Bootcamp IGTI: Tecnologias de Big Data - Processamento de dados massivos

Desafio

Módulo 3	**Processamento de Dados Massivos**

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Entendimento da manipulação de dados com Spark.
- ✓ Uso do Spark SQL.
- ✓ Otimização de processamentos com Spark.
- ✓ Deploy e desenvolvimento de aplicações.
- ✓ Uso do Spark em conjunto dos serviços do GCP.

Enunciado

Suponha que, como engenheiro de dados, foi apontada uma necessidade de utilizar informações de bases de dados públicas para enriquecer os dados da sua companhia. Deve ser estabelecido então uma aplicação de processamento e escrita de dados no Data Lake da empresa, hoje todo disponibilizado na plataforma da nuvem do Google, o GCP. Os dados a serem utilizados são disponibilizados publicamente no site http://200.152.38.155/CNPJ/, e dizem respeito a base completa de CNPJs já registrados no Brasil. Sua tarefa é juntar as tabelas de estabelecimentos, CNAEs e municípios em uma única tabela analítica, que será consumida por áreas estratégicas da empresa.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1 – Download dos Dados

Os arquivos necessários para a realização da prática estarão disponíveis em um bucket público do GCS, cujas informações seguem abaixo:

- **Região:** us-east1.
- **Bucket:** desafio-final.
- **Link:** https://console.cloud.google.com/storage/browser/desafio-final.

Durante a prática, os alunos devem acessar esses dados diretamente do bucket, utilizando as ferramentas apresentadas em aula.

As tabelas utilizadas são:

- CNAE.
- Municípios.
- Estabelecimentos de CNPJ.

Um dicionário de dados das tabelas pode ser encontrado abaixo:

- Dicionário de Dados.

Os arquivos estão disponibilizados no formato csv.

Caso os haja problemas no acesso ao bucket, será necessário acessar a fonte e realizar o download dos seguintes arquivos:

- Tabela de CNAE: F.K03200$Z. {CODIGO}.CNAECSV.zip
- Tabela de Municípios: F.K03200$Z.{CODIGO}.MUNICCSV.zip
- Tabela de Estabelecimentos*: K3241.K03200Y{N}. {CODIGO}.ESTABELE.zip

Obs.: nos nomes acima, o {CODIGO} indica que pode ser qualquer código referente a atualização da tabela. Essa parte do nome apenas indica a data da última atualização dos dados.

*A tabela de estabelecimentos é bastante grande, dispondo de mais de 46 milhões de registros, e por causa disso ela é particionada em 9 arquivos (até o momento). Assim, é necessário realizar o download de **todas** essas partições, indicadas pelo nome em comum. A parte do nome {N} indica qual das partições o arquivo se refere. Isso também significa que as partições precisam ser **unidas** durante a prática.

Todos os arquivos estão comprimidos e devem ser descomprimidos antes de continuar o restante da prática. No trabalho prático foi indicado uma forma de realizar a descompressão, caso haja dúvidas.

2 – Criação da conta no GCP e disponibilização dos serviços

Os passos para a criação da conta no GCP e a habilitação dos serviços foram apresentados ao longo de todo o capítulo 8. Os alunos devem assistir as aulas e tomar a apostila como referência.

Nessa prática devem ser usados o Google Cloud Storage (GCS) e o Dataproc. Lembre-se de habilitar corretamente os componentes do cluster para que seja acessada a interface do Jupyter.

3 – Tarefas Específicas

Durante a prática, o aluno deve:

- Criar um bucket no GCS e disponibilizar um cluster no Dataproc. **Atenção:** é recomendado utilizar um cluster com mais nós (máquinas), uma vez que o volume de dados é bastante intenso. Abaixo o comando para gerar o cluster recomendado:

  ```
  gcloud dataproc clusters create cluster-teste
  --enable-component-gateway
  --region ${REGION} --zone ${REGION}-c
  --master-machine-type n1-standard-4
  ```

```
--master-boot-disk-size 500
--num-workers 5
--worker-machine-type n1-standard-4
--worker-boot-disk-size 500
--image-version 2.0-debian10
--optional-components JUPYTER
--project ${PROJECT_ID}
```

Segundo a própria documentação do Dataproc, esse cluster geraria um custo de US$ 0,48 caso ficasse ligado 2 horas. O custo é calculado da seguinte forma:

Nº de vCPUs * horas * $0,01

Apesar de o limite gratuito ser mais do que suficiente para que a prática seja realizada, esteja ciente de que cada hora gasta implica em um custo. Use o cluster com cuidado e lembre-se de **excluí-lo** ao final de cada sessão de uso.

- Ler as tabelas disponibilizadas (**Dica:** utilize a opção escape = "\"").
- Construir o schema e ajustar os tipos (**Atenção:** não é porque o dado deve ser de um determinado tipo que uma simples operação de “cast” irá convertê-lo corretamente. Observe a forma como os dados estão dispostos).
- Escrever as tabelas **individuais** em um bucket do GCS, em uma pasta chamada “trusted”.
- Realizar operações subjetivas nas tabelas da pasta trusted, de forma a tornar as tabelas mais interessantes e confiáveis para o usuário final. Utilize os conhecimentos aprendidos ao longo do curso!
- Juntar as tabelas, consolidando as informações em um único DataFrame.
- Escrever os dados em um bucket do GCS, chamado “refined”.
- Após isso, responder as questões do desafio. As questões são sobre o **processo de manipulação** e sobre o **DataFrame final**.

A nomenclatura das pastas utilizadas nessa prática é a mais comum em estruturas de Data Lake reais. Para isso, deve-se considerar que a camada “raw” é o bucket público disponibilizado. Caso o aluno tenha realizado o download dos dados direto da fonte, é ele deve salvá-los na pasta “raw” do bucket e lê-los de lá.

4 – Deploy da aplicação

Por se tratar de uma prática em que é necessário realizar várias operações exploratórias, é natural utilizar o Jupyter Notebook para responder as questões. No entanto, fica o desafio opcional para o aluno de executar o processo final utilizando um script Python e o recurso de lançamento de aplicações Spark do Dataproc, visto em aula. Essa parte não será avaliada diretamente, mas é um passo importante para o entendimento e consolidação dos conhecimentos desenvolvidos.