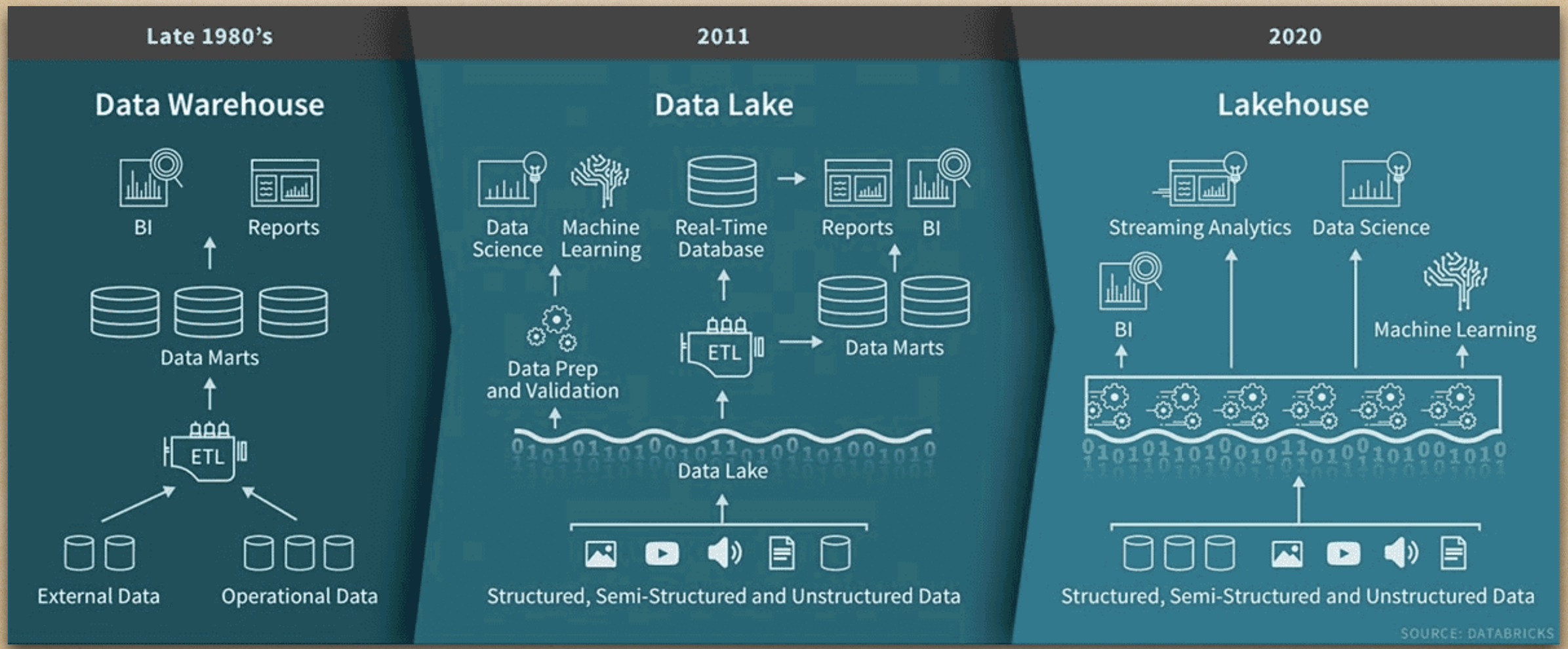
LAKEHOUSE Architecture

Credits : databricks.com

Different Types of Architectures from 1980 To Till



1980's Traditional Data Warehouse Projects

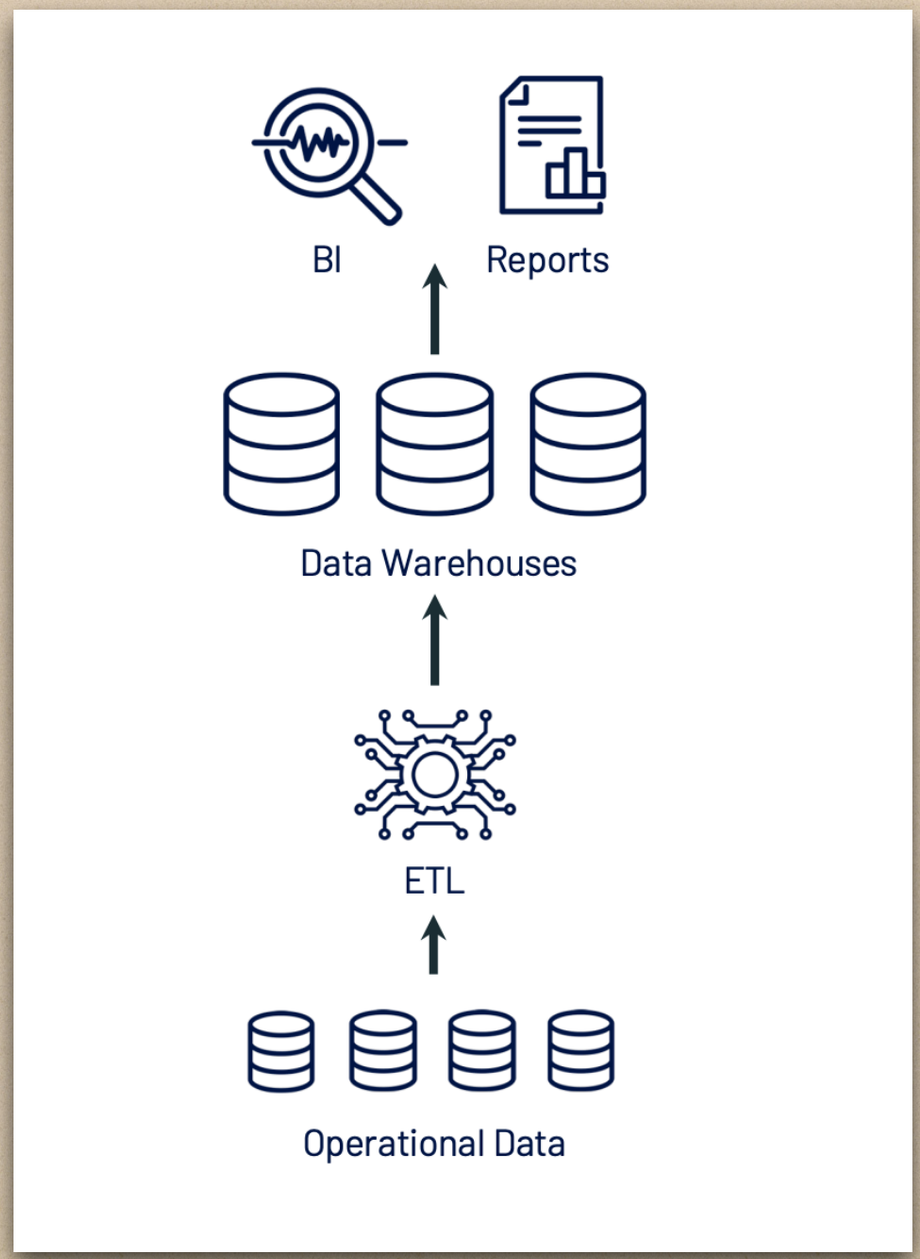
ORACLE ERP

SAP ERP

techcloud ERP solutions unlimited

Tally ERP.9

SAP Business One



ORACLE DATABASE

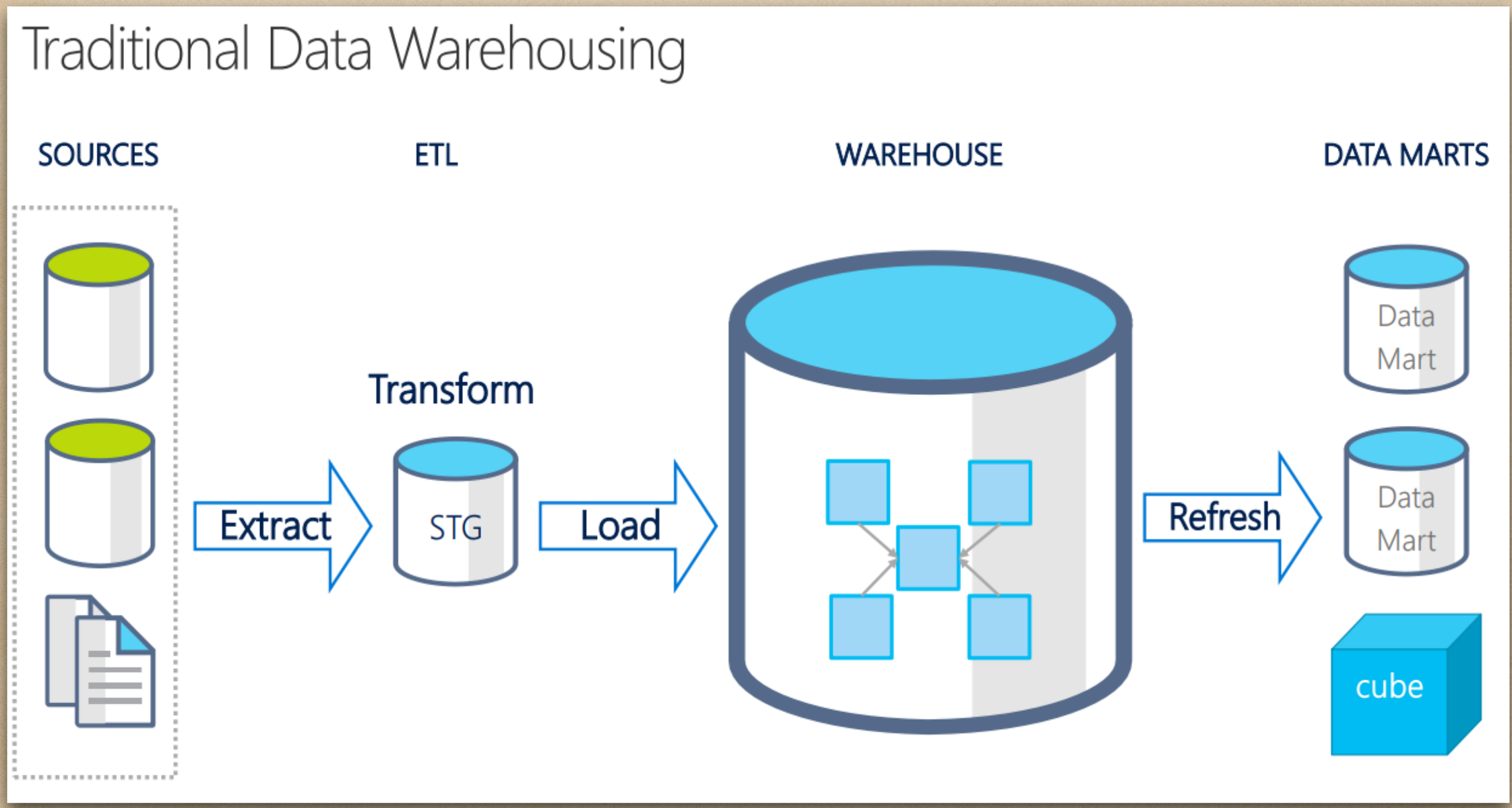
teradata.

IBM DB2

Microsoft SQL Server

MySQL

Traditional Data Warehouse Architecture



What is Data Warehouse?

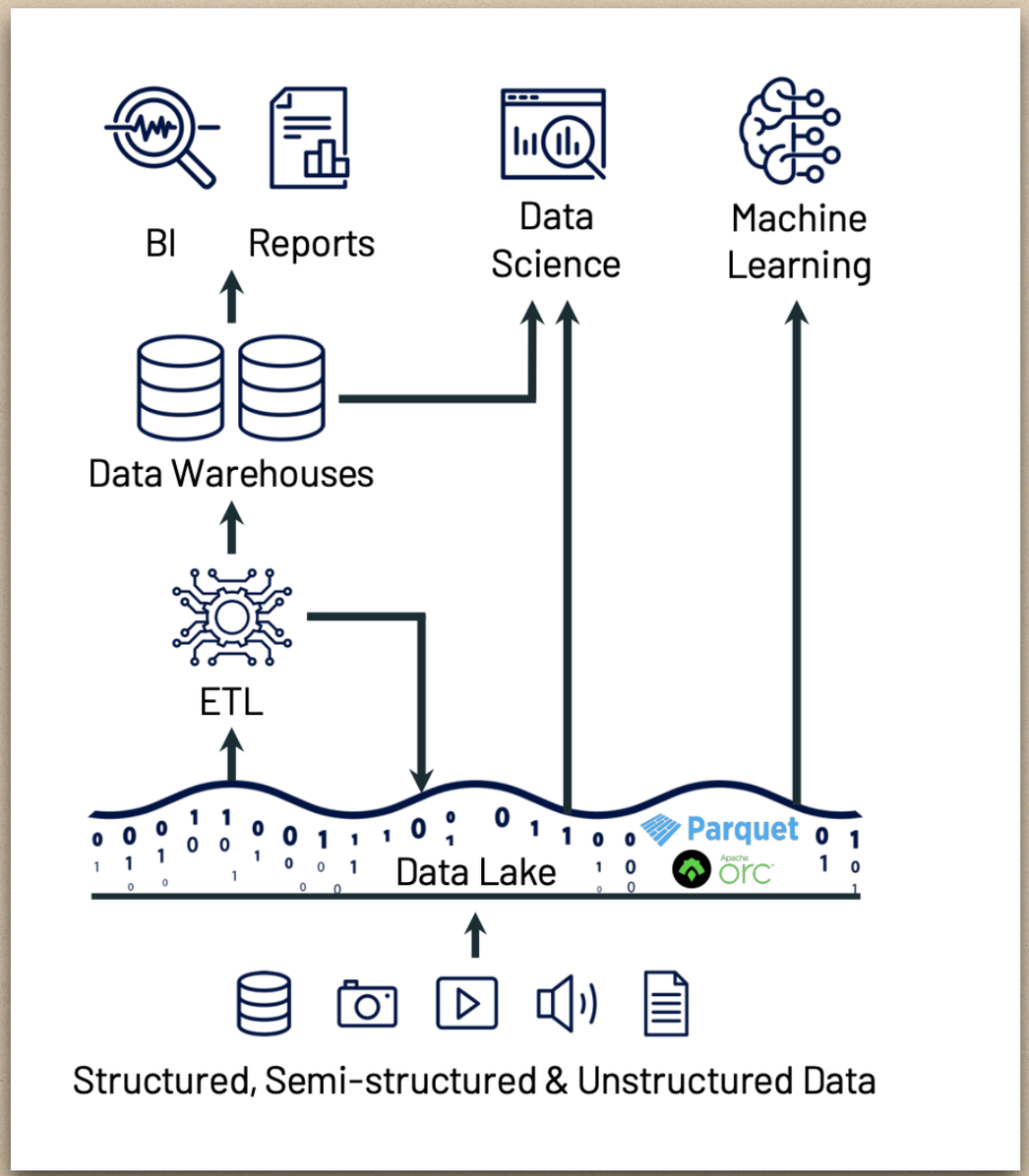
1. The Data Warehouse is your company's central data store.
2. A Data Warehouse is required for all companies that wish to make data-driven choices since it serves as the "Single Source of Truth" for all data in the company.
3. Analyzing Existing Business Data
4. New Business Models Launched Based on Existing Data Analysis
5. Enhance customer service and interaction
6. boost quality and resource utilisation

Are there any gaps in the Data Warehouse?

1. Data Warehouses only support Structured Data.
2. There is no support for video, audio, or text.
3. There is no support for data science or machine learning, and there is only limited support for streaming.
4. As a result, the majority of commercial and social media data is kept in data lakes and blob storage.

Current Data Lake + Data Warehouse Projects

Azure Data Lake Storage Gen2



Big Data Modern Data Lake + Cloud Warehouse

What is Data Lake?

A datalake is a centralized repository for storing, processing, and securing massive volumes of organized, semistructured, and unstructured data. It can store data in its native format and handle any type of data, regardless of size.

ADVANTAGES OF DATA LAKE

1. The primary benefit of data lakes is the concentration of several information sources.
2. At a cheaper cost, distributed and infinite storage
3. There are image, video, audio, json, csv, and other data storage types accessible.
4. **Volume and Variety**: A data lake can hold the massive amounts of data required by Big Data, artificial intelligence, and machine learning. Data lakes can manage the volume, diversity, and velocity of data absorbed in any format from numerous sources.
5. **Ingest Speed**: Format is unimportant during ingest. It employs schema-on-read rather than schema-on-write, which implies that data is not processed for use until it is required. Data can be swiftly written.
6. **Cost savings**: In terms of storage expenses, a data lake can be substantially less expensive than a data warehouse.
7. This enables businesses to acquire a broader range of data, such as unstructured data like rich media, sensor data from the Internet of Things (IoT), email, or social media.
8. **Greater Accessibility**: Data stored in a data lake makes it simple to open copies or subsets of data for multiple users or user groups to access. Data access might be restricted, while businesses can allow more accessibility.
9. **Advanced Algorithms**: Data lakes enable businesses to run complicated queries and deep learning algorithms to identify trends.

Disadvantages Of DataLake

1. Data appending is difficult.
   Adding additional data results in improper readings.
2. It is tough to modify existing data.
   GDPR/CCPA necessitates fine-grained adjustments to current data lakes.
3. Jobs that fail in the middle result in corrupted files.
   Half of the data is present in the data lake, but the other half is missing.
4. Operation in real time
   Inconsistency results from combining streaming and batch.
5. Expensive to preserve older data versions.
   Reproducibility, auditing, and governance are all required in regulated situations.
6. Difficulty in dealing with vast amounts of metadata
   The metadata itself becomes challenging to handle in vast data lakes.
7. "Too many files" issues
   Data lakes struggle to handle millions of little files.
8. It is difficult to obtain excellent performance.
   Data partitioning for performance is error-prone and difficult to modify.
9. Problems with data quality
   It is a continual challenge to guarantee that all data is correct and of high quality.
10. Incomplete ACID Transactions
    Insert, update, and delete operations are not possible due to a lack of ACID Properties.

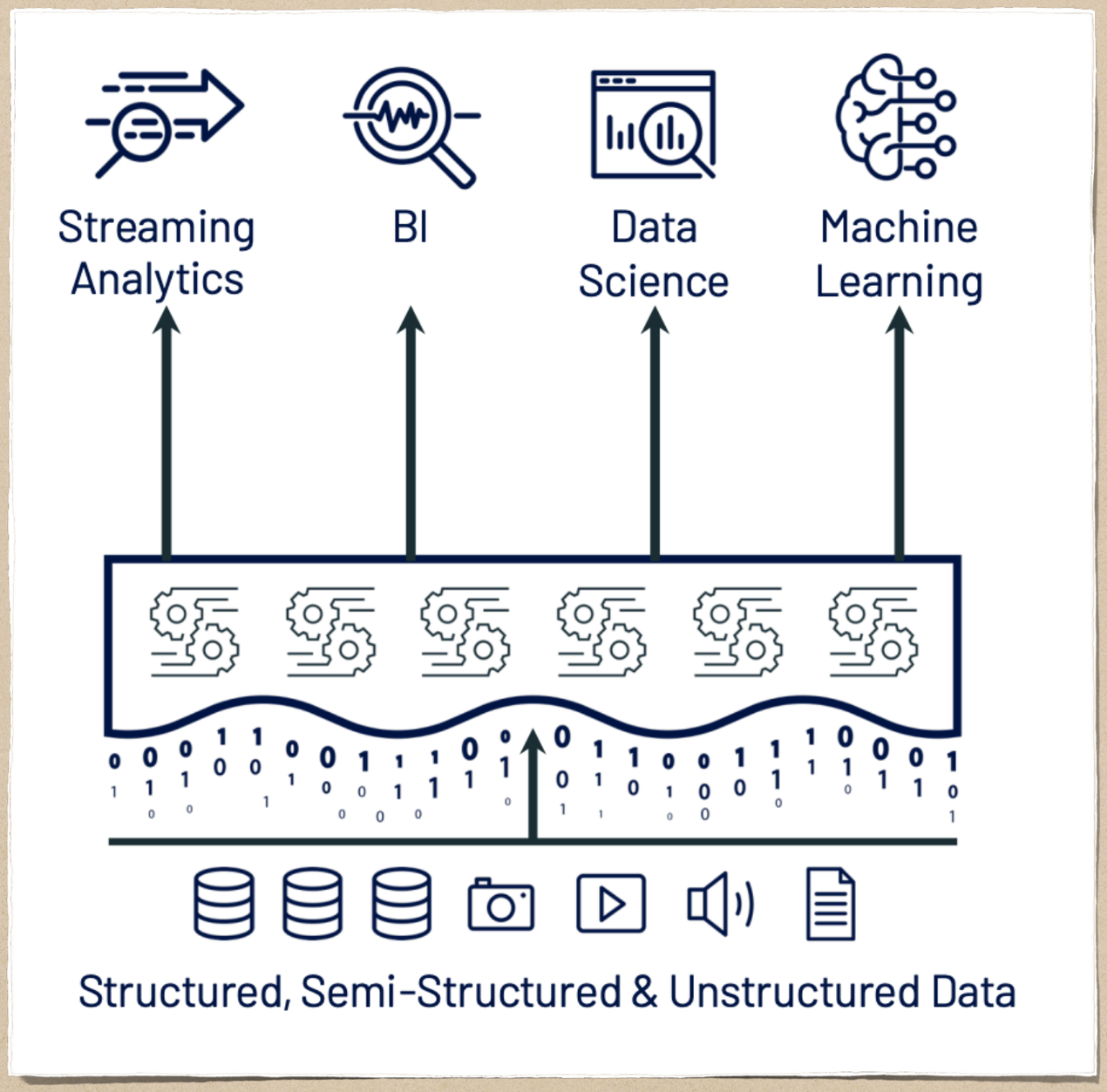
11 We require extra Warehouse and ETL processes to load from DataLake to DataWarehouse in order to satisfy Incomplete ACID Transactions and Metadata Catalog maintenance.

DataLake + DataWarehouse = LAKEHOUSE

Google Cloud Storage

Azure Data Lake Storage Gen2

amazon S3



databricks

DELTA LAKE

PySpark

Spark SQL

What exactly is DeltaLake?

Delta Lake is an open-source storage framework that allows for the creation of a Lakehouse architecture with computation engines such as Spark.
DeltaLake enables ACID transactions, scalable #metadata management, and the unification of streaming and batch data processing. Delta Lake is fully compatible with Apache Spark APIs and runs on top of your existing datalake.

1) Because it is ACID compliant, it provides guaranteed consistency.
2) Strong data storage. Data is stored in versioned parquet files.
3) It is intended to work with Apache Spark.
4) The audit logs will be kept in json format.
5) Time Travel Support

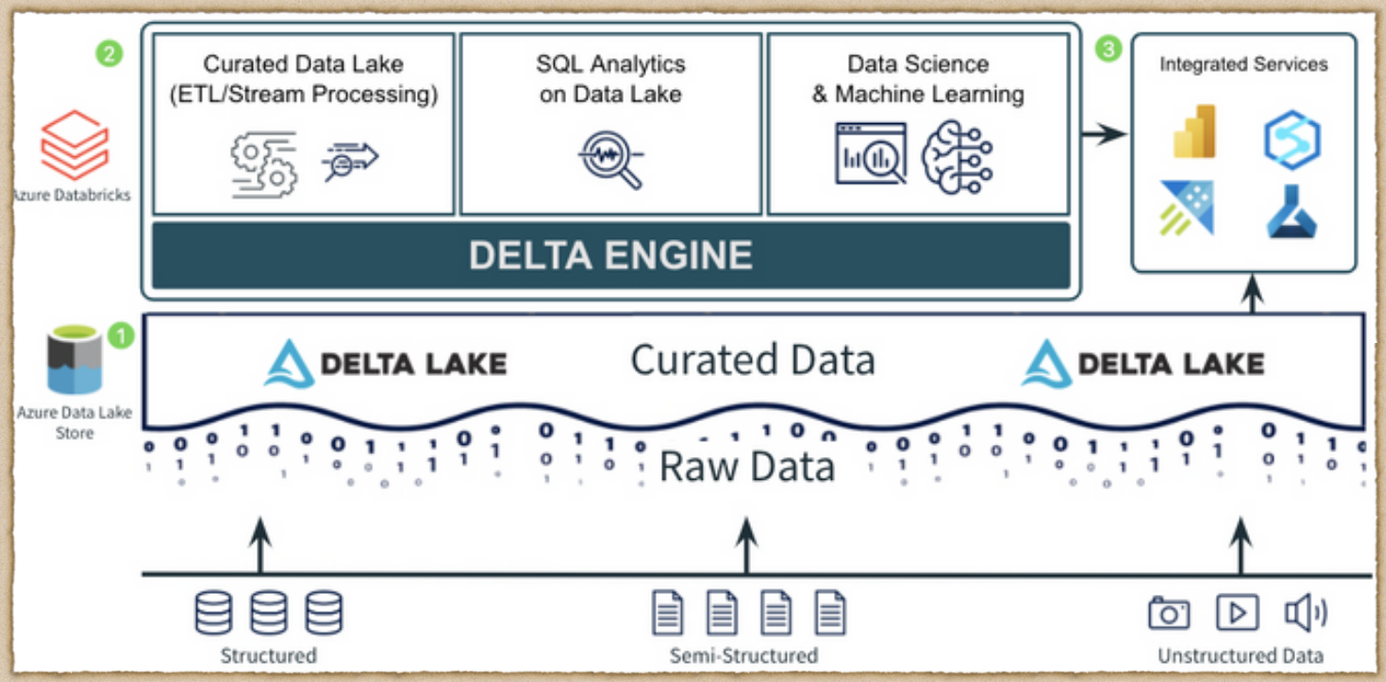
A lakehouse has the following features:
1) support for numerous data kinds and formats
2) data reliability and consistency help with a range of jobs (BI, data science, machine learning, and analytics)
3) direct use of BI techniques to source data
4) Metadata layers for data lakes
5) new query engine designs that enable high-performance SQL execution on data lakes
6) improved access to data science and machine learning tools
7) Data reading and writing at the same time.
8) Schema support combined with data governance methods.
9) Source data can be accessed directly.
10) Storage and computation resources are separated.
11) Storage formats that are standardised.
12) Structured and semi-structured data types, including IoT data, are supported.
13) End-To-End Batch and Streaming Loads.

Lakehouse Architecture

Google Cloud Storage

Azure Data Lake Storage Gen2

amazon S3



DELTA LAKE

Delta Lake Alternative.....

ICEBERG

Apache hudi

All The Best ✍️

Learn and Lead 👍

https://www.youtube.com/@TRRaveendra