Data Integration with Python & Spark

www.vinsys.com

Data Integration with Python & Spark

Course Overview

This 4-day course is designed to provide participants with a strong foundation in building scalable and efficient data integration workflows using Python and Apache Spark. Through practical, hands-on sessions, participants will learn how to read, transform, and write structured and unstructured data from various sources. The course emphasizes real-world scenarios including ETL operations, data cleaning, data preparation for analytics, and performance optimization. By the end of the course, participants will be equipped with the skills to handle large-scale data processing tasks using PySpark in production environments.

Prerequisites

- Good understanding of Python programming (data types, functions, and libraries like pandas)
- Basic knowledge of SQL and data structures
- Familiarity with concepts of data processing and file formats (CSV, JSON, etc.)
- No prior knowledge of Spark required, but helpful

Target Audience

- Data engineers and analysts looking to automate data workflows
- Python developers expanding into big data and ETL
- Technical professionals working on data platforms and pipelines

Course Objectives

- Understand the architecture and components of Apache Spark
- Use PySpark for reading, transforming, and writing large-scale datasets
- Build robust ETL pipelines integrating multiple data sources
- Optimize Spark jobs for better performance and scalability
- Apply best practices in data integration, handling schema changes and errors

Course Outline

Module 1: Introduction to Data Integration

- The role of data integration in modern analytics
- Overview of ETL (Extract, Transform, Load) processes
- Introduction to data formats and source types (CSV, JSON, SQL, APIs)
- Challenges in data integration: Volume, Variety, Velocity

Module 2: Python for Data Engineering

- Data manipulation with Pandas
- Reading and writing data (CSV, Excel, JSON, etc.)
- Handling missing data and data cleaning basics
- Python scripting for automation

Module 3: Apache Spark & PySpark Fundamentals

- What is Apache Spark and when to use it
- Spark architecture and components (Driver, Executors, Cluster Manager)
- Introduction to PySpark
- SparkSession and RDDs vs. DataFrames

Module 4: Working with Spark DataFrames

- Creating and manipulating DataFrames
- Reading data from files (CSV, JSON, Parquet)
- Data transformation with PySpark (filter, join, groupBy, agg)
- Writing results to various formats

Module 5: Data Integration from Multiple Sources

- Connecting to relational databases (PostgreSQL, MySQL) using JDBC
- Consuming REST APIs and processing JSON/XML
- Combining data from multiple heterogeneous sources
- Scheduling and automating data integration jobs

Module 6: Building End-to-End ETL Pipelines

- Designing scalable data workflows
- Example: Pipeline from CSV + API to cleaned Spark DataFrame
- Logging and error handling in data jobs
- Performance tuning and optimization in Spark

Module 7: Final Project & Review

- Real-world integration scenario (e.g., integrating sales + customer data)
- Building the full pipeline using Python + Spark
- Walkthrough, feedback, and wrap-up
- Q&A and follow-up guidance

VINSYS®

www.vinsys.com