

DATA ENGINEERING PLATFORMS (MSCA 31012)

sbharadwaj@uchicago.edu | salerno@uchicago.edu | jchan530@uchicago.edu

Objective

- End to end process of gathering, preparing and storing data in databases
- Collaborating with the team on business use case, data preparation and analysis
- Connecting to databases, analyzing the data and deriving valuable insights
- Creation of reports and dashboards based on the business case to communicate insights

Project Timelines

- Week 2: Form project teams, research and socialize project ideas
- Week 4: Define scope and finalize project data sources and datasets
- Week 6: Create the data models and get started on the data preparation process
- Week 8: Iterate on data loads, automated data pipelines along with definition of insights
- Week 10: Upload artifacts for grading

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 or a non-relational database, analysis and visualization, by identifying datasets, analyzing it, and providing recommendations to your client.

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

- Executive summary
- Research objective(s)
 - The problem to be solved and datasets you plan on using
- Data Ingestion, analysis and preparation
 - Gathering data and preparing it for storage and Analysis
- Methodology and various tools used in the process
 - Evaluation of analytical or transactional data stores for the use cases
 - Automation methodology for the End to End pipeline.
 - At-least 5 database tables, ER diagram, normalization, use of SQL queries
 - Analytical: Star/snowflake data model (as applicable)

- ETL: Scripts relating to data ingestion and transformations and/or converting relational to star/snowflake schema (as applicable)
- Applicability of a document database, graph database for the dataset and the corresponding data model
- Reporting
 - At least 5 reports along with a dashboard.
- Recommendations
 - Design considerations including reasons for choosing one data store over the other.
 - Corrective measures and scope for improvement
- Lessons Learned
- References

Data

Students have the flexibility to can use any public dataset. The following URLs can also be used to refer for additional datasets

- Enron emails dataset (<https://www.cs.cmu.edu/~./enron/>)
- <https://pushshift.io/kavanaugh-twitter-dataset/>
- <https://toolbox.google.com/datasetsearch/>
- <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/>
- IRI Dataset
 - NDA and data dictionary available at Modules > final Project > datasets > IRI
 - NDA signed by every member of the team and sent to gguevara@uchicago.edu
 - Once permission granted, login to midway. Data located at /project/databases/IRIData

Submissions

- Students will work in teams of 2 to 4 people.
- Single submission per team.
- Following artifacts to be submitted as a single submission per team in canvas:
 - Enhanced Entity Relationship (EER) model (sql workbench file or screenshot)
 - All scripts file(SQL/Python/R) containing all analysis/modeling queries
 - Visualization Dashboards/Reports – Tableau, Excel or PowerBI, etc. (raw files)

- Final Presentation slides (as PPT)

Grading Rubric

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

- Business Use Case - 10%
 - Understanding the business problem and articulating projects goals
- Data Ingestion, Analysis & Preparation - 25%
 - Data Ingestion, cleaning, transforming to the target structure
- Data Modeling & Design – 25 %
- Tools / Database concepts – 20%
- Design of Dashboards/reports to communicate insights - 20%