Data Engineering Platforms
MSCA 31012

FINAL PROJECT

Objectives

- End to end process of gathering, preparing and storing data in a relational database
- Analyzing data and creating reports and dashboards to enable business decision making
- Articulation of business case and team collaboration on data preparation and analysis

Project

The goal behind the final project is to 'put it all together' by developing a coherent, concise, and realistic analysis in the form of a report and presentation to an executive audience (your client). The project will provide you with the opportunity to apply your knowledge and understanding of data collection, storage in a relational database, analysis and visualization, by identifying a dataset, analyzing the data, and providing recommendations to your client.

The project report should contain the following sections and be written for the intended executive audience in mind:

- Executive summary
- Research objective(s)
    - The problem to be solved and data you plan on using
- Methodology and various tools used in the process
    - At least 5 relational database tables (3NF+), EER diagram
    - Additionally you may also use a No-SQL database to store some of the data
- Data analysis and Visualization
    - Insights using at least 4 moderately complex SQL queries
    - Insights using at least 4 reports and a dashboard/storyboard
- Recommendations
    - Corrective measures and scope for improvement
- Lessons Learned

Data

Students may use any moderately sized public dataset (see samples below):

- https://data.cityofchicago.org/
- https://opendata.cityofnewyork.us/
- https://data.gov.in/catalogs
- https://github.com/awesomedata/awesome-public-datasets
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/

Project Timelines

- Week 1: Form project teams
- Week 2: Research and socialize project ideas and scope
- Week 3: Finalize project data sources and datasets
- Week 4: Create conceptual and logical ER model
- Week 5: Finalize and review EER model
- Week 6: Complete loading data into relational (and non-relational if any) schema
- Week 7: Create and test SQL queries
- Week 8: Create visualization and dashboards
- Week 9: Present findings and recommendations
- Week 10: Present findings and recommendations
- Upload all submissions to course portal within 3 days of final presentation

Submissions

- Students will work in teams of 3 to 4 members
- Teams may be asked to present their projects in the final sync session through video/screenshare
- Following artifacts to be subitted as a single submission per team:
    - Enhanced Entity Relatonship (EER) model (sql workbench file or screenshot)
    - SQL script file containing all important analysis queries
    - Visualization Dashboards/Reports – Tableau, Excel or PowerBI, etc. (raw files)
    - Final Presentation slides (as pptx or pdf)

Grading Rubric

The final team project accounts for 40% of your overall grade, and project grade will be determined based on:

- Business Use Case - 20%
    - Understanding the business problem and articulating projects goals
- Data Preparation - 20%
    - Data extraction, cleaning, normalization
- Relational Modeling - 20%
    - Conceptual and logical ER model
- Relational Database Implementation - 20 %
    - EER, database creation and data loading
- Reporting and Visualization -20%
    - Insights gained through SQL analysis, reports and dashboards