Uncovering Data Warehouse Issues and Challenges in Big Data Management

Rohit Kr Batwada(✉), Namita Mittal(✉), and Emmanuel S. Pilli(✉)

Malaviya National Institute of Technology, Jaipur, India
{2017rcp9070,nmittal.cse,espilli.cse}@mnit.ac.in

Abstract. With the Advancement in Information & Communication Technology, there is an enhancement in Cloud based systems & mobile devices. With the increasing availability & usages of those device, huge data is flowing through various communication channels through different data sources. New demand from academic & industry includes analyzing generated data effectively to come up with fruitful insights that can be actionable either by systems or by humans. Main Goal is to go beyond the questions 'what is happening' and 'why it has happened' to those such as 'what is needed in the future' and 'what are the recommended actions'. Data Lake, which is an extent to Data Warehouse & doing a great job in managing Big Data, providing answers to known questions. In this Paper we are describing issues with traditional data warehouse systems and challenges while managing Big Data with such systems. Also various challenges identified which occurs while understanding the interconnected Data stories in various Data repository and to prepare them for advanced analytics. A novel Data Lake architecture which could work as a Decision support system for Big Data Management has been provided after comparing with other existing.

Keywords: Big Data · Data Lakes · Data Warehouse · Data discovery · Data analysis · Metadata Management

1 Introduction

Human-beings started storing information long back but there was no true relation between the datasets that's how the relation database system came into existence. In the 19th century where people adhered to Mainstream Relational database management systems (RDBMS), entering into the 20th century where the volume & variety of data grew, people started realizing that, their database would not help them to make better decisions and that's where the Data warehouse & BI (Business intelligence) tools came into the picture. Big Data, which has laid down the foundation of all the mega trends happening around us from social to mobile to the cloud to everything. Now, looking at the future, where the emergence of various data sources & "Big Data", brings the demand of Data Lake. *Data lake is emerging as a great asset for any organization due to the huge data which is generating at a faster pace.*

With the advancement of Big Data Management, Data lake which is a considerable prospect as a flexible repository as compared to Data warehouse. The term 'Data Lake'

R. Patgiri et al. (Eds.): BigDML 2019, CCIS 1317, pp. 48–59, 2020.
https://doi.org/10.1007/978-3-030-62625-9_5

itself depicts a 'reservoir, only for Data' [1]. Raw, unstructured data has been kept in a Data Lake in its original format. We need a system, it could be a flat file system where data is moved for processing. Generally Hadoop File System which has already gained popularity due to its speedy processing with huge data sets in the Big Data ecosystem, Data lake is most filled into it to being used.

Due to Data Lake's capability to support storing data in native format, the benefit is, if we have everything to get all known and unknown facts, known facts is being used today but may be things which might not be valuable currently could be turned out interesting in the near future. Adding more to it, Data Flow is exponentially increasing and we never know, data which we lose, will never be captured again so better to make that proof as a future aspect as compared to traditional Data warehouses.

2 Motivation and Background

There are various ways to store Big Data where data storage & data analysis plays an important role to categorise the business use case.

The origin of the project proposal is based on the research article published by the IEEE Computer Society in one of its Journal IEEE Intelligent Systems [1] where Daniel E. O'Leary, compares with existing Enterprise data warehouse (EDW). Investigation using various applications with comparison of Data Lake & Data warehouse has been in place where data sources those required to be integrated from various data sources, facilitating them with Data management.

James Serra, big data and data warehousing solution architect at Microsoft, shares his vision towards using Data Lake over Data warehouse, which also gives the clear directions to think about beyond existing traditional databases systems [2].

Another opinion behind working on Data lake is in industries there is a need to get a 360° view of customer so that business can be improved in multiple ways. Hence there are lot much to explore in Data lake & to deep dive to create such a platform which can extend a Data warehouse Capabilities to meet with Modern day needs. Such platforms are high in demand at domestic as well as international level to store, manage and analyze a variety of data. A new Data Lake architecture approach has been proposed which takes care of issues associated with the Existing Data Warehouses & challenges with existing Data Lake Solutions.

2.1 Data Warehouse and Data Lake

Data warehouses, enormously, a great set of management system to store specially structured data, cleaned up the data & to keep things organised for all business needs. It serves the data for Data Integration & Business Intelligence tools for all their purposes. But now with the addition of various Big Data sources, to make them enable to store all structured & unstructured data in their original form, Data Lake work as a middle layer before data moves into Data warehouse in its structured format (Fig. 1).

It's time for turning data into a service over cloud based platform, hence the questions is quite clear to choose between data warehouse & data lake or else use them together considering data movement flexible from one system to another keeping business analysis on priority (Table 1).

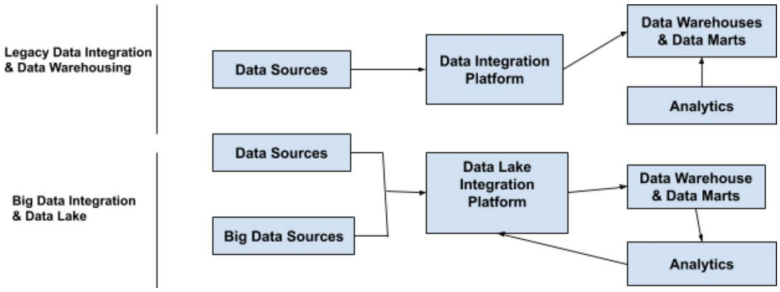

Fig. 1. Data Warehouse & Data Lake design concept

Table 1. Comparison in between Traditional Data warehouse & Modern Data Repository (Data Lake)

Data Warehouse	***Data Lake***
Data structured in Heavily structured Schemas	Data Structured as is (structured, semi-structured & Unstructured formats)
Pre-Processed data ingestion	Data ingestion is quite rapid
Retails Structured Data	Retains all data
Rigid: hard to Change	Agile : Relatively Easy to Change
Expensive and proprietary	Cheap & open source
Schema-on-load to support historical reporting	Scheme-on-query to support the rapid data exploration & hypothesis testing
Matured in terms on Security	Maturing
Users are likely to be Business Professionals	Data Scientists
Accurate results of past events and performance	'Good Enough' prediction of future events & performance

Whenever there is a debate on both the repositories, Data lake vs data warehouse, which is right for me?, *Few Organizations often need both but in the future,* when all small and medium organization will be facing the Big Data related Challenges, Data lakes which is born out of the need to harness big data and benefit from the raw, granular structured and unstructured data is going to be considered as a best fit.

2.2 Big Data and Data Lake

These two terminology Data lakes and big data analytics in the Data space goes along hand in hand. To make a big data application succeed you need at least two things: knowing what (blended) actionable data you need for your desired outcomes and getting the right data to analyze and leverage in order to achieve those outcomes. Big Data is rapidly growing and becoming the best suitable resources for academics & data industries. Curated & refined data helping companies to improve their functions & helping academic organizations to uncover insights with their research work. But as Data which is now a days flows from multiple sources, it is very important to prevent to get converted into Data Silos & Data Swamps.

Data Lake, an architecture which plays an important role here by pooling all data in a central repository. It act as a storage space of Data for data scientists and other business users that can access it for future use. Data Lake & Big Data has shown a great bond together, with the capability of managing larger quantities of Data, open source Hadoop emerged as a great processing platform. In the era where storage cost is decreasing day by day, the main feature of all big data tools & technologies like Spark, Hadoop etc. that storage cost is slightly low if compared with Data Warehouse considering the solutions these tools can provide. There is a huge demand in terms of Managing Big Data Application & with the support of Data Lake it is going to be a great combination for Data Analyst to work on such centralised repositories to get insights for specific domains due to availability of the actual Origin point of Data.

A data lake can be used as a platform where users are allowed to perform various experiments on set of Data models for specific domains. Data Lake can be specific to a domain where Big Data sets generated to that domain can be received in Raw area. Then it send to staging area where with various transformation techniques this could supplied to Data warehouse to produced curated data to get some known value. But both Data Lake & Data warehouses can be optimized for multiple purposes & main goal should be to design a system which suits all level of enterprises need to get the best solution in each possible way.

2.3 Data Warehouse Issues

Since the growth of Traditional Databases, Data warehousing has become a general practice for many businesses. It becomes a very important aspect in the Modern business era & it heavily depends on modern database models like how they support in business development.

Today, Data consolidation is needed at one location from where it can be easily accessible to all business units for analysis purpose. Insights generated from such data can be applied further to improve business process. But there are some challenges identified which needed for improvement on the existing traditional databases & data warehouse systems.

Structuring of Data

Structuring of Data is required considering future needs in mind as we add more & more data into system, structuring of data becomes difficult & its resulting into slow down the processing speed significantly.

Information Driven Analysis

Traditional Data Warehouses driven by the already placed schema. But Now a days there is a need to drive modern data warehousing by the information we provide while writing the data in it. Analysis needs to be done during the early stages of Implementation, where

it should be designed in a way so that data analyst gets enough time on understanding & documenting the business Needs.

Best Selection of Modern Data Warehouse
There is a huge Gap choosing the right set of operational Modern Database for building a data warehouse out of available tools. Choosing a custom warehouse can save time but pre-configured warehouse could save your time from initial configuration. Although selection of Modern Database depends on business Model & specific goals.

End User Expectations: As more and more information added to a data warehouse, end user expect refined results as analysis considering data stored is from Various resources. At the same time performance is definitely decreased when Data Volume Increases & its leading to reduction in Speed & efficiency.

Investment in Data Governance
Information is one of the Major assets for every organization hence it should be monitored closely. Investment on Data governance allows to define task ownership and ensure that data remains consistent and accurate

3 Related Work

Fang [3] research on Data Lake describes the concept of Data lake around Big Data Era. Data Lake is rapidly growing as a way to organize and build the next generation repository to master new big data challenges. Further he shares the key differentiators between Data Lake and Enterprise Data Warehouse. He suggested four model to validate the deployment with the Hadoop. But in terms of Data Lake, support is given to three type of deployment model those are, EDW with Hadoop, Hadoop growing with Data Lake & Data Lake Cloud.

As per the Shahrokni and Soderberg [4] Data lake is a new concept that has the ability to secure, convert and process the data, which make the data can be consumed with speed and value required by the user even though that operation is quite complicated to run. Instead of moving data in a data store, place it into a data lake in its native format. Prieto and Bregon [5] mentioned the primary goal for setting up a data lake to permit ingestion, storage of large amounts of raw data (all formats) for further transformation and integration. Proposed approach emphasizes Data governance considering data quality while continuously tracking Data & Manipulations before delivering the particular analytics.

As per Rusitschka [6], Big data has been considered in the context of its main 3v's volume, velocity, variety including veracity suggested the few attributes of Data Lake Architecture. In the physical form of Data Lake, many servers are using the distributed file system with a layer of Analytics & Data is catalogued on entry as well as during transformation as per need by the data analyst. This approach contrasts with the current traditional ETL methodology whereas rather than going with 'Extract Transform Load',

the process could be followed as 'Extract Load & Transform'. This can be achieved with the 'Schema on Read' approach where, Predefined schema is not playing any role data capturing process.

Ahmad F. [7] Proposes, QoS Lake, its architecture for implementing the QaaS (Quality as a service) model in the same line as PaaS and SaaS using big data technologies. For converting it into service model, approach suggested that Quality of Service in Data lake will be supporting in Analytics & emphasized on Prediction, recommendation & knowledge discovery with Security & Defect Tolerant challenge. Maccioni [8], talks about the Data Preparation part where data scientists needed enough "time-to-action" is exceedingly & it becomes worse when big data sets increased for a Data Lake even no metatags are available to associate & those are included in the ingestion pipeline. KAYAK, a Data Management Framework was proposed to accomplish data preparation in a data lake. It works on metadata cataloging, which keeps track of Meta-Information like how datasets are related to each other.

Suriarachchi and Plale [9] works on Data provenance model which refers to records of the inputs & processes that influence data of interest by providing a historical record of the data and its origins. The paper's three main contributions are related to identification of the data management and traceability issues, Second, reference architecture to overcome the challenges associated in Data Provenance. Third, an evaluation of the proposed architecture to reduce the overhead of Data Provenance.

Raju, Mital and Finkelsztein [10] suggested AIR Traffic Management Data Lake, where The Data Lake had a Cloud component running inside Amazon Web Services (AWS). The data inside the Data Lake is divided into Zones Raw, processed & Refined. Russom P. [11] discussed a tool which is built for inspecting & managing data lakes. Motivation behind the tool development was around Schema discovery solution, identify data security issues & discover data curation process. There are few Challenges faced here in terms of improving the accuracy in joins in tables & gaps were found around integration of Users & tasks.

Sabitha and Vijayalakshmi [12], talks about all happenings around Big Data including recent tools, technologies & challenges those occured due to shift from the traditional systems. Further added that, Devices related to IoT, social websites, automated & smart devices considering as a fuel for the explosion of Data in the near future. Ideal purpose of this paper is to share & discussion the opinions of different researchers, all the available tools for Big Data Process including storage, overall management & analytics. Challenges associated with fee specific domains are highlighted.

Surabhi and Ravinarayana [13] discussed about the need of Data Lake in their survey paper where it is mentioned that to handle the Big Data, Database system processing capabilities required to be extended as per the current capacity. The way Data moves & arrives in Variety & velocity, relational databases architecture is not a perfect fit. The concept of Data Lake has been introduced to take up the challenges arise for Big Data Management.

4 Data Lake Potential and Challenges

Whenever in industries or academics, there is a need to capture information about the client interaction, CRM (Customer Relationship Management) Plays a role here. User

record all the details regarding Sales prospective, customer feedback and other information in database. In this case predefined tables representing customer & all associated details.

Nowadays, There is a need to store the information first which is gathered from various resources which can provide a 360° view of Customer & businesses. Current traditional databases are not that much capable to store & capture every piece of data which is getting generated is unstructured, semi-structured & structured format.

Main Goal here is to leverage the capabilities of Data Lake Platform which can help to migrate and upgrade from Traditional database systems to upgraded Modern Cloud Based Data Lake services [14, 15].

When we talk about Big Data Management, Data Lake is not only designed based on the Big Data Platform on Hadoop but it should be designed using multiple technologies which can help Data Lake to design & provide sustainability like Data Warehouse while offering all Capabilities like DW and more to make it analytics compatible.

4.1 Challenges

All the organizations whether it is a Government firm or Private firm or Academics, all have started looking & investing in Data Lake Architecture. With the current flow of Big Data in every type of organizations, there is a need felt for the new data strategy around the generated data & analytics.

Based on the need, Data Lake architecture contains four major components: Data Extraction & Ingestion, Metadata Management & Data Storage, Query Processing Management & Data Lake Management including analytics engine. When we consider the design section, it has various Challenges to address those are described as below. In this paper, Challenges related to each & every component of Data Lake is described in brief. Every issue associated with each component is altogether a separate entity.

4.2 Data Ingestion Challenges

To ingest data in a central repository from various different sources like Social Media, IoT Devices, sensors etc. is required to ingest into system. Few Major Challenges highlighted in current set of Traditional Database systems & existing Data Lake applications.

1. Ingestion of Multiple Source Files, as current tools required a Common Format to Insert.
2. Problem occurs while subsequently Change the loaded data (Support for Merging & updating is an issue in Current Big Data Ingestion Tools), scaling up of the Data is altogether a Challenge for Big Data Organizations.
3. Parallelizing the ingestion process.
4. Completeness, Correctness & Consistency of your data in ingestion pipeline.

    a. Scalability while Data Collection.
    b. Need to identify missing tuples.
    c. Improvement in Data Lineage required to validate Data from source to destination.

4.3 Metadata Management Challenges

Metadata management plays an important role in the design of a Data Lake Architecture. Metadata management is required to understand the Data Lineage issues while validating the data quality. It should be efficient enough to extract the meta tags from the data sources & making the raw data meaningful while associating the metadatas with it. There are few challenges associated with Metadata management while designing the Data lake.

1. Proper governance required while applying Metadata Management.
2. Reduce confusion while data gets ingested into repo.
3. Streamline data interpretation.
4. Reduce the level of effort required to integrate and prepare data.

4.4 Transformation and Query Processing Challenges

There are various challenges occurs in data lake while transformation & query processing i.e.:

1. Loading of multiple tables & so row & columns hence loading time increased.
2. Amount of data fetched by query if used voluminous data for reference hence processing gets slow down.
3. Analysing the data involved complex queries, sub queries & business logics, sometimes could cause timeout errors & get fails.

To simplify the transformation & query processing and at the same time the speed of the data processing is the main challenge which needs to be considered while working on Data Lake's query processing & analytics engine.

5 Proposed Data Lake Architecture

All the organizations whether it is a Government or Private firm or Academics Institutes, all have started looking & investing in building an in-house Data Lake. With the current flow of Big Data in every type of organizations, there is a need felt for the new data strategy around the generated data for analytics purposes. As per the capabilities which an integrated data lake can fulfill, it's architecture contains four major components which includes Data Extraction & Ingestion as part of Data Staging, Metadata Management & Data Storage, Query Processing Management & overall Data Lake Management to ensure smooth functioning from One Component to another component.

Functional & Technical scope of the proposed Data Lake covers below main items where each & every process has its own significance.

5.1 Data Ingestion and Capture

In our approach data ingestion would be categorised in two steps Extraction of Data & Loading of Data.

In Data Lake priority remains as to keep this automated as there are a lot of overheads when companies perform the task to convert incoming data into single or standardized format manually. So for data lake architecture, in this approach, derived models would be based on a platform which can perform the automation of the process along with conducting other tasks like data quality check of the incoming data by managing the overall data ingestion life.

5.2 Storage and Metadata Management for Various Data Formats

Metadata management if effective, becomes the integral art of any data lake. Data lake handled various types of data but there is a huge changes of converting that data into data swamps. Efficient metadata management of any Data warehouse/Data store & Data Lake needs to be capable enough to capture data about data. It would supports during data conversion from Raw data to refined data in proposed data lake.

5.3 Data Processing and Transformations

Data Processing part in proposed approach will take the data input from the Ingestion section, then after processing the data & this will get stored & there are some ways to keep that data in partitioned way.

As per Fig. 2 mechanism need to define where Raw data can be stored to make Data lake meaningful by capturing the origin point of data. Also along with it will keep curated data as well for analytics purpose.

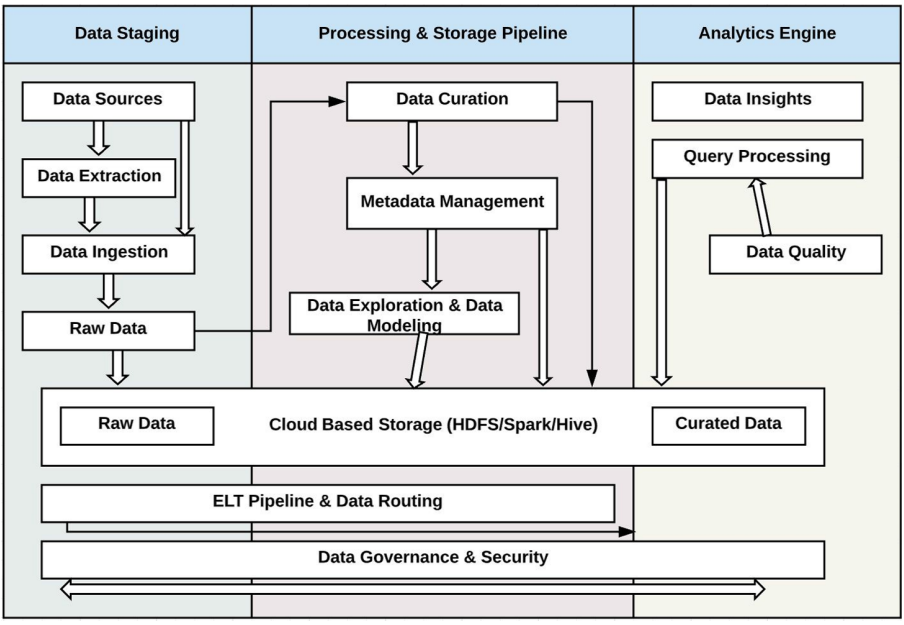

Fig. 2. Proposed Data lake architecture.

5.4 Workflow Management and Scheduling

Workflow management & scheduling refers here for ELT pipeline where after extraction, data gets loaded & then transformed. Main Job here in this step to execution of this pipeline with proper scheduling.

5.5 Data Discovery and Preparation

With Data discovery solution, data lake can be leveraged with adding compatibility for better understanding of data relationships & overall data modeling which would be used for data analysis & if prepared well, data discovery solution module can act as a guide for advanced analytics function for a Data Lake.

5.6 Analytics

Query Processing can be part of the data discovery or analytics section where the soulful purpose of Data lake is to get meaningful insights from the data which gets load in the data lake.

5.7 Data Governance and Data Quality

With Data Quality measures, if data policies, governor limits identified for any data lake, a lot of existing challenges of converting data units into data swamp can be reduced. Whenever there is no existing assurance & strategy about data accuracy & data quality, trust would not be able to generate hence an effective data governance would be needed for any data lake.

6 Discussion

An extensible Data Lake Framework for Big Data Management, handling existing issues will be the key outcome of this Project. This Data Lake Framework will be able to manage structured, semi structured & unstructured Data. It will provide ingestion mechanism, providing a way to use & enrich the metadata with schema designs specific to Domains. It will have a unique storage mechanism which can be a boon for capturing Raw Data as well as Curated Data along with Data Quality. Data governance will also be covered here with basic security & provenance mechanisms. Query Processing & response mechanism from such centric repository will be used for insightful actions.

7 Conclusions and Future Work

A Single Data Store for all Raw Data that anyone could need for analysis could be a great idea to achieve success for Big Data Management. By looking at the needs of a centric data repository organisation has started creating their in- house Data Lakes, but this has not been standardized yet. There are few existing frameworks like Kylo, Zaloni's Data lake etc. are already stepped in the market & few other Kayak, Constance, CoreKG

are the research based Frameworks those are yet to get standardised. All Components have a wide scope to work upon in the future. Data Ingestion, Metadata Management, Query Processing on Modern Data Bases & analytics on Data Lake Repository all has wide scope all around. We will limit our approach to a unique mechanism to ingest the data, to provide metadata tagging along with data curation process followed with Query Processing with Great Data Quality parameters.

In Future New Approach those acknowledge the challenges of volume, velocity, and variety towards society is going to be a great step. Lot of Proposals are already in the pipeline relevant to use of Data Lake into various segments of society like Healthcare Management, Smart City Planning, Crime Investigations, Export/Import of the items on which countries are heavily dependant on other countries but can mitigate the gap with proper planning based on investigations needed on available resources. Scientific community actually facing lot of Challenges with the Current Databases systems, in contrast to that available Modern Databases are not fulfilling all the needs.

References

1. O'Leary, D.: Embedding AI and crowdsourcing in the big data lake. IEEE Intell. Syst. **29**, 70–73 (2014)
2. Serra, J.: Why use a data lake?. https://www.jamesserra.com/archive/2015/12/why-use-a-data-lake/
3. Fang, H., Zhang, Z., Wang, C., Daneshmand, M., Wang, C., Wang, H.: A survey of big data research. IEEE Network **29**, 6–9 (2015)
4. Shahrokni, A., Soderberg, J.: Beyond information silos challenges in integrating industrial model-based data. In: CEUR Workshop Proceedings, vol. 1406, pp. 63–72 (2015)
5. Martinez-Prieto, M., Bregon, A., Garcia-Miranda, I., Alvarez-Esteban, P., Diaz, F., Scarlatti, D.: Integrating flight-related information into a (Big) data lake. In: 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC) (2017)
6. Rusitschka, S., Curry, E.: Big data in the energy and transport sectors. In: Cavanillas, J.M., Curry, E., Wahlster, W. (eds.) New Horizons for a Data-Driven Economy, pp. 225–244. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21569-3_13
7. Ahmad, F., Sarkar, A., Debnath, N.: QoS lake: challenges, design and technologies. In: 2017 International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom) (2017)
8. Maccioni, A., Torlone, R.: Crossing the finish line faster when paddling the data lake with KAYAK. Proc. VLDB Endow. **10**, 1853–1856 (2017)
9. Suriarachchi, I., Plale, B.: Crossing analytics systems: a case for integrated provenance in data lakes. In: 2016 IEEE 12th International Conference on e-Science (e-Science) (2016)
10. Raju, R., Mital, R., Finkelsztein, D.: Data lake architecture for air traffic management. In: 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC) (2018)
11. Russom, P.: Data lakes purposes, practices, patterns, and platforms. https://info.talend.com/rs/talend/images/WP_EN_BD_TDWI_DataLakes.pdf
12. Sabitha M.S., Dr. Vijayalakshmi, R.M., Rathikaa, S.R.E.: Big data – literature survey. IJRASET (2015)
13. Hedge, S.D., Ravinarayana: Survey paper on data lake. Int. J. Sci. Res. **5**, 1718–1719 (2016)
14. Karambelkar, H.: Scaling Big Data with Hadoop and Solr, 2nd edn. (2015)
15. Hegde, S.D., Ravi Narayana, B.: Survey paper on data lake. Int. J. Sci. Res. **5**(7), 2319–7064 (2016)

16. Meena, D., Meena, V.: Data lakes - a new data repository for big data analytics workloads. Int. J. Adv. Res. Comput. Sci. **7**(5) (2016)
17. Huang, P., Chen, Yi.: Enhancing the data privacy for public data lakes. In: Proceedings of IEEE International Conference on Applied System Innovation, pp. 1065–1068 (2018). 978-1-5386-4342-6
18. Zuo, C., Shao, J., Liu, J.K., Wei, G.: Constance: an intelligent data lake system. IEEE Trans. Inf. Forensics Secur. **13**, 186–196 (2018)
19. Klettke, M., Awolin, H., Storl, U., Muller, D., Scherzinger, S.: Uncovering the evolution history of data lakes. In: 2017 IEEE International Conference on Big Data (Big Data) (2017)
20. Amado, A., Cortez, P., Rita, P., Moro, S.: Research trends on Big Data in Marketing: a text mining and topic modeling based literature analysis. Eur. Res. Manage. Bus. Econ. **24**, 1–7 (2018)
21. Kaur, N., Sood, S.: Efficient resource management system based on 4Vs of big data streams. Big Data Res. **9**, 98–106 (2017)