Course Length: 3 days

Course Description

Get to grips with processing large volumes of data and presenting it as engaging, interactive insights using Spark and Python.

Overview

Processing big data in real time is challenging due to scalability, information inconsistency, and fault tolerance. Big Data Analysis with Python teaches you how to use tools that can control this data avalanche for you. With this course, you'll learn effective techniques to aggregate data into useful dimensions for posterior analysis, extract statistical measurements, and transform datasets into features for other systems.

The course begins with an introduction to data manipulation in Python using Pandas. You'll then get familiar with statistical analysis and plotting techniques. With multiple hands-on activities in store, you'll be able to analyze data that is distributed on several computers by using Dask. As you progress, you'll study how to aggregate data for plots when the entire data cannot be accommodated into memory. You'll also explore Hadoop (HDFS and YARN), which will help you tackle larger datasets. The course further covers Spark and its interaction with other tools.

By the end of this course, you'll be able to bootstrap your own Python environment, process large files, and manipulate data to generate statistics, metrics, and graphs.

- Learning Objectives
- Use Python to read and transform data into different formats
- Generate basic statistics and metrics using data on the disk
- Work with computing tasks distributed over a cluster
- Convert data from various sources into storage or querying formats
- Prepare data for statistical analysis, visualization, and machine learning
- Present data in the form of effective visuals

Scope

This course is designed for Python developers, data analysts, and data scientists. This course is not for beginners.

Target Audience

Big Data Analysis with Python is designed for Python developers, data analysts, and data scientists who want to get hands-on with methods to control data and transform it into impactful insights. Basic knowledge of statistical measurements and relational databases will help in understanding various concepts explained in this course.

Technical Requirements

Hardware:

For the optimal student experience, we recommend the following hardware configuration:

- Processor: Intel or AMD 4-core or better
- Memory: 8 GB RAM
- Storage: 20 GB available space

Software:

- Any of the following operating systems:
- Browser: Google Chrome or Mozilla Firefox
- Conda
- Jupyter lab

Course Outline

Lesson 1: The Python Data Science Stack

- Python Libraries and Packages
- Using Pandas
- Data Type Conversion
- Aggregation and Grouping
- Exporting Data from Pandas
- Visualization with Pandas

Lesson 2: Statistical Visualizations

- Types of Graphs and When to Use Them
- Components of a Graph
- Which Tool Should Be Used?
- Types of Graphs
- Pandas DataFrames and Grouped Data
- Changing Plot Design: Modifying Graph Components
- Exporting Graphs

Lesson 3: Working with Big Data Frameworks

- Hadoop
- Spark
- Writing Parquet Files
- Handling Unstructured Data

Lesson 4: Diving Deeper with Spark

- Getting Started with Spark DataFrames
- Writing Output from Spark DataFrames
- Exploring Spark DataFrames
- Data Manipulation with Spark DataFrames
- Graphs in Spark

Lesson 5: Handling Missing Values and Correlation Analysis

- Setting up the Jupyter Notebook
- Missing Values
- Handling Missing Values in Spark DataFrames
- Correlation

Lesson 6: Exploratory Data Analysis

- Defining a Business Problem

- Translating a Business Problem into Measurable Metrics and Exploratory Data Analysis (EDA)
- Structured Approach to the Data Science Project Life Cycle

Lesson 7: Reproducibility in Big Data Analysis

- Reproducibility with Jupyter Notebooks
- Gathering Data in a Reproducible Way
- Code Practices and Standards
- Avoiding Repetition

Lesson 8: Creating a Full Analysis Report

- Reading Data in Spark from Different Data Sources
- SQL Operations on a Spark DataFrame
- Generating Statistical Measurements