

**Introduction to Data Science on the AWS Platform**

**Course Number:** DATA-106  
**Duration:** 5 days

**Overview**

Scale your data science workloads on Amazon Web Services to take advantage of on-demand delivery of compute power, database services, storage, applications, and IT resources, as well as tools that are unique to the AWS platform.

This in-person or online Data Science on the AWS Platform training course teaches engineers, data scientists, statisticians, and other quantitative professionals how to use AWS (Amazon Web Services) with Jupyter notebooks for data science to create scalable data analytics solutions.

Did you miss our live webinar? You can still view the [AWS and Data Science with Python webinar](https://youtu.be/4QzEIDuM8MM) recording.

**Prerequisites**

All students must have experience with data science or statistical programming (any language).

**Materials**

All AWS for Data Science training students will receive comprehensive courseware.

**Software Needed on Each Student PC**

A modern web browser and an Internet connection.

**Objectives**

* Use AWS SageMaker (a managed Jupyter notebook service from AWS)
  + Use the interface to run different notebook kernels and virtual machines in SageMaker
  + Explore AWS sample notebooks and new use cases of data science on the cloud
  + Use the GitHub integration and Git via the graphical JupyterLab interface
  + Write notebooks and use the SageMaker Papermill integration to schedule and parallelize running notebooks as parameterized compute jobs
* Use Open Datasets on AWS
  + Gain experience working with large datasets in the cloud (GB and TB scale)
  + Use the AWS CLI to explore collections of files and buckets within Amazon S3
  + Copy, sync, and move data to and from SageMaker for analysis
  + Implement and build upon steps described in tutorial notebooks from the Registry of Open Data
  + Write a tutorial notebook explaining a use case you are interested in
* Explore and test AWS Machine learning APIs
  + Explore using Amazon Rekognition,the state of the art in computer vision
  + Explore using Amazon Comprehend to obtain valuable insights from text within documents
  + Test and analyze the behavior of these machine learning services on your own data using AWS SageMaker
  + Write an analysis notebook
  + Explain unique insights into the performance of the ML services and demonstrate by testing on data

**Outline**

* Introduction
  + Notebook Computing
  + Project Jupyter
  + Data science environments
  + Managed notebook services
  + Amazon SageMaker Studio
* Cloud Concepts
  + Definition of a web service
  + Cloud providers
  + Six advantages of cloud computing
  + Different types of cloud computing models (e.g. IAAS, PAAS, SAAS)
  + 5 Principles of cloud computing
  + A new computing paradigm
* JupyterLab Interface
  + Jupyter notebook format
  + JupyterLab notebook model
  + Kernels
  + Instances
  + GitHub integration
  + Cloning repositories
* AWS Cloud Security and Billing
  + Shared responsibility model
  + AWS IAM
  + IAM users, groups, policies, and roles
  + AWS pricing model
  + Securing a new AWS account
  + AWS Console
  + AWS Billing and Cost Explorer
  + Setup Amazon CloudWatch Billing Alarms
  + AWS Cloud Shell
* Cloud Prerequisites
  + Common Linux distributions on AWS
  + YUM and APT
  + Basic commands such as ls, cp and chmod
  + JSON
  + RESTful APIs
* AWS Services
  + Main AWS service categories and core services
  + Regional and Zonal services
  + Services with no charge
  + AWS APIs
  + AWS CLI
  + AWS Python SDK
* Amazon Simple Storage Service (S3)
  + Block storage versus object storage
  + S3 overview
  + S3 storage classes
  + IAM policies
  + Bucket URLs (two styles)
  + Three common use cases
  + S3 pricing
  + AWS CLI commands for S3
  + Python boto3 for S3
  + Registry of Open Data on AWS
* AWS Machine Learning APIs
  + Amazon Rekognition (computer vision service)
  + Amazon Comprehend (NLP service)
  + Amazon Translate
  + Amazon Transcribe (speech-to-text service)
  + Amazon Polly (text-to-speech service)
* Amazon Elastic Compute Service (EC2)
  + Example use cases
  + EC2 overview
  + Amazon Machine Image
  + Instance types
  + User data scripts
  + Storage options
  + Tagging
  + Security group settings
  + EC2 pricing
  + Four pillars of cost optimization
* Amazon Elastic Container Registry (ECR)
  + Container basics
  + What is Docker
  + JupyterLab on EC2 via Docker
  + Amazon ECR overview
  + SageMaker Docker images for deep learning
* AWS Lambda
  + Serverless AWS services
  + Benefits of Lambda
  + Event sources
  + Lambda function configuration
  + AWS Lambda limits
  + Use Lambda to execute and schedule notebooks
* Conclusion