Alphanove Academy

Unlock your potential with Alphanove Academy - where diversity meets opportunity.

Course Contents

Module 1: Big data engineering foundations

- Introduction
- The Data Engineering Landscape and Lifecycle
- Data Pipelines
- Data Ingestion and Data Storage
- Enterprise Data Warehouses
- Batch vs Real-Time Processing
- Structured, Unstructured and Complex Data

Module 2: Python Programming

- The Python environment
- Debugging
- Arithmetic Variable Assignment and Strings
- Lists and Sets
- Dictionaries, Tuples and Operators
- Control Flow
- Loops
- Functions
- Object Oriented Programming
- Advanced Python
- Error Handling

Module 3: Data Storage and Retrieval

- Introduction to databases and data storage
- Relational databases (SQL)
- NoSQL databases (e.g., MongoDB, Cassandra)
- Data warehousing (e.g., Amazon Redshift, Google BigQuery)
- Data lake vs. data warehouse

Module 4: Data Formats and Processing Libraries

- JSON, CSV, XLSX and YAML
- Tabular Data
- Intro to Pandas, NumPy

- Pandas Dataframes
- Data Cleaning in Pandas

Module 5: Data ingestion

- Principles of Data Ingestion
- Batch Processing
- Real-Time Data Processing
- APIs and Requests
- Kafka Essentials
- Kafka-Python
- Streaming in Kafka

Module 6: Data transformation and Data orchestration

- Data Transformations: ELT & ETL
- Apache Spark and Pyspark
- Distributed Processing with Spark
- Integrating Spark & Kafka
- Integrating Spark & AWS S3
- Spark Streaming
- Apache Airflow
- Integrating Airflow & Spark

Module 7: Data Quality and Governance

- Introduction to data quality
- Data validation and verification
- Data cleaning and normalization
- Data governance and compliance
- Introduction to data cataloging

Module 8: Infrastructure and Deployment

- Cloud vs. on-premise infrastructure
- Introduction to containerization (e.g., Docker, Kubernetes)
- Deployment strategies (e.g., blue-green, canary)
- Monitoring and logging