BIG DATA STORAGE - AIRBNB CASE STUDY

Group 3
Victoriya Mokhovikova 20230796
Vladislav Botnev 20230795
Yan Sidoryk 20222004
Christian Deluca 20241264
Nikolaos Kanellopoulos 20242109

PROJECT OBJECTIVE AND BUSINESS CONTEXT

Main Goal:

Efficiently store and analyze large-scale Airbnb data from Lisbon using MongoDB

Why this project matters?

The dataset reflects real-world, semi-structured data. MongoDB lets us extract insights thanks to its flexible schema and scalability.

Business applications

- optimize pricing strategies
- Understand user and host behavior
- Identify high-performing listings and districts

WHY MONGODB & DATASET OVERVIEW

Dataset

- Short-term rental listings in Lisbon
- includes property details, hosts, neighborhoods, and guest reviews

Why this dataset

- Realistic, diverse and widely used platform
- Dynamic data with user reviews, prices, amenities
- Valuable for business insights and modeling practice

Why MongoDB

- Flexible schema → ideal for heterogeneous data
- Handles nested structures like reviews and amenities
- Great for semi-structured, real-world datasets
- Trade-off: no joins, requires thoughtful design

DATABASE STRUCTURE & DESIGN DECISIONS

Collections

- **listings:** main dataset with property info, host id, prices, availability and more
- **hosts:** information about hosts(name, start date, number of listings, photos)
- **neighbourhoods**: textual and grouped info about Lisbon’s districts
- **neighbourhoods_geo:** geographic features (GeoJSON format) for mapping
- **reviews:** user comments and review metadata (listing_id, reviewer_id, date)

Design Decisions

- Flexible schema to support heterogeneous and nested data
- Embedded fields in listings (e.g. URLs, amenities)
- Referenced links across collections using host_id, listing_id, etc.
- Applied data validation: e.g. valid dates, enum values (room_type), coordinates
- Separated geographic and descriptive neighborhood data for clarity

DATA VALIDATION RULES

Applied to Listings

- Required fields: id, name, price, room_type, host_id
- latitude and longitude must be valid coordinates
- room_type must match one of 4 enum values
- last_scraped must not be in the future
- Bathrooms, bedrooms, beds must be non-negative
- Availability values must follow logical order (30 ≤ 60 ≤ 90 ≤ 365)

Applied to Neighbourhoods & Geo

- Required: neighbourhood, neighbourhood_group (non-empty strings)
- In neighbourhood_features: geometry coordinates must be valid arrays

Applied to Hosts

- Required: host_name, host_since, host_is_superhost, host_listings_count
- host_name must be string
- host_since must not be in the future
- host_is_superhost must be a Boolean
- host_listings_count must be non-negative

Applied to Reviews

- Required fields: date, reviewer_id, comments, listing_id
- comments must be a non-empty string
- date must be not in the future

QUERIES FOR BUSINESS INSIGHTS

- Top 10 listings ranked by estimated annual revenue
- Most expensive districts by average price of listings
- Districts with the largest number of listings
- Districts with the largest average annual revenue
- Top 5 districts by average number of bedrooms
- Average revenue of hosts vs. superhosts
- Average availability by price chunk (<100, 100-199, 200-299, 300 – 499, 500+)
- First 10 hosts on the platform
- Top 5 listings by reviews

INSIGHTS

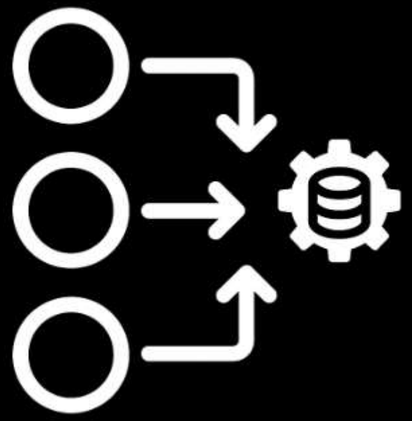
- Top listing by average revenue: Cascais Bay Terrace I, Casa do Largo - Central Lisbon Apartment
- Districts with high average revenue: Azambuja Meca, Cacm e So Marcos
- Districts with biggest number of listings: Santa Maria Maior(3406), Misericrdia(2474), Arroios (2133), Cascais e Estoril(1510)
- Districts with the highest annual revenue: Cardosas, Azambuja, Santa Maria Maior
- Avg revenue: superhost - 16068.2, host - 6722.5
- The lowest availability is in 100-199 chunk: availability_60 = 25.7, availability_90 = 41.7

Conclusions

For new host: it is better to be a superhost, most profitable price range is 100-200, the best districts are Azambuja, Santa Maria Maior, Cascais, Estoril, Arroios

For Airbnb: analyzing the availability Airbnb can make approproate rocomendation to balance the demand

AGGREGATIONS USED FOR ANALYSIS



- Compared average price, rating and listing count per property type
- Linked host data to calculate avg. price and rating per host
- Aggregated neighbourhood scores (cleanliness, value, location)
- Measured presence and impact of amenities (e.g. Wi-Fi, hot tub)

QUERIES FOR INDEXING RESEARCH

Indexes Created

- Single-field index on neighbourhood_cleansed
- Compound index on neighbourhood_cleansed & room_type

Performance Measurement

- Measured query performance before & after indexing
- Evaluated efficiency with explain("executionStats")
- Results: 41 ms → 20 ms; 35 ms → 3 ms

BENEFITS & LIMITS OF CHOICES

Benefits

- Flexible schema adapts to varied property data
- Indexes significantly improve query performance
- Aggregations provide versatile and powerful analysis
- Data validation ensures consistent and accurate data

Limits

- No predefined structure may lead to inconsistency
- Index misuse can degrade performance
- Complex pipeline can slow down aggregations
- Data validation requires careful definition and testing

THANK YOU