Parte 3 - Projeto para a vaga de Engenheira de Dados

Pergunta 1

Devido à necessidade de disponibilização dos dados de clientes para prestadores de serviço, pode-se optar em remover os campos sensíveis. Esta solução, entretanto, poderia influenciar na análise desses dados, visto que as mesmas podem conter informações importantes para o resultado final do modelo. Outra opção seria a anonimização dessas informações. Ou seja, criar uma tabela para converter os dados sensíveis em valores que o parceiro não possa identificar a quem se refere àquele dado.

Pergunta 2

O projeto pode ser dividido em 2 etapas: (1) Infraestrutura de Dados e (2) Produto. Na primeira fase, será utilizado o Apache Sqoop, uma ferramenta para transferir informações de um banco de dados relacional para o ambiente Hadoop. Após a extração, será utilizado o Sistema de Arquivos Distribuídos do Hadoop (HDFS), que armazena dados de forma otimizada. Com isto, esses dados serão inseridos diretamente na ferramenta de busca, Elasticsearch, e estarão disponíveis em formato JSON, podendo ser consumidos de diferentes formas. A figura 1, ilustra o funcionamento da arquitetura proposta.

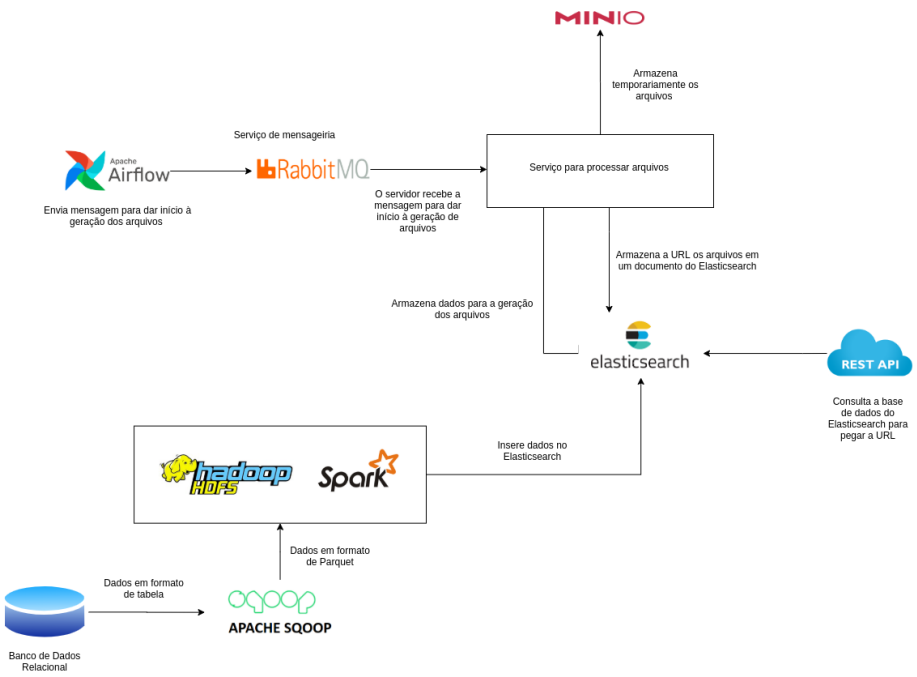

Figura 1: Arquitetura geral

Logo, após a disponibilização dos dados em um formato mais compacto e de fácil manipulação, utiliza-se dois serviços principais para a geração dos arquivos. O primeiro é a Rest API, que disponibilizará os dados para o usuário. Espera-se que nesta aplicação seja implementado inicialmente os seguintes *endpoints*:

- baseurl/api/v1/users/login
- baseurl/api/v1/data/downloadlastfile

baseurl/api/v1/data/login

A necessidade de criação desse método visa a segurança da aplicação e evitar que usuários sem autentição tenham acesso aos dados de arquivos. Neste momento, foi apresentado uma arquitetura inicial. Logo, para gerar robustez ao serviço, será necessário implementar uma sessão para controle de acesso com *endpoints* para cadastro, consulta, atualização de informações do usuário.

- Método: POST;
- Descrição: O usuário terá que informar um seu nome de usuário e senha para ter acesso aos dados;
- Retorno: Um *json* contendo o campo accesstoken.

baseurl/api/v1/data/searchdownloadfile

- Método: GET;
- Descrição: O *endpoint* retornará um link para o download do arquivo gerado na semana corrente.
- Retorno: Um *json* contendo o campo urlfile.

O segundo serviço que faz parte dessa arquitetura é o servidor de processamento de arquivos. Optou-se por esta abordagem para evitar um alto consumo da API. Neste ponto, a mesma tem como objetivo servir de interface entre a geração de arquivos e a disponibilização dos mesmos.

O Airflow possui um *job* para inicializar o processamento dos arquivos a cada uma semana. Este envia uma mensagem para o servidor através de uma fila utilizando o RabbitMQ. Este recebe a mensagem e inicia os seguintes passos:

- Realiza uma consulta dos dados no elasticsearch, buscando pelos documentos inseridos durante a semana.
- Com o resultado dessa consulta, utiliza a biblioteca *dash* do python para a conversão desses dados em csv. Após isso, utiliza-se outros métodos para a compressão desse arquivo.
- Armazena o arquivo comprimido em um serviço chamado MinIO, servidor de armazenamento em nuvem compatível com a Amazon S3.

- O MinIO possui uma biblioteca que disponibiliza formas para ter acesso ao arquivo armazenado. Ele já cria uma url para download do arquivo.
- Utiliza-se a url criada pelo MinIO e armazena a mesma em um documento do elasticsearch com o campo *urlfile*.

Pergunta 4

Utilizaria as seguintes linguagens de programação:

- Scala: Para processamento dos arquivos no ecossistema Hadoop.
- Python: Fácil manipulação durante as análises de dados e possui muitas bibliotecas, bem como, documentação acessível.

Tecnologias para o processamento de arquivos:

- Docker: Configuração do ambiente.
- Apache Sqoop: Ferramenta para transferir informações de um banco de dados relacional para o ambiente Hadoop.
- Hadoop: O ecossistema Hadoop é completo e possui diversas ferramentas para a manipulação dos dados. Estes seriam trabalhados no Sistema de Arquivos Distribuídos do Hadoop (HDFS), que armazena dados de forma otimizada
- Elasticsearch: Ferramenta de busca e armazenamento de dados. Utilizado para otimização de consultas tendo em vista sua robustez.
- Airflow: Controle do pipeline de dados. Seria usado para controlar a ingestão de dados no data lake

Tecnologias para o deploy:

- Docker: Configuração do ambiente.
- Rancher: Gerenciador de infraestrutura Docker. Baseado no Kubernetes, ajudará na orquestração dos containers de cada aplicação que faz parte da arquitetura de dados.
- Digital Ocean: Fornece espaço para adicionar a arquitetura em produção e possui valor abaixo de outras plataformas.