

Seu próprio Google BigQuery

Vamos implementar um sistema para manipulação de dados conhecido como ELT (Extract, Load, Transform). Essa arquitetura permite que sistemas trabalhem em conjunto para buscar, armazenar e transformar dados.

Existem várias formas de implementar um sistema ELT, desde softwares que executam as etapas de extração e armazenamento simultaneamente até sistemas que cuidam de todo o processo em um único lugar (como o Google BigQuery). Neste projeto, vamos utilizar softwares específicos para cada etapa, ainda que eles possam realizar outras funções além das principais.

Objetivo

O objetivo deste projeto é proporcionar uma experiência prática no desenvolvimento de um sistema ELT, utilizando ferramentas que fazem parte do dia a dia do profissional de dados. Você aprenderá a:

- Escolha uma API pública ([sugestão de API](#))([outras sugestões](#))
- Criar um Banco de Dados para armazenar os dados consumidos da API
- Manipular dados (extraíndo, armazenando e transformando).
- Configurar infraestrutura com Docker.
- Integrar sistemas.
- Conhecer sistemas de coleta de dados como o Airbyte.
- Armazenar dados para fins analíticos usando o ClickHouse.
- Criar visualizações de dados com o Metabase.
- Desenvolver habilidades em análise de dados e automação de processos.

Instruções

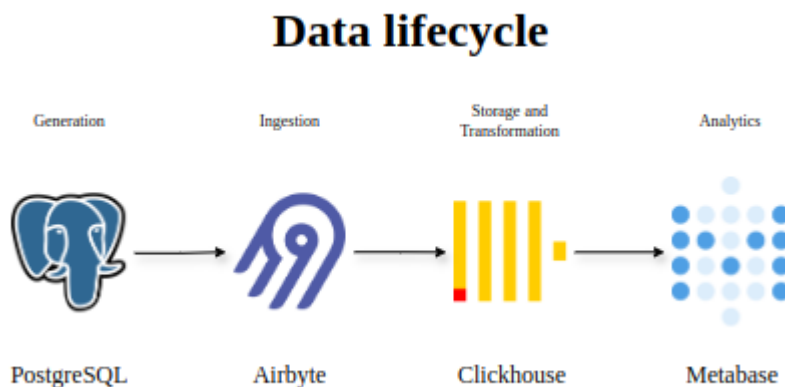
Para construir nosso ELT, selecionamos algumas ferramentas conhecidas no ecossistema de dados: Airbyte, ClickHouse e Metabase. Com essas ferramentas configuradas, montaremos nosso sistema de ELT.

O que faremos:

1. Subiremos uma infraestrutura com 3 softwares: Airbyte, ClickHouse e Metabase
 - Airbyte: Coletará os dados do PostgreSQL que foi abastecido com os dados da API escolhida
 - ClickHouse: Armazenará os dados.
 - Metabase: Criará visualizações dos dados armazenados, permitindo que você possa analisá-los.
2. A infraestrutura será configurada usando Docker localmente.
 - Docker Compose: Utilizaremos Docker Compose para a nossa configuração. É ideal verificar imagens já prontas e utilizá-las.
3. Criaremos views no ClickHouse e análises no Metabase:
 - No ClickHouse, após os dados serem inseridos, vamos criar algumas [views](#):
 - Ao menos 4 views de acordo com os dados escolhidos, por exemplo se fossemos analisar um jogo (crie as views de acordo com os dados escolhidos):
 1. Jogadores com maiores pontuações.
 2. Histórico de pontuações ao longo do tempo.
 3. Conquistas por jogador.
 4. Pontuações médias por jogador.
 - No Metabase, para cada view, vamos criar um [dashboard](#) com as visualizações (dashboards criados de acordo com o exemplo acima):
 - Gráfico de barras: Jogadores com as maiores pontuações em ordem decrescente.
 - Gráfico de dispersão: Pontuações em diferentes datas.

- Gráfico de barras: Quantidade de conquistas por jogador.
- Gráfico de barras: Pontuação média de cada jogador em ordem decrescente.
- É interessante escolher APIs que tenham uma boa quantidade de dados. Quanto mais, melhor!

Abaixo, um diagrama que explicita a infraestrutura de ELT:



Algumas recomendações

- Verifique a documentação do Docker e Docker Compose para entender como utilizá-los.
- Utilize imagens prontas dos softwares que vamos usar. Você não precisa criar a imagem você mesmo, ainda que seja uma experiência de aprendizado interessante.
- Considere usar um software como o [DBeaver](#) para interagir com o ClickHouse de maneira mais fácil, sem precisar usar a CLI.

Submissão

Ao finalizar o projeto, envie os arquivos do Docker neste repositório.