Implementing XGBoost Models in Databricks

Janani Ravi
Co-founder, Loonycorn

www.loonycorn.com

Overview

An overview of gradient boosting algorithms using XGBoost

Implement machine learning models using XGBoost on Databricks

An overview of machine learning using Apache Spark

Train XGBoost models using Apache Spark pipelines

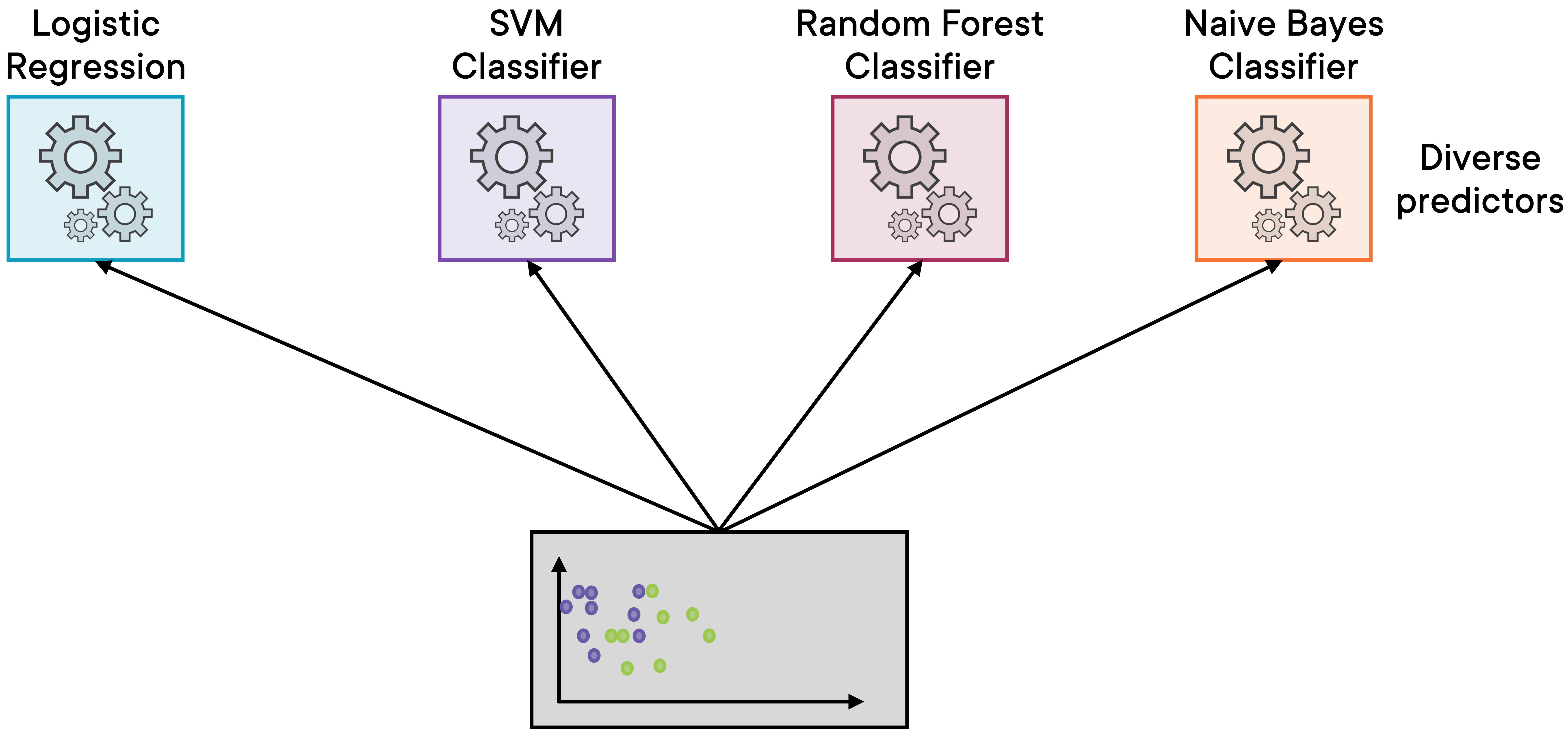
An Overview of XGBoost

XGBoost (eXtreme Gradient Boosting) - an ensemble learning technique that uses boosted tree algorithms

Ensemble Learning

Machine learning technique in which several learners are combined to obtain a better performance than any of the learners individually.

Ensemble Learning



Averaging and Boosting

Averaging

Train predictors in parallel and average scores of individual predictors

Boosting

Train predictors in sequence where each predictor learns from earlier mistakes

Averaging and Boosting

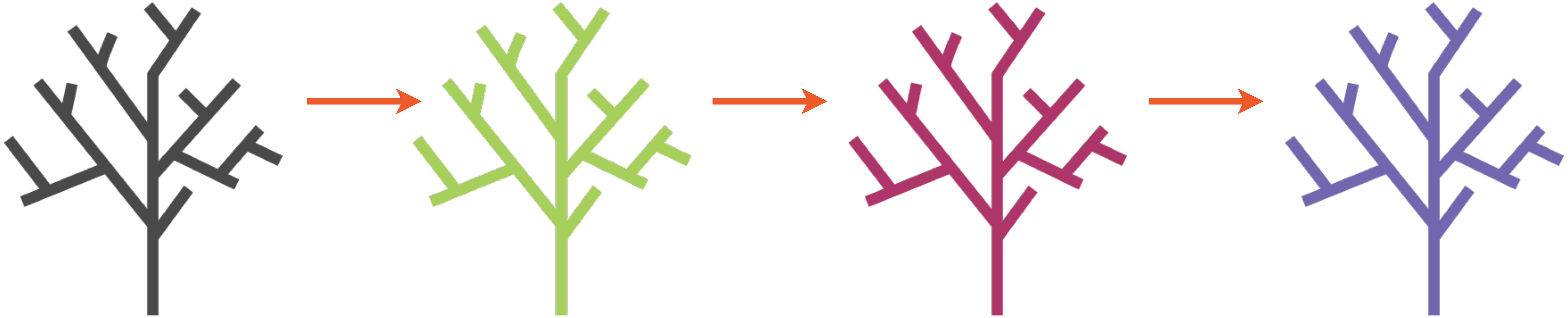
Averaging

Train predictors in parallel and average scores of individual predictors

Boosting

Train predictors in sequence where each predictor learns from earlier mistakes

Gradient Boosting



Many machine learning models come together to work on the training data (weak learners)

Gradient Boosting

Each individual weak learner may have residuals (data that it was unable to learn from)

Gradient Boosting

Model 1 fails to learn something from the underlying data (residuals)

Gradient Boosting

Model 2, the next model in the sequence, will learn from the previous model's residuals

Gradient Boosting

The sequence of models will together extract as much information as possible from the underlying data

Gradient Boosting

The combined sequence will produce a strong learner

XGBoost

Supports different gradient boosting algorithms:

- Gradient boosting
- Stochastic gradient boosting
- Regularized gradient boosting

XGBoost Features

Parallelization

Distributed computing

Out-of-core computing

Cache optimization

Demo

Building and training a classification model in Databricks using XGBoost and MLflow

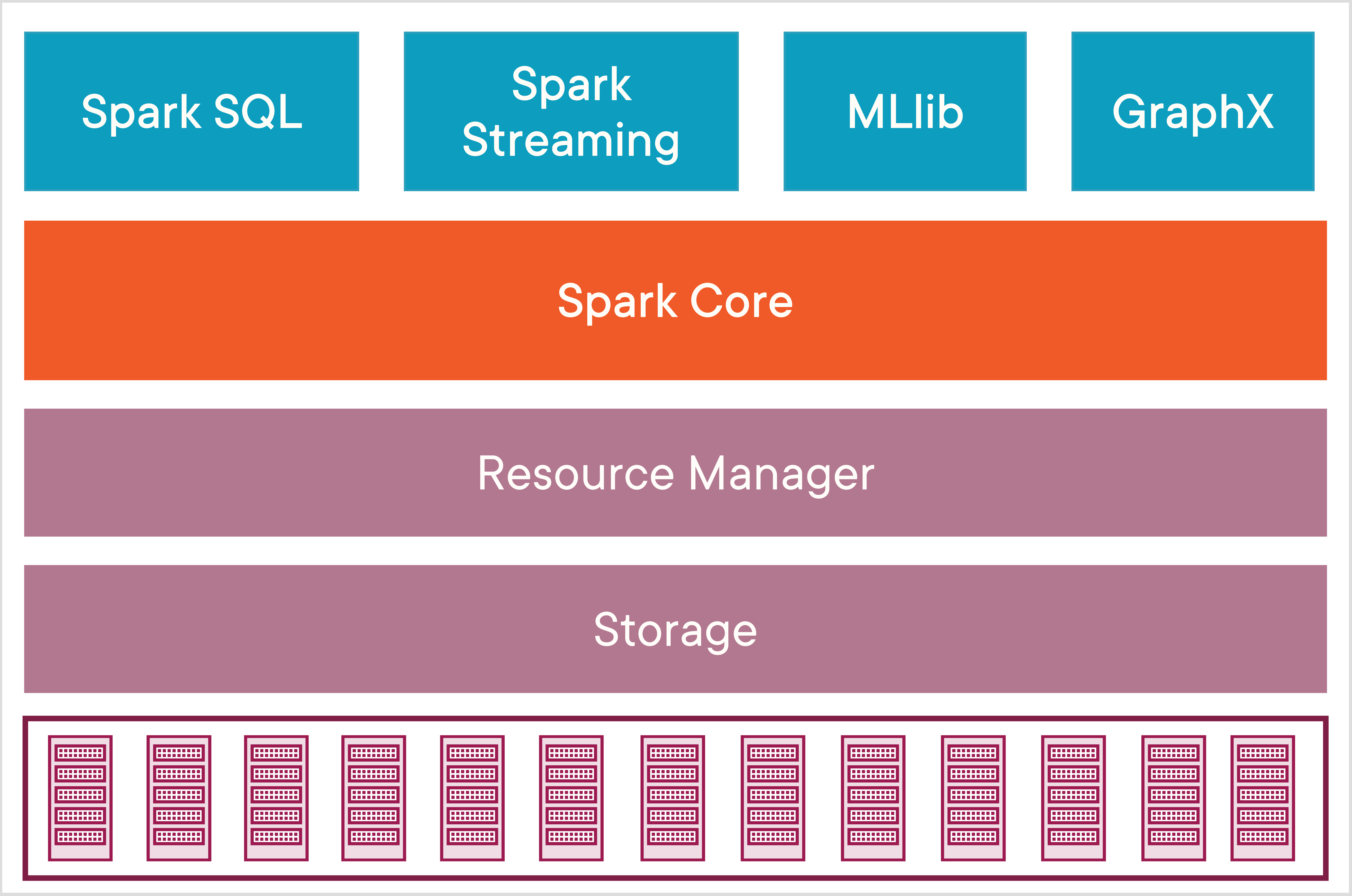
Machine Learning on Apache Spark

Apache Spark

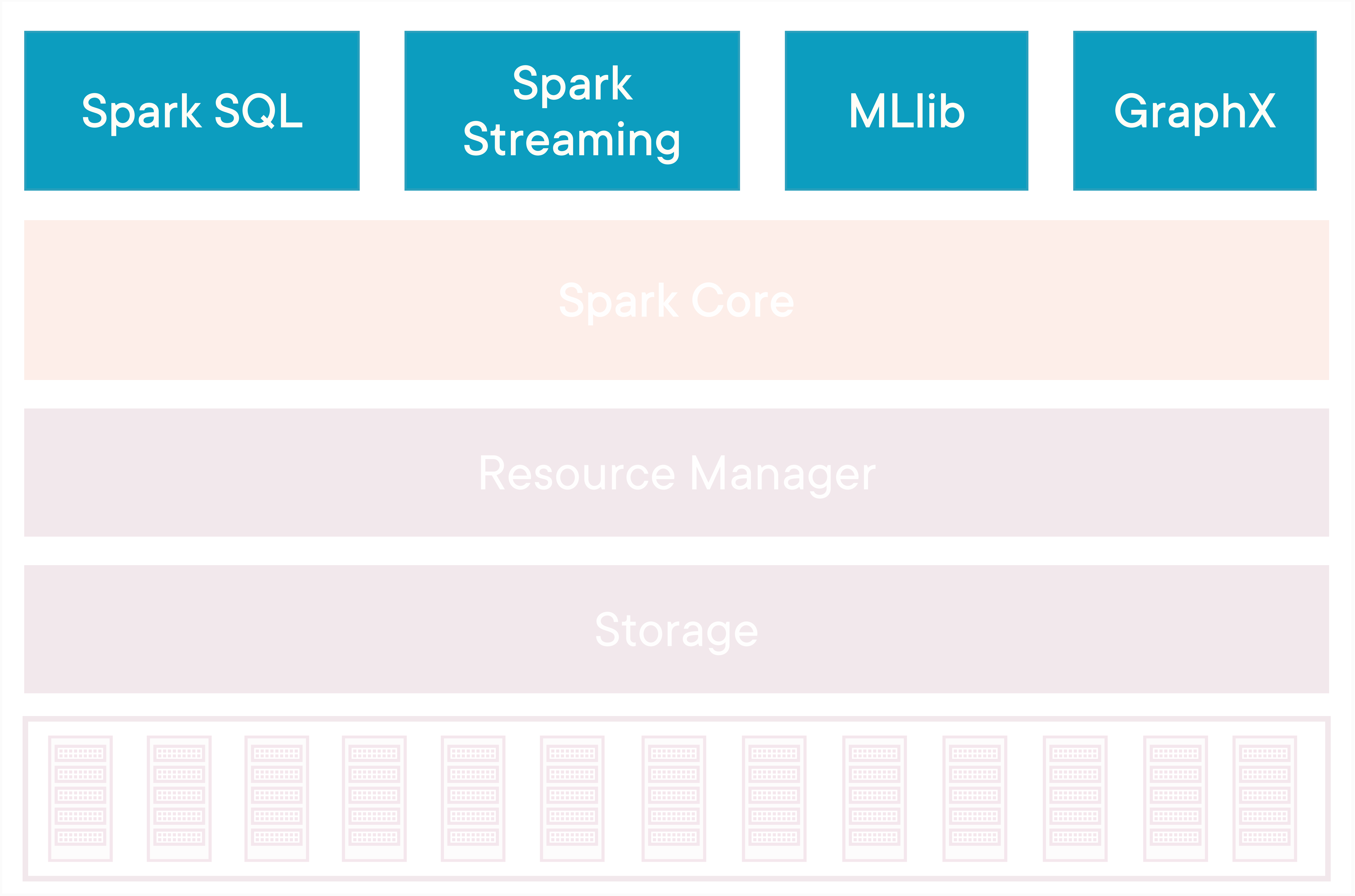
A unified analytics engine for large-scale data processing.

https://spark.apache.org/

Apache Spark

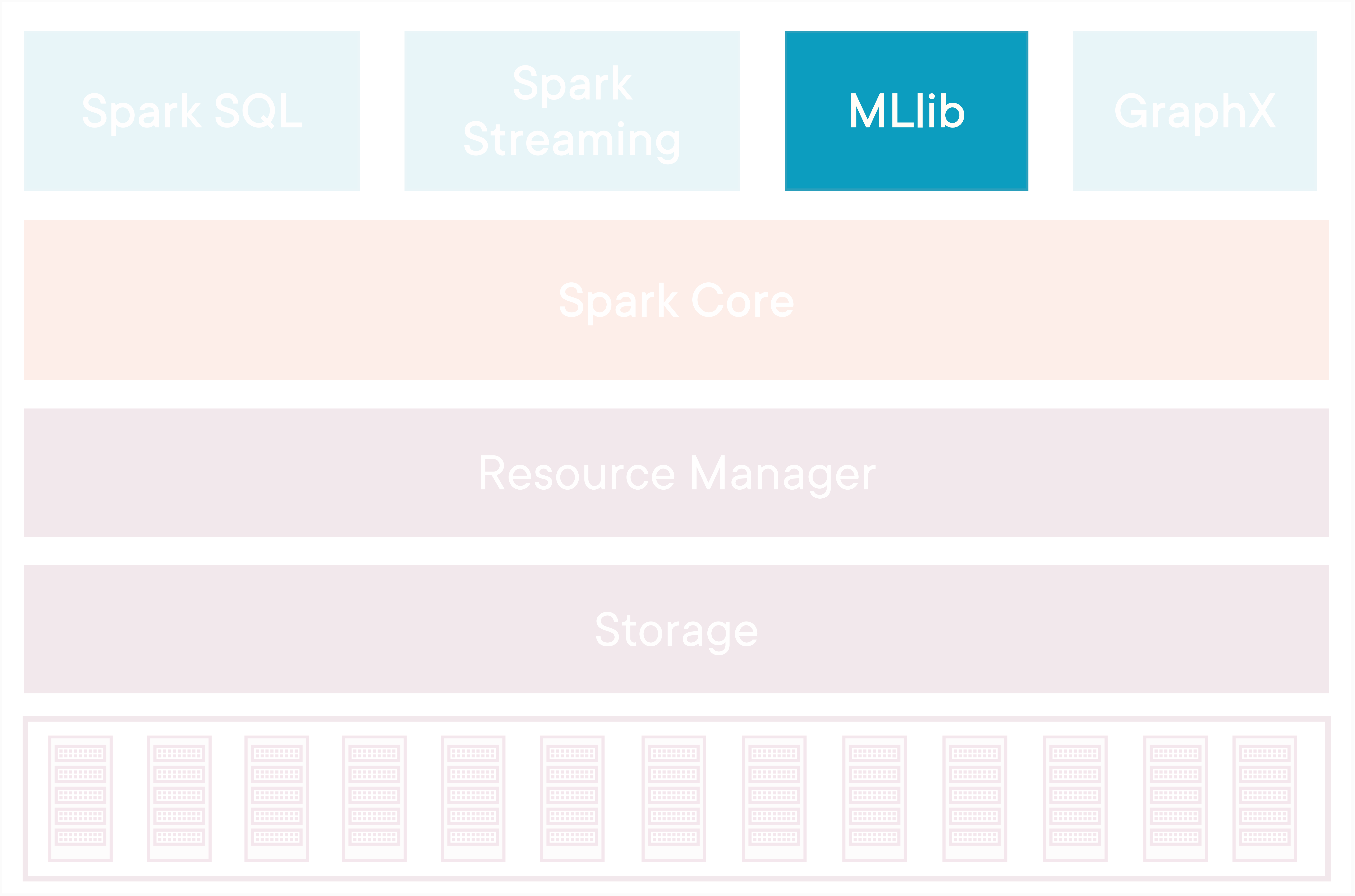


Apache Spark



Spark libraries

Machine Learning Library (MLlib)



Machine Learning Library (MLlib)

Makes practical machine learning scalable and easy.

https://spark.apache.org/docs/latest/ml-guide.html

MLlib Tools

Machine learning algorithms:

- Classification, regression, clustering, collaborative filtering

Featurization:

- Feature extraction, transformation, dimensionality reduction, selection

MLlib Tools

Pipelines:

- Constructing, evaluating, and tuning ML pipelines

Persistence:

- Save and load algorithms, models, and pipelines

Utilities:

- Linear algebra, statistics, and data handling

ML models built using MLlib take advantage of Apache Spark's distributed processing framework

XGBoost models can be trained using Spark ML pipelines

Demo

Building and training a regression model using Spark ML and XGBoost

Summary

An overview of gradient boosting algorithms using XGBoost

Implement machine learning models using XGBoost on Databricks

An overview of machine learning using Apache Spark

Train XGBoost models using Apache Spark pipelines

Up Next:
Hyperparameter Tuning for Machine Learning Models