databricks

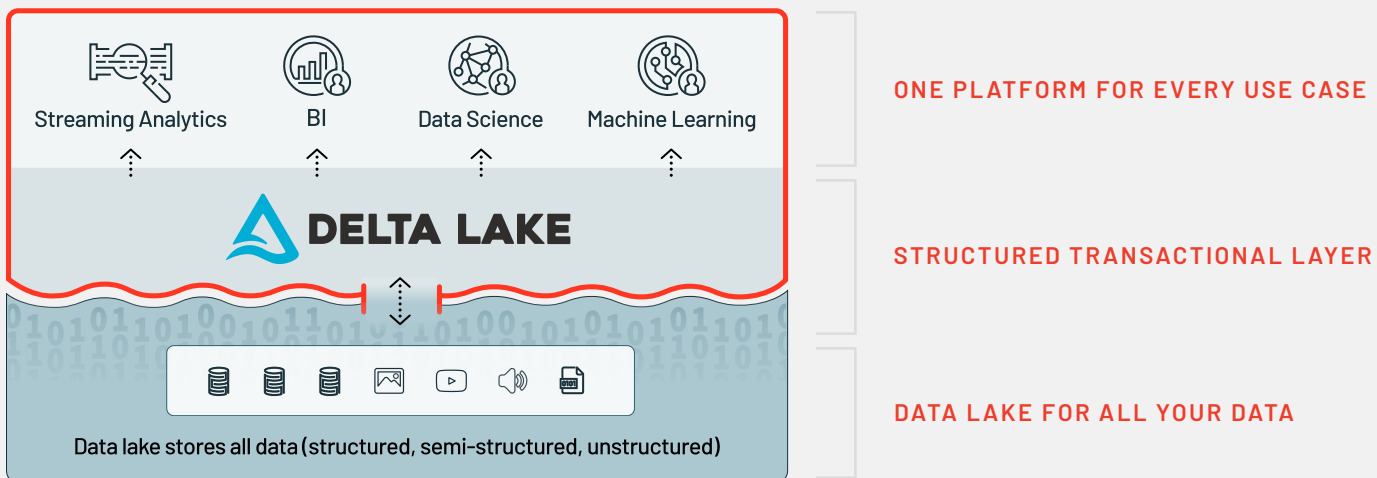# Lakehouse Build-Out Package

Implement your lakehouse vision with Databricks Delta Lake

Executive summary

Combining all your data in a traditional data warehouse is an anti-pattern that requires a lot of ETLs on a rigid platform that mainly supports structured data geared toward mission-critical reporting and BI use cases. To leverage ML and AI effectively and be a better, more data-driven organization, you need a unified repository supporting large and diverse data sets, including semi and unstructured, in an open format with enterprise-grade reliability and performance. Welcome to lakehouse, a new architecture that combines the best elements of data lakes and data warehouses. This packaged offering from Databricks will accelerate implementing one simplified platform for data analytics, data science and ML — setting the foundation for your lakehouse vision.

Key outcomes

- Modern multi-hop pipeline reference architecture per your chosen scenario
- A reference implementation for 1–2 scoped data pipelines of your choice
- Reference consumer layer that integrates with SQL Analytics or orchestration
- Extended options for additional pipeline scenarios plus performance optimizations

Streaming Analytics   BI   Data Science   Machine Learning

DELTA LAKE

Data lake stores all data (structured, semi-structured, unstructured)

ONE PLATFORM FOR EVERY USE CASE

STRUCTURED TRANSACTIONAL LAYER

DATA LAKE FOR ALL YOUR DATA

Overview

Build or optimize a Delta Lake as the structured transaction layer on top of low-cost storage following the most up-to-date best practices, laying the foundation of your lakehouse vision. The package addresses many of the common problems found with data lakes by leveraging Delta Lake and its associated architecture patterns. The package offers three tiers: Foundation, Extended, and Optimized. Milestones and the outcome for each tier are produced by our prescriptive methodology, and each tier can be chained for greater impact on your journey of realizing your lakehouse vision. See the Resources and schedule section for details.

Common challenges with data lakes and data warehouses

DATA LAKE

- Data integrity (failed writes)
- Lack of consistency (multiple readers/writers)
- Schema mismatch
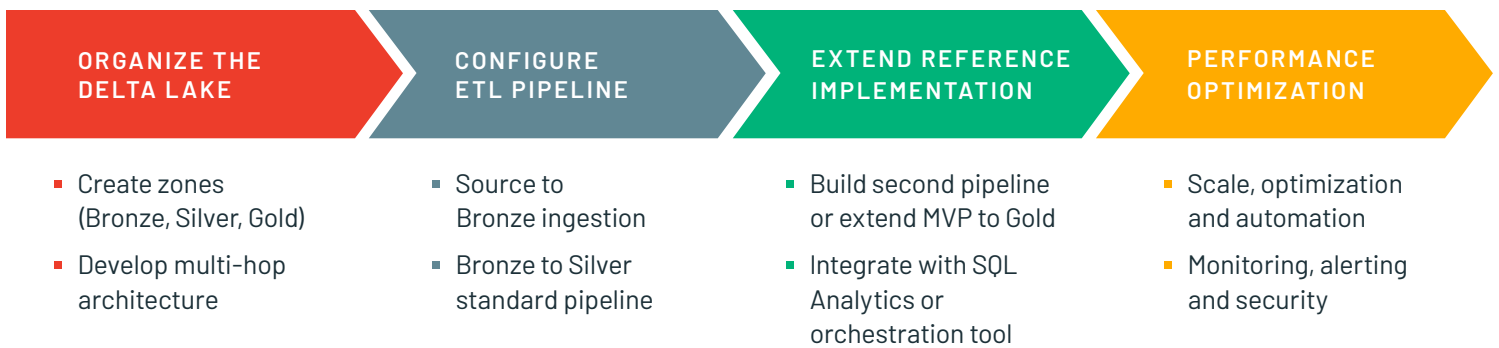- Complex Lambda architecture

DATA WAREHOUSE

- Mainly supports structured data
- Not the most cost-efficient
- Lacks open standards and formats
- Largely only stores refined data sets

Key benefits of lakehouse with Delta Lake

- Greater data reliability and scalability
- Unified batch and streaming
- Modern data platform ready for your ML and AI initiatives

- Data teams can collaborate and innovate faster
- Simplifies data and AI with a single platform
- Enjoy the data lake economics

Databricks lakehouse build-out process

ORGANIZE THE DELTA LAKE	CONFIGURE ETL PIPELINE	EXTEND REFERENCE IMPLEMENTATION	PERFORMANCE OPTIMIZATION

- Create zones (Bronze, Silver, Gold)
- Develop multi-hop architecture

- Source to Bronze ingestion
- Bronze to Silver standard pipeline

- Build second pipeline or extend MVP to Gold
- Integrate with SQL Analytics or orchestration tool

- Scale, optimization and automation
- Monitoring, alerting and security

Resources and schedule

FOUNDATION
2 WEEKS, $40K*

- Reference architecture
- Reference implementation
- Consumer enablement

EXTENDED
2 WEEKS, $45K*

- Additional pipeline or Gold
- SQL Analytics or orchestration integration
- Optimized compute defined

OPTIMIZED
2 WEEKS, $45K*

- Scaled performance optimization
- CI/CD and automation
- Monitoring, alerting and security

*Up to 4 resources supporting the activity over a 2-week sprint
Prior to kickoff, be sure to review the readiness checklist and complete required tasks

Out of scope

- Configuration and integration of non-Databricks products and systems
- Data cleansing and solving data quality issues