North South University

Department of
Electrical and Computer Engineering

ASSIGNMENT ON DATA WAREHOUSE (DW)

Submitted By:

Md. Saif Ahammod Khan

ID# 2315333650

Graduate Student

Submitted To:

Dr. Abu Sayed Md. Latiful Hoque

North South University

Course Code: CSE512

Course Title: Distributed Database

Section: 01

Date of Submission: 15th May, 2023

Analysis:

1. Analysis of operational database: Here our goal is to analyze the operational database from all the given sources and select the appropriate entities and attributes. Later on all the selected item will be upload to the datab warehouse. After doing a Analysis we found the following entities and attributes

i. supplier : sup-id, name, product_type, address (city, district)

ii. customer : cus_id, name, NID, address(Houseno, street, thana, city, district, division), ~~age~~ birthdate

iii. transiction : tran_id, transiction_type, quantity, unit_price, total_price

iv. timestamp : time_id, time, day_of_ week, date, week, month, year

v. Item : item_id, name, type, manufacture_country

vi. store : store_id, address (thana, city, district, division)

After analyzing all the mentioned entity and attributes we will select to upload into the database.

2. Analysis of activities to design the wrappers: For design the wrapper for upload data to data ware house following activities are required:

i) Understanding the source data: Here we are going to take data from the source so at first we have to understand the source data. Here our main source of data is all the super store. At first we have to observe and understand the structure of all source data. Are all source data same or different source data has different structure.

ii. Data extraction: Next our job is to extract data from the source site. Which which entity and attribute we will take and which will drop we will decide that and next we will extract those data from the source.

iii. Data pre-processing: Next step is pre processing the data. The data we will extract from the source there will be a lot of issues like different format or missing values we have to pre process the data and remove or fill the null values by following statistis.

iv. Data integration: Our forth step is data integration. Here we will integrate our data into a common schema and data will be integrated by maintaining consistency and integrity

v. Data upload: Our final step is data ~~integration~~ upload where we will load our integrated data into the data ware house

To design a wrapper we have to go through all these activity.

Design:

Task 1: Design of warehouse architecture An architecture of a data warehouse has three major layer. Those are:

i. Source Layer
ii. Integration Layer
iii. Presentation Layer

i. Source layer: The source layer is the layer where data generates. Here we will collect our data from the stores.

- Data source: In our architecture our major source of data is all the superstore all over the country.
- Data extraction: In our architecture all our shop have a database from there we will gather data using standard Query Language (SQL). This is how we will extract data from source
- Data profile: During processing SQL query we will give our data a profle so that later on we can understand the nature of data during data integration.

ii. Integration Layer: In integration layer the data from the source layer, go through some steps and the integrated into a common schema.

- Data preprocesing: At first in this layer we will preprocess our data by removing all the unnecessary data and clean the data.

- Noise reduction: We will remove any duplicate data and handle missing value by replacing the missing value with statistical analysis. We cannot remove an entire entity fora singel missing value as that will cause a major mulfunction in transiction and inventory.

- Transform: Now we will re-format our data according to our schema and change the data type if necessary

- Data quality: Now governmen policy and data quality rule will be applied on data in this step to increase the data quality and maintain the quality.

iii. Presentation Layer: In this layer data is stored and presented to user for visualization and analysis.

- Data modeling: In our architecture we will use multidimential modeling to model data.
- Data storage: In this architecture we will use star schema to store the modeled data into the data ware house

There are anothe step between the layer of integration and presentation layer. That step is uploading the data. In this step preprocessed data is uploaded ~~to~~ into the presentation layer. For this job we will use ETL tool. Which is extract transform and load tool. As atfirst after data ~~ex~~ extration we will transform the data first and then we will load the data next.

■ Task 2: Design star schema: This star schema will have the following dimensions

supplier_dim: sup_id, name, product_type, city, district

/

customer_dim: customer_id, name, nid, house_no, street, thana, city, district, division

transiction_dim: tran_id, transiction_type, quentity, price, total_price

timestamp_dim: time_id, time, day_of_week, date, week, month, year

Item_dim: item_id, name, type, manufacture_c

store_dim: store_id, thana, city, district, division

In fact table we will have all the id of dimention table.

In the data warehouse data from the superstore will be collected by source driven. As we want to keep updated the ware hous so when source have any data they will push it to the data warehouse.

■ Task 3: Mapping: The DW given in VIS it has five dimention table and a fact table. In my designed DW there are six time stamp.

In vis data warehouse there are no supplier time stamp. Without it we cannot made any analysis on the supplier end site. We can not analysis the supply chain.

In my opinion supply dimention is important for a national based super shop as like customer supplier is also important.

Beside that my ware hose fact table is more complex than the VIS fact table. VIS fact tabe is better.