

# Databricks - Scalable Machine Learning with Apache Spark

---

<b>Code:</b>	DB301-2
<b>Length:</b>	2 days
<b>URL:</b>	<a href="#">View Online</a>

---

- This course guides students through the process of building machine learning solutions using Spark. You will build and tune ML models with SparkML using transformers, estimators, and pipelines. This course highlights some of the key differences between SparkML and single-node libraries such as scikit-learn. Furthermore, you will reproduce your experiments and version your models using MLflow.
- You will also integrate 3rd party libraries into Spark workloads, such as XGBoost. In addition, you will leverage Spark to scale inference of single-node models and parallelize hyperparameter tuning. This course includes hands-on labs and concludes with a collaborative capstone project. All of the notebooks are available in Python, and in Scala as well where available.

## Skills Gained

- Create data processing pipelines with Spark
- Build and tune machine learning models with SparkML
- Track, version, and deploy models with MLflow
- Perform distributed hyperparameter tuning with Hyperopt
- Use Spark to scale the inference of single-node models

## Who Can Benefit

- Data scientist
- Machine learning engineer

## Prerequisites

Intermediate experience with Python Beginning experience with the PySpark DataFrame API (or have taken the Apache Spark Programming with Databricks class) Working knowledge of machine learning and data science

---