

Guia de Setup - Como Replicar este Projeto



O que você vai precisar

- Conta no Databricks
- Cluster Spark ativo
- Os dados já estão disponíveis no DBFS do workspace!



Dados Disponíveis

Com base na estrutura do workspace, os dados estão em:

Taxi Amarelo

/Volumes/workspace/raw-zone/taxi_amarelo/

```
|— taxi_amarelo_abr.parquet
|— taxi_amarelo_fev.parquet
|— taxi_amarelo_jan.parquet
|— taxi_amarelo_mai.parquet
|— taxi_amarelo_marc.parquet
```

Taxi Verde

/Volumes/workspace/raw-zone/taxi_verde/

```
|— taxi_verde_abr.parquet
|— taxi_verde_fev.parquet
|— taxi_verde_jan.parquet
|— taxi_verde_mai.parquet
|— taxi_verde_marc.parquet
```



Como Executar

Passo 1: Clone o repositório

```
git clone [url-do-seu-repositorio]
```

Passo 2: No Databricks Workspace

2.1. Importe os notebooks

1. No Databricks, vá em **Workspace**
2. Clique em **Import**
3. Importe os arquivos **.py** do repositório como notebooks
4. Ou copie e cole o código em novos notebooks

2.2. Verifique se os dados existem

Execute este comando para confirmar que os dados estão lá:

```
# Verificar taxi amarelo
```

```
dbutils.fs.ls("/Volumes/workspace/raw-zone/taxi_amarelo/")
```

```
# Verificar taxi verde
```

```
dbutils.fs.ls("/Volumes/workspace/raw-zone/taxi_verde/")
```

Passo 3: Execute os pipelines NA ORDEM

3.1. Pipeline Taxi Amarelo (Trusted)

```
# Notebook: raw-trusted-taxi-amarelo.ipynb
```

```
# Lê: /Volumes/workspace/raw-zone/taxi_amarelo/*.parquet
```

```
# Cria: trusted-zone.tb_corrida_taxi_amarelo
```

3.2. Pipeline Taxi Verde (Trusted)

```
# Notebook: raw-trusted-taxi-verde.ipynb
```

```
# Lê: /Volumes/workspace/raw-zone/taxi_verde/*.parquet
```

```
# Cria: trusted-zone.tb_corrida_taxi_verde
```

3.3. Pipeline Taxi Unificado (Refined)

```
# Notebook: trusted-refined.ipynb
```

```
# Lê: trusted-zone.tb_corrida_taxi_amarelo + trusted-zone.tb_corrida_taxi_verde
```

```
# Cria: refined-zone.tb_corrida_taxi_unificado
```

Estrutura do Repositório

```
ifood-case/
```

```
|— src/
|   |— raw-trusted-taxi-amarelo.ipynb
|   |— raw-trusted-taxi-verde.ipynb
|   |— trusted-refined.ipynb
|— docs/
|   |— README.md
|   |— taxi_amarelo_trusted_documentation.md
|   |— taxi_verde_trusted_documentation.md
```

```
| └─ taxi_corrida_refined.md
| └─ analysis/
|   └─ analytics.ipynb
|   └─ README.md
| └─ requirements.txt
```

Ajustes Necessários

Se os dados estiverem em outro local

Caso os dados não estejam no caminho `/Volumes/workspace/raw-zone/`, ajuste as variáveis:

```
# No início de cada pipeline, altere:
path = "/seu/caminho/para/os/dados/"
```

Se usar outro workspace

Os caminhos podem variar. Use este comando para encontrar seus dados:




```
dbutils.fs.ls("/") # Listar raiz
dbutils.fs.ls("/Volumes/") # Listar volumes
```

Observações Importantes

- **Ordem de execução importa:** Trusted antes de Refined
- **Cluster ativo:** Certifique-se que o cluster está rodando
- **Permissions:** Você precisa ter acesso de escrita nos schemas
- **Dados originais:** Os arquivos parquet já estão organizados por mês

Resultado Esperado

Depois de executar tudo:

-  `trusted-zone.tb_corrida_taxi_amarelo` - Dados limpos taxi amarelo
 -  `trusted-zone.tb_corrida_taxi_verde` - Dados limpos taxi verde
 -  `refined-zone.tb_corrida_taxi_unificado` - Dados unidos (Jan-Mai 2023)
-

Nota: Este projeto foi desenvolvido usando Databricks Community Edition com dados locais no DBFS. Não foi possível conectar via S3 por limitações da edição free.