Apache Spark Eğitimi

Açıklama

Bu uygulamalı eğitim, geliştiricilerin Apache Spark ile yüksek performanslı paralel uygulamalar geliştirmek için ihtiyaç duydukları ana kavramları ve uzmanlıkları sunar. Katılımcılar, yapılandırılmış verileri sorgulamak için Spark SQL'in; çeşitli kaynaklardan gelen veri akışı üzerinde gerçek zamanlı işlem gerçekleştirmek için ise Spark Streaming'in nasıl kullanılacağını öğrenirler. Geliştiriciler ayrıca ETL işleme ve yinelemeli algoritmalar gerçekleştirmek için çekirdek Spark'ı kullanan uygulamalar yazma alıştırmaları yapacaklar. Eğitim, dağıtık dosya sisteminde saklanan büyük veri kümeleriyle nasıl çalışılacağını ve bir Hadoop kümesinde Spark uygulamalarının nasıl çalıştırılacağını içerir. Katılımcılar, bu eğitimi aldıktan sonra gerçek dünyadaki zorluklarla yüzleşmeye, daha hızlı ve daha iyi karar vermeyi sağlayan uygulamalar geliştirmeye ve çok çeşitli kullanım senaryoları, mimarileri ve sektörlere uygulanabilen etkileşimli analizler oluşturmaya hazır olacaklar.

Bu eğitimde neler öğreneceksiniz?

- Verileri Hadoop kümesinde dağıtma, saklama ve işleme
- Spark uygulamalarını bir küme üzerinde yazma, yapılandırma ve dağıtma
- Etkileşimli veri analizi için Spark kabuğunu kullanma
- Spark SQL kullanarak yapılandırılmış verileri işleme ve sorgulama
- Canlı veri akışını işlemek için Spark Streaming'i kullanma

Kimler Katılmalı?

Bu eğitim programlama deneyimi olan geliştiriciler ve mühendisler için tasarlanmıştır. Hadoop ve/veya Spark hakkında önceden bilgi sahibi olunmasına gerek yoktur.

Eğitim İçeriği

Introduction to Apache Hadoop and the Hadoop Ecosystem

- Introduction to Apache Hadoop and the Hadoop Ecosystem
- Apache Hadoop Overview
- Data Ingestion and Storage
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to the Hands-On Exercises

Apache Hadoop File Storage

- Apache Hadoop Cluster Components
- HDFS Architecture
- Using HDFS

Distributed Processing on an Apache Hadoop Cluster

- YARN Architecture
- Working With YARN

Apache Spark Basics

- What is Apache Spark?
- Starting the Spark Shell
- Using the Spark Shell
- Getting Started with Datasets and DataFrames
- DataFrame Operations

Working with DataFrames and Schemas

- Creating DataFrames from Data Sources

- Saving DataFrames to Data Sources
- DataFrame Schemas
- Eager and Lazy Execution

Analyzing Data with DataFrame Queries

- Querying DataFrames Using Column Expressions
- Grouping and Aggregation Queries
- Joining DataFrames

RDD Overview

- RDD Overview
- RDD Data Sources
- Creating and Saving RDDs
- RDD Operations

Transforming Data with RDDs

- Writing and Passing Transformation Functions
- Transformation Execution
- Converting Between RDDs and DataFrames

Aggregating Data with Pair RDDs

- Key-Value Pair RDDs
- Map-Reduce
- Other Pair RDD Operations

Querying Tables and Views with Apache Spark SQL

- Querying Tables in Spark Using SQL
- Querying Files and Views
- The Catalog API
- Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark

Working with Datasets in Scala

- Datasets and DataFrames
- Creating Datasets
- Loading and Saving Datasets

- Dataset Operations

Writing, Configuring, and Running Apache Spark Applications

- Writing a Spark Application
- Building and Running an Application
- Application Deployment Mode
- The Spark Application Web UI
- Configuring Application Properties

Distributed Processing

- Review: Apache Spark on a Cluster
- RDD Partitions
- Example: Partitioning in Queries
- Stages and Tasks
- Job Execution Planning
- Example: Catalyst Execution Plan
- Example: RDD Execution Plan

Distributed Data Persistence

- DataFrame and Dataset Persistence
- Persistence Storage Levels
- Viewing Persisted RDDs

Common Patterns in Apache Spark Data Processing

- Common Apache Spark Use Cases
- Iterative Algorithms in Apache Spark
- Machine Learning
- Example: k-means

Apache Spark Streaming: Introduction to DStreams

- Apache Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Streaming Applications

Apache Spark Streaming: Processing Multiple Batches

- Multi-Batch Operations
- Time Slicing
- State Operations
- Sliding Window Operations
- Preview: Structured Streaming

Apache Spark Streaming: Data Sources

- Streaming Data Source Overview
- Apache Flume and Apache Kafka Data Sources
- Example: Using a Kafka Direct Data Source