



# Base: Inside Airbnb


Link: <http://insideairbnb.com/get-the-data.html>

## Equipe de Projeto

- Elizabete Yamaguti
- Fabiane Bugarin
- Luciane Altimari
- Rodolfo Bugarin



# Tópicos

- A. Objetivo da proposta
  - B. Dados utilizados
  - C. Arquitetura da Solução
  - D. Análise
  - E. Modelagem
  - F. Recomendações
- 



**A. Objetivo da proposta:** Prever a taxa de ocupação dado o preço e as características do imóvel

- Permitir ao locatário (Host) definir o preço do aluguel de seu imóvel de acordo com a taxa de ocupação desejada, maximizando o potencial de retorno.

## B. Dados utilizados

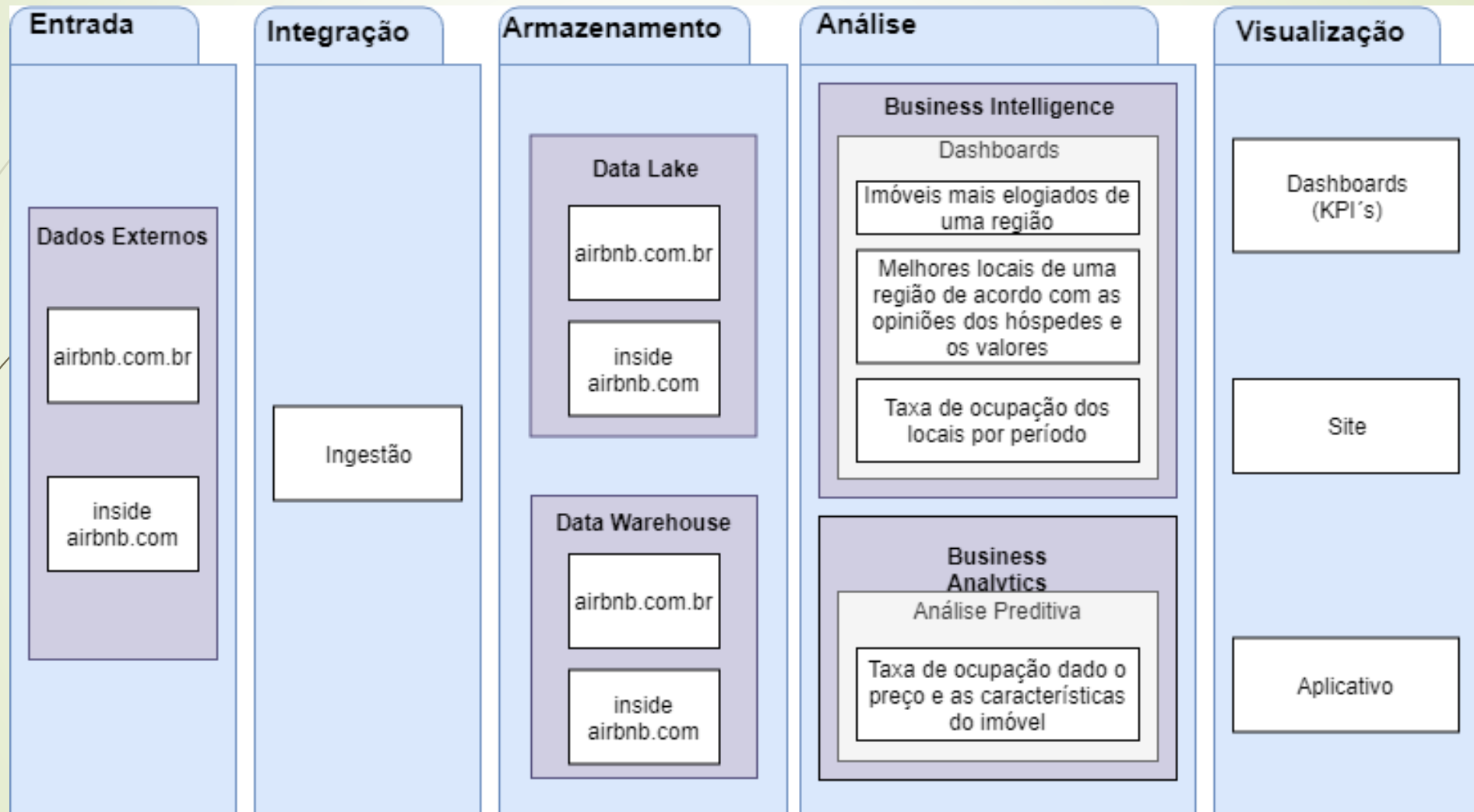
Todas as colunas das bases abaixo foram analisadas na Análise Exploratória

- ➡ listings.csv
- ➡ listings\_detailed.csv
- ➡ neighbourhoods.csv
- ➡ neighbourhoods.geojson
- ➡ calendar.csv
- ➡ reviews.csv
- ➡ reviews\_detailed.csv

## B. Dados utilizados (continuação)

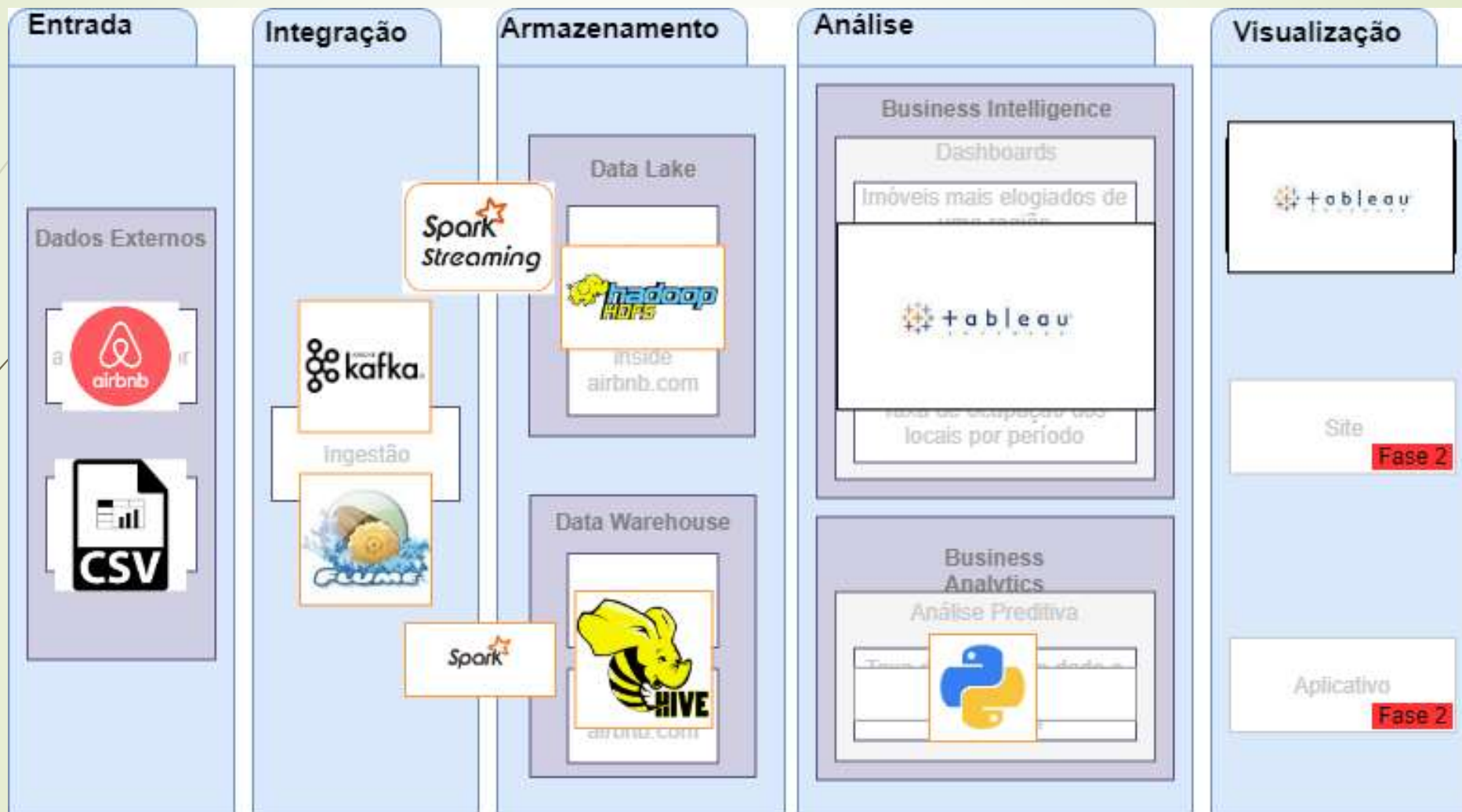
- Utilizamos em nosso modelo as colunas quantitativas e qualitativas (incluindo as dummies) somente da base `listings_detailed.csv`;
  - Todas as colunas da base `listings.csv` estão presentes na base `listings_detailed.csv`;
  - Todas as colunas da `reviews.csv` estão presentes na base `reviews_detailed.csv`. A única informação relevante (pontuação dada pelos reviewers) já está presente na base `listings_detailed.csv`;
  - A base `neighbourhoods.csv` possui todos os bairros da cidade e esta informação já está presente na base `listings_detailed.csv`.
  - A base `neighbourhoods.geojson` lista todas as latitudes e longitudes que definem cada bairro e esta informação já está presente na base `listings_detailed.csv`.
  - Para estimar a taxa de ocupação necessitamos das características atuais do listing (como os amenities, o bairro e o preço atual) e dados de seu passado (como a taxa de ocupação média mensal e a avaliação dos hóspedes). Como a base `calendar.csv` apresenta dados futuros (preço de aluguel a ser praticado em cada dia nos meses subsequentes), esta base não é necessária. Foi utilizada o preço de aluguel atual da base `listings_detailed.csv`;
- Não utilizamos campos descritivos pois já temos as informações com a característica dos listings no campo amenities;
- A análise descritiva de cada coluna está no código (Aribnb Analise Exploratoria e Modelo.ipynb).

## C. Arquitetura da Solução – Conceitual

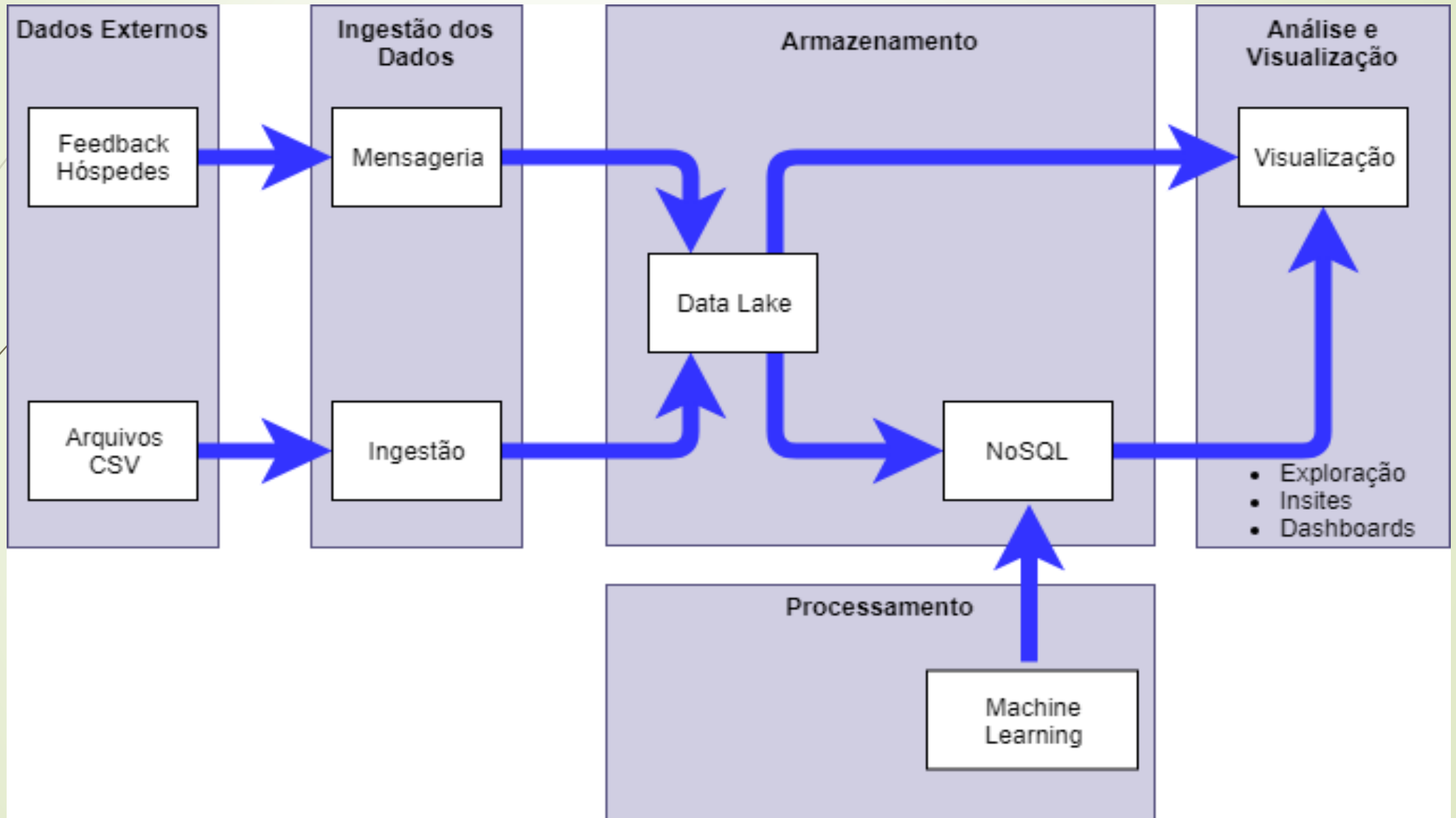




## C. Arquitetura da Solução – Tecnológica

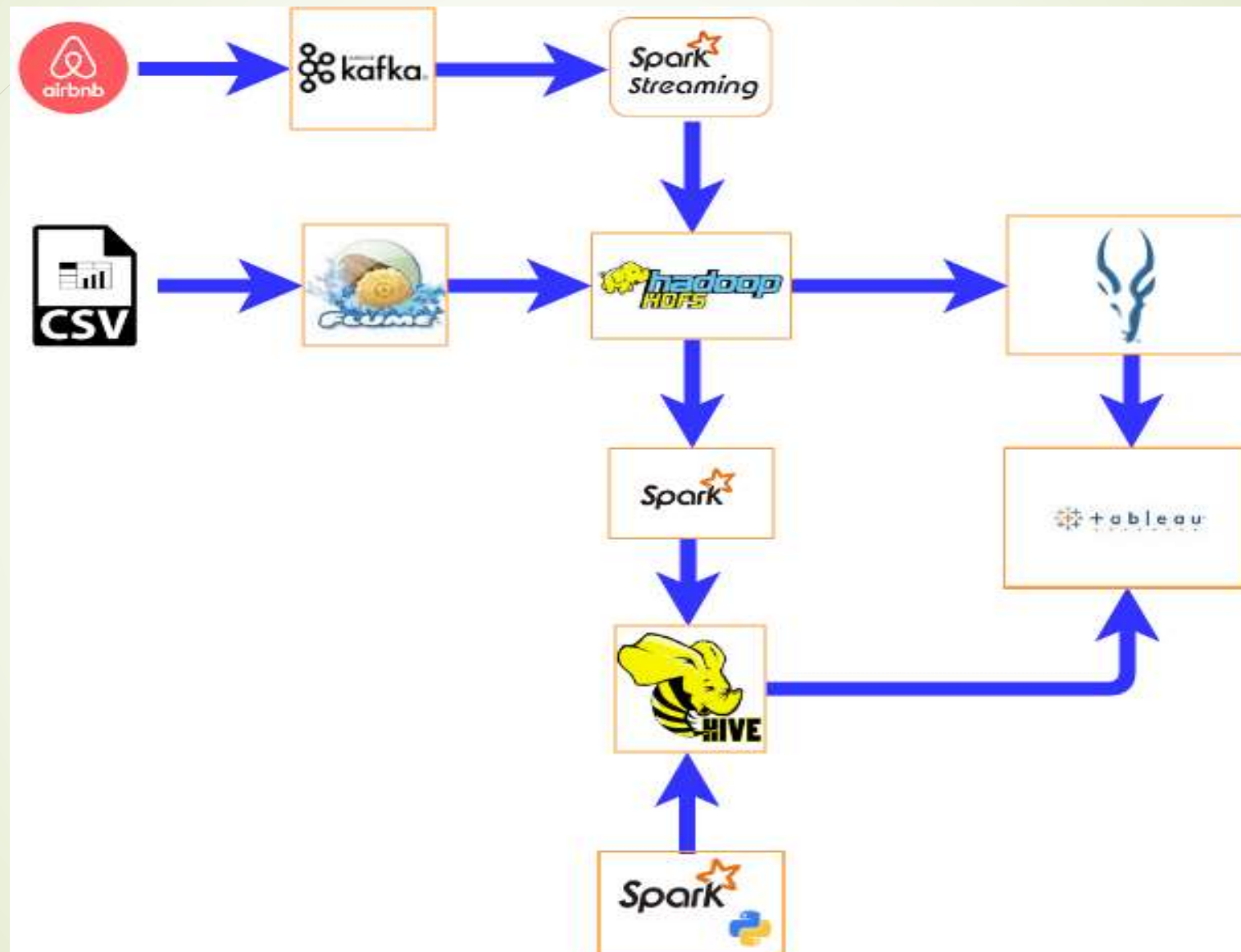


## C. Arquitetura da Solução – Fluxo Completo





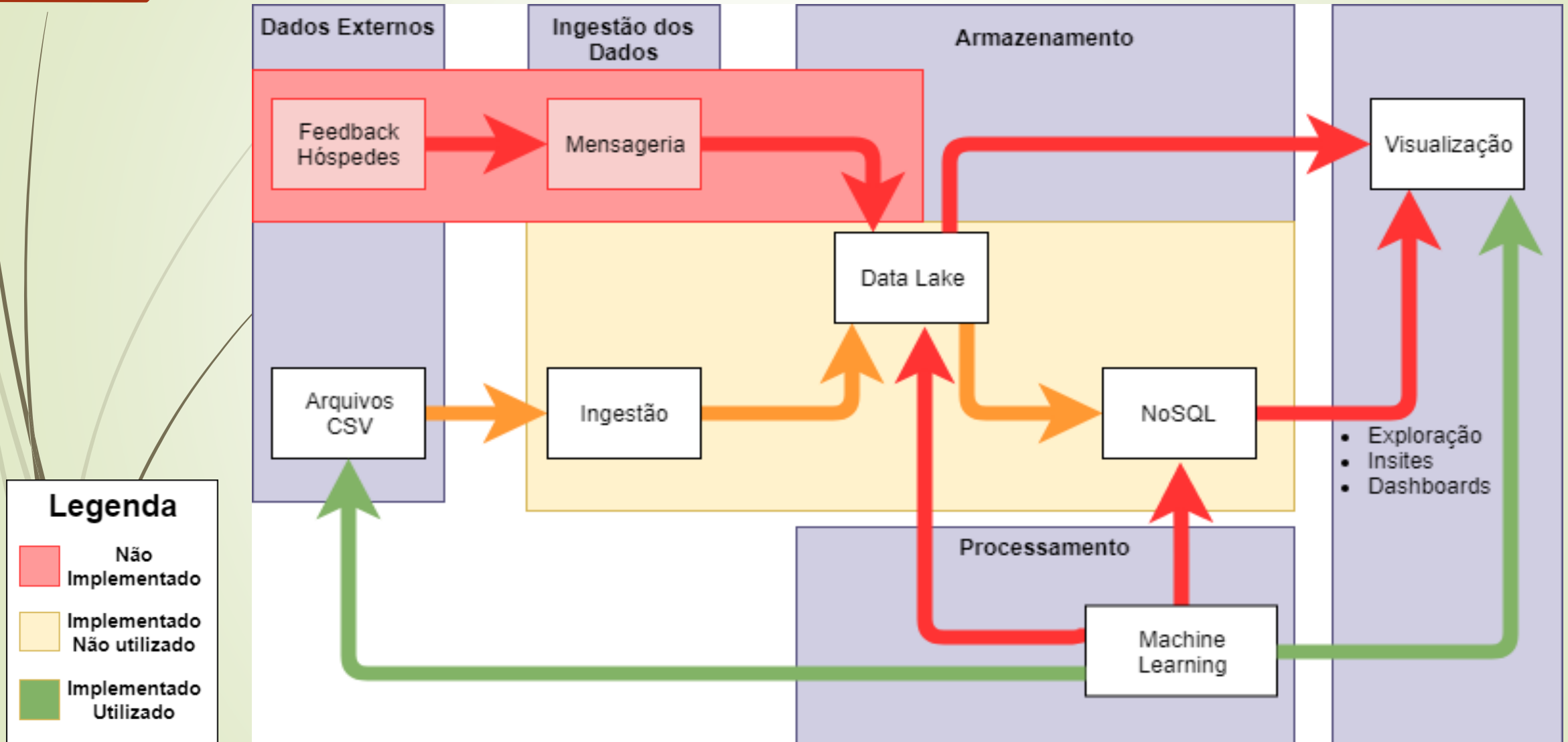
## C. Arquitetura da Solução – Fluxo Tecnológico Completo



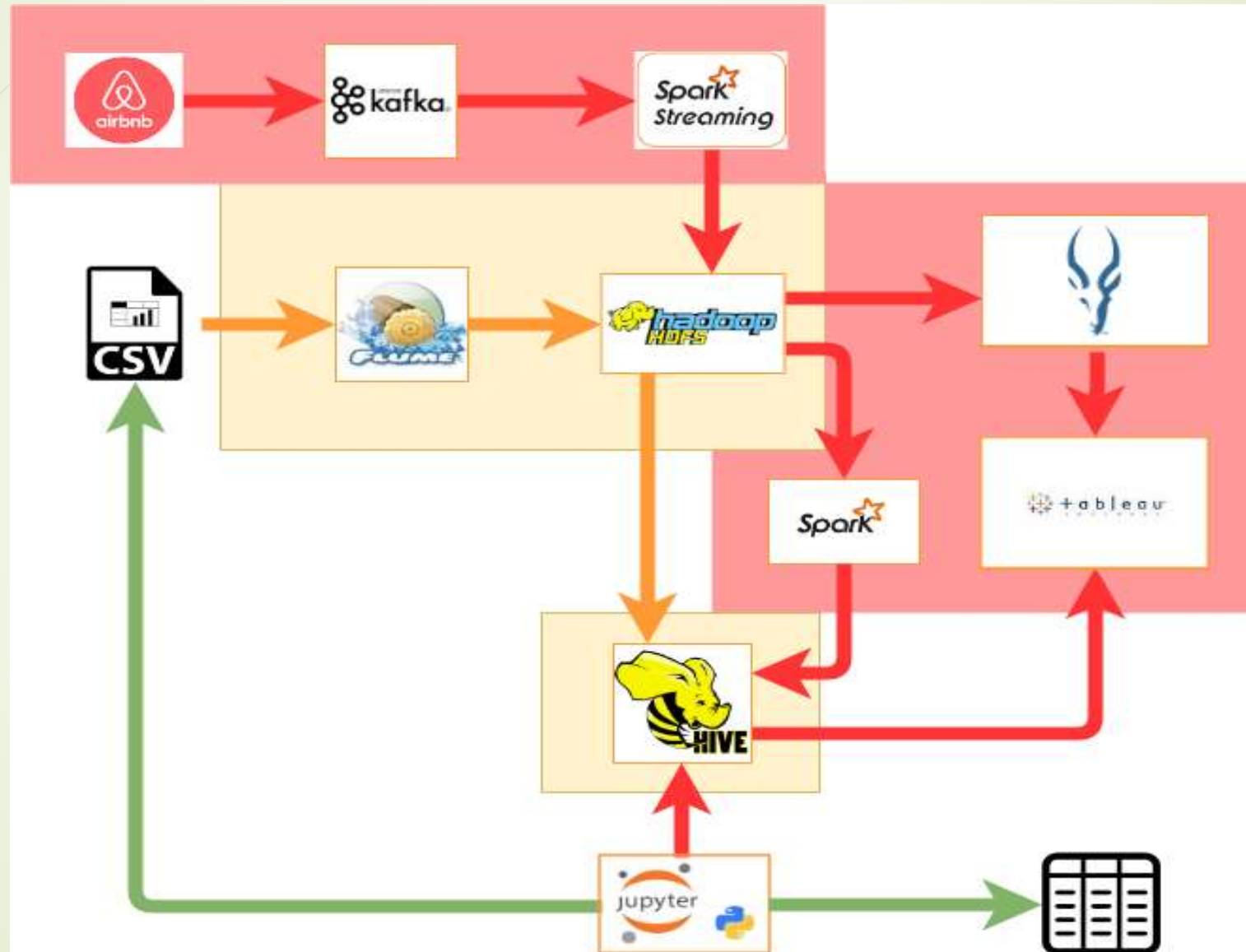
### Legenda

-  Não Implementado
-  Implementado Não utilizado
-  Implementado Utilizado

## C. Arquitetura da Solução – Fluxo Desenvolvido



## D. Arquitetura da Solução – Fluxo Tecnológico Desenvolvido



## D. Análise Exploratória

- ❑ O site Inside Airbnb não fornece um dicionário de dados. Para interpretarmos as colunas, criamos um usuário do Airbnb para verificar cada campo disponível. Adicionalmente, verificamos o conteúdo das colunas, comparamos os valores entre colunas e interpretamos o nome de cada coluna.
- ❑ Criação de colunas dummies para o conteúdo da coluna Amenities, onde cada listing tem diferentes “amenidades” que são dummies no modelo;

	id	amenities
0	3344	{TV,"Cable TV",Internet,Wifi,"Air conditioning",Kitchen,Doorman,Gym,Breakfast,Elevator,"Buzzer/wireless intercom",Heating,"Suitable for events",Washer,Dryer,"Smoke detector","First aid kit","Fire extinguisher",Essentials,Shampoo,"Lock on bedroom door","24-hour check-in",Hangers,Iron,"Laptop friendly workspace","Hot water","Long term stays allowed","Host greets you","Paid parking on premises")
1	3362	(TV,Internet,Wifi,"Air conditioning",Kitchen,"Paid parking off premises","Free street parking",Heating,"Family/kid friendly",Washer,Dryer,"Smoke detector","Carbon monoxide detector","First aid kit",Essentials,Shampoo,"Lock on bedroom door","24-hour check-in",Hangers,"Hair dryer",Iron,"Laptop friendly workspace","Self check-in",Keypad,"Private entrance","Hot water","Luggage dropoff allowed","Long term stays allowed")

- ❑ Conversão de strings contendo símbolos de moeda, separador de milhar e percentual para variáveis numéricas, removendo os caracteres especiais;
- ❑ Tratamento dos missings com base em regras de negócio para cada coluna. Exemplo: para atualizar para hospedagens semanais e mensais foi utilizado preço da diária vezes 7 e 30, respectivamente.

## E. Modelagem

- ❑ Bairros (neighbourhood) foram convertidos em dummies, o que faz com o que o modelo deve ser gerado por cidade. Exemplo: a cidade de Nova Iorque tem bairros que não existem em Washington (e vice versa). E duas cidades podem ter bairros com mesmo nome mas que são locais diferentes o que tem impactos diferentes na previsão.
- ❑ Foram treinados modelos de Regressão Linear, Árvore de Decisão, Random Forest e Boosting com os seguintes resultados:
  - ❑ Regressão Linear - MSE: 0,0011 - R2: 87,26%
  - ❑ Árvore de Decisão - MSE: 0,0002 - R2: 97,51%
  - ❑ Random Forest - MSE: 0,0034 - R2: 60,26%
  - ❑ Boosting - MSE: 0,0001 - R2: 98,48%