University of Pittsburgh
School of Information Sciences
Graduate Information Science &Technology Program

INFSCI 2711: Advanced Topics in Database Management

Spring 2017

PROJECT: Data Warehousing Strategies

Purpose of the project

You are to analyze the requirements for, design, implement, demonstrate, and document a Data Warehouse aggregates an enterprise data. *As an advanced project option*, your data warehouse can be implemented and replicated in a Virtualized Data Center (*Local Cloud*).

First, you should design and implement a database with the following generic data specification:

- **Customers:** customer ID, name, address (street, city, state, zip code), kind (home/business). If business, then business category, company gross annual income, etc. If home, then marriage status, gender, age, income.
- **Products:** product ID, name, inventory amount, price, product kind w.r.t. some classification.
- **Transactions:** record of product purchased, including order number, date, salesperson name, product information (price, quantity, etc.), customer information.
- **Salespersons:** name, address, e-mail, job title, store assigned, salary.
- **Store:** store ID, address, manager, number of salespersons, region.
- **Region:** region ID, region name, region manager.

You should refine the set of tables and their content depending on specific application of your choice. At this stage you can re-use a database that you developed in you INFSCI2710 project. You may also choose to implement another application after discussing it with instructor.

Then, you should design and implement a Data Warehouse System for business decision support. The Data Warehouse system must provide the following decision support queries:

- What is the ratio of business to home customers?
- What stores are increasing in sales?
- Maintain every day the aggregate sales and profit of the top 5 and the bottom 5 products.
- Maintain every day the top 2 customer categories (highest sales) and the top product categories.
- How do the various regions compare by sales volume?
- Which businesses are buying given products the most?
- What is the demand curve for each product category (i.e., the curve, that measures the propensity of customer to pay for a product as the price changes)?
- Develop a direct marketing data; for each product, a list of customers that buy the product more than 2 times per week.

- Other interesting aggregate values that you will come up with.

If you choose to implement another application, please prepare your application-specific list of decision support queries and discuss them with instructor.

After implementing a relational Data Warehouse you should provide two alternative implementations of the Data Warehouse backend using NoSQL database systems. We suggest you to use Neo4j and MongoDB systems, but your group may choose another tool provided that is approved by the instructor.

Rules of the game

- **Groups:** The project is to be done in groups of 6 students. A roster for each group must be submitted to the instructor by the date specified in the ``Due Dates" section of this assignment.
- **Assumptions:** In cases where the above description of the application is incomplete, it is acceptable to make assumptions about the application providing that: 1) they are explicitly stated in the final report, 2) they don't conflict with any of the requirements specified above, and 3) they are "reasonable". If you have a question about the acceptability of any of your assumptions, check with the instructor or GSA. Interesting questions should be raised in class.
- **Report:** A final report should be handed in for grading at the end of the term. The report must be formatted in a reasonable manner (i.e., using a text processor and a decent printer). The final report is due during class on the "Project Due" date specified in the class schedule.
- **Implementation:** The project requires a working implementation of the system to be built, tested, and demonstrated. A large part of the project grade depends on the quality of this implementation.

Report Requirements

The final report must contain:

1. A short overview of the system including identification of the various types of users, administrators, etc. who will be accessing the system in various ways.
2. A list of assumptions that you have made about the system.
3. A description of the data that will be maintained in your system.
4. A description of STAR schema design. Specification of FACT table and tables for the Data Warehouse dimensions (including corresponding DDL statements). The SQL statements to populate the STAR schema.
5. Specification of pre-aggregated summary tables (including corresponding DDL statements). The SQL statements for creation and populating of the summary tables. Specification of nightly scheduled batch job to summarize data.
6. A description of Data Warehouse queries and front-ends required for the warehouse.
7. Some example scenarios of how various types of users will interact with the system.
8. A description of alternative implementation of your DW system using two NoSQL platforms.
9. A detailed comparison of relational and NoSQL implementations with explanation of advantages and disadvantages of each approach

Advanced project option:

10. A specification of your virtualization configurations and justification of your choices.
11. A specification of your system architecture that integrates your virtualization and replication solutions.
12. A detailed step-by-step specification of your virtualization setting up process for each of your chosen virtualization model with example screen shots.

13. A description of your testing efforts and erroneous cases that you had to handle while setting up your virtualization.
14. A specification of the replication models that you considered and justification of your choices.
15. A detailed step-by-step specification of your replication setting up process for each of your chosen replication model with example screen shots.
16. A description of your testing efforts and erroneous cases that you had to handle while setting up your replication.

In addition, a demo of the working system will be required. All members of the group must attend this demo, and must be prepared to explain and demonstrate those aspects of the project for which they were responsible. The source code for the project should be available on-line during the demonstration.

Go back to 2711 Home Page.

Last Modified: February 8, 2017.