Distributed Machine Learning

Durasi : 4 Hari

Overview

Sebagian besar pekerjaan machine learning pada dunia nyata melibatkan kumpulan data dalam jumlah atau ukuran yang sangat besar melampaui keterbatasan CPU, memori, dan penyimpanan di satu komputer. Distributed Machine Learning adalah Engine yang dapat mengolah data terdisitribusi di banyak mesin komputer secara parallel, sehingga Analisa data menjadi jauh lebih cepat dan handal.

Apache Spark adalah framework open source yang memanfaatkan komputasi cluster dan penyimpanan terdistribusi untuk memproses set data dan mempekerjakan algoritma machine learning dalam jumlah atau ukuran yang sangat besar secara efisien dan efektif. Oleh karena itu pengetahuan terapan untuk bekerja dengan Apache Spark adalah aset dan pembeda potensial yang sangat bagus untuk engineer machine learning.

Pelatihan ini akan meningkatkan keterampilan peserta, memanfaatkan dan mengimplementasikan Analisa Data menggunakan algoritma Machine Learning yang scalable dan tersebar pada Big Data menggunakan Apache Spark.

Objectives

Setelah mengikuti Pelatihan ini peserta akan dapat:

- Mempraktikan Apache Spark MLib, dan menerapkannya untuk memecahkan masalah machine learning yang melibatkan data dalam jumlah yang kecil maupun besar
- Menulis kode paralel yang mampu berjalan bersamaan pada ribuan CPU.
- Memanfaatkan cluster komputasi skala besar untuk menerapkan algoritma machine learning pada Petabytes data menggunakan Apache SparkML Pipelines.
- Menghindari kesalahan kehabisan memori yang dihasilkan oleh framework machine learning tradisional saat jumlah data tidak mencukupi dengan memori utama komputer.
- Uji ribuan model ML yang berbeda secara paralel untuk menemukan kinerja terbaik

- Menjalankan SQL pada set data yang sangat besar menggunakan Apache SparkSQL dan DataFrame Apache.

Prerequisites

- Mampu membuat program dengan Python
- Memahami Machine Learning
- Keterampilan SQL

Outline

Hari I:

Sesi 1: Data Analytics dan Machine Learning

- Memahami Data Analytics
- Memahami Machine Learning
- Kuis

Sesi 2: BigData dan Distributed Machine Learning

- Konsep dan Arsitektur Hadoop BigData
- Konsep dan Arsitektur Spark Distributed Machine Learning
- Hadoop dan Spark EcoSystem
- Lab: Persiapan dan Instalasi Hadoop

Sesi 3: Instalasi Apache Spark

- Apache Spark MLib Library
- Python PySpark, Anaconda dan Jupyter-lab
- Lab: Install Apache PySpark

Hari II: Tipe Data dan Statistik

Sesi 4: Tipe Data

- Local Vector
- Labeled point

- Local Matrix
- Distributed Matrix
- Lab: Tipe Data

Sesi 5: Statistik

- Mean, median, modus
- Distribution
- Analysis of Variance (Anova)
- Standard normal curve
- Z-Score calculation and T-test
- Probabilities
- Confidence intervals, prediction intervals, p-values
- Kuis

Sesi 6:

- Lab: Statistik

Hari III:

Sesi 7: Spark SQL dan Spark MLLib

- Memahami Spark SQL
- Memahami Spark MLlib
- Studi Kasus dengan Spark MLlib
- Kuis

Sesi 8: Customer Churn Prediction

- Memahami Customer Churn
- Menggunakan Logistic Regression
- Lab: Memprediksi Customer Churn

Sesi 9: Customer Segmentation

- Memahami Customer Segmentation
- Menggunakan K-Means Clustering
- Menggunakan Hierarchical Clustering
- Lab: Customer Segmentation

Hari IV:

Sesi 10: Movie Recommendation

- Memahami Movie Recomendation
- Menggunakan Singular Value Decomposition (SVD)
- Lab: Movie Recommendation

Sesi 11: Sentiment Analysis

- Memahami Sentiment Analysis
- Menggunakan Transformer
- Menggunakan Estimator
- Menggunakan Pipelines
- lab: Sentiment Analysis

Sesi 12:

- Review and Summary