Trabalho final da disciplina
Processamento de Dados Massivos em Nuvem 2020/2

Durante o semestre foram propostos dois exercícios de análise de dados direcionados, para introduzir a turma ao processo de processamento desse tipo de dados. No trabalho final, cada estudante (ou dupla) deverá elaborar um relatório de análise de dados para uma base de sua escolha. Para facilitar, separamos algumas bases que já estão disponíveis no cluster da disciplina e que são descritas neste documento.

Entregas:

O trabalho será dividido em duas entregas: a primeira, para o dia 15/03/2021, deve ser uma pequena proposta de análise; a segunda, para o dia 19/03/2021, será o relatório final com os resultados obtidos.

Proposta (5 pontos):
No dia 15/03/2021 (até as 23:59) deve ser entregue uma proposta em uma página (PDF, sem formato obrigatório) contendo:

- Nome do/a estudante ou dupla
- Base de dados escolhida
- Descrição do problema que se pretende abordar/perguntas que devem ser respondidas

O tipo de análise esperada deve ser pelo menos do mesmo nível dos exercícios de programação. Certamente, análises mais aprofundadas são encorajadas.

A análise deve ser realizada idealmente usando Spark. Caso se deseje usar outros recursos isso deve ser indicado e justificado na proposta.

Relatório final (restante da nota):
No dia 19/03/2021 (até as 23:59) cada estudante/dupla deve entregar um relatório em PDF com o resultado da análise proposta. Idealmente, todas as perguntas mencionadas na proposta devem ser respondidas, mas alterações (remoção/adição de perguntas) podem ser possíveis caso haja motivos válidos durante o processo. O relatório também deve incluir pelo menos uma seção que descreva os princípios gerais da metodologia utilizada na avaliação. Não é para incluir o código usado, mas sim descrever quais técnicas foram empregadas, algum problema particular que precisou ser tratado, alguma transformação mais específica que foi necessária, etc. Arquivos com o código (Spark) usado para a análises devem ser entregues junto com o relatório, como um arquivo .zip ou .tar.gz.

O relatório deve ser formatado usando o padrão da SBC: https://www.sbc.org.br/documentos-da-sbc/summary/169-templates-para-artigos-e-capitulos-de-livros/878-modelosparapublicaodeartigos (o padrão está disponível no Overleaf: https://www.overleaf.com/latex/templates/sbc-conferences-template/blbxwjwzdngr). Não há um número esperado de páginas, o importante é um conteúdo relevante, uma análise que demonstre o domínio da matéria. Eu diria que 8 páginas pode ser considerado o limite superior, mas ninguém é obrigado a "encher linguiça" para atingir o limite.

Descrição das bases de dados disponíveis

ID 1 - Covid (Twitter)

Fonte: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april e https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april

Extração de tweets entre 29/03/2020 e 30/04/2020 partir das principais hashtags relacionadas: #coronavirus, #coronavirusoutbreak, #covid19, #covid_19, #coronavirusPandemic, #ihavecorona, #StayHomeStaySafe} e #TestTraceIsolate.

A base foi disponibilizada em arquivos CSV, separados por dia, foi processada e contém alguns campos do formato tradicional de um tweet.

Descrição da API: https://dev.twitter.com/overview/api/tweets

Tamanho da base: 4.9 GB
Número de registros: 32 milhões de tweets
Localização no HDFS: /datasets/covid19/

ID 2 - ENEM (Twitter)

Base coletada em JSON utilizando a API do Twitter a partir de busca por termos relacionados como “enem”, “vestibular” e “inep” em perfis públicos. O período de coleta foi em maio de 2014, no entanto, podem conter resultados de antes desse período.

Descrição da API: https://dev.twitter.com/overview/api/tweets

Tamanho da base: 5.7GB
Número de registros: 1.821.897
Localização no HDFS: /datasets/enem/tweets_enem.json

ID 3 - Dengue (Twitter)

Amostra de uma base coletada em JSON utilizando a API do Twitter a partir de busca por termos relacionados como “dengue”, “chikungunya” e “zika” em perfis públicos em março de 2019.

Descrição da API: https://dev.twitter.com/overview/api/tweets

Tamanho da base: 1.2 GB
Número de registros: 200.000
Localização no HDFS: /datasets/dengue/dengue.json

ID 4 - Last.fm

Fonte: http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html

Extração das listas das músicas escutadas por 992 usuários do site Last.fm, um site popular de música, em 2010.

O conjunto de dados contém informações dos 992 usuários (userid-profile.tsv) e seus históricos completos de músicas até maio de 2009 (*userid-timestamp-artid-artname-traid-traname.tsv*). Sobre o arquivo de músicas, cada linha contém as tuplas de informação de cada evento no formato *<id do usuário, timestamp, id do artista, nome do artista, id da música, nome da música>*. Já o arquivo de usuários, as tuplas estão no formato *<id do usuário, sexo, idade, país, data de inscrição>*.

Tamanho da base: 2.4 GB
Número de registros: 19.150.868
Localização no HDFS: /datasets/last_fm/

ID 5 - Homofobia (Facebook)

Base coletada em JSON utilizando a API do Facebook a partir de busca por termos relacionados à homofobia em junho de 2014, no entanto, podem conter resultados de antes desse período.

Descrição da API: https://developers.facebook.com/docs/graph-api/reference/post/

Tamanho da base: 6.6 GB
Número de registros: 2 milhões
Localização no HDFS: /datasets/homofobia/facebook_homofobia_01_2M.json

ID 6 - Conjunto de dados de traços de mobilidade de ônibus no Rio de Janeiro, Brasil

Fonte: http://www.crawdad.org/coppe-ufrj/RioBuses/20180319/

Dados de posição em tempo real informados por ônibus, atualizados a cada minuto, da cidade do Rio de Janeiro, Brasil.

O arquivo é CSV, contendo a data, hora (formato 24h), ID do ônibus, linha do ônibus, latitude, longitude e velocidade de mais de 12.000 ônibus. Os dados coletados são de outubro de 2014, em um período de um mês.

Tamanho da base: 6.3 GB
Número de registros: 115.667.479
Localização no HDFS: /datasets/gps_bus_rio/bus-rio.csv

ID 7 - Conjunto de dados de traços de mobilidade de ônibus em Curitiba, Brasil

Dados de posição em tempo real informados por ônibus, atualizados a cada cinco segundos, da cidade de Curitiba, Brasil.

O arquivo está no formato JSON, contendo os campos: VEIC, código de identificação do ônibus; LAT, latitude; LON, longitude; DTHR, data e tempo; e COD_LINHA, código da linha de ônibus. Os dados foram coletados entre os dias de 04 a 18 de maio de 2017.

Tamanho da base: 6.2 GB
Número de registros: 63.969.501
Localização no HDFS: /datasets/gps_bus_curitiba/bus-gps.json

ID 8 - Qualidade do ar da cidade de Sófia (Bulgária)

Fonte: https://www.kaggle.com/hmavrodiev/sofia-air-quality-dataset

Amostra de dados dos sensores (temperatura, umidade e pressão) de qualidade do ar coletados na cidade de Sófia na Bulgária no período de julho/2017 a dezembro/2018.

A base foi processada e disponibilizada em arquivos CSV, separados por meses, com os campos: sensor_id, lat, lon, timestamp, pressure, temperature, and humidity.

Tamanho da base: 4.6 GB
Número de registros: 61.512.638
Localização no HDFS: /datasets/air_quality/

ID 9 - Bitcoins (Twitter)

Fonte: https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329

Extração de tweets entre 01/01/2016 e 29/03/2019 a partir de posts contendo as palavras Bitcoin ou BTC.

A base foi processada e disponibilizada em arquivos CSV com os campos *user, fullname, tweet-id, timestamp, url, likes, replies, retweets e text*, do formato tradicional de um tweet.

Descrição da API: https://dev.twitter.com/overview/api/tweets

Tamanho da base: 4.0 GB
Número de registros: 16 milhões de tweets
Localização no HDFS: /datasets/bitcoin/

ID 10 - UFMG (Facebook e Instagram)

Coleta de posts públicos a partir do termo ufmg. O período de coleta foi de setembro de 2014, no entanto, podem conter resultados de antes desse período.

Para os posts do Facebook, a base foi coletada em JSON utilizando a API do Facebook. Já para os posts do Instagram, a base foi coletada em JSON utilizando a API do Instagram.

Descrição da API do Facebook: https://developers.facebook.com/docs/graph-api/reference/post/

Tamanho da base: 885 MB (Facebook) e 111 MB (Instagram)
Número de registros: 305.599 (Facebook) e 37.699 (Instagram)
Localização no HDFS (Facebook): /datasets/ufmg/

Outras fontes

Além de outras fontes mais conhecidas como Kaggle, você pode procurar por novas bases em:

- Registry of Open Data on AWS - **https://registry.opendata.aws/**
- No repositório Awesome Public Datasets - **https://github.com/awesomedata/awesome-public-datasets**
- Órgãos do governo, como o INEP e o IBGE