BSc (Hons) in Information Technology
Specialising in Data Science

IT3021 – Data Warehousing and Business Intelligence
Year 3

Assignment 1

Semester 1, 2021

Complete following tasks and demonstrate the same with SQL Server (any version & edition). Additionally, document the steps followed in completing the tasks. Include the screen shots of the steps you followed with a short description for each step in the report and submit before the deadline.

Step 1: Data set selection (5 marks)

- Select a data set of your choice. The data set should be of a transactional type data set. Not a data set from a data warehouse design.
    - Restrain from selecting a data set which we discussed during practical sessions such as customer/order scenario
    - do not use **AdventureWorks** data sets either
    - Look for a novel scenario
- Ensure the data set you select has at least around one year's data and sufficient number of records and attributes.
- Ensure your data set will have sufficient data to demonstrate following:
    - Multiple sources
    - Data warehouse design
    - A rich set of ETL tasks
    - Enough data to put into SSAS Cubes
        - Hierarchies
        - Dimensions
        - Aggregates
    - Sufficient data to create reports

Documentation: provide a description of the data set you chose. You may use ER-diagrams to aid your description.

Step 2: Preparation of data sources (10 marks)

- Prepare the data set for data extraction. You may consider separating your data into multiples source types.
    - Ex: if you get a customer information along with order information in a sales scenario in a single file, order related details can be separated from the text file and load into one or more database tables and keep customer data in a

text file, so that you have two types of data sources to work with. Customer dimension can be loaded from the text file and order details can be loaded from the database. You can introduce data such as product category to enrich your data to include hierarchies if they are not available in your data set. (This is only an example)

- You should have at least two (2) types of data sources.
    - Example data source types: CSV, database, text, excel, etc.
- Remember, you will need sufficient data to demonstrate all the DWH concepts that you have learnt, in assignment 1 and 2.

Documentation: describe your sources. "What information is available in which format" should be explained.

Step 3: Solution architecture (5 marks)

- Design a high-level DW & BI solution architecture.

Documentation: provide an architectural diagram to describe the components of your DW & BI solution. Provide a summarized description for each component of the solution.

Step 4: Data warehouse design & development (35 marks)

- Design a data warehouse schema (a dimensional model) for the data set you selected in step 1.
    - This may be of star schema or snowflake
    - Ensure you have at least two (2) dimensions apart from common dimensions such as ***Date***
    - Ensure you have at least one (1) fact table
    - Ensure you have at least one (1) slowly changing dimension
    - Implement the data warehouse schema in SQL Server

Documentation: describe your data warehouse dimensional model. Use a relational diagram to support your description. You should provide any assumptions you made for the design.

Step 5: ETL development (45 marks)

- Develop the ETL using SSIS for data extraction, transformation and loading
    - As mentioned in step 2, you should have at least two (2) types of data sources.
        - Example data source types: CSV, database, text, excel, etc.
- Ensure you have sufficient SSIS tasks to demonstrate your capabilities around ETL process and implementation. Following are some sample transformation tasks you can use:
    - Look ups

- Derived columns
- Splitting
- Merge
- Union
- Sort

Documentation: describe the steps in the ETL process. You should consider the order of execution of ETL tasks when loading data to the data warehouse.

You may use any internet resources (MSDN recommended) to get an idea about how to develop above components.

Marks will be based on:

- Complexity of the data set
- Types of data sources used as inputs to ETL process
- Completeness of the Data Warehouse (number of facts, dimensions)
- Completeness and complexity of the ETL process (number of data sources, transformations)
- Correctness of all of the above