1 2 FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE D COIMBRA 9 0

DEI DEPARTAMENTO ENGENHARIA INFORMÁTICA

Project Information

12th November 2025

Advanced Infrastructures for Data Science

Pedro Neves, pedroneves@dei.uc.pt, 2025/2026

Master in Data Science and Engineering (MDSE) Course

Project Objective

- **Objective**
  - Simulate a **real-world data engineering pipeline** using Docker containers and Kubernetes

- **Pipeline components**
  - Data ingestion (Apache Kafka or Apache Spark)
  - Data storage (postgres, MinIO, etc.) or optional distributed storage (HDFS)
  - Processing engine (Apache Spark)
  - Optional visualization – Grafana or Superset dashboard
  - Optional Monitoring – Prometheus/Grafana

- **Deployment target:**
  - Minikube (local Kubernetes environment) for development and testing
  - Optionally, deployment to a cloud-based Kubernetes cluster for advanced teams

Smart City Data Analytics

- Possible Projects:
  - Real-time traffic congestion detection using streaming
  - Predicting air quality based on meteorological data
  - Smart energy consumption analysis across city districts

- Open Datasets:
  - Open Data Portal – City of Lisbon
  - EU Open Data Portal – Air Quality
  - Transport for London (TfL) API

Climate & Environmental Data

- Possible Projects:
  - Temperature trend analysis using NASA's global datasets
  - Extreme weather prediction using ML pipelines
  - CO_2 emissions clustering by country/sector

- Open Datasets:
  - NOAA Climate Data Online
  - NASA EarthData
  - Our World in Data – CO_2 Emissions

Social Media & Sentiment Analysis

- Possible Projects:
  - Twitter sentiment analysis for major events (sports, elections)
  - Topic clustering from Reddit or news feeds using NLP pipelines
  - Hashtag popularity tracking via Kafka streams

- Open Datasets:
  - Kaggle – Twitter US Airline Sentiment
  - Reddit Comments Dataset
  - News API (Free Tier)

Health & Life Sciences

- Possible Projects:
  - Predict disease spread patterns using public health records
  - Analyze patient readmission data to optimize hospital resources
  - Gene expression clustering (bioinformatics)

- Open Datasets:
  - WHO Global Health Observatory
  - UCI Machine Learning Repository – Heart Disease, Diabetes, etc.
  - Kaggle – COVID-19 Open Research Dataset (CORD-19)

Finance & Fraud Detection

- Possible Projects:
  - Fraud detection using transaction patterns
  - Stock trend analysis with real-time streams
  - Risk modeling using distributed ML training

- Open Datasets:
  - Credit Card Fraud Dataset (Kaggle)
  - Yahoo Finance API
  - World Bank Financial Indicators

Project Proposal (1 slide)

- Project Title & Team Members
  - Short title + names of team members
- Problem / Objective
  - One sentence on what the project aims to demonstrate, indicating the expected output/insight
- Dataset(s)
  - Source, type, and approximate data size
- Proposed Architecture (Diagram)
  - Visual diagram showing data flow between components
  - (e.g., Kafka → Spark → MinIO → Dashboard)
- Technologies / Tools
  - Bullet list of main technologies used (e.g., Docker, Kubernetes, Spark, HDFS, etc.)

Project Deliverables

1. Technical report (5 pages)
   - Problem description, architecture, technologies, and evaluation

2. Deployment files
   - Dockerfiles, Kubernetes manifests, Spark job scripts, …

3. Presentation & Demonstration (10 to 15 mins + Q&A)
   - Presentation and demonstration of the data pipeline and key insights

Delivery Guidelines

- Project proposal submission & presentation: **14th November**
  - Submission platform: InforDocente
- Final project submission: **11th December**
  - Submission platform: InforDocente
- Presentation and demo: **12th December & 19th December** (during theoretical and practical classes)

Evaluation Criteria

- Architecture and design quality (10%)
  - Architectural decisions, scalability, and clean integration of components
- Technical implementation (30%)
  - Working, reproducible system using appropriate distributed technologies (Docker, Kubernetes, Spark, etc.)
- Data handling and analytics (30%)
  - Effective data ingestion, transformation, storage, and/or analytics pipeline
- Technical report (15%)
  - Clarity, structure, and depth of the written report — including problem definition, design choices, challenges, and results
- Presentation and communication (15%)
  - Clear, concise demonstration of the project, ability to explain technical aspects and results