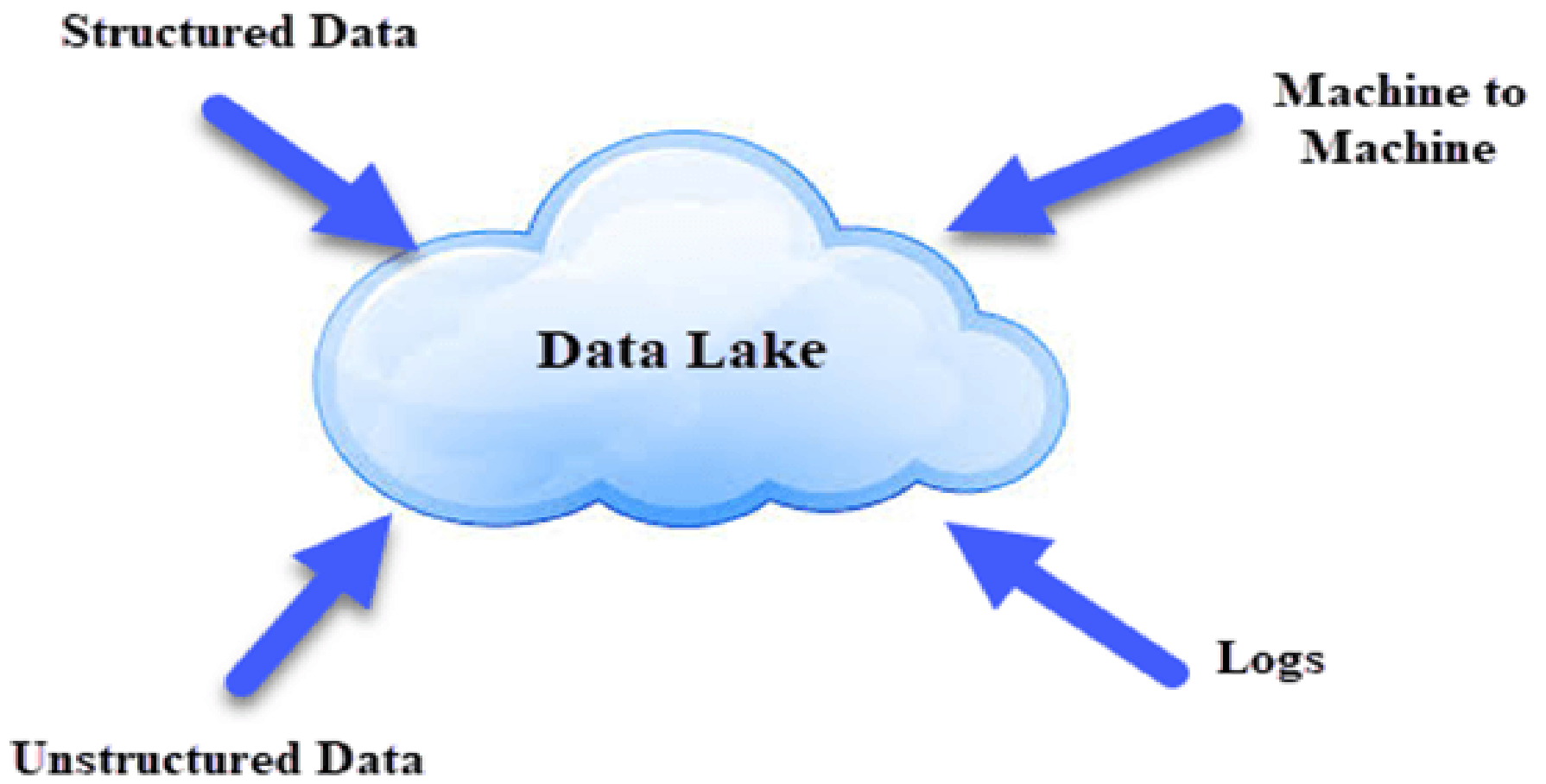
Virtuoso Infotech Pvt. Ltd.

About Virtuoso Infotech

- Fastest growing IT firm; Offers the flexibility of a small firm and robustness of over 30 years experience collectively within the leadership team
- Technology expertise & passionate team
- Successful client engagements across India, USA, UK, Australia and Argentina
- Handle enterprise solutions that involve **30,000 active users**, more than 20 servers, **data volume as big as 5 million entries per day**


Structured Data
Machine to Machine
Data Lake
Logs
Unstructured Data

Agenda

- What is Data Lake?
- Why Data Lake?
- Data Lake Architecture
- Key Data Lake Concepts
- Maturity stages of Data Lake
- Data lakes Vs Data warehouse
- Advantages
- Disadvantages

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.

Why Data Lake?

- 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.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

Data Lake Architecture

- **Ingestion Tier**: The tiers depict the data sources. The data could be loaded into the data lake in batches or in real-time
- **Insights Tier:** The tiers represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.
- **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.

- **Distillation tier:** takes data from the storage tire and converts it to structured data for easier analysis.
- **Processing tier:** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
- **Unified operations tier:** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

Key Data Lake Concepts

- **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.

- **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.

Maturity Stages of Data Lake

Stage 1: Handle and ingest data at scale

This first stage of Data Maturity Involves improving the ability to transform and analyse 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 analyse 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.

Data Lake Vs 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.

Advantages

- 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

Disadvantages

- 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

Thank You!

Virtuoso InfoTech Pvt. Ltd.
4th Floor, Victory Landmark, Opp. D-Mart,
Behind Dominos Pizza, Baner, Pune.

+91 20 6050 1318
support@virtuositech.com

virtuoso INFOTECH

www.virtuosoitech.com