Pandemia COVID-19 e o Mundo do Trabalho

Análise Inteligente em Sistemas de Big Data

1 Introdução

O **caso de uso** apresentado visa analisar métricas que investiguem o impacto que o aparecimento do Covid-19 teve nos estabelecimentos profissionais e na vida dos trabalhadores. Pretende-se avaliar como é que a evolução da pandemia e da vacinação afetaram as regras aplicadas a estabelecimentos profissionais e como é que isso influenciou o apoio financeiro a pessoas que perderam o seu trabalho.

2 Metodologias

A pesquisa dos datasets tomou uma abordagem **bottom-up**, isto é, a partir dos datasets chegamos ao caso de uso.
A pesquisa das ferramentas foi dividida em **4 categorias**: junção dos datasets, para o armazenamento dos dados, para o processamento e tratamento do dataset e para visualização dos dados. Nesta pesquisa encontraram-se 8 ferramentas, 2 para cada categoria.

3 Estado de Arte

Concatenação

Package Pandas e **Package Numpy** da linguagem Python. Ambas as ferramentas utilizam métodos de análise e merge que permitem juntar os datasets.

Armazenamento

HDFS do Hadoop e a base de dados NoSQL, **MongoDB.** Ambas as ferramentas possuem funcionalidades para armazenar o dataset concatenado.

Processamento

Apache Spark e **MapReduce** do Hadoop. As ferramentas utilizam métodos que permitem aplicar o ETL ao dataset concatenado.

Visualização

Os software **PowerBI** e **Tableau.** Estas ferramentas possuem diversos gráficos que se enquadram no problema em questão.

4 Dataset

Base - Who Covid

Dataset CSV fornecido pelos docentes que representa os **dados globais associados à pandemia** Covid-19, desde 3 de janeiro de 2020 até 1 de março de 2022 em cerca de 236 países.

Vacinação

Este dataset JSON representa a evolução do **processo de vacinação** desde 22 de janeiro de 2021 até 27 de março de 2022 em 235 países. Cada observação do dataset possui atributos como:

- nº de doses administradas;
- nº de pessoas com pelo menos uma dose;
- nº de pessoas totalmente vacinadas;
- nº de doses de reforço;
- entre outros...

Fecho dos locais de trabalho

Neste dataset CSV encontra-se associada a cada dia (1 de janeiro de 2020 a 21 de março de 2022) e a cada um dos 186 países, as **restrições aplicadas aos locais de trabalho**. Estas medidas são caracterizadas por valores de 0 a 3.

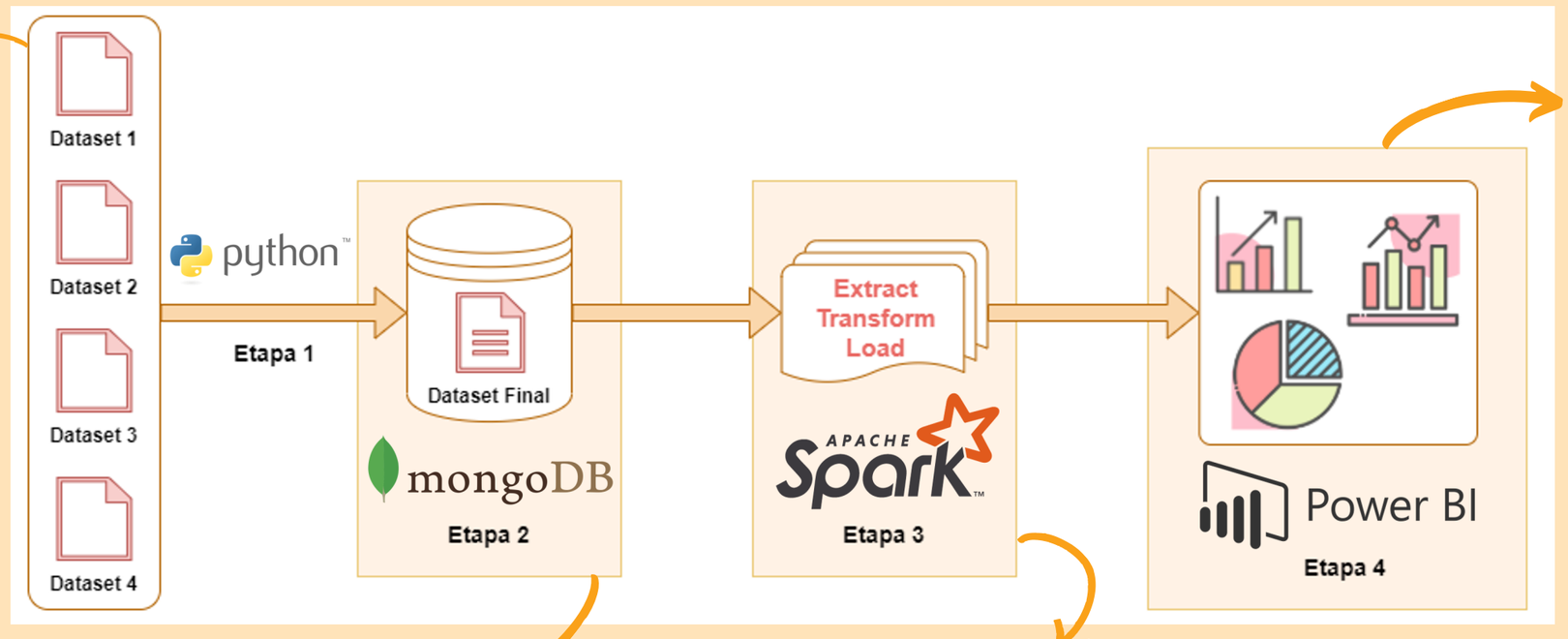
- 0 : inexistência de medidas;
- 1 : recomendação de fecho;
- 2 : fecho do local para uma categoria de trabalhadores;
- 3 : fecho do local exceto se este prestar serviços essências.

Apoio no salário

Neste dataset CSV encontra-se associada a cada dia (1 de janeiro de 2020 a 21 de março de 2022) e a cada um dos 186 país, uma métrica que simboliza **se o governo está a cobrir os salários a pessoas que perderam os seus empregos ou não podem trabalhar**. Esta métrica é caracterizada por valores de 0 a 2.

- 0 : inexistência deste tipo de apoio;
- 1 : governo apoia pessoas com menos de 50% do salário perdido;
- 2 : governo apoia com 50% ou mais do salário perdido.

5 Pipeline



ETAPA 1 - Junção dos datasets:
Nesta fase vamos utilizar a **biblioteca Pandas** pois, há familiaridade com os métodos que utiliza e como pertence à linguagem python são mais fáceis de usar juntamente com esta linguagem. É uma biblioteca que utiliza dataframes permitindo ao utilizador maior facilidade na manipulação destes objetos. Nesta fase vamos concatenar os datasets utilizando o método merge e a biblioteca pycountry para permitir o merge.

ETAPA 2 - Armazenamento do dataset:
Nesta fase utilizamos o **MongoDB**, pois armazena documentos do tipo JSON ou CSV de forma mais robusta e flexível. Além disso, tem uma linguagem estilo query que permite retirar informação do documento original facilmente. Nesta fase vamos armazenar o dataset concatenado para ser usado nas fases posteriores.

ETAPA 3 - Processamento e tratamento dos dados:
Na fase de ETL será utilizada a ferramenta **Apache Spark** pois, oferece aos utilizadores a **API PySpark** que aplica as funcionalidades desta ferramenta em ambiente Python, mantendo desta forma consistência com as bibliotecas de Python que também serão utilizadas. Além disso, implementa métodos iterativos mais complexos de forma mais eficiente. Nesta fase vamos gerar valores sintéticos de forma a preencher os valores a null, mas também vamos retirar observações que possuam muitos desses valores.

ETAPA 4 - Visualização dos dados:
Nesta fase utilizamos **PowerBi** pois, é muito intuitivo e fácil de utilizar e como os dados que serão utilizados para visualização não são obtidos em tempo real, não há necessidade de utilizar uma ferramenta tão complexa como o Tableau. Nesta fase vamos criar gráficos que avaliem a influência do número de mortes e do número de vacinas nas restrições aplicadas aos locais de trabalho e nos apoios do governos aos salários dos trabalhadores.

Trabalho Individual - PG46983 - Ana Luísa Lira Tomé Carneiro