KOENIG SOLUTIONS · COURSE OUTLINE

Python, SQL & PySpark: RDDs, Joins & Streaming

Learn to write Python programs, query and manage data with SQL, and process large-scale data with PySpark RDDs, DataFrames and streaming.

12 Days	12	147	10
DURATION	MODULES	TOPICS	HANDS-ON ACTIVITIES

COURSE OVERVIEW

This course begins with core Python programming, covering syntax, data structures, functions, file handling, exception handling, and external libraries such as NumPy, Pandas and Matplotlib. It then moves into SQL, covering statement types, sorting and filtering, joins and subqueries, built-in functions, GROUP BY and HAVING, data modification, triggers, stored procedures and indexes. The final section covers PySpark, including RDD creation and transformations, actions, DataFrames, collaborative filtering, Spark Streaming, and an ETL capstone project.

WHO THIS COURSE IS FOR

Python Developers | **Data Analysts** | **Database Developers** | **Data Engineers** | **SQL Developers**

LEARNING OBJECTIVES

- Write Python programs using operators, slicing and string formatting
- Control program flow using if statements, loops and exception handling
- Define functions with parameters, arguments and docstrings
- Apply external libraries such as NumPy, Pandas and Matplotlib
- Query and filter data using SQL SELECT statements and predicates
- Combine tables using inner, outer, cross and self joins
- Summarise data using GROUP BY, HAVING and built-in functions
- Build and manage RDDs, DataFrames and Spark Streaming pipelines

COURSE SYLLABUS

12 Modules, 147 Topics

Python (4 Days)

01

Day 1

TOPICS

Installing and setting up python

WRITING YOUR VERY FIRST PROGRAM IN PYTHON

Printing Hello World

Operators and Expressions

SLICING

Negative slicing · Using step in slicing · Slicing backwards

STRINGS

String operators · String formatting

PROGRAM FLOW CONTROL IN PYTHON

if statement · elif · for loop · continue and break · while loop

02

Day 2

LIST AND TUPLES

mutable vs Immutable objects · List · Sorting a list · Removing items from list · Replacing items in list · What are tuples · Performing basic functions to a tuple

Dictionary and Sets

FUNCTIONS

Defining a function · Parameters and arguments · Returning values · Docstring · *args

03

Day 3

INPUT AND OUTPUT IN PYTHON

Reading and writing to a text file · Appending to a file · Object persistence using shelve

Exception handling in python · Generators, Decorators and lambda expression

04

Assessment

Day 4

TOPICS

Introduction to external libraries in Python

DEEP DIVE INTO LIBRARIES

NumPy, Pandas and Matplotlib

SQL (4 Days)

05

Hands-on

Day 1

INTRODUCTION TO SQL

- Introduction
- Work with Schemas
- Explore the structure of SQL Statements DDL, DML, DCL
- Examine the SELECT statements
- Work with data types
- Handle NULLs

06

3 Hands-ons

Day 2

SORT AND FILTER RESULTS IN SQL

Sort your results · Limit the sorted results · Page results · Remove duplicates · Filter data with predicates

COMBINE MULTIPLE TABLES WITH JOINS IN SQL

Understand joins concepts and syntax · Use Inner joins · Use Outer joins · Use Cross joins · Use Self joins

WRITE SUBQUERIES IN SQL

Understand Subqueries · Use scalar or multi-valued subqueries · Use self-contained or correlated subqueries

07

2 Hands-ons

Day 3

TOPICS

- Use built-in functions and GROUP BY in SQL

CATEGORIZE BUILT-IN FUNCTIONS

- Use aggregate functions - AVG SUM MIN MAX COUNT
- Use Mathematical functions - ABS, COS/SIN, ROUND RAND
- Use Ranking functions - RANK, DENSE-RANK
- Use Analytical function - LAG, LAST_VALUE, LEAD, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK
- Use Logical functions - CHOOSE, GREATEST, LEAST
- Summarize data with GROUP BY
- Filter groups with HAVING

MODIFY DATA WITH SQL

- Insert data
- Generate automatic values
- Update data
- Delete data
- Merge data based on multiple tables

08

2 Hands-ons | Assessment

Day 4

TOPICS

Triggers

STORED PROCEDURE

Stored procedures · Create · Modify · Delete · Execute · Specify parameters

INDEXES

Heaps (Tables without Clustered Indexes) · Clustered & Non-Clustered Indexes

PySpark (4 Days)

09

Day 1

FUNDAMENTALS OF PYSPARK

- A Brief Primer on PySpark
- Brief Introduction to Spark
- Apache Spark Stack
- Spark Execution Process
- Newest Capabilities of PySpark
- Cloning GitHub Repository

RESILIENT DISTRIBUTED DATASETS

- Resilient Distributed Datasets
- Creating RDDs
- Schema of an RDD
- Understanding Lazy Execution
- Introducing Transformations – .map(…)
- Introducing Transformations – .filter(…)
- Introducing Transformations – .flatMap(…)
- Introducing Transformations – .distinct(…)
- Introducing Transformations – .sample(…)
- Introducing Transformations – .join(…)
- Introducing Transformations – .repartition(…)
- **Project 1:** Count Data Project (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition)
- **Project 2:** Weather Temperature Crunch (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition on instream data)

10

Day 2

RESILIENT DISTRIBUTED DATASETS AND ACTIONS

- Introducing Actions – .collect(…)
- Introducing Actions – .reduce(…) and .reduceByKey(…)
- Introducing Actions – .count()
- Introducing Actions – .foreach(…)
- Introducing Actions – .aggregate(…) and .aggregateByKey(…)
- Introducing Actions – .coalesce(…)
- Introducing Actions – .combineByKey(…)
- Introducing Actions – .histogram(…)
- Introducing Actions – .sortBy(…)
- Introducing Actions – Saving Data
- Introducing Actions – Descriptive Statistics
- Project 3: 10 Tasks in Students/Professor University Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions.)
- Project 4: 8 Tasks in Customer Data Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions through specified applicability)
- **Project 5:** Movie ratings

DATAFRAMES AND TRANSFORMATIONS

- Creating DataFrames
- Specifying Schema of a DataFrame
- Interacting with DataFrames
- The .agg(…) Transformation
- The .sql(…) Transformation
- Creating Temporary Tables
- Joining Two DataFrames
- Performing Statistical Transformations
- The .distinct(…) Transformation
- **Project 6:** CompanyMegaData (doing all the transformation logics, columunal logic and aggregation and exploratory data analysis)
- **Project 7:** University Data (end to end pyspark execution of insight delivery on University Data)

11

Day 3

COLLABORATIVE FILTERING AND TECHNIQUES

- Collaborative filtering
- Utility Matrix
- Explicit and Implicit Rating
- Expected Results
- Dataset
- Joining Dataframe
- Train and Test Data
- ALS model
- Optimization Hyperparameter tuning and cross validation
- Best model and evaluate prediction
- **Project 8:** IMDB Rating project (Optimization logics focused on the project with extensive pyspark logic and clever techniques of manipulation)

SPARK STREAMING

- Introduction to spark streaming
- Spark streaming with RDD
- Spark streaming Context
- Spark streaming Reading Data
- Spark streaming Cluster Restart
- Spark streaming RDD Transformation
- Spark streaming DF and Display
- Spark streaming DF Aggregation
- **Project 9:** Streaming Crunch Dataset (orchestration of a stream pipeline project of end to end execution of the ingestion of live data)

12

Day 4

SPARK ETL AND CAPSTONE PROJECT

- Introduction to ETL
- ETL Pipeline
- Dataset
- Preprocessing, extraction, transformation
- Loading Data and cleaning
- RDS Networking

Day 4 (continued)

SPARK ETL AND CAPSTONE PROJECT (CONTINUED)

- Downloading PostGres
- Configuration and execution
- **Project 10:** Completion of Capstone Project (Full end to end project Streaming Crunch Dataset of entire pyspark concepts from data exploratory to applying techniques and finding out the logics to the requirement of the dataset along with applying multiple ways to solve a solution and figuring out the correct and most optimized way and efficient way)

HANDS-ON ACTIVITIES

Practise What You Learn

#	MODULE	ACTIVITY
04	**Day 4**	**Assessment:** Assessment
05	**Day 1**	**Hands-on:** Work with SELECT statements
06	**Day 2**	**HANDS-ON** ■ Sort and filter query results ■ Query multiple tables with joins ■ Use Subqueries
07	**Day 3**	**HANDS-ON** ■ Use built-in functions ■ Modify data
08	**Day 4**	**HANDS-ON** ■ Stored procedure ■ Indexes **Assessment:** Assessment

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