Bootcamp: Engenheiro(a) de Dados Cloud

Desafio

Módulo 1	Fundamentos em Arquitetura de Dados e Soluções em Nuvem

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Arquiteturas de Dados em Nuvem;
- ✓ Implementação de Data Lake em solução Cloud de Storage;
- ✓ Implementação de Processamento de Big Data;
- ✓ Esteiras de Deploy, utilizando o Github;
- ✓ IaC com Terraform.

Enunciado

Você é Engenheiro(a) de Dados de uma. A Startup está expandindo seu negócio para outras áreas do Brasil. A principal fonte de dados para entender o cenário econômico e de trabalho atual é a RAIS, uma base de dados desafiadora!

Você deve fazer a ingestão da RAIS 2020 em uma estrutura de Data Lake na AWS (ou em outro provedor de sua escolha). Depois disso, você deve utilizar alguma tecnologia de Big Data para converter os dados para o formato *parquet*. Em seguida, disponibilize os dados para consulta no AWS Athena (ou outra engine de Data Lake de outra nuvem ou no BigQuery, no caso do Google Cloud) e faça uma consulta para demonstrar a disponibilidade dos dados. Por fim, utilize a ferramenta de Big Data ou a engine de Data Lake para realizar investigações nos dados e responder às perguntas do desafio.

Atenção! Toda a infraestrutura em nuvem deve ser implantada utilizando o Terraform (ou outra solução de IaC de sua escolha) e esteiras de deploy no Github (ou Gitlab, ou

Bitbucket, ou outro de sua escolha). O dado que vamos trabalhar no desafio é grande. Evite fazer consultas desnecessárias.

DIVIRTA-SE!

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Realizar a ingestão dos dados de VÍNCULOS PÚBLICOS da RAIS 2020 no AWS S3 ou outro storage de nuvem de sua escolha. Dados disponíveis em: http://pdet.mte.gov.br/microdados-rais-e-caged (não vamos trabalhat com dados de ESTABELECIMENTOS). O método de ingestão é livre. Os dados devem ser ingeridos na zona *raw* ou zona *crua* ou zona *bronze* do seu Data Lake.

2. Tratar o dataset da RAIS 2020 e seguir os seguintes passos (na dúvida, consulte o código em: https://github.com/neylsoncrepalde/igti_edc_mod1_desafio_final_rais/blob/main/etl/emr-rais.ipynb e lembre-se de que ele foi pensado para rodar em EMR):

    a. Modifique os nomes das colunas, troque espaços por "_", retire acentos e coloque todas as letras minúsculas;

    b. Construa a coluna "uf" com o seguinte comando: *rais = rais.withColumn("uf", f.col("municipio").cast('string').substr(1,2).cast('int'))*

    c. Modifique as colunas de remuneração para que sejam do tipo *double*.

3. Transformar os dados no formato parquet e escrevê-los na zona *staging* ou zona *silver* do seu Data Lake.

4. Fazer a integração com alguma engine de Data Lake. No caso da AWS, você deve:

a. Configurar um Crawler para a pasta onde os arquivos na staging estão depositados;

b. Validar a disponibilização no Athena.

5. Caso deseje utilizar o Google, disponibilize os dados para consulta usando o Big Query. Caso utilize outra nuvem, a escolha da engine de Data Lake é livre.

6. Use a ferramenta de Big Data ou a engine de Data Lake (ou o BigQuery, se escolher trabalhar com Google Cloud) para investigar os dados e responder às perguntas do desafio.

7. Quando o desenho da arquitetura estiver pronto, crie um repositório no Github (ou Gitlab, ou Bitbucket, ou outro de sua escolha) e coloque o código IaC para a implantação da infraestrutura. **Nenhum recurso deve ser implantado manualmente.**

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: