

Data Warehousing & Business Intelligence

Assignment 1

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

- 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 traditional data set such as typical customer/order scenario (do not use AdventureWorksDW). Look for a novel scenario.
- Ensure the data set you select has
 - o At least around one year data
 - o Sufficient number of records and attributes.
- Ensure your data set will have sufficient data to demonstrate following
 - o Multiple sources
 - o Data warehouse design
 - o A rich set of ETL tasks
 - o Enough data to put into SSAS Cubes
 - Hierarchies
 - Dimensions
 - Aggregates
 - o Sufficient data to create reports
- Ensure your data set contains sufficient number of records and attributes to demonstrate above features.

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

Step 2: Preparation of Data Sources

- 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. (This is only an example). You can introduce data such as product category to enrich your data to include hierarchies (if they are not available in your data set)

- You should have at least two (2) types of data sources.
 - o 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

- Design a high-level BI solution architecture.
- To keep the design and the process simple you may skip the “Staging Layer” and directly integrate data sources with the data warehouse.
- Architectures are discussed in lecture 2: slides 4 to 15.

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

Step 4: Data warehouse design & development

- 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.
- Warehouse dimension models are discussed in lecture 3: slides 8 to 15.

Documentation: Describe your data warehouse design. You may use Schema Diagrams to support your description. You should provide any assumptions you made for the design.

Step 5: ETL development

- 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.
 - o 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:
 - o Look ups
 - o Derived columns
 - o Splitting
 - o Merge
 - o Union
 - o 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.

Due Date: 28th April 2018.

Submission Method: To be notified.