NoSQL & Big Data Exam – 2h

Subject: NoSQL Database Modeling, Data Quality Check, Migration, and Querying
Duration: 2 hours
Environment: Docker, Python, Git, Report
Dataset: Airbnb Seasonal Rentals (CSV format provided)
NoSQL Models: Key-Value (Redis), Document Store (MongoDB), Column-Family (HBase)

▯ Objectives

You are tasked with designing and implementing a NoSQL database solution for the seasonal rentals dataset. You now need to:

1. Choose the most appropriate NoSQL model.
2. Propose a suitable data modeling for your choice.
3. Clean and validate the data before migration.
4. Write a script to perform the migration.
5. Deploy the NoSQL system using Docker Compose.
6. Interact with the data using Python.
7. Justify your design choices and answer analysis questions.

All files needed are available in this link : here

▯ Instructions

1. Choose a NoSQL Model

- Pick one of the following types:
  - Key-Value Store (Redis)
  - Document Store (MongoDB)
  - Column-Family Store (HBase)
- Justify your choice according to the data structure, queries to be supported, and scalability concerns.

2. Design the NoSQL Schema

- Describe how you will structure the dataset in your chosen database system.
  - Create a schema proposal & present informations that will be saved in your database.
- Highlight any denormalization decisions or redundancy you accept.

3. Data Cleaning and Quality Check

- Analyze the CSV file for missing values and invalid entries.
- Write a **Python script** that:
  - Loads the CSV file.
  - Checks and logs data quality issues.
  - Cleans and transforms the data to fit your NoSQL model.

4. Data Migration

- Write a **Python migration script** that:
  - Loads the cleaned data.
  - Connects to the NoSQL system.
  - Inserts the records using the model you designed.

5. Docker Compose

- Create a `docker-compose.yml` to deploy:
  - Your selected NoSQL database.
  - Any required services (Jupyter, GUI client, etc.).
  - Your migration script.

6. Querying with Python

- Using Python, write scripts or notebooks to answer the following questions after data migration:

```
 1. How many listings are there for each type of property?

2. How many listings were made on June 12, 2024, for the city of Paris?

3. What are the 5 listings with the highest number of reviews? How many reviews do they have?

4. What is the total number of unique hosts?

5. How many instant-bookable rentals are there? What proportion of the listings do they represent?

6. Are there any hosts with more than 100 listings on the platform? If so, who are they, and what percentage of hosts do they repres

7. How many unique superhosts are there? What percentage of hosts do they represent?
```

7. Deliverables

You must submit:

- A link to a GitHub repository containing:
  - Docker configuration
  - Python scripts
  - Data cleaning and migration code
  - Queries and results (as code or notebooks)
  - A report (pdf or markdown) with:
    - Justification for your NoSQL model choice
    - Schema design explanation
    - Screenshots of Docker containers running
    - Explanations and answers to the query section
    - Optional: Data quality statistics

□ Evaluation Criteria

Criterion	Weight
Justification of model choice	10%
Schema adequacy and clarity	10%
Data cleaning & validation	20%
Migration script functionality	25%
Docker setup	10%
Query accuracy	15%
Report completeness	10%

Good luck!