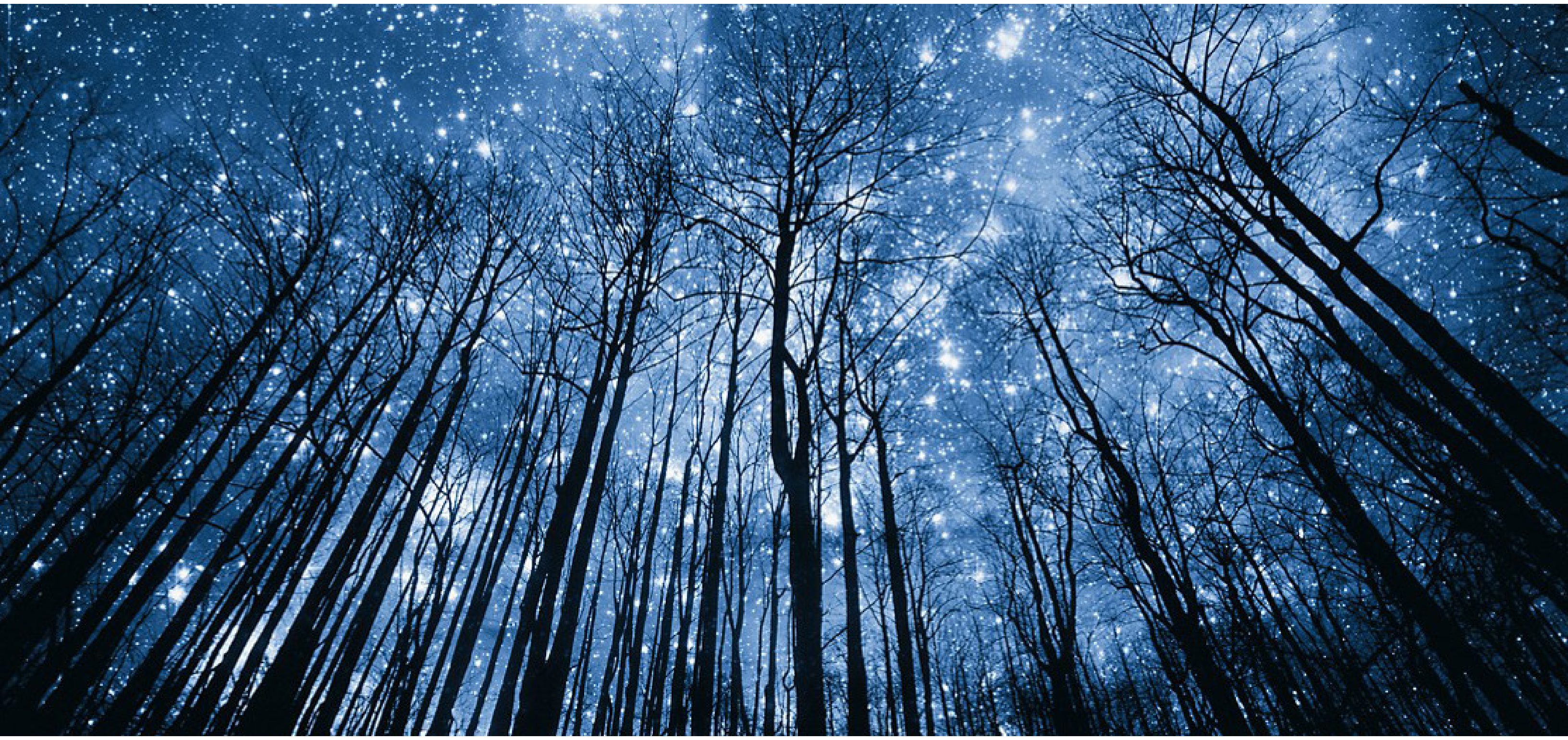
DEMYSTIFYING DATA LAKEHOUSE - A NEW PARADIGM!



Abdul Mannan
Senior Systems Engineer
Dell Technologies
Abdul.mannan@dell.com

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world’s most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Introduction and Brief History

In 1902, Julia Davis Chandler published a culinary recipe in "*The Boston Cooking School Magazine of Culinary Science & Domestic Economics*" that had an enormous effect on our taste buds. It was the first recipe for a peanut butter and jelly sandwich. I think we can all agree that peanut butter and jelly taste delicious on their own; at the same time, it is difficult to argue that they are a fantastic, flavorful combo. What does this have to do with the topic of ***Data Lakehouse***? What are kitchen recipes doing in our data management discussion? Well, the latest buzzword in data management called ***Data Lakehouse*** is also a lethal combo analogous to the peanut butter and jelly recipe wherein we have a combination of the best features of both data warehouse and data lakes, thereby giving us best of both worlds. Before we delve into the nitty gritty of Data Lakehouse, let us first investigate the history of data management systems and how they came to be.

The importance of data and its consolidation for analysis purposes has been recognized for centuries. The need for systems that can provide decision-based roles and functions goes back to times earlier than the first database models. Any system designed for an operational or transactional purpose had its own functions that could not go beyond that purpose for which it had been built. These systems could handle a specific amount of data for only a limited amount of time. Especially, as the historical data was of no value to these operational systems, they were not architected and designed for long-term retention. Over time it became increasingly important to have insight and visibility into the business functioning via data usage. Thus, historical data became more important. This, along with the introduction of RDBMS-based environments (early 1960s), led technicians to find ways and means to transfer and copy the data from these transactional environments to different databases either through manual or automated means and then use them for historical reporting and analysis. Contrary to transactional systems where data are constantly changing, it was not the case in the historical reporting DBs. Instead, their whole purpose was to store as much data as possible. This in turn gave rise to the term "**Data Warehouse**" as these DBs or repositories would act like the warehouse of the data. Bill Inmon, also known as "***the father of Data warehousing,***" coined this term in the early 1970s and had done enormous work on the data modeling and its usage while publishing quite a few research papers on this topic.

In the late 1970s, ACNielsen, a marketing and rating-based company, provided their client a system called "***DataMart***" to improve their sales functions. But the main transformation came in the late 1980s when IBM-based researchers **Paul Murphy** and **Barry Devlin** introduced the idea of "***Business Data warehouse***" (though the same term was also used earlier in 1970s by Bill Inmon). If we go by the abstract of this publication, Paul and Barry had aptly described the current problem and its solution pertaining to the data management systems.

> ***"The operational databases that use the transaction processing systems was the main target of computerization at the beginning. At the same time companies have started using more and more data for their day-to-day reporting and data analysis purposes. Within IBM, the information systems are getting computerized, driven by business needs and the availability of new and improved tools for accessing the data."***

The late 1980s were the age when enterprises began to build decision support systems with the sole purpose of storing data and using that data for reporting. This practice became prominent and was commonly referred to as “***Data warehousing***”. With the rapid improvement and advancements in these systems from both a configuration and performance points of view during the late 1990s and early 2000s, data warehouse started becoming part and parcel of all the enterprise systems, especially their core IT group. The importance was of such a high value that most vendors (Teradata, Vertica NCR, etc.) created customized solutions in terms of hardware, etc. to manage data warehouse applications and systems for different customers. Data warehouse started to become a high priority for all the high-profile companies and businesses.

Data warehouse was yielded to collect the structured enterprise data under one roof

The rise in Internet and online systems ushered in dramatic growth of data creation. A single DB was not sufficient to accommodate the huge data inflow. Consequently, organizations resorted to multiple DBs, sometimes each DB specific to an organized line of business within an enterprise. Thus, there were multiple DBs each dedicated to different types of data, but all disconnected from each other, giving rise to data silos and decentralization of data across the organization. This decentralization of data and hence the inability of converting the information in these data silos into actionable insights that would help the business grow and operate in an efficient and effective manner led to the **Data warehouse**, which united the disarrayed DBs across the company.

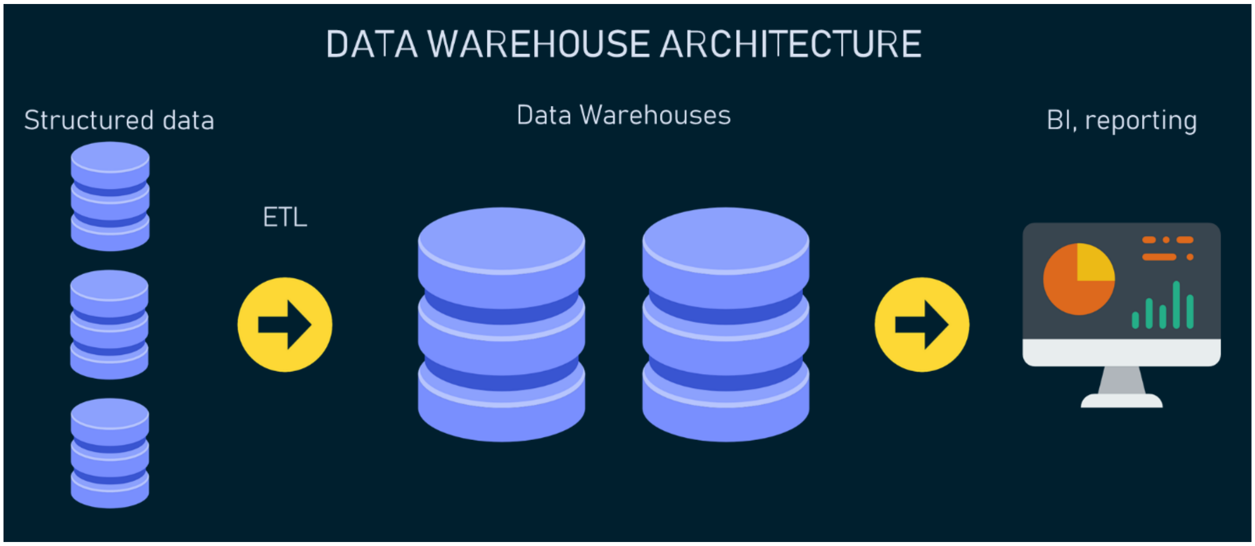


Figure 1: Data Warehouse Architecture

Data warehouse-based systems came to the rescue by collating the disparate collection of relational databases under a single roof with the ability of querying and viewing the data as a whole. Beginning in the 1990s, data warehouse became the most prevalent data architecture for big organizations. The main advantages of this architecture included:

- Ability to **integrate** different data sources.
- Data **optimization**, in particular for read access.
- Ability to run **fast and high-performance** analytical queries.
- Data **reporting**, **auditing** and **governance**.

The importance of these systems was also realized in enterprises when different departments in an organization were able to gain insights and inferences that can be actioned into business decisions, which in turn helps the organizations to drive operational excellence and create additional revenue generating opportunities.

Data warehouse fulfilled much of the needs, but over time the shortcomings of this technology became increasingly evident. As the internet flourished and with the introduction of Internet of Things (IOT) based devices there was a data boom. With millions of IoT based new devices getting added to the ecosystems every day and each of them creating different types of data. And as such data warehouse-based systems were unable to withstand this revolution of data and hence unable to store the raw and unstructured data generated by these devices. At the same time this architecture was based on propriety based expensive hardware and software systems, making it difficult to scale the systems especially due to the tight coupling between the computational power and the storage.

Apache Hadoop facilitated unstructured data analysis and thereby lead to the data lakes

With the advent of "***Big Data***" in the earlier part of this century, the importance of data grew more and more, as did the volume of data generated. Aptly, the term Big Data corresponded to the 3 big V's of **Velocity**, **Veracity** and **Volume** (this later changed to Big 5's) that defined the state of data generation during this period. One of the primary causes of this was the increase in the no. of devices connected to the internet and the realization of organizations of the plethora of information that was hidden in this data which could eventually lead them to the generation of more revenue by helping them to make better decisions and make the functioning of these organizations effective and efficient. But this huge amount of data at hand and the demanding computational power made these companies realize that a single host machine or server was not enough to carry out these analytical tasks. Furthermore, the format and type of the data that needed to be analyzed was not neat and clean all the time- these organizations required to use the unstructured data as well. To fill all these gaps and to overcome the concerns of high cost and vendor monopoly of data warehouse, **Apache Hadoop**, an open-source distribution system came to the forefront.

Apache Hadoop is often defined as ***the open-source architecture apt for big data analytical and reporting tasks that processes large sets of data using parallel combination of clusters of high computational computers***. It consists of 3 main entities including **Hadoop Distributed File systems (HDFS)-** *allows the data transparently to be spread across different storage devices*, **Hadoop Map Reduce-** *the algorithm that powers the analytical tasks deciding how to split and divide a computational task into smaller ones to be run parallel across the computers* and **Yet Another Resource Negotiator (YARN)**- *helps to process the batch and streaming data across the HDFS.* The innovation of Hadoop was revolutionizing in itself especially due to 2 main factors. Firstly, it broke the organizations from the shackles of propriety architecture of hardware and software of data warehouse and secondly, it made it feasible to process and analyze huge amounts of data (both structured and unstructured) which was not possible in the past. With organizations having now the capabilities and means of processing raw, unstructured and structured data and realizing the importance of information this data is drenched in, collecting and storing it became of utmost vitality and thereby giving rise to the concept of ***Data Lakes***.

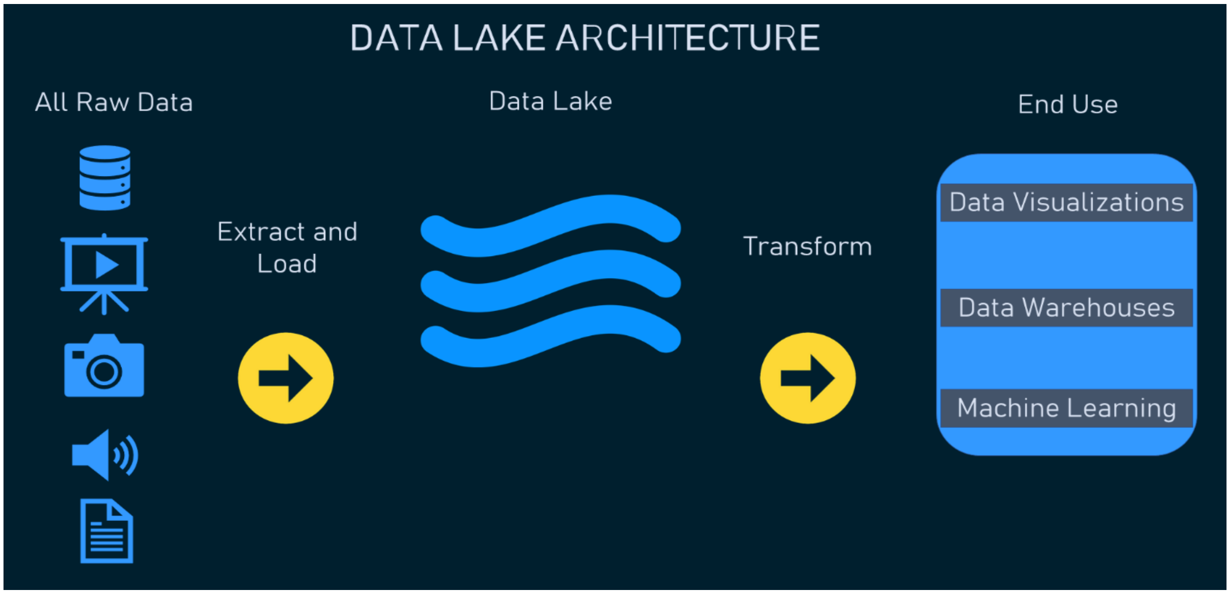


Figure 2: Data Lake Architecture

Data lakes *centralized and consolidated the raw, unstructured, semi-structured or for that matter even the structured data, collected from multiple and varied sources even without any predefined schema or structure*. Compared to the legacy data warehouse systems, data lakes architecture was quite different especially owing to the fact that data warehouse stored data in hierarchical manner in the form of folders and files while as data lakes architecture used flat object-based storage format to store data. Here, object-based storage was unique due to the fact that it stores data with tagged metadata and unique identifiers. This helped to identify and retrieve the data from the huge collection of data spread across different storages in a fast and highly performance-oriented manner. In addition to this, Data lakes enabled applications to take advantage of the data populating less expensive open format object storages as compared to its predecessor (Data Warehouse). With Data lakes, organizations began to consolidate their data in a centralized location in its raw format irrespective of data source or the schema. The best part about data lakes was that structured data can be stored alongside the raw unformatted data with the ability to process all kinds of data including document, audio files, video files or images- all important in today age of Machine learning and data analytics.

Another important benefit of data lake is its **agility.** Due to the lack of schema or structure, it is much easier to carry out modifications in the models and their respective queries. In this case, Data lakes are more flexible and agile as reconfigurations can be done as per the needs while it was more difficult and time consuming to do changes in the schema in the case of Data warehouse. Furthermore, it eliminated the problem of data silos which earlier suffered with duplication of data, implementation of security policies on different data storages which in turn used to make the collaboration difficult. Data lakes also provided a common landing pad for the new data, thereby keeping it always up to date by collecting and storing any type of data indefinitely. From the user's perspective data lakes was a complete package as it offered them all sets of tools and languages to perform all types of analytical and reporting tasks.

Shortcomings of the current Data Lake architecture and the need for change

As the saying goes, "**nothing is perfect**," Same was the case with Data lakes. As more enterprises started using data lakes, they realized that it was missing some of the critical and vital features such as lack of implementation of **data governance** or **data quality**, **mediocre performance optimization** and **no support for transactions**. In addition to these enterprises realized that with the increase in the size of data in the data lakes, the performance of queries suffered especially due to metadata handling and lack of proper partitioning of the data. Data governance was another aspect where the data lake was missing the point as it had no clarity on data updating or data deletion. This made compliance and regulatory requirements difficult to achieve for these organizations. Furthermore, storing the data in data lakes without any oversight of its contents and the need to have a process to arrange and secure this data resulted in an unsecure and ungoverned "**Data Swamp**."

With individual challenges of both Data warehouse and data lakes, the best option available was to use multiple data systems comprising of an ecosystems of data lakes and multiple data warehouses. Some of the major challenges faced by the above 2-tiered systems were as below:

Data Reliability and Staleness

Storing consistent data in data lakes and data warehouse is not a straightforward process. It is costly and complex at the same time. ETL'ing the data between the 2 systems for high performance business intelligence and decision support requires heavy engineering resources. Such multi-tiered ETL processes are susceptible to bugs and failures especially due to the varied difference and disconnect between the data warehouse and data lakes engines.

These systems also suffered from ***data staleness,*** especially for Data warehouse where the new data takes days to load into the system. This was averse to the idea of data analytics on live and fresh data.

Lack of support for Machine Learning and Advanced Analytics Toolsets

With the rise in Machine learning and its related tools, the business expectations also increased. For meeting these expectation, advanced machine learning toolsets were the answer, but none of these tools including XGBoost, TensorFlor or PyTorch were fully compatible with this setup. This was the case because these tools work on large datasets and execute complex SQL-based queries. A workaround of running these tools dedicatedly on the data lakes datasets resulted on missing other data management features of indexing, data versioning or ACID transactions.

Propriety Formats

In addition to this, Data warehouse characteristically locks the data in specific formats which in turn makes it difficult to migrate to other systems and hence makes the migration a costly affair. Moreover, in case we choose the option of directly accessing the data via SQL queries, it makes it very slow and expensive.

Due all these factors, legacy data lakes systems are not enough to fulfil the needs of the organizations that are looking into the implementation of new and leading-edge innovative technologies, thereby forcing them to operate in complicated architectural environments. Simplifying these complex architectures is the main motive of all these organizations that shoot for harnessing the power of modern-day technologies like Data Analytics and Machine Learning.

Data Lakehouse – The best of both of worlds

The solution to the above challenges was in the introduction of state of art architecture of ***Data Lakehouse***. Data Lakehouse overcame the shortcomings of Data lakes by incorporating the transactional layer at the top of the hierarchy. **Data Lakehouse, as the name suggests, is a combination of Data warehouse and the Data lakes by exploiting features of data management and data structure from the Data warehouse architecture and operating them on cloud-based data lakes.** This combination of Data Lakehouse architecture ensured that modern day machine learning, analytics and data science approaches all coexisted in an open format and a common ecosystem.

With the best of both worlds, Data Lakehouse introduces new set of use cases covering enterprise scale **Machine Learning**, **Business Intelligence** (**BI**) and **Data Analytical** projects helping the organizations to unlock the informational treasure in the data, and thereby creating a monumental business value for these organizations. Data Lakehouse has opened new opportunities for all kind of technology experts with data analysts now being able to gain insights from the data by using SQL based query on the data residing in Data Lakehouse, Machine learning (ML) engineers can create ML based models with better performance and accuracy while BI experts can create dynamic visual dashboards using faster and simpler toolsets. All this can be executed simultaneously on Data Lakehouse without any modification in the data structure and with constant streams of new data coming in from different and varied sources.

Why and what of Data Lakehouse?

It was Databricks that was the main proponent and introduced the concept of Data Lakehouse to the technological world in the year 2020 (though AWS in 2019 started to use the terms “Data house” with respect to the server Amazon redshift Spectrum that primarily enabled users of AWS data warehouse service to execute commands and queries on the data stored in AWS S3 service). As per Databricks, **Data Lakehouse** can be defined as “***a data platform that merges the best characteristics of data lakes and data warehouse- with performance and data management typical of data warehouse while at the same time enjoying the low-cost and flexible object-based storage characteristic of Data Lakes***.” Owing to the fact that more than 90% of the enterprise data nowadays resides on Data lakes, the foundation of Data Lakehouse is built on data lakes itself. Thereby eliminating the process of copying the data subsets to a siloed data warehouse ecosystem to run legacy BI toolsets.

Data Lakehouse architecture is open and enabled by a new design that has copied the data mgmt. and data structure architecture from the data warehouse running on low-cost cloud based open formatted storage. In other words, if we would have been given an opportunity today to redesign and reconstruct the data warehouse architecture in the preset modern world with highly reliable and cheap storage, we would get something similar to the Data Lakehouse. Here, the idea was to introduce the benefits of structured analytics of data warehouse to the data stored in low-cost and cheap cloud-based data lakes, particularly true for data types that are not relational in nature.

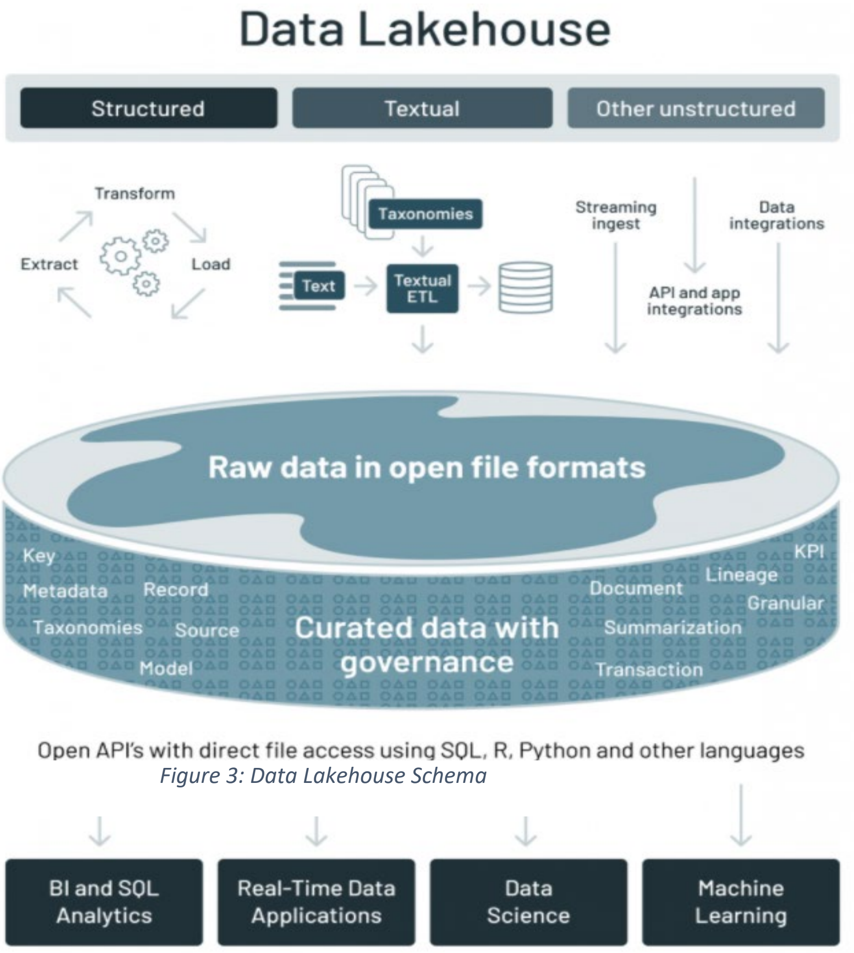


Figure 3: Data Lakehouse Schema

Compared to data warehouse, Data lakes had features of agility and flexibility that were suitable to the ever-evolving business and data but still suffered to create a foot hold due to the missing data engineering processes and governance, thereby resulting in its inaccessibility to the business owners and general users. Data Lakehouse came to the rescue with the boundaries between data lakes and data warehouse being blurred by keeping the promise of low cost, persistent and flexible data storage of the cloud while at the same time maintaining the data structure for an advanced and stimulated business decisions and analysis.

Open Source

Data Lakehouse is based on open-sourced architecture built on standardized and open formatted files including Optimized Row Columnar (ORC) and Apache Parquet that facilitates the access of data directly via open APIs without the need of vendor locked engines. In addition to SQL, it also supports other tools and languages for Machine Learning and Python libraries.

Compatibility with Modern Day Tools

Data Lakehouse can provide effective and efficient support for a variety of data types including videos, audio files textual data or structured data for applications to process. This includes direct access of huge chunks of data to be processed by non-SQL applications and Machine Learning tools and libraries. To further facilitate the Machine Learning systems, Data Lakehouse also provides APIs for the data abstraction and preparation using Dataframes that can be used to access and query data via ML tools like Pytorch, XGBoost and TensorFlow.

Data Lakehouse also provides versioning capabilities for ML and Data science teams to revert to earlier versions of data if needed. This is also important for audits and roll back procedures.

Reliability, Cost and Performance

With Data Lakehouse, new performance-oriented optimization features were introduced that included caching, data skipping, clustering. It also supports robust auditing and governance mechanisms that leverage Atomicity, Consistency, Reliability and Durability (ACID) based transactions for the purpose of consistency as different users read and write the data in a concurrent manner.

Built on a low-cost architecture, Data Lakehouse uses cloud-based storage including Amazon S3, Google Cloud storage and Azure Blobs with cost efficient models that further lower the cost of the overall system.

Data Lakehouse – Architecture at-a-glance

It is a known fact that it is always beneficial for the enterprise to collate and collect their entire data from all the sources to a common ground. This way they can get greater and deeper insights into the data at hand. The past practice was to either store the data in silos of datastores in order to get quick outputs on SQL based queries on primarily structured data- this was the case of **Data warehouse systems** or the other option was to collect the vast data from different silos and converge it to a single store in order to carry out Machine Learning and Data Analytics based operations- this was the case of **Data Lakes**. Now, to move the data across these ecosystems, it has become a complicated, cumbersome process with lots of hurdles and to overcome this issue and facilitate the data movement, Data Lakehouse comes to the rescue.

As far as the architecture of Data Lakehouse is concerned, it is built on the requirement of integrating and merging the dispersed data warehouse, data lakes and other data management services into a single coherent entity. Data Lakehouse architecture allows us to have a unique space on which we can execute our data analytics and machine learning based jobs while at the same time give utmost performance, thereby enabling us to have live dashboards and real time reporting use cases.

To discuss the architecture of Data Lakehouse, we must know that AWS and Databricks have been two of the main advocates of Data Lakehouse, so while discussing the architecture, we will consider both vendors. Below are some of the key elements on which the foundation of Data Lakehouse is laid:

- Data Services that are purpose built.
- Scalability
- High performance and low cost
- Smooth Data Movement
- Unified Governance.

As far as the real-world approach to customer data is considered, the data needs to be mobile and able to move across different data stores and data analytical services. This data movement can be ***an Inside-out movement*** defined by customers storing their data in data lakes and then moving a subset of that data to a specific store for the purpose of performing data analytical or machine learning jobs. Or on the other hand it can be ***outside-in movement***, wherein customer data stored in entities like data warehouse is moved to a data lake to execute data analytical jobs on it. In other cases, it can be ***around the perimeter***, which integrates the data warehouse, data lakes and its data stores in a seamless and smooth manner. The below diagram helps us in understanding the above data movement cases in a better manner.

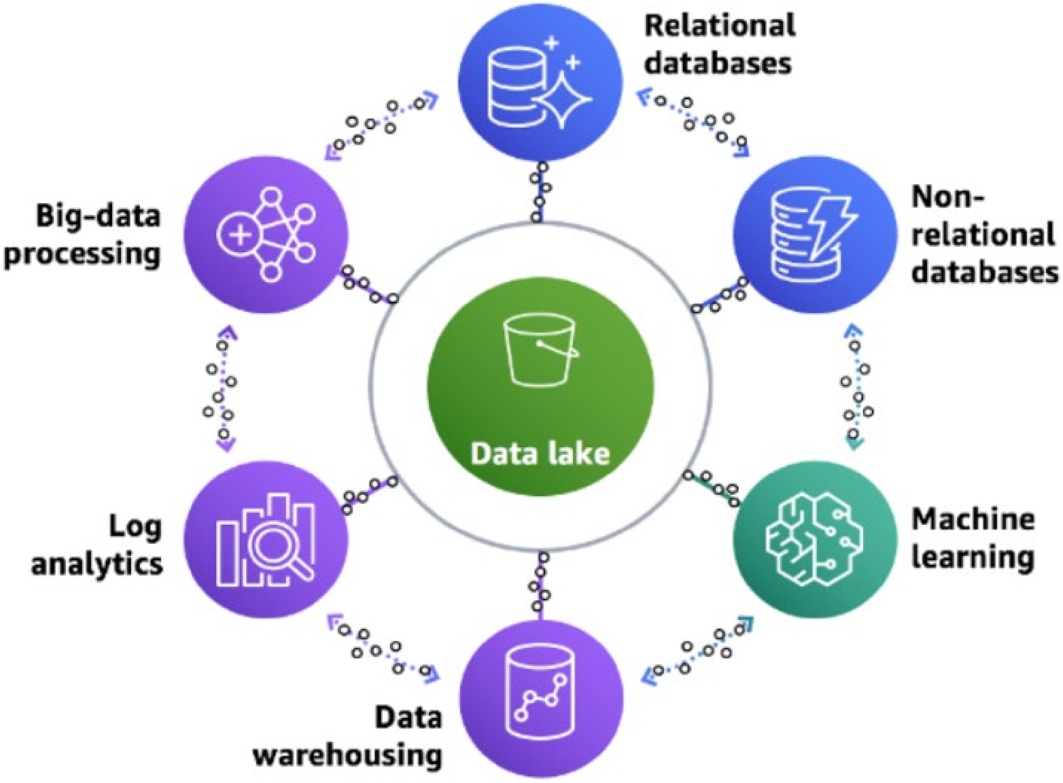


Figure 4: Data Movement and Uses

Before delving into the architectural aspect of Data Lakehouse, we need to understand the Data sources and their functionality. One of the foremost features of Data Lakehouse is its ability to operate on data from various sources. These sources represent different business applications within an organization including Enterprise Resource Planning (ERP) applications, Line of Business (LOB) applications or even Customer Relationship management (CRM) based data that is defined by a structured format. Apart from structured formatted data types, unstructured or semi structured data types from web-based sources like social media and IOT devices like mobiles, sensors, edge devices etc. can also be a major source of data for Data Lakehouse.

The architecture of Data Lakehouse is based on the idea of layering and componentization, which facilitates us to use the right approach and tools for the right tasks and at the same time ensure that we can build the system architecture with agility and incrementally. This ensures that we have the flexibility and agility to fulfil current and future demands and needs as new use cases are introduced resulting in new data sources with new requirements and new analytical methods. Considering this, the Data Lakehouse architecture is composed and arranged in stack of layers, with each layer having a specific purpose and set of components built to address specific use cases and requirements.

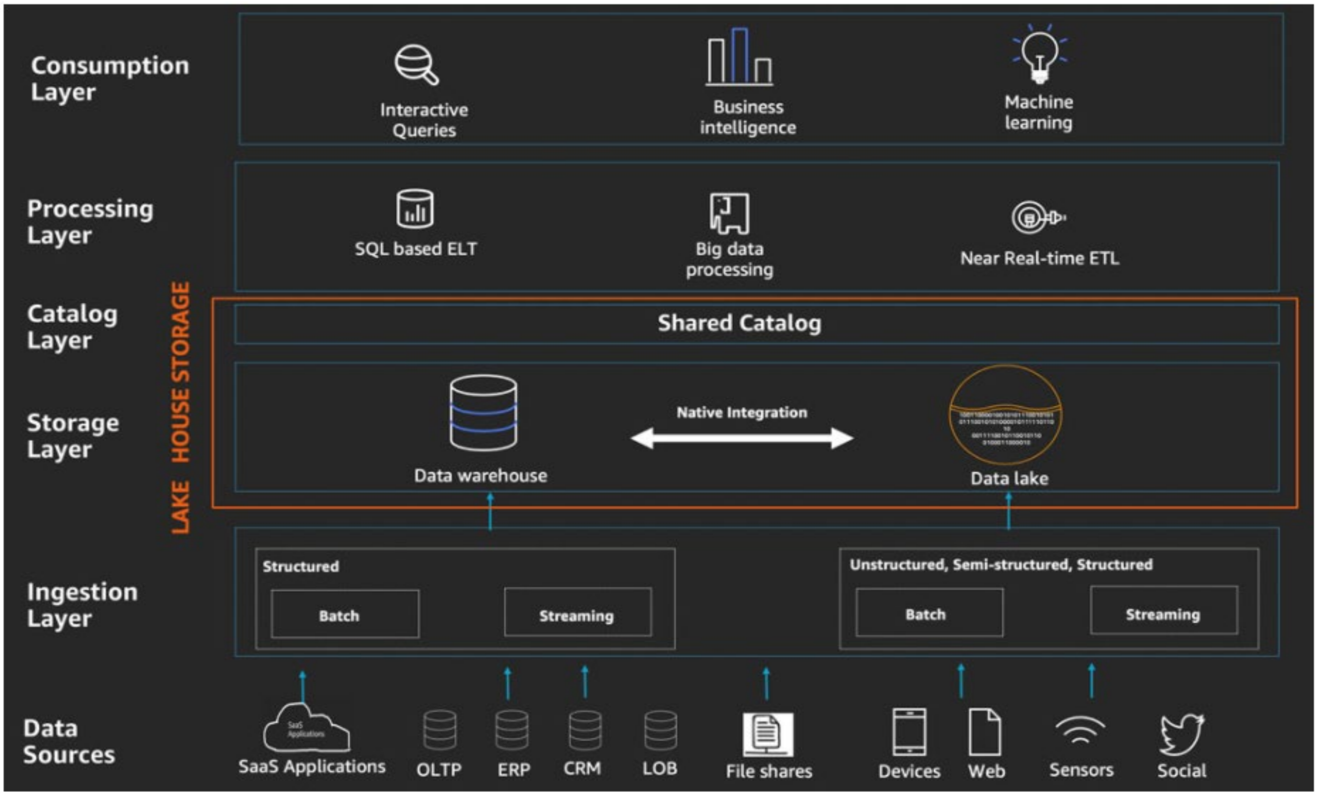


Figure 5: Data Lakehouse layer-based Architecture

Let us discuss each component layer that comprises the modern-day Data Lakehouse:

Data Ingestion Layer

Ingesting data into the Data Lakehouse is the first and the foremost step in the data pipeline and primarily is categorized under two main categories viz streaming data and batch processed data. Usually, batch data is loaded into the Data Lakehouse using powerful compute and storage-based clusters that make it possible to ingest petabytes of data within very less time. All this is happening and is processed in the ingestion layer of the data lake ecosystem.

Ingestion layer is ***the first layer of the Data Lake stack and is primarily responsible for the intake of the data from different data sources and then presenting it to the storage layer***.

With capabilities that include consolidating the batch and streaming processes, different connectivity protocols are used to connect to different internal and external data sources that include RDMs, CRM applications, file shares, social media, IOT devices, websites, databases, and SaaS based applications. The layered approached architecture ensures that the data is accessible for queries as soon as it lands on the Data Lakehouse. It also ensures that the most recent data is on the top and readily available.

Components including Apache Kafka for data streaming, Data Migration Service (DMS) from AWS for importing the data etc. are vital and important in this phase.

Storage Layer

The main objective of data storage layer is to provide sustainable, low cost, scalable and durable storage for huge quantity of data to reside. data warehouse and data lakes integrate natively together to accomplish a cost effective and sustainable layer of storage that is compatible with all kinds of data formats including structured, unstructured and raw data. It is architected in such a way that it can store the data as objects in low-cost object stores making it feasible for users to directly access these objects using open file formats, this facilitates the other layer components to access and use the data at the same time.

The data stored in data warehouse stores is characterized by a proper structure and trusted data source from internal or external sources including transactional or relational DB systems. These data stores are high performance storages having the ability to store exabytes of data in compressed columnar format providing ultra-low latency to the data queries owing to the Multi parallel processing (MPP) transactions. At the same time, data lakes conform to a centralized repository of Enterprise data supporting all formats of data including structured, semi-structured and unstructured one and with the ability to scale in an automated manner up to petabytes of capacity. Depending on the consumption readiness of the data at hand, data lakes are segmented into different zones including curated, trusted, raw and landing zone. The ingestion of the data in data lakes happens irrespective of the data schema, this helps to lower the time needed for ingestion. Implementation of big data methodologies of Machine Learning and data preparation and data processing facilitates the analysis of the varied data stores.

This native integration between these two entities results in reduction of the overall storage cost by facilitating the move of large amount of cold, archivable historical data from data warehouse storage to low-cost data lakes.

Metadata Layer (Catalog Layer)

Metadata layer, often called Catalog layer, is one of the core layers in the architecture of Data Lakehouse instrumental in setting it apart compared to its predecessors. This layer is responsible for storing the metadata (both technical and business) of the data hosted in the storage layer. In other words, it is an integrated catalog that holds the metadata for all the objects residing in the data lakes and is shared by both data warehouse as well as data lakes. Enabling the same queries to be run simultaneously on data residing on both data warehouse as well as data lakes. This cataloging is the backbone of the below features that feature in the Data Lakehouse:

- ACID based queries to facilitate concurrent transactions with a consistent database image.
- Caching for files from cloud store
- Zero-copy cloning
- Indexing
- Data versioning

In addition to this, the catalog layer also improves the efficiency of the entire data pipeline by facilitating in application and implementation of the Data warehouse schemas and auditing and governance functionalities directly on the data lakes. Part of these functionalities include the ***scheme enforcement*** by virtue to which any writes that are not compatible with the schema of the tables are rejected to ensure data quality and integrity. It also includes ***schema evolution*** that allows the schema dynamic change in the schemas with the change in data format along with a single interface for the management and auditing purpose.

Databricks with their Delta lakes and Apache Iceberg have already exploited these capabilities for their optimization and performance enhancements.

The close integration between the data warehouse and data lakes at both storage as well as metadata layer level helps in providing a common interface of Data Lakehouse to the next two layers of processing and consumption. Unified interfaces like SQL and Spark can facilitate the consumption of data stored in the Data Lakehouse storage layer by the processing and consumption layers. This avoids the data movement between data warehouse and data lakes and enables direct and transparent data access of the data in the Data Lakehouse. This integration between the two data storing entities provides the following features:

- Storing of data in all formats in cost effective and flexible data lakes while as hot and structured data is stored in data warehouse.
- Irrespective of the type of data, whether it is structured or unstructured, it relies on a single framework that can prepare and analyze the data in a common pipeline.
- Combination of flat RDBMS based data in data warehouse with the hierarchical data in data lakes to create native ELT or ETL based pipeline.

Data Processing Layer

Data transformation into a state that is consumable via normalization, cleanup, validation and enrichment is the main attribute of this layer. This layer provides components that are specially designed to provide data transformations that include big data processing, real time ETL jobs etc. This layer is responsible for providing the right tools and components that match the characteristics like size, format, structure, current tasks etc. This layer is responsible for enabling the handling of large data chunks in a cost-effective manner and at the same time fulfill the requirements for supporting the portioning of the datasets, on read, on write process of the schema with varied data formats. The processing layer can access the storage layer (Data) and the catalog layer (Metadata) by having direct contact with the unified Data Lakehouse storage interfaces. Its advantages are as below:

- Undue and unneeded data movements, data duplication and ETL code redundancies are avoided, especially when working on data warehouse and data lakes separately.
- Time to market is considerably lowered.

Data Consumption Layer

This layer is responsible for offering the components that are scalable and performance oriented and using the Data Lakehouse interfaces to access the data and metadata that has been stored in the storage layer and metadata layer respectively. This layer hosts analytics and visualization tools like Tableau, Power BI etc. to facilitate ML based analytics, dashboard for visualization and interactive SQL queries.

The different components in this layer are enabled to:

- Carry out jobs based on Machine Learning and Data Analytics that access and consume data from both data warehouse-based schemas as well as tables hosted in data lakes.
- Compatibility with open file formatted datasets including ORC, Parquet and Avro and their storage.
- Ability to partition and prune large datasets hosted in data lakes and thereby optimize the cost and elevate the performance.

AWS based Data Lakehouse Architecture and Services

The mantra of most of the organizations and enterprises is to reduce costs and at the same time increase their efficiency by embracing modern day applications. These applications are defined by 24x7 round the globe access, data processing involving large volumes of critical storage, scale up and scale down at the push of the button and ensuring high availability and performance. One of the most important components in this modernization of applications and infrastructure revolution is the drift from monolithic to microservices based applications. This shift allows application modules and their services to develop, deploy and be managed as independent entities. The move from monolithic nature to microservice-based applications has lots of advantages but comes with a tradeoff, complicating the overall data architecture of the ecosystem. For this Data Lakehouse comes to the rescue and presents a precise and perfect architecture that inculcates decentralized microservices and thereby ensures smooth data movement especially between the data warehouse to data lakes and vice versa.

Amazon Web service (AWS) has been one of the main proponents of Data Lakehouse architecture and first started using the term '**Lakehouse**' for its service (Amazon Redshift Spectrum) that was used to execute queries from Amazon Redshift based Data warehouse to the data that was stored in the Amazon S3 storage. As the use cases aligned to this structure started increasing, AWS widened its preview beyond the Data warehouse case. This resulted in a full-fledged AWS based Data Lakehouse architecture with different AWS services and components playing a vital role in each and every layer of the Lakehouse.

Considering this and the importance of AWS based services in the current ecosystem, we will be discussing how different AWS based services help in creating robust reference architecture for consumption.

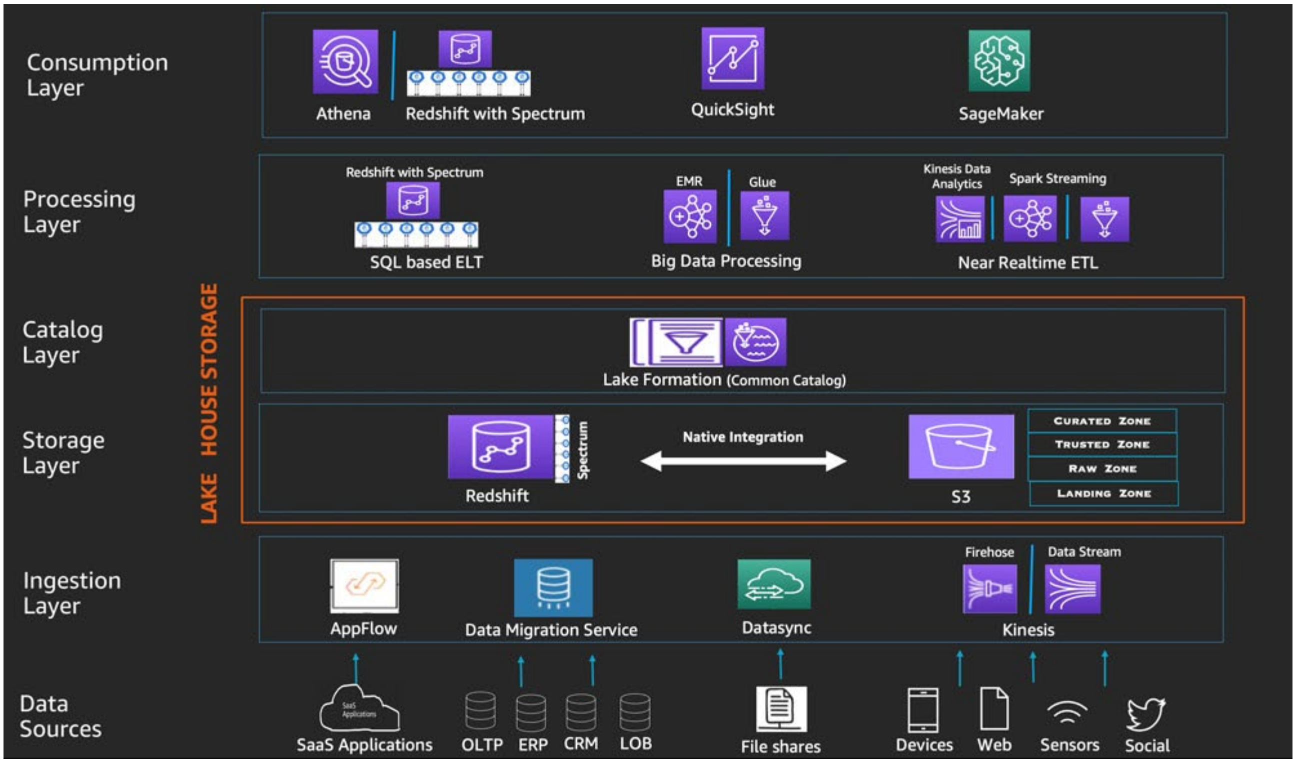


Figure 6: AWS Services integration with Data Lakehouse

Ingestion layer

As far as the data sources are concerned, these are varied and different, and accordingly there are multiple and varied AWS based services that cater to specific type of sources to ingest data into the Data Lakehouse storage, serving data directly to the data warehouse and data lake datastores. These AWS based services are compiled based on different properties of the data sources including the format of the data, its connectivity, the speed at which the data is getting ingested and above all the structure of the data at hand.

AWS-based Data Migration Service (DMS) connects and ingests data from databases that can be relational DBs (RDBMS) or NoSQL-based databases. DMS is used to copy or ingest data by creating an import session at the initial phase while as the changes are replicated continuously to the Amazon S3 storage whether in data warehouse or data lake datasets.

As far as the ingestion of SaaS application-based data into the storage layer of AWS based Data Lakehouse is concerned, ***Amazon AppFlow*** is the most appropriate service that can be used for this use case. AppFlow is an AWS based fully managed service that facilitates the data transfer from applications like SAP, Salesforce or ServiceNow to AWS based storages like S3 in a secure and smooth manner. Here the AppFlow service can directly connect to the SaaS applications and ingest data and at the same time transfer it to the storage layer comprising of S3 based datastores either in the data lake or the AWS Redshift data warehouse tables. To automate the data ingestion process from these applications to the storage layer, scheduling or event-based triggers can also be used.

Another AWS based service called ***AWS DataSync*** can be used to ingest file-based data that is hosted on NAS based filesystems either in structured or unstructured forms. This helps us to ingest petabytes of data from these file shares (NFS or SMB) into the storage layer of Data Lakehouse ecosystem. DataSync as a service is intelligent enough to handle scheduling, scripting or the network performance optimization tasks in an automated manner by initially doing the one-time bulk transfer of the data and then based on the monitoring and data validation confirmation syncs only the changes into the Data Lakehouse system.

One of the major and important portions of the modern data production ecosystem is from the streaming data. Streaming data is produced by sources as varied as IOT based devices to infrastructure logs to click stream application to OTT based services like Netflix, Amazon Prime etc. In order to facilitate the ingestion of such data into the Data Lakehouse schema, AWS based ***Kinesis Data Firehouse*** is the best fit, with the ability to collect and load the data into the storage layer and ensure near real-time analytics with visualization-based dashboards and BI Tools. The service is efficient and effective, especially because it buffers the ingested data into batches by compressing and encrypting the data into S3-based objects to Data Lakehouse-based Data lakes or Redshift Data warehouse datastores. From the administration point of view, this service requires no administration as it is a low cost serverless service where we pay only for the data that has been transferred and processed. Kinesis Data Firehouse can scale up and down based on the amount of data being handled.

Storage Layer

Two of the main components of the AWS Data Lakehouse storage layer are **Amazon Redshift** and **Amazon S3**, both of which incorporate a unified storage layer. While as Redshift corresponds to the structured data that is highly secure, curated into standard schemes (Redshift can now store Semi-structured data as well), S3 based storage provides storages for all forms of open formatted data including structured, semi-structured or unstructured data forms. With Amazon S3 and open formatted files support, multiple processing of the files can happen at the same time.

Combining **AWS Redshift** with **AWS Redshift Spectrum** facilitates a common SQL based interface that can process the SQL queries that can be executed on the combined datastores of data lakes and data warehouse datasets. AWS Redshift Spectrum can be used to build pipelines that can ingest only the recent hot data into the data warehouse systems while keeping the huge chuck of old and cold data in the data lake systems, thereby processing the data both hot as well as cold at the same time without the need to move or transfer the data across any of the datastores. AWS redshift Spectrum can also discharge big volumes of cold data from the data warehouse storage to the less costly data lake storage without any compromise on the performance or execution time of Redshift queries.

As far as the way data is stored in AWS Redshift is concerned, it is characterized by a highly structured and compressed manner with columnar and distributed format across high performance nodes configured in a redundant and highly available cluster with each node capable of providing up to 64TB of fast and efficient storage. It fulfils the requirements for cases where we need to have highly concurrent, fast and low latency access to the storage for complex queries governed by BI and Data analytics-based Dashboards.

Use cases that need Machine Learning, Big Data and Data science-based processing usually is handled by AWS S3 hosting structured, semi structured and unstructured data while as high performance, trusted and interactive executions are typically powered by structured data in AWS Redshift.

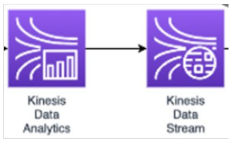
Metadata Layer (Catalog Layer)

In a normal Data Lakehouse ecosystem, hundreds of datasets are hosted by ingesting data from varied sources. Having a central repository governed by a catalog that facilitates metadata-based search enables self-service based discovery process of the data in the Data Lakehouse. Having two separate layers for catalog and storage provides "**schema-on-read**" for the rest of the layers (processing and consumption layers) and their components (like Redshift Spectrum). This metadata layer corresponds to the ***Amazon Lake Formation*** service in AWS that facilitates a central catalog storing entire dataset related metadata irrespective of AWS Redshift or AWS S3 storage of Data Lakehouse. Enterprises use Lake Formation service to store all kinds of metadata including the portioning information, schema of the versioned tables, location of the data, timestamps or the data ownership details and other business-related information.

Due to the constant evolving nature of the data and their corresponding data partitioning in data lakes, AWS has come up with another service called ***AWS Glue Crawlers*** that helps in creating metadata for the data in S3 to be arranged as tabled database and at the same time tracks the constantly changing schemes and partitioning information of the data stored in datasets by updating the lake formation catalog.

Lake Formation enables permissions to be assigned at a granular level of tables and columns so that users and other AWS services have access to only authorized tables for the purpose of processing. This ensures that the data is safe and secure with restricted and need-based permissions only.

Data Processing Layer



There are variety of AWS based services that can be used to exploit different use cases for processing of the data in Data Lakehouse architecture, each component suitable for specific data structure and volume and speed of data that is being ingested from various sources. All these components can read and write the data from AWS S3 or AWS Redshift-based storage datasets for processing different jobs in the processing layer using interfaces. Purpose built components can be exploited to create new pipelines that can be used for data transformation, especially for the implementation of below:

ETL for SQL Based Data (AWS Redshift Spectrum)

SQL based semantics can be used to create the right ETL pipelines for the transformation of the structured data in the storage layer of Data Lakehouse. AWS based services like ***Redshift spectrum*** and ***Redshift*** can be used by these pipelines in an MPP fashion to boost the performance and throughput to facilitate the spin up of large count of nodes for scaling to process the data in rush. This can also be used to process flat and structured data ingested by DMS or AppFlow into the table format of AWS Redshift.

Processing of Big Data (AWS Glue and EMR)

AWS based services like ***AWS Glue*** or ***AWS EMR*** can be used to run Apache Spark based big data processing tasks in cases where large amount of data (structured, semi-structured or unstructured) hosted in Data Lakehouse storage layer (Redshift and S3) is to be processed. Open source as well as Native connectors based on Apache Spark can be used for these jobs including the merger of flat and complex data stored in AWS redshift and data characterized by hierarchal and structured data on AWS S3, with the ability to deliver the processed data back into the S3 or Redshift based data warehouse storage datasets.

ETL functions with pay per use serverless model having the capability to process petabytes of varied data can be provided by services like ***AWS Glue***. This can all be done without the need to install and manage physical or virtual servers or host clusters. AWS Glue not only generates code for ETL but at the same time enhances the overall development process with the intrinsic property of processing the data in both S3 data lake as well as Redshift data warehouse. ***AWS Glue crawlers*** help in populating the common Lake Formation service, which is in turn used by AWS Glue tasks to access Redshift and S3 based datastores. In addition to this, AWS Glue facilitates the process of portioning the data in an incremental manner and at the same time works with triggers and workflows to create an overall data pipeline to process the data.

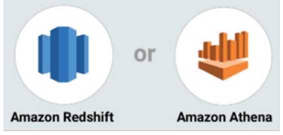
Similarly, AWS-based EMR clusters can auto scale to fulfill the ever-changing demands of the big data pipelines, which uses AWS EC2 spot instances to create a cost- optimized and performance-oriented clusters. The EMR service is a managed platform service used to run big data frameworks including spark and Hadoop to facilitate the data analysis on a larger scale. AWS EMR service via the Spark ETL jobs can connect to the lake formation service to read the schema on Data lakes. This all happens by granting the AWS Glue access to the underlying objects on S3 storage using AWS IAM service.

ETL or Real Time data (AWS Kinesis Data Analytics, Glue and Kinesis Data Firehose)

Modern applications are nowadays relying more and more on real time analysis with live dashboards that are getting updated constantly with the ever-changing data in a real time manner. From the Data Lakehouse perspective, this all happens by ingesting data that is generated at high frequency, which then needs to be validated, prepared, and make it available for use in the data Lakehouse storage. From AWS services point of view, the processing of large quantity of data that is generated in a highly fast manner can be accomplished by ***AWS Kinesis Data Analytics (SQL)***, ***Glue or EMR (Streaming Data)*** and ***Kinesis Data Firehose*** in combination with ***AWS Lamda***.

All the above services help in creating a real time pipeline for data processing without any overhead of managing or creating servers or hosting compute. Data Source throughput is the basis for the auto scaling of services like Kinesis Data Analytics and Data Firehose while AWS EMR and Glue can be scaled using modifications in the operating parameters within no time. AWS Kinesis Data Analytics and Spark streaming based pipelines read records from Kinesis Data Stream (ingestion Layer), transforms the data, rewrite the records on a separate Data stream which in turn connects with Kinesis Data Firehose, that provides data to the AWS S3 and Redshift as part of the storage Layers. AWS Lamda functions in conjunction with these services can make the overall operation simple and more effective by delivering the data in micro batches.

Data Consumption Layer



Consumption of data across the AWS Data Lakehouse consumption layer to facilitate different use cases involving Analytics, Machine Learning and Business Intelligence is enabled by different customized AWS based services that use the unified interfaces of the Data Lakehouse to access the data and catalog layer-based metadata that respectively is available in AWS S3/Redshift and the lake Formation services. These services are flexible in terms of the type of data they can consume, which can include AWS Redshift tables based flat data or objects stored in AWS S3 in the form of complex and unstructured data.

Interactive SQL (SageMaker and SageMaker Studio)

Data Scientist and Data analysists can explore Data Lakehouse storage via interactive SQL queries using AWS services like **Redshift** and **Athena**. These services can be used to execute queries on datastores on AWS Data Lakehouse (Lake formation) and apply the same on read schemas. Athena as a service is serverless where we must pay only for the executed queries and can be used to run SQL based queries on the storage layer without the need of loading the data onto the database. It also exploits and uses the metadata information pertaining to the dataset portioning in the catalog layer to deliver lower cost and faster results by reducing the overhead time for data scanning.

Owing to the versatile and robust query optimizer in Redshift, it can run fast online data analytical processes on OLTP based large volumes of data stored in the Data Lakehouse storage layer. These queries, if needed, can be facilitated by scaling of node-based clusters to support massive MPP based complex and performance oriented parallel queries.

Machine Learning Based Use Cases

AWS based service like AWS Redshift and Athena contribute on a larger account for the simplification and acceleration of processes like data exploration, data wrangling and feature engineering that would otherwise take ML based engineers extended hours while training the ML models. These services take advantage of different features to make these processes simpler and faster. Some of them include:

- Discovery and search option for the data residing on the storage layer of the Data Lakehouse via metadata of lake Formation service in the catalog layer.
- Transformation of data in the Data Lakehouse datasets using interactive SQL queries.
- Transformation of datasets on Data Lakehouse into feature sets using Unified Spark based processes.

Subsequently, ***AWS SageMaker*** is used by these data scientists to prepare, model and train the data and eventually deploy these models to connect and consume this data residing on Data Lakehouse. ***Amazon SageMaker*** is a fully managed service provided by AWS for preparing, building, training and finally deployment of the Machine Learning based Models using the interactive features of ***SageMaker Studio*** service. ***AWS SageMaker Studio*** is an interactive and visual interface that is used by Data scientist and Machine Learning engineers to carry out different ML based modelling tasks like data upload, train and test new models, create new notebooks, analyze results from different model and then eventually deploy the models into the production environment.

In addition to this, SageMaker as a service provides different preconfigured notebooks and models that are easy and ready to be deployed in a fast and efficient manner. These notebooks have the latest frameworks especially on the deep learning areas including PyCarat, TensorFlow, Gluon etc. SageMaker service runs on cost efficient compute instances like EC2 Spot instances that makes the usage of this service cost efficient and highly optimized. These instances can easily scale up and down based on the demand of the training models with high powered GPU acceleration. SageMaker also ensures that automatic and optimized hyperparameter based tunning of the machine Learning Models is enabled.

Business Intelligence (AWS QuickSight)

Amazon QuickSight powered by ***AWS Redshift*** and ***Athena*** provides easy solution for the creation of interactive and rich dashboards for Business Intelligence. QuickSight is Amazon based cloud-based BI service that is used to deliver insights that are easy to comprehend and understand, apt for the audience irrespective of the fact where they reside. Here, data can be either processed from the Data Lakehouse or QuickSight can be used to directly connect to the database sources or SaaS applications. QuickSight enables **SPICE**-an in-memory caching and processing engine that helps in creating fast and interactive dashboards and at the same time replicates the data in an automated, highly redundant and highly available manner to provide users with reports and analysis that are fast, user-friendly and interactive at the same time. These dashboards are characterized by Machine Learning and BI based insights like Anomaly detection, predictive analysis and forecasting. The service can scale for hundreds of users in a fraction of time by providing a cost-effective model and at the same time an interface that can be accessed from different devices including web portals, mobile apps or web applications.

Industry based use cases

Different vendors and IT organizations across the board are collaborating and working with each other using innovation, agile development and modernization as the model to deliver enterprise class solutions. Similar collaboration and consultations have happened on the Data Lakehouse development front as well, to helping customers to replace the scattered silos of data with scalable, secure and agile Data Lakehouse solutions that have the ability to store, process and consume data irrespective of its form or structure across the business and at the same time setting it up for analysis.

With data driven analysis powered by modern day technologies like Machine Learning, forecasting, predictive analysis and visual analytics especially on real-time data being adopted by each and every organizations, Data Lakehouse as a concept is being embraced by each and every industry across the board, all finding different use cases for the implementation of this modern- day data management concept to fulfil different needs of the business.

Banking and Financial Sector

In the present world, financial institutions are facing some major challenges from the data management point of view, which makes their function less efficient and less effective. Some of these obstacles are:

- This sector is quite unique and different when we talk of the **compliance and regulatory guidelines**. These guidelines are difficult to meet and adhere to, while lack of features like data agility and reproducibility make it even harder to achieve.
- Customer-related **information and data is spread across** different businesses and each unit has its own data silo. This segregation of the customer data prevents them to have a holistic view or insights into the end-to-end customer behavior and further business opportunities. This in turn makes it unmanageable for the organizations to create personalized deliverables for the customers. This is further hampered by the inability of legacy technologies to **store and process unstructured or raw data** and thereby missing the opportunity to deliver fresh insights using this unstructured data.
- Organizations are **not able to use all the tools** in the systems that are managed by specific vendors and operated by vendor locks and thus lack the ability to make better decisions.

With the adoption of Data Lakehouse, Financial institutions have started reaping the benefits of the ecosystem such as:

- A **unified environment** for Data and AI uses cases that collate together different forms of data from different and varied sources enhancing the transformation, productive innovation, and other financial services of these organizations.
- With **no vendor locks** in place, different IT vendors are delivering customized solutions that help in accelerating the data driven offering for these financial institutions. Thereby providing deeper and thorough insights into the customer's behavior.

- These solutions offered by different vendors are equipped with **strong and robust tools** customized for different use cases that in turn help in stimulating the overall offering and services of these financial organizations.
- From the **regulatory and compliance** point of view, Data Lakehouse helps in better data governance and compliance practices by its secure and simple data management practices by streamlining the processes of data acquisition, preparation and processing.
- With the ability to **ingest data from various sources** irrespective of their structure, Data Lakehouse enables the financial institutions to have customized services and products for a personalized experience for the customers, by enabling access to the customer data across different businesses and organizations (including external data). This collection of vast data from both internal as well as external sources further enhances in creating increasingly innovative services that rely on the data driven and analytical capabilities of the Data Lakehouse ecosystem.
- The **highly efficient architecture** coupled with massive and powerful MPP provides the capability to ingest high velocity data at a rapid rate, thereby enabling these institutions to process the information in a **real-time** manner. This helps them to make quick decisions based on the patterns governing the investment opportunities or even fraud detections.

Considering the above advantages, there is an increasing number of financial organizations that are using Data Lakehouse systems to power their Analytical and Artificial Intelligence based plans. These institutions use Data Lakehouse to process petabytes of complex, raw and scattered data to create insights and personalized interfaces for their clients. In addition to this, Data Lakehouse also enhances their efforts towards the modernization and consumption of the data by direct delivery of this data into the workspaces to enable users to work on critical and specific data without exploring the entire data.

Healthcare and Life Sciences

Since the eruption of the pandemic, there has been extra focus on the healthcare sector with the aim of boosting this industry and equipping it with modernized and data driven toolsets that help in making the world a better place to live. The intent is to accelerate the research and technological advances that improve the lifecycle of the patients with an open and well architected platform that uses data and AI to base their decisions. Before going ahead with these advances, lets discuss some of the shared challenges and hurdles that are being faced by the healthcare industry in the current form:

- One of the most important facets of the healthcare industry is to understand the journey of patients with the records of their health history, personal information etc. In most of the current cases, this **information is scattered** in disparate silos equipped with unstructured data that is having limited support, hence resulting in fragmented patient information.
- The **cost of management and scaling** of current on-premises architecture is swelling day by day. This is especially important owing to the huge data inrush from different IoT based devices and the growth in fields like gene sequencing and imaging.
- **Legacy tools** incorporated with data warehouse cannot withstand the performance and throughput required for real**-time treatment** and patient care. This prevents

understanding the real-time patient insights and health patterns which prevent on-the-spot delivery of decisions and remedies.
- The field of drug research that involves modern sciences like genome sequencing, predictive analytics etc. is not achievable on **legacy data management architecture**. It needs modern systems that are equipped with high compute and robust storage that is able to withstand the tsunami of data of different forms.

The above challenges are overcome by capitalizing on the **open**, **modernized** and **collaborative data** and **AI** based platform provided by **Data Lakehouse architecture**. This robust architecture enhances the healthcare and life science-based sectors in numerous ways. Some of them are below:

- **Consolidation of all the data** from patients, operations, and research irrespective of the form and format (structured, unstructured, or semi-structured) on to a common unified platform for Machine Learning and real-time analytics. This approach provides a **360-degree view** of the patients' lifecycle and its journey, thereby helping the organization to provide better personalized therapeutics.
- With patient data increasing rapidly, there is an expectation of a fast and accelerated analysis of the data requiring a high computational system that can scale with agility and speed. Data Lakehouse, with its **scalable and agile platform,** provides organizations with a scalable cloud-based platform that can help them to gain insights and complete overview of patient heath patterns and thereby deliver better decisions and treatments.
- Healthcare applications ranging from the ones that control the production and distribution of drugs or the ones that oversee the hospital bed allotments all require analysis that is to be done on a **real-time** manner. These applications are powered by fast and high-volume data being ingested from different fast changing sources and then processed and consumed at a rapid rate. All to deliver insights based on **real-time data analytics and AI decisions**. This is possible with the modern-day architecture of Data Lakehouse that can withstand the speed and volume of the data produced.
- Data Lakehouse also empowers applications to unleash the capabilities of AI and ML to understand and comprehend the different aliments and diseases and predict the healthcare requirements. This is possible only with the unified interface and tools provided by the Data Lakehouse ecosystem.

Retail and Consumer Industry

The retail industry is one of the most important sectors contributing to the overall economy of the nations. The sector has been growing at a rapid rate in the last decade and is riding on the bandwagon of modernization and tech revolution. It is estimated that retail-based sales would be rising from $23 trillion (about $71,000 per person in the US) in 2018 to a staggering growth of 3.58% to $26.6 billion (about $82 per person in the US) by the end of 2022. With the ability to tap into new channels and markets, new data driven technologies are driving this sector and making it easier for the consumer and retails sector to tap on emerging and other international markets across the globe. With the revolution of data and the velocity and volume at which retail consumers are producing data of all kinds, it has become one of the most important catalysts for the growth of this sector, especially to understand the shopping trends and habits of the consumers and at the same time attract new customers towards their business.

Expanding the existing market and capturing the new ones using data driven technologies has become challenging in the modern retail and consumer sector. Some of these challenges are as below:

- The competitive world that we live in demands quick, efficient and effective decisions that are based on **real-time and agile** data sources. The retail industry is facing a similar challenge to make itself onboard for **real-time analysis** that would complement its supply demand equations and at the same time help customers to gain a better shopping experience.
- Retail and consumer industry needs to adopt new methods that overcome the limitations in the **old and archaic toolsets** thereby enhancing overall agility, speed and flexibility.
- The legacy architecture that governed data warehouse-based systems are **not up to the mark** for today's data and the use cases that are demanded by the industry. So, the need of the hour is to have a system that has an architecture that can facilitate the delivery of any data of any form or type.
- Retail and consumer industry is expanding like anything, with new players in the form of startups joining the sector at a rapid rate. One of the most important factors for their success is to **lower the overall operation cost**. Currently, the data management systems, being one of the most important backbones of the sector, is featured by complexity and high cost. Ask of the hour is to get these costs to lower down for the overall success of the industry.

With the retail and consumer sector looking each and every day for new ways and means to merge the gap between the products and the data, Data Lakehouse is facilitating the industry to narrow this gap and at the same time creating meaningful insights for the customers and partners to have a better collaboration and harness the potential of the data by using the integrated and unified platform in the form of Data Lakehouse. Below are some of the ways in which adoption of Data Lakehouse has enhanced the growth in this sector:

- Customer experience is one of the most important pillars of the retail industry. Adoption of Data Lakehouse architecture has created opportunities that deliver comprehensive customer experiences. The sector uses **fast and real-time** transactional data from the customer usage to ingest into the Data Lakehouse ecosystem and deliver real AI based insights across the value chain in a scalable and real-time manner.
- Getting rid of the data silos that were previously based on different data structures and consolidating them into a centralized and unified data management system in the form of Data Lakehouse has enabled the consumer and retail industry to get a **holistic 360-degree view** of the customer business and at the same time providing **better and richer insights** to achieve meaningful results.
- The **low-cost model of Data Lakehouse** enables this sector, especially the new breed, to facilitate collaboration with vendors, partners and manufacturing teams to deliver greater innovation with better efficiency.
- Accurate forecasting and predictive analysis play an important role in the success of the retail sector. With Data Lakehouse powered by **modern algorithms and AI based toolsets** that have the ability to analyze and predict based on the social media and web browsing data has become the latest success in this sector.

Conclusion

The world is going through an era of the 4th Industrial revolution, riding on the bandwagon of technological modernization and innovation. A phase in history, where digital technology is transforming each and every facet of human existence, with products fast moving from luxury to necessity. These advances have been happening in each and every sector of industry and business, with changes happening at a rapid rate. This has accelerated, especially owing to the recent pandemic, where industries are trying to make their businesses more and more profitable and at the same time giving a smooth and on the click user experience.

Introduction of innovative technologies like 5G, Edge computing, IOT and adoption of microservices based architecture is creating a big impact. This has in turn resulted in a data explosion, where businesses have started to realize the power behind the data and its uses. This has resulted in a domino effect, thereby giving rise to AI, ML and Data Analytical based functions that are being adopted not only for understanding the trends and patterns of the business but also to create better planning and effective and efficient business operations. This has resulted in tremendous growth of these organizations and better customer satisfaction, which is paramount.

As data generation accelerates, industry is looking for optimal means to store and analyze this huge volume of data. The need of the hour is for data management solutions that can withstand the resiliency, latency, variety and the volume of data that is being generated by these organizations. Different options had been tried to use the existing legacy technologies like Data warehouse or Data lakes to adopt one another's qualities like being interactive like Data lake or storing structured data like data warehouse, but all approaches have failed and resulted in inflated costs and unhappy users.

The introduction of Data Lakehouse architecture has helped the organizations to overcome most of these hurdles by consolidating the siloed and separate data management systems. With this thought of having a single access point for all the data irrespective of the form or format of organizations, "Data Lakehouse" is the new buzz word in the world of data management systems. The new architecture of Data Lakehouse is now complementing the existing Data lakes with the capabilities they are powered by SQL performance and at the same time saving cost for the organizations by overcoming any redundant or inconsistent data as part of the earlier multiple tiered systems. Now the organizations and their businesses are adopting new methods for transforming their data management and data analytical use cases that were based on Business Intelligence to the ones that are based on Artificial Intelligence and Machine Learning. All this is being done by taking into consideration the compliance, reliability and security of the data at hand.

Compared to its predecessors like Data warehouse or Data lakes, the planning and implementation of Data Lakehouse ecosystem is much simpler and faster, this is especially due to the fact that there are multiple vendor free out of the box solutions offered by different companies. These solutions can range from Delta Lake architectural solutions offered by Databricks or the multiple services offered by AWS or Google cloud for their own version of Data Lakehouse. These solutions are based on open protocols, native connectors and open APIs that makes the communication between the system and the databases and applications much smoother and more efficient thereby helping the data engineers to create and deliver data pipelines in an effective manner.

Considering all the hype and buzz around Data Lakehouse, we need to keep under consideration that the concept and acceptability of Data Lakehouse is still in its infancy and there is still lot of work to be done on this and the concept is still "**work under progress**" status. This is true especially when considering the monolithic structure of the solution that might be challenging to maintain in future. There are some experts of the opinion that a 2-tiered solution of data warehouse with Data lakes was more efficient and effective, giving us the power and tools of both worlds.

Taking into consideration the above views, it seems that there is still some time left till we see the dawn of wider acceptability of the Data Lakehouse architecture. Design and adoption of agile open file formatted data structures and improved caching solutions would be some of the main catalysts that can power the new phase of application data and thereby accelerating the adoption of Data Lakehouse systems. With all this I can see a bright and clear future for Data Lakehouse and the technology is going to stay for a long time. Proving to be a strong and new paradigm shift in the history of Data Management systems!

Bibliography

1. AltexSoft. 2022. Data Lakehouse: Concept, Key Features, and Architecture Layers. [online] Available at: *https://www.altexsoft.com/blog/data-lakehouse*
2. Amazon Web Services, Inc. 2022. Modern Data Architecture on AWS | Amazon Web Services. [online] Available at: *https://aws.amazon.com/big-data/datalakes-and-analytics/modern-data-architecture*
3. Amazon Web Services. 2022. Build a Lake House Architecture on AWS | Amazon Web Services. [online] Available at: *https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws*
4. Amazon Web Services. 2022. Harness the power of your data with AWS Analytics | Amazon Web Services. [online] Available at: *https://aws.amazon.com/blogs/big-data/harness-the-power-of-your-data-with-aws-analytics*
5. Bajda-Pawlikowski, K., 2022. Why Data Lakehouse Architecture Now?. [online] Blog.starburst.io. Available at: https://blog.starburst.io/why-data-lakehouse-architecture-now>
6. Databricks. 2022. History and evolution of data lakes - Databricks. [online] Available at: *https://databricks.com/discover/data-lakes/history*
7. Databricks. 2022. Introduction to Data Lakes - Databricks. [online] Available at: *https://databricks.com/discover/data-lakes/introduction#:~:text=A%20data%20lake%20is%20a,storage%20to%20store%20the%20data*
8. Databricks. 2022. The Data Lakehouse Solution for Retail and CPG - Databricks. [online] Available at: *https://databricks.com/solutions/industries/retail-industry-solutions*
9. DATAVERSITY. 2022. A Brief History of Data Lakes - DATAVERSITY. [online] Available at: https://www.dataversity.net/brief-history-data-lakes
10. Devlin, B., 2022. Weaving Architectural Patterns II – Data Lakehouse. [online] Data Virtualization blog - Data Integration and Modern Data Management Articles, Analysis and Information. Available at: *https://www.datavirtualizationblog.com/data-lakehouse-weaving-architectural-patterns-ii*
11. Fivetran.com. 2022. What Is a Data Lakehouse? | Blog | Fivetran. [online] Available at: *https://www.fivetran.com/blog/what-is-a-data-lakehouse*
12. Google Cloud Blog. 2022. Open data lakehouse on Google Cloud | Google Cloud Blog. [online] Available at: *https://cloud.google.com/blog/products/data-analytics/open-data-lakehouse-on-google-cloud*
13. Ibm.com. 2022. Data Lake Solutions. [online] Available at: *https://www.ibm.com/analytics/data-lake?utm_content=SRCWW&p1=Search&p4=43700068838236895&p5=e&gclid=CjwKCAiA1JGRBhBSEiwAxXblwdZG6n1Jp9CMWBATH0DDraya-4xBzyCf8rFNDBHEgfJmm-9V4yBhPRoC7hQQAvD_BwE&gclsrc=aw.ds*
14. Inmon, B. and Levins, M., 2022. Evolution to the Data Lakehouse. [online] Databricks. Available at: *https://databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html#:~:text=Building%20the%20Data%20Lakehouse.,the%20data%20warehouse%2C%20Bill%20Inmon*
15. Kutay, J., 2022. Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns | Striim. [online] Striim. Available at: *https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview*
16. Linkedin.com. 2022. Lake House Architecture - must watch "Data Paradigm" in 2021. [online] Available at: *https://www.linkedin.com/pulse/lake-house-architecture-must-watch-data-paradigm-2021-vamsi-behara*

17. Lorica, B., Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M., 2022. What Is a Lakehouse? - The Databricks Blog. [online] Databricks. Available at: *https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html*
18. Marr, B., 2022. What Is A Data Lake? A Super-Simple Explanation For Anyone. [online] Forbes. Available at: *https://www.forbes.com/sites/bernardmarr/2018/08/27/what-is-a-data-lake-a-super-simple-explanation-for-anyone/?sh=1d38e0de76e0*
19. Snowflake. 2022. What is a Data Lakehouse?. [online] Available at: *https://www.snowflake.com/guides/what-data-lakehouse#:~:text=A%20data%20lakehouse%20is%20a,cost%2Deffective%20for%20data%20storage*
20. Spglobal.com. 2022. So, the data lakehouse is now officially a 'thing' what is it and why should you care. [online] Available at: *https://www.spglobal.com/marketintelligence/en/news-insights/blog/so-the-data-lakehouse-is-now-officially-a-thing-what-is-it-and-why-should-you-care*

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes, or methodologies.