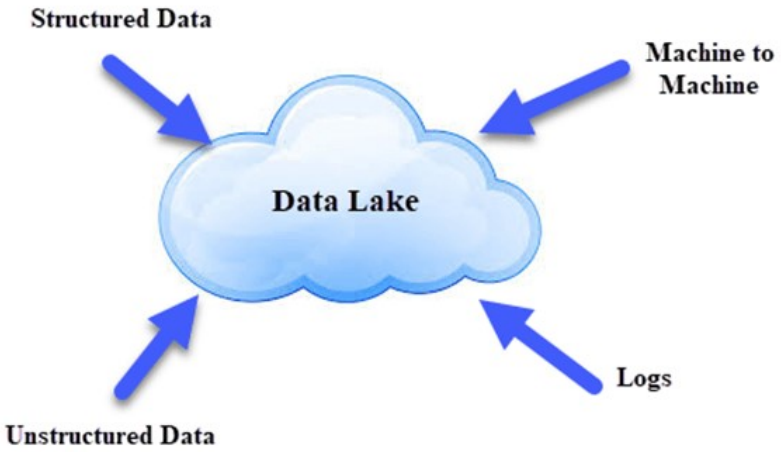
What is Data Lake? It's Architecture

What is Data Lake?

A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data. It is a place to store every type of data in its native format with no fixed limits on account size or file. It offers high data quantity to increase analytic performance and native integration.

Data Lake is like a large container which is very similar to real lake and rivers. Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time.



The Data Lake democratizes data and is a cost-effective way to store all data of an organization for later processing. Research Analyst can focus on finding meaning patterns in data and not data itself.

Unlike a hierarchal Dataware house where data is stored in Files and Folder, Data lake has a flat architecture. Every data elements in a Data Lake is given a unique identifier and tagged with a set of metadata information.

In this tutorial, you will learn-

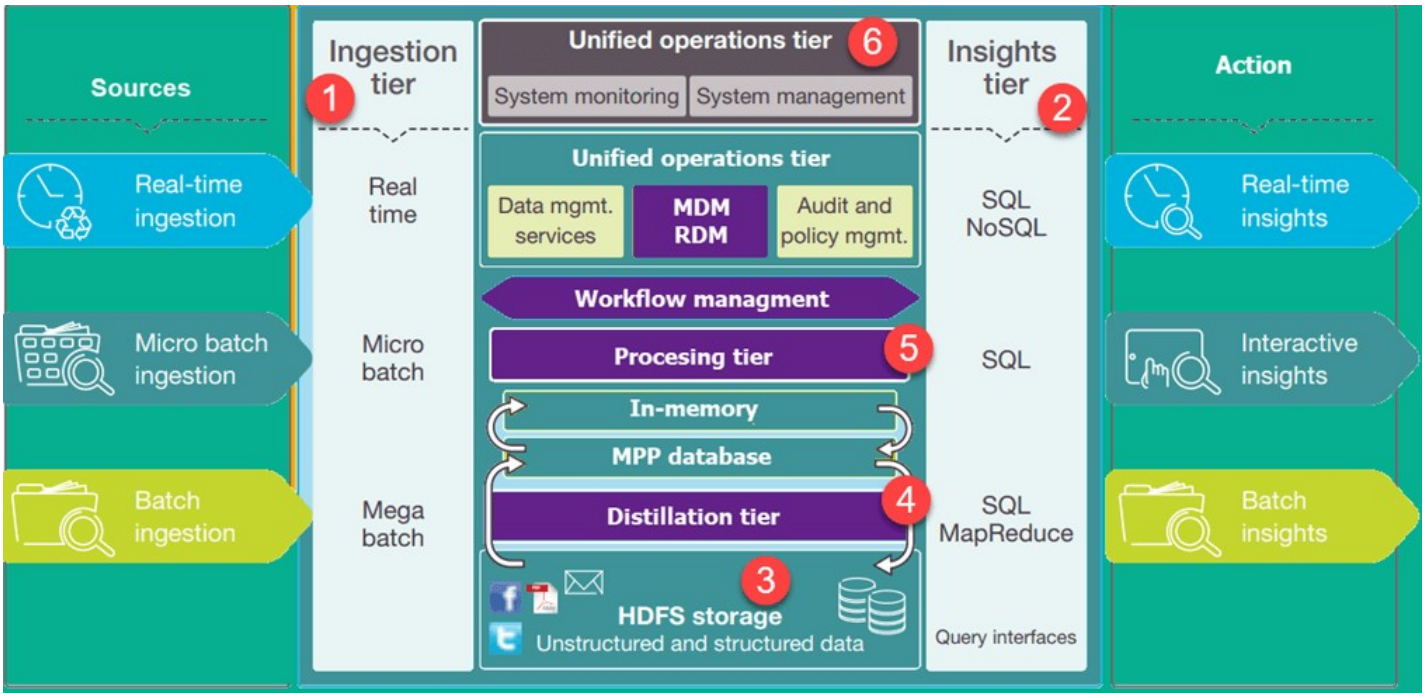
Why Data Lake?

The main objective of building a data lake is to offer an unrefined view of data to data scientists.

Reasons for using Data Lake are:

- With the onset of storage engines like Hadoop storing disparate information has become easy. There is no need to model data into an enterprise-wide schema with a Data Lake.
- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

Data Lake Architecture



The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency. Following are important tiers in Data Lake Architecture:

1. **Ingestion Tier**: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time
2. **Insights Tier:** The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.
3. **HDFS** is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.
4. **Distillation tier** takes data from the storage tire and converts it to structured data for easier analysis.
5. **Processing tier** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
6. **Unified operations tier** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

Key Data Lake Concepts

Following are Key Data Lake concepts that one needs to understand to completely understand the Data Lake Architecture

Data Ingestion

Data Ingestion allows connectors to get data from a different data sources and load into the Data lake.

Data Ingestion supports:

- All types of Structured, Semi-Structured, and Unstructured data.
- Multiple ingestions like Batch, Real-Time, One-time load.
- Many types of data sources like Databases, Webservers, Emails, IoT, and FTP.

Data Storage

Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.

Data Governance

Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.

Security

Security needs to be implemented in every layer of the Data lake. It starts with Storage, Unearthing, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards.

Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.

Data Quality:

Data quality is an essential component of Data Lake architecture. Data is used to exact business value. Extracting insights from poor quality data will lead to poor quality insights.

Data Discovery

Data Discovery is another important stage before you can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the Data lake.

Data Auditing

Two major Data auditing tasks are tracking changes to the key dataset.

1. Tracking changes to important dataset elements
2. Captures how/ when/ and who changes to these elements.

Data auditing helps to evaluate risk and compliance.

Data Lineage

This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination.

Data Exploration

It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration.

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

Maturity stages of Data Lake

The Definition of Data Lake Maturity stages differs from textbook to other. Though the crux remains the same. Following maturity, stage definition is from a layman point of view.

Stage 1: Handle and ingest data at scale

This first stage of Data Maturity Involves improving the ability to transform and analyze data. Here, business owners need to find the tools according to their skillset for obtaining more data and build analytical applications.

Stage 2: Building the analytical muscle

This is a second stage which involves improving the ability to transform and analyze data. In this stage, companies use the tool which is most appropriate to their skillset. They start acquiring more data and building applications. Here, capabilities of the enterprise data warehouse and data lake are used together.

Stage 3: EDW and Data Lake work in unison

This step involves getting data and analytics into the hands of as many people as possible. In this stage, the data lake and the enterprise data warehouse start to work in a union. Both playing their part in analytics

Stage 4: Enterprise capability in the lake

In this maturity stage of the data lake, enterprise capabilities are added to the Data Lake. Adoption of information governance, information lifecycle management capabilities, and Metadata management. However, very few organizations can reach this level of maturity, but this tally will increase in the future.

Best practices for Data Lake Implementation:

- Architectural components, their interaction and identified products should support native data types
- Design of Data Lake should be driven by what is available instead of what is required. The schema and data requirement is not defined until it is queried
- Design should be guided by disposable components integrated with service API.
- Data discovery, ingestion, storage, administration, quality, transformation, and visualization should be managed independently.
- The Data Lake architecture should be tailored to a specific industry. It should ensure that capabilities necessary for that domain are an inherent part of the design
- Faster on-boarding of newly discovered data sources is important
- Data Lake helps customized management to extract maximum value
- The Data Lake should support existing enterprise data management techniques and methods

Challenges of building a data lake:

- In Data Lake, Data volume is higher, so the process must be more reliant on programmatic administration
- It is difficult to deal with sparse, incomplete, volatile data
- Wider scope of dataset and source needs larger data governance & support

Difference between Data lakes and Data warehouse

Parameters	Data Lakes	Data Warehouse
Data	Data lakes store everything.	Data Warehouse focuses only on Business Processes.
Processing	Data are mainly unprocessed	Highly processed data.

Type of Data	It can be Unstructured, semi-structured and structured.	It is mostly in tabular form & structure.
Task	Share data stewardship	Optimized for data retrieval
Agility	Highly agile, configure and reconfigure as needed.	Compare to Data lake it is less agile and has fixed configuration.
Users	Data Lake is mostly used by Data Scientist	Business professionals widely use data Warehouse
Storage	Data lakes design for low-cost storage.	Expensive storage that give fast response times are used
Security	Offers lesser control.	Allows better control of the data.
Replacement of EDW	Data lake can be source for EDW	Complementary to EDW (not replacement)
Schema	Schema on reading (no predefined schemas)	Schema on write (predefined schemas)
Data Processing	Helps for fast ingestion of new data.	Time-consuming to introduce new content.
Data Granularity	Data at a low level of detail or granularity.	Data at the summary or aggregated level of detail.
Tools	Can use open source/tools like Hadoop/ Map Reduce	Mostly commercial tools.

Benefits and Risks of using Data Lake:

Here are some major benefits in using a Data Lake:

- Helps fully with product ionizing & advanced analytics
- Offers cost-effective scalability and flexibility
- Offers value from unlimited data types
- Reduces long-term cost of ownership
- Allows economic storage of files
- Quickly adaptable to changes
- The main advantage of data lake is the **centralization** of different content sources
- Users, from various departments, may be scattered around the globe can have **flexible access** to the data

Risk of Using Data Lake:

- After some time, Data Lake may lose relevance and momentum
- There is larger amount risk involved while designing Data Lake
- Unstructured Data may lead to Ungoverned Chao, Unusable Data, Disparate & Complex Tools, Enterprise-Wide Collaboration, Unified, Consistent, and Common
- It also increases storage & computes costs
- There is no way to get insights from others who have worked with the data because there is no account of the lineage of findings by previous analysts
- The biggest risk of data lakes is security and access control. Sometimes data can be placed into a lake without any oversight, as some of the data may have privacy and regulatory need

Summary:

- A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data.
- The main objective of building a data lake is to offer an unrefined view of data to data scientists.
- Unified operations tier, Processing tier, Distillation tier and HDFS are important layers of Data Lake Architecture
- Data Ingestion, Data storage, Data quality, Data Auditing, Data exploration, Data discover are some important components of Data Lake Architecture
- Design of Data Lake should be driven by what is available instead of what is required.
- Data Lake reduces long-term cost of ownership and allows economic storage of files
- The biggest risk of data lakes is security and access control. Sometimes data can be placed into a lake without any oversight, as some of the data may have privacy and regulatory need.