Business Intelligence I

Fall Semester 2024/2025

Project Assignment Handout

This handout details the rules for the mandatory practical project for Business Intelligence I, to be developed and completed during the academic calendar of the BI I class.

Project Summary

The project for this class aims to reinforce both the conceptual and practical knowledge to be acquired from your BI course. As such, you are required, as part of a group, to design, implement and explain (in a report) a **fully working Data Warehouse solution**.

Note that a Data Warehouse is only one of the major components of a complete Business Intelligence solution; should you go onto the follow-up class in the next semester (BI II), you'll get the chance to build the remaining components (analysis, reporting and dashboarding layers) and finish off your BI solution.

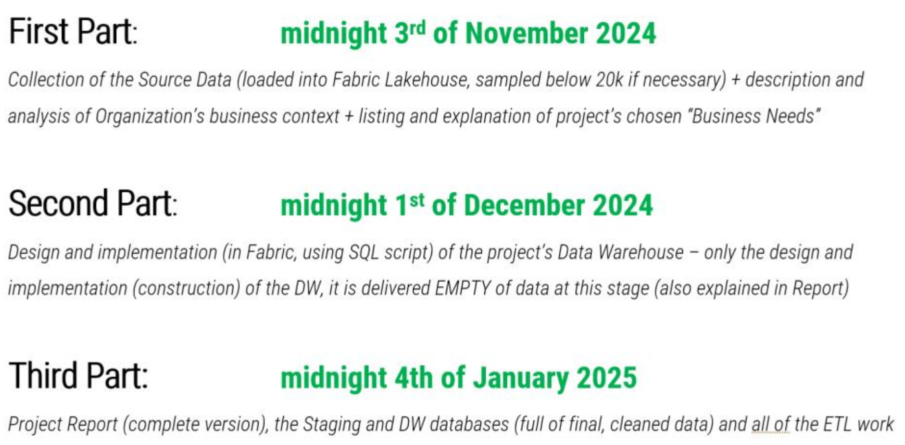
In this practical project, each group is expected to develop a proper Data Warehouse (using Microsoft Fabric Data Warehouse) and then implement a varied set of Extraction, Transformation and Loading processes (using Microsoft Fabric Data Engineering tools) to "clean" and "shape" the source data from transactional (OLTP) operational systems and load the project's data warehouse with properly structured data. Your project report will explain and detail the work that your group carried out.

Group Rules

The project should be done in a **group of between one (1) to four (4) students**; we consider this the ideal size for group work. Groups larger than four will NOT have project graded.

You are allowed to have group members from different BI I classes, including a mix of students from Daytime and Nighttime classes.

Important Project (partial work) Delivery Dates:



First Part: **midnight 3rd of November 2024**

Collection of the Source Data (loaded into Fabric Lakehouse, sampled below 20k if necessary) + description and analysis of Organization's business context + listing and explanation of project's chosen "Business Needs"

Second Part: **midnight 1st of December 2024**

Design and implementation (in Fabric, using SQL script) of the project's Data Warehouse – only the design and implementation (construction) of the DW, it is delivered EMPTY of data at this stage (also explained in Report)

Third Part: **midnight 4th of January 2025**

Project Report (complete version), the Staging and DW databases (full of final, cleaned data) and all of the ETL work

Project Starting Point – The Source Data

We strongly encourage each group to choose a real-life business/organizational problem that can be used as the foundation of this practical project. In most real-life cases, BI projects usually involve large sets of source data, from where you'll be extracting the relevant data for your project. Typically, this source data is obtained from a relational (transactional or operational) database, but this is not required for your project – most of the projects we see start from "flat" (txt or csv) files obtained from online datasets.

We expect that your source data should contain, at the very minimum, the following features (some of these will become clear as the course progresses, and you become more familiar with business intelligence terminology):

i. Some sort of transactional records that can represent quantifiable facts, to be used in your future data warehouse; these need not be extensive in volume, but should represent a few thousand records of business transactions, and must require some sort of "cleaning" or "normalizing" actions during the ETL process.

ii. Enough attributes (characteristics) so that you can extract at least five dimensions from the source data, according to the dimensions criteria for the project.

iii. If your source data comes from a relational database, you will have to extract the data from whatever tables you deem necessary and save this data into "flat" files (delimited text data files), so they can be loaded into Microsoft Fabric (cloud lake).

iv. The quantifiable data that will make up the facts of your data warehousing solution should represent several years of transactional history (at minimum, more than one year), so that analysis can be done across multiple years.

v. The characteristics that you'll be setting up as dimensions of your DW model should provide you with the ability to set up different hierarchies in the different dimensions ("*what is an hierarchy?*" will be explained and demonstrated in class).

vi. The submitted solution, to be shared with all Lab teachers, must be self-contained inside Microsoft Fabric, not requiring any additional data or tools to run.

What to do if your source data does not meet the minimum criteria above?

In essence, your group has two choices:

- Rework whatever source data you already have, so that it meets the above criteria. Note that this will not count for project credit, so we are not concerned in how you achieve this – you can use any tools or techniques to fix and improve your source data (your lab teacher will try to suggest possible strategies to help you fix your source data, but the responsibility lies entirely with the Group).

- Your teachers will point you to public datasets that are freely available online, and from which you can obtain some sort of acceptable starting data source; however, you will still be required to "imagine" and develop a problem story that fits this data, to serve as the "original data problem" that your BI solution will address.

Expected Deliverables

Your group must build and deliver a Data Warehouse solution that contains the following items:

i. A Data Warehouse, using a dimensional model (Star Schema is mandatory).

ii. The Data Warehouse must feature:

  a. At least 5 dimensions, one of which must be a Date dimension.

  b. At least 5 hierarchies (average depth of three levels when all considered).

  c. At least one Fact table, featuring at least 2 distinct primitive measures.

  d. At least one Fact table must have a minimum of 5 thousand rows (the only exception will be for real-life projects, and subject to teacher approval).

iii. You are expected to employ as source data:

  a. A varied set of source data, to come from transactional databases, different types of flat files (text, comma-separated files, excel, etc.), cloud hosting or some other "usable" sources.

  b. Your Source Data must not have any individual dataset (dimension or facts) larger than 20 thousand rows (teachers will help sample down).

  c. This source data must be able to be "transferred" to delimited text-files (typically, CSV) so as to be ingested into Microsoft Fabric data lake.

iv. You are required to develop and use a Staging Area:

  a. The staging area should be developed as a different database than the Data Warehouse.

  b. The staging area database may employ tables and columns that are not part of the final DW model, and are used to make the ETL processes easier and more efficient to execute, or to check data quality.

v. A set of ETL (Extract, Transform and Load) processes:

  a. All ETL processes are to be developed using Microsoft Fabric Data Engineering functionality and should follow best practices seen in Labs.

  b. There should be two set of independent (different) ETL processes: one process to load Staging Area, and another to load Data Warehouse.

  c. As much as possible, the ETL processes should employ a varied set of techniques and transformations, reflecting the different aspects learned in the laboratory sessions.

vi. In addition to these deliverables, you are expected to provide a complete and detailed project report that clearly describes all the work the group carried out.

IMPORTANT NOTE: Please do NOT USE as your source data a database, dataset or some other type of data that is usually employed in training / teaching / demonstration settings, as these typically have Data Warehousing solutions already available. For example, do not use AdventureWorks, Northwind, WorldWideImporters, Chinook, Sakila, etc.

Detailed Evaluation Criteria for Project Delivery

1. Report – the report is worth **3 points** (out of 20) of the final project grade, in accordance with the following criteria:

   1.1. Structure: *is the report developed with an appropriate and proper structure, reflecting the recommendations from this Handout?*

   1.2. Content: *is the report content clear and objective? Does it meet with expected guidelines for a professional BI project report? Is it efficient (not too much, nor too little) and is it effective (does it clearly explain the work done by the Group)?*

   1.3. Analysis of project work: *did the group outline the methodological choices made throughout the project? Did they describe problems encountered, as well as the solutions developed to overcome those problems? Did they justify the dimensional model adopted for the Data Warehouse solution, in view of the organization's original problem?*

   1.4. Conclusion: *are there concluding points, as well as a critical assessment of the project?*

2. Data Warehouse – the design, implementation and use of the data warehouse is worth **8 points** (out of 20), as detailed below:

   2.1. Source Data: *did the group make available the original source data that was used? Is it in "usable" format? Is it documented in the report? Are all the databases, files and/or connections also available and documented?*

   2.2. Dimensional Model: *Is the dimensional model used for the Data Warehouse correct and properly developed (is the Star Schema correctly designed and implemented? Is it adjusted for the Group's specific Business Needs)?*

   2.3. Staging Area: *Is the group employing a Staging Area database? Is it properly setup so as to make the ETL processes easier and more effective? Is the use of the Staging Area documented and justified, in the report?*

   2.4. DW Fact table: *is the Data Warehouse's Fact table correctly designed and developed? Are the measures adequate, properly designed and configured? Are the relationships with dimensions properly configured/verifiable?*

   2.5. DW Dimensions: *are the Data Warehouse's Dimension tables correctly designed and developed? Are the attributes properly designed and configured? Do the dimensions reflect the necessary attributes to enable the required hierarchies to be employed at a later stage?*

3. Extraction, Transformation and Loading (ETL) Processes – the ETL processes of the project are worth **8 points** (out of 20), as detailed below:

   3.1. Documentation: *Is the ETL adequately documented and explained in Report? Preferably, in a visually rich and descriptive manner, making clear the ETL approach taken by the group.*

   3.2. General ETL Organization:

      - *Has the workspace been delivered after a successful execution of the entire ETL process, with all STG, DW and Data Quality data fully loaded into the tables?*

- *Are the general ETL artefacts (pipelines, activities, dataflows, etc.) in Fabric properly organized and correctly named (as per Labs best practices)?*
- *Has Group met requirement in terms of mandatory use of BOTH Copy Data and Dataflows: at least ONE (1) of each MUST BE USED anywhere in the ETL.*
- *Has the Group organized their ETL into separate and properly segmented Fabric Pipelines, using distinct pipelines for different ETL objectives (i.e. STG, DW, etc.)?*

3.3. Data Quality:

- *Is the Group correctly employing Data Quality Validation checks within their ETL, fulfilling, AS A MINIMUM, the FOUR (4) Data Quality checks taught in Labs?*
- *Is the Group's ETL correctly registering and storing Data Quality validation checks into a LOG table, as specified in Labs and kept in the project's STG AREA?*

3.4. Data Transformations and Cleansing:

- *Is the project correctly employing a varied, non-trivial and technically-correct range of data transformations (e.g. cleansing) through the use of Dataflows?*
- *Are the transformations described above useful and appropriate to the project's needs? Do they make sense in relation to source data and DW design / tables?*

3.5. ETL Execution:

- *Is the STG's ETL design and implementation correct? Executes both preparatory and actual ETL tasks? Run in effective and efficient manner? Covers all STG needs?*
- *Is the DW's ETL design and implementation correct? Correctly executes preparatory tasks, SKs, lookups and complete transfer of the STG data into DW?*

4. Extra ETL (Original work) – worth 1 point (out of 20), based on complexity and originality of the extra (creative) work:

- Extra (original / creative) ETL development # 1
- Extra (original / creative) ETL development # 2
- Extra (original / creative) ETL development # 3
- Extra (original / creative) ETL development # 4 (max of 4; more ignored)

Report Structure

The structure of the project report should follow, as much as possible, the following structure:

i. Presentation of business / organization / problem scenario

   a. Presentation of the organization, providing readers (teachers) with a general context of what organization is being used, what it does and why it is a suitable candidate for the Group's BI project.

   b. Explanation of the informational problem that this organization is facing; in other words, what (general) business problem is the BI project solving?

   c. Identification of Business Needs; a more detailed – and very objective – listing of the specific business questions that are going to be driving the development of the Group's dimensional model (star schema in the DW). Please list the BN's but also provide a brief and objective explanation of each one.

   d. Typically, we expect this section to take around 3 to 5 "normal" pages of your Report, but varies considerably, depending on presentation choices.

ii. Original Data Sources

   a. Description of the structure and data in the organization's source data, as well as of any additional supporting flat files developed by Group.

   b. It is far more important to explain, in general terms, the source data than to simply provide a list of the original data fields.

   c. Typically, we expect this section to take around 1 to 2 "normal" pages of your Report, but varies considerably, depending on presentation choices (any detailed lists must go to the back of the Report, as Appendices).

iii. Staging Area

   a. Description of the development of the Staging Area and why it is used.

   b. Typically, we expect this section to take around 1 to 2 "normal" pages of your Report, but varies considerably, depending on presentation choices.

iv. Data Warehouse

   a. Presentation and description of the DW developed, including the methodology employed in its design, as well as how it serves to meet the business needs; must include a clear diagram of the DW, of course!

   b. More important than repeating "useless" (will get you zero grade points) Data Warehousing or "dimensional modelling" theory, is that you provide a clear explanation of how your choice of facts and dimensions make sense for your organization, and its informational problem.

   c. Typically, we expect this section to take around 2 to 6 "normal" pages of your Report, but varies considerably, depending on presentation choices.

v. ETL Processes

a. Description of the development of the ETL processes, including problems and solutions found.

b. This is one section of your Report where "less is more"; it is quite important that you are exhaustive and cover the entire ETL process, but do not simply show, in repeated fashion, routine and common ETL tasks alongside trivial and inconsequential comments…

c. Instead, try to show, using printscreens, diagrams and other visual artefacts, how your ETL design works, why you've designed it this way and where the "complex" (or problematic) parts are present and how they work.

d. Typically, we expect this section to be the largest in the entire Report (anything from 4 or 5 pages, all the way up to 10 or more); however, the best way to expose the ETL work is by using visual imagery, to which you will add strategic and carefully considered commentary, explaining in objective but brief manner each part/component of your ETL pipelines.

vi. Listing of Extra (original) ETL Work

a. Additional and original (innovative) ETL developments that were not demonstrated in the Lab classes; may have been spoken about in Labs, but if not directly demonstrated, these techniques will count as "extra work".

b. This "original" work MUST be detailed in a separate section (make sure to title it appropriately), so the teachers have no chance of missing it when evaluating the Report; if we miss it, you get zero for this work.

vii. Critical Review / Lessons Learned

a. A short summary of what lessons and experience the Group, as a whole, took from this Project; Group should describe main difficulties along with positive lessons learned from having done this project.

b. MOST IMPORTANT: this is separate from the Conclusion, as it covers what the Group students learned from doing this project… What "life lessons" about BI are you going to take with you, having done this project?

viii. Conclusions

a. A short summary of the business value that has resulted from the development of this BI project; Group should describe, from the perspective of their Project organization, the business (additional) value that is now available to the firm, from having successfully built the Group's new BI platform.

Project Delivery Methodology

Your Project will be evaluated across **three (3) partial but cumulative deliveries**.

This both encourages each group to start working on the BI project as soon as possible, but also ensures that you have some means of responding to feedback and (partially) correct and realign your BI journey, in future deliveries.

In each delivery, you MUST BOTH submit part of your Project Report (in pdf format) in the respective Moodle submission section, as well as make available (by sharing with all Lab teachers) your Project Fabric Workspace, containing the work to be delivered:

- Delivery #1 – due at 23h59 of 3rd of November of 2024:
  - In this delivery, you must submit all the work referred to in section 2.1 of the Detailed Evaluation Criteria (in essence, the Source data).
  - You must also submit your Project Report, which should – at this stage – contain sections (i) and (ii) referred to in the Report Structure.
  - At this stage, your Report's presentation and other global aspects (structure, conclusions, etc.) will NOT yet be evaluated; however, the content mentioned above will be GRADED, representing the definitive partial grade for this delivery requirements.
- Delivery #2 – due at 23h59 of 1st of December of 2024:
  - In this delivery, you must submit all the work referred to in section 2 of the Detailed Evaluation Criteria (all of the DW work) EXCEPT the Staging Area (section 2.3); your DW is to be delivered EMPTY, at this stage.
  - You must also submit your Project Report, which should – at this stage – add section (iv) referred to in the Report Structure, to previous delivery.
  - At this stage, your Report's presentation and other global aspects (structure, conclusions, etc.) will NOT yet be evaluated; however, the content mentioned above will be GRADED, representing the definitive partial grade for this delivery requirements.
- Delivery #3 - due at 23h59 of 4th of January of 2024:
  - In this delivery, you must submit all the remaining work referred to in the Detailed Evaluation Criteria; in other words, please complete and deliver anything that has not yet been evaluated in previous deliveries.
  - You must also submit your Project Report, which should now be complete and in its final presentation format.
  - At this stage, ALL of your Report's presentation and other global aspects (structure, conclusions, etc.) will be evaluated.

Please note that there is no REGRADING of Intermediate delivery parts; in other words, the criteria that counts for Intermediate evaluation are only graded once; you can, of course, *change and improve your project after Intermediate delivery, but those parts are NOT REGRADED.*

Applicable for ALL deliveries, in cumulative fashion:

- Failure to deliver on time will incur a 0.5 point penalty for each late day (for example, 4 late days will accrue a 2.0 point penalty), across any of the deliveries (in cumulative fashion).

- Failure to comply with Moodle and/or Fabric delivery instructions (missing student identification, no proper naming of objects, duplicated or unclear files, improper folder configuration, etc.) will meet with a 0.5 point penalty at each delivery where this failure is present and detected.

Questions and Clarifications

Should it be necessary, we will provide further clarifications and answers to questions from students, updating the respective Moodle project forum as appropriate.

Good luck with your project!