Course duration

- 2 days

Course Benefits

- Quick Intro to Spark / PySpark
- Applying Spark SQL / DataFrames to problems that lend themselves to being solved using SQL and Pivot tables
- Exploratory Data Analysis (EDA)-visual analysis using graphs

Course Outline

1. Introduction to Apache Spark
    1. What is Apache Spark
    2. The Spark Platform
    3. Spark vs Hadoop's MapReduce (MR)
    4. Common Spark Use Cases
    5. Languages Supported by Spark
    6. Running Spark on a Cluster
    7. The Spark Application Architecture
    8. The Driver Process
    9. The Executor and Worker Processes
    10. Spark Shell
    11. Jupyter Notebook Shell Environment
    12. Spark Applications
    13. The spark-submit Tool
    14. The spark-submit Tool Configuration
    15. Interfaces with Data Storage Systems
    16. Project Tungsten
    17. The Resilient Distributed Dataset (RDD)
    18. Datasets and DataFrames
    19. Spark SQL, DataFrames, and Catalyst Optimizer
    20. Spark Machine Learning Library
    21. GraphX
    22. Extending Spark Environment with Custom Modules and Files
    23. Summary
2. The Spark Shell
    1. The Spark Shell
    2. The Spark v.2 + Command-Line Shells
    3. The Spark Shell UI
    4. Spark Shell Options

5. Getting Help
6. Jupyter Notebook Shell Environment
7. Example of a Jupyter Notebook Web UI (Databricks Cloud)
8. The Spark Context (sc) and Spark Session (spark)
9. Creating a Spark Session Object in Spark Applications
10. The Shell Spark Context Object (sc)
11. The Shell Spark Session Object (spark)
12. Loading Files
13. Saving Files
14. Summary

3. Introduction to Spark SQL
    1. What is Spark SQL?
    2. Uniform Data Access with Spark SQL
    3. Hive Integration
    4. Hive Interface
    5. Integration with BI Tools
    6. What is a DataFrame?
    7. Creating a DataFrame in PySpark
    8. Commonly Used DataFrame Methods and Properties in PySpark
    9. Grouping and Aggregation in PySpark
    10. The "DataFrame to RDD" Bridge in PySpark
    11. The SQLContext Object
    12. Examples of Spark SQL / DataFrame (PySpark Example)
    13. Converting an RDD to a DataFrame Example
    14. Example of Reading / Writing a JSON File
    15. Using JDBC Sources
    16. JDBC Connection Example
    17. Performance, Scalability, and Fault-tolerance of Spark SQL
    18. Summary
4. Practical Introduction to Pandas
    1. What is pandas?
    2. The Series Object
    3. Accessing Values and Indexes in Series
    4. Setting Up Your Own Index
    5. Using the Series Index as a Lookup Key
    6. Can I Pack a Python Dictionary into a Series?
    7. The DataFrame Object
    8. The DataFrame's Value Proposition
    9. Creating a pandas DataFrame
    10. Getting DataFrame Metrics
    11. Accessing DataFrame Columns
    12. Accessing DataFrame Rows
    13. Accessing DataFrame Cells
    14. Using iloc
    15. Using loc
    16. Examples of Using loc
    17. DataFrames are Mutable via Object Reference!

18. Deleting Rows and Columns
19. Adding a New Column to a DataFrame
20. Appending / Concatenating DataFrame and Series Objects
21. Example of Appending / Concatenating DataFrames
22. Re-indexing Series and DataFrames
23. Getting Descriptive Statistics of DataFrame Columns
24. Getting Descriptive Statistics of DataFrames
25. Applying a Function
26. Sorting DataFrames
27. Reading From CSV Files
28. Writing to the System Clipboard
29. Writing to a CSV File
30. Fine-Tuning the Column Data Types
31. Changing the Type of a Column
32. What May Go Wrong with Type Conversion
33. Summary

5. Data Visualization with seaborn in Python
    1. Data Visualization
    2. Data Visualization in Python
    3. Matplotlib
    4. Getting Started with matplotlib
    5. Figures
    6. Saving Figures to a File
    7. Seaborn
    8. Getting Started with seaborn
    9. Histograms and KDE
    10. Plotting Bivariate Distributions
    11. Scatter plots in seaborn
    12. Pair plots in seaborn
    13. Heatmaps
    14. Summary

Class Materials

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Python class:

- Knowledge of SQL.
- Familiarity with Python (or the ability to learn the basics of a new language).