Iverson Associates Sdn Bhd (303330-M)
Suite T113 – T114, 3rd Floor, Centrepoint, Lebuh Bandar Utama
Bandar Utama, 47800 Petaling Jaya, Selangor Darul Ehsan
Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Module Title : Big Data on AWS
Duration : 3 days

Course Description

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.

Course Objectives

This course is designed to teach 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 and Spark SQL on Amazon EMR
- Choose appropriate AWS data storage options
- Identify the benefits of using Amazon Kinesis for near real-time big data processing
- Define data warehousing and columnar database concepts
- Leverage Amazon Redshift to efficiently store and analyze data
- Comprehend and manage costs and security for Amazon EMR and Amazon Redshift deployments
- Identify options for ingesting, transferring, and compressing data
- Use visualization software to depict data and queries
- Orchestrate big data workflows using AWS Data Pipeline

Intended 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

Prerequisites

We recommend that attendees of this course have:

- Basic familiarity with big data technologies, including Apache Hadoop and HDFS
- Knowledge of big data technologies such as Pig, Hive, and MapReduce is helpful but not required
- 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

Hands-On Activity

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

Course Outline

This course will cover the following concepts on each day:

Day 1

- Overview of Big Data
- Ingestion, Transfer, and Compression
- Storage Solutions
- Storing and Querying Data on DynamoDB
- Big Data Processing and Amazon Kinesis
- Introduction to Apache Hadoop and Amazon EMR
- Using Amazon Elastic MapReduce

Day 2

- Hadoop Programming Frameworks
- Processing Server Logs with Hive on Amazon EMR
- Processing Chemistry Data Using Hadoop Streaming on Amazon EMR
- Streamlining Your Amazon EMR Experience with Hue
- Running Pig Scripts in Hue on Amazon EMR
- Spark on Amazon EMR
- Interactively Creating and Querying Tables with Spark and Spark SQL on Amazon EMR
- Managing Amazon EMR Costs
- Securing your Amazon EMR Deployments

Day 3

- Data Warehouses and Columnar Datastores

- Amazon Redshift and Big Data
- Optimizing Your Amazon Redshift Environment
- Big Data Design Patterns
- Visualizing and Orchestrating Big Data
- Using Tibco Spotfire to Visualize Big Data