Preparing data at scale with Amazon SageMaker Studio Notebooks

Sumedha Swamy

Principal Product Manager
Amazon Web Services

Agenda

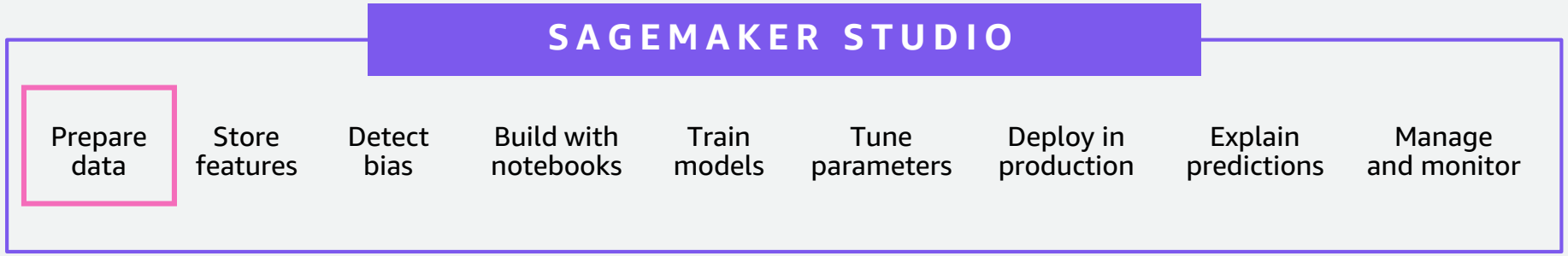
Introduction to SageMaker Studio

Introduction to Amazon EMR and AWS Glue Interactive Sessions

Demo

Amazon SageMaker Studio

FULLY INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) FOR MACHINE LEARNING



Amazon EMR

EASILY RUN SPARK, HIVE, PRESTO, HBASE, FLINK, AND MORE BIG DATA APPLICATIONS ON AWS

Latest versions

Updated with latest open source frameworks **within 30 days**

Support for popular OSS like **Flink, Hudi**

Great performance at a low cost

Spark workloads run up to **3x faster** compared to other open-source tools

Up to **50–80% reduction** in costs with EC2 Spot and reserved instances **per-second billing** for flexibility

Use Amazon S3 storage

Process data in Amazon S3 **securely** with **high performance** using the EMR File System (EMFRS) connector

Scale compute and storage independent of each other

Easy and scalable

Fully managed, no cluster setup, node provisioning, or cluster tuning

Vertical and horizontal auto scaling to suit workload demands

AWS Glue interactive sessions with Studio

NEXT-GENERATION INTERACTIVE DATA EXPLORATION AND JOB DEVELOPMENT

Use the **built-in AWS Glue kernels** in SageMaker Studio

Time to first Spark query: ~**30 seconds**

On-demand, highly scalable, and shuts off when idle

No up-front configuration needed

Dedicated resources for no noisy neighbors

Thank you!

Sumedha Swamy