

## Big Data on AWS

### Overview

---

Big Data on AWS introduces you to cloud-based big data solutions such as Amazon Elastic MapReduce (EMR), Amazon Redshift, Amazon Kinesis and the rest of the AWS big data platform. In this course, we show you how to use Amazon EMR to process data using the broad ecosystem of Hadoop tools like Hive and Hue. We also teach you how to create big data environments, work with Amazon DynamoDB, Amazon Redshift, Amazon QuickSight, Amazon Athena and Amazon Kinesis, and leverage best practices to design big data environments for security and cost-effectiveness.

### Prerequisite Comments

---

We recommend that attendees of this course have the following prerequisites:

Basic familiarity with big data technologies, including Apache Hadoop, HDFS, and SQL/NoSQL querying.

Students should complete the Big Data Technology Fundamentals web-based training or have equivalent experience.

Working knowledge of core AWS services and public cloud implementation.

Students should complete the AWS Essentials course or have equivalent experience.

Basic understanding of data warehousing, relational database systems, and database design.

### Target Audience

---

This course is intended for:

Individuals responsible for designing and implementing big data solutions, namely Solutions Architects and SysOps Administrators.

Data Scientists and Data Analysts interested in learning about big data solutions on AWS.

### Course Objectives

---

This course teaches you how to:

Fit AWS solutions inside of a big data ecosystem

Leverage Apache Hadoop in the context of Amazon EMR

Identify the components of an Amazon EMR cluster

Launch and configure an Amazon EMR cluster

Leverage common programming frameworks available for Amazon EMR including Hive, Pig, and Streaming

Leverage Hue to improve the ease-of-use of Amazon EMR

Use in-memory analytics with Spark on Amazon EMR

Choose appropriate AWS data storage options

Identify the benefits of using Amazon Kinesis for near real-time big data processing

Leverage Amazon Redshift to efficiently store and analyze data

Comprehend and manage costs and security for a big data solution

Identify options for ingesting, transferring, and compressing data

### Course Outline

---

### 1 - Day 1

Overview of Big Data, Apache Hadoop, and the Benefits of Amazon EMR  
Amazon EMR Architecture  
Using Amazon EMR  
Launching and Using an Amazon EMR Cluster  
Hadoop Programming Frameworks

### 2 - Day 2

Using Hive for Advertising Analytics  
Using Streaming for Life Sciences Analytics  
Overview: Spark and Shark for In-Memory Analytics  
Using Spark and Shark for In-Memory Analytics  
Managing Amazon EMR Costs  
Overview of Amazon EMR Security  
Data Ingestion, Transfer, and Compression  
Using Amazon Kinesis for Real-Time Big Data Processing

### 3 - Day 3

Using Amazon Kinesis for Real-Time Big Data Processing  
AWS Data Storage Options  
Using DynamoDB with Amazon EMR  
Overview: Amazon Redshift and Big Data  
Using Amazon Redshift for Big Data  
Visualizing and Orchestrating Big Data  
Using Tableau Desktop or Jaspersoft BI to Visualize Big Data

---