Data Lake Model to Modern Educational Organizations

Palanivel K[1], Suresh Joseph K[2]

[1]Computer Centre, Pondicherry University, Puducherry – 605014, INDIA.
[2]Department of Computer Science, Pondicherry University, Puducherry – 605014, INDIA.

---***---

Abstract - *Nowadays, educational organizations have a much greater collection of more relevant data than ever before. This includes a diverse range of sources, internal and external data, cloud-based applications, and machine-generated data. Unfortunately, the data warehouse architecture continues to strain under the burden of extremely large, diverse data sets. Business analysts often wait for more for data to flow into the data warehouse before it is available for analysis. They can wait even longer for complex queries to run on that data. In many cases, the storage and compute resources required to process and analyze that data are insufficient. This leads to systems hanging or crashing which results in even longer delays. Hence, it is proposed to introduce an alternative approach called a data lake to educational organizations. Data lake allows educational organizations to scale the storage capacity as the data volume grows to meet the data processing requirements. Educational organizations can move to data lake to effectively shift its culture and creating a data-driven approach to problem solving.*

Key Words: Data lake, data lake architecture, data warehouse, data analytics, educational institutions.

1. INTRODUCTION

Educational institutions decision-makers are keen to utilize the vast and still growing volumes of data on learning objects, students, faculty, staff, and institutions themselves. Vast quantities of data offer the possibility of greater student success and more effectively managed institutions, educational leaders must consider how data analytics can be most effectively harnessed. They must also consider how strategies for good data governance and organizational strategies can support informed decision making [28] with the issues of privacy and security.

Data-driven organizations like educational institutions seek to extract all the insight from all their data to optimize every aspect of their business and better serve their customers (i.e., students, teachers, parents, recruiters and other institutions). They collect and analyze more and more data from traditional sources, such as ERP, CRM, Network, Wi-Fi and e-Learning server, as well as from newer data sources such as for Weblogs, click-stream applications, IoT devices, and social media. Educational institutions increasingly use data from third-party sources, or shared data sets, to enrich their data and analytics. However, it has historically been impossible for an organization to load all its data into a traditional data warehouse [21].

Data from newer sources (i.e., Weblogs, click-stream applications, IoT devices, social media, etc.) often arrives in semi-structured formats that require the data to be transformed and processed before it is loaded. Furthermore, the cost and complexity of storing large quantities of raw, unrefined data in a traditional data warehouse from an increasing number of sources are prohibitive.

1.1 Problems in Existing Systems

Traditional enterprise data warehouse (EDW) architecture has been used for educational organizations. There are data sources, data is extracted, transformed and loaded (ETL) to educational applications for processing [27]. It does some kind of structure creation, cleansing etc. It predefines the data model in EDW and then creates departmental data marts for reporting, online analytical processing (OLAP) cubes for slicing, dicing, and self-service business intelligence (BI) [23]. This architecture is ubiquitous and has served us well for a long time now. However, some inherent challenges in this architecture cannot scale in the era of Big Data in education. Some of the challenges in the existing traditional EDW are listed below.

i. Most of educational organizations setup is completely cloud-based with a huge amount of data.
ii. Educational organizations needs to understand the data first. It should support structured, semi-structured and unstructured data format from various data sources (i.e., social media, sensors & IoT devices).
iii. Educational model supports learning analytics as business requirements and there some anomalies in data so on and so forth. This is tedious and complex work.
iv. Educational organizations have to make choices and compromises on which data to store and which data to discard in the smart learning environment.
v. A lot of time is spent upfront on deciding on store and transformation of data. Lesser time is required on performing data discovery, uncovering patterns, or creating a new hypothesis for business value add.

The existing data warehouse architectures failed to overcome the above challenges and they could not able to handle huge volume of data for processing. In addition, the existing architectures did not scale and therefore were not prepared to handle the velocity or volume of data expected in the future. However, some power users within the educational organization were executing overwhelmingly complex queries that exceeded system limitations. Finally, the data systems themselves were housed and managed by disparate business units with minimal integration. It is recommended to do several changes in the existing data warehouse system. The changes including the client restructure the data warehouse, create a better data process, help them to identify and use the right data types. In addition, it requires a new model and architecture that allows retrieving desired data from the lake without bringing the entire system to a halt.

1.2 Proposed Solution

To create the maximum value out the educational organization's data landscape, traditional data warehouse system is no longer adequate. New architectural patterns need to be developed to harness the power of educational data. To fully capture the value of using Big Data, organizations need to have flexible data architectures and able to extract maximum value from their data ecosystem.

The data lake is an emerging technology for data repository with flexible data architecture. Its main goal is to be a scalable, low-cost data repository for storing raw data from a diverse set of sources so it can be explored and refined. Then, subsets of the refined data are moved to other systems, including a data warehouse, to support high-performance analytics and reporting. The data lake has emerged as the recognized mechanism to enable educational organizations to define, manage and govern the use of various Big Data technologies [9]. This represents an evolution of Big Data towards the use in an enterprise and the associated focus on the management of such assets.

The objective of this paper is to introduce a data lake approach and propose a data lake model to educational organizations. It is the advanced version of the traditional data warehouse concept in terms of data source type, processing type, and structure that operates for learning analytics solutions. It supports new changes of data variants through the iterative approach of enhancements of the architecture adds values to the organization, which implements a data lake.

This research article is organized with the following structure. *Section 2* introduced the state-of-art-technologies required to prepare this article. *Section 3* discussed the literature review and the methodology. *Section 4* presented the proposed data lake model to educational organizations. Finally, this article is concluded in the *Section 5* with future enhancements.

2. STATE-OF-ART-TECHNOLOGIES

Technology helps change the responsibilities of education supporting organizations. They spare more efforts to build the smart learning environment and upgrade hardware facilities and technologies for educational institutions. To optimize education management decisions and processes, a big data platform can be established to deepen education data use based on technologies including data warehouse, data digging and big data analysis.

2.1 Data Warehouse

Data warehouse system is probably the system to which academic communities and industry bodies have been paying the greatest attention among all the decision support systems (DSSs). Data warehousing [7] can be informally defined as follows: "Data warehousing is a collection of methods, techniques, and tools used to support knowledge workers to conduct data analyses that help with performing decision-making processes and improving information resources." The essential properties of a data warehouse system are separation, scalability, extensibility, security and administerability. Data warehouses are regularly updated from operational data and keep on growing. Data is never deleted from data warehouses and updates are normally carried out when data warehouses are offline.

The different classifications adopted for data warehouse architectures are single layer, two-layer and three-layer architecture. A single-layer architecture goal is to minimize the amount of data stored; to reach this goal, it removes data redundancies. The two-layer architecture consists of four subsequent data flow stages - source layer, data staging layer, data warehouse layer and data analysis layer. The three-layer architecture has an additional layer called reconciled data layer or operational data store. This layer materializes operational data obtained after integrating and cleansing source data. Fig.1. shows the three-layer data warehouse architecture. The three-layer architecture can be a *centralized or federated architecture.*

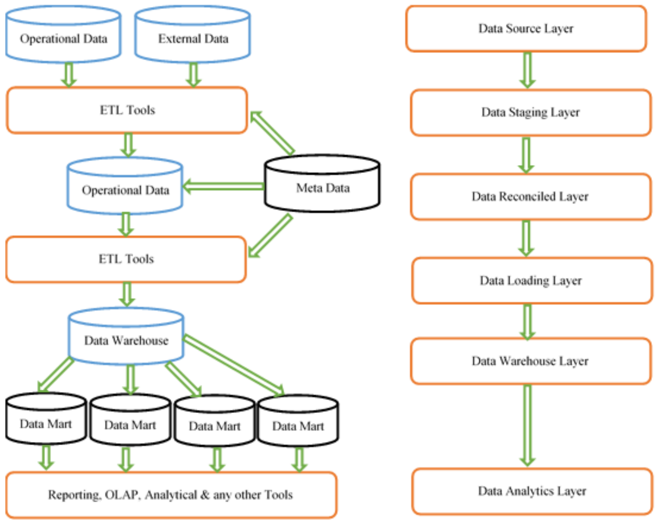

Fig.1. The three-layer data warehouse architecture

2.2 Modern Data Warehouse

Today, the business world is experiencing a data tsunami, with data available from a variety of sources and other sources too numerous and varied to list [12]. The volume and variety of this data can quickly overwhelm a conventional, on-premises data warehouse and often causes data processing and analysis to hang or even crash the system, due to an overload of users and the workloads they process at any given time. Adapting to the exponential increase of data requires a fresh perspective. The conversation must shift from how big an organization's data warehouse must be to whether it can scale cost-effectively, without friction, and about magnitude necessary to handle massive volumes of data.

Technology trends, including the educational cloud and big data, have created enormous amounts of data in educational organizations, which is commonly stored in data warehouse. New technology trends, including the Internet of Things (IoT), Artificial Intelligence (AI), and live streaming data are desired to analyze it in new and different ways. The traditional enterprise data warehouse reached maturity, the enterprise data warehouse died, laid low by a combination of big data and the cloud. Hence, a modern data warehouse is required to meet challenges of big data and cloud technology.

A modern data warehouse enables organizations to efficiently store, manage, access, and generate value out of data stored in both on premise storage infrastructures as well as in the cloud. Modern data warehouse enable data to be consolidated in the right manner, allowing organizations to apply next-generation analytics and AI technologies to generate value from this data. Moreover, a data lake can be the foundation of a modern data platform.

2.3 Data Lake

Data lakes are being accepted by business organizations that want to modernize their data platforms. A Data lake is a storage repository that can store a large amount of structured, semi-structured, and unstructured data. It is a place to store every type of data in its native format with no fixed limits on account size or file. It offers high data quantity to increase analytic performance and native integration. A data lake [3] is a large storage repository that holds a vast amount of raw data in its native format until it is needed. An enterprise data lake (EDL) is simply a data lake for enterprise-wide information storage and sharing.

A data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time as shown in Fig.2.

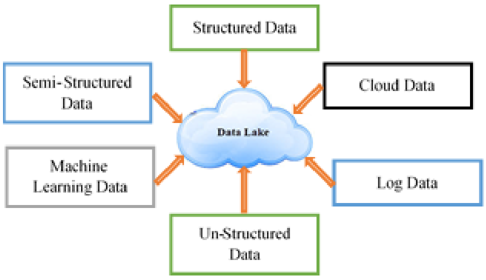

Fig.2. Data types in Data Lake

The Data lake democratizes data and is a cost-effective way to store all data of an organization for later processing. Data lake has a flat architecture. Every data elements in a Data lake is given a unique identifier and tagged with a set of metadata information. The key data lake concepts are data auditing, discovery, exploration, ingestion, governance, lineage, quality, storage and security [13]. One needs to understand these key data lake concepts that to completely understand the data lake architecture 15]. One of the main goals of data lakes is to help promote self-service from the perspective of the business analyst and business user.

Data lakes must incorporate other technologies that were designed for data lakes. This is where the traditional BI tools have fallen short when it comes to BI on big data platform. Most traditional BI tools treat the data lake like any other data store. The sophisticated users need to run the system, which takes away from the goal of achieving self-service BI for business users.

2.4 Data Lake Components

Fig. 3 shows the key data lake concepts that one needs to understand to completely understand the data lake architecture.

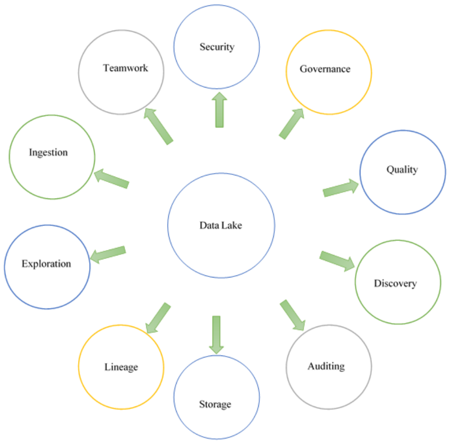

Fig. 3. Components of Data Lake

Data Ingestion allows connectors to get data from a different data sources and load into the data lake. *Data storage* should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats. *Data governance* is a process of managing availability, usability, security, and integrity of data used in an organization. *Data security* needs to be implemented in every layer of the data lake. The basic need is to stop access for unauthorized users. Authentication, accounting, authorization and data protection are some important features of data lake security.

Data quality is an essential component of data lake architecture. Extracting insights from poor quality data will lead to poor quality insights. *Data discovery* is another important stage before it can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the data lake. *Data auditing* tasks are tracking changes to the key dataset. Data auditing helps to evaluate risk and compliance. This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination. *Data exploration* is the beginning stage of data analysis. It helps to identify right dataset is vital before starting data exploration.

All the above components need to work together to play an important part in data lake building easily evolve and explore the environment.

2.5 Maturity Stages of Data Lake

The data lake maturity stages are first, second, third and fourth stage. This first stage involves improving the ability to transform and analyze data. Here, business owners need to find the tools according to their skillset for obtaining more data and build analytical applications. The second stage involves improving the ability to transform and analyze data. In this stage, organizations use the tool, which is most appropriate to their skillset and they start acquiring more data and building applications. The third stage involves getting data and analytics into the hands of as many people as possible. In fourth stage, enterprise capabilities are added to the data lake. Adoption of information governance, information lifecycle management capabilities, and metadata management.

A data lake platform (DLP) is meant to address the data lake challenges [20] of complexity, sluggish time to value, and overall messiness. A DLP offers a single platform that takes care of everything from data management, ETL and storage to processing. It improves performance and resource utilization throughout storage, processing and serving layers. DLPs enable developers without an extensive big data background to create a complete pipeline from incoming data streams to structured data, that can be queried using SQL or other analytic tools. By doing so, DLPs enable organizations to generate more business value from their data lakes, at a faster pace [11].

2.6 Data Lake Architecture

Data Lake architecture is all about storing large amounts of data, which can be structured, semi-structured or unstructured, e.g. Web server logs, RDBMS data, NoSQL data, social media, sensors, IoT data and third-party data. A data lake is a function and architecture decision [4]. When it does come to architecture and technologies that are traditionally used, such as relational databases, object stores, or Hadoop, they are complementary architectures and technologies. A data lake has flat architecture [1] to store data and schema-on-read access across huge amounts of information that can be accessed rapidly. A data lake gives structure to an entity by pulling out data from all possible sources into legitimate and meaningful assimilation. Adopting data lake architecture means developing a unified data model, explicitly working around the existing system without affecting the business applications, alongside solving specific business problems.

Data lakes are helpful when working with streaming data such as by IoT devices, video conferencing devices, clickstream tracking, or product/server logs. Typically, these are small records in very large quantities, in a semi-structured format. The deployments of data lakes usually addresses business use cases - *business intelligence and analytics, data science* and *data serving*.

The key benefits [1] of a data lake are scalability, high velocity of data, structure, storage and scheme. The other benefits are centralization, security, flexible access, normalize and enrich. Information is power and a data lake puts enterprise-wide information into the hands of many more employees to make the organization as a completely smarter, more agile, and more innovative. Common challenges [20] include in a data lake are technical complexity, slow time-to-value and data swamps [17].

2.7 Related Terms

A data lake and a data hub [32] are vastly different at their core. Data warehouses and data lakes are endpoints for data collection that exist to support the analytics of an enterprise while data hubs serve as points of mediation and data sharing [30]. Data warehouses, data lakes, and data hubs are not interchangeable alternatives [8]. Nevertheless, they are complementary and together they can support data-driven initiatives and digital transformation. The other related terms are Data virtualization [30], Data Swamp [16], and Data Cube [9] and Data Mar*t* [2].

3. LITERATURE REVIEW AND METHODOLOGY

Data lake for educational organizations, an exciting resource for education researchers, policymakers, and innovators. The data lake securely houses data gathered from across the colleges and universities. This includes learning management system (LMS), campus IT systems, student surveys, research study results, and much more. The data lake allows academicians and researchers to follow how its students and alumni learn throughout their lifetimes. This is an unprecedented opportunity to optimize the long-term impact the university has on its students. Educational data lake continually works with stakeholders including faculty, students, and advisors to update its data ethics and data management policies and processes.

Higher education institutions take advantage of Big Data to improve student performance and raise teachers' effectiveness, while reducing administrative workload [31]. Higher educational institutions can benchmark their student, teachers and curriculum performance against like universities, yielding yet new insight into potential for improvement.

The purpose is to frame a data lake [27] and provide context around how they fit within enterprise data strategies. The rapid growth of data and demand for increasingly versatile analytic use cases (such as reporting, machine learning, and predictive analytics) could result in educational organization outgrowing its data infrastructure much sooner than you currently foresee [6].

Designing data lake architecture [13] is critical for laying down a strong data foundation. This architecture aims at storing the maximum data possible in its raw form for an extended period; the lack of design planning can result in the lake being transformed into a data swamp. Building a data-driven, inquisitive and adaptive culture in higher education is not simple. Data lake technology [5] may be combined with more traditional data warehousing technologies provide great flexibility for faster data discovery and analysis.

The Big Data-powered applications for higher education institution [31] are student acquisition, student course major selection, student performance effectiveness, student work groups, student retention, teacher's effectiveness, student lifetime value/booster effectiveness, student advocacy and bookstore effectiveness.

Modern Big Data infrastructures [10] act as components of data lake architectures. These platforms offer various services, including *Amazon S3, Azure Databricks, Hadoop and Microsoft's Azure Data Lake.* Data Lakes are majorly implemented through Cloud providers [25] and architected with several data storage and data processing tools and managed services based services are associated to process and maintain the data infrastructure for Data Lake.

Modern capabilities, including predictive analytics and ML [24] enable organizations to leverage large amounts of data from social media, online journeys, the IoT and other sources to enable data-driven decisions across an organization. Leveraging a data lake [12] to store the necessary information for powering predictive analytics and machine learning workloads empowers staff across an organization to analyze data, test theories and drive changes to business processes, the customer experience and products.

3.1 Methodology

The methodology includes literature survey in last five years. The choice of the review period was a practical one and took into consideration the fact that data lake is a rather recent phenomenon. It was not expected that there would be any reference before 2017. Beside the period of publication, it is used two inclusion criteria for the literature search: full article publication and relevance to the topic. About ten articles focusing solely on data lake design. The literature survey followed a systematic approach. This was done in three steps.

1) In the first step, it searched using all combinations of two groups of keywords of which the first group addresses "Data Lake" (i.e., "Model" and "Architecture") and the second group refers to "Educational" (i.e. "Organizations" or "Institutions").
- The two databases were chosen because of their wide coverage of relevant literature. From these two databases, about 5 peer-reviewed articles were retrieved. These were scanned for relevance by identifying passages that were relevant to "Data Lake to Educational Organizations".
- In screening the literature, it first used the search function to locate the paragraphs containing the above key words and then read the text to see whether they can be related to the proposed topic. As a result, all were considered most relevant.
- It is found that the number of relevant peer reviewed literature not very high which could be explained because "Data Lake" and "Educational Institutions" are relatively new concepts. For that purpose, it has used the search engine for reports, magazines, blogs, and web-items related to the groups of keywords.
- This has resulted in 53 including reports, magazine articles and blogs. Each of the above was evaluated on relevance based on its title of the article. It was removed possible duplications in the resulted articles. The result containing the title of the article that were evaluated by further reading. Consequently, about 50 topics have been considered as containing relevant information for further analysis.

2) In the second step, it read the selected literature in detail to extract the information relevant to the title.
3) In the third step, the extracted information was analyzed and synthesized as described in Section 4.

4. PROPOSED SYSTEM

In educational organizations, the terms data-informed and data-driven are describing how data analytics supports organizational decision-making. The availability of effective data reporting, data analysis and data visualization tools allow harnessing the power of structured, semi-structured, and unstructured data resources through data architecture designs such as a data lake. Data lake's self-dependent mechanisms to create process cycle to serve enterprise data to help them in consuming applications.

4.1 Requirements

The foundation of any data lake design is physical storage. The core storage is used for the primary data assets. Typically, it will contain raw and/or lightly processed data. The key considerations when evaluating technologies for cloud-based data lake storage principles and requirements [26] are high durability, high scalability, independence from the fixed schema, separation from computing resources and support for different types of data.

4.2 Quality Attributes & Design Factors

The *quality attributes* considered the proposed data lake system are data engineering, modelling, visualization, platform, infrastructure, security, transparency, flexibility and scalability. The data engineering attributes able to integrate data from any data source. The data models accelerate cross-functional analyses that can be tailored to the educational institution's needs. Data visualization can integrate with existing institutional BI and data visualization tools and provide new visualizations. For more flexibility and scalability can be gained by separating storage and compute capacity into physically separate tiers, connected by fast network connections [18]. This can allow scaling the storage capacity and independently scaling the compute capacity to meet the processing requirements.

The *design factors* [14] considered during the design of data lake architecture are innovation, speed, self-service, single view, streaming of data, storage option, analytics engine, BI and data catalog. The data platform able to integrate with, extend and augment core technologies to address multiple use cases as they are identified and evolve. Infrastructure requires high-performance, scalable and secure cloud-based data platform. The security secures data in transit and at rest. Educational institution's subject-matter-experts will have transparently access and visibility into the code and logic of predictive models.

4.3 Data Lake Model

The primary value of a data lake is enabling flexibility, through a scalable platform for analysis of complex data sets. Many different technologies will go into this analysis, including predictive analytics tools, data modeling, data quality and machine learning. The first part of any analytical workflow is the data process. The steps commonly followed to ingest, cluster, index and ultimately analyze data within a data lake. These steps are key to ensuring that high quality data is brought together, associated properly and organized to enable data scientists to analyze the prepared data. The proposed data lake model to educational organizations is shown in Fig. 4. The educational institutions consist of many different departments, sections, and users. They have access to many data sources from various business systems and applications installed in their smart devices and their locations. It includes various data sources.

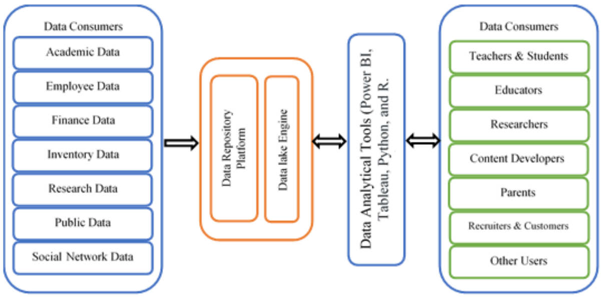

Fig.4. Data Lake model to Educational Organizations

Academic data includes programmes, courses, admission, seminars, conferences, workshops, events, etc. *Employee data* includes teachers, students, staff, parents, recruiters, other educational institutions and research organizations, etc. *Finance data* includes employee salary, scholarships, payment, settlement, etc. *Inventory data* includes the items purchased and their maintenance to the educational organizations. *Research data* includes publications, projects, research analytics, research runs, test results, equipment data, etc. *Customer support data* includes tickets and responses. *Social Networks data* includes data collected from various social networks and interactions online.

These data focus on making structured and unstructured data searchable from a central data lake. The goal is to provide data access to end-users in near real-time and improve visibility into the manufacturing and research processes. The enterprise data lake collects and processes all the raw data in one place, and then indexes that data into search (i.e., Cloudera, Impala, and HBase) for a unified search and analytics experience for end-users.

As suggested by [13], the proposed model consists of multiple user interfaces (or APIs). They are being created to meet the needs of the various user communities. Simple search user interfaces (UIs) and sophisticated UIs allows more advanced searches to be performed. Some UIs will integrate with highly specialized data analytics tools. The security requirements will be respected across UIs. By design, this model is supposed to be well abstracted from the services that consume information that resides within them.

4.4 Data Lake Architecture

Data Lake architecture should be flexible to educational organizations. It relies on a comprehensive understanding of the technical requirements with sound business skills to customize and integrate the architecture. Educational organizations may prefer to build the data lake customized to their need in terms of the business, processes and systems. The data lake architecture is efficiently designed to support the security, scalability, and resilience of the data. The proposed data lake architecture to a smart education system is shown in Fig. 5. This architecture consists of data intake tier (data ingestion), data processing (data management) tier and data consumption (application Tier).

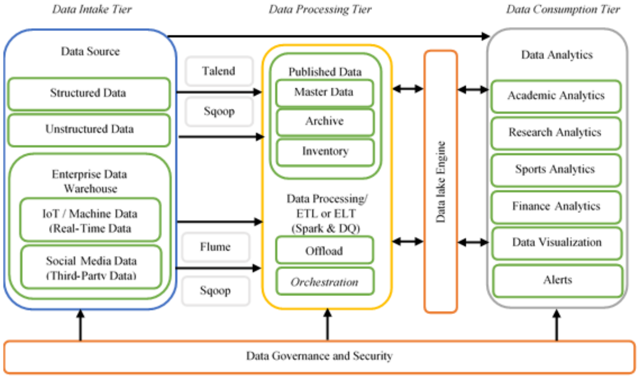

Fig.5. Data Lake Architecture to Educational Organizations

The Data Intake Tier (or Data Ingestion Tier). The objective of data intake tier is to ingest data into raw as quickly and as efficiently as possible. The data source can be homogeneous or heterogeneous, or both. The data here is not ready to be used; it requires a lot of knowledge in terms of appropriate and relevant consumption. In this tier, the educational data should remain in its native format. It does not allow any transformations at this stage. No overriding is allowed. Raw data still needs to be organized into folders. The end-users should not be granted access to this layer. The end-users are dashboard, BI tools, data science, e-Commerce and mobile apps. Examples are transaction business applications, EDW, Multiple documents (.csv and .txt,) Software-as-a-Service (SaaS) applications, device logs and IoT sensors.

The Data Processing (or Data Management) Tier. The data processing tier comprises of DataStore, Metadata store and the replication to support the high availability of data. The index is applied to the data for optimizing the processing. The best practices include including a cloud-based cluster for the data processing tier. This tier is efficiently designed to support the security, scalability, and resilience of the data. In addition, proper business rules and configurations are maintained through educational administrators. There are several tools and cloud providers that support this data processing layer. The data processing tier has two sublayers called *standardized data tier* and *cleansed data tier*.

The objective of *standardized data sub-tier* is to improve performance in data transfer from raw to cleanse. Both daily transformations and on-demand loads are included. While in raw, data is stored in its native format, in standardized it chooses the format that fits best for cleansing. The structure is the same as in the previous layer but it may be partitioned to lower grain if needed. In *cleansed (or Curated) data tier*, data is transformed into consumable data sets and it may be stored in files or tables. It should expect cleansing and transformations before this tier. Usually, end users are granted access only to this layer. Examples are Apache Spark, Azure Databricks, and Data lake solutions from AWS. *Data lake engines* run between the systems that manage the educational data and the tools you use to analyze, visualize, and process data for different data consumer applications. Rather than moving data from sources into a single repository, data lake engines are deployed between existing data sources and the tools of data consumers such as BI tools and data science platforms.

The Data Consumption (or Data Application) Tier. After data processing tier, data lake provides the processed data to the target systems or educational applications. Several systems consume data from data lake through an API or connectors. It sourced from cleansed and enforced with any needed business logic. If any of education applications use ML models that are calculated on data lake. The structure of the data will remain the same, as in cleansed.

Data Governance and Security fixes the responsibility for governing the right data access and the rights for defining and modifying data. The *Information Lifecycle Management* (ILM) ensures that rules are governing what the end-users can or cannot store in the data lake. The *Metadata* captures vital information about the data as it enters the data lake and indexes this information so that users can search for metadata before they access the data itself. Well-built metadata will allow organizations to harness the potential of the data lake, deliver the self-service and data provisioning mechanisms to the end-users to access data, and perform analytics.

The proposed architecture allow the end-users to transform raw data into structured data that is ready for SQL analytics with low latency. A centralized data lake eliminates problems with data silos (like data duplication, multiple security policies, and difficulty with collaboration), offering downstream users a single place to look for all sources of data. All data types including batch and streaming, video, image, and binary files, and more can be collected and retained indefinitely in a data lake. Data lakes are incredibly flexible, enabling users with completely different skills, tools, and languages to perform different analytics tasks all at once.

4.5 Components of Proposed System

Educational organizations mature through the different levels, there are technology, people and process components. The data lake provides a platform for execution of advanced technologies, and a place for staff to mature their skill sets in data analysis and data science. To effectively manage data in a data lake requires an in-depth understanding of issues around data governance, metadata management, lineage tracking, indexing/searching, security, and others. The logical components [20] included in the proposed solution is discussed below.

All above-discussed components work together and play a vital role in data lake to create an environment where end users can discover and explore valuable insights out of the data in a secured and managed environment. The benefits of the proposed architecture are self-service, speed, innovation, advanced analytics and cost. The challenges of building a data lake are a huge volume of data, volatile data and governance & support. It is difficult to deal with sparse, incomplete and volatile data. Hence, a wider scope of dataset and source needs larger data governance & support. This multi-layer data lake architecture introduces many challenges [20], which include flexibility, complexity, IT-centric, costs, data governance and data freshness.

Data lakes and data warehouses are both design patterns [29]. The other patterns are data science lab, advanced analytics, data pipeline and stream analytics. The above solution patterns shown here support many different data lake use cases. These patterns can be combined for different purposes with all accessing data from a common object store.

4.6 Case Study

The best practices for data lake deployment are discussed in detail. Architectural components, their interaction and identified products should support native data types. The design of the data lake should be driven by what is available instead of what is required. The design should be guided by disposable components integrated with service API. The data lake architecture should be tailored to a specific industry. It should ensure that capabilities necessary for that domain are an inherent part of the design. Data discovery, ingestion, storage, administration, quality, transformation, and visualization should be managed independently. In a data lake, faster on boarding of newly discovered data sources is important. It should support existing enterprise data management techniques and methods. The use case of the best practices [22] for data lake deployment are data lake architecture for Biopharmaceuticals and High Tech.

5. CONCLUSION & FUTURE ENHANCEMENT

A data lake is a storage repository that can store a large amount of structured, semi-structured, and unstructured data. The main objective of building a data lake is to offer an unrefined view of data-to-data scientists. In this article, it is proposed to design a data lake architecture to educational institutions. Data ingestion tier, data processing tier and application tier are important layers of data lake architecture. The design of a data lake should be driven by what is available instead of what is required.

Data lake reduces the long-term cost of ownership and allows economic storage of files. The biggest risk of data lakes is security and access control. Data can sometimes be placed into a lake without any oversight, as some of the data may have privacy and regulatory need. Combining Data Lake with machine learning (Sherry Tiao, 2018) can make predictions, find fraud, make recommendations and more - the technology is constantly changing and getting more exciting. In future, it is proposed to introduce a data hub architecture to educational institutions.

REFERENCES

[1] Ajit Singh, 2019. A Data Science Foundation White Paper. Data Science Foundation, Copyright 2016 – 2017, www.datascience.foundation

[2] Angela Bonifati, Fabiano Cattaneo, et al. 2001. Designing data marts for data warehouses, ACM Transactions on Software Engineering and methodology. DOI:https://doi.org/10.1145/384189.384190

[3] Carlos Maroto, 2016. A Data Lake Architecture with Hadoop and Open Source Search Engines.

[4] Cassandra Balentine, 2019. The Modern Data Lake, *Software Magazine.*

[5] David Kieffer, 2019. Building an Enterprise Analytics and BI Practice in Higher Education.

[6] Eran Levy, 2018. Understanding Data Lakes and Data Lake Platforms. https://www.upsolver.com/blog/understanding-data-lakes-and-data-lake-platforms

[7] Golfarelli, Rizzi, 2018. Data Warehouse Design: Modern Principles and Methodologies (ISBN-13: 978-0071610391).

[8] Geoffrey Craig, 2016. What Is the Difference Between Data Lakes, Data Marts, Data Swamps, And Data Cubes? https://intersog.com/blog/what-is-the-difference-between-data-lakes-data-marts-data-swamps-and-data-cubes/

[9] IBM 2016. IBM Industry Models. IBM Industry Model support for a data lake architecture Version 1.0, Copyright IBM Corporation 2016.

[10] Jennifer Zaino, 2020. Data Lakes Prove Key to Modern Data Platforms, https://biztechmagazine.com/article/2019/02/data-lakes-prove-key-modern-data-platforms-perfcon

[11] Jennifer Zaino, 2017. Data Lakes Prove Key to Modern Data Platforms.

[12] Joe Krayanak, David Baum, 2020. Cloud Data Warehousing Dummies, 2nd Snowflake Special Edition, John Wiley & Sons, Inc, 2020.

[13] Joerg Stephan, 2020. The Charm of Security-Driven Data Lake Architecture.

[14] Joey Jablonski, 2019. Building a Platform for Machine Learning and Analytics. https://www.cloudtp.com/doppler/building-platform-machine-learning-analytics/

[15] Kiryl Halozyn, 2020. Data Lake Architecture.

[16] Michelle Knight, 2018. Data Lake vs. Data Swamp: Leveraging Enterprise Data. https://www.dataversity.net/data-lake-vs-data-swamp-leveraging-enterprise-data/#

[17] Nate Feldmann, 2020. Data Lake or data swamp? https://www.nvisia.com/insights/data-swamp

[18] Neil Stokes, 2019. Architecting Your Data Lake for Flexibility.https://www.datanami.com/2019/10/08/architecting-your-data-lake-for-flexibility/

[19] Pablo Álvarez, 2018. Architecting a Virtual Data Lake.

[20] Paweł Mitruś, 2020. Understanding Data Lakes and Data Lake Platforms, 2018. https://www.upsolver.com/blog/understanding-data-lakes-and-data-lake-platforms

[21] Pradeep Menon, 2017. Demystifying Data Lake Architecture

[22] Rick Mullin, 2018. How pharmaceutical research is navigating the data lakem, A trend in large-scale data storage rigs cloud computing with advanced analytic software, Informatics, 96(40).

[23] Sen S, Datta D, & Chaki N, 2012. "An Architecture to Maintain Materialized View in Cloud Computing Environment for OLAP Processing," *Intel. Conf. on Computing Sciences*, 360-365, DOI: 10.1109/ICCS.2012.13.

[24] Sherry Tiao, 2018. Machine Learning and the Modern Data Lake.

[25] Snowflake 2020. Build a true data lake with a cloud data platform, A single source of truth that is secure,

governed, and fast, Snowflake, Inc. http://www.snowflake.com.

[26] Sudi Bhattacharya, Neal Matthews, 2019. Enterprise data lake architecture: What to consider when designing.

[27] Thomas Spicer, 2019. Data Lakes? Big myths about architecture, strategy and analytics. www.openbridge.com.

[28] Webber K. L, Zheng H. 2019. Data Analytics and the imperatives for data-informed decision-making in higher education. Research Projects Series, 2019-004.

[29] Wes Prichard, 2018. Four Data Lake Solution Patterns for Big Data Use Cases. https://blogs.oracle.com/bigdata/data-lake-solution-patterns-use-cases.

[30] Youssra El Harrab, 2020. How to differentiate a Data Hub, a Data Lake and a Data Warehouse. https://blog.semarchy.com/how-to-differentiate-a-data-hub-a-data-lake-and-a-data-warehouse.

[31] Bill Schmarzo, 2014. What Universities Can Learn from Big Data – Higher Education Analytics

[32] Larry Dignan, 2018. Why AI and machine learning are driving data lakes to data hubs.