Assignment on Data warehouse (DW)

CSE 512, sp 2013

Student name: Elias Hossain

ID: 2315382050

Analysis-01

Analyze the operational database in different sources and select the appropriate entities and attributes for uploading to the data warehouse.

1. Analyzing the operational database

Solution: The process of designing and implementing the e-commerce data warehouse for the nationwide chain of superstores will involve the selection of relevant entities and attributes, data integration, data extraction, preprocessing and uploading to the DW. Moreover, based on the given scenario, we can identify the following entities and attributes:

* Suppliers: SUP_id, name, type of products, address (street, city, district)

* Superstores: store_id, location, transaction_id, transaction_type (cash or card), timestamp_id, time of the day, day of the week, date, week, month, year, quantity, unit price and total price

* Customers: customer id, name, NID, address (house no., street, thana, city, district, division and age group)

* Items: Item id, name, type, country of manufacture

[Please turn over]

We will select the above entities and their attributes for uploading to the data warehouse.

Analysis-02: Activities required for designing the wrappers for data integration, data extraction, pre-processing, and uploading to the DW:

(A) Data Integration: We will integrate the data from the different sources such as supplier data, superstore data, customer data, and item data into a single database for analysis.

(B) Data Extraction: We will extract the relevant data from the operational database using ETL (Extract, Transform, Load) tools such as Talend, Informatica, or pentaho

(C) pre-processing: We will perform data cleaning and data transformations to remove any inconsistence or errors in the data. We will also convert the data into a suitable format for analysis & loading into the data warehouse.

(D) uploading: In this step, we will upload the pre-processed data into the data warehouse using ETL tools.

Design-

Task-01 – Design the architecture of the warehouse and explain the sources, preprocessing, noise reduction, transformation and uploading.

My response:

For the e-commerce data warehouse for the nation wide chain of superstores in Bangladesh, we can use a traditional Kimball-style data warehouse architecture that includes the following components:

(A) Source systems: These are operational systems that contain the data we need to load into the data warehouse. In this case, the source systems would include the various superstores systems that track transactions and inventory, as well as the supplier systems that track orders and shipments.

(B) Data integration layer: This layer includes the ETL process that extract data from the source systems, transform it into a format that is suitable for the data warehouse, and load it into data warehouse.

(C) Data warehouse: This is central repository where all of the data is stored. We can use a star schema design for this warehouse.

[Please turn over]

(D) Business intelligence layer:

This layer includes the tools and technologies that enable users to access and analyze the data stored in the warehouse.

The following preprocessing, noise reduction, transformation, and uploading steps can be performed during the data integration layer:

(A) preprocessing: This involves identifying and resolving any inconsistencies or anomalies in the data before it is loaded into the warehouse, For example, we may need to clean up customer addresses or reconcile product name and types across different systems.

(B) Noise reduction: This involves filtering out any irrelevent or redundant data that we don't need in the warehouse. For example, we may not need to store every single transaction record from the superstore systems, and can instead aggregate the data by day or week

[Please turn over]

④ Transformation: This involves transforming the data into a format that is suitable for the data warehouse. To give an example, we may need to standardize units of measurement or calculate derived metrics such as revenue or profit margins.

⑤ Uploading: This process involves loading and transformed data into the warehouse, which can be done incrementally or on regular schedule.

To sum up, the architecture and processes should be designed to ensure that the data in the warehouse is accurate, consistent, and up-to-date.

Task-02 – Design the star schema for the warehouse using the scenario and the data set and explain how the data of the superstore database will be collected to the DW (source driven or destination driven).

My response: For the star schema design of the data warehouse, we can use the following dimensions:-

[please turn over]

(A) Time dimension:

This dimension includes attributes such as timestamp, date, week, month, and year, which will allow us to analyze sales trends and patterns over time.

(B) Customer dimension:

This dimension includes attributes such as customer ID, name, NID, and address, which will allow us to analyze customer behaviour and demographics.

(C) Product dimension: This dimension includes attributes such as product ID, name, type, and country of manufacture, which will allow us to analyze product sales and performance.

(D) Supplier dimension: This dimension includes attributes such as supplier ID, name, type of product supplied, and address, which will allow us to analyze supplier performance and relationships.

Please turn over

- The fact table for the star schema will contain the transaction data, including attributes such as transaction ID, transaction type, ~~timestamp~~ timestamp, customer ID, product ID, supplier ID, quantity, unit price and total price.

- The data from the superstore database can be collected using a source-driven approach, where the data integration layer pulls data from the source systems and loads it into the data warehouse. This can be done using ETL processes that are designed to extract data from the various superstore systems, transform it into the appropriate format for the data warehouse, and load it into the warehouse.

Task-03: Prepare a mapping of your design with the DW given in VIS and propose any change (if required).

My response:

In terms of fact table, both designs have similar column such as customer key, item key, quantity, unit price, and total price. However, the VIS DW design has additional columns for payment key, time key, and store key which are not present in the e-commerce data warehouse design.

In the case of dimension tables, both designs have similar dimensions such as item, customer, and item. However the VIS DW design has an additional dimension for store, which is not present in the e-commerce data warehouse design. Additionally, the VIS DW design has a transaction dimension for payment type and bank name, which is not present in the e-commerce data warehouse design.

[please turn over]

Based on this comparision, if we were to map the e-commerce data warehouse design to the VIS DW, we would need to add dimension tables for store and transaction and add columns for payment key, time key, and store key in the fact table.

It should be mentioned that the design of a data warehouse is highly dependent on the specific requirements and goals of the particular organization, therefore, any changes should be carefully considered in the context of those factors.

Overall, both data warehouse have similar structures with fact tables and dimension tables, but there are some differences in the specific columns included in each table. To put it more simply, the e-commerce data warehouse includes some information that is not present in the VIS data warehouse, such as the type of products supplied by each supplier and the age group of each customer, while the VIS data warehouse includes information about the type of transaction and bank name, as well as the Upazila for each store.