International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

Evolution of Data Warehouses to Data Lakes for Enterprise Business Intelligence

Sidharth S Prakash

M.Tech in Software Systems, Department of Information Technology, School of Engineering, CUSAT, Cochin University of Science and Technology, Cochin, India

ABSTRACT: Enterprise data warehouses were built to store the historical structured online transactional data for later to be used for online analytic processing in business intelligence. The insights obtained after the analytics process was used by the business to take strategic decisions so that the business runs better. But with the advent of big data, the data warehouses are not capable of storing the huge volume and unstructured data that is produced by the business. Thus data lakes were envisioned which store the data in its raw format and essentially support storage of big data. In this paper the traditional data warehousing concept will be explored and the reasons why data lakes are opted by most of the business in this modern era of big data and machine learning will be explained.

KEYWORDS: Data Lakes, Business Intelligence, Data Warehouses, Big Data, Machine Learning

I. INTRODUCTION

Business intelligence is the domain that utilizes the historical data patterns to formulate new strategies to run the business better. In traditional systems, business intelligence was facilitated by data warehouses and report dashboards. The data warehouses store the historical transactional data and the reports provided insights to the business on how their business is running at present through graphical plots derived from the data stored in data warehouses. This structure for business intelligence was running successfully till the advent of big data and techniques like machine learning. With huge amount and variety of data getting generated, the traditional data warehouses which were built on top of relational data bases were incapable of storing and providing them to the reporting layer. Thus data lakes were developed which store the data in its raw format and essentially support storage of big data and its processing from a business intelligence perspective [1].

The following paper is organized as follows. The section II describes the traditional data warehouse architecture. This is followed by Section III which discuss on big data. Section IV illustrates the data lake architecture and tabulates the differences of data lakes from data warehouses. Section VIdescribes the challenges in deploying a data lake. Finally this paper is concluded in the section VI.

II. TRADITIONAL DATA WAREHOUSE ARCHITECTURE

Traditional data warehouse system is derived using relational database management systems by organizing the data into fact and dimension tables [2]. Fact tables store the measurable values like revenue, cost etc. and the dimensions are the perspective in which we see the factual data like date, region, product etc. The fact and dimensional data are populated using through the ETL (Extract, Transform and Load) process [3]. The enterprise data warehouse will have ‘n’ number of sources like ERP systems, flat files, CRM etc. These heterogeneous data from various sources is captured by the ETL tool in the extraction process. The captured data is cleaned and enriched in the transformation stage and finally this enriched data is moved to the fact and dimension tables.

The data in the warehouse is later used for reporting and data mining activities as part of business intelligence process. The reports generated from the data warehouse is analysed by the data analyst as well as the top level management of the enterprise. The observed insights are then used by the stakeholders to make efficient and effective business level decisions that can reshape and reform the business output of the organization.

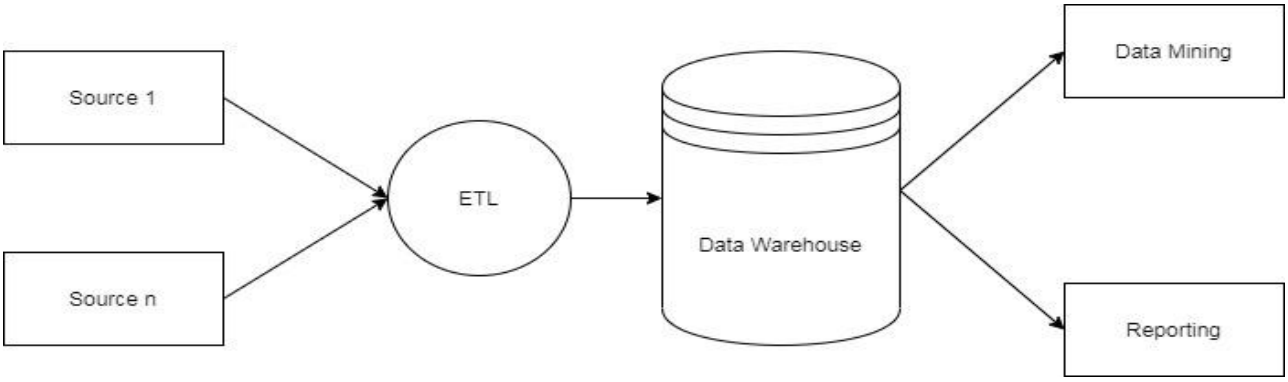

Fig.1. Data Warehouse Architecture

Relational databases are built to store and query structured data [4]. But with the onset of data getting bigger, faster and unstructured, the above data warehouses created on top of relational database management systems no longer proved to be effective.

III. INTRODUCTION TO BIG DATA

Big data is the term used to denote very large amount of data, which are mostly unstructured and difficult to analyse [5]. Big data is usually denoted by four V's:

- Volume: The main characteristic that makes data a big data is the volume of information produced. With the enterprise business expansion to social media, the enterprise business information is not just constrained to business transactions alone, but also to each facebook likes and twitter tweets. Hence a large amount of storage units are required to store and process this information
- Variety: Another characteristic of big data is that it doesn't have a homogenous structure. A structured data like bank statement can be easily fit into a relational model. But now, big data can encompass anything from image, audio, video, GPS tags etc.
- Velocity: The speed at which the huge amount of data getting generated is also an important perspective when we deal with big data. Social media are in always in active mode, IoT devices are always sensing the environment and hence the world of processing has moved to real time. Businesses are now relying on the right amount of real time data to make effective decisions.
- Veracity: This characteristic deals with the quality and integrity of data. As many devices and sources are generating data, the overall quality of the data can fall down in certain instances. Data needs to be cleaned properly to avoid the inherent discrepancies and this requires huge processing capability.

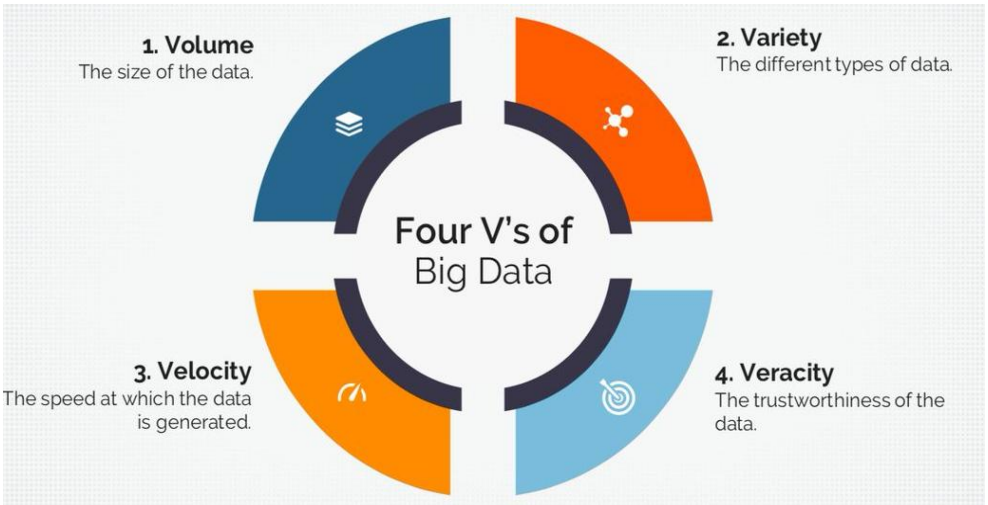

Fig. 2. The Four V's that encompass Big Data

IV. DATA LAKE ARCHITECTURE

The data lake approach is preferred nowadays by enterprises because with storage environments like hadoop, big data can be stored and processes easily [6]. This provides better agility to the business to stand out in the competitive world. Data lakes can also support artificial intelligence algorithms which can perform machine learning on the data lakes. This further enhances the productivity of the business.

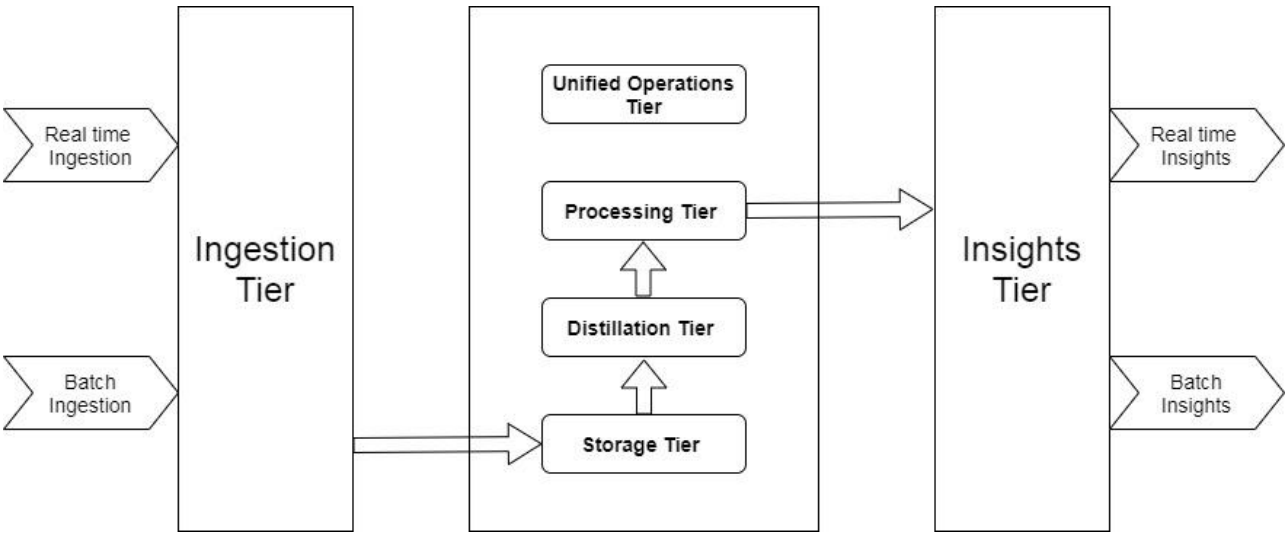

Fig. 3. Data Lake Architecture

The data lake architecture which is capable of dealing with enterprise big data is as follows:

- Ingestion tier: Data flows into the data lake through the ingestion tier either in real time or in the traditional batch format.
- Storage tier: Hadoop distributed file system (HDFS) provides most robust and economical way of storing big data. This layer acts as the landing place for all the data that flows into the data lake.
- Distillation tier: For ease of analysis, the distillation layer captures the data from underlying storage tier and converts to a structured format.
- Processing tier: Real time and batch user queries are run by processing layer and provides the data in a structured format to facilitate analytics.
- Unified operations tier: The overall management of the data lake is taken care by the unified operations tier. The critical operations like the processing control and resources management are done by this layer.
- Insights tier: Data provided by the processing layer will move to the insights tier from which it is directed to the analytics dashboards. These dashboards provide insights to the raw, unstructured data through graphical representations.

Below table illustrates the major differences between a data warehouse and a data lake.

Perspective	Data Warehouse	Data Lake
Workload	Generally data is loaded through batch processing	Supports both batch processing and advanced real time processing
Data	Cleaned	Raw
Schema	Schema is well defined before the data load	Usually schema is defined after the data load is completed as the data values are not known beforehand
Scalability	Scalable to support large data volumes for business in moderate or high cost	Scalable to support large data volumes to business at low or moderate cost
Access Methods	Data accessed through seek method	Data accessed through scan method using

	using standarised SQL and BI tools	SQL-like tools
Complexity	Data processing is relatively simpler and allows complex join statements during retrieval	Data processing is complex and does not allow complex queries during retrieval
Efficiency	Efficiency is acceptable with respect to its cost	Higher efficiency in terms of failover and processing

The potential business benefits of using data lakes for enterprise data are as follows. A data lake allows the processing to happen on several high performance servers and put forth super scalability when dealing dynamic data [7]. They also support parallelization using the popular programming languages like Java, C++, Python etc. The support of high level programming frameworks like Pig and Hive are extremely promising in dealing with big data. Moreover the data lakes also support artificial intelligence and machine learning workloads also which are extremely useful for the modern day business intelligence.

V. CHALLENGES IN DATA LAKE ENVIRONMENT

Even though the concept of data lake is promising to the enterprise business, there could be different challenges in the deploying a reliable model. The main challenges in building and deploying a data lake for an enterprise are:

- Data Reliability: Without proper sophisticated tools, it is difficult of the end users and data scientists to make the best use of the data available in the data lake. Data validation techniques are difficult to enforce in a data lake environment without the help of an expert. As data lakes deal with both batch data and real time data, it is difficult to maintain capture and update of large amount of historical data simultaneously [8].

- Query Performance: As a data lake holds huge volume of data, there can be many bottlenecks hampering its interactive query performance [9]. Limitations in I/O throughput of the data lake can be a reason for slowing down the performance when multiple small files are present in the data lake. The modern data lakes which are created using cloud storage technologies will not purge the deleted files for a fixed amount of time. These deleted files can slow down the query result output even when they are deleted from user the perspective.

- Privacy: Data privacy is not maintained in a typical data lake as the data lake was created for a business purpose and each parties using the data lake share a common data store [10]. If there exist within a business ecosystem where privacy needs to be maintained across some teams or group of people – say top level management, then the data privacy feature needs to be added additionally by the data lake developers.

VI. CONCLUSION

A data lake is a centralized enterprise data repository which is capable of storing both the traditional structured data and the unstructured data of today in its raw format. Most of the data lakes utilize the cloud object storage capabilities to address the data storage on hadoop, unlike the traditional data warehouse which use the relational database management systems for data storage. The modern enterprises are migrating to data lake architecture for its business intelligence operations from data warehouse as a data lake provides efficient and scalable storage. Moreover data lakes provides analytic platform with low cost for both real time and batch data. Moreover the raw data stored in the data lake can leverage the artificial intelligence and machine learning workflows that are pre-built into the cloud storage environment and makes the data analytics much more intuitive. Hence the modern enterprises prefer to leverage the help of data lakes over data warehouses to drive the overall business operations to maximum efficiency.

REFERENCES

1. Huang Fang, “Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem”, IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, June 2015.
2. Rodrigo Cofre Loyola ; Angelica Urrutia Sepulveda ; Manuel Wilson Hernandez, “Optimization slowly changing dimensions of a data warehouse using object-relational”, 34th International Conference of the Chilean Computer Science Society, November 2015.
3. Papa Senghane Diouf ; Aliou Boly ; Samba Ndiaye, “Variety of data in the ETL processes in the cloud: State of the art”, IEEE International Conference on Innovative Research and Development, May 2018.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

4. Yu Haiyan ; Li Jingsong ; Chen Huan ; Zhang Xiaoguang ; Tian Yu ; Yang Yibing, "Performance Evaluation of Post-Relational Database in Hospital Information Systems", Second International Workshop on Education Technology and Computer Science, March 2010.
5. Salman Salloum ; Joshua Zhexue Huang ; Yulin He, "Random Sample Partition: A Distributed Data Model for Big Data Analysis", 15(11) pp 5846-5854, November 2019.
6. Sean Rooney ; Daniel Bauer ; Luis Garcés-Erice ; Peter Urbanetz ; Florian Froese ; Sasa Tomic, "Experiences with Managing Data Ingestion into a Corporate Datalake", IEEE 5th International Conference on Collaboration and Internet Computing, December 2019.
7. Jayesh Patel, "An Effective and Scalable Data Modeling for Enterprise Big Data Platform", IEEE International Conference on Big Data, December 2019.
8. Sun Park ; ByungRae Cha ; JongWon Kim," Design and Implementation of Connected DataLake System for Reliable Data Transmission", 23rd International Computer Science and Engineering Conference, November 2019.
9. Ajay Dholakia ; Prasad Venkatachar ; Kshitij Doshi ; Ravikanth Durgavajhala ; Stewart Tate, "Designing a high performance cluster for large-scale SQL-on-hadoop analytics", EEE International Conference on Big Data, December 2017.
10. Yi-Hua Chen ; Hsin-Hsin Chen ; Po-Chun Huang, "Enhancing the data privacy for public data lakes", IEEE International Conference on Applied System Invention, April 2018.