

**Mestrado em Engenharia Informática-Internet das Coisas**

Mestrado, 2º Ciclo

Plano: Despacho n.º 13495/2022 - 18/11/2022

**Ficha da Unidade Curricular: Análise de Grande Volume de Dados**

ECTS: 10; Horas - Totais: 260.0, Contacto e Tipologia, TP:30.0; PL:30.0; OT:30.0;

Ano | Semestre: 1 | S1

Tipo: Obrigatória; Interação: Presencial; Código: 390913

Área Científica: Ciências e Tecnologias da Programação

**Docente Responsável**

Ana Cristina Barata Pires Lopes

Professor Adjunto

**Docente(s)**

Luís Miguel Lopes de Oliveira

Professor Adjunto

Ana Cristina Barata Pires Lopes

Professor Adjunto

Renato Eduardo Silva Panda

Professor Adjunto Convidado

Paulo Sérgio Correia Monteiro

Professor Adjunto Convidado

**Objetivos de Aprendizagem**

Capacitar os estudantes com competências práticas e teóricas em processamento de grandes volumes de dados, utilizando ferramentas modernas de data science e big data num contexto aplicado com projetos práticos.

**Objetivos de Aprendizagem (detalhado)**

Esta unidade curricular visa capacitar os estudantes com competências práticas e teóricas nas áreas de processamento e análise de dados em larga escala. Serão abordadas técnicas modernas de engenharia de dados, preparação, armazenamento, transformação, análise exploratória e visualização de dados, com recurso a ferramentas como Pandas, Dask, Spark e

Streamlit, integradas num contexto de projetos aplicados.

## **Conteúdos Programáticos**

1. Introdução à ciência de dados e Big Data
2. Ambiente de desenvolvimento
3. Python para análise e visualização
4. Aquisição e armazenamento de dados
5. Engenharia de dados e EDA
6. Dashboards interativos
7. Processamento de dados em larga escala

## **Conteúdos Programáticos (detalhado)**

1. Introdução à ciência de dados e Big Data
  - 1.1 Conceitos base, ciclo de vida de projetos, papéis em data science
  - 1.2 Os 5Vs do Big Data: volume, velocidade, variedade, veracidade, valor
  - 1.3 Ética, privacidade, transparência e impacto social
  - 1.4 Reprodutibilidade, documentação, controlo de versões
2. Ambiente de desenvolvimento
  - 2.1 Jupyter, Python e VS Code
  - 2.2 Docker, DevContainers e ambientes reprodutíveis
  - 2.3 Isolamento de dependências (pip, conda)
  - 2.4 Gestão de ambientes com requirements.txt
3. Python para análise e visualização
  - 3.1 Revisão de sintaxe Python, estruturas de dados e scripts básicos
  - 3.2 NumPy e manipulação de arrays
  - 3.3 Pandas para análise tabular
  - 3.4 Visualização com Matplotlib, Seaborn, Plotly
4. Aquisição e armazenamento de dados
  - 4.1 Acesso a dados locais e remotos (CSV, JSON, Parquet)
  - 4.2 APIs REST, autenticação e tratamento de erros
  - 4.3 Web scraping com requests e BeautifulSoup
  - 4.4 Armazenamento em MongoDB (document store) e Redis (key-value)
5. Engenharia de dados e EDA (Exploratory Data Analysis)
  - 5.1 Construção de pipelines ETL
  - 5.2 Limpeza e transformação de dados
  - 5.3 Exploração inicial e análise descritiva
  - 5.4 Otimização com formatos eficientes (Parquet, compressão)
6. Dashboards interativos
  - 6.1 Introdução a Streamlit e Dash
  - 6.2 Construção de interfaces com filtros, tabelas e visualizações
  - 6.3 Integração com pipelines e APIs
  - 6.4 Aplicações práticas nos projetos da UC
7. Processamento de dados em larga escala
  - 7.1 Introdução ao paradigma MapReduce
  - 7.2 Arquiteturas: Hadoop, Spark, Dask

- 7.3 Estratégias de paralelização e chunking
- 7.4 Operações com RDDs e DataFrames em PySpark
- 7.5 Introdução ao MLlib (Spark) e machine learning distribuído

### **Metodologias de avaliação**

A avaliação é contínua e baseada em três projetos práticos obrigatórios:

- Projeto I (35%) – Reorganização de um conjunto de dados, criação de loaders e construção de dashboards interativo
- Projeto II (35%) – Desenvolvimento de pipeline ETL com análise exploratória e visualização
- Projeto III (30%) – Processamento em larga escala com Dask e Spark

A entrega de todos os projetos é obrigatória, assim como a defesa dos mesmos.

### **Software utilizado em aula**

Python (Anaconda, Jupyter Notebooks), Pandas, NumPy, Matplotlib, Seaborn, Plotly, MongoDB, Redis, Dask, PySpark, Streamlit, Dash, Docker, Git, Visual Studio Code

### **Estágio**

Não aplicável

### **Bibliografia recomendada**

- Marr, B. (2022). *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things..* Kogan Page. USA
- McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython..* O'Reilly. USA
- Rioux, J. (2022). *Data Analysis with Python and PySpark..* Manning. USA
- Santos, M. e Costa, C. (2019). *Big Data – Concepts, Warehousing, and Analytics. ..* FCA. Lisboa
- Triguero, I. e Galar, M. (2023). *Large-Scale Data Analytics with Python and Spark..* Cambridge University Press. UK

### **Coerência dos conteúdos programáticos com os objetivos**

Os conteúdos programáticos da unidade curricular foram elaborados para garantir a coerência com os objetivos delineados. Tal é demonstrado abaixo:

Objetivo 1 – Compreender o ciclo de vida em projetos de dados:

- O conteúdo 1 fornece a base conceptual, apresentando o ciclo de vida da ciência de dados, os 5Vs e a importância da ética.

Objetivo 2 – Dominar ferramentas para aquisição, transformação e análise de dados:

- Os conteúdos 2 a 5 abordam desde ambientes de desenvolvimento até pipelines de ETL, com práticas de scraping, APIs, MongoDB, Redis e Pandas.

Objetivo 3 – Aplicar técnicas de exploração e visualização de dados:

- Desenvolvido nos conteúdos 3, 5 e 6, com bibliotecas como Matplotlib, Seaborn, Plotly e uso de Streamlit/Dash para interfaces interativas.

Objetivo 4 – Compreender os paradigmas de processamento distribuído:

– O conteúdo 7 introduz o paradigma MapReduce e frameworks como Hadoop, Spark e Dask, com foco nas arquiteturas modernas e paralelização.

Objetivo 5 – Processar grandes datasets com estratégias escaláveis:

– Incluído nos subtemas 7.3 e 7.4, que exploram chunking, paralelização e processamento tanto em Python como usando Spark e Dask.

Objetivo 6 – Desenvolver dashboards interativos:

– Aplicado no conteúdo 6, onde os alunos constroem no âmbito dos projecto dashboards interativos ligados a pipelines de dados.

Objetivo 7 – Noções de aprendizagem automática em Big Data:

– Introduzido no conteúdo 7.5 com Spark MLlib, contextualizando o uso de ML em ambientes de big data.

### **Metodologias de ensino**

Aulas teórico-práticas para introdução dos conceitos, com exemplos de implementação. Aulas laboratoriais orientadas com uso de notebooks Jupyter e scripts Python. Projetos aplicados ao longo do semestre consolidam os conhecimentos e permitem a aplica

### **Coerência das metodologias de ensino com os objetivos**

As metodologias de ensino adoptadas asseguram o cumprimento eficaz dos objetivos de aprendizagem. As aulas teórico-práticas oferecem o enquadramento necessário, enquanto as demonstrações práticas e notebooks exemplificam na prática os conceitos introduzidos. As aulas PL permitem a experimentação com ferramentas de data science e big data, promovendo o desenvolvimento de competências técnicas. Os projetos práticos ao longo da disciplina articulam os diferentes temas (da aquisição de dados à construção de dashboards e processamento distribuído), reforçando a aprendizagem alinhada com os objetivos.

### **Língua de ensino**

Português

### **Pré-requisitos**

Não aplicável

### **Programas Opcionais recomendados**

Não aplicável

### **Observações**

Objetivos de Desenvolvimento Sustentável:

- 4 - Garantir o acesso à educação inclusiva, de qualidade e equitativa, e promover oportunidades de aprendizagem ao longo da vida para todos;  
9 - Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação;
- 

**Docente responsável**

---