Project requarements

The project can be done in pairs(group of 2 people) or individually. It should have the code itself and presentation in video format.

Stages of the project:

1. **Find a source of data. Ideally, several sources should be combined, but in the case of public datasets - it is quite hard to find several related to each other, so a single dataset can be used but not recommended. In addition, to fetch data into the system, different scrappers can be used(e.g., Beautiful Soup can be used to parse data from eBay, or Twitter API can be used to fetch data from it). As a last resort case, the mock-generated data can be used.**
2. **Define the Data model that will be used. Describe it. And according to it - load data to the system.**
3. **Create a consumption app for the data in the system. It can be done in one of 3 ways:**
    1. **The report shows some insights from data (Superset report of properly formatted Jupiter notebook)**
    2. **Web server. It requires an additional pipeline, which will push data to the OLTP database, which will be used by the web server.**
    3. **In case you have knowledge in the AI/ML area - train the model on top of collected data. Depending on your case - the model can produce new data as well as enrich existing data (i.e., an unsupervised training model can flag a data record anomaly or not, or a regression model can build forecasting for new data)**
4. **Create a presentation and present a project. It should include:**
    1. **What data did you pick, and why**
    2. **Data model**
    3. **How do you ingest data?**
    4. **How you transform data**
    5. **How data consumed**

Submitting project:

The project should be submitted before the penultimate lecture. Submission should include a PDF with slides, video speech, source code, and a way to run it(ideally with a make file, but it can be just plain text instruction or both). During the dedicated lecture/lab slot, teams in order of natural queue will present their work(max 5 minutes) and answer questions (2 minutes). For teams who will not have enough time during the timeslot to present - grading will be done offline(according to the submitted video presentation).

Grade mechanism:

?

15% what source was selected, why.

60% software implementation

25% presentations

Places where public datasets can be found:

- **Google: https://datasetsearch.research.google.com**
- **Kaggle: https://www.kaggle.com/datasets?topic=trendingDataset**
- **Open data from EU https://ec.europa.eu/eurostat**

Project examples:

- **Use dataset https://www.kaggle.com/datasets/nnekaekwemuka/fitbit-fitness-tracker-dataset and visualize aggregation for each axis: step number, calory, etc. Show splitting by hours and day of week**
- **NSE-stock-market-historical-data https://www.kaggle.com/datasets/bhaktij/nse-stock-market-historical-data - show the most volatile stocks, stocks with best dividends. Or find data to enrich stock with company information as a dividend and show industry with best devident**
- **Brazilian e-commerce company: OLIST https://www.kaggle.com/datasets/erak1006/brazilian-e-commerce-company-olist find a product with biggest sales, most reviewed product(ratio between reviews number and order number)**
- **Investigate StackOveflow data and build some reports on top of it (https://www.kaggle.com/datasets/stackoverflow/stackoverflow): which tags receive more comments, tag popularity per year, etc.**

Ostatnia modyfikacja: środa, 7 maja 2025, 15:56

Pobierz aplikację mobilną