

Apache Spark ile Büyük Veri İşleme Eğitimi

Süre: 3 Gün

Açıklama

Apache Spark ile Büyük Veri İşleme Eğitimi, büyük ölçekli verilerin dağıtık sistemler üzerinde hızlı ve etkin şekilde işlenmesini sağlayan Apache Spark teknolojisinin temel bileşenlerini ve uygulama geliştirme yöntemlerini ele alan teknik bir eğitim programıdır. Eğitim kapsamında Apache Spark mimarisi, dağıtık veri işleme prensipleri, PySpark, RDD ve DataFrame API, Spark SQL, veri okuma ve yazma işlemleri, veri dönüşümleri ve temel performans optimizasyonu konuları uygulamalı olarak incelenmektedir. Katılımcılar, Apache Spark kullanarak farklı veri kaynaklarından veri okuma, veri temizleme, dönüştürme, analiz etme ve temel ETL süreçleri geliştirme becerileri kazanacaktır. Katılımcıların Apache Spark mimarisini ve temel çalışma prensiplerini öğrenmelerini, PySpark ve Spark SQL kullanarak büyük ölçekli veriler üzerinde veri işleme ve analiz uygulamaları geliştirebilmelerini sağlamaktır. **Kimler Katılmalı**

- Data Engineer
- Big Data Developer
- Data Analyst
- ETL Developer
- Software Developer
- Veri işleme ve büyük veri teknolojileriyle çalışan teknik ekipler

Eğitim İçeriği

1. Introduction to Big Data & Apache Spark

- Big Data kavramı ve temel özellikleri
- Büyük veri işleme teknolojilerine genel bakış
- Apache Hadoop ve Apache Spark karşılaştırması
- Apache Spark'ın kullanım alanları
- Apache Spark ekosistemi
- Spark Core, Spark SQL, Structured Streaming ve MLlib bileşenlerine genel bakış
- Apache Spark'ın avantajları ve kullanım senaryoları
- Open Data Table (Delta Lake, Iceberg)

2. Apache Spark Architecture

- Apache Spark mimarisi
- Driver ve Executor kavramları
- Cluster Manager
- Spark Application yaşam döngüsü
- Jobs, Stages ve Tasks
- Distributed Computing
- Lazy Evaluation
- Transformations ve Actions
- SparkSession ve SparkContext

3. Apache Spark Development Environment

- Apache Spark kurulumu ve yapılandırılması
- Python ve PySpark geliştirme ortamı
- SparkSession oluşturma
- Spark uygulamalarının çalıştırılması
- Local Mode ve Cluster Mode
- Spark Shell ve PySpark kullanımı
- İlk Apache Spark uygulamasının geliştirilmesi

4. Resilient Distributed Datasets (RDD)

- RDD kavramı
- RDD oluşturma yöntemleri
- RDD Transformations
- RDD Actions
- Map, Filter ve Reduce işlemleri
- Key-Value RDD
- RDD Persistence ve Caching
- RDD ve DataFrame karşılaştırması
- Uygulama: RDD ile veri işleme

5. PySpark & DataFrame API

- DataFrame kavramı ve mimarisi
- DataFrame oluşturma
- Schema ve Data Types
- Veri seçme ve filtreleme
- Column Operations
- Veri dönüştürme işlemleri
- GroupBy ve Aggregations
- Join işlemleri

- Missing Values yönetimi
- Duplicate Records yönetimi
- Uygulama: PySpark DataFrame işlemleri

6. Spark SQL

- Spark SQL mimarisi
- SQL ve DataFrame API ilişkisi
- Temporary Views oluşturma
- Spark SQL sorguları
- SELECT, WHERE, GROUP BY ve ORDER BY
- JOIN işlemleri
- Aggregate Functions
- Window Functions
- SQL ile veri analizi
- Uygulama: Spark SQL ile veri sorgulama

7. Data Sources & File Formats

- Spark ile veri okuma ve yazma
- CSV veri formatı
- JSON veri formatı
- Parquet veri formatı
- ORC veri formatına genel bakış
- Schema Inference
- Explicit Schema tanımlama
- JDBC üzerinden veri tabanı bağlantıları
- Partitioned Data
- Veri okuma ve yazma seçenekleri
- Uygulama: Farklı veri kaynaklarıyla çalışma

8. Data Processing & ETL with Spark

- ETL kavramı ve veri işleme süreçleri
- Batch Data Processing
- Veri temizleme
- Veri dönüştürme
- Veri birleştirme
- Veri doğrulama
- Aggregation işlemleri
- ETL Pipeline oluşturma
- Uygulama: PySpark ile temel ETL süreci geliştirme

9. Spark Performance Fundamentals

- Spark uygulamalarında performans kavramı
- Partitioning
- Repartition ve Coalesce
- Caching ve Persistence
- Shuffle Operations
- Broadcast Variables
- Temel Join optimizasyonu
- Spark UI kullanımı
- Execution Plan inceleme
- Temel performans iyileştirme yöntemleri

10. Introduction to Structured Streaming

- Batch ve Streaming veri işleme
- Structured Streaming kavramı
- Streaming DataFrame
- Streaming Sources ve Sinks
- Micro-Batch Processing
- Streaming sorgularının çalıştırılması
- Checkpointing kavramı
- Uygulama: Basit Structured Streaming uygulaması

11. Hands-on Workshop: Apache Spark Data Processing

- Örnek veri setinin hazırlanması
- PySpark ile veri okuma
- DataFrame oluşturma
- Veri temizleme ve dönüştürme
- Spark SQL ile veri analizi
- Farklı veri kaynaklarının birleştirilmesi
- Temel ETL Pipeline geliştirme
- Sonuçların Parquet formatında kaydedilmesi
- Spark UI üzerinden uygulama incelemesi

Ön Koşullar

- Temel Python programlama bilgisi
- SQL sorgulama bilgisi
- Veri tabanları ve veri işleme süreçleri hakkında temel bilgi
- Büyük veri teknolojileri hakkında ön bilgi sahibi olunması önerilir