

Big Data on AWS (AWS-BIG-DATA)

Course Summary

Description

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.

Objectives

After taking this course, students will learn to:

- 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

Topics

- | | |
|----------------------------------------------------------------------------|------------------------------------------------------------|
| • Overview of Big Data | • Web Interfaces on Amazon EMR |
| • Big Data Ingestion and Transfer | • Lab 5: Running Pig Scripts in Hue on Amazon EMR |
| • Big Data Streaming and Amazon Kinesis | • Apache Spark on Amazon EMR |
| • Lab 1: Using Amazon Kinesis to Stream and Analyze Apache Server Log Data | • Lab 6: Processing NY Taxi data using Spark on Amazon EMR |
| • Big Data Storage Solutions | • Using AWS Glue to automate ETL workloads |
| • Big Data Processing and Analytics | • Amazon Redshift and Big Data |
| • Lab 2: Using Amazon Athena to Query Log Data From Amazon S3 | • Visualizing and Orchestrating Big Data |
| • Apache Hadoop and Amazon EMR | • Lab 7: Using TIBCO Spotfire to Visualize Data |
| • Lab 3: Storing and Querying Data on Amazon DynamoDB | • Managing Big Data Costs |
| • Using Amazon EMR | • Securing Your Amazon Deployments |
| • Hadoop Programming Frameworks | • Big Data Design Patterns |
| • Lab 4: Processing Server Logs With Hive on Amazon EMR | |

Big Data on AWS (AWS-BIG-DATA)

Course Summary (cont'd)

Audience

This course is intended for:

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

Prerequisites

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

- Basic familiarity with big data technologies, including Apache Hadoop, MapReduce, HDFS, and SQL/NoSQL querying
- Students should complete the free 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 Technical Essentials course or have equivalent experience
- Basic understanding of data warehousing, relational database systems, and database design

Duration

Three Days