Guidelines for setting your MD Project

2021/22

Types of Projects

- Data modeling and/or Technological approaches comparisons
- Case study with data integration from different sources and different data types
- Multidimensional Modeling of a specific theme
- Enriching a Data Model with Semantic Web (LOD)
- Surveys

General Examples

- Padrões e Tendências de Crimes em Chicago
- Processamento e análise de indicadores sociais e ecológicos à escala global
- Survey on geospatial data tools
- Multidimensional Modelling for .....
- Comparison of Relational Databases and Graph Databases
- Criminalidade na cidade de Chicago - Análise e comparação de diferentes abordagens de modelação

General Examples

- Estudo da indústria cinematográfica através da comparação de diferentes modelos e tecnologias para modelação de dados
- Database enrichment using the Semantic Web
- LPG and RDF Triple Stores Comparison
- Análise do Comunica: a Modular SPARQL Query Engine for the Web
- Dataset Query Performance Analysis: MySQL, RDF, Neo4j
- Bases de Dados em Grafos Representação Social VS Espacial

Concrete Examples

- Comparação de Grafos vs Documentos - Taxis de NY
- Cassandra vs BigQuery - Acidentes de viação EUA
- Relacionais vs Grafos - AirBnB
- Relacionais vs Grafos - Eventos meteorológicos
- Multidimensional - YELP
- Cypher vs Gremlin - Rede Ferroviária de Alta Velocidade ES
- Comparação de Grafos vs Relacional - Taxis de NY

Concrete Examples

- Semantic web implementation of Hospital attendance data in UK
- Relacionais vs Grafos - YELP
- Relacionais vs Grafos - Filmes e Series
- Essay on Query Engines for the Web
- Relacionais vs Grafos - Voos EUA
- Relacionais vs Graphos - Bikes Chicago
- Multidimensional modelling: United Kingdom Road Accidents
- Neo4J vs Jena vs PostgreSQL vs Elastic Search - Noticias em PT

Looking for Data?

- Consider the following article "**The 50 Best Public Datasets for Machine Learning**".
  - https://medium.com/datadriveninvestor/the-50-best-public-datasets-for-machine-learning-d80e9f030279
- The most well know source of public datasets, **Kaggle**
- From **google** https://datasetsearch.research.google.com
- **Open data from Lisbon**:
  - http://lisboaaberta.cm-lisboa.pt/index.php/pt/
- **Pordata** - Contemporary Portugal Database

Main Requirements

- The subject has to be relevant in terms of MD syllabus
- Your proposal should be accepted by the teachers
- Elements for you proposal
  - Title
  - Abstract
  - Justification of relevance for MD
  - Data
  - Used Technologies

Write your proposal

- Upload a PDF to the Team shared folder
  - GNN-Subject-Registration-2021.10.dd.PDF
- Ask for feedback
- Iterate until you get an approval

Computador não é poderoso?!

- Amostra para estudos iniciais
  - Com amostrar?
- Índices e outros configurações
- Carregamentos BULK, Sequências
- SSD vs HDD
- Azure

Análises de desempenho?

- Queries
- Samples de diferentes dimensões dos Dataset
- N execuções
- Métricas de desempenho que medem a média
- Gráficos da evolução de desempenho

Ingredientes essenciais

- **Dados**
  - Algum volume
  - Alguma complexidade
    - Inerente aos DS escolhido ou por integração de várias fontes (por exemplo Semantic Web)
- **Queries ou perguntas que gostariam de explorar**
  - Cada “Query” tem que ser caracterizada e justificada
    - Filtragem; Agregação; Cálculo; Join
- **Tecnologias**