Departamento de Informática

Título do Projeto

Análise de dados em SparkR

Orientador: Paula Prata (pprata@di.ubi.pt)

Co-orientador: Maria Eugénia Ferrão (Departamento de Matemática)

Objetivos

Numa sociedade que produz cada vez maiores quantidades de dados, a análise desses dados de forma a deles retirar conhecimento útil é crucial. A análise de enormes quantidades de dados exige grande capacidade de processamento, sendo a programação paralela uma via óbvia para acelerar o processo. Pretende-se analisar os dados do Exame Nacional do Ensino Médio (ENEM) [1] do Brasil, com cerca de 8 milhões de registos, mostrando os resultados dessa análise graficamente. Por exemplo, mostrar o aproveitamento dos estudantes por nível socioeconómico, mostrar a distribuição dos estudantes por raça ou por estado federal, etc. Para o processamento dos dados deve ser usado o SaparkR [3]. Sendo o Apache Spark um ambiente de execução paralela e distribuída para análise de dados [2], o SparkR fornece uma interface para usar o Spark a partir da linguagem R [4].

Tarefas a Realizar e Cronologia

T1 - Estudo da framework Spark e da interface SparkR (1 mês);

T2 – Análise preliminar dos dados e especificação das análises de interesse (0,5 mês);

T3 – Implementação (1 mês);

T4 – Testes e escrita de relatório (1 mês).

Requisitos Técnicos / Académicos

Gostar de programar;

Ter aprovação às disciplinas de programação.

Elementos de Avaliação a Entregar

- Relatório de projecto;

- Código desenvolvido.

Resultados Esperados

- Uma aplicação funcional;

- Um relatório de projeto.

Referências Bibliográficas

[1] http://inep.gov.br/en_US/web/guest/enem

[2] https://spark.apache.org/

[3] https://spark.apache.org/docs/latest/sparkr.html

[4] https://www.dataxone.com/installing-sparkr-windows-rstudio/