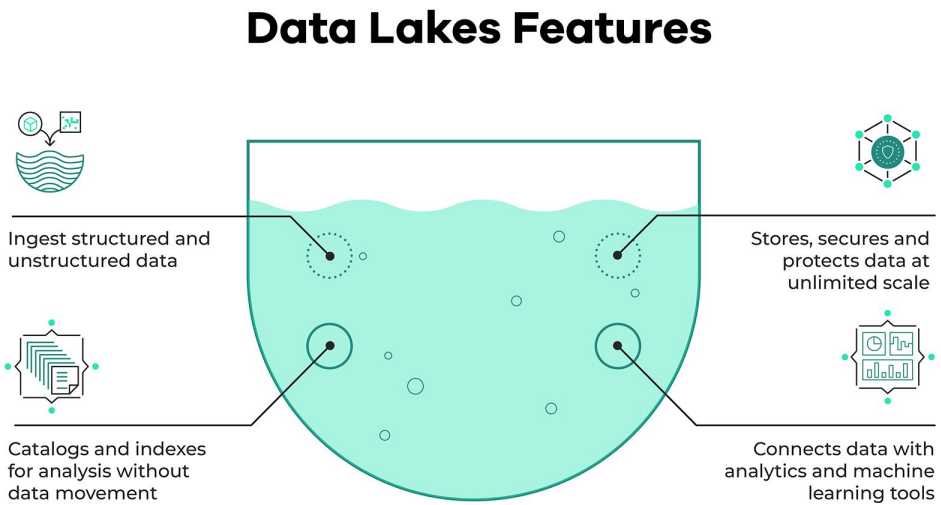
Data Lake

Index

- What is a Data Lake?
- Data Lake vs Data warehouse
- Gotcha of Data Lake
- ETL vs ELT
- Cloud provider Data Lake

Data Lake is?


Data Lakes Features
Ingest structured and unstructured data
Stores, secures and protects data at unlimited scale
Catalogs and indexes for analysis without data movement
Connects data with analytics and machine learning tools

DATA LAKE

Data

unstructured

Users

Data Scientists,
Data Analysts

Use cases

Stream Processing,
Machine Learning,
Real time analysis

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

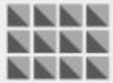
Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

vs

DATA WAREHOUSE

Data



Structured

Users

Business Analysts

Use cases

Batch Processing,
BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

How did it start?

- Companies realized the value of data
- Store and access data quickly
- Cannot always define structure of data
- Usefulness of data being realized later in the project lifecycle
- Increase in data scientists
- R&D on data products
- Need for Cheap storage of Big data

ETL vs ELT

- Extract Transform and Load vs Extract Load and Transform
- ETL is mainly used for a small amount of data whereas ELT is used for large amounts of data
- ELT provides data lake support (Schema on read)

Gotcha of Data Lake

- Converting into Data Swamp
- No versioning
- Incompatible schemas for same data without versioning
- No metadata associated
- Joins not possible

Cloud provider for data lake

- GCP - cloud storage
- AWS - S3
- AZURE - AZURE BLOB