# Cross-Database Data Integration and Analysis Project

**Project Overview:**

In this complex data engineering and analysis project, your team will tackle the integration of data from three different database systems - SQL, MongoDB, and Neo4j. The aim is to combine structured, semi-structured, and graph data sources for a comprehensive analysis. You will utilize PySpark for data processing and manipulation, and then perform in-depth data analysis to derive meaningful insights.

**Project Tasks:**

1. Select the data you want to work on (any topic you are interested in, any source, ex. – Sales transactions). I suggest using different APIs. Load the different data into MySQL, MongoDB and Neo4J database in Python. If you don’t have any ideas, let me know – I will find the data for you.

**In Python (example for the Web Store data):**

1. Data Extraction:

- Extract data from a SQL database containing sales transactions, a MongoDB database with customer reviews, and a Neo4j graph database storing social network relationships.

2. Data Transformation:

- Use PySpark to clean and transform the extracted data into a unified format suitable for analysis.

- Handle data type conversions, null values, and inconsistencies.

3. Data Integration:

- Combine the transformed data from the three databases into a common storage or data frame structure. (in Python – examples, in Pandas or you can store all this data locally in the files)

- Design a schema that accommodates the heterogeneous data types and structures.

4. Data Analysis:

- Perform exploratory data analysis (EDA) to understand the characteristics of the integrated dataset (average values, standard deviation, etc.).

- Using Time Series, Linear regression or any other approach to identify patterns, trends, and correlations.

5. Graph Analysis:

- Extract relevant data from the Neo4j graph database to analyse social network influences on customer behaviours.

- Identify key influencers, communities, and connections

6. Data Visualization:

- Create interactive visualizations using libraries like Matplotlib, Plotly, or Seaborn.

- Visualize sales trends, customer sentiment distribution, and social network graphs to present findings effectively.

7. Insights and Reporting:

- Summarize the insights gained from the data analysis, including correlations between social connections, sentiments, and sales performance.

- Prepare a detailed report and presentation for stakeholders, highlighting actionable recommendations.

8. Performance Optimization:

- Optimize data processing tasks in PySpark for improved performance, considering data partitioning, caching, and parallel processing.

Additional:

9. Sentiment Analysis:

- Apply natural language processing techniques using PySpark to the customer reviews from MongoDB.

- Perform sentiment analysis to categorize reviews as positive, negative, or neutral.

- Correlate sentiment with sales data to understand the impact on business performance

AAAAAAAAAAAAAAAAAAAAABKIpgEAAAAA15BZ6twc1M2Y1jDyUOEJWkFtrwc%3D0MNMzafx9XVeBmAsYdnFd3ll7qiXa2hdA8HTjHK3WKKByXwH6L Bearer Token

1654876589992165378-ljTQ9A4KYNS3iMYvKJ6PeYhth9L5BN Access Token

OSovSOAeP2cXjdIKPvyTdfihbktQQNuELiwzgbIsU6yVM Access Token Secret

r1vUV774ooNgLradEKQ5Ft4cV – API key

jgNGyg55czNZWYXXEOgA6gEM29CUjZ3T1xOUE4cYpV2Ja1v9cr API Key Secret