databricks

# Hadoop Migration Package

**Migrate your on-premises Hadoop workloads to the Databricks Unified Data Analytics Platform**

## Summary

The Databricks Hadoop Migration Package gives you a concrete migration outcome through a prescriptive approach that maximizes the value of your existing data and pipeline investments — and delivers a trouble-free migration. The Databricks Unified Data Analytics Platform enables your data and AI team to holistically analyze data from data warehouses, data lakes and other data stores by utilizing a single engine for batch, ML, streaming and real-time queries. By migrating to the Unified Data Analytics Platform, you give your teams access to data in one place and the ability to automate the most complex data pipelines, running and maintaining machine learning at scale — so you can innovate faster.

## Key outcomes

A Findings and Best Practices recommendation report tailored for your Databricks environment, configuration, architecture and processes, covering key areas:

- Build a phased migration plan from migration inventory discovery and analysis
- Build target reference architecture
- Build reference implementation — one migration scenario co-selected by you and Databricks
- Extend with more migration scenarios and productionization

## Strategy

We tackle your key migration scenarios, ensuring the journey has a sound foundation and robust plan. And reference implementation serves as a design pattern for similar workloads. The packages include the following:

- Migrate data, code, ETL pipeline and job orchestration
- Execute pilot for ingestion and integration
- Performance testing
- Productionization hardening best practice
- Code and performance optimization

| HADOOP | | DATABRICKS |
|---|---|---|
| Data Eng, ML (Spark) | → | Databricks jobs/Delta Lake |
| ETL, SQL (Hive/Impala) | → | Databricks jobs/Delta Lake |
| Real-time event processing (Storm/Spark) | → | Databricks Structured Streaming |
| Batch Process (MapReduce) | → | Databricks Spark jobs |
| Scalable apps on columnar store (HBase) | → | Databricks Spark integrates with HBase on cloud (alternatively use cloud data stores) |

### Challenges and risks of on-premises Hadoop

- Heavy DevOps burden
- Inflexible clusters and sizing
- Data reliability and performance issues
- Data and AI team siloed
- Missed opportunity

### Key benefits

- Focus on business not infrastructure
- Save cost from elastic scaling
- Improve productivity
- Separate storage and compute
- Faster time to market
- Accelerate your innovation
- No lock-in to aging hardware
- No more upgrade hassles

### Out of scope

- Data cleansing and solving data quality issues

databricks

# Databricks migration methodology


## DISCOVERY, PLANNING AND ROAD MAP

- Gather all necessary background on the current Hadoop environment
- Build an inventory
- Develop a phased migration plan based on data and pipeline priority, dependency and project criticality

## TARGET DELTA ARCHITECTURE

Ensure target data architecture leverages data lake best practices with scalability and flexibility supporting varying workloads (e.g., batch, streaming)

## MIGRATION

- Data migration
- Cluster migration planning
- Set up Databricks environment
- Metastore migration
- Migrate code/pipeline
- Data validation
- CI/CD and automation readiness

## VALIDATION

Validate the outcome of the migration from Hadoop to Databricks

## Resources and schedule

| FOUNDATION* 4 weeks | EXTENDED** 2 weeks | OPTIMIZED** 2 weeks | MIGRATION DELIVERY TEAMS |
|---|---|---|---|
| ▪ Planning<br>▪ Reference architecture<br>▪ Reference implementation | ▪ Implement second migration scenario<br>▪ Orchestration integration or BI tool | ▪ Scaled performance optimization<br>▪ CI/CD and automation<br>▪ Monitoring/alerting/security | ▪ Implement migration in bulk following references developed in earlier phases<br>▪ Price determined at engagements |

* Up to 4 resources supporting the activity over 2 sprints, each with a 2-week duration
** 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