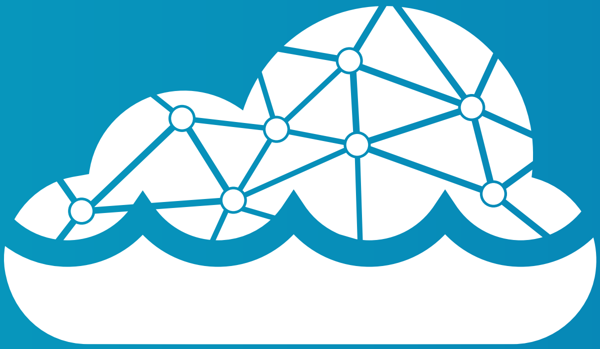
WHAT IS AN OPEN DATA LAKE?



A data lake is a system or repository that stores data in its raw format as well as transformed trusted datasets and provides both programmatic and SQL based access to this data for diverse analytics tasks such as machine learning, data exploration, and interactive analytics.

The data stored in a data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video).

A data lake, where data is stored in an open format and accessed through open standards-based interfaces, is defined as an Open Data Lake. This adherence to an open philosophy, aimed at preventing vendor lock-in, permeates through every aspect of the system, including data storage, data management, data processing, operations, data access, governance, and security.

We define an open format as a format that is based on an underlying open standard, developed and shared through a publicly visible and community-driven process without vendor-specific proprietary extensions. For example, an Open Data Format is a platform-independent, machine-readable data format such as ORC or Parquet, whose specification is published to the community, such that any organization can create tools and applications to read data in the format.

A typical data lake has the following capabilities:

- Data Ingestion and storage
- Data processing and support for continuous data engineering
- Data Access and consumption
- Data Governance - Discoverability, Security and Compliance
- Infrastructure and operations

In the following sections, we will describe openness requirements for each capability.

Data Ingestion and Storage

An Open Data Lake ingests data from sources such as applications, databases, data warehouses, and real-time streams. It formats and stores the data into an open data format, such as ORC and Parquet, that is platform-independent, machine-readable, optimized for fast access and analytics, and made available to consumers without restrictions that would impede the re-use of that information.

An Open Data Lake supports both the pull and push based ingestion of data. It supports pull-based ingestion through batch data pipelines and push-based ingestion through stream processing. For both these types of data ingestion, an Open Data Lake supports open standards such as SQL and Apache Spark for authoring data transformations. For batch data pipelines, it supports row-level inserts and updates — UPSERT — to datasets in the lake. Upsert capability with snapshot isolation — and more generally, ACID semantics — greatly simplifies the task, as opposed to rewriting data partitions or entire datasets.

The ingest capability of Open Data Lake ensures zero data loss and writes exactly-once or at-least-once; handles schema variability; writes in the most optimized data format into the right partitions, and provides the ability to re-ingest data when needed.

Data Processing and support for Continuous Data Engineering

Open Data Lake stores the raw data from various data sources in a standardized open format. However, use cases such as data exploration, interactive Analytics, and Machine Learning require that the raw data is processed to create use-case driven trusted datasets. For Data Exploration and Machine Learning use cases, users continually refine data sets for their analysis needs. As a result, every data lake implementation must enable users to iterate between data engineering and use cases such as interactive analytics and Machine Learning. We call this “Continuous Data Engineering”.

Continuous Data Engineering involves the interactive ability to author, monitor, and debug data pipelines. In an Open Data Lake, these pipelines are authored using standard interfaces and open source frameworks such as SQL, python, Apache Spark, and/or Apache Hive.

Data Governance: Discoverability, Security and Compliance

When data ingestion and data access are implemented well, data can be made widely available to users in a democratized fashion. When multiple teams start accessing data, data architects need to exercise oversight for governance, security, and compliance purposes.

Data Discovery

Data itself is hard to find and comprehend and not always trustworthy. Users need the ability to discover and profile datasets for integrity before they can trust them for their use case. A data catalog enriches metadata through different mechanisms, uses it to document datasets, and supports a search interface to aid discovery.

Since the first step is to discover the required datasets, it's essential to surface metadata to end-users for exploration purposes, to see where the data resides and what it contains, and to determine if it is useful for answering a particular question. Discovery includes data profiling capabilities that support interactive previews of datasets to shine a light on formatting, standardization, labels, data shape, and so on.

Open Data Lake provides an open metadata repository. As an example, Apache Hive metadata repository is an open metadata repository that prevents vendor lockin for metadata.

Security

Increasing accessibility to the data requires data lakes to support strong access control and security features on the data. An Open Data Lake does this through non-proprietary security and access control APIs. As an example, deep integration with open source frameworks such as Apache Ranger and Apache Sentry can facilitate table, row and column level granular security. This enables administrators to grant permissions against already-defined user roles in enterprise directories such as Active Directory etc. By basing access control on open source frameworks, the Open Data Lake avoids vendor lock-in through proprietary security implementation

Compliance

New or expanded data privacy regulations, such as GDPR and CCPA, have created new requirements around "Right to Erasure" and "Right to Be Forgotten". These govern consumers' rights about their data and involve stiff financial penalties for non-compliance (as much as 4% of global turnover), so they must not be overlooked. Therefore, the ability to delete specific subsets of data without disrupting a data management process is essential. An Open Data Lake supports this ability on open formats and open metadata repositories. In this way, they enable a vendor agnostic solution to compliance needs.

Infrastructure and Operations

Whether the data lake is deployed in the cloud or on-premises, each cloud provider has specific implementation to provision, configure, monitor, and manage the data lake as well as the resources it needs.

Open Data Lake is cloud-agnostic and is portable across any cloud-native environment including public and private clouds. This enables administrators to leverage benefits of both public and private cloud from economics, security, governance and agility perspective.

Conclusion

The increase in volume, velocity and variety of data, combined with new types of analytics and machine learning is creating the need for an open data lake architecture. Across our 200+ customers including market leaders like Expedia, Disney, Lyft, Adobe and more, we find that the Open Data Lake is becoming a common feature alongside the Data Warehouse. While the Data Warehouse has been designed and optimized for SQL analytics, the need for an open, simple and secure data lake platform, that can support new types of analytics and machine learning is driving the Open Data Lake adoption. Unlike the Data Warehouse's world of proprietary formats, proprietary SQL extensions, proprietary metadata repository and lack of programmatic access to data, the Open Data Lake ensures no vendor lock-in while supporting a diverse range of analytics. The Open Data Lakes provide a robust and future-proof data management paradigm to support a wide range of data processing needs including data exploration, interactive analytics, and machine learning.