Bootcamp: Engenharia de Dados

Desafio Prático

Módulo 5: Desafio Final

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Bootcamp:

1. Pipelines de Dados.
2. Containers.
3. Bases de Dados SQL e NoSQL.
4. Conexão a APIs.
5. ETL.
6. Data Lake.
7. Processamento de dados distribuído.

Enunciado

Você foi contratado pela empresa (fictícia) *#VamosJuntos - Desenvolvimento Social* para desenvolver o seu primeiro projeto de Dados. Em seu trabalho investigativo preliminar, você já identificou que as principais fontes de dados necessárias são:

- Uma Database MongoDB disponível na nuvem para consulta.
- Uma API do IBGE (https://servicodados.ibge.gov.br/api/docs/localidades) para extração de informação de regiões, mesorregiões e microrregiões no Brasil.

Você deve, portanto, construir um pipeline de dados que faça a extração dos dados no MongoDB e na API do IBGE e deposite no Data Lake da empresa. Após a ingestão dos dados no Data Lake, você deve disponibilizar o dado tratado e filtrado apenas para o público de interesse da empresa em um DW. Com os dados no DW, você vai realizar algumas consultas e extrair resultados importantes para a *#VamosJuntos.*

Atividades

Você deverá desempenhar as seguintes atividades:

1. Subir o Airflow localmente em uma estrutura de containers, usando docker-compose para utilização mais robusta (https://github.com/neylsoncrepalde/docker-airflow);
2. Criar uma conta *free tier* na AWS para realização das atividades;
3. Criar um bucket no serviço S3 com o nome igti_bootcamp_ed_2021_<numero_da_sua_conta>;
4. Criar uma instância RDS de banco de dados relacional de sua escolha (pode criar a instância de DEV de 1CPU e 1GB de RAM, pois ela faz parte do free tier);
5. Construir um pipeline que faz a captura de dados do MongoDB e da API do IBGE e deposita no S3;
6. O pipeline também deve fazer a ingestão na base de dados SQL que estará servindo como DW;
7. Para persistir os dados no DW, você deve ingerir apenas os dados referentes ao público-alvo da *#VamosJuntos*, a saber, mulheres de 20 a 40 anos;

8. Conectar seu cliente favorito no DW e realizar consultas para responder às perguntas do desafio.

Informações relevantes:

O cluster MongoDB foi disponibilizado pelo professor para consulta de todos os alunos participantes da atividade. Trata-se de um cluster pequeno apenas para testes. Desse modo, faça apenas as requisições necessárias. Não queremos correr o risco de ter indisponibilidade no serviço.

As informações necessárias para conectar no MongoDB:

- host: unicluster.ixhvw.mongodb.net
- database: ibge
- collection: pnadc20203
- username: estudante_igti
- password: SRwkJTDz2nA28ME9

Para conectar em seu ambiente AWS, você vai precisar de duas chaves disponibilizadas no seu usuário no serviço IAM: *access_key_id* e *secret_access_key*. Você pode criar suas chaves no serviço IAM acessando a aba usuários e depois a aba credenciais de segurança. Para conectar direto do seu código python, utilize a biblioteca boto3, o SDK oficial da AWS. Informações sobre instalação e utilização com o S3, aqui: https://boto3.amazonaws.com/v1/documentation/api/latest/guide/s3-uploading-files.html

Docker compose

Para criar um ambiente docker já preparado para suas atividades, clone o repositório https://github.com/neylsoncrepalde/docker-airflow. No terminal, digite:

git clone https://github.com/neylsoncrepalde/docker-airflow.git

Em seguida, entre na pasta do repositório baixado e edite o arquivo *docker-compose-CeleryExecutor.yml*. Todas as linhas que tiverem o seguinte comando:

image: neylsoncrepalde/airflow-docker:latest

devem ser substituídas por esta linha abaixo:

image: neylsoncrepalde/airflow-docker:2.0.0-pymongo

que já possui as dependências necessárias para a realização da atividade.

Depois disso, no terminal, dentro da pasta do repositório (docker-airflow), execute:

docker-compose -f docker-compose-CeleryExecutor.yml up -d

Divirta-se!