Data Mining

Data Lakes in the Data-Driven Decisions Pipeline

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

The Need for Data-Driven Decisions

- Modern organizations generate massive amounts of data:
  - Transactions, IoT sensors, logs, clickstreams
  - Images, documents, videos, social media
- Business competitiveness depends on timely, data-informed decisions
- Traditional Data Warehouses cannot always keep up:
  - Structured-only focus
  - Expensive and rigid schemas
  - Hard to scale to all enterprise data

Key Challenges Today

- Data silos: Marketing, Sales, Operations all keep separate datasets
- Volume: Petabytes of logs, sensor streams, customer interactions
- Variety: Structured (SQL tables), semi-structured (JSON, XML), unstructured (text, video)
- Velocity: Real-time decision-making requires streaming ingestion
- Value: Data is underutilized without proper consolidation

What is a Data Lake?

- A Data Lake is a centralized repository that stores all types of data:
  - Structured, semi-structured, and unstructured
  - Raw format, at scale
- Data is stored with minimal transformation (*schema-on-read*)
- Enables advanced analytics, AI/ML, and self-service exploration

Data Lake vs Data Warehouse

Data Warehouse

- Schema-on-write
- Structured, curated data
- Optimized for BI reporting
- Expensive storage

Data Lake

- Schema-on-read
- All data types (raw + curated)
- Supports BI + ML + advanced analytics
- Cheap, scalable storage (cloud, HDFS, S3)

Motivation for Data Lakes

- Ingest all enterprise data without prior modeling
- Provide a foundation for data science and AI
- Complement (not replace) warehouses: - Warehouse: curated BI reporting - Lake: exploration + ML/AI + diverse data
- Support the Data-Driven Decisions Pipeline end-to-end

From Data Sources to Decisions

Raw data sources → storage → processing → insights → actions

- A Data Lake is often the first stop for raw enterprise data
- Enables both:
  - Exploratory analytics (data scientists)
  - Operational dashboards (managers)

Pipeline Requirements

- Scalable ingestion (batch + streaming)
- Governance and security (access controls, lineage)
- Multi-modal processing: SQL, ML, streaming
- Low-latency insights for decision-making

Benefits for Management

- Single source of truth for all enterprise data
- Faster, evidence-based decision-making
- Support for advanced KPIs and predictive analytics
- Flexibility to support new business cases

Benefits for ITC Personnel

- Cheap, elastic storage of petabyte-scale datasets
- Unified platform: avoid proliferation of silos
- Easier data integration from multiple sources
- Enablement of hybrid workloads: BI + ML + AI

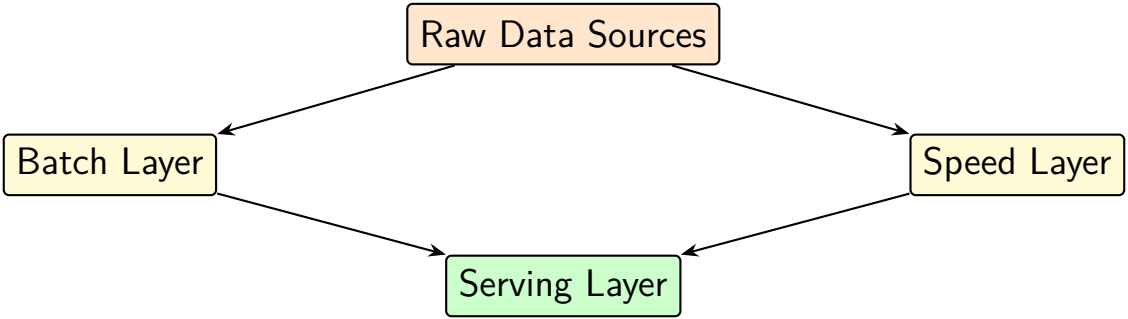
Challenges of Data Lakes

- Risk of becoming a data swamp without governance
- Data quality issues if ingestion is uncontrolled
- Metadata management is critical
- Security and compliance must be enforced

Governance in Data Lakes

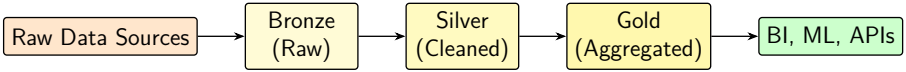
- Metadata cataloging (Glue, Hive Metastore, Data Catalog)
- Role-based access control and auditing
- Data lifecycle management (retention, archival)
- Integration with data quality frameworks

Lambda Architecture



Pros: Completeness + freshness Cons: Two code paths, complexity

Lakehouse / Medallion Architecture



Pros: Unified, ACID, governance Cons: Still maturing

Evolution Timeline

Use Cases of Data Lakes

- Customer 360 degrees view across all touchpoints
- Real-time fraud detection in financial services
- Predictive maintenance in manufacturing
- Omnichannel personalization in retail
- Genomic and medical data analysis in healthcare

Retail Example

- Raw data: POS transactions, loyalty cards, clickstream
- Data Lake: Store structured + unstructured data at scale
- Warehouse: Curated sales dashboards
- Advanced analytics: Predict promotions, optimize inventory

Why Data Ingestion Matters

- Ingestion is the entry point to the Data Lake
- Quality of ingestion directly affects:
  - Data completeness
  - Latency of insights
  - Data quality and trust
- Must handle both batch and streaming sources

Types of Data Ingestion

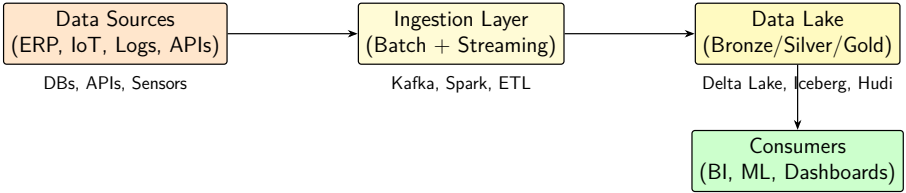
Batch Ingestion

- Periodic loads (e.g., nightly ETL)
- Best for stable, large datasets
- Example: ERP exports to Data Lake

Streaming Ingestion

- Continuous, real-time flow
- Best for logs, IoT, clickstreams
- Example: Kafka, Kinesis, Pulsar

Ingestion Pipeline Architecture



Best Practices in Data Ingestion

- Use a unified ingestion framework to avoid "ad–hoc" methods for each application
- Apply data quality checks early in the pipeline
- Design for scalability (cloud-native, serverless)
- Ensure security and compliance (PII masking, GDPR)
- Makes data discoverable, reusable, and governed
- Support schema evolution gracefully

Conclusion

- Data Lakes are critical in the modern data-driven pipeline
- They enable unified storage of all enterprise data
- Together with Lakehouse architectures, they support BI, ML, and AI
- Governance is essential to avoid “data swamps”

What to do in an organisation

- Establish a scalable (cloud-based?) Data Lake
- Define governance, security, and metadata cataloging
- Integrate warehouse and lake for a unified Lakehouse strategy
- Train analysts, engineers, and business users in usage