Data Lakes & Data Warehouses

Data Management optimized for analytics and decisions

Lance Kataria, Shreya Mahajan, Vaibhav Patel, Julien Sanchez, Jonathan Thomas

AGENDA

01 BACKGROUND
- What is a Data Warehouse/Lake
- Business Context
- Key Terms
- Comparison

02 PRODUCT OVERVIEW
- Core Features
- Leading Products
- Research

03 TECHNICAL DETAILS
- Deep Dive into Two Products
- Differences
- Research

04 SAMPLE APPLICATIONS
- Real-World Uses
- Case Study
- Research
- Impact Overall

05 Future Trends
- How it will Evolve
- Challenges
- Breakthroughs

01

Background

Introduction

As businesses and research institutions become increasingly data-driven, the need for scalable, efficient, and intelligent data storage architectures has grown. **Data Warehouses** and **Data Lakes** are two dominant solutions for managing and analyzing vast amounts of structured and unstructured data.

These systems serve distinct but complementary purposes, supporting business intelligence, predictive analytics, AI-driven decision-making, and regulatory compliance.

While both serve as data management platforms, they cater to different requirements.

Data Warehouses are optimized for **structured, fast querying** with a focus on reporting, compliance, and BI.

Data Lakes provide **highly scalable, raw data storage** that supports big data processing, AI, and machine learning applications.

Understanding their roles requires a look at key terminology, and their impact on business operations.

Key Terms and Concepts

Data Warehouse: A centralized, structured repository that **integrates data from multiple sources**, storing it in a **predefined schema** for fast analytical queries and business intelligence.

Data Lake: A **highly scalable, flexible storage system** that holds **raw, unstructured, semi-structured, and structured data** in its native format for advanced analytics and machine learning.

Schema-on-Write vs. **Schema-on-Read**

- Data Warehouses use **Schema-on-Write**, meaning data is structured before storage, ensuring optimized querying.
- Data Lakes use **Schema-on-Read**, allowing data to be stored in its raw form and structured only when queried.

Structured, **Semi-Structured**, and **Unstructured Data**

- **Structured Data**: Data stored in **relational databases** (SQL tables, customer records).
- **Semi-Structured Data**: Data with **some structure** but no rigid format (JSON, XML).
- **Unstructured Data**: **Raw data** without predefined schemas (videos, images, text).

ETL (Extract, Transform, Load) vs. **ELT** (Extract, Load, Transform)

- **ETL** (Used in Data Warehouses): Data is extracted, transformed into structured format, and then loaded into storage.
- **ELT** (Used in Data Lakes): Data is loaded in raw form and **transformed when needed**, allowing flexibility in analytics.

Purpose and Business Context

Data Warehouses are critical for structured business analytics, providing:

- **Business Intelligence & Decision-Making**: Enables fast, SQL-based analytics for enterprise reporting.
- **Regulatory Compliance**: Ensures auditable data storage for industries like finance and healthcare.
- **Historical Trend Analysis**: Optimized for long-term storage and analytical queries.
- **Real-Time Data Insights**: Used in fraud detection, stock market analysis, and operational intelligence.

Data Lakes provide scalability and flexibility, supporting:

- **Big Data & AI Applications** – Stores vast raw datasets for deep learning, NLP, and AI model training.
- **IoT & Streaming Analytics** – Handles real-time data from IoT sensors, web traffic, and logs.
- **Cybersecurity & Threat Detection** – Stores and analyzes network logs for identifying cyber threats.
- **Data Science & Research** – Supports exploratory data analysis without requiring pre-structured schemas.

Businesses increasingly adopt a ***'Lakehouse'*** Model, which blends the structured querying power of warehouses with the flexibility of lakes. This allows:

- **Unified Data Storage** – A single repository for both structured BI reports and unstructured AI workloads.
- **Advanced AI/ML Workflows** – Allows direct ML model execution on stored data.
- **Multi-Cloud & Hybrid Architecture** – Data can be accessed across AWS, Azure, and Google Cloud.

Data Warehouse vs Data Lake

Feature	Data Warehouse	Data Lake
Data Type	Structured & semi-structured	Structured, semi-structured, and unstructured
Schema	Schema-on-Write (predefined)	Schema-on-Read (flexible)
Processing	Optimized for SQL-based queries	Supports SQL + AI/ML processing
AI/ML Support	Supports in-database ML for structured data	Supports large-scale AI/ML training on raw data
Performance	Fast queries, structured processing	High flexibility, batch & real-time scalability
Use Cases	Business intelligence, compliance, reporting	Big data analytics, AI/ML, IoT & real-time processing

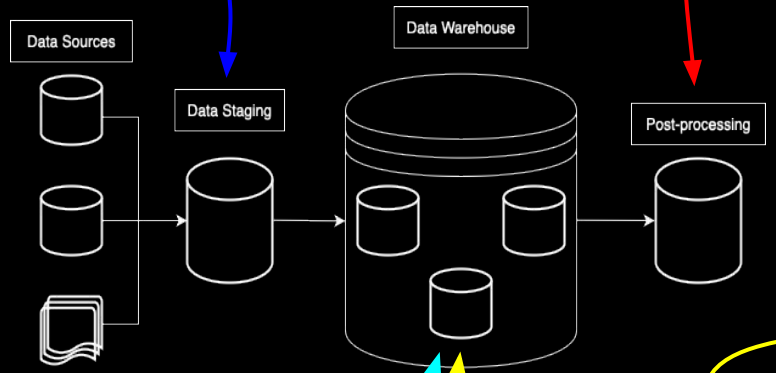
Key Terms and Concepts

ETL

Prepares and moves data to the warehouse.
(Extract, Transform, Load)

Columnar Storage

Stores data efficiently for analytical queries



Real-time Analytics

Processes data instantly for insights

Concurrency Staging

Adjusts resources based on demands

02

Product Overview

Features and Functions: Warehouses

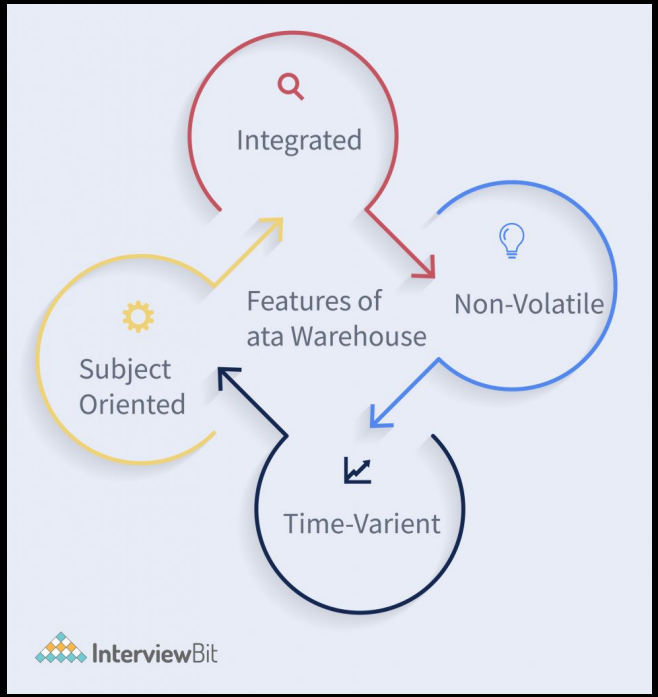
Scalability and Performance

Data Integration

Security and Compliance

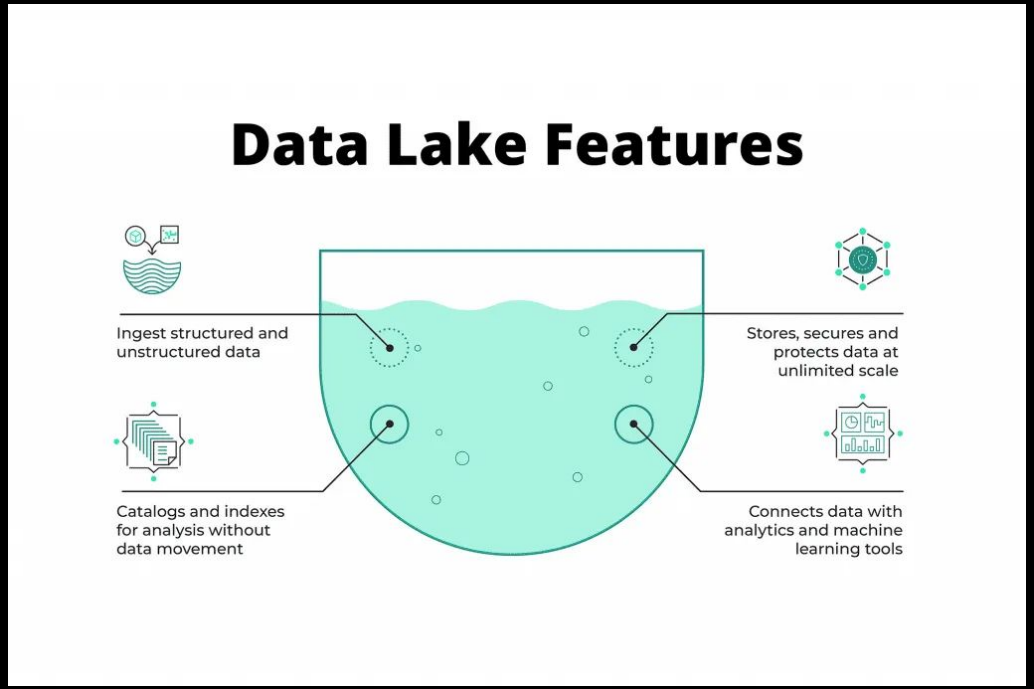
Subject-oriented

Non-Volatile



Source:
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.interviewbit.com%2Fblog%2Fcharacteristics-of-data-warehouse%2F&psig=AOvVaw3XGeLSE2BwN24s2hMilLOx&ust=1741209180705000&source=images&cd=vfe&opi=89978449&ved=0CBQ0jRxqFwoTCMjH3_er8YsDFQAAAAAdAAAAABAE

Features and Functions: Lakes



Source: https://www.google.com/url?sa=i&url=https%3A%2F%2Flakefs.io%2Fblog%2Fdata-lakes%2F&psig=AOvVaw2LH_LCjPxfmExx68uK8vRg&ust=1741209293010000&source=images&cd=vfe&opi=89978449&ved=0CBQQjRxqFwoTCLitr52s8YsDFQAAAAAdAAAAABAg

Warehouses	Architecture	Server Management	Deployment	Performance	Scalability
Amazon Redshift	Shared-nothing MPP	More self-managed	Cloud-based	Good	Vertical & Horizontal
Snowflake	Hybrid	More serverless	Cloud-based	High	Vertical & Horizontal
Google BigQuery	Shared-nothing MPP	Serverless	Cloud-based	Good	Vertical & Horizontal
Microsoft Azure	Shared-disk MPP	More self-managed	Cloud-based	High	Vertical & Horizontal

Table Info, Wording, and Source: https://www.altexsoft.com/blog/snowflake-redshift-bigquery-data-warehouse-tools/

Lakes	Lakehouse Architecture	Multi-cloud	Support for Mult. Prog. Languages	Decoupled Storage & Compute	Pipeline Service
Amazon S3	Yes	No	Yes	Yes	AWS Glue
Snowflake	Yes	Yes	Yes	Yes	Snowpark
Databricks	Yes	Yes	Yes	Yes	Delta Engine
Microsoft Azure	Yes	No	Yes	Yes	Azure Data Factory

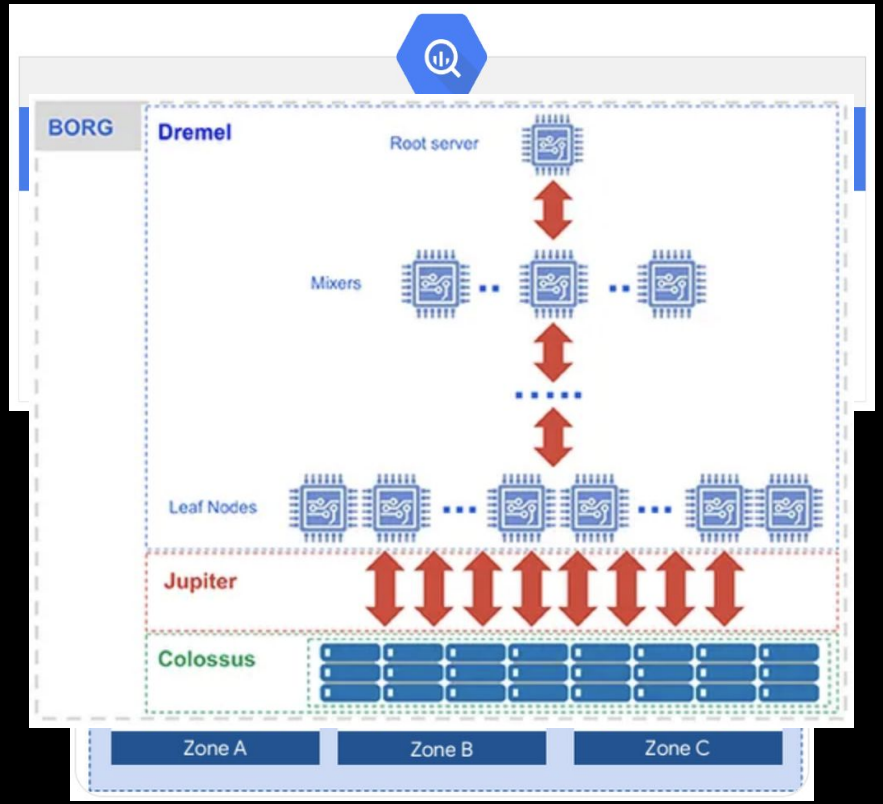
Table Info, Wording, and Source: https://www.polestarllp.com/blogcms/documents/cloud-data-lake-comparison-guide.pdf

03

Technical Details

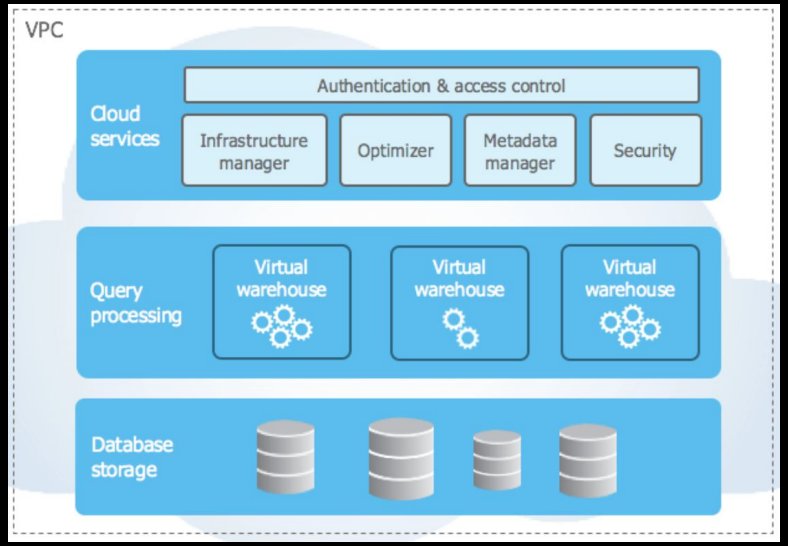
Google BigQuery

- Storage Layer
  - Data is reorganized and compressed into Capacitor Files
  - BigQuery stores data in the Colossus File System
- Compute Layer
  - Dremel Query Execution Engine makes use of multiple layers and nodes
  - Slots are distributed by Google's orchestration layer
  - "Mixers" rewrite queries and aggregate results, which are generated by leaf servers
- Orchestration
  - Invisible to users
  - Handles metadata, resource allocation, results caching, etc.
  - Cluster management system handles slot provisioning, assignment, and fault tolerance
  - Jupiter network enables rapid data transfer (13 PB/sec bisectional bandwidth)



Snowflake

- Multi-cloud platform; storage and compute resources
  - Users can choose from AWS, Google Cloud Platform, or MS Azure
- User-managed “virtual warehouses” operate independently to handle queries, and each cache relevant data to improve performance
  - multiple sizes (XS, S, M, L, XL, and so on up to 6XL in select regions)
  - Virtual warehouses are independent of one another, and allow you to run multiple workloads concurrently
- Cloud services layer handles query parsing and optimization
  - Metadata is stored in internal database and used to improve data lookup performance
  - Result caching
  - Role based and discretionary access control
- When data is loaded into Snowflake, the data is reorganized it into a columnar format and divided into smaller “micro-partitions”



Snowflake

- Multi-cloud platform in a single service
  - Storage and compute resource performance depends on underlying cloud-platform
- Compute resources are allocated in clusters as “virtual warehouses” and offers more user control
- Pricing is based on amount of compressed data being stored and warehouse usage time

Google BigQuery

- Single-cloud platform
  - Google BigQuery Omni can be used to run analytics on data stored in AWS and/or Azure
- Compute resources - dubbed “slots” - are allocated by Google as needed by the query being run
- Gives the option of either flat-rate or on-demand pricing for compute costs and an additional storage cost

Use Cases

- Can be used to store and analyze structured and semi-structured data for analytics, machine learning, and/or data sharing
  - Forbes uses BigQuery, BI tools, and AI/ML tools to analyze user interactions and use AI/ML to expedite and improve the writing process
  - Samsung Ads uses Snowflake to store and analyze raw (semi-structured) data, and for data sharing (Snowpark/Snowsight and Snowgrid)

04

Sample Applications

Use Case 1: Uber

- petabyte scale data from rides, Uber Eats, transactions, and interactions
- combination of lakes and warehouses

- real-time analytics
- ML & AI
- efficiency
- scalability
- cost effectiveness
- fraud detection
- risk analysis

Warehouse Architecture Data Flow

- data sources
  - rides, gps logs, payments, Uber Eats orders, customer interactions
- ingestion layer
  - Apache Kafka and Hadoop
- storage layer
  - Apache Hadoop
  - Apache Hive & Presto
  - Apache Hudi
- processing layer
  - Apache Spark & Hive
- querying
  - Apache Presto & Pinot

THE APACHE® SOFTWARE FOUNDATION

Use Case 2: Netflix

- 200 million+ subscribers
- data lakes and warehouses

- personalized content
- stream performance
- real-time analytics
- scalability
- cost efficiency
- open source

Warehouse Architecture Data Flow

- data sources
  - play/pause/search actions, content metadata
- ingestion layer
  - Apache Kafka and AWS Glue
- storage layer
  - AWS S3
  - Apache Iceberg
- processing layer
  - Apache Spark and AWS EMR
  - Apache Flink
- querying
  - Trino
  - Apache Druid

05

Future Trends

Prognosis and Future Evolution

- Traditional data warehouses are evolving to handle real-time analytics and AI driven insights

- Rise in cloud data warehouses
  - Scalability, reduced costs

- Companies switching to decentralized model called data mesh
  - Different teams can manage their own data independently instead of relying on a single large database

Future Directions

Real Time Data Processing

- Businesses need real-time data analysis to make quick decisions
    - Technologies like Apache Kafka and Apache Flink help process data instantly

Security and Privacy

- Increasing use of AI in detecting security risks and preventing data leaks
- Researchers developing stronger encryption methods to keep data safe

Convergence and Automation

- Data lakes and warehouses merging into “lakehouse” architectures
    - Combining structured and unstructured data for better performance
- AI automating data management and quality checks

Cost-effective and Scalable Storage

- Cloud services like AWS and Google cloud offer pay-as-you-go models
- Companies use efficient data formats (Parquet, ORC) -> faster to search for data

Challenges

Large-Scale Queries

- Large scale queries can become very expensive
- These queries can also become much slower
- New solutions improving cross-cloud query engines

Data Quality and Consistency

- Data lakes store raw unstructured data
  - Can be harder to ensure consistency and accuracy
- New solutions creating AI-driven data validation, automated schema enforcement

AI and ML Integration

- Traditional SQL-based warehouse not designed for complex ML workflows
- New tools like BigQuery ML and Databricks make it easier to run ML models in data platforms

THANK YOU