Bootcamp: Engenheiro(a) de Dados Cloud

Desafio Final

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Kubernetes.
- ✓ Pipelines de Dados.
- ✓ Orquestração.
- ✓ Processamento de Big Data.

Enunciado

Você é Engenheiro(a) de Dados na Edu++ (pronuncia-se Edu “mais mais”) uma empresa de estratégia educacional voltada para o ensino médio no Brasil. A Edu++ está preparando o report mais completo sobre o ensino médio para sua carteira de clientes. A fonte de dados mais atualizada para entender esse cenário hoje é o ENEM 2020. Para isso, será necessário implementar um pipeline completo de ingestão, tratamento e disponibilização dos dados do ENEM 2020 para consulta dos clientes e demais analistas de negócio.

Você deve fazer a ingestão dos microdados do ENEM 2020 em uma estrutura de Data Lake na AWS (ou em outro provedor de sua escolha). Depois disso, você deve utilizar o Spark Operator para, dentro do Kubernetes, converter os dados para o formato *parquet* e escrever os resultados em uma outra camada do Data Lake. 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. Se desejar tornar o seu desafio ainda mais bacana, monte uma visão utilizando alguma ferramenta de BI (Superset, Metabase etc.) com as respostas para as perguntas.

Atenção! Todo o pipeline desde a ingestão de dados, processamento e disponibilização para consulta devem ser realizados no Kubernetes. Fique à vontade para escolher a melhor forma de implementar os processos (manifestos, argo CD, Airflow etc.).

DIVIRTA-SE!

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Criar um cluster Kubernetes para a realização das atividades (local ou baseado em nuvem. Recomendamos utilizar um cluster baseado em nuvem para comportar o volume de dados trabalhado).

2. Realizar a instalação e configuração do Spark Operator conforme instruções de aulas.

3. Realizar a instalação e configuração de outas ferramentas que se deseje utilizar (Airflow, Argo CD etc.).

4. Realizar a ingestão dos dados do ENEM 2020 no AWS S3 ou outro *storage* de nuvem de sua escolha. Dados disponíveis em <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. **Os dados devem ser ingeridos de maneira automatizada** na zona *raw* ou zona *crua* ou zona *bronze* do seu Data Lake.

5. Utilizar o SparkOperator no Kubernetes para transformar os dados no formato parquet e escrevê-los na zona *staging* ou zona *silver* do seu data lake.

6. 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.

7. 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.

8. 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.

9. Se desejar, utilize alguma ferramenta de BI (também implantada no Kubernetes) para responder de maneira visual as perguntas do desafio.

10. Quando o desenho da arquitetura estiver pronto, crie um repositório no Github (ou Gitlab, ou Bitbucket, ou outro de sua escolha) e coloque os códigos de processos Python e implantação da estrutura Kubernetes.