

## Course: IL - Big Data on AWS

### Overview

In this course, you will learn about cloud-based Big Data solutions such as Amazon EMR, Amazon Redshift, Amazon Kinesis, and the rest of the AWS Big Data platform. We will show you how to use Amazon EMR to process data using the broad ecosystem of Hadoop tools like Hive and Hue. We will 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. This course will prepare for: AWS Certified Big Data - Specialty

### Course Details

- Duration: 3 Days
- Level 300

### Who this course is designed for:

- Solutions architects
- SysOps administrators
- Data scientists
- Data analysts

### Course Objectives

- Fit AWS solutions inside a big data ecosystem
- Leverage Apache Hadoop in the context of Amazon EMR
- Identify the components of an Amazon EMR cluster, then launch and configure an Amazon EMR cluster
- Use common programming frameworks available for Amazon EMR, including Hive, Pig, and streaming
- Improve the ease of use of Amazon EMR by using Hadoop User Experience (Hue)
- Use in-memory analytics with Apache 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
- Leverage Amazon Athena for ad-hoc query analytics
- Use AWS Glue to automate extract, transform, and load (ETL) workloads
- Use visualization software to depict data and queries using Amazon QuickSight

### Pre-Requisites

- Familiarity with big data technologies, including Apache Hadoop, Hadoop Distributed File System (HDFS), and SQL/NoSQL querying
- Working knowledge of core AWS services and public cloud implementation
- AWS Technical Essentials or equivalent experience
- Understanding of data warehousing, relational database systems, and database design

## Course Outline

### Course Outline

Overview of Big Data

Data Ingestion, Transfer, and Compression

AWS Data Storage Options

Using DynamoDB with Amazon EMR

Using Kinesis for Near Real-Time Big Data Processing

Introduction to Apache Hadoop and Amazon EMR

Using Amazon Elastic MapReduce

The Hadoop Ecosystem

Using Hive for Advertising Analytics

Using Streaming for Life Sciences Analytics

Using Hue with Amazon EMR

Running Pig Scripts with Hue on Amazon EMR

Spark on Amazon EMR

Running Spark and Spark SQL Interactively on Amazon EMR

Using Spark and Spark SQL for In-Memory Analytics

Managing Amazon EMR Costs

Securing your Amazon EMR Deployments

Data Warehouses and Columnar Datastores

Introduction to Amazon Redshift

Optimizing Your Amazon Redshift Environment

The Big Data Ecosystem on AWS

Visualizing and Orchestrating Big Data

Using Tibco Spotfire to Visualize Big Data