Project Title:

Data Integration and Analysis Using a Data Warehouse and NoSQL

Objective:

The goal of this project is to design and implement a Data Warehouse (DW) and explore the use of NoSQL technologies for managing heterogeneous data, leveraging Neo4j and Cypher for graph-based data representation. The project emphasizes data integration, querying, visualization, clustering analysis, and handling Big Data.

General Instructions:

- Students can work in groups of 2, 3, or 4 members. Each group must inform the instructor about the composition of the team and the chosen dataset.
- When submitting the dataset, clearly indicate its complexity and the challenges associated with its storage and integration.

Project Outline:

1. **Dataset Selection and Domain Analysis**
   - Choose and analyze a dataset relevant to a specific application domain.
   - Highlight the context, challenges, and peculiarities of the dataset (e.g., from UCI, Kaggle, data.gouv.fr).
   - Identify and justify the use of heterogeneous data sources, where applicable, and discuss how NoSQL can address challenges with unstructured or semi-structured data.

2. **Logical Design of the Data Warehouse**
   - Create a schema for the Data Warehouse using a **star schema** or **snowflake schema**.
   - Add hierarchies and aggregates where relevant.
   - For heterogeneous data, model relationships using Neo4j and design queries with Cypher to connect and analyze data.

3. **Physical Implementation**
   - Physically create the Data Warehouse.
   - Populate the DW with data; random data generation is acceptable if necessary.
   - Load unstructured and semi-structured data into Neo4j, leveraging its graph database capabilities to capture relationships among entities.

4. **Querying the Data Warehouse and NoSQL Database**
   - Use OLAP queries (e.g., CUBE, ROLLUP) to generate output matrices with rows and columns.
   - Develop and execute Cypher queries on the Neo4j database to explore relationships in the heterogeneous data.

   - Combine insights from the relational and graph databases to address the application domain's challenges.

5. **Big Data Integration and Processing**
   - Discuss how Big Data tools (e.g., Hadoop, Spark) can enhance scalability and performance for large datasets.
   - Optional: Implement a pipeline to process and analyze Big Data using the selected technologies. You can use Pyhton, or PySpark.

6. **Data Analysis and Visualization**
   - Load the generated matrix (or matrices) into Python.
   - Visualize the data using scatter plots or other appropriate methods.
   - Perform clustering using the k-means algorithm and visualize the clusters.
   - Interpret and explain the results, discussing patterns and insights.

7. **Report Preparation**
   - Compile a comprehensive report covering all project phases:
     - Objective and context of the DW and NoSQL integration.
     - Schema design and implementation.
     - Description and motivation for the queries.
     - Analysis of results and visualization.
     - Clustering analysis and conclusions.
     - Discussion of the role of Big Data and NoSQL in addressing project challenges.
   - Submit the report on Moodle.

Deliverables:

- A fully implemented Data Warehouse.
- A Neo4j database with queries using Cypher.
- A Python script for loading, visualizing, and clustering the data.
- A detailed report explaining the design, queries, analysis, and insights.