ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 5, May 2017

Enhancement in Business Intelligence using Data Lake Approach

Ramesh Kumar C

Department of Computer Science and Engineering, Galgotias University, Yamuna Expressway Greater Noida, Uttar Pradesh, India

Email Id: C.RAMESH@Galgotiasuniversity.edu.in

ABSTRACT: The data lake strategy has developed as a promising method to deal with huge volumes of organized and unstructured data. Big data technology empowers enterprises to significantly improve its Business Intelligence. Be that as it may, there is an absence of exact examines on the utilization of data lake strategy in the enterprises. This paper gives the aftereffects of an exploratory investigation intended to improve the comprehension of the utilization of data lake strategy in the enterprises. It talked with 12 specialists who had executed this approach in different enterprises and distinguished three significant reasons for executing data lakes: (1) as organizing sources or regions to data warehouses, (2) as a stage for experimentation for analysts and data scientists and (3) as an immediate source for self-administration business intelligence. The examination likewise recognizes a few saw advantages and difficulties of data lake strategy. The outcomes might be gainful for the two practitioners and academics.

KEYWORDS*:* Big Data, Data Lake Strategy, Data Warehouse, Decision Making Strategy and Enterprises

I. INTRODUCTION

BI i.e. Business Intelligence is a contemporary methodology that consolidates processes, methodologies, technologies and architectures to change crude information into significant data for basic leadership. BI can assume a crucial job in improving hierarchical performance by distinguishing new chances, featuring potential dangers, uncovering new business bits of knowledge, and improving decision making forms. In this manner, BI is the top need for associations in many industries. Generally, BI centres fundamentally around organized and inner enterprise data, disregarding possibly important data inserted in unstructured and outer information. This could bring about a fragmented perspective on the real world and one-sided enterprise decision making. Accelerated development and unavoidable advancement of web, cloud and internet technologies have given new which means to the expression "data over-burden".

These innovative advances have prompted the age of remarkable volumes and gatherings of information. Huge and complex information are frequently depicted by the idea of big data[1]. As the big data become progressively accessible, the test of breaking down huge and developing data sets is developing increasingly dire. Along these lines, BI now faces new difficulties, yet in addition energizing opportunities. Big data was enormous trendy buzzwords. The principal associations to grasp the big data were start up and online companies. As per the researchers, organizations like Facebook, Google and eBay were worked around the big data from earliest starting point.

Big data altered the manner in which enterprises manipulated information, giving not just new chances to deal with data, yet in addition better approaches to utilize and increase the value of huge measures of data originating from IOT i.e. Internet of Things, web logs, sensors and social media. Big data additionally underpins the data supply as an asset that associations can employ. Big data has likewise prompted the rise of present day advances such as data lakes that empower enterprises to store also, handle enormous volumes of organized and unstructured data in its local arrangement[2]. Be that as it may, notwithstanding the pervasiveness of this innovation, its writing search yielded just a bunch of studies talking about the data lakes. One investigation talked about the data lakes in the superficial way, while another examined a portion of the difficulties of the data lakes in detailed style. In any case, it found no exact examinations on the utilization of the data lakes in the enterprises. The primary goals of the investigation are to

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 5, May 2017

comprehend the job of the data lake in the architecture of BI and how the data lake is employed practically speaking by the enterprises[3]. The accompanying examination questions have guided the exploration:
What are the reasons for actualizing the data lake into the architecture of BI?
How do the data lakes influence the architecture of BI of an organization?
What are the advantages and difficulties of executing the data lake in the architecture of BI?
Since the subject has not been exactly inspected in earlier research, this investigation directed exploratory research of BI specialists from different ventures. In the following area of this paper, it talk about theoretical background to this study[4]. At that point, it show the exploratory examination approach by portraying the data analysis and data collection analysis. Along these lines, it present the aftereffects of this exploratory investigation. The paper closes with a talk of inquire about discoveries, and an end.

II. BACKGROUND

The big data alludes to the tremendous development of data that associations are as of now encountering. Big data can likewise allude to technology advancements in data processing and data storage that make it conceivable to deal with exponential increments in information volume in a configuration. Another perceived meaning of the big data is in view of 3-V model that includes three elements of difficulties in growth of data: volume, variety and velocity.

- Volume alludes to the developing measure of data.
- Velocity depicts speed of the data accessibility and speed of new information creation for additional investigation.
- At long last, variety portrays the scope of various data types and sources.

All the more as of late, researchers have proposed the fourth V: value that focuses on the significance of accomplishing something significant with the data. BI is firmly interrelated with the big data since BI gives the technological and methodological capacities to the data analysis. The BI is an all-encompassing term for choice supportive frameworks which utilization data analysis and data integration to enhance decision making. Subsequently, it is generally used to depict a wide range of applications of data analysis that assist informed decision making dependent on more extensive knowledge[5]. A commonplace architecture of BI involves ETL i.e. extract transform load layer, data warehouse layer, metadata layer, data source layer, mend user layer. Of these layers, data warehouse layer is the most significant.
Data warehousing includes moving information from a lot of source frameworks into an objective storehouse. The extracted information are sent to transitory stockpiling called data staging zone. The change of data portrays the procedure by which the data are changed over utilizing a lot of the business rules into predictable formats for analysis and reporting. These changed data are then stacked into data warehouse. Hence, data warehouse can likewise be characterized as the focal capacity that stores and gathers data from external and internal data sources to help strategic and tactical decision making. The big data was instituted to portray the changing innovation scene that brought about huge amounts of data, multiple data sources, multiple data formats, and a continuous data flow[6].
Information are the fundamental asset for BI. Ventures across different enterprises are starting to put its information into the data lakes without playing out any data changes. The surviving writing contains hardly any examinations on the data lake innovations. Researcher led an examination in which it characterized the concept of Data Lake. It contended that technology of Data Lake has developed as new kind of information storehouses that empowers processing power and storage to help the investigation of huge unstructured information sets. The ERP employed Data Lake with the goal that data can be gathered once during the stored centrally, updated and initial transaction in real time. In any case, no investigations have yet observationally inspected the utilization of the data lakes in the enterprises. What's more, the BI writing has been quiet on how the data lakes influence BI models.

III. METHODOLOGY

In this exploratory examination, expert interview strategy by researchers was utilized. Information were gathered from 14 semi-organized meetings with the BI specialists from various businesses in Norway. The specialists were recognized utilizing LinkedIn dependent on its propriety as witnesses for this investigation. What's more, a snowballing strategy was utilized in which every source was approached to suggest other conceivable informants.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 5, May 2017

All the meetings were translated and investigated utilizing NVivo. To lead data analysis, thematic analysis guidelines of researchers were utilized, which characterize six phases of investigation. In the primary stage, the creator acquaints herself with data. In this stage, the information were perused and re-read while taking note of down starting thoughts.
Second stage includes creating starting codes. The intriguing highlights of information were coded in an efficient style over the whole informational set and information applicable to every code were gathered. The third stage includes looking for subjects[7]. The codes were grouped into potential subjects and every one of the information pertinent to every potential topic were accumulated. Fourth stage is looking into subjects.
Here, the creator checked whether subjects worked in connection to coded extracts from first stage and the whole data set from second stage. The fifth stage includes characterizing and naming subjects. In this stage, the general examination was inspected to produce clear definitions also, names for each subject. At long last, a report of the examination was generated, which is displayed in the outcomes area.

IV. RESULTS

This section shows the aftereffects of the meetings. To start with, it present how the witnesses characterize data lake strategy, trailed by the apparent advantages of data lakes. It at that point analyse the reasons for data lakes in the enterprises what's more, investigate its difficulties. The witnesses characterized data lakes from the two points of view: a business perspective and technology perspective. From technology point of view, one witness expressed that Data Lake is the assortment of advancements with data which it has to store in a particular format.
Along these lines, data lake isn't the one data lake; numerous advancements that serve need of data. Most witnesses additionally clarified that Data Lake is the focal vault of any kind of data what's more, a focal store of truth. Nonetheless, a couple of witnesses additionally characterized the data lake from the business point of view. For example, one of the witnesses referenced that Data Lake is an ability of business where it can get crude, unaltered information that are from various source frameworks. This witness likewise expressed that Data Lake is a place where it can get every one of data in its enterprise.

- *Perceived Advantages of Data Lakes:*

The witnesses accentuated few perceived advantages of the data lakes: decrease of direct front exertion through data capacity, fast access to crude data, preservation and better data acquisition. First, a greater part of the witnesses underlined that data lake decreases in advance exertion since it ingest information in any of the format without requiring an underlying pattern. It clarified this late processing and early ingestion of data is the developments of the data lakes. Another advantage of the data lakes that few of the witnesses recognized was that it make getting new data simple.
One of the witnesses noticed that, "In data lake, it simply state, it simply dump every one of the information in there. It take every one of information from sources it put into data lake in light of the fact that this is a lot quicker than doing this work to rebuild data. The witnesses additionally noticed that the data lake can deposit a wide range of data, bringing about less exertion during the data acquisition[8]. Its interviews noticed that another advantage of the data lake is it give speedy access to crude information.
Most witnesses contended that having speedy access to crude information is useful to enterprise. For instance, one witness noticed that, "With Data Lake, as a matter of first importance, data will as have now be there. So that implies, when business clients pose an inquiry, the analysts or data scientists could go in there, bring data, and do its data transformation, so it will relate with business question. Finally, numerous witnesses considered protecting data in its local structure to be the advantages of the data lakes. A large portion of the witnesses underlined the significance of approaching crude or immaculate data.

- *Intentions of Data Lakes:*

The meetings uncovered three intentions behind the data lakes: as the staging sources or areas to data warehouses, as the stage for experimentation for analysts or data scientists, and as the direct source to the self- service BI, as showed in Fig. 1. First, most witnesses focused on the significance of employing the data lakes as the staging areas or sources to data warehouses. As referenced before, a staging source is an impermanent area between a data warehouse and a data source. This is shown by the accompanying statement from one witness:
The staging source is a capacity [area], normally relational database, to incidentally keep a duplicate of source information as step while in data warehouse way. In the augmentation, staging zone is additionally employed to store brief outcome sets from estimations and changes as a piece of ETL forms[9]. The principle reason for staging area is

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 5, May 2017

for keep away from potential overload and heavy processing of source framework that may be basic for organizations while changing the information while in transit for data warehouse. The witnesses called attention to a drawback of the staging zones. It expressed that:

At the point when the sensors and IOT i.e. Internet of Things become possibly the most important factor, it need somewhere to store these different information that originates from new innovation. To have the option to store that information, the relational databases, as SQL, will not be fit to this intention.

Second, a few informants discussed utilizing the data lakes for putting away archiving or histories. It clarified that the data lakes can likewise be utilized for offloading chronicled information from the data warehouses. Along these lines, all informants contended that the data lake is the helpful part in any architecture of data warehouse and that it tends to be viewed as an augmentation of the idea of BI[10].

At long last, a few witnesses referenced that the data lakes could be utilized as immediate assets for self-administration BI. The witness noticed that, "On the off chance that it need another report, at that point it can construct that straightforwardly on Data Lake. So it employ self-service BI legitimately on Data Lake, in addition to working together with data warehouse.

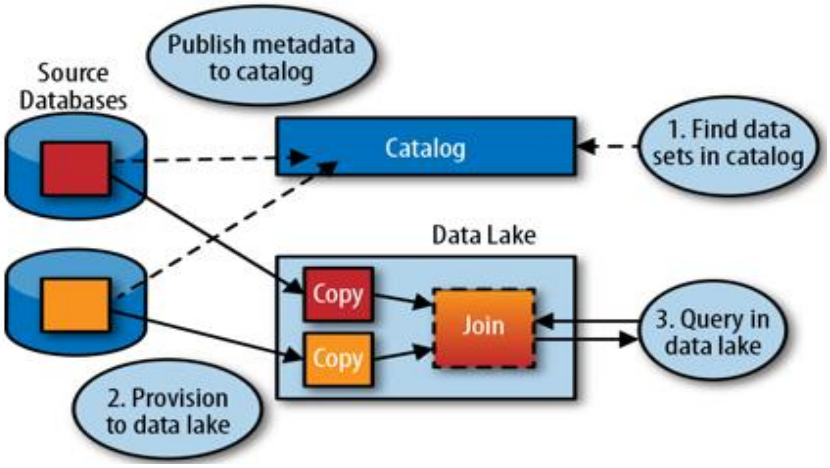

Fig.1: Different Purposes of Dara Lakes

- *Challenges:*

The meetings additionally uncovered a few challenges identified with the data lakes, including challenges identified with data governance, data quality, data retrieval, data stewardship and skills required for analytical purposes.

In the first place, the vast majority of the witnesses called attention to data stewardship is the most significant difficulties of the data lakes. The informants expressed that, "what needs from Data Lake is information stewardships. It is significant to comprehend what this information is. Indeed, even unstructured information could be dumped into this. In any case, on the off chance that it has clickstreams which is coming into it, at that point it ought to be well-characterized this is the website level. Another test concerns the abilities expected to utilize the information in the data lakes. A large portion of the witnesses likewise distinguished information quality as a significant test. One of the witnesses expressed that, "So it has a few difficulties there in the setting of information quality, too. At last, data retrieval represents another test identified with the data lakes. The witnesses clarified:

The distinction between data warehouse and Data Lake is that, in the data warehouse, it change the information before it store it in data warehouse.

V. CONCLUSION

This paper researched the abilities of the data lakes in the enterprises. An exploratory examination was directed to comprehend the technologies of data warehouse and gave bits of knowledge into the apparent advantages and reasons for the data lakes. This investigation discover that the data lakes coordinate flawlessly with an assortment of data warehouses and data sources. Despite the fact that data warehouses keep on meeting clients' data needs and give significant incentive to the enterprises, the data lakes give rich wellsprings of information for analysts, self-service

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 5, May 2017

information customers, data scientists, while likewise serving the necessities of big data and BI. This paper creates three commitments for BI writing: the data lakes are utilized as the staging region to the data warehouse; the data lakes fill in as a stage for experimentation to analysts and data scientists; and the data lakes could be employed as an immediate source to self-administration BI. Basically data lakes don't supplant the data warehouses; rather, it enlarge or supplement the architecture of data warehouse. Thus, the data lakes ought to be viewed as augmentations of the architecture of BI. The investigation likewise recognized a few challenges identified with the data lakes. A more profound familiarity with these difficulties could profit associations trying to set out on the projects of Data Lake. Like any investigation, this examination has a few constraints. Despite the fact that this exploratory examination drew on specialists with experience and knowledge in the data lakes, the specialists came uniquely from enormous enterprises. Along these lines, every one of the outcomes depend on experiences of specialists from huge enterprises. Besides, this examination speaks to just a single exploratory investigation; in this way, it has constrained generalizability.

REFERENCES

[1] A. R. C. Maita, L. C. Martins, C. R. López Paz, S. M. Peres, and M. Fantinato, "Process mining through artificial neural networks and support vector machines: A systematic literature review," *Business Process Management Journal*. 2015.

[2] S. Yadav, G. Shroff, E. Hassan, and P. Agarwal, "Business data fusion," in *2015 18th International Conference on Information Fusion, Fusion 2015*, 2015.

[3] Q. Chen and C. H. Chang, "Enhancement of kernel dependency estimation with information generalization and a case study on skewed data," *Appl. Intell.*, 2014.

[4] 37–41. http://doi.org/10.1037/a0022390 Tuma J. M. & Pratt J. M. (1982). Clinical child psychology practice and training: A survey. \ldots of Clinical Child & Adolescent Psychology 137(August 2012) *et al.*, *Detecting diseases in medical prescriptions using data mining tools and combining techniques*. 2016.

[5] M. Maita, A. R. C., Martins, L. C., López Paz, C. R., Peres, S. M., &Fantinato, "Process mining through artificial neural networks and support vector machines," *Bus. Process Manag.J.*, 2015.

[6] J. Macko, "Formal concept analysis as a framework for business intelligence technologies," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[7] A. Shashwat, D. Kumar, and L. Chanana, "Message Level Security Enhancement For Service Oriented Architecture," in *International Conference on "Computational Intelligence and Communication Technology", CICT 2018*, 2018.

[8] K. R. Prasad, "Big data sentiment analysis using distributed computing approach," in *Advances in Intelligent Systems and Computing*, 2019.

[9] D. Ojakaa*et al.*, "CSOs HSS support proposal," *World Heal. Organ.*, 2014.

[10] 37–41. http://doi.org/10.1037/a0022390 Tuma, J. M., & Pratt, J. M. (1982). Clinical child psychology practice and training: A survey. \ldots of Clinical Child & Adolescent Psychology, 137(August 2012) *et al.*, "Stanovich and West's (2007) Actively Open-Minded Thinking Scale: An Examination of Factor Structure," *Annu. Meet. Assoc. Psychol. Sci.*, 2011.

- RS Venkatesh, PK Reejeesh, S Balamurugan, S Charanyaa, "Further More Investigations on Evolution of Approaches and Methodologies for Securing Computational Grids", International Journal of Innovative Research in Science, Engineering and Technology , Vol. 4, Issue 1, January 2015
- V M Prabhakaran, S Balamurugan, S Charanyaa, "Developing Use Cases and State Transition Models for Effective Protection of Electronic Health Records (EHRs) in Cloud", International Journal of Innovative Research in Computer and Communication Engineering, 2015
- VM Prabhakaran, S Balamurugan, S Charanyaa, " Entity Relationship Looming of Efficient Protection Strategies to Preserve Privacy of Personal Health Records (PHRs) in Cloud", International Journal of Innovative Research in Computer and Communication Engineering, 2015
- ManjotKaur, Tanya Garg, RitikaWason and Vishal Jain, "Novel Framework for handwritten Digit Recognition Through Neural Networks", 3C Technology, Glosses of innovation applied to the SME, ISSN: 2254-4143, Vol. 29, Issue 2, page no. 448 – 467.
- Muhammad Noman, Muhammad Iqbal, Muhammad TalhaAlam, Vishal Jain, HiraMirza, Kamran Rasheed, "Web Unique Method (WUM): An Open Source Blackbox Scanner For Detecting Web Vulnerabilities", International Journal of Advanced Computer Science and Applications (IJACSA) having ISSN No. 2156-557, Vol. 8, No. 12, December, 2017.