Maslak Mahallesi, Maslak Meydan Sk. No:5. Spring Giz Plaza. Maslak/İstanbul
+90 212 282 7700 - info@bilginc.com

Big Data on AWS

Learn via: **Classroom/Virtual**

Duration: **3 Days**

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, and Amazon Kinesis, and leverage best practices to design big data environments for security and cost-effectiveness.

Target Audience

- 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

Delivery Method

This course is delivered through a mix of:

- Instructor-Led Training (ILT)
- Hands-On Labs

Hands-On Activity

This course allows you to test new skills and apply knowledge to your working environment through a variety of practical exercises.

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 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

What You Will Learn

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
- Leverage Amazon Athena for ad-hoc query analytics
- Leverage AWS Glue to automate ETL workloads.
- Use visualization software to depict data and queries using Amazon QuickSight

- Orchestrate big data workflows using AWS Data Pipeline

Outline

Day 1

- Overview of Big Data
- Ingestion
- Big Data streaming and Amazon Kinesis
- Using Kinesis to stream and analyze Apache server logs
- Storage Solutions
- Querying Big Data using Amazon Athena
- Using Amazon Athena to analyze log data
- Introduction to Apache Hadoop and Amazon EMR

Day 2

- Using Amazon Elastic MapReduce
- Storing and Querying Data on DynamoDB
- Hadoop Programming Frameworks
- Processing Server Logs with Hive on Amazon EMR
- Streamlining Your Amazon EMR Experience with Hue
- Running Pig Scripts in Hue on Amazon EMR
- Spark on Amazon EMR
- Processing New York Taxi dataset using Spark on Amazon EMR

Day 3

- Using AWS Glue to automate ETL workloads
- Amazon Redshift and Big Data
- Visualizing and Orchestrating Big Data
- Visualizing
- Managing Amazon EMR Costs
- Securing Big Data solutions
- Big Data Design Patterns