KOENIG
step forward

KOENIG SOLUTIONS · COURSE OUTLINE

PySpark & Airflow: ETL DataFrames & DAG Orchestration

Learn to build ETL pipelines with PySpark DataFrames and orchestrate them using Airflow DAGs.

8	45
MODULES	TOPICS

COURSE OVERVIEW

This course covers extract, transform and load processes using PySpark, starting with file transformations across formats such as CSV, JSON, Parquet and Avro, and progressing through DataFrame creation, common operations and advanced techniques including window functions, UDFs and Catalyst optimisation. It then covers working with tables in Hive or Delta Lake, including CRUD operations, partitioning and PySpark SQL queries. The course moves on to Apache Airflow, covering DAGs, operators and tasks, configuring Airflow to run PySpark jobs, and designing and scaling workflows. It concludes with orchestrating multiple chained PySpark jobs and building an end-to-end ETL pipeline that combines PySpark transformations with Airflow orchestration.

WHO THIS COURSE IS FOR

Data Engineers | **ETL Developers** | **Big Data Developers** | **Data Pipeline Architects** | **Workflow Orchestration Engineers**

LEARNING OBJECTIVES

- Explain the components of a PySpark ETL workflow
- Apply file transformations across CSV, JSON, Parquet and Avro formats
- Create and manipulate DataFrames using joins, filters and aggregations
- Implement advanced DataFrame operations such as window functions and UDFs
- Manage tables in Hive or Delta Lake using PySpark SQL
- Configure Airflow DAGs, operators and tasks to run PySpark jobs
- Design and monitor workflows with scheduling and task dependencies
- Build an end-to-end ETL pipeline orchestrated with Airflow

COURSE SYLLABUS

8 Modules, 45 Topics

01

Introduction to ETL with PySpark

TOPICS

Overview of ETL (Extract, Transform, Load) · Advantages of using PySpark for ETL processes · Key components of a PySpark ETL workflow

02

File Transformations in PySpark

READING AND WRITING VARIOUS FILE FORMATS

- CSV
- JSON
- Parquet
- Avro
- Handling nested and semi-structured data
- File partitioning strategies for large datasets
- Best practices for schema management during file transformations

03

DataFrame Handling in PySpark

TOPICS

- Introduction to DataFrames in PySpark

CREATING AND INITIALIZING DATAFRAMES FROM

- Structured data files
- RDDs
- Databases

COMMON DATAFRAME OPERATIONS

- Selecting, filtering, and sorting data
- Aggregations and group-by operations
- **Joins:** Inner, outer, left, right, and cross joins
- Managing schema evolution and schema enforcement

04

Advanced DataFrame Operations

TOPICS

- Window functions and their use cases
- Handling nulls and missing data in DataFrames
- Writing efficient UDFs (User Defined Functions)
- Broadcasting variables and its implications in distributed systems
- Optimizing PySpark jobs using Catalyst Optimizer

05

Working with Tables in PySpark

TOPICS

- Reading and writing tables in Hive or Delta Lake
- Performing CRUD operations on tables
- Partitioning and bucketing tables
- Managing table metadata and catalogs
- Using PySpark SQL for complex table queries

06

Creating DAGS in Airflow to Orchestrate PySpark Jobs

TOPICS

- Overview of Apache Airflow and its architecture
- **Key concepts:** DAGs, Operators, and Tasks

SETTING UP AIRFLOW TO RUN PYSPARK JOBS

- Creating custom operators for PySpark
- Configuring connections and variables in Airflow

DESIGNING ROBUST WORKFLOWS WITH AIRFLOW

- Scheduling and triggering DAGs
- Setting task dependencies and execution sequences
- Monitoring and debugging failed DAG runs
- Best practices for scaling Airflow workflows

07

Orchestration of Multiple PySpark Jobs

TOPICS

- Strategies for dividing ETL processes into modular PySpark jobs

CHAINING PYSPARK JOBS USING AIRFLOW

- Passing data and metadata between jobs
- Managing interdependencies and avoiding race conditions
- Handling retries and fault tolerance in job orchestration
- Logging and tracking progress across multiple jobs

08

Building an End-to-End Pipeline

IMPLEMENTING AN END-TO-END ETL PIPELINE USING PYSPARK AND AIRFLOW

- Data extraction from multiple file formats
- Transformations involving schema standardization and data cleansing
- Loading processed data into a database or data lake
- Orchestrating the pipeline with Airflow DAGs

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training | Live Online Classes (ILO) | Fly-Me-A-Trainer (FMAT) | Flexi (Self-Paced) | 1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com | info@koenig-solutions.com