Overview PF --> Students

Final Projects

About the final project

This stage involves simulating a real working environment. The goal is to enhance both technical and soft skills, promoting teamwork both independently. The group must coordinate, leverage individual strengths to achieve synergy, and progress in the development of one or more data products, responding to the requirements set by a Product Owner, based on a product proposal that must originate from the group.

As a starting point, the group of students must analyze the theme and available data to formulate a project proposal to develop. The proposal they make will be accepted or modified by the PO, who is also responsible for supervising progress, redirecting efforts if necessary (weekly).

In addition, throughout the development, groups will have daily guidance from an HM, who can provide support on project generalities and, eventually, technical advice. However, coordination, decisions, and task execution are entirely the responsibility of the team itself.

Milestones and deliverables

Each week of project development, there will be a set of tasks to perform, many of which you will have to define based on your objectives. However, some of these tasks (including all possible subtasks) will be critical to move on to the next week/stage of the PF. For example, without completing the full ETL, the Analytics stage cannot begin. For this reason, we identify the completion of the ETL, which we know includes many subtasks, as a milestone for the first week.

Some tasks result in products, whether tangible or intangible, providing an indication of the completion of that task. These products are what we identify as deliverables.

In summary

Milestones can be defined as the actions that, once completed, provide an indication of the degree of progress in achieving the objectives. While deliverables are the product of the completion of specific tasks throughout the project and can be tangible or intangible.

Important

Documentation is a fundamental item in the development of any data project, and it is expected that throughout the entire PF, you provide documentation for each stage. It should be as detailed and clear as possible. Each week will include specific elements to include in the documentation, but in general, you should always document what you did, with what tools, and justify all decisions regarding the project.

Sprint #1: Project Kickoff and Data Work

In this week, you must analyze the selected project and available data. Based on the understanding gained from the theme, you should propose how to approach it, providing a solution or tools developed by yourselves to approach this solution.

This proposal should include the following items:

Current situation understanding

The proposal should demonstrate a proper handling of the issue, you should be able to contextualize it and express possible analyses/solutions around it.

Objectives

Objectives should be concrete actions (verbs) that clearly describe what you aim to achieve with the project. Develop, create, do, etc.

Scope

Themes are often broad and may allow much more comprehensive treatments in scope and magnitude than can be achieved during the project's development. Therefore, you should delimit your work by defining the scope and tasks/developments that may be considered important for the project's integrity but are out of reach due to complexity or time. You can propose these as possibilities for the project's continuation.

Associated Objectives and KPIs (proposal)

From the understanding of the issue, questions will arise that you will seek to address with the work or tools developed. These questions, formulated as objectives, will allow the creation of KPIs to evaluate their achievement. This is a comprehensive and specific task concerning both the issue and the chosen approach.

Github Repository

Set up a Github repository to work collaboratively with the entire group. It must be public so that both the mentor and the Product Owner can view it. You will have to carry out different branches and version controls of your own work.

Proposed Solution

You must detail the tasks you will perform to achieve the previously proposed work objectives and how you will do it (work methodologies, organization method, task distribution, roles of each team member, etc.). Also, detail the products that will arise

from your work and at what stage you will present them, taking into account the general requirements (expected deliverables) for each stage of the project.

You must also make a time estimation for each task, considering overall execution times and the milestones planned for each week, and present that estimation in a Gantt chart. A crucial part of the proposed solution is with what tools (technological stack) you will build the project architecture. For this, you will have to select a small portion of the available data and perform a cleaning and transformation process using the tools you plan to implement. This will give you an idea of how they will work in the complete project and allow you to have a better approach for future tasks. Keep in mind that, as this item will be a preview presentation of what you will work on in the second sprint, the PO can give the OK or determine the best path for you to take. This will allow you to advance work for the second week since you won't have to wait until the second demo to verify if the architecture meets the PO's requirements.

Finally, as working with quality data is very important in Data, you must include in your report an analysis of the data you will work with (metadata), detailing it as much as possible: sources and reliability, what each column of each dataset represents, types of data, acquisition method, acquisition date, and last update, etc.

Milestones

4 KPIs

Project scope documentation

Data Exploratory Analysis (EDA)

Github repository

Implementation of the technological stack

Work methodology

Detailed design

Team - Roles and responsibilities

General schedule - Gantt

Preliminary data quality analysis

Deliverables

Documentation

Chosen stack and justification

Workflow

Sprint #2: Data Engineering

In the continuation of the first week, you are expected to work on setting up the infrastructure of your project, with pipelines to perform the ETL process aiming at Data Warehouse, Datalake, or Datalakehouse structures, considering incremental data loading.

You must use big data tools and/or cloud services of your preference. If the group consists of five people, its use is mandatory. If there are fewer members, they are also expected to use these tools. If necessary, the PO will indicate whether the project's architecture is in line with expectations. For example, working with data solely using Python and storing it locally is not acceptable.

In the case of using relational models in their storage structures, they must deliver an appropriate and detailed design of the entity-relationship model, specifying tables, relationships, and adopted data types. If they are going to use non-relational models, they should explain why they consider their implementation over other models, always supporting the decisions they make.

Continuing the idea of advancing work, as in sprint 1, in this second sprint, students will have to conduct a comprehensive analysis on a representative sample of data. This may include identifying outliers, variable distributions, and preliminary correlations. This will give them an idea of the characteristics and peculiarities of the complete data, and feedback can be provided.

Dashboard Design (Proof of Concept):

Create a simplified version of the dashboard planned for implementation, including the connection with DW. Incorporate some preliminary visualizations and sample data. This will allow you to explore the interaction and structure of the dashboard before working with complete data. Again, the PO can give approval before the final demo (after which no further modifications can be made based on recommendations).

The other alternative would be ML Products:

Before implementing interactive products using Streamlit (or the chosen tool), create test versions with sample data. This will help identify potential usability and functionality issues before presenting the final version.

Milestones

Complete ETL

Implemented data structure (DW, DL, etc.). You can use any service.

Automated ETL pipeline

ER Model Design

Pipelines to feed the DW

Data Warehouse

Automation

Data validation

Documentation

Detailed ER diagram (tables, PK, FK, and data type)

Data dictionary

Workflow detailing technologies

Analysis of sample data

MVP/Proof of Concept of ML product or MVP/Proof of Concept of Dashboard

Deliverables

Documentation and report produced

Sprint #3: Data Analytics + ML

In the last week, you are expected to create an interactive dashboard, along with an analysis of the data you have worked on. Include the KPIs that you determined as important for the performed analysis and prepare a storytelling with it.

For groups of five people, or if the project requires it, it will be mandatory to have at least implemented models, and work is starting on an ML product.

Final Demo: Final Considerations

*The final demo is on the Friday of the third sprint.

For the last demo of the Final Project, it is expected that students arrive with their projects completed, implementing the final touches/features that students deem necessary or that the PO has requested. It is important to allocate a part of the last week to work on presentations, storytelling, verify that everything works correctly, and complete documentation. Repositories should be well-organized, and with a comprehensive readme presenting the project.