

Advanced Big Data Analytics for Technical Professionals

Machine learning, databases, Spark, NLP and deep-learning workflows

FA-0588 · 24 days

Data & Analytics · Software Development · DevOps, Cloud & Infrastructure

Develop analytics projects across supervised and unsupervised learning, database design, Spark, NLP and PyTorch using guided datasets.

Introduction

This extended technical course brings together machine learning, database design, distributed analytics, text processing and deep learning. Participants work through nine sections with guided projects and selected demonstrations, building on existing Python, SQL and introductory machine-learning experience.

Exercises use prepared public, synthetic or appropriately de-identified datasets. Malaysian contexts can inform illustrative examples, but no project implies access to a named organisation's records or a validated lending, medical, election or investment system. Emphasis is on model evaluation, data quality, engineering trade-offs and communicating limitations rather than becoming an expert by completing the course.

Learning outcomes

The course teaches participants to:

- Implement and evaluate ensemble models, tune hyperparameters and address feature and class-imbalance issues.
- Explore clustering, dimensionality reduction, anomaly detection and recommendation examples.
- Design relational schemas and compare MongoDB/Cassandra data-modelling approaches.
- Process data with Spark DataFrames/SQL and explore a selected Structured Streaming workflow.
- Develop text-processing, classification and entity-extraction pipelines using spaCy and compatible tools.

- Build and compare introductory PyTorch neural-network, CNN and sequence-model examples.
- Explore topic modelling and translation/summarisation demonstrations with appropriate models rather than assuming spaCy includes language generation.
- Organise a reproducible project with Git, Docker and notebooks.

Prerequisites

- Working Python programming skills and basic command-line experience.
- Intermediate understanding of machine-learning algorithms and basic statistical evaluation.
- Basic SQL, relational databases and general database concepts.
- Introductory NLP knowledge is helpful before the advanced text-processing section.
- A compatible prepared Python/database/Spark environment and suitable compute for selected deep-learning exercises; larger examples can be demonstrated using prepared results.
- Use only approved public, synthetic or de-identified learning data.

Training outline

1. Days 1–5 — Supervised Machine Learning

Days 1–2 — Ensemble Methods

- Random Forests
- Gradient Boosting
- Guided classification exercise using a public or synthetic dataset; discuss why retrospective fit is not reliable election forecasting.

Days 3–4 — Hyperparameter Tuning

- Grid Search
- Random Search
- Illustrative credit-scoring model tuning with de-identified or synthetic data, evaluation and fairness limitations.

Day 5 — Features and Imbalanced Data

- Feature Engineering
- Handling Imbalanced Data

- Illustrative transaction-anomaly/fraud classification project with class imbalance and false-positive trade-offs.

2. Days 6–8 — Unsupervised Machine Learning

Day 6 — Clustering

- K-Means Clustering
- Hierarchical Clustering
- Customer-segmentation exercise using a prepared synthetic dataset.

Day 7 — Dimensionality Reduction

- Principal Component Analysis (PCA)
- t-SNE for exploratory visualisation; compare its purpose with PCA feature reduction.
- Feature-reduction exercise using a suitable de-identified or synthetic dataset, without clinical conclusions.

Day 8 — Anomalies and Recommendations

- Anomaly Detection
- Recommender Systems
- Illustrative recommendation example for attractions or similar items.

3. Days 9–10 — Relational Database Design

Day 9 — Schema Design and SQL

- Normalization and Denormalization
- Advanced SQL Functions
- Design a schema for a fictional e-commerce application.

Day 10 — Integrity and Security

- Database Integrity and Transactions
- Database Security
- Review integrity and access-control choices for a fictional nonprofit-style database; do not claim complete security.

4. Days 11–12 — NoSQL Essentials

Day 11 — NoSQL Models and MongoDB

- Types of NoSQL Databases
- When to Use NoSQL
- MongoDB exercise using prepared, licensed sample text rather than collecting personal social-media records.

Day 12 — Modelling and Cassandra

- Data Modeling in NoSQL
- Introduction to Cassandra
- NoSQL modelling exercise using a public weather dataset.

5. Days 13–14 — Apache Spark Analytics

Day 13 — Spark Architecture and DataFrames

- Spark Architecture
- Compare RDD concepts with DataFrame/SQL operations for the selected analytics workflow.
- Analyse a prepared, appropriately licensed news-text dataset.

Day 14 — ML and Streaming Demonstrations

- Introductory Spark ML pipeline example and evaluation.
- Structured Streaming concepts and a prepared DataFrame-based example.
- Illustrative traffic-stream analysis using prepared or simulated events.

6. Days 15–17 — NLP Foundations with spaCy

Day 15 — Text Processing

- Text Processing
- Tokenization
- spaCy tokenisation and text-analysis exercise using sample news material.

Day 16 — Classification and Sentiment

- Text Classification Algorithms
- Sentiment-analysis task formulation and a compatible classifier or pipeline; sentiment is not an inherent guarantee of every spaCy model.
- Review-analysis exercise using synthetic or suitably licensed sample customer text.

Day 17 — Entities and Representations

- Named Entity Recognition (NER)
- Word Embeddings
- Extract named entities from public sample documents and evaluate errors.

7. Days 18–20 — Deep Learning with PyTorch

Day 18 — Neural-Network Foundations

- Neural Networks
- PyTorch Basics
- Small neural-network example using an approved image dataset.

Day 19 — Convolutional Networks

- CNN Architecture
- Image Classification
- Illustrative wildlife/image classification project with dataset and performance limitations.

Day 20 — Sequence Models

- RNN and LSTM Basics
- Sequence Prediction
- LSTM sequence-prediction demonstration on historical or synthetic data with chronological evaluation and baseline comparison; no investment-performance claim.

8. Days 21–23 — Advanced NLP Workflows

Day 21 — Topic Modelling

- Latent Dirichlet Allocation using an appropriate topic-modelling library, with spaCy where useful for preprocessing.
- Non-negative Matrix Factorisation and topic interpretation.
- Topic-modelling exercise using a suitably licensed public speech/text dataset; inferred topics require interpretation.

Day 22 — Translation Models

- Sequence-to-sequence model concepts using a compatible translation model.
- Language-model roles and resource/licence constraints; spaCy supports processing but does not itself supply general translation generation.
- English–Bahasa Malaysia translation demonstration using an appropriate model, with bilingual quality review.

Day 23 — Ethics and Summarisation

- NLP Ethics
- Text summarisation using a suitable model or bounded extractive approach, with source checking.
- Summarise sample news and examine omissions, unsupported statements and ethical concerns.

9. Day 24 — Data-Science Development Tools

- Git for Version Control
- Introduction to Docker
- Jupyter Notebook Best Practices
- Set up a small reproducible analytics project using Git, Docker and Jupyter.
- Review reproducibility, model/data limitations and the distinction between a learning project and a validated operational system.