Amazon SageMaker Studio – Amazon EMR
End-to-End ML and Data Science workflows using Apache Spark on Amazon EMR with SageMaker Studio

Sumedha Swamy, Damon Cortesi
April 27, 2022
Learning Objectives
Learning Objectives

• Brief introduction to the capabilities of Amazon SageMaker Studio and Amazon EMR

• Understand benefits of a universal notebook for data analytics, data preparation and machine learning

• Understand though live product demonstrations how you can easily incorporate scalable big data workloads using EMR as a part of your ML workflows on SageMaker Studio
Amazon SageMaker Overview

**PREPARE**

- **SageMaker Ground Truth**
  Label training data for machine learning

- **SageMaker Data Wrangler**
  Aggregate and prepare data for machine learning

- **SageMaker Processing**
  Built-in Python, BYO R/Spark

- **SageMaker Feature Store**
  Store, update, retrieve, and share features

- **SageMaker Clarify**
  Detect bias and understand model predictions

**BUILD**

- **SageMaker Studio Notebooks**
  Jupyter notebooks with elastic compute and sharing

- **Built-in and Bring your-own Algorithms**
  Dozens of optimized algorithms or bring your own

- **Local Mode**
  Test and prototype on your local machine

- **SageMaker Autopilot**
  Automatically create machine learning models with full visibility

- **SageMaker JumpStart**
  Pre-built solutions for common use cases

**TRAIN & TUNE**

- **One-click Training**
  Distributed infrastructure management

- **SageMaker Experiments**
  Capture, organize, and compare every step

- **Automatic Model Tuning**
  Hyperparameter optimization

- **Distributed Training**
  Training for large datasets and models

- **SageMaker Debugger**
  Debug and profile training runs

- **Managed Spot Training**
  Reduce training cost by 90%

**DEPLOY & MANAGE**

- **Fully Managed Deployment**
  Fully managed, ultra low latency, high throughput

- **Kubernetes & Kubeflow Integration**
  Simplify Kubernetes-based machine learning

- **Multi-Model Endpoints**
  Reduce cost by hosting multiple models per instance

- **SageMaker Model Monitor**
  Maintain accuracy of deployed models

- **SageMaker Edge Manager**
  Manage and monitor models on edge devices

- **SageMaker Pipelines**
  Workflow orchestration and automation

---

SageMaker Studio

Integrated development environment (IDE) for ML

---

Not a comprehensive list. Visit [https://aws.amazon.com/sagemaker](https://aws.amazon.com/sagemaker) for the latest information

© 2022, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Amazon SageMaker Studio Notebooks

- **Quick start**: Start your notebook without spinning up compute resources.
- **Elastic**: Easily dial up or down the available resources. Changes take effect transparently in background.
- **Customizable**: Bring your own images, packages, extensions. Automate customization with Lifecycle configurations.
- **Managed**: Administrators manage access and permission to the fully managed and secure environment.
- **Collaborative**: Easily share notebooks with coworkers with a complete snapshot of your work.
- **Integrated**: Run end to end data prep and ML workflows in purpose built, performance optimized runtimes.
Amazon EMR
Amazon EMR

Easily Run Spark, Hive, Presto, HBase, Flink, and more big data apps on AWS

**Latest versions**
- Updated with latest open source frameworks within 30 days
- Support for popular OSS like Flink, Hudi

**Best Performance at Lowest cost**
- Spark workloads run 2.4x faster compared to Open Source
- 50–80% reduction in costs with EC2 Spot and Reserved Instances
- Per-second billing for flexibility

**Use S3 storage**
- Process data in S3 securely with high performance using the EMRFS connector
- Scale Compute and Storage independent of each other

**Easy & Scalable**
- Fully managed, no cluster setup, node provisioning or cluster tuning
- Vertical and Horizontal Auto-Scaling to suit workload demands
Universal Notebook
Universal notebook for data prep and ML: Drivers

- Data preparation and analytics are foundational components of ML workflows
- Switching between multiple notebooks, tools, and interfaces reduces productivity
- Security and access control need to be consistent across analytics and ML services
Product Demo
Product Demos
SageMaker Studio: Universal notebook for interactive analytics and ML

Demo: Create and Manage Amazon EMR clusters from SageMaker Studio
Demo: Discover and connect to EMR clusters in a cross-account scenario from SageMaker Studio
Demo: 1-click monitoring and debugging Spark jobs from SageMaker Studio
Demo: End to end workflow to prepare data, train and deploy SageMaker models from SageMaker Studio
Resources

• **AWS Blog**: Create and manage Amazon EMR Clusters from SageMaker Studio to run interactive Spark and ML workloads

• **AWS Documentation**: Prepare Data at Scale with Studio Notebooks

• **IAM Permissions** to enable the functionality we saw today