CS 6400 Database Project: **BuzzBuy Data Warehouse**

Summer 2024

Project Overview

The purpose of this project is to analyze, specify, design, and implement a data warehouse.

The BuzzBuy Data Warehouse

BuzzBuy is a retail business, selling all kinds of products, with stores throughout the United States. Your team has been tasked with designing and building a data warehouse to determine how stores are doing. This section describes in detail the requirements for the BuzzBuy Data Warehouse (BBDW).

A data warehouse is a database system used for reporting, analysis, and other tasks required for decision support. Unlike *transactional* databases which are generally designed to record repetitive day-to-day business transactions (e.g., point of sale, buy and sell stock orders, online shopping carts, etc.), *data warehouses* are especially suited for reporting and analysis over millions of records to support enterprise-wide decision making. As an example, a large online merchant like amazon.com or bestbuy.com relies on a transactional (also called *operational*) database system for recording customer orders and payments in real time. A data analyst tasked with generating a report that compares sales of a certain product among the different regions of the United States will typically query a purposely designed data warehouse for the report instead of accessing the transactional databases directly. This is for several reasons: the data warehouse consolidates data from multiple transactional databases, the data warehouse schema easily supports complex queries aggregating millions of rows, and querying the data warehouse will not impact the performance of the source transactional databases.

For this project, you will design the database schema for BuzzBuy's data warehouse along with a rudimentary user interface. You need not be concerned with the transactional databases that we assume exist to support the point-of-sale system at each store. Instead, you will design the schema to support a consolidated view of the products offered and sold in all stores across the country. What follows is a description of the requirements for the data warehouse in terms of what information must be stored to support a set of reports defined by the executive team.

Even though some amount of redundancy is typically acceptable in a data warehouse schema, ***for this project you should create a normalized schema with as little redundancy as possible***.

> When reading through this project description, please make the following assumptions: unless otherwise specified as optional, all attributes are required; unless otherwise specified, if given a list of potential values, choices should be limited to that list; If a set of values is listed with "and/or", combinations of those values are possible, while "or" indicates only a single value is possible; that you should create normalized schemas, and minimize the use of NULL attributes whenever and wherever possible; ensure that you store non-numeric data that appear as numbers (such as street numbers, phone numbers, postal codes, etc.) as strings and not numeric data types; and avoid "catch-all" forms with unnecessary inputs that the user would leave empty or NULL. You also do not need to be concerned about handling concurrent operations that could conflict and introduce inconsistencies in your database.

You may implement the project UI as a traditional standalone application (e.g., Java GUIs or Python's TkInter) or as a web application (e.g., web scripting languages like PHP or JSP). Your project will not be graded on its aesthetic appeal, but on its functionality.

Do not create any additional functionality that is not mentioned in this specification or attempt to enhance your final product beyond what the specification requires. Adding unwanted functionality can and will impact your grade!

Data Requirements

The BuzzBuy Data Warehouse (BBDW) maintains information about each *store*, including a uniquely pre-assigned *store number*, and the store's *phone number*. Each store must be assigned to a *district*, which is comprised of one or more stores, and each district has already been assigned a *district number*.

BBDW should also maintain information about each store's *city*, including the *city name*, the *state* in which the city is located, and the *population* of the city. It is possible for multiple stores to be located in the same city.

BBDW contains information about every *product* stocked at BuzzBuy stores. Products have a pre-assigned numeric unique identifier (*PID*), similar to a UPC barcode, as well as the *name* of the product. Assume that all products are available and sold at all stores—that is, there is no need to specify that a certain product is only available at a certain store.

Each product is related to a single *manufacturer*. Each manufacturer has a *name*. All manufacturer names can be assumed to be unique. It is possible that multiple products are made by the same manufacturer.

To help identify the kinds of products that are popular, each product is assigned one or more predefined *categories*. Each category has a *name*, which we assume to be unique. Every product must be in at least one category.

Every product has a retail *price*. The retail price is in effect unless there is a temporary, promotional *discount*. BBDW maintains the discount *date* and discount *price* of any product that has a discount. If a product is discounted for multiple days in a row, then a record is stored in the data warehouse for each day. It is possible that the same product is discounted multiple times (i.e., different days) with different discounted prices. **If a product is discounted, it has the same price in all stores—i.e., there are no store-specific discounts.**

The BuzzBuy executive team would like the ability to compare sales data on *holidays* versus non-holidays, so BBDW should maintain information about which specific dates are holidays. The specific name of the holiday is also required. You do not need to worry about a date having several holidays associated with it – it is only possible for a date to have one holiday.

Finally, BBDW stores information about which products are *sold*, including the *store* where it is sold, the sold *date*, and the *quantity* of the product purchased. The total is not stored explicitly but can be derived based on the date purchased and the quantity. Sales tax does not count as revenue and does not need to be tracked. Also, the data warehouse is not required to store which products were purchased together during a single sales transaction.

BuzzBuy's DBAs are working on an extract of sample data from their point-of-sale system for you to test in your data warehouse, however, they estimate it will take at least two months before it can be made available to you. Due to this, you must ensure that your schema design matches the data as described here so that any transformation needed to load the data is kept to a minimum.

Along with the sales data that will be stored in the data warehouse, some additional data must be kept in the data warehouse as it will be used to allow users access to the system, to ensure they can only view data according to their access level, to track whenever a holiday is added to the system, and to log each time a user views a report.

All *users* in the data warehouse will be identified by their pre-existing *employee ID*, which is a maximum of 7 characters and may have leading zeroes (an example employee ID number is 0010423). Instead of a password, they will authenticate by entering the *last 4 digits* of their Social Security number, followed by a dash, followed by their properly capitalized *last name*. To greet the user upon login, their *first name* will also be stored. Users will be manually configured by the database administrator, and the appropriate details for the record will be given to the DBA by BuzzBuy HR when setting up a new user.

To track permissions in the database, each user is assigned to one or more *districts*. Like user data, the DBA will manually manage these assignments. A district sales director could only have access to their one assigned district, while a corporate employee will be assigned to all districts, or someone responsible for marketing for three districts would be assigned to those three districts. ***It is important to note that this district assignment will control what data and reports are available to the logged-in user. This will be further detailed in the "reports" section.***

Users that have been granted access to all districts have the ability to add holidays to the data warehouse. All holidays will be associated with the user that created it.

Finally, at the request of the BuzzBuy cybersecurity, an *audit log entry* must be created each time a report is viewed. Each log entry should have the *user's employee ID*, their *last name* and *first name*, the *timestamp* (date and time local to the data warehouse), and the *name* of the viewed report. To ensure consistency and to save storage space, the list of all reports will be stored in the database, maintained by the DBA, and the report names should match how they are named in this specification. Users who are allowed to view the audit log within the data warehouse UI will have a special flag set on their account by the DBA.

BuzzBuy Data Warehouse User Interface

All reports must be accessible from a UI that must be developed.

Users will log in with their credentials as previously described in the "Data Requirements" section. An appropriate screen prompting them for their employee ID, and their "password" (the last four digits of their Social Security Number followed by a dash, followed by their properly capitalized last name).

After successful login, users will be shown a main menu screen which can be used to access all functionality of the system that has been described in this specification. On this main menu, there should be a message welcoming the user by their first and last name, such as "Welcome, George Burdell!" and the following statistics should be displayed: the count of stores, cities, districts, manufacturers, products, categories, and holidays. (These statistics should be shown for all data in the data warehouse, regardless of the user's access level.) Finally, the reports available for the user should be listed with a button, link, or other navigation option to view that report. (The reports and their definitions will be listed in the following section of this specification.)

In addition to the reports, there are some relatively simple interfaces you should design and provide as part of maintaining the data warehouse. First, you must provide an interface for viewing existing and adding holiday information within the application. All users can view holidays, but remember, only users who have been granted access to all districts can add holidays, and the system must also record which user added it. Obviously, if a holiday already exists for a date, you cannot allow the user to create a new one on the same date.

Finally, an interface for viewing audit log must be made available to users who have audit log permissions. On the audit log page, the most recent 100 audit log records should be displayed in a table, showing the timestamp, employee ID, last name and first name separated by a comma and space (an example of this is "Burdell, George"), and report name, ordered by timestamp descending and employee ID ascending. In addition, to assist in detecting suspicious user behaviors, if the user for an audit log entry has been assigned to all districts, that log entry row should have a yellow background.

BuzzBuy Data Warehouse Reports

Your team must also develop the queries for the following reports, accessed via the user interface that you will create. The reports are grouped into three types: general, district, and corporate. Each report group description will provide more specific details.

Some of the report queries are expensive to run given the size of the sales data. Therefore, whenever possible you should include the filter conditions specified. For example, some reports ask for data from only a certain time period. If you leave off this filtering condition, the query will likely take a long time to return any results.

General Reports

General reports are reports available to all users that provide general information about the data stored in the data warehouse. Because of this, they should return all data for all users regardless of their district assignment.

Report 1 – Manufacturer's Product Report

For each manufacturer, return the manufacturer's name, total number of products offered by the manufacturer, average retail price of all the manufacturer's products, minimum retail price, and maximum retail price. Ignore all discount days (do not take into account the days the product is discounted). Sort the results by average price with the highest average price appearing first, for only the top 100 manufacturers based on average price.

This report should also have "drill-down" detail (in other words, each line in the master report should have a method for loading its detail, such as a hyperlink or a button) for the manufacturer, which shows in the report header the manufacturer's name, the summary information from the parent report, and lists for each of the manufacturer's products' its product ID, name, category (or categories), and price, ordered by price descending (high to low). If a product has multiple categories, it must not show up as multiple rows on the report, but as a single row with multiple categories concatenated together, with a comma and space delimiting each category.

Report 2 – Category Report

For each category, return the category name, total number of products in that category, total number of manufacturers offering products in that category, and the average retail price (not including discount days) of all the products in that category, sorted by category name ascending.

District Reports

District reports are also available to all users but provide more specific information regarding sales. These reports should return data based on the district(s) assigned to the logged-in user.

Report 3 – Actual versus Predicted Revenue for GPS units

BuzzBuy's marketing team wants to predict whether offering items at a discount actually helps to increase revenue by encouraging a higher volume of sales. This report compares how much revenue was actually generated from a product's sales to a predicted revenue if the product was never discounted. After speaking with some marketing consultants, the team has learned that product discounts introduce on average a 25% increase in volume (quantity sold). Therefore we assume that if an item that was offered at a discount were instead offered at the retail price, the quantity of items sold would be reduced by 25%. However, it is still possible that the predicted revenue would be higher since the reduced volume of products would be sold at a higher price per product. The first version of the report should only be for products in the GPS category.

Here is a simple example:

Assume that Product Z has a retail price of $10. Assume that it was offered at a discount for on 6/1/2012 and 6/2/2012. Also assume the following transaction data for Product Z:

Date	**Price**	**Quantity**	**Actual Revenue**
5/1/2012	10.00	5	50.00
6/1/2012	8.00	10	80.00
6/2/2012	7.00	5	35.00
TOTALS		**20**	**$165.00**

Table 1 - Actual Revenue

The predicted revenue is calculated by assuming that the product is never offered at a discount and only 75% of the original quantity was actually sold on discounted days. Note that because this is just a predicted average, we assume that it is possible to sell a fraction of a product (e.g., 7.5 DVD players).

Date	**Price**	**Quantity**	**Predicted Revenue**
5/1/2012	10.00	5	50.00
6/1/2012	~~8.00~~ 10.00	10 * .75 = 7.5	75.00
6/2/2012	~~7.00~~ 10.00	5 * .75 = 3.75	37.50
TOTALS		**16.25**	**$162.50**

Table 2 - Predicted Revenue

In this example, the discounted prices resulted in slightly more revenue due to the higher volume of sales ($2.50 more).

Generate the following report: For each product in the GPS category, return the product ID, the name of the product, the product's retail price, the total number of units ever sold, the total number of units sold at a discount (i.e., during a discount day), the total number of units sold at retail price, the actual revenue collected from all the sales of the product, the predicted revenue had the product never been discounted (based on 75% volume selling at retail price), and the difference between the actual revenue and the predicted revenue. If the difference is a positive number, it means that the discounts worked in favor of BuzzBuy because the predicted revenue is less than the actual revenue collected. If it is a negative number, it indicates that BuzzBuy would have been better off not offering the product discounts. Only predicted revenue differences greater than $200 (positive or negative) should be displayed and sorted in descending order.

Report 4 – Air Conditioners on Groundhog Day?

Some customer associates have noticed that air conditioner purchases seem to spike on Groundhog Day (which falls on February 2 each year). They surmise that this is because customers begin thinking about the warm spring weather ahead. The BuzzBuy marketing team would like to confirm this, so they have requested the following report.

For each year, return the year, the total number of items sold that year in the air conditioning category, the average number of units sold per day (assume a year is exactly 365 days), and the total number of units sold on Groundhog Day (February 2) of that year. Sort the report on the year in ascending order. The marketing team will use the report to determine if the total number of units sold on Groundhog Day each year is significantly higher than the average number of units sold per day.

Corporate Reports

Corporate reports provide information that should only be available to users who have been granted access to all districts. Due to this, the information displayed should accordingly reflect data for all districts. Your user interface should ensure that logged-in users who have not been granted access to all districts cannot access these reports and not displayed in any report listing.

Report 5 – Store Revenue by Year by State

This report shows the revenue collected by stores per state grouped by year. The states available for querying should be presented in a drop down box. For example, the user would select "New York" and the system would show each store in New York state, show the store ID, city name, year, and total revenue. Be sure the revenue calculation takes into account items that were sold at a discount. Sort the report first by year in ascending order and then by revenue in descending order.

Report 6 – District with Highest Volume for each Category

BuzzBuy management is planning to recognize all stores in the district that sell the greatest number of units for each category. They want to view this monthly, so the user interface must allow choosing a year and month from the available dates in the database before running the report. The report will return for each category: the category name, the district that sold the highest number of units in that category (i.e., include items sold by all stores in the district), and the number of units that were sold by stores in that district. This output shall be sorted by category name ascending. Note that each category will only be listed once unless two or more districts tied for selling the highest number of units in that category. *The report can take a significant time to run, which may require tuned indices for the final implementation, but do not focus on their creation until the final phase*.

This report should also have "drill-down" detail (in other words, each line in the master report should have a method for loading its detail, such as a hyperlink or a button) using the criteria of

district, category, and year/month to provide the IDs, states, and cities of all the stores. This sub report should be ordered by store ID ascending and the header should include the original criteria from the parent report (category, year/month, district).

Report 7 – Revenue by Population

To help forecast expansions into other cities, BuzzBuy management would like to see what the average revenue is for specific population categories, and to see if there is a trend for growth, the revenue should be broken down on an annual basis. The categories for city size are: Small (population <3,700,000), Medium (population >=3,700,000 and <6,700,000), Large (population >=6,700,000 and <9,000,000) and Extra Large (population >=9,000,000). There is some flexibility in formatting this report, in that it could be "pivoted" to present it with either years or city category as columns or as rows, so ensure that both elements are arranged in ascending order (oldest to newest for years, smallest to largest for city size) so that no matter how it is formatted it is properly organized and understandable.

Revision History

Version	**Notes**	**Date**
1.0	First version	5/28/24
1.0.1	Removed references to address from reports, clarified that DBA will set audit log viewer flag	6/10/24