Documentação Técnica – Projeto Final DataGirls

Por Marcela Andrade

# 1. Arquitetura do Pipeline

O pipeline do projeto foi construído utilizando Apache Airflow (v2.7.2) para orquestração e automação das tarefas, executando em contêineres Docker. O ambiente é composto por três serviços principais: Webserver, Scheduler e banco de dados Postgres.  
  
O processamento de dados segue o modelo ETL (Extract, Transform, Load):  
- Extração dos dados a partir da API do Kaggle.  
- Transformação com pandas, salvando os dados tratados no formato Parquet.  
- Carga dos dados transformados para um bucket no Google Cloud Storage (GCS).  
  
O pipeline foi desenhado para ser compatível com posterior ingestão no BigQuery e criação de visualizações no Power BI ou Looker Studio.

# 2. Execução do Pipeline

A execução é feita localmente através do comando `docker compose up`, que inicia os serviços definidos no arquivo docker-compose.  
Após isso, a interface do Airflow pode ser acessada em http://localhost:8080, onde a DAG `etl\_datagirlspfinal` pode ser ativada.  
A DAG está configurada com `schedule\_interval='@daily'`, permitindo a execução automática diária do pipeline, ou manual conforme desejado.

# 3. Transformações Realizadas

Durante o processo de transformação dos dados, as seguintes ações foram aplicadas:  
  
- Remoção das colunas `EmployeeCount`, `Over18`, `StandardHours` e `EmployeeNumber` por conterem valores constantes ou sem relevância analítica.  
- Eliminação de duplicatas e valores nulos para garantir integridade.  
- Padronização dos nomes das colunas para snake\_case, com remoção de espaços e caracteres especiais.  
- Conversão de todas as colunas textuais para letras minúsculas, facilitando análises posteriores.  
  
Essas escolhas garantem que os dados estejam limpos, organizados e prontos para análise.

# 4. Integração com GCP e Visualização

Após a transformação, o arquivo `dados\_transformados.parquet` é enviado para o bucket GCS nomeado `etl\_datagirlspfinal`.  
O próximo passo do projeto foi realizar a carga destes dados no BigQuery, onde forão organizados em tabelas e disponibilizados para análise.  
O dashboard foi desenvolvido utilizando Looker Studio consumindo os dados diretamente do BigQuery.

https://lookerstudio.google.com/reporting/4c51f089-1699-4ecb-bb05-adedb49ba6d5

# 5. Perguntas Norteadoras de Negócio

1. Como a empresa pode monitorar a rotatividade de funcionários semanalmente?

Através do agendamento do pipeline com frequência semanal ou diária, a empresa pode manter os dados atualizados no BigQuery. Assim, gráficos no dashboard podem refletir tendências de rotatividade quase em tempo real.

2. Quais informações devem ser atualizadas em tempo real ou periodicamente?

Informações como desligamentos (`Attrition`), promoções, mudanças de cargo e dados salariais devem ser atualizadas periodicamente para garantir a acurácia das análises de RH.

3. Como garantir que os dados estejam prontos para análises de forma confiável?

Com um processo ETL automatizado via Airflow, o pipeline garante limpeza, padronização e consistência dos dados, reduzindo erros e promovendo reprodutibilidade dos resultados.

4. É possível criar um modelo incremental com essa base?

Sim, desde que a extração futura inclua apenas registros modificados ou novos. Isso pode ser implementado com filtros baseados em datas, e adaptado no processo de transformação.