of data retrieval (i.e. query time). When data are retrieved, user will transform that data according to the parts of the enterprise to acquire business insight.

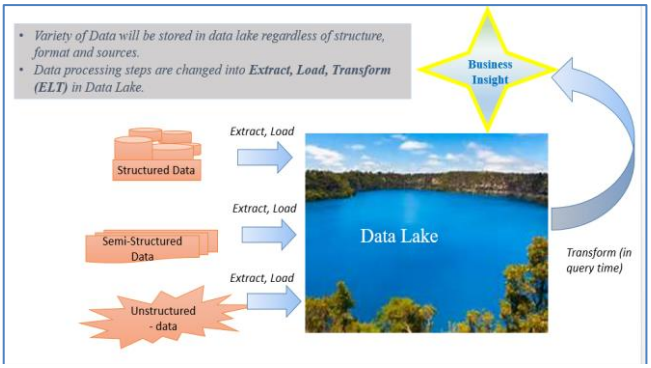

Fig. 1. A Simplified View of Data Lake

Many implementation of Data Lake are originally based on Apache Hadoop. Variety of data from heterogeneous data stores will be extracted to be stored in the Hadoop Cluster. HADOOP (Highly Available Object Oriented Data Platform) is a widely popular big data tool especially suitable for batch processing workload of big data. Hadoop have two main components – HDFS (Hadoop Distributed File System) and MapReduce engine. HDFS File System [17] handle the single point of failure and scalability by replicating multiple copies of data blocks in different nodes of the cluster. All data stored in these data block will be processed in MapReduce approach [10]. Data will be retrieved as a list of key-value pairs i.e. Map phase. The same keys of data will be shuffled, sorted and listed into groups to perform necessary operations i.e. Reduce phase. All data produced by an enterprise will be dumped into the Data Lake Hadoop Cluster [2].

For the real-time load, later data lakes are using stream processing framework such as Apache Spark, Apache Flink. The required data will be transformed according to the need of the analytics systems on the fly in the query time. Saving data with various format and structures from heterogeneous sources, and handling different data velocity (i.e. different processing speed of big data) request the careful consideration in building data pipe for carrying data into the lake [4]. Data Lake can often include a semantic DB, a conceptual model and add a layer of context to define the meaning of data its interrelationship with other data [12]. It can be said that Data Lake strategy includes storing all types of data (data variety) from SQL and NoSQL databases as well as combining the concepts of OLTP with OLAP. SQL databases are used to store structured data. NoSQL databases (Key-value, Columnar, Document, and Graph Stores) are used to store semi-structured and unstructured data. However, they can also be used to store structured data. All the transactional data from these databases (**Extract - E**) will be stored (**Load - L**) into the data lake without changing their format. When data are required (query time), data in the lake will be transformed (**Transform – T**) according to the parts of the enterprise system. Necessary works for query operations have to perform at the application level.

2.3 Suggestion for data lake improvement

As data lake use a flat architecture, each data element has a unique identifier and metadata tags. Although data lakes do not need to follow the strict pre-built schema for manipulation of different types and shapes of data, the order of data arrival time must be maintained. Because a range of data - from historical data to new data generated in near real time with different structure are stored in one place. Data and schema requirement are not predefined in advance until data are queried. This is the schema-on-read property of Data Lake [12].

Ref [12] also suggestion that the separation of data lake tier can be handled in two ways. The first approach is grounded on data structure, and another is centered on time lifetime. Based on data structure, Data Lake can be divided into three tiers - raw data, augmented daily datasets and third-party information respectively. Based on data lifetime, data can be categorized into data which has lifetime less than 6 months, older but still active data, and archived data that need to be retained although they are no longer used (suggested to move and maintain in slower, less expensive media) [12]. Metadata management is an important aspect in Data Lake. As Data Lakes do not have pre-defined schema like data warehouses, they have to rely on metadata during the query time for the analysis process. Metadata are added when data are stored. Ref [1] suggests a content metadata management framework for alignment in data lake construction using openML Data Lake with a prototype. Keys for creating a successful data lake are suggested in [4] as follow.

1. Align innovation initiative with corporate strategy. The priority of the enterprise strategy are Business Acceleration, Operational Efficiency, Security and Risk. Data Lake implementation should focus on the key priority of the corporate strategy.

2. Apply solid data integration strategy. The technology for data integration may be changing overtime in big data. The first Data Lake solutions are based on Hadoop. To handle real-time and streaming data, many DL solutions are now using Streaming framework such as Spark, Flink. Data Lakes need to keep track of evolving best-practices for metadata management. Data analytic pipeline need to automate the process of data extracting, loading, cleaning, transforming, and performing analytics.

3. Establish a modern onboarding strategy. Data Lake can get filled in batch load or in trickle feed. Simplify the process for loading data (regardless of types, sources or complexity) into the lake by enabling and establishing repeatable processes. Meanwhile maintain the appropriate level of data governance. Pay attention to the process of on the fly metadata injection.

4. Embrace new data management strategies by adopting of early ingestion and adaptive execution processing such as MapReduce, Spark, or Flink that allow for flexibility. Derive Metadata at onboarding (loading) time. Create the analytical model on the fly (automate the creation process). Extend data management processing and strategies to all data. Data analytics should be able to apply at anywhere in the data pipeline. Modernize the data integration infrastructure.

5. Apply Machine learning algorithms to drive real business value. The workflow for applying Machine learning algorithms should be repeatable. Data flow should go along with data preparation, engineering data feature, and manipulation of data sets.

3 Differention of data lake and data warehouse

Both data warehouse and data lake are data repositories. However, they are different in many aspects from concepts, structures, and implementation. Data warehouses have well-defined regulatory functions and storage capacity. In theory, Data Lakes have no limit for storage capacity. Any kind of data with any amount can be loaded into the data lake storage repository. “Data Lakes enable enterprises to look past the type and structure of data, giving them the chance to collect as much data as they desire” [15]. Ref [12] gives the data lake distinctions in contrast to data warehouse’s processing of highly structured data, pre-built design before query time, slowly changing data as follows.

- Rapid arrival of unstructured data volumes
- Use of dynamic analytical applications (for query),
- Data become accessible as soon as it is created (as data are transformed based on query operation and application domain).

The detail of differentiation between data lake and data warehouse is explained in next sub sections.

3.1 Deviation of data lake from data warehouse

Data Lake concepts obviously deviate from data warehouse by means of processing of data in the "E-L-T" order and utilizing "Schema-on-Read" approach. Data warehouse approaches follow the traditional ETL process. First, data from operational databases are extracted (E). Then, the data are processed, cleaned and transformed (T) before loading (L) them into the data warehouses or data marts. Data warehouse are especially designed to handle read-heavy workload for analytics. Data warehouse need to define their schema in advance before data are loaded. Therefore, they are considered as "Schema-On-Write" approach.

Different from the traditional idea of ETL of data warehouses, Data Lakes make the different ordering in processing data. The data will be stored in its original format. The preprocessing step will not be handled until the data are required by the application or in query time. Therefore, it breaks the traditional ETL rules of data warehouse. Instead, DL promote the new idea of ELT (Extract, Load, Transform) the change of order for data processing **(ELT)**. There is no predefined data schema in DL. When data are extracted from the source to the data lake, the required metadata are especially added in the data. In this way, Data Lake can handle both write-heavy workload and read-heavy workload (transaction and analytics) recombining two separation of concerns (write and read). The data are not transformed until the applications call them. Only when the data are required and called by the application for query, the data are transformed into appropriate form using the metadata which is added earlier on. In this way, expensive data pre-transformation can be avoided in Data Lake. The transform operations will only be performed when data are read from the data lake. Therefore, Data Lake approach is called "Schema-on-Read" approach.

3.2 Comparison for data lake and data warehouse

From business domain, enthusiasts of data lake concepts [14] provide the summarization for the difference between Data Lake and data warehouse as in Table [9]. They can be further can be explained as follow.

Table 1. Comparison of Data warehouse and Data Lake

Comparison	Data Warehouse	Data Lake
Data	Structured, processed data	Structured/semi-structured, unstructured data, raw data, unprocessed data
Processing	Schema-on-write	Schema-on-read
Storage	Expensive, reliable	Low cost storage
Agility	Less agile, fixed configuration	High agility, flexible configuration
Security	Matured	Maturing
Users	Business professional	Data Scientists (especially those familiar with domain)

1. **Data** – A popular say in business domain is only 20% of the data are structured. Data warehouse is the cream of the crop as it only accepted ETL-ed and consolidated extremely summarized structured data. The other 80% of the semi-structured and unstructured data

(there is no exact suggestion how many % each of them take place - only largest portion is unstructured, and semi-structured portion is larger than structured data. Other experts predict as 70% of unstructured+ semi-structured data.) As DL store data in unstructured and unorganized way, they can be more easily manipulated and handled in a variety of ways which is surprisely more suitable for big data.

2. **Processing** – As mentioned before, data stored in data warehouse are carefully selected through Extract-Transform-Load process. They are intended to use only for analytical purpose – especially for the decision maker level. They passed through the Extract-Transform-Load order of data processing with Schema-on-Write approach. However, data warehouses are limited by their very summarized and structured nature. They are not able to answer out-of-box questions of decision makers or questions that need to extract transactions data and/or combined with unstructured data. User query can only performed limitedly on highly defined structured data. However, Data Lake can handle these kind of user query. Only when user query the data, the data will be transformed according to the user applications for the analytics level in the order of (Extract-Load-Transform) processing applying "Schema-on-read" approach. This gives more flexibility for data scientist who are familiar with domain to retrieve value from previously untapped or unexplored data such as raw or binary data combining with structured data. It is also a reason why data scientists are able to tolerate the imperfections of Data Lake and want to adopt and refine Data Lake.

3. **Cost** – Many data lake solutions are implemented in open-source framework and designed for commodity servers. Therefore, compared to the high-licensing fees of data warehouse storage, it is relatively much cheaper. According to [11], the cost of Data warehouse can be recovered within a year. However, the cost of Data Lake and its ROI is still a questionable debate.

4. **Agility** - Data warehouse design is made in advance before data are loaded (schema-on-write). By definition, it is a highly-structured definition with highly governed data management. Although it is possible to change the data warehouse design, it is very time consuming and require enormous effort as it is tied to many business processes. Sometimes, the whole design reconsideration is required leading to significant maintenance cost. Moreover, Data warehouse cannot handle the request of enterprises that want to analyze a wide variety of data. If they are forcibly enforced, data already loaded into the data warehouse may be corrupted or the Warehouse design may be damaged. Data Lake lacks the explicitly defined structure of data warehouse. Therefore, they are more flexible and agile. It gives the developers and data scientists the ability to easily configure the models, queries, and apps on-the-fly [14] [15]. Because of Data Lake agility and flexibility (as mentioned before), Data Lakes are worth to explore.

5. **Security** - Data warehouses have been there for decades and have well-defined security. However, data lakes are still lacking the answer to the complete security "when" question. Security of data in Data Lake are still left as open research area.

6. **Users** - Until now, Data Lake is still the most suitable for data analysts and data scientists, not analytics for everyone due to the reasons as mentioned above. Many experts in current market are more familiar with data warehouse procedures and consider Data Lake as incompetent and cumbersome. However, data scientists who are interested in the concepts of Data Lake are researching and building the data lake (especially test building of Data Lake in small and medium size enterprises, or building as sample prototypes). They are the main resources which has been providing feedbacks for data lake improvement.

4 Concerns and challenges of data lake

4.1 Data lake concerns

There are two primary concerns for Data Lake. From business point of view, Data Lake may be another marketing hype for Hadoop. From Technical point of view, Data Lake can easily be changed into data swamps [18].

Marketing Hype: Adversary argument is that Data Lake in reality is just a Hadoop's marketing hype of the Business Intelligence solution developers. Gartner point out that Data lake concepts is emerged from Agility and accessibility need of data analysis. Data Lake can definitely provide value to various parts of organization, but they are not the solution ready for enterprise-wide data management [6]. However, the accuse of "Hadoop marketing hype concept" has been slowly evaporating as current data lake solutions are implementing using other frameworks like streaming engines such as Spark, Flink, etc.

Data Swamp: Even for the supporter of data lake noticed the pitfalls of Data Lake. One of the biggest one is becoming a data swamp [7]. No one knows what will be put into the lake. Moreover, there is no procedures from preventing them such as entering wrong data, repeated data, or incorrect data [18]. Data feed into the data lake do not guarantee the veracity since their extraction. If no one knows what kind of data resides in the lake, they might not be able to find out that some data in the lake are corrupted until it's too late. As organizations have started using this technology without sophisticated security measures, these shortcomings become important. Moreover, security compromises are yet to be addressed [16]. Data Lake might have a tendency to data pollution and chances of becoming a data swamp are high. For data lake advocates, how to prevent a data lake from becoming a data swamp turn out to be an interesting topic.

4.2 Challenges of data lake

Data warehouse can guarantee governance, performance, security and access controls. They also have defined semantic consistency. None of these can be guaranteed by Data Lake. Data Lakes lack satisfactory guarantee in areas such as Security, Metadata management, and performance. As data are vaguely structured and most technologies are open-sourced, sensitive data may be compromised, and security can easily be compromised. Many users lack the ability to manage the metadata as Data Lakes do not provide specific mechanism for handling them [15]. Data extractor needs to start from scratch even when other analytics have been found value from these data as there is no metadata maintained or category classified to trace them back. Gaining value from Data Lake is particularly difficult. There are no unique identifiers in Data Lake for advanced searching of large data volume [16]. Unlike data warehouses, "Performance" of Data Lakes have not been benchmarked or proved. These lack of guarantee come from the fact that current Data lakes focus on storing disparate data and ignore how or why data are used, governed, defined and secured.

Based on [6], data lake challenges can be considered as follow.

1. Data lack the ability to determine data quality or the lineage of findings. Other data analysts have found out them in the same data lake but cannot provide for later analysts.
2. Data Lake accepts any data without oversight and governance.
3. There is no descriptive metadata or a mechanism to maintain metadata leading to data swamp.
4. Data need to analyze from scratch every time.
5. Performance cannot be guaranteed.
6. Security (privacy and regulatory requirements) and Access control (weakness of metadata management) as data in the lake can be replaced without oversight of the contents.

5 Discussion and conclusion

Data Lakes try to solve two problems – data silos (old problem) and challenges imposed by big data initiatives (new problem) [6]. Instead of having independently data collections, all data to be stored are collected in Data Lake to handle the old silos problem. The new problem is handling the challenges of big data era i.e. data lakes try to solve the challenges imposed by big data V's characteristics – volume, velocity, verity, variety and value.

If data generated or produced from different departments within an organization are stocked only in their data stores, the chances of becoming data silos are very likely. Data Lake tries to integrate data from these different stores in a single place attempting to end the possibility of data silos. Traditional data warehouse with structured format cannot handle variety of data with different latency need. In big data context, Data Lake may be able to tackle volume and variety if aforementioned challenges were handled.

Conceptually, Data Lake accepts any data structure and volume of data. Every data can simply store into the data lake. Data Lake considers that all required necessary data can be added into the lake as much as necessary (e.g. more nodes will be added in the Hadoop solution ensuring scalability). When data are retrieved from Data Lake, stored metadata and extra stored data can be used as a support for data transformation process and data analytic process.

Retrieving variety of data from heterogeneous sources and the need for handling different velocity processing speed request the careful construction of data pipelines. The data pipelines for feeding data into the lake have to paid attention for handling these Vs' requests when they are built. They need to carry not only all data generated and/or extracted, but also metadata of them and extra data for later use.

Data Lake may suffer by due to data velocity of big data. Data warehouses have definite specification in their operations. They will be loaded in batch at predefined time. Instead, there is no definite processing speed specification for Data Lake. Data Velocity is the speed of data needed to be extracted, cleaned, stored, transformed, loaded or processed. Data velocity can be differentiated into batch, near real-time, real-time and streaming. Since Data Lake tries to replicate the functionality of data warehouse, batch loading is possible for Data Lake. However, undefined flow of stream processing and the need of precise response in near-real time and real time system will ask for definite specification in handling in Data Lake. As security and metadata management are weak in Data Lake, data veracity cannot be assured.

However, Data Lake can provide unique value (untapped, unexplored and surplus) as data for profit which are important for organization. There are optimistic supporters of DL even when turning into data swamp. They argue that even data retrieved from data swamp can still be used to a point in applications and query where all data correctness is not necessary for instance analyzing customer shopping cards abandoning. Fraction of data extracted from Data Lake (even a swamp) can provide unique insights.

As mentioned before, Data Lake has one very important mission i.e. reconciliation of concerns. Analytical and Transactions are separated in the current data landscape. Daily operational data are stored in databases and handled by transactional systems for performing basic CRUD operations and simple query. Data from these transactional data systems are still need to be extracted, transformed, and loaded into highly summarized and consolidated data warehouses to perform ad-hoc analytical queries. The noble mission of data lake is to combine these transaction and analytic into one mechanism. Expected Scenario is that DL will store all transaction data in their naïve format, and allow extracting data on the fly necessary for analysis. Query and Transforms will be performed based on the need of application domain.

A Data Lake may need to pass through five maturity levels [2]. They are - 1. Consolidated and categorized raw data, 2. Attribute-level metadata tagging and linking such as joins, 3. Data set extraction and analysis, 4. Business-specific tagging, synonym identification, and links, and 5. Convergence of meaning within context. As data lake maturity level increases, the usage of Data Lake across enterprise and value of analytics will increase.

Data lakes are also increasing in popularity because of IoTs (Internet of Things) boom [16]. However, currently Data Lakes are not threatening to replace data warehouses as they have not handled the mentioned problems and challenges yet. Ref [16] also gives very interesting opinions of Data Lake. If the methods were discovered that will make ensure data lake solutions are worth replacing in place of warehouse, then the answer would be the evolution of the big data warehouses. It also coincide the opinion for tableau big data forecast [13]. The forecast predicts that Data Warehouse and Data Lake concepts may be combined in near future i.e. totally, Data Warehouse and Data Lake can become only one concept once again by enhancing and adding each other's capabilities. If Data Lake can successfully handle the challenges caused by big data and end the problems of data silos, the whole landscape of data storage architecture may change again in coming future.

Author expresses her gratitude to her supervisor Professor Wang Zhao Shun for providing encouragement in completing this paper. Author thanks to persons (both anonymous and entitled) who provided many good advices in completing this paper. The work is performed under the grant No 2017YFB0202303, National Key Research and Development plan 2017 for High Performance Computing.

References

[1] Ayman Alserafi, Alberto Abell´o, Oscar Romero, Toon Calders,Towards Information Profiling: Data Lake Content: Metadata Management, 2016 IEEE 16th International Conference on Data Mining Workshops.

[2] Brian Stein, Alan Morrison," The enterprise data lake: Better integration and deeper analytics, Technology Forecast: Rethinking integration", Issue 1, 2014, Retrieved 25, Aug. 2017: www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf.

[3] Chris Campbell, Top five difference between data lakes and data warehouses, JAN 26, 2015, Blue Granite, Retrieved Aug 25, 2017. https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses.

[4] Chuck Yarbrough, 5 Keys creating killer data lake, Retrieved July 21, 2017. http://www.pentaho.com/blog/5-keys-creating-killer-data-lake.

[5] Dan Wood, Big data requires a big new architecture, Forbes, Retrieved Aug 8, 2017. https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/#66609cb61157.

[6] Gartner.Inc, Gartner Says Beware of the Data Lake Fallacy, STAMFORD, Conn., July 28, 2014, Retrieved 29 Aug, 2017. http://www.gartner.com/newsroom/id/2809117.

[7] Hassan Alrehamy Coral Walker, Personal Data Lake With Data Gravity Pull, 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, 26-28 Aug. 2015, Dalian, China.

[8] Huang Fang, Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.

[9] James Dixon, Pentaho, Hadoop and Data Lakes, Retrieved 10 Aug 2017. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
[10] Jeffrey Dean, Sanjay Ghemawat, MapReduce: Simplified Processing on large cluster, Communication of the ACM, Vol. 51, No. 1, Jan 2008.
[11] Michael Stonebraker, Uğur Çetintemel, One Size Fits All: An Idea Whose Time Has Come and Gone, Proceedings. 21st International Conference on Data Engineering, April, 2005. ICDE 2005, Tokyo, Japan.
[12] Natalia Miloslavskaya, Alexander Tolstoy, Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues, 2016 4th International Conference on Future Internet of Things and Cloud Workshops.
[13] Tableau, Top 10 Big Data Trends for 2017, Retrieved 30, Aug.2017: https://www.tableau.com/resource/top-10-big-data-trends-2017.
[14] Tamara Dull, Data Lake Vs Data Warehouse: Key Differences, Retrieved Sep 26, 2017 http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html.
[15] Timothy King in Best Practices, Data Warehouse vs. Data Lake; What's the Difference? June 9, 2016, Retrieved Sep 10, 2017: https://solutionsreview.com/data-management/data-warehouse-vs-data-lake-whats-the-difference/.
[16] Timothy King "The Emergence of Data Lake: Pros and Cons", March 3, 2016, Retrieved Sep 15, 2017: https://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/
[17] Tom White, Hadoop: The Definitive Guide, 4th ed., Storage and Analysis at Internet Scale, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, USA, 2015.
[18] Kumar Srivastava, Four Common Mistakes That Can Make For A Toxic Data Lake, Forbes, Nov 25, 2014, Retrieved Aug 24, 2017: https://www.forbes.com/sites/ciocentral/2014/11/25/four-common-mistakes-that-make-for-toxic-data-lakes/