INSTITUTO UNIVERSITÁRIO DE LISBOA

# Processamento de Big Data

Projeto final

13 de março de 2023

## 1 Introdução

Este projeto tem como objetivo principal a consolidação de conhecimentos práticos no desenho e implementação de uma solução computacional para fazer face a um problema de análise de dados, de grande dimensão.

O projeto visa também fomentar a experimentação em outros ambientes de trabalho e execução, para além do computador pessoal. Assim, vão ser utilizados serviços em ambiente *cloud*, da Amazon para contexto académico – AWS Academy – na qual será disponibilizada uma classe de ensino para os(as) alunos(as) da unidade curricular.

Em termos de ferramentas, o projeto deve ser implementado recorrendo sobretudo a funcionalidades disponibilizadas pela plataforma Apache Spark e linguagem de programação Python. Refira-se que o requisito de utilização de Apache Spark e Python é mandatório.

## 2 Contextualização do problema

Pretende-se que seja implementada uma solução computacional para estudo e análise de dados de grande dimensão. Nesse sentido, deverá ser contruído um modelo de análise e processamento de dados baseado em métodos e algoritmos lecionados nas aulas.

A escolha do domínio de dados e respetivo conjunto de dados a utilizar, bem a formulação do próprio problema em estudo, será da responsabilidade dos autores do trabalho.

### 2.1 Domínio de dados

Os dados associados ao problema a formular devem ser obtidos a partir da lista de fontes especificada em anexo, na tabela 1. Alguns ficheiros de dados estão associados a competições no sítio `https://www.kaggle.com` , eventualmente

com requisitos adicionais no acesso aos mesmos. Nesses casos, se necessário, podem contactar o(s) docente(s) tendo em vista a obtenção dos dados.

Os autores do trabalho devem selecionar uma fonte de dados e informar o(s) docente(s) sobre a decisão tomada, até ao dia **21 de março de 2023**.

### 2.2 Algoritmos

Os algoritmos a utilizar que estarão na base da construção do modelo de análise e processamento de dados devem fazer parte da plataforma Apache Spark.

## 3 Implementação

A implementação da solução deve ser modular, ou seja, deve ser composta por mais do que um notebook ou módulo Python. Compete aos autores do trabalho estruturar de forma criteriosa o código implementado. Por outro lado, chama-se a atenção para os seguintes aspetos, também já referidos ao longo das aulas:

- A escolha do domínio de dados, bem como a formulação do problema em estudo, são da maior importância para o sucesso do projeto como um todo. Estas fases não devem ser menosprezadas, em termos relativos.
- Por questões de produtividade, devem ser considerados dois conjuntos de dados aquando do desenvolvimento da solução. Assim, para além dos dados originais, deve ser utilizado um conjunto de dados de menor dimensão, para o caso de tarefas intensivas e frequentes, inerentes ao próprio desenvolvimento da solução.
- Por norma, o ambiente da plataforma na *cloud* AWS Academy não é para ser utilizado em tarefas intrinsecamente associadas ao desenvolvimento da solução. Isto porque existem limites temporais na utilização destes recursos.

## 4 Material a entregar

O trabalho, a ser realizado em grupos constituídos por duas pessoas, deve ser submetido de acordo com as seguintes regras:

- A submissão consiste num arquivo em formato zip (extensão zip e não outra) com os seguintes elementos de avaliação: (i) relatório e (ii) notebooks e/ou módulos Python.
- O prazo de submissão é **12:00 de 8 de abril de 2023**, com o respetivo arquivo zip a ser submetido na plataforma de ensino Moodle. O *link* a utilizar será indicado em momento oportuno.
- O relatório deve ser um documento sucinto e em formato pdf, com o máximo de oito páginas. No relatório, os autores devem:

  - enunciar o problema em estudo e respetivos dados utilizados, abordar os aspetos mais relevantes sobre as decisões tomadas, bem como experiências e testes realizados;
  - ter em consideração os ambientes de desenvolvimento e teste utilizados – computador pessoal e plataforma na *cloud* AWS Academy;
  - incluir uma análise crítica sobre os resultados obtidos, tendo em consideração o problema formulado;
  - incluir uma referência explícita à localização na web dos ficheiros de dados selecionados;
  - incluir outra informação que considerem relevante.

- Os notebooks e/ou módulos Python constituem a solução computacional. Assume-se que os mesmos são auto-explicativos, contendo comentários com nível de detalhe apropriado.

- A submissão do trabalho não pode conter ficheiros de dados.

Refira-se ainda que, de acordo com as regras de avaliação da unidade curricular, este projeto tem uma ponderação de 40% na nota final da unidade curricular.

# 5 Apresentação do trabalho

O trabalho será apresentado oralmente, em local e hora a indicar após submissão do mesmo e de acordo com a disponibilidade dos membros do grupo e dos docentes. Como nota, relembra-se que o resultado da avaliação do trabalho é individual.

Tabela 1: Lista de fontes de informação para seleção de dados.

| Número | URL |
|---|---|
| 1 | https://github.com/otto-de/recsys-dataset |
| 2 | https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london |
| 3 | https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews |
| 4 | https://www.kaggle.com/datasets/erikbiswas/higgs-uci-dataset |
| 5 | https://www.kaggle.com/datasets/dasgroup/rba-dataset |
| 6 | https://www.kaggle.com/datasets/skeller/2021-us-federal-award-data |
| 7 | https://www.kaggle.com/datasets/salikhussaini49/chicago-crimes |
| 8 | https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022 |
| 9 | https://www.kaggle.com/competitions/predict-student-performance-from-game-play |
| 10 | https://www.kaggle.com/competitions/amex-default-prediction |
| 11 | https://www.kaggle.com/competitions/g-research-crypto-forecasting |
| 12 | https://www.kaggle.com/competitions/vsb-power-line-fault-detection |
| 13 | https://www.kaggle.com/competitions/reducing-commercial-aviation-fatalities |
| 14 | https://www.kaggle.com/competitions/microsoft-malware-prediction |
| 15 | https://www.kaggle.com/competitions/ga-customer-revenue-prediction |
| 16 | https://www.kaggle.com/competitions/home-credit-default-risk |
| 17 | https://www.kaggle.com/competitions/outbrain-click-prediction |
| 18 | https://www.kaggle.com/competitions/expedia-hotel-recommendations |