Course Outline

AWS Academy Data Engineering

CONFIDENTIAL – DO NOT DISTRIBUTE

Course version

This course outline applies to version 1.0 of AWS Academy Data Engineering in English.

Description

AWS Academy Data Engineering is designed to help students learn about and get hands-on practice with the tasks, tools, and strategies that are used to collect, store, prepare, analyze, and visualize data for use in analytics and machine learning (ML) applications. Throughout the course, students will explore use cases from real-world applications, which will enable them to make informed decisions while building data pipelines for their particular applications.

Curriculum objectives

This course prepares students to do the following:

- Summarize the role and value of data science in a data-driven organization.
- Recognize how the elements of data influence decisions about the infrastructure of a data pipeline.
- Illustrate a data pipeline by using AWS services to meet a generalized use case.
- Identify the risks and approaches to secure and govern data at each step and each transition of the data pipeline.
- Identify scaling considerations and best practices for building pipelines that handle large-scale datasets.
- Design and build a data collection process while considering constraints such as scalability, cost, fault tolerance, and latency.
- Select a data storage option that matches the requirements and constraints of a given data analytics use case.
- Implement the steps to process structured, semistructured, and unstructured data formats in a data pipeline that is built with AWS.
- Explain the concept of MapReduce and how Amazon EMR is used in big data pipelines.
- Differentiate the characteristics of an ML pipeline and its specific processing steps.
- Analyze data by using AWS tools that are appropriate to a given use case.
- Implement a data visualization solution that is aligned to an audience and data type.

Duration

The course duration is approximately 40 hours when delivered synchronously by an educator. The course is designed to be delivered over one semester. Actual delivery times will vary from class to class and depending on delivery format. This course must be delivered over a period of at least 8 weeks.

Intended audience

This intermediate (level 200) course is intended for students at AWS Academy member institutions who seek expertise on the tasks, tools, and strategies that are used to collect, store, prepare, analyze, and visualize data for use in analytics and ML applications. This course is most aligned to a data engineer role but would also be appropriate for data analysts; data scientists; extract, transform, and

load (ETL) developers; or ML practitioners who want to understand how the data that they use in their analyses and predictions is prepared for analysis.

Student prerequisites

This course requires a strong foundation in IT concepts and skills. To ensure success in this course, students should have the following:

- Completed the AWS Academy Cloud Foundations course or have equivalent experience
- Worked with Structured Query Language (SQL)
- Worked with databases
- Introduced to general networking concepts
- Understanding of decision-making knowledge in math, probability, and statistics

Delivery methods

Learning materials are provided to support synchronous or asynchronous learning. Lecture slides and an instructor guide are provided for instructor-led training. Recorded lectures and demos are provided for independent learning. The educator can determine the preferred delivery method for each module.

Educator prerequisites

This course does not have any prerequisites for educators. However, prior to facilitating this course, educators are recommended to complete this course, complete the AWS Academy Cloud Foundations course, and pass the AWS Certified Cloud Practitioner exam.

Learning resources

- Lecture slides
- Student guide
- Instructor guide
- Practical activities
- Lab exercises
- Instructor lab sandbox environment
- Recorded lectures
- Recorded demos
- Module knowledge checks
- Course assessment
- Capstone project

Course timing

This table provides the suggested durations for all course activities. Note that the total classroom time for all the modules in this course is 2,400 minutes (40 hours). Items that are not applicable are marked NA.

Title	Lecture (Minutes)	Activity/Lab /Demo (Minutes)	Knowledge Check (Minutes)	Total Classroom Timing (Minutes)	Digital Lecture (Minutes)
Module 1: Welcome to AWS Academy Data Engineering	30	NA	NA	30	NA
Module 2: Data-Driven Organizations	75	70	10	155	25
Module 3: The Elements of Data	75	30	10	115	30
Module 4: Design Principles and Patterns for Data Pipelines	85	130	10	225	33
Module 5: Securing and Scaling the Data Pipeline	90	NA	10	100	52
Module 6: Ingesting and Preparing Data	90	NA	10	100	40
Module 7: Ingesting by Batch or by Stream	115	100	10	225	54
Module 8: Storing and Organizing Data	85	100	10	195	32
Module 9: Processing Big Data	105	200	10	315	44
Module 10: Processing Data for ML	140	65	10	215	53
Module 11: Analyzing and Visualizing Data	75	120	10	205	23
Module 12: Automating the Pipeline	50	130	10	190	18

Title	Lecture (Minutes)	Activity/Lab /Demo (Minutes)	Knowledge Check (Minutes)	Total Classroom Timing (Minutes)	Digital Lecture (Minutes)
Module 13: Bridging to Certification	30	NA	NA	30	NA
Capstone Project	NA	240	NA	240	NA
Course Assesment	NA	NA	60	60	NA
Total Course Timing	**1,045**	**240**	**170**	**2,400**	**404**

Module sections

This section lists the module sections in this course.

Module 1: Welcome to AWS Academy Data Engineering

- Course prerequisites and objectives
- Course overview

Module 2: Data-Driven Organizations

- Data-driven decisions
- The data pipeline – infrastructure for data-driven decisions
- The role of the data engineer in data-driven organizations
- Modern data strategies
- Lab: Accessing and Analyzing Data by Using Amazon S3
- Knowledge check

Module 3: The Elements of Data

- The five Vs of data – volume, velocity, variety, veracity, and value
- Volume and velocity
- Variety – data types
- Variety – data sources
- Veracity and value
- Activities to improve veracity and value
- Activity: Planning Your Pipeline
- Knowledge check

Module 4: Design Principles and Patterns for Data Pipelines

- AWS Well-Architected Framework and Lenses
- Activity: Using the Well-Architected Framework
- The evolution of data architectures
- Modern data architecture on AWS
- Modern data architecture pipeline: Ingestion and storage
- Modern data architecture pipeline: Processing and consumption
- Streaming analytics pipeline
- Lab: Querying Data by Using Athena
- Knowledge check

Module 5: Securing and Scaling the Data Pipeline

- Cloud security review
- Security of analytics workloads
- ML security
- Scaling: An overview
- Creating a scalable infrastructure
- Creating scalable components
- Knowledge check

Module 6: Ingesting and Preparing Data

- ETL and ELT comparison
- Data wrangling introduction
- Data discovery
- Data structuring
- Data cleaning
- Data enriching
- Data validating
- Data publishing
- Knowledge check

Module 7: Ingesting by Batch or by Stream

- Comparing batch and stream ingestion
- Batch ingestion processing
- Purpose-built ingestion tools
- AWS Glue for batch ingestion processing
- Scaling considerations for batch processing
- Lab: Performing ETL on a Dataset by Using AWS Glue
- Kinesis for stream processing
- Scaling considerations for stream processing
- Ingesting IoT data by stream
- Knowledge check

Module 8: Storing and Organizing Data

- Storage in the modern data architecture
- Data lake storage
- Data warehouse storage
- Purpose-built databases
- Storage in support of the pipeline
- Securing storage
- Lab: Storing and Analyzing Data by Using Amazon Redshift
- Knowledge check

Module 9: Processing Big Data

- Big data processing concepts
- Apache Hadoop
- Apache Spark
- Amazon EMR
- Managing your Amazon EMR clusters
- Lab: Processing Logs by Using Amazon EMR
- Apache Hudi

- Lab: Updating Dynamic Data in Place
- Knowledge check

Module 10: Processing Data for ML

- ML concepts
- The ML lifecycle
- Framing the ML problem to meet the business goal
- Collecting data
- Applying labels to training data with known targets
- Activity: Labeling with SageMaker Ground Truth
- Preprocessing data
- Feature engineering
- Developing a model
- Deploying a model
- ML infrastructure on AWS
- SageMaker
- Demo: Preparing Data and Training a Model with SageMaker
- Demo: Preparing Data and Training a Model with SageMaker Canvas
- AI/ML services on AWS
- Knowledge check

Module 11: Analyzing and Visualizing Data

- Considering factors that influence tool selection
- Comparing AWS tools and services
- Demo: Analyzing and Visualizing Data with AWS IoT Analytics and QuickSight
- Selecting tools for a gaming analytics use case
- Lab: Analyzing and Visualizing Streaming Data with Kinesis Data Firehose, OpenSearch Service, and OpenSearch Dashboards
- Knowledge check

Module 12: Automating the Pipeline

- Automating infrastructure deployment
- CI/CD
- Automating with Step Functions
- Lab: Building and Orchestrating ETL Pipelines by Using Athena and Step Functions
- Knowledge check

Module 13: Bridging to Certification

- AWS Certification overview

IoT Use Case (Optional)

This supplemental PowerPoint deck presents a sample use case for building an Internet of Things (IoT) data pipeline. The PowerPoint file includes sections for each of the main pipeline layers (ingestion and processing, storage, and analysis and visualization).

Capstone Project

The Capstone Project provides an integrative project-based learning experience that reinforces technical skills that are taught in this course. The capstone offers students an opportunity to demonstrate critical thinking, problem solving, the software development lifecycle, and communication skills.