Big Data Processing with Apache Spark

Duration: 2 Days Course Code: LO035435

Overview:

Processing big data in real-time is challenging due to scalability, information consistency, and fault tolerance. This course shows you how you can use Spark to make your overall analysis workflow faster and more efficient. You'll learn all about the core concepts and tools within the Spark ecosystem, like Spark Streaming, the Spark Streaming API, machine learning extension, and structured streaming.
You'll begin by learning data processing fundamentals using Resilient Distributed Datasets (RDDs), SQL, Datasets, and Dataframes APIs. After grasping these fundamentals, you'll move on to using Spark Streaming APIs to consume data in real time from TCP sockets, and integrate Amazon Web Services (AWS) for stream consumption.

Target Audience:

This course is aimed at IT professionals seeking to learn Spark to process big data. This course is get you up and running with Apache Spark and Python. You'll integrate Spark with AWS for real-time analytics. Finally, you'll apply processed data streams to machine learning APIs of Apache Spark.
Big Data Processing with Apache Spark is for you if you are a software engineer, architect, or IT professional who wants to explore distributed systems and big data analytics.

Objectives:

- By the end of this course, you'll not only have understood how to use machine learning extensions and structured streams but you'll also be able to apply Spark in your own upcoming big data projects.
- After completing this course, you will be able to:
- Write your own Python programs that can interact with Spark
- Implement data stream consumption using Apache Spark
- Recognize common operations in Spark to process known data streams
- Integrate Spark streaming with Amazon Web Services
- Create a collaborative filtering model with Python and the movielens dataset
- Apply processed data streams to Spark machine learning APIs

Prerequisites:

No prior knowledge of Spark is required, however previous experience of working with Python is recommended.

Content:

Lesson 1: Introduction to Spark Distributed Processing

- Introduction to Spark and Resilient Distributed Datasetsll
- Operations Supported by the RDD API
- Self-Contained Python Spark Programs
- Introduction to SQL, Datasets, and DataFrames

Lesson 2: Introduction to Spark Streaming

- Streaming Architectures
- Introduction to Discretized Streams
- Windowing Operations
- Introduction to Structured Streaming

Lesson 3: Spark Streaming Integration with AWS

- Spark Integration with AWS Services
- Integrating AWS Kinesis and Python
- AWS S3 Basic Functionality

Lesson 4: Spark Streaming, ML, and Windowing Operations

- Spark Integration with Machine Learning

Further Information:

For More information, or to book your course, please call us on 00 971 4 446 4987

training@globalknowledge.ae

www.globalknowledge.com/en-ae/

Global Knowledge, Dubai Knowledge Village, Block 2A,First Floor, Office F68, Dubai, UAE