Pontifícia Universidade Católica do Rio de Janeiro
Departamento de Informática - Pós-Graduação em Data Science & Analytics
Disciplina: Engenharia de Dados
Relatório Final
Aluno: João Felipe Maia Barbosa
Data: Dezembro/2025

---

1. Introdução e Objetivo

Este projeto tem como objetivo a construção de um pipeline de Engenharia de Dados ponta a ponta (*End-to-End*) para analisar o mercado de aluguéis de curta temporada (*Short-Term Rentals* - STR). O foco principal é a cidade do **Rio de Janeiro**, utilizando **New York City** como base comparativa de um mercado maduro.

A abordagem adotada foi híbrida, equilibrando a construção técnica de um Data Lakehouse com a entrega de valor analítico. O MVP buscou responder a perguntas fundamentais sobre performance de anúncios, precificação e impacto da profissionalização dos anfitriões (*hosts*).

2. Fonte de Dados e Arquitetura

Para a execução do projeto, foram utilizados dados públicos do portal **Inside Airbnb**. A arquitetura foi implementada no **Databricks Community Edition**, utilizando o **Unity Catalog** para governança e gestão de volumes.

2.1. Dados Brutos

Foram selecionados os snapshots mais recentes dos seguintes datasets para ambas as cidades:

- **listings.csv.gz:** Atributos detalhados dos anúncios.
- **calendar.csv.gz:** Disponibilidade, ocupação e preços diários.
- **reviews.csv.gz:** Histórico de avaliações.
- **neighbourhoods.geojson:** Geometria espacial dos bairros.

2.2. Pipeline de Dados (Medallion Architecture)

O fluxo de dados foi estruturado em três camadas lógicas:

1. **Bronze:** Ingestão bruta (*Raw Data*), preservando o formato original.
2. **Silver:** Limpeza, tipagem e padronização (*Refined Data*).
3. **Gold:** Modelagem dimensional (*Star Schema*) para consumo analítico (*Business Data*).

---

3. Desenvolvimento do Pipeline de Engenharia

3.1. Ingestão e Camada Bronze

A ingestão enfrentou desafios técnicos relacionados às limitações do ambiente *Serverless* do Databricks Community.

- **Armazenamento:** Utilização de *Managed Volumes* no Unity Catalog para persistência dos arquivos CSV.
- **Tratamento de GeoJSON:** O arquivo de geometrias (`.geojson`) não era nativamente tabular. Foi desenvolvido um script Python local para extrair as coordenadas e converter o arquivo para CSV antes do upload.
- **Leitura:** Utilização da função `read_files` com parâmetros robustos (`multiLine`, `quote`, `escape`) para tratar quebras de linha dentro de campos de texto descritivos, garantindo que não houvesse desalinhamento de colunas.

3.2. Tratamento e Camada Silver

Na camada Silver, o foco foi a qualidade e o enriquecimento dos dados sem perda de informação (*soft delete*).

- **Limpeza de Texto:** Remoção de tags HTML (`<br>`, tags de formatação) e normalização de caracteres especiais nos campos de comentários e descrições.
- **Tipagem:** Conversão de strings de moeda e data para tipos nativos (`DOUBLE`, `DATE`), criando colunas sufixadas (ex: `price_numeric`, `last_review_date`).
- **Enriquecimento:** Criação de colunas booleanas para categorização (ex: `host_is_superhost_bool`, `is_entire_home`).
- **Decisão de Projeto:** Optou-se por manter registros com preços nulos na camada Silver para garantir rastreabilidade, delegando a filtragem para a camada Gold.

3.3. Modelagem e Camada Gold (Star Schema)

A camada final foi modelada seguindo a metodologia dimensional de Kimball (*Star Schema*), otimizada para consultas analíticas.

Tabelas Dimensionais:

- **dim_listings:** Tabela central contendo características do imóvel, flags de performance e dados do anfitrião.
- **dim_hosts:** Dados consolidados dos anfitriões, incluindo classificação de "profissionalismo" baseada no tamanho do portfólio.
- **dim_neighbourhoods:** Dados geográficos e geometria dos bairros.

Tabelas Fato:

- **fact_calendar:** Granularidade diária (*listing* x *data*). Contém disponibilidade e métricas de receita estimada.
- **fact_reviews:** Granularidade por avaliação (*listing* x *review*). Permite análise de sentimento e frequência.

Foram aplicadas *constraints* de Chave Primária (PK) e Chave Estrangeira (FK) no Unity Catalog para garantir a integridade referencial do modelo.

---

4. Qualidade dos Dados e Catálogo

Antes da análise final, foi executada uma etapa rigorosa de **QA (Quality Assurance)**:

- Validação de integridade referencial entre Fatos e Dimensões (zero órfãos).
- Verificação de consistência temporal e lógica (ex: datas de reviews dentro de ranges válidos).
- Documentação completa em um **Catálogo de Dados**, detalhando grão, origem, tipo de dado e regras de negócio para cada tabela da camada Gold.

---

5. Análise de Dados e Resultados

A etapa final consistiu na Análise Exploratória (EDA) utilizando SQL e PySpark, respondendo às perguntas de negócio definidas no início do projeto.

5.1. Fatores de Alta Performance

Identificou-se que anúncios de "Alta Performance" (Top 10% em ocupação/receita) possuem características claras:

- Prevalência de imóveis do tipo *Entire home/apt* de porte pequeno a médio.
- Alta frequência e recência de *reviews* (prova social é determinante).
- Gestão por anfitriões que, embora não necessariamente tenham portfólios gigantescos, demonstram comportamento profissional (Superhosts).

5.2. Dinâmica de Preço e Receita

- **Influência:** O preço é determinado majoritariamente pelas características físicas (número de quartos/capacidade).
- **Correlação:** No Rio de Janeiro, observou-se uma correlação positiva relevante (≈0.33) entre ocupação e receita anual, indicando que a estratégia de volume é eficaz. Em NYC (mercado saturado), a correlação é menor (≈0.07), sugerindo que o preço da diária tem peso maior na receita final do que a taxa de ocupação.

5.3. Impacto da Localização

A análise geoespacial confirmou a concentração de valor:

- **RJ:** Leblon, Ipanema e Barra da Tijuca concentram a maior produtividade financeira.
- **NYC:** A região de Midtown Manhattan domina em preço e receita.
- Apesar da concentração, a variância de ocupação dentro dos mesmos bairros sugere que a qualidade do anúncio e do serviço (reviews) é o diferencial competitivo final.

---

6. Conclusão

O MVP cumpriu integralmente os requisitos acadêmicos e técnicos propostos. Foi entregue um pipeline robusto, documentado e funcional, capaz de transformar dados brutos e complexos em insights de negócio estratégicos. A arquitetura escolhida (Databricks + Medallion) provou-se adequada, garantindo escalabilidade e organização para futuras expansões do projeto.