
🎓 **Projeto Final da Trilha de Engenharia de Dados:**

ETL Automatizado de dados de Funcionários

👋 **Introdução e Boas-Vindas**

Parabéns, participante do **Bootcamp [RE]Start – Trilha de Engenharia de Dados**! Você já percorreu uma jornada incrível – da leitura e manipulação de dados até a construção e monitoramento de pipelines completos. Agora chegou o momento de aplicar todo esse aprendizado em um **projeto** prático, completo e alinhado às demandas do mercado.

Neste desafio, você irá conduzir todas as etapas de um projeto real de engenharia de dados com base no dataset **"IBM HR Analytics Attrition & Performance"**, disponível no Kaggle (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data). O objetivo é construir um pipeline de **ETL automatizado** usando um conjunto de dados relacionado ao **turnover** (*attrition*) de funcionários. A ideia é preparar e transformar os dados para torná-los prontos para análises, modelagem ou consumo por ferramentas de BI que ajudem a empresa fictícia **Data Girls S.A** a tomar decisões estratégicas de RH.

Vamos juntas transformar dados em impacto! 💜

⚙️ **Etapas Orientadoras do Projeto**

1. Extração de Dados

- Importar os dados do CSV (local ou Kaggle API - https://www.kaggle.com/docs/api#getting-started-installation-&-authentication).

2. Transformação de Dados

- Pode utilizar SQL, python ou pyspark.

3. Armazenamento dos Dados em Nuvem

- Salvar os dados processados em formato .parquet, .csv ou em banco de dados relacional.
- Estruturar o pipeline com organização em pastas ou banco com tabelas separadas por domínio.
- DICAS: AWS S3 ou RDS, Azure Blob Storage ou SQL Database, Google Cloud Platform, **Databricks.**
- DICAS: A nuvem **sugerida** para ser utilizada nesse projeto **é AWS**, mas **pode utilizar outra caso prefira como Azure ou GCP**: Através do console da AWS, crie (caso ainda não tenha) um par de **Access Key ID** e a **Secret Access Key** para o seu usuário (IAM —> Security credentials —> Create access key), guarde em segurança, elas vão ser usadas como variáveis de ambiente na DAG.

4. Automação do Pipeline

- Criar scripts ou notebooks executáveis com agendamento simulado:
    - Pode ser feito com **Airflow**
- Incluir logging básico e validações durante o processo.
- DICA: Você pode utilizar o mesma estrutura de projeto de airflow que utilizaram em aula, pode fazer um fork do repositório e criar uma nova dag na pasta dags/, para o projeto completo, sua dag vai conter 3 tasks: 1 para extrair os dados da API, 1 para o processamento com pyspark e 1 para fazer o upload do arquivo final no S3

5. Documentação Técnica

- Explicar a arquitetura do pipeline e como executá-lo.
- Justificar as transformações e escolhas feitas.

🎯 **Perguntas Norteadoras de Negócio**

1. Como a empresa pode monitorar a rotatividade de funcionários semanalmente?
2. Quais informações devem ser atualizadas em tempo real ou periodicamente?
3. Como garantir que os dados estejam prontos para análises de forma confiável?
4. É possível criar um modelo incremental com essa base?

🧠 **Opcional – Bônus**

Integração com um Painel no Power BI

- Conectar seu banco ou arquivos tratados com uma visualização (exportar dashboard);
- Criar um **dashboard simples com métricas como**:
    - Total de funcionários ativos;
    - Percentual de rotatividade;
    - Engajamento por área.

📌 **Avaliação do Projeto**

Critério	Peso	Detalhes
Pipeline ETL completo	25%	Extração, transformação e carga coerente e funcional

Critério	Peso	Detalhes
Automação	25%	Execução agendada e simulada
Qualidade dos dados	25%	Dados limpos, coerentes, estruturados
Documentação técnica	10%	Explicação clara das etapas e decisões tomadas, incluindo perguntas norteadoras respondidas
Organização do repositório	10%	Pastas, nomes e arquivos bem estruturados
Criatividade	5%	Criatividade no desenvolvimento da solução
Bônus	+10%	Desenvolveu o opcional e incluiu na documentação
TOTAL	**100% (110%)**	**NOTA 10 (podendo alcançar 11)**

🗓️ Prazo e Entrega

- **Data limite**: 10/08/2025, às 23h59
- **Entregáveis**:
    - Código/Script do pipeline;
    - Dados tratados;
    - Documentação em PDF ou README;
    - Colocar todos os entregáveis no GitHub e submeter o link no formulário de entrega: https://forms.gle/Ln4CBzFUM9jfLybQA

💡 Dicas Finais

- Pode usar prints, imagens, vídeos para ajudar a explicar o pipeline ou mostrar o agendamento rodando, lembrando de coloca-los na documentação;
- Organize seu código e use comentários para pontos importantes;
- Simule agendamento, mesmo que seja em ambiente local;
- Confira os outros projetos presentes na página do dataset no Kaggle.
- Troque ideias com as outras participantes no Discord.

Estamos muito felizes com sua evolução até aqui e animadas para ver sua entrega final! Essa é sua chance de brilhar com as habilidades de uma verdadeira engenheira de dados.

Mãos à obra e boa sorte!

Com carinho,

Equipe [RE]Start & Mentoras Data Girls 💜