DCT DATA SCIENCE ASSOCIATE

Gain practical data science skills in this 40-hour DCT Associate course covering data collection, analysis, visualization, and basic machine learning using Python and real-world datasets.

Duration: 5 Days

Course Code: DCT - DS - DSA

Overview:

- The Data Science Associate course is a practical, hands-on course designed to introduce participants to the fundamentals of data science, from data collection and cleaning to analysis, visualization, and basic machine learning. Over 40 hours, participants will learn how to work with real-world datasets using Python, explore data trends, build simple predictive models, and understand how data science integrates with cloud and data engineering workflows. This workshop is ideal for students, cloud professionals, database developers, and administrators looking to expand their skills into the growing field of data science.

Course Objectives

By the end of this course, participants will be able to:

- Understand the data science lifecycle and key roles.
- Prepare, clean, and explore datasets effectively.
- Apply exploratory data analysis (EDA) techniques using Python.
- Build and evaluate basic machine learning models.
- Use visualization tools to communicate findings.
- Integrate cloud-based data tools for scalable analysis.
- Understand the link between data engineering and data science workflows.

Target Audience

- Aspiring data engineers
- Cloud professionals expanding into analytics
- Database developers and administrators
- Students beginning their data science journey

Prerequisites

- Basic programming (Python preferred)
- Understanding of databases and SQL
- Familiarity with cloud or Linux environment is a plus

Course Outline

Module 1: Introduction to Data Science and the Data Ecosystem

- What is Data Science?
- Data Science vs. Data Engineering vs. Machine Learning
- The Data Science Lifecycle
- Key tools and technologies (Python, Jupyter, Pandas, scikit-learn)
- Overview of roles in data teams
- Setting up the data science environment

Lab Exercises

- Setting up Jupyter or VS Code notebooks
- Exploring a sample dataset (CSV)
- Simple data queries and transformations

Module 2: Data Acquisition, Cleaning, and Preparation

- Types and sources of data (databases, APIs, files, cloud storage)
- Loading data with Pandas and SQL
- Data wrangling: handling missing data, duplicates, and outliers
- Data types and conversions
- Feature extraction and normalization
- Intro to ETL and integration with data engineering workflows
- Lab Exercises
- Loading data from CSV and SQL databases
- Cleaning and transforming a messy dataset
- Writing and executing data cleaning scripts in Python

Course Outline:

Module 3: Exploratory Data Analysis and Visualization

- Descriptive statistics and data summaries
- Correlation, covariance, and feature relationships
- Visualizing data using Matplotlib and Seaborn
- Identifying patterns and trends
- Introduction to hypothesis testing

Lab Exercises

- Plotting and summarizing dataset features
- Detecting correlations and relationships visually
- Building a small exploratory data analysis report

Module 4: Introduction to Machine Learning

- Machine learning fundamentals
- Supervised vs. Unsupervised Learning
- Model training, validation, and evaluation
- Classification and regression algorithms
- Model performance metrics (accuracy, precision, recall, RMSE)
- Introduction to scikit-learn

Lab Exercises

- Building a regression and classification model
- Evaluating model accuracy
- Feature scaling and model tuning basics

Module 5: Cloud Data Science and Capstone Project

- Cloud data science overview (AWS, Azure, GCP)
- Managed services: SageMaker, Vertex AI, Azure ML
- Using cloud storage and data warehouses (BigQuery, Redshift, Synapse)
- Integrating notebooks with cloud platforms
- Real-world workflow: from data ingestion to insight

Capstone Lab (Comprehensive Project)

- Acquire, clean, and analyze a real dataset
- Perform EDA and build a predictive model
- Visualize and present results in a Jupyter notebook
- Discuss deployment and next steps

Tools and Technologies

- Programming: Python 3.x
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, scikit-learn
- Environment: JupyterLab or VS Code
- Databases: SQLite / PostgreSQL
- Visualization: Matplotlib, Seaborn, or Plotly