Data Engineering Project A↓

1. Project Description

The aim of the project is to offer you hands-on experience for an end-to-end data engineering project. For this, you will synthesize and apply prior and new knowledge gained in this course in technologies, tools, and methods used in the data pipeline. More specifically, you will work on the course's custom pipeline comprising open-source tools and commercial platforms to complete a data engineering project.

There will be teams of 1-3 students. Project ideas and datasets will be provided. Of course, you may bring your project idea and dataset; you just have to properly document its goal and outcomes and discuss it with us.

Given the plethora of projects, datasets, approaches, methods, and other related resources that you can use, the challenge is to remain focused on your project, to finish it in time.

In this project we will be using Python; please be mindful of any proprietary or privacy issues regarding the data, as you would need to submit everything (code + data) along with your final report.

Project Management and collaboration have a pivotal role in this activity. The project timeline has milestones and deliverables acting as checkpoints.

Finally, the project should have a GitHub repository. I encourage you to track all your code changes by pushing them to your GitHub repository throughout the semester, but this is not mandatory. Assistance will be provided if needed.

At the least, the final version of the project should include all the code along with a "README" file of your project describing its function, dataset used, architecture, and final results in your GitHub repository before the submission date. It should be in a state so that for anyone to replicate your project on the pipeline provided by this class, they would just need to download the final GitHub repository.

1.1 Project Ideas and Datasets

This project gives you the opportunity to work with Data Engineering methods and tools while building an application that interests and excite you.

The progress of the project will follow the course's modules. Here are some project ideas (along with the instance pipeline they follow).

Pipeline: Batch-ML-Visualization

1. Create a movie genre prediction model
   Dataset: **https://www.kaggle.com/code/tillmandegens/movie-prediction-project/notebook** ⤓ **(https://www.kaggle.com/code/tillmandegens/movie-prediction-project/notebook)**
2. Predict number of COVID-19 related deaths and graph the predicted deaths in the future
   Dataset: **https://github.com/nytimes/covid-19-data** ⤓ **(https://github.com/nytimes/covid-19-data)**
3. Predict the future of Rice, Wheat or Corn Prices and graph the predicted price(s) in the future
   Dataset: **https://www.kaggle.com/datasets/timmofeyy/-cerial-prices-changes-within-last-30-years** ⤓ **(https://www.kaggle.com/datasets/timmofeyy/-cerial-prices-changes-within-last-30-years)**
4. Predict Job title based on Job description
   Dataset: **https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor** ⤓ **(https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor)**
5. Predict the length of an Audiobook and compare them to other works by the same narrator
   Dataset: **https://www.kaggle.com/datasets/snehangsude/audible-dataset** ⤓ **(https://www.kaggle.com/datasets/snehangsude/audible-dataset)**
6. Predict the title of a movie based on the plot
   Dataset: **https://github.com/markriedl/WikiPlots** ⤓ **(https://github.com/markriedl/WikiPlots)**
7. Detect whether a transaction is false or not and plot the number of faulty transactions compared to legitimate ones
   Dataset: **https://www.kaggle.com/datasets/kartik2112/fraud-detection** ⤓ **(https://www.kaggle.com/datasets/kartik2112/fraud-detection)**
8. Build a model to Categorize YouTube videos based on their comments and statistics.
   Dataset: **https://www.kaggle.com/datasets/datasnaek/youtube-new** ⤓ **(https://www.kaggle.com/datasets/datasnaek/youtube-new)**

Pipeline: Stream-Visualization

1. Visualize real-time traffic incidents based on location
   Dataset: **https://developer.tomtom.com/traffic-api/documentation/product-information/introduction** ⤓ **(https://developer.tomtom.com/traffic-api/documentation/product-information/introduction)**
2. Visualize the live scores of multiple football matches
   Dataset: **https://www.api-football.com/** ⤓ **(https://www.api-football.com/)**

Note: You can of course use your own project idea. In this case, you can email me to discuss it in advance. Be mindful of the Python version. Usually, it is advised to use the latest stable version available, however, certain libraries might have specific Python version dependencies.

1.2 Project Deliverables

- <mark>October 8</mark>: Project Proposal due [10%]
- <mark>October 29</mark>: Project Milestone due [20%] (covers modules 1-5)
- <mark>December 8</mark>: Final project due (report pdf and GitHub Repository) [70%]

2. Deliverables

The project progress will follow the sequence of the course modules and the project implementation & delivery activities:

- Define the scope of the project
  - Choose the dataset
  - Choose the pipeline
  - Set-up the pipeline
- Explore, assess, and transform the data, as needed
- Choose your analysis model, as needed
- Run the pipeline
- Show the results
- Document the Project

2.1 Project Proposal

Submit a pdf description of your proposed project (about 1-2 pages). Include the following sections:

- Title: What is your project title?
- Purpose: What problem are you trying to solve? What questions will your project address? How will the project show/visualize the results?
- Pipeline and Methods: Which pipeline will you use? What dataset will you use? Depending on the pipeline, what methods and tools will you use?
- The final result you are expecting to achieve

2.2 Project Milestone (covers Modules 1-5)

This milestone acts as a progress marker; the project progress should follow the sequence of the course modules. Submit a pdf report <mark>(around 3 pages)</mark> written as an *early draft* of your final report, including:

- What have you done so far?
- What pipeline, project, and dataset are you using?
- Does your data need data cleaning or preprocessing? If so, what?
- Will you perform Exploratory Data Analysis? Which methods are you going to use?
- What information about data provenance have you listed? Answer the characteristic data provenance questions addressed in Module 5
- Are there any untested assumptions or other reasons that would prevent you from completing your project?
- What are your next steps?
- What is left to do?

2.3 Final Deliverable and Project Report

GitHub: (example: https://github.com/ankurchavda/streamify ▱ (https://github.com/ankurchavda/streamify))

The project should have a GitHub repository. I encourage you to track all your code changes by pushing them to your GitHub repository throughout the semester, but this is not mandatory. It should be in a state so that for anyone to replicate your project on the pipeline provided by this class, they would just need to download the final GitHub repository as a zip file and should be able to run it.

Readme:
The readme should describe (no more that 3-4 lines for the following topics)

Title: What is the title of your project?

Project's function: *This is an overall description of your project:* What is objective of the project? What is the problem you are trying to solve?

Dataset: Briefly describe your dataset

Pipeline / Architecture: Which pipeline did you use? Which tools?

Data Quality Assessment: Describe the quality status of the data set and the way you assessed it

Data Transformation Models used: Briefly describe the transformations and models used

and final results that you were able to achieve. If there are any special instructions needed to execute your code (e.g., signing up to a specific API to access the dataset that is needed) those need to be listed as well.

Infographic: A simple infographic describing the architecture of your data pipeline including datasets, storage, and tools used along with another final infographic describing the results of the engineering task accomplished. Examples can be provided if needed.

Code: A link to GitHub Repository

Thorough Investigation: *This critically assesses the viability of your idea:* Based on the results of this project (your pilot project, your prototype, etc), from a technical leadership point of view, what are your conclusions or recommendations for continuing this project in terms of scaling it up? How would you assess the innovativeness of your project? Any technical or platform concerns, difficulties, or limitations of the pipeline for the project? Based on your experience and results, what next step would you recommend to take this project to the next level/phase?