Scalable data preparation using Amazon SageMaker Studio notebooks

Sumedha Swamy, Sean Morgan

July 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 that you can try yourself how you can easily incorporate scalable big data workloads using EMR as a part of your ML workflows on SageMaker Studio
Amazon 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 co-workers with a complete snapshot of your work.

**Integrated**
Run end to end data prep and ML workflows in purpose built, performance optimized runtimes.

© 2022, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
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 up to 3x 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

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