

Universidade de São Paulo  
Instituto de Ciências Matemáticas e de Computação  
Departamento de Ciências de Computação

## **Exercicio 3,4 e 5**

**Luca Gomes Urssi - 10425396 - [lucaurssi@usp.br](mailto:lucaurssi@usp.br)**

**Vinicius Flnke Raiter José - 9791052 - [vinicius.jose@usp.br](mailto:vinicius.jose@usp.br)**

**Gabriel Santos Brito - 10284250**

**Prof. Dr. Caetano Traina Júnior**

### Exercício 3 - Preparar tabelas e carregar dados:

Foi utilizado as bibliotecas **panda**, **psycopg2** e **sqlalchemy** para extrair as colunas do arquivo csv e transformar as colunas em sql.

Foram criado as tabelas com os seguintes comandos:

```
import pandas as pd
```

```
%sql DROP TABLE IF EXISTS listings CASCADE;
```

```
df = pd.read_csv('listings.csv', sep=',')
```

```
df.to_sql("listings", engine)
```

```
%sql DROP TABLE IF EXISTS calendar CASCADE;
```

```
df = pd.read_csv('calendar.csv', sep=',')
```

```
df.to_sql("calendar", engine)
```

```
%sql DROP TABLE IF EXISTS reviews CASCADE;
```

```
df = pd.read_csv('reviews.csv', sep=',')
```

```
df.to_sql("reviews", engine)
```

### Exercício 4 - Normalize as 3 tabelas para BCNF:

```
Listings = { index, id, listings_url, scrape_id, last_scraped, source, name, description,
neighborhood_overview, picture_url, host_id, host_url, host_name, host_since, host_location,
host_about, host_response_time, host_response_rate, host_acceptance_rate,
host_is_superhost, host_thumbnail_url, host_picture_url, host_neighbourhood,
host_listings_count, host_total_listings_count, host_verifications, host_has_profile_pic,
host_identity_verified, neighbourhood, neighbourhood_cleansed,
neighbourhood_group_cleansed, latitude, longitude, property_type, room_type, accommodates,
bathrooms, bathrooms_text, bedrooms, beds, amenities, price, minimum_nights,
maximum_nights, minimum_minimum_nights, maximum_minimum_nights,
minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm,
maximum_nights_avg_ntm, calendar_updated, has_availability, availability_30, availability_60,
availability_90, availability_365, calendar_last_scraped, number_of_reviews,
number_of_reviews_itm, number_of_reviews_i30d, first_review, last_review,
review_scores_rating, review_scores_accuracy, review_scores_cleanliness,
review_scores_checkin, review_scores_communication, review_scores_location,
review_scores_value, license, instant_bookable, calculated_host_listings_count,
calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms,
calculated_host_listings_count_shared_rooms, reviews_per_month }
```

```
Reviews = { index, listing_id, id, date, reviewer_id, reviewer_name, comments }
```

Calendar = { index, listing\_id, date, available, price, adjusted\_price, minimum\_nights, maximum\_nights }

Dado que as 3 tabelas atendem a primeira forma normal, é necessário apenas identificar as chaves candidatas e aplicar a terceira forma normal em cada chave.

- Listings: index e id.
- Reviews: index, id e (reviewer\_id, date, listing\_id)
- Calendar: index e (listing\_id, date)

*Calendar* atende a terceira forma normal.

A tabela *review* não está na terceira forma normal, pois reviewer\_name depende de reviewer\_id, portanto é necessário separar em uma nova tabela.

- Reviews = { index, listing\_id, id, date, reviewer\_id, comments }
- Reviewer = { id, nome }

*Listings* não atende a terceira forma normal.

Listings = { index, id, listings\_url, scrape\_id, last\_scraped }

Scrape = { ID, source, name, description, neighborhood\_overview, picture\_url, host\_id, neighbourhood, neighbourhood\_cleansed, neighbourhood\_group\_cleansed, latitude, longitude, property\_type, room\_type, accommodates, bathrooms, bathrooms\_text, bedrooms, beds, amenities, price, minimum\_nights, maximum\_nights, minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_maximum\_nights, maximum\_maximum\_nights, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm, calendar\_updated, has\_availability, availability\_30, availability\_60, availability\_90, availability\_365, calendar\_last\_scraped, number\_of\_reviews, number\_of\_reviews\_itm, number\_of\_reviews\_i30d, first\_review, last\_review, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, licence, instant\_bookable, reviews\_per\_month }

Host = { ID, url, name, since, location, about, response\_time, response\_rate, acceptance\_rate, is\_superhost, thumbnail\_url, picture\_url, neighbourhood, listings\_count, total\_listings\_count, verifications, has\_profile\_pic, identity\_verified, calculated\_listings\_count, calculated\_listings\_count\_entire\_homes, calculated\_listings\_count\_private\_rooms, calculated\_listings\_count\_shared\_rooms }

## Exercício 5 - Diferença das tabelas e carregar dados:

Em **vermelho** o que não é compatível.

### Airbnb:

**Listings** = { **index**, id, listings\_url, scrape\_id, last\_scraped }

**Scrape** = { **ID**, source, name, description, neighborhood\_overview, picture\_url, host\_id, neighbourhood, neighbourhood\_cleansed, neighbourhood\_group\_cleansed, latitude, longitude, property\_type, room\_type, accommodates, bathrooms, bathrooms\_text, bedrooms, beds, amenities, price, minimum\_nights, maximum\_nights, minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_maximum\_nights, maximum\_maximum\_nights, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm, calendar\_updated, has\_availability, availability\_30, availability\_60, availability\_90, availability\_365, calendar\_last\_scraped, number\_of\_reviews, number\_of\_reviews\_itm, number\_of\_reviews\_i30d, first\_review, last\_review, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, licence, instant\_bookable, reviews\_per\_month }

**Host** = { **ID**, url, name, since, location, about, response\_time, response\_rate, acceptance\_rate, is\_superhost, thumbnail\_url, picture\_url, neighbourhood, listings\_count, total\_listings\_count, verifications, has\_profile\_pic, identity\_verified, calculated\_listings\_count, calculated\_listings\_count\_entire\_homes, calculated\_listings\_count\_private\_rooms, calculated\_listings\_count\_shared\_rooms }

**Reviews** = { **index**, listing\_id, id, date, reviewer\_id, comments }

**Reviewer** = { **id**, nome }

**Calendar** = { **index**, listing\_id, date, available, price, adjusted\_price, minimum\_nights, maximum\_nights }

### MR :

**Localização** = { país, estado, cidade, bairro }

**Interesses** = { país, estado, cidade, interesse }

**Usuário** = { ID, nome, sobrenome, telefone, dataNasci, sexo, email, senha, numero, rua, cidade, CEP, país, estado, cidade\_Localização }

**Tipo\_Usuario** = { Usuario, hospede, dono }

**Propriedade** = { ID, nome, numero, rua, cidade, CEP, país, estado, cidade\_localização, classificação, nroBanheiros, nroQuartos, maxHospedes, diaria, limpeza, min\_noite, max\_noite, hr\_checkin, hr\_checkout, disponibilidade\_inicio, disponibilidade\_fim, nrocamas, tipocamas, comodidades, regras, dono }

**Avalia** = { timestamp, hospede, propriedade, mensagem, limpeza, comunicação, localização, valor }

**fotosAvaliação** = { timestamp, hospede, propriedade, foto }

**Trocagem** = { Timestamp, use1, user2, mensagem }

**Fotos** = { timestamp, user1, user2, foto }

**Locação** = { Datareserva, hospede, propriedade, confirmação, checkin, checkout, imposto, desconto, cod\_promo, custoEstadia, preçoTotal }

**ContaBancaria** = { ID, roteamento, tipo, dono }

#### **Comentarios:**

A forma de guardar a **Avaliação** de cada propriedade é incompatível, pois o Airbnb está armazenando o número de avaliações e as médias das notas, mas no mapeamento deste trabalho foi guardado cada avaliação de forma separada. Assim impossibilitando a conversão das notas.

A **Locação**, **Conta Bancária** e **Troca de Mensagem** não estão incluídos nos arquivos do Airbnb. Também não foram incluídas fotos durante a avaliação.

O **endereço** da propriedade não foi incluído, apenas a cidade e estado do "Host" foram incluídos.