

Escola Superior de Tecnologia de Tomar

Ano letivo: 2023/2024

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

Ricardo Nuno Taborda Campos

Colaborador

Docente(s)

Ricardo Nuno Taborda Campos

Colaborador

Objetivos de Aprendizagem

Esta UC visa a introdução de noções fundamentais acerca da aquisição, processamento, armazenamento e recuperação de elevados volumes de dados com recurso ao paradigma map-reduce e a frameworks de processamento de dados de larga escala

Objetivos de Aprendizagem (detalhado)

Esta UC tem por objetivo introduzir os alunos à aquisição, processamento, armazenamento e recuperação de dados em larga escala como suporte a tarefas de Ciência de Dados (CD).

No final da UC o aluno deverá saber

1. listar os passos envolvidos num projeto de CD em larga escala e descrever as funções de cada um;
2. conhecer as principais ferramentas de desenvolvimento de um projeto de CD;
3. estar familiarizado com os conceitos fundamentais dos grandes volumes de dados;
4. saber aplicar métodos de aquisição de dados com recurso a pacotes de software python, APIs e web scraping;

5. dominar o processo de armazenamento e recuperação de dados em larga escala;
6. conhecer e saber aplicar de forma adequada as estratégias de processamento de dados em larga escala;
7. entender o paradigma map-reduce;
8. conhecer os fundamentos das principais frameworks de processamento de dados em larga escala;
9. saber usar, programar e processar dados em larga escala com recurso à framework Spark.

Conteúdos Programáticos

1. Introdução à Ciência de Dados
2. Ferramentas no contexto da Ciência de Dados
3. Introdução ao Big Data
4. Aquisição de Dados em Larga Escala
5. Armazenamento e Recuperação de Dados em Larga Escala
6. Estratégias de Processamento de Dados em Larga Escala
7. Programação de Aplicações com base no Paradigma Map-Reduce
8. Frameworks de Processamento Dados
9. Processamento Dados com Spark

Conteúdos Programáticos (detalhado)

1. Introdução à Ciência de Dados
 - Definição de ciência de dados em larga escala
 - Visão geral das competências de um cientista de dados em larga escala
 - Etapas de um projeto de ciência de dados
 - A importância da ciência de dados em ambientes com grandes volumes de dados
 - Desafios e oportunidades em ciência de dados e big data
 - Áreas de atuação da ciência de dados em larga escala
 - Tópicos emergentes
 - Repositórios de dados
 - Data lakes
2. Ferramentas de desenvolvimento no contexto da Ciência de Dados
 - Git
 - Github
 - Docker
 - Python (Anaconda - Jupyter Notebook)
 - Google Colab
3. Introdução ao Big Data
 - Definição de Big Data
 - Evolução histórica
 - Características
 - Vantagens
 - Aplicações práticas com grandes volumes de dados
 - Arquitetura de um sistema de Big Data

- Plataformas para processamento de dados em larga escala

4. Aquisição de Dados em Larga Escala

- Formatos de dados (estruturados; semi-estruturados; não-estruturados)
- Extração de Informação a partir de ficheiros
- Extração de Informação com recurso a packages
- Extração de Informação com recurso a APIs
- Extração de Informação com recurso a Web Scraping
- Princípios e ética do web scraping

5. Armazenamento e Recuperação de Dados em Larga Escala

- Bases de dados NoSQL
- Vantagens
- NoSQL vs SQL
- Tipos de bases de dados NoSQL
- Bases de dados NoSQL open-source

6. Estratégias de Processamento de Dados em Larga Escala

- Quantos dados são muitos dados?
- Visão geral das estratégias para processamento de dados em larga escala (compressão de dados; bases de dados; chunking; scale-up (expansão de recursos); scale-out (data parallelism) e big data)
- A importância dos GPUs no contexto da ciência de dados em larga escala

7. Programação de Aplicações de Larga Escala com base no Paradigma Map-Reduce

- Visão geral do paradigma map-reduce
- História do map-reduce
- Funcionamento
- Vantagens
- Frameworks

8. Frameworks de Processamento de Dados em Larga Escala (Hadoop, Spark, Dask)

Hadoop

- Visão geral do Hadoop
- História e evolução
- Características
- Arquitetura
- Ecossistema

Spark

- Visão geral do Spark
- História e evolução
- Características
- Arquitetura
- Spark vs Hadoop Map-Reduce

Dask

- Visão geral do Dask

- Características
- Arquitetura
- Dask vs PySpark

9. Processamento de Dados em Larga Escala com Spark

- Introdução aos conceitos fundamentais de Spark
- RDDs (Resilient Distributed Datasets)
- Spark DataFrames
- Spark Streaming

Metodologias de avaliação

Avaliação Periódica

- P1 - Projeto I (trabalho de grupo): 40%
- P2 - Projeto II (trabalho de grupo): 40%
- F - Frequência: 20%

A classificação final da UC resulta da média ponderada das classificações obtidas nas componentes de avaliação definidas. O aluno obtém aprovação à UC, estando dispensado de Exame, no caso de obter uma nota igual ou superior a 9.5 valores.

Avaliação Final

- Exame: 100% (prova realizada em computador com consulta parcial dos conteúdos)

Requisitos de admissibilidade à frequência e ao exame:

- Mínimo de 70% de assiduidade às aulas durante o período de ensino-aprendizagem (exceto trabalhadores estudantes);
- Nota mínima de 6 valores em AE, onde $AE = ((P1 * 40\%) + (P2 * 40\%) + (F * 20\%))$

O incumprimento de qualquer um destes itens (incluindo a submissão de projetos fora do prazo) impede o aluno de se submeter à frequência e ao exame.

Software utilizado em aula

Python: Anaconda e Jupyter Notebooks; PySpark

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
- Triguero, I. e Galar, M. (2023). *Large-Scale Data Analytics with Python and Spark..* Cambridge

Coerência dos conteúdos programáticos com os objetivos

Esta unidade curricular visa a introdução de noções fundamentais acerca da ciência de dados em larga escala. Em particular, pretende-se que os alunos compreendam em profundidade os objetivos, desafios, etapas, estratégias, ferramentas e a arquitetura necessária ao desenvolvimento e à implementação de um projeto de ciência de dados em larga escala, nomeadamente a partir da aquisição, armazenamento, processamento e recuperação de dados com recurso ao paradigma map-reduce e a frameworks de processamento de dados de larga escala. Os conteúdos programáticos estão em coerência com os objetivos da unidade curricular, atendendo a que:

- O ponto 1 pretende concretizar o ponto 1 dos objetivos
- o ponto 2: objetivo 2
- o ponto 3: objetivo 3
- o ponto 4: objetivo 4
- o ponto 5: objetivo 5
- o ponto 6: objetivo 6
- o ponto 7: objetivo 7
- o ponto 8: objetivo 8
- o ponto 9: objetivo 9

Metodologias de ensino

Exposição dos conteúdos programáticos com recurso ao método expositivo e demonstrativo. Análise e resolução de casos práticos através de notebooks. Os conhecimentos adquiridos serão avaliados através da realização e apresentação de projetos e testes

Coerência das metodologias de ensino com os objetivos

Os objetivos de aprendizagem da UC são atingidos pelo acompanhamento dos estudantes no decurso da realização dos exercícios práticos e na implementação dos projetos, permitindo desta forma que os alunos solidifiquem as competências adquiridas no decurso da UC.

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;

Docente responsável

Assinado por: **Ricardo Nuno Taborda Campos**
Num. de Identificação: 11022050
Data: 2024.03.13 02:19:26+00'00'

