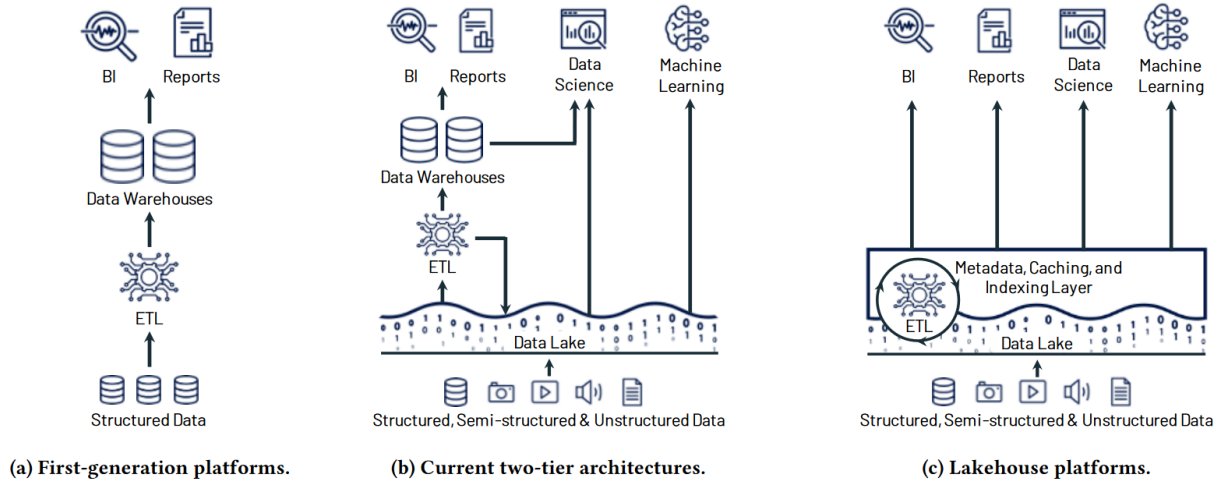


01 Lakehouse

Lakehouse Architecture



Introduction

The Lakehouse architecture represents a paradigm shift in modern data engineering, combining the best features of traditional data warehouses and data lakes. This architecture has become fundamental to virtually every SaaS application and data engineering solution today.

Historical Context: Evolution of Data Architecture

1980s: The Era of Data Warehousing

Key Characteristics:

- **Primary Technology:** SQL-based data warehouses
- **Data Type:** Exclusively structured data (rows and columns)
- **Data Model:** Dimensional Data Modeling
- **Architecture:** Facts and Dimensions tables

Dimensional Data Modeling

- **Definition:** A technique for organizing data into fact tables (containing metrics) and dimension tables (containing descriptive attributes)
- **Components:**
 - **Fact Tables:** Store quantitative data (sales amounts, quantities, counts)
 - **Dimension Tables:** Store descriptive attributes (customer info, product details, time periods)
- **Schema Types:**
 - Star Schema (denormalized dimensions)
 - Snowflake Schema (normalized dimensions)

- Galaxy Schema (multiple fact tables)

Advantages of Traditional Data Warehouses:

- High query performance for structured data
- ACID compliance
- Strong consistency
- Mature SQL ecosystem
- Excellent for business intelligence and reporting

The Big Data Challenge: The Three V's Problem

Volume

- **Definition:** Exponential growth in data size
- **Drivers:**
 - Internet proliferation
 - Social media expansion
 - IoT device proliferation
 - Digital transformation
- **Scale:** From gigabytes to petabytes and exabytes

Velocity

- **Definition:** Speed of data generation and processing requirements
- **Characteristics:**
 - Real-time data streams
 - High-frequency data ingestion
 - Near real-time analytics requirements
- **Examples:** Stock trades, sensor data, clickstreams

Variety

- **Definition:** Diversity of data types and formats
- **Types:**
 - **Structured:** Traditional relational data
 - **Semi-structured:** JSON, XML, Avro, Parquet
 - **Unstructured:** Images, videos, audio, text documents, logs

Extended V's (Modern Additions):

- **Veracity:** Data quality and reliability
- **Value:** Business value extraction from data
- **Variability:** Inconsistency in data flows

Early 2000s: The Data Lake Revolution

What is a Data Lake?

A centralized repository that allows storage of structured, semi-structured, and unstructured data at any scale without requiring upfront schema definition.

Key Advantages:

1. **Cost Efficiency:** Lower storage costs compared to traditional warehouses
2. **Flexibility:** Schema-on-read approach
3. **Scalability:** Horizontal scaling capabilities
4. **Variety Support:** Can store any data format
5. **Raw Data Preservation:** Maintains original data fidelity

Popular Data Lake Technologies:


- **Cloud:** Amazon S3, Azure Data Lake, Google Cloud Storage
- **On-premises:** Hadoop HDFS
- **Processing:** Apache Spark, Hadoop MapReduce

Data Lake Limitations:


1. **Performance Issues:** Slower query performance for analytical workloads
2. **Data Governance:** Lack of ACID transactions
3. **Schema Evolution:** Difficult to manage schema changes
4. **Data Quality:** Risk of becoming "data swamps"
5. **Security:** Limited fine-grained access controls

Stakeholder Analysis: The Performance Dilemma

Data Scientists

- **Requirements:** Large volumes of raw data, flexibility, ML/AI capabilities
- **Data Lake Satisfaction:**  High - appreciated the flexibility and variety
- **Use Cases:** Feature engineering, model training, exploratory data analysis

Data Analysts/Business Intelligence Developers

- **Requirements:** Fast query performance, reliable data, structured access
- **Data Lake Satisfaction:**  Low - poor report performance
- **Use Cases:** Dashboards, reports, business intelligence
- **Impact:** Slower decision-making for business leaders

Business Leaders

- **Requirements:** Real-time insights, reliable reporting, data-driven decisions
- **Challenge:** Cannot make quick decisions with slow data access

The Lakehouse Solution: Best of Both Worlds

Definition

A Lakehouse is a data architecture that combines the flexibility and cost-effectiveness of data lakes with the performance and reliability of data warehouses.

Core Concept

Lakehouse = Data Warehouse + Data Lake

Key Benefits

From Data Lakes:

- ☒ Support for all data types (structured, semi-structured, unstructured)
- ☒ Cost-effective storage
- ☒ Scalability
- ☒ Raw data preservation
- ☒ Machine learning and AI workloads

From Data Warehouses:

- ☒ High-performance SQL queries
- ☒ ACID transactions
- ☒ Data governance and quality
- ☒ Business intelligence capabilities
- ☒ Concurrent user support

Technical Implementation: Open Table Formats

What Enables Lakehouse Architecture?

The key to combining data warehouse capabilities with data lake storage lies in **open table formats** and **metadata layers**.

Delta Lake (Databricks)

- **ACID Transactions:** Ensures data consistency
- **Schema Evolution:** Handle changing data schemas
- **Time Travel:** Access historical versions of data
- **Unified Streaming and Batch:** Single platform for all processing
- **Data Versioning:** Track changes over time

Other Open Table Formats:

- **Apache Iceberg:** Netflix-originated, supports schema evolution
- **Apache Hudi:** Uber-originated, focuses on incremental processing
- **Parquet + metadata:** Column-oriented storage with rich metadata

Key Technical Features:

1. **Metadata Management:** Rich metadata for performance optimization
2. **Indexing:** Advanced indexing for faster queries
3. **Caching:** Intelligent caching mechanisms
4. **Vectorized Processing:** Optimized compute engines
5. **Data Skipping:** Skip irrelevant data during queries

Modern Lakehouse Platforms

Databricks Lakehouse Platform

- **Unity Catalog:** Centralized governance
- **Delta Engine:** High-performance query engine
- **MLflow:** Machine learning lifecycle management
- **Auto Scaling:** Dynamic resource management

Other Major Players:

- **Snowflake:** Cloud data platform with lakehouse capabilities
- **Amazon Redshift Spectrum:** Query data in S3
- **Google BigQuery:** Serverless data warehouse with lake capabilities
- **Microsoft Synapse:** Unified analytics platform

Implementation Considerations

Architecture Patterns:

1. **Lambda Architecture:** Separate batch and stream processing
2. **Kappa Architecture:** Stream-first approach
3. **Modern Lakehouse:** Unified approach for all workloads

Data Organization:






- **Bronze Layer:** Raw, unprocessed data
- **Silver Layer:** Cleaned and validated data
- **Gold Layer:** Business-ready, aggregated data

Governance Framework:






- **Data Catalog:** Searchable metadata repository
- **Lineage Tracking:** Data flow documentation
- **Access Controls:** Fine-grained permissions
- **Quality Monitoring:** Automated data quality checks

Benefits and Challenges

Benefits:

-  Unified platform for all analytics workloads
-  Reduced data movement and duplication
-  Lower total cost of ownership
-  Faster time-to-insight
-  Support for emerging technologies (AI/ML)

Challenges:

-  Complexity in implementation
-  Skills gap and training requirements
-  Integration with legacy systems
-  Vendor lock-in considerations
-  Data security and compliance

Future Outlook

Emerging Trends:

- **Real-time Analytics:** Streaming analytics capabilities

- **AI/ML Integration:** Native AI/ML support
- **Multi-cloud:** Cross-cloud data sharing
- **Edge Computing:** Processing at the edge
- **Quantum Computing:** Future computational paradigms

Industry Adoption:

- **Current State:** Rapidly growing adoption across industries
 - **Future:** Expected to become the standard data architecture
 - **Investment:** Major cloud providers investing heavily
-

Conclusion

The Lakehouse architecture represents the natural evolution of data platforms, addressing the limitations of both traditional data warehouses and data lakes. By providing a unified platform that supports diverse workloads while maintaining performance and governance, Lakehouses have become essential for modern data-driven organizations.

Key Takeaway: The Lakehouse isn't just a technology choice—it's a strategic approach to building scalable, flexible, and performant data platforms that can adapt to evolving business needs.