Data Engineering Project: House and Terrain Cost Prediction Database

Project Overview:
We're excited to invite you to participate in a data engineering project as part of your application process for the Data Engineer position at InDatalytics. This project is designed to assess your skills and creativity in designing and building a database system for predicting the cost of houses and terrains in Nuevo Leon, Mexico. It's a fantastic opportunity for you to demonstrate your abilities in a practical data engineering scenario.

Project Description:
Your task involves designing and implementing a comprehensive database system that facilitates the analysis and prediction of housing and terrain costs in Nuevo Leon. The main objective is to create a versatile platform that empowers accurate cost predictions based on various property features, including location, size, amenities, and market trends. The emphasis is on developing a scalable, efficient, and organized database structure to accommodate the anticipated growth of data.

Project Steps:

Data Collection and Cleaning: Begin by sourcing relevant datasets containing information about houses and terrains in Nuevo Leon. These datasets should encompass property details, historical price data, geographic information, and pertinent socioeconomic indicators. Rigorously clean and preprocess the data to ensure uniformity and data quality.

Database Design: Devise a relational database schema that adeptly stores and organizes the collected data. Define relationships that exist between different entities, such as properties, locations, market trends, etc. Additionally, specify the suitable data types and constraints for each attribute.

ETL Process: Develop a versatile ETL (Extract, Transform, Load) process to populate the database with the cleansed data. Consider automating the ETL process to accommodate regular data updates and maintain data accuracy.

Query Optimization: Construct SQL queries that facilitate data scientists in retrieving pertinent information for the purpose of cost prediction analysis. Prioritize query optimization to ensure the efficient retrieval of data.

Scalability and Performance: Take into account the database system's ability to manage escalating data volumes. Implement strategies such as indexing and partitioning to uphold performance levels as the dataset expands.

Documentation: Create a comprehensive documentation package that elucidates the database schema, ETL process, and other pertinent details. Ensure that the project's components can be comprehended and upheld by fellow team members.

Software and Language Flexibility:

You are encouraged to leverage any software tools, technologies, and programming languages that you deem suitable for this project. The focus is on delivering a robust and functional solution.

Submission:
Kindly submit your completed project as soon as possible. You have the option to share your work via a GitHub repository or by providing a compressed file containing all the requisite code, documentation, and sample queries. Additionally, you can include any supplementary notes that enhance the clarity of your work.

Should you have any questions or require further clarifications about the project, please do not hesitate to get in touch with us.