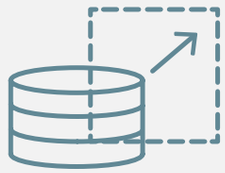
CHAPTER 5: Making Organizations Successful with ML

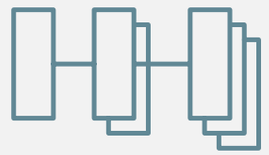
Standardizing the ML lifecycle with MLflow is a great step to ensure that data scientists can share and track experiments, compare results, reproduce runs and productionize faster.

In addition to increasing data science team productivity and collaboration and applying good engineering practices to machine learning, organizations also need to do the following:



Reliably ingest, ETL and catalog big data

Work with state-of-the-art ML frameworks and tools

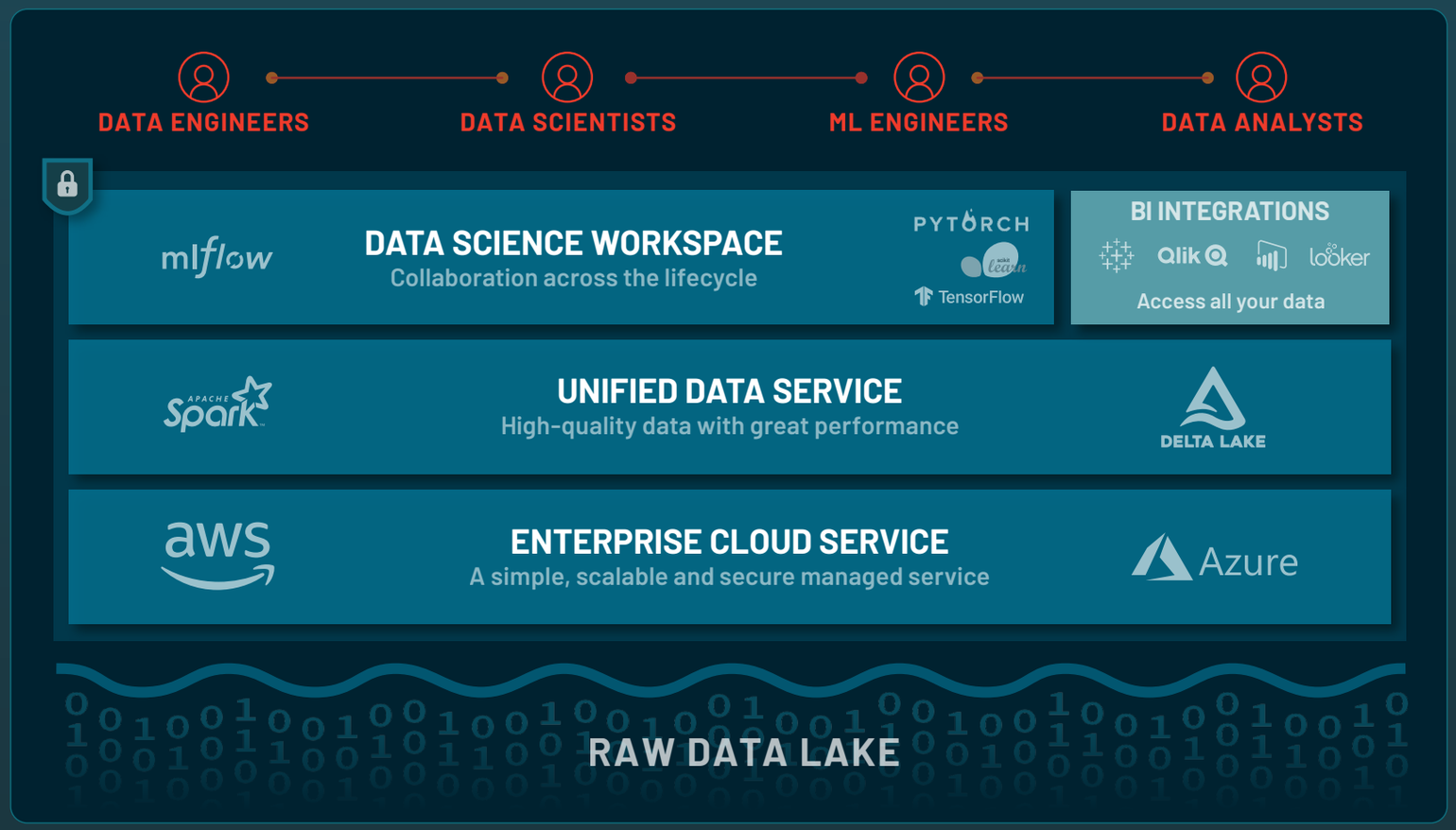


Easily scale compute from single to multi-node

Databricks excels at all the above. Learn more at databricks.com

CHAPTER 6: Introducing the Unified Data Analytics Platform

Databricks accelerates innovation by unifying data science, engineering and business. Through a fully managed, cloud-based service built by the original creators of Apache Spark, Delta Lake and MLflow, the Databricks Unified Data Analytics Platform lowers the barrier for enterprises to innovate with AI and accelerates their innovation.



Data engineering

Speed up the preparation of high-quality data, essential for best-in-class ML applications, at scale

Data science

Collaboratively explore large data sets, build models iteratively and deploy across multiple platforms

Providing managed MLflow on Databricks

MLflow is natively integrated with the Databricks Unified Data Analytics Platform so that ML practitioners and engineers can benefit from out-of-the-box tracking, packaging, deployment and management capabilities for ML models with enterprise reliability, security and scale.

By using MLflow as part of Databricks, data scientists can:

WORKSPACES

Benefit from a streamlined experiment tracking experience with Databricks Workspace and collaborative Notebooks

BIG DATA SNAPSHOTS

Track large-scale data that fed the models, along with all the other model parameters, then reproduce training runs reliably

JOBS

Easily initiate jobs remotely, from an on-premises environment or from Databricks notebooks

SECURITY

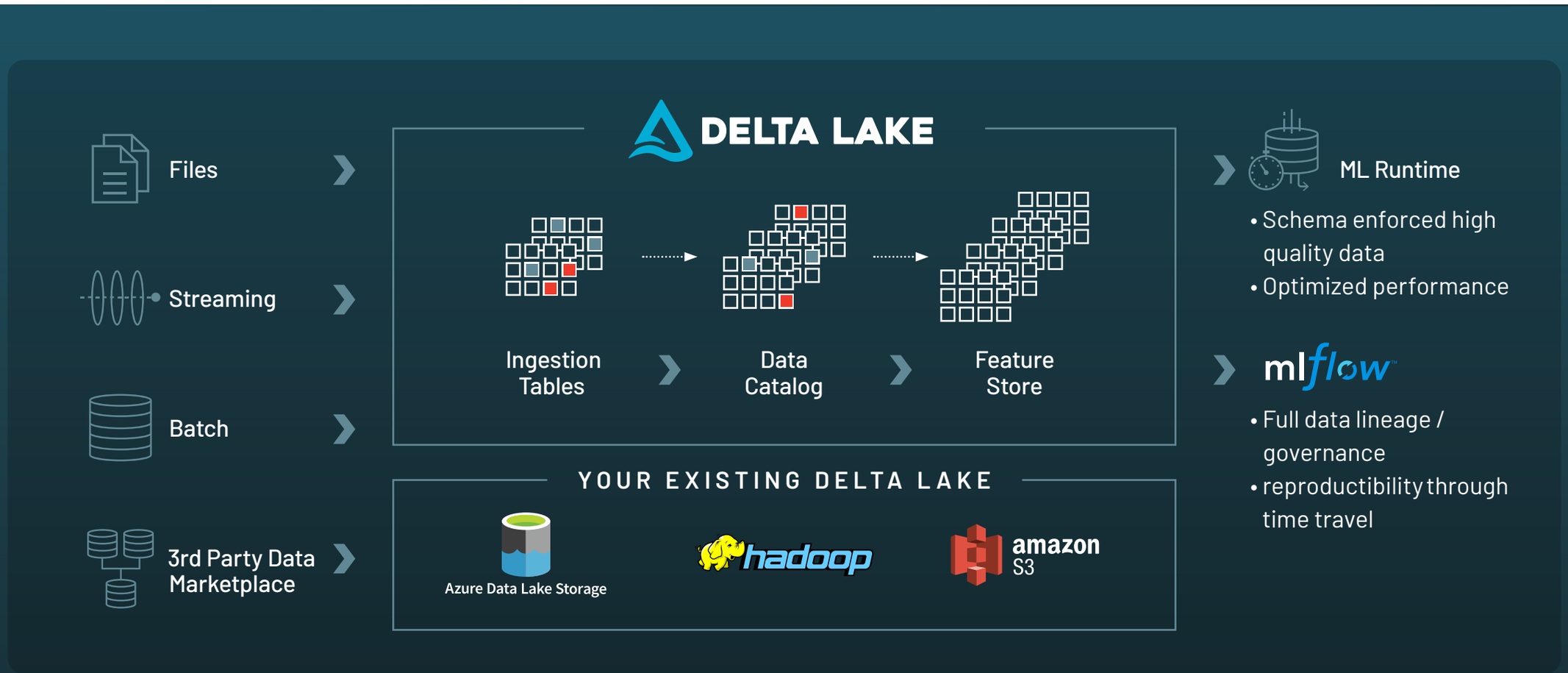
Take advantage of one common security model for the entire machine learning lifecycle

Read our blog to learn more about these integrations.

Getting data ready for ML with Delta Lake

Delta Lake is a storage layer that brings reliability to data lakes. Delta Lake provides ACID transactions and scalable metadata handling, and it unifies streaming and batch data processing. Delta Lake runs on top of your existing data lake and is fully compatible with Apache Spark APIs.

By using Delta Lake, data engineers and data scientists can keep track of data used for model training.



Ready-to-use ML environments

Databricks Runtime for Machine Learning provides data scientists and ML practitioners with on-demand access to ready-to-use machine learning clusters that are preconfigured with the latest and most popular machine learning frameworks, including TensorFlow, Keras, PyTorch, scikit-learn, XGBoost and Horovod.

By using the Databricks Runtime for ML, data scientists can get to results faster with one-click access to ML clusters, optimized performance on popular ML algorithms, and simplified distributed deep learning on Horovod and GPUs. It also supports Conda for further customization.

PACKAGES AND OPTIMIZES MOST COMMON ML FRAMEWORKS

BUILT-IN OPTIMIZATION FOR DISTRIBUTED DEEP LEARNING

Distribute and Scale any Single-Machine ML Code to thousands of machines

BUILT-IN AUTOML AND EXPERIMENT TRACKING

Auto ML and Tracking / Visualizations with MLflow

CUSTOMIZED ENVIRONMENTS USING CONDA