Big Data Trabalho Grupo

Carlos Caravalho, Rui Vieira† and Francisco Franco†

Engenharia Informática,Universidade do Minho, Gualtar, Braga, Portugal.
Grupo 11.

Contributing authors: pg47092@alunos.uminho.pt; pg47635@alunos.uminho.pt; pg47187@alunos.uminho.pt;
†These authors contributed equally to this work.

Abstract

Este trabalho surge na sequência do trabalho individual, em que cada um dos elementos do grupo, fundamentou e estruturou metodologias e o estado de arte resultante da procura e investigação de datasets e ferramentas big-data de forma a conseguir verificar se a pandemia Covid-19 resultou numa menor mobilidade humana a nível mundial e, consequentemente, numa alteração positiva nos níveis de qualidade do ar. Assim cada um dos elementos estruturou o que pensava ser a melhor arquitetura para este projeto. Este trabalho de grupo vem interligar as 3 arquiteturas definidas por os elementos presentes no grupo procurando assim da melhor forma conseguir estruturar e fundamentar uma arquitetura BigData e aplica-la.

Keywords: BigData, Arquitetura, Processamento, Armazenamento, Visualização, Estado de Arte, Metodologias, ferramentas

1 Introdução

Este trabalho parte então da crença do possível impacto da pandemia de COVID-19 nos níveis de poluição do ar ao redor de todo o mundo. Esta crença deve-se ao facto de, durante os dois anos de pandemia ter existido menos mobilidade populacional, subentenda-se mobilidade como todo o tráfego rodoviário recorrente num dia a dia de grandes cidades, assim como a utilização de qualquer outro meio de transporte poluente.

De forma a tentar comprovar esta hipótese, com os datasets recolhidos na fase anterior no trabalho individual, vamos dar algum feedback sobre as ferramentas escolhidas pelo grupo e demonstrar como as utilizamos.

Por fim com os dados obtidos e tratados pelas ferramentas que achamos adequadas vamos procurar responder à nossa hipótese e identificar através da ferramenta de visualização quais (em caso de realmente existirem) as alterações na qualidade do ar em redor do mundo.

Posto isto, será feita uma análise dos resultados relativa às duas fases anteriormente abordadas.

2 Ferramentas

Tendo em conta as ferramentas utilizadas, decidimos enquanto grupo procurar o denominador comum nas ferramentas identificadas na fase individual deste projeto, assim relativamente à ferramenta de processamento decidimos optar pelo MongoDB, que é uma ferramenta que todos já tínhamos utilizado e é uma ferramenta que consideramos acessível e relativamente simples de trabalhar.

Relativamente à ferramenta de processamento apesar de não ser comum a todos os elementos do grupo depois de uma análise mais cuidadosa, optamos pelo Apache Spark, no caso utilizamos o PySpark de forma a conseguir utilizar de melhor forma a ferramenta.

Relativamente á ferramenta de visualização, seguimos mais uma vez pelo denominador comum e identificamos o tableau como ferramenta, uma vez que este premite uma fácil visualização dando acesso a uma rápida amostragem e uma muito estruturada e cuidada criação de gráficos, à parte isso é uma ferramenta bastante preparada para BigData.

3 Datasets e tratamento de dados

3.1 Covid-19

Este dataset foi fornecido pelo pelos docentes da Unidade Curricular, nele constam campos/parâmetros como:

- Date_reported
- Country_code
- Country
- WHO_region
- New_cases
- Cumulative_cases
- New_deaths
- Cumulative_deaths

A partir destes dados podemos deduzir o número de pessoas em isolamento com o número acumulativo de casos numa determinada data, assim, será possível retirar para o nosso caso de estudo alguma informação para relacionar o número de infetados em cada país, com as taxas de mobilidade respectivas, também a cada pais. Apesar de não existirem muitos parâmetros com a possibilidade de limpeza ou remoção, optamos por excluir colunas como o Country e Who_region que consideramos não ser necessário para a obtenção da resposta à nossa hipotese, uma vez que todos os datasets tem uma coluna Country_code não sendo necessária a coluna com o nome do País a que os dados dizem respeito.

Á parte isso alteramos alguns nomes dos parametros, de forma a facilitar a junção dos datasets.

```
covid = spark.read.csv("Covid.csv", header=True)
covid = covid.drop(f.col("WHO_region"))
covid = covid.drop(f.col("Country"))
covid = covid.withColumnRenamed("Date_reported", "Date")
#covid.show()
```

Fig. 1 PySpark Treatment

3.2 Global Mobility during Covid-19

No dataset relativo à mobilidade humana global durante a pandemia, podemos encontrar parâmetros:

- countryregioncode
- country_region
- subregion1
- subregion2
- metro_area

- iso31662_code
- censusfipscode
- place_id
- date
- retailandrecreationpercentchangefrombaseline
- groceryandpharmacypercentchangefrombaseline
- parkspercentchangefrombaseline
- transitstationspercentchangefrom_baseline
- workplacespercentchangefrombaseline
- residentialpercentchangefrombaseline

Dados os parâmetros relativos a este dataset, tal como no anterior enunciado, existia, de facto, alguma limpeza a ser feita a este, desde dados em falta ou repetidos, como algumas colunas, que para o nosso caso de estudo são um pouco irrelevantes.

Assim como podemos verificar na imagem a baixo, eliminamos várias colunas, as subregion porque não achamos necessidade de mostrar os dados por cidade, áparte isso eliminamos algumas colunas cujo os dados eram em sua grande maioria nulos.

Numa prespetiva de tentar relacionar a mobilidade com o número de casos de covid, usamos a relação entre data e o código do país a que os dados são referentes, de forma a tentar perceber se a diferença com o normal pré-covid tem relação com o número de casos de infetados num determinado país.

```
mobility = spark.read.csv("Mobility.csv", header=True)

mobility = mobility.dropDuplicates(['country_region_code', 'date'])

mobility = mobility.drop(f.col("sub_region_1"))
mobility = mobility.drop(f.col("sub_region_2"))
mobility = mobility.drop(f.col("metro_area"))
mobility = mobility.drop(f.col("iso_3166_2_code"))
mobility = mobility.drop(f.col("census_fips_code"))
mobility = mobility.drop(f.col("place_id"))


mobility = mobility.withColumnRenamed("country_region_code", "Country_code")
mobility = mobility.withColumnRenamed("date", "Date")
mobility = mobility.drop(f.col("country_region"))
#mobility.show()
```

Fig. 2 PySpark Treatment

3.3 Airquality

Relativamente ao dataset de qualidade do ar, necessitou de um pouco de mais trabalho, uma vez que este é formado por um conjunto de datasets divididos

por quartil desde 2019 até 2022, assim numa primeira análise tivemos de juntar todos estes datasets num único para posteriormente o melhor trabalhar.

Para esta junção utilizamos o pandas como é demonstrado na figura a baixo.

```
Final19 = pd.concat([Q1_19, Q2_19], axis=0)
Final19 = pd.concat([Final19, Q3_19], axis=0)
Final19 = pd.concat([Final19, Q4_19], axis=0)

Final20 = pd.concat([Q1_20, Q2_20], axis=0)
Final20 = pd.concat([Final20, Q3_20], axis=0)
Final20 = pd.concat([Final20, Q4_20], axis=0)

Final21 = pd.concat([Q1_21, Q2_21], axis=0)
Final21 = pd.concat([Final21, Q3_21], axis=0)
Final21 = pd.concat([Final21, Q4_21], axis=0)

Final1 = pd.concat([Final19, Final20], axis=0)
Final = pd.concat([Final1, Final21], axis=0)
```

Fig. 3 Pandas concat

Os parâmetros deste conjunto de datasets são então:

- Date
- Country
- City
- Specie
- count
- min
- max
- median
- variance

Este dataset exigiu um tratamento um pouco mais cuidado, uma vez que truncamos os dados das cidades de forma a ficarmos unicamente com os dados relativos ao país, uma vez que estes são os dados mais relevantes.

Abaixo mostramos então o tratamento executado em spark para este dataset final.

```
airquality = spark.read.csv("Final.csv", header=True)

airquality = airquality.drop(f.col("_c0"))
airquality = airquality.withColumnRenamed("Country", "Country_code")

airquality = airquality.withColumn("count",f.col("count").cast('double'))
airquality = airquality.withColumn("min",f.col("min").cast('double'))
airquality = airquality.withColumn("max",f.col("max").cast('double'))
airquality = airquality.withColumn("median",f.col("median").cast('double'))
airquality = airquality.withColumn("variance",f.col("variance").cast('double'))

airquality = airquality.groupBy("Date", "Country_code", "Specie").agg(sum("count").alias("Count"),
                                                                      min("min").alias("Min"),
                                                                      max("max").alias("Max"),
                                                                      mean("median").alias("Median"),
                                                                      mean("variance").alias("Variance"))
```

Fig. 4 PySpark Treatment

3.4 Junção dos datasets

Relativamente á junção dos datasets, como fomos referindo procuramos fazer esta junção dos datasets maioritariamente através do código do país e da data em que os dados foram obtidos, desta forma obtemos um dataset final com todos os dados agrupados.

```
mobility = spark.read.csv("Mobility.csv", header=True)

mobility = mobility.dropDuplicates(['country_region_code', 'date'])

mobility = mobility.drop(f.col("sub_region_1"))
mobility = mobility.drop(f.col("sub_region_2"))
mobility = mobility.drop(f.col("metro_area"))
mobility = mobility.drop(f.col("iso_3166_2_code"))
mobility = mobility.drop(f.col("census_fips_code"))
mobility = mobility.drop(f.col("place_id"))


mobility = mobility.withColumnRenamed("country_region_code", "Country_code")
mobility = mobility.withColumnRenamed("date", "Date")
mobility = mobility.drop(f.col("country_region"))
#mobility.show()
```

Fig. 5 PySpark Treatment

4 Metodologia abordada

Relativamente a metodologia abordada, como anteriormente referido, fizemos então o tratamento dos datasets escolhidos com a utilização de PySpark, de seguida passamos o dataframe final para o MongoDB onde temos mantemos todos os nossos dados guardados e por fim conectamos a base de dados do Mongo ao Tableau através da ferramenta MongoDB BI Connector, e a partir dai partimos á análise de resultados que poderemos ver mais a frente neste relatório.

5 Dificuldades Encontradas

Ao longo deste projeto deparamos-nos com uma grande dificuldade, uma vez que a quantidade de dados a tratar é considerável e que algumas operações do Spark no tratamento dos dados eram consideravelmente complexas estas demoravam algum tempo a serem realizadas, este problema refletiu-se uma vez mais no import dos dados do dataframe final para o MongoDb.

Por outro lado, a maior dificuldade que encontramos foi mesmo o facto de todas estas ferramentas serem "desconhecidas" para todos os elementos do grupo nunca tendo propriamente tido interagido com estas frameworks e ferramentas, levando assim o seu tempo na aprendizagem e procura de informação sobre as mesmas.

6 Análise dos Resultados

No final do tratamento no PySpark e do armazenamento com o MongoDB dos dados deste era necessário realizar a visualização dos mesmos para, desta forma, ser possível uma análise mais clara. Desta maneira, como referido anteriormente, este processo foi realizado com o auxílio da ferramenta Tableau ligando esta diretamente com o MongoDB através do MongoDB BI Connector.

Uma vez que a ligação estava concluída, começamos por filtrar os dados à chegada no que toca à análise da qualidade do ar, tendo decidido que o fator mais relevante seria a quantidade de partículas p_1, $p_{2,5}$ e p_{10} presentes no ar.

De seguida começamos por determinar quais as formas de expressar os dados de forma a ser possível realizar uma análise minuciosa e ser possível retirar conclusões. Criamos, com este intuito, 5 formas de expressar diferentes aspetos dos dados.

- Mapa com taxa de ocupação de parques
- Mapa com os novos casos de COVID
- Mapa com a quantidade de partículas presentes no ar
- Gráfico com comparação dos diferentes fatores a nível mundial
- Gráfico com comparação dos diferentes fatores a nível de cada país

Nos 3 primeiros usamos um mapa para haver um contexto geográfico que está presente em todos os datasets iniciais e ainda com a possível seleção de uma data para ser possível visualizar a evolução temporal dos diversos

elementos a analisar. Os últimos dois servem para conseguirmos realizar a análise das relações dos diferentes fatores dos nossos datasets para chegar a resultados mais concretos, um deles ainda é possível visualizar esta análise por país.

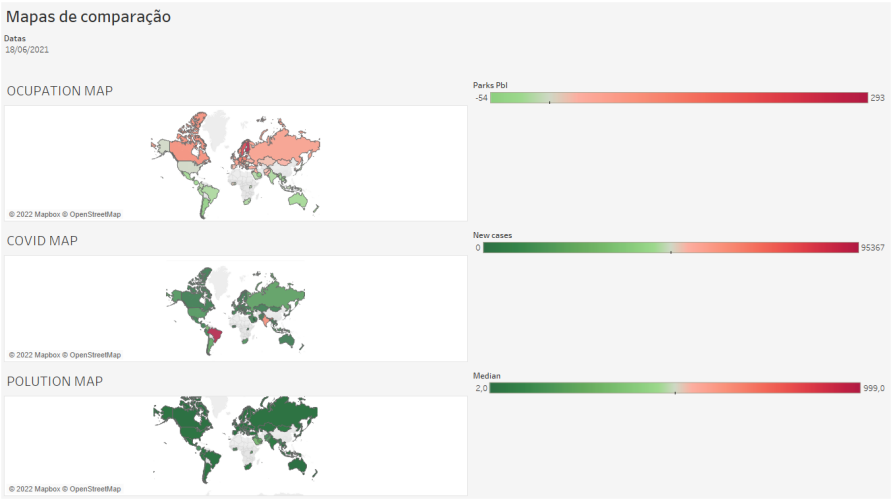

Fig. 6 Global Comparison

Na Figura 6 é possível ver a taxa de ocupação de parques, novos casos de covid e quantidade de particulas no ar no mundo inteiro no dia 18 de junho de 2021. Conseguimos ver que os países pintados de verde no mapa do covid, são os que têm menor número de novos casos e os mesmos apresentam coloração vermelha no mapa de ocupação, que significa que os parques estão mais lotados. Isto reflete-se pois havendo poucos casos o país não se encontra em fase de isolamento obrigatório e teletrabalho.

Em relação ao mapa da poluição, não há alterações visíveis que permitem tirar conclusões. Estávamos a espera de ver uma redução no valores de partículas no ar com a diminuição da taxa de ocupação de parques mas isso não acontece de forma visível.

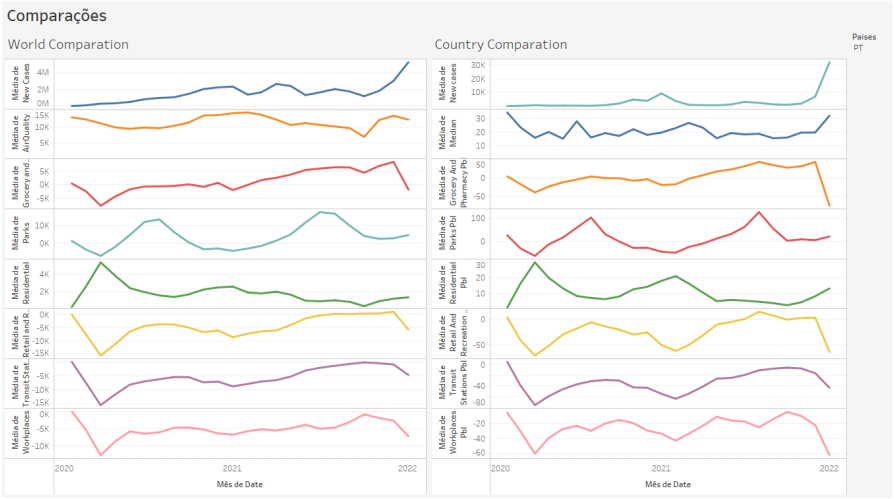

Fig. 7 Global vs Portugal

Na Figura 7 temos os dados relativos a taxa de ocupação de vários setores a nível mundial e de Portugal desde o início da pandemia em 2020 até à atualidade. Os valores de novos casos são em milhões e os de airQuality são em milhares. Os restantes são percentagens de ocupação com base num valor médio. Primeiramente o mais relevante a analisar é o o pico que se nota em março/abril de 2020 tanto no mundo como em Portugal. Esta baixa das taxas de ocupação em todos os setores exceto no residencial, podem ser explicadas pois a OMS declarou o covid-19 como uma pandemia mundial a 11 de março de 2020. Foi nesta altura que todos os países começaram a apresentar vários novos casos de covid levando os países a declarar recolher obrigatório como forma de contenção do vírus. A taxa de recidenciais é inversamente proporcional ás outras taxas, o que faz sentido.

Analisando apenas o gráfico relativo a Portugal, voltamos a detetar uma quebra nas ocupações em janeiro de 2021 e em 2022, altura em que tivemos mais casos diários de covid-19 como mostra a primeira linha do gráfico.

Mais uma vez, a nível de poluição e qualidade do ar, não é possível tirar conclusões acertivas, pois os valores não têm grandes variações, ou as variações não permitem tirar conclusões com o avanço da pandemia tanto a nível mundial como em Portugal.

7 Conclusão

Este trabalho no âmbito da disciplina de BigData surge na tentativa do grupo aplicar de melhor forma uma arquitetura BigData definida pelo grupo, assim procuramos de certa forma seguir a arquitetura que achamos mais adequado, arquitetura esta concebida pelos elementos do grupo na fase individual do Projeto.

Com a realização desta arquitetura, tínhamos então como objetivo verificar se existiu ou não uma alteração significativa na qualidade do ar, face à menor mobilidade durante o período de pandemia.

Para concluir, estamos satisfeitos com o trabalho final. As nossas expectativas iniciais eram então observar alguma alteração na qualidade do ar com a redução das taxas de ocupação pois à partida há menos emissões de gases poluentes dos meios de transporte e das fábricas que encerravam por causa do isolamento. Os gráficos de Portugal mostram bastante bem a evolução da pandemia que vivemos o que nos leva a crer que tanto os datasets tinham dados verídicos como o tratamento e merge destes foi bem afetuado. Em relação ás ferramentas, escolhemos, após alguma discussão, as que foram pesquisadas pelos 3 elementos do grupo no primeiro trabalho da UC de forma a montar a melhor arquitetura para a boa execução deste projeto.

Apesar de tudo isto, e mesmo considerando que conseguimos cumprir o objetivo temos conhecimento de alguns aspetos a melhorar a realização deste projeto, mais incidentes á fase de processamento dos dados.