R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

ISTE-DW-201
The Product Company
Final Data Mart Development Report

Overview

For this course's final experience, your data mart development team will put together a report on your final data mart design and your implementation process. You will then demonstrate your data mart implementation to your course instructor(s) at a formal project review scheduled for the final exam week.

Your report document must minimally include the following:

- A definition of the universe of discourse (UoD) of your data mart.
- An Information Package chart that includes:
  - A definition of the process being modeled
  - The grain of the data
  - The dimensions and their attributes
  - The facts
- A definition of the entity represented by each dimension in *genus differentia* format (ref. https://en.wikipedia.org/wiki/Genus%E2%80%93differentia_definition)
- A dimensional model (DM) of your final physical data mart schema showing how the dimension and fact entities are implemented with:
  - Attributes – meaningfully named
- Primary keys and relationships with cardinality and participation, including.
- A data-staging extract report lists each of the data extract file(s) from which you obtained your source data along with the format of each extract file.
- A data staging dictionary that defines the source-to-target mapping of each attribute in the dimensional model. Follow the same format as indicated in "The Data Warehouse ETL Toolkit" by Kimball & Caserta, Fig. 3.1 on page 60, which is available on Books 24x7:
  - Target:
    - Table name
    - Column name
    - Data type
    - Table type
    - SCD type
  - Source:
    - Database or file name
    - Table name
    - Column name
    - Data type
  - Transformation
- The SQL code used to create the data mart schema and the supporting constraints

- A report on the following data-staging activities:
  - Data cleansing activities:
    - Which attributes were cleansed
    - The cleansing process – including a description of any manual processes or code
    - What data problems were encountered – by extract file and field
    - How the problems were resolved
  - Transformation processes:
    - How the table images were created – including process and any programs used
  - Table population activities
    - How table data was loaded (including programs and commands, copies of any programs written, Pentaho .ktr and .kjb files, and any parameter files used)
- End-user application(s):
  - SQL queries that answer the user reporting needs from Lab #2
    - A minimum of three (3) queries are required
    - Should use SQL99 as discussed in class
    - Include the code for any index built to support the queries
    - Include sample output
  - At least one view that addresses one or more of the users' reporting needs.
    - Include the code and sample output
  - Additional queries using Aggregated Data Marts.
    - Using the aggregation methods described in the lecture, add and populate each of the following three types of aggregated data marts:
      - Lost Dimension
      - Shrunken Dimension
      - Collapsed Dimension
    - Specify the method you used for each aggregate fact table.
    - Describe types of summary queries with a Use Case description (Not necessarily UML Use Cases) you can generate from each data mart. Include two example queries for each aggregated fact table. A use case is to describe when/what/why users would use this type of query.
    - Include MySQL dump files to be able to reproduce each aggregated data mart.
    - Update your ERD with three aggregated data marts. Submit as a pdf file.
- Consider how you would implement Slowly Changing Dimensions (SCD)
  - Consider a minimum of five attributes in your dimension tables
    a) What type of slowly changing dimension would you implement? Why? (There should be at least two different SCD types other than Type 0)
    b) Create a set of sample source data (minimum 25 records) and make a copy of dimension tables to demonstrate your choices of slowly changing dimension types. Don't apply this to your original data mart. Use Pentaho Kettle steps or any other tool that you have used in the project. Describe how the slowly changing dimension types were implemented and include the code. The Pentaho transformation or any other tools should be generic so that you can apply to any other source datasets.
    c) **Extra Credit**: implement SCD Type 6 to an attribute.
      - Follow the same directions as b) in the above.

- *Investigate Additional Design Options*

  - A STAR schema is built based upon 1-to-many (1:N) relationships between a set of facts (the fact table) and one or more dimensions that are useful for analyzing the facts (the dimension tables). Many-to-many relationships between dimensions are typically

expressed through the fact table (as an associative entity). One-to-many relationships between components of a dimension are modeled as dimensional hierarchies.

- Interestingly, however, when modeling real-life many-to-many (N:M) relationships can sometimes be found between dimension components or between facts and dimensions. Suppose TPC is interested in evaluating the performance of its suppliers. They understand a supplier hires multiple shipping companies as subcontractors. A shipping company supports more than one supplier. In other words, there is a many-to-many relationship between supplier and shipping company. They require creative modeling approaches. You may need to research the literature for physical modeling strategies to support your logical design. Be sure to document any references you use appropriately. Attach a copy of any interesting approaches (chapter, articles, or papers) that you find on the subject and submit them with your final report.

    - What implementation strategy do you propose? What are the other options? Justify your choice of implementation.
    - Create another ER diagram to include supplier and shipping company dimension tables. Update your team's ER diagram to implement the N:M relationships. Submit as a pdf file.
- **Appendix**: Describe how/what you have corrected/changed your data mart developed in Lab #3.
- **PowerPoint Presentation**: Include the grade items shown in the Gradesheet.
- During the Final Project/Exam presentation/defense, your instructor will evaluate each team member individually, asking a question(s), which is (are) either directly related to the team project or basic data warehousing concepts covered in the semester. Reviewing the midterm exam study guide is a way of checking the DW concepts.

Formatting Requirements:

- An **MS/Word** document; double-spaced and spell-checked.
- Text written in the document must be complete, well structured, sentences (except for table and figure descriptions).
- The report should be compiled with separate sections for each of the requirements above clearly identified
    - Include relevant screenshots, printouts, etc.
    - Any reference material used in the implementation must be annotated in APA format and included on a separate page at the end of the relevant section(s).

Your instructor will supply you with a Word document with the required report format.

Report Submission:

The **electronic version** of your report and other files: ER diagram; Data staging dictionary; SQL code to create & populate the data mart & constraints (a dump file); Data Staging activity files; End-user query files with aggregate data marts; Handling SCD; updated ER diagram to implement the N-M relationship; and Presentation PPT & Peer Evaluation will be submitted to the MyCourses dropbox **by 11:59 PM 11/29/20(Sun)**. Refer to the grade sheet for the required files/code to submit. You won't receive any Final Project/Exam credits without submitting a Peer Evaluation Form.