databricks

Data Lake Optimization Package

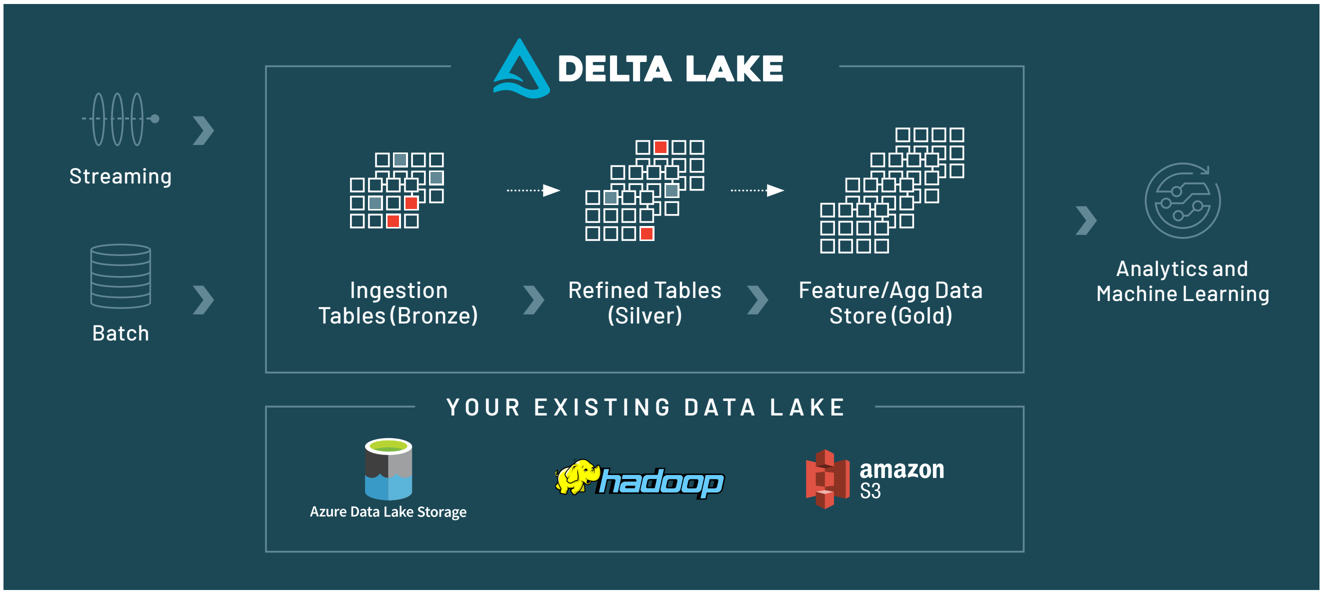
Solve your big-data challenges with Databricks Delta Lake

Summary

Combining all of your data in a traditional data warehouse is an anti-pattern that requires a lot of ETLs. This rigid approach is limited, as it mainly supports structured data geared toward mission-critical reporting and BI use cases. To leverage ML and AI effectively and be a better data-driven organization, you need a unified repository that supports large and diverse data sets, including semi and unstructured data in an open format with enterprise-grade reliability and performance. This packaged offering from Databricks will accelerate building and optimizing your data lake modernization effort.

Key outcomes

- Build modern multi-hop reference architecture for a co-selected data pipeline or scenario
- Build a reference implementation for scoped data pipelines of your choice
- Build reference consumer layer with integrations with BI or orchestration tools
- Extend, adding pipeline scenarios plus performance optimizations



Strategy

Build and optimize a data lake with the most up-to-date best practices, guided by experts. The package addresses many of the common challenges faced with data lakes by leveraging Delta Lake and its associated architecture patterns. The package offers three tiers: **Foundation**, **Extended** and **Optimized**. Milestones and outcomes for each tier are produced by our prescriptive methodology, and each tier can be chained for greater impact on your data lake modernization effort. See the **Resources and schedule** section for details.

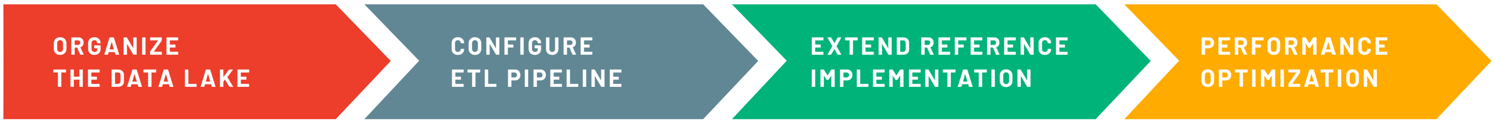
Common problems and challenges with data lakes

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

Key benefits

- Greater data reliability and scalability
- Unified batch and streaming
- Modern data lake ready for your ML and AI initiatives
- Faster insights from your data
- Drive data lake usage via optimized self-service

Databricks data lake build and optimization process



- 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 BI or Orchestration tool

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

Resources and schedule

FOUNDATION 2 weeks, $40K	EXTENDED 2 weeks, $45K	OPTIMIZED 2 WEEKS, $45K
▪ Reference architecture ▪ Reference implementation ▪ Consumer enablement	▪ Additional pipeline OR Gold ▪ BI or Orchestration integration ▪ Optimized compute defined	▪ Scaled performance optimization ▪ CI/CD and automation ▪ Monitoring/alerting/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

© Databricks Inc. 2020. All rights reserved.