What is the Data Lakehouse Architecture?

Definition
Data lakehouse aims to combine both functional and non-functional features of the traditional big data analytics architecture having data lake and data warehouse capabilities under single platform.

This definition has expanded over time to accommodate more analytical services. Cloudera expects this trend to continue in the future and include scope for real-time streaming analytics and operational data stores.

Origin
The term data lakehouse first entered the Enterprise Data Platform lexicon back in 2017. It was used by an organisation called Jellyvision to describe and combine structured data processing (data warehouse) with a schemaless system (data lake). Combining these two architectural paradigms, led to the data lakehouse. Since then, the term’s definition has evolved to include additional analytical services such as machine learning (ML), but also greater support for the management features of traditional data warehouses.

Qualities
A modern data lakehouse should bring together the benefits of a data lake and a data warehouse at a low total cost of ownership (TCO). It should therefore possess the following key qualities:

- Open, flexible, and performant file and table formats e.g., Apache Parquet, Iceberg
- ACID transactions, table versioning, snapshots, and sharing at petabyte scale
- Multifunction analytics across an open ecosystem
- Strong data management (security, governance, and lineage)
- Strong data quality and reliability
- Best in class SQL performance
- Direct and declarative access for non-SQL interactions

Why A Data Lakehouse Architecture Is Essential

Limitations of Data Lake and Data Warehouse
To understand why the Data Lakehouse architecture is growing in popularity, we need to consider the architectures it replaces and their limitations. Data is first ingested into a data lake by an export transform load (ETL) operation from each source system. Historically, these sources would mainly be operational systems containing structured data. However, today more than half of the data ingested is semi-structured or unstructured data.

Data is then loaded into a data warehouse with another ETL operation. Data is conformed into a given logical data model, often on an underlying proprietary storage layer. SQL can then be used to query the data and we get to benefit from DBMS features of the Data Warehouse such as support for transactions, table versioning and snapshots.
While these architectures provide the economic benefits of cheap, scalable storage, they suffer from three main challenges:

- **Data duplication**
- **Reliability/Quality**
- **Limited support for analytical services**
- **Support for all kind of data**

<table>
<thead>
<tr>
<th>Data Lake</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros:</strong></td>
<td><strong>Pros:</strong></td>
</tr>
<tr>
<td>🔄 Open for broad ingestion</td>
<td>🔄 Structured, Well curated data</td>
</tr>
<tr>
<td>🔄 Semi-structured and unstructured data</td>
<td>🔄 Precise analytics, good for reporting</td>
</tr>
<tr>
<td>🔄 Good for artificial intelligence and machine learning</td>
<td>🔄 Well governed</td>
</tr>
<tr>
<td>🔄 Easy to add diverse analytic engines</td>
<td>🔄 Built for SQL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cons:</strong></th>
<th><strong>Cons:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>🔴 Data quality</td>
<td>🔴 Inflexible</td>
</tr>
<tr>
<td>🔴 Varying degree of accuracy</td>
<td>🔴 Does not support unstructured data so it struggles with AI/ML capabilities</td>
</tr>
</tbody>
</table>

**Our Solution: Building the Open Data Lake House**

Cloudera and Tech Mahindra are working together to resolve limitations of the traditional big data analytics architecture having data lake and a data warehouse.

In this section we introduce the Cloudera data platform (CDP). We then summarize the key logical service components that support Cloudera's open data lakehouse. We describe how Apache Iceberg provides a flexible, scalable table format to support schema-based access to multiple analytical services across the data lifecycle.
Overcoming Limitations While Delivering Qualities

<table>
<thead>
<tr>
<th>Avoiding Data Duplication</th>
<th>Supporting an Open Ecosystem of Analytical Engines</th>
<th>Flexible Hybrid Deployment Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instead of data being copied from source systems into a data lake and then again into a data warehouse, the data lakehouse provides a layer of abstraction to the underlying data in the lake. This transactional metadata layer on top of the underlying Data Lake provides support for common Data Warehouse management features such as transactions, data versioning and snapshots. Reducing the number of ETL steps to get data into the data warehouse reduces the likelihood of errors, improved efficiency and potential inconsistencies in processing engines.</td>
<td>Providing an efficient SQL interface for BI and reporting is necessary but insufficient and quite limiting when supporting ML. Systems that implement the data lakehouse architecture therefore need to be able to provide direct access to the underlying data in the lake. At the same time, ML frameworks must be able to take advantage of metadata to simplify the process of importing data into data frames for data science pipeline building and model creation.</td>
<td>To significantly reduce the TCO, we must move away from expensive and proprietary hardware. Modern implementations of the Data Lakehouse architecture decouple compute and storage and opt for cloud-native architectures. This allows each to scale independently and simplifies running multiple analytical workloads across shared data. Data lakehouses are required on premises on commodity hardware and in the public cloud. There are advantages for adopting cloud native hybrid solutions that can leverage object storage and managed container services across each environment.</td>
</tr>
</tbody>
</table>

The Cloudera Data Platform (CDP)

CDP is a hybrid data platform designed to provide the freedom to choose any cloud, any analytics, and any data. CDP delivers fast and easy data management and data analytics for data anywhere, with optimal performance, scalability, and security. CDP provides the freedom to securely move applications, data, and users bi-directionally between data centers and multiple data clouds, regardless of where data resides.

Solution is scalable to upgrade the existing legacy Cloudera set up into a data lakehouse.

Apache Iceberg—An Open Table Format Enabling the Lakehouse Architecture

As highlighted earlier in the document, a data lakehouse possesses a set of qualities. Those qualities are the union of those from a data lake and those from a data warehouse. We cannot simply bring together a processing engine and a flexible table format, and say it implements a data lakehouse. We must also integrate the qualities of a data lake.

Iceberg provides a flexible and open storage format that supports petabyte scale tables. It does this by storing both the data and metadata in the data lake. Data is typically stored in Apache Parquet format and the associated metadata in Apache Avro format. Entries in the Data Catalog are then a pointer to the manifest file on the Data Lake.

Iceberg also supports many of the management features of a traditional Data Warehouse. These include transactions, data versioning and snapshots. Iceberg supports flexible SQL commands to...
merge new data, update existing rows, and perform targeted deletes. Time travel enables reproducible
queries that use the same table snapshot, or lets users easily examine changes. Version rollback
allows users to quickly correct problems by resetting tables to a previously known state.

Ease of Using APACHE ICEBERG in CDP

### Use Cases

**Use Case** | **Description**
---|---
Real-Time Data Analytics | Real-time data analytics is the instant processing of all enterprise data for analysis as soon as it enters an organization’s information system. The streaming data and historical data in lakehouse servers for real-time analytics like bank fraud detections.
Predictive Analytics | Many organizations are looking to improve their customer service through ML programs like NLP. A data lakehouse can be used to store huge amounts of structured and unstructured data that can be analyzed using NLP platforms.
Demand Forecasting | The historical data stored in lakehouse can be used for cases like demand forecasting for retail, procurement, and dynamic pricing.
Predictive Maintenance | Sensor data about machines working conditions in a manufacturing facility can be brought in Lakehouse to analyze and indicate about any machine performing below par and its cause.
Medical Diagnosis | AI/ML based workloads can run through a huge set of image files stored in Lakehouse to predict a medical condition like cancer.
Reports and Dashboards | The structured data and its metadata stored in lakehouse serves very well for BI reporting cases like sales report.
Benefits

### Platform Benefits

- Reduced complexity of handling BI, AI, ML workloads with CML, Cloudera data visualization services
- Optimize processing
- Reduce TCO and maintenance overheads
- Cloud native
- Scalability
- Much easier to implement data security and governance on a single storage tier using SDX
- Avoid data duplication, data staleness, and increase data quality by easily maintaining a single source of truth

### Apache Iceberg Capabilities

- Open storage format that is flexible to be used by multiple data processing engines
- High performance over concurrent read and write of huge volume and variety of data
- Engine agnostic
- Snapshot isolation
- Time travel
- Hidden partitioning

Success Stories

1. Improve fraud detection, customer relationship management, network quality and operational efficiency for a leading German Telco
   - *Gained ~20% reduction in revenue losses due to fraud*

2. IoT enabled predictive maintenance to help fleet and truck owners minimize vehicle downtime for an International Vehicle Manufacturer
   - *Gained ~30% reduction in vehicle maintenance costs*

3. Big Data Analytics for holistic view and data analysis enabling improvement is profitability and saving lives for a Global Pharma Major
   - *5PB of Data across 10 domains and 2000 silos combined*

About Tech Mahindra

Tech Mahindra offers innovative and customer-centric digital experiences, enabling enterprises, associates, and society to Rise. We are a USD 6 billion organization with 157,000+ professionals across 90 countries helping 1290 global customers, including Fortune 500 companies. We are focused on leveraging next-generation technologies, including data analytics, 5G, blockchain, cybersecurity, artificial intelligence, and more, to enable end to end digital transformation for global customers.

About Cloudera

At Cloudera, we believe data can make what is impossible today, possible tomorrow. Cloudera taught the world the value of big data, creating an industry and ecosystem powered by the relentless innovation of the open-source community. We empower our customers, leaders in their industries, to transform complex data into clear and actionable insights. Through our hybrid data platform, organizations are able to build their data-driven future by getting data - no matter where it resides - into the hands of those who need it. Learn more at Cloudera.com.

Ref: https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/the-open-data-lakehouse.pdflanding.html