Business Intelligence I

Fall Semester 2021/2022

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 is meant to reinforce the conceptual knowledge that you'll be acquiring throughout the 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 and fully-working 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 (OLAP solution and a reporting and dashboarding solution) and finish off your BI solution.

During the course of the practical project, each group is expected to develop a proper Data Warehouse (using Microsoft SQL Server Management Studio) and then implement a varied set of Extraction, Transformation and Loading processes (using Microsoft SQL Server Integration Services) so as to obtain data from original business sources and load the project's data warehouse with clean and organized data. Your project report will serve to describe, in detailed fashion, the work that your group carried out.

Group Rules

The project should be done in a **group of between two (2) to five (5) students**; we consider this the ideal size for group work, and do not allow other sizes.

Should you constitute groups of size other than 2-5 members, you will be subjected to the following penalties:

- For each member in excess or below the stated size (2 to 5 members): penalty of **1.5 points** (out of 20 total points) in final project grade.

For example: should you choose to constitute a group of 7 members, your final project grade will be penalized in 3.0 points (out of a possible maximum grade of 20 points).

The only allowed exceptions that will be made regarding group size will be for situations beyond the control of the group members (for example, one of the students dropping out of the course) and will be evaluated on a case-by-case basis.

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

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 organized in a relational database, but this is not mandatory for your project – you can use other types of data sources (flat files, online datasets, etc.) but it must be in some sort of "usable" format for your project.

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 terms):

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 you provide a relational database, it should be in Microsoft SQL Server 2019 (or compatible) version, or able to be converted to this format by your group (it is up to the group to ensure this conversion).

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).

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.

vi. The submitted solution, to be run by your Lab teacher, must also provide clear and direct access to your source data, without requiring some sort of custom software.

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

In essence, your group has two choices:

- Rework whatever organizational 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 are free to do whatever you feel is more effective, in getting your source data to the required state!

- Your teachers will point you to public datasets that are freely available online, and from which you can "setup" some sort of acceptable starting data source; however, you will still be required to "imagine" and develop a problem story that fits this data, and that will 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 scheme is recommended)

ii. The Data Warehouse must feature:

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

   b. At least 5 hierarchies, with average depth of three levels

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

iii. You are expected to employ as source data:

   a. One principal source database (in SSMS format; having a source database is heavily recommended)

   AND/OR

   b. Other source data, to come from some type of flat files (text, comma-separated files, excel, etc.), cloud hosting or some other "usable" sources

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

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

   a. All ETL processes are to be developed using Microsoft SQL Server Integration Services tool, and should follow best practices seen in Labs

   b. There should be two set of independent (different) ETL processes: one SSIS package 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

   d. At least two different dimensions should employ a "type 2" (or greater) method of dealing with Slowly-Changing Dimensions (SCD)

vi. In addition to these deliverables, you are expected to provide a complete and detailed project report that aptly describes all the work that your 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

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?*

   1.2. Content: *is the report content clear and objective? Does it meet with the expected guidelines? Is it efficient (not too much, nor too little) and is it effective (does it carry its message across)?*

   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 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 (ideal is Star Schema, but other models accepted if correctly justified by the group)?*

   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 it's use 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 properly designed and configured? Are the relationships with dimensions properly setup?*

   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. Are all ETL processes properly documented and explained, and do they tie in properly with the original sources of data provided by the organization? Is the development and methodological choices of these processes fully explained in the project report?

   3.2. Well-organized, simple and clear setup of all ETL processes, within the SS Integration Services tool, and that serve to clean and/or organize the source data.

   3.3. Proper use of containers and other techniques, to organize and better manage the layout and flow of ETL processes.

3.4. ETL processes demonstrate a varied and appropriate set of data transformations, aimed at improving and/or correcting the quality of the source data.

3.5. Preference for the use of shared connections to source data, whenever possible.

3.6. Demonstrate a correct and useful use of variables, within at least one of the ETL processes.

3.7. Development of incremental, rather than integral, loading processes; if integral is required, group is able to correctly justify this option.

3.8. Both the staging area and data warehouse ETL processes work correctly and to completion.

3.9. Other (extra) developments: is there evidence and documentation of extra developments, going beyond strictly what was shown in class? These ETL developments may be worth up to a maximum of 1.0 points (out of 20).

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, making use of statistical data and descriptive aspects
   b. Identification of Business Questions (including the context behind the informational problem that the DW/BI system is going to solve)

ii. Original Data Sources
   a. Description of the structure and data in the organization's source database, as well as of the supporting (auxiliary) flat files

iii. Staging Area
   a. Description of the development of the Staging Area, including reason for using it

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

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

vi. Conclusions and Lessons Learned

Intermediate Delivery: Data Warehouse Design and Implementation

As a means of making sure that all groups are on the right path, in terms of their Data Warehouse design and development work – and to encourage each group to start working on the BI project as soon as possible – we require that each group provides, approximately half-way through the semester, the first part of their BI practical project, namely the Data Warehouse design and implementation (including the report sections that explain and detail the Data Warehouse design).

This delivery will take place through a Moodle submission and the date for submission is the **28th of November** (deadline of midnight, as it'll be announced in Moodle).

This initial delivery will count for all evaluation points for the Data Warehouse part of the project, as outlined in criteria 2.1 to 2.5 above – with the notable exception of item "2.3 Staging Area", which we do not expect to be developed at this stage, and will only be evaluated at the final delivery.

Otherwise, everything outlined in the Criteria up to item 2.5 counts for the intermediate delivery – including this initial part of your Project Report, which must document and explain the work you have done to this point.

Final (Report and ETL) Project Delivery Guide

The delivery of the second part of the project will also be done through a Moodle submission, and the delivery date is always **before midnight of the last Saturday before the exam**. Each group must meet the following requirements in preparing its project deliverables:

- One zip file, named as "**GroupXX.zip**" where XX is the group number (note the English use of 'group' and the fact that there are no spaces)
- In the zip file, there must be the following content:
  - Project Report: "**Report_GroupXX.pdf**" (note pdf format)
  - A folder named as "**1. Source Data**", containing the following file:
    - "**GroupXX_SOURCE_DB.bacpac**" (note the SQL Server format)

    and/or

    - all the source flat files (text, comma-separated, excel, etc.), properly named as used within the project
  - A folder named as "**2. Project Databases**", containing the following two database files: a SQL (text-based script) file for the Staging Area database and another SQL file for the Data Warehouse database, which should be named as "**GroupXX_STAGING.sql**" and "**GroupXX_DW.sql**"
  - A folder named as "**3. SSIS ETL**", containing the Integration Services solution directory, which should be named as "**SSIS_GroupXX**"

- Remember: when the zip file is extracted, it should recreate the 3 folders mentioned above, leaving outside these folders the final project report in pdf format!

Final Notice: **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).

Failure to comply with the delivery guide (no proper naming of objects, duplicated or unclear files, improper folder configuration, etc.) will meet with a once-only 0.5 point penalty.

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!