Database, Data Warehouse Technology for U.S. Chain-Restaurant Menu Item Nutrition Data Analysis

Foroozan Akhavan, Minh-Tam Pham, Zhe Li, Chloe Ngo, and Taoli Zhen

**Abstract**—These instructions give you guidelines for preparing papers for IEEE Computer Society Transactions. Use this document as a template if you are using Microsoft Word 6.0 or later. Otherwise, use this document as an instruction set. Please note that use of IEEE Computer Society templates is meant to assist authors in correctly formatting manuscripts for final submission and does not guarantee how the final paper will be formatted by IEEE Computer Society staff. This template may be used for initial submissions; however, please consult the author submission guidelines for formatting instructions as most journals prefer single column format for peer review. An abstract should be 100 to 200 words for regular papers, no more than 50 words for short papers and comments, and should clearly state the nature and significance of the paper. Abstracts *must not* include mathematical expressions or bibliographic references. Please note that abstracts are formatted as left justified in our editing template (as shown here).

**Index Terms**—Keywords should be taken from the taxonomy (http://www.computer.org/mc/keywords/keywords.htm). Keywords should closely reflect the topic and should optimally characterize the paper. Use about four key words or phrases in alphabetical order, separated by commas (there should not be a period at the end of the index terms)

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

————————————————

1. F.A. Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.
2. M.P. Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.
3. T.C. Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.
4. Z.L Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.
5. C.N. Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.
6. T.Z. Author is with the San Jose State University, San Jose, CA 95192. E-mail: author@ sjsu.edu.

—————————— ◆ ——————————

# 1 Introduction

N

owadays dining out is the integral part of city life.

More than ⅓ of Americans dine out at chain restaurants. Studies [1] show that the fast-food chain restaurants' contribution to obesity is significant. For people who don’t have time to prepare meals by themselves, ordering meals that can satisfy daily macro nutritional needs while having calories and unhealthy fat under the daily consumption limit become the key to address the obesity issue. But can people dine out in chain-restaurants achieve the nutrition requirement to address the obesity issue?

At the meanwhile, as more and more people started to care about the nutrition composition in the food they consume and started to learn to read nutrition labels, some major chain-restaurants also started to include some healthy options on their menu. In addition, some major restaurants voluntary labeled their menus with calorie information and provided a fair amount of nutrition infomatin on their website since 2008. As mentioned in [1], Back in 2010, the U.S. government passed the Affordable Care Act with a rule requiring chain-restaurant to display calorie information on menus. With this rule and the voluntary labeling in major chain-restaurants, there is a sufficient amount of menu nutrition data available from menustat.org [2] for us to explore in this topic. The dataset enables us to investagate in what ways chain-restaurants had altered their menu items in order to provide healthier food over the years.

In this paper, we will go through and explore the technologies that enables data analytics on our topic, as well as data cleansing procedures and operations that are needed to store and manage the menu nutrition data. These operations and database technologies enables us to perform efficient analytics on the menu dataset while keeping the cost at minimum. We will employ a cloud database management platform to store and organize the data we gathered, which the data can be replicated into an analytics platform to improve the efficiency when performing On-Line Analytical Processing.

# 2 Methods and Implementation

## 2.1 Data Cleansing and Transform

The dataset files from menustat.org was in csv format and the menu nutrition data is stored in saperate files by year. A star schema or possibly a galaxy schema would be suited for the dataset. Detailed code for extracting the data, normalizing the data, and transforming the data is in the data\_transform.ipynb notebook file, which is in the Appendix Github link.

Diagram

Description automatically generated

Fig. 1. Galaxy Schema for the Menu Nutrition Data. Serve as a data warehouse schema.

The 2018 (latest) dataset was used as the base case for all other years. All attritebute were in one single large table, which was in the 1st normal form. The first step is to transform the table to 3rd normal form for the galaxy schema. A Python package Pandas was used in the cleansing and processing stage.

For our schema, there are a restaurant dimension, a food category dimension, an item information dimension, a nutrition fact table and a combo data fact table.

For the restaurant and food category dimension, the first step was to extract restaurant name and food category name from the main table. For each of the dimension, we only retained the number of unique restaurant and food category names and assigned a primary key (surrogate key) for each unique name. These dimension tables were ready to load after the above processing.

For the item information dimension, we selected the item\_id, year, item\_name, and item\_deescription column from the main table and retained the unique rows only. The item\_id and year attribute are the composite key for this dimension.

Once we have the meta data transformed into dimension tables, we can transform the main table into the nutrition fact table. First, we joined the restaurant dimension and food category dimension with the main table to obtain the restaurant id and food category id for each item. With the item\_id, year, restaurant\_id and category\_id as the composite key in the fact table. Restaurant\_id and category\_id is also the foreign key to the corresbonding dimensions. Each transaction records the nutrition record for a menu item for the specific year.

Finally, for the combo meal fact table, we have combo\_id, year, item\_id and category\_id as the composite primary key. And the attribute ‘builds’ records whether the item in a combo is a main item or an accompanying item. Each combo can have multiple main items and accompanying items, the logic for the combo is that people can only choose one main item and one accompanying item for each food category.

Text

Description automatically generated

Fig. 2. Example sql query to set up the replication filter function.

With the base case data extracted and transformed, we created a transform\_data(dat, year) function reusing the code for 2018 year to transform the rest of the years’ data. Note that there was a small amount of NA values (under 200 rows) in restaurant\_id column in the nutrition fact table in year 2008, 2010, and 2012; which means that these restaurants were no longer in the menu data record in later years. Thus, these NA values can be ignored as we can’t compare the nutrtion data with later years.

## 2.2 Loading Data to MySQL

Initially we loaded the data to the local MySQL server using sqlalchemy. To enable the ability for team members to connect to the database remotely and optimize the data warehouse for analytics, we decided to employ MariaDB and SkySQL platform to run our data warehouse. To migrate the database to MariaDB Cloud, we exported the menu nutrition database from MySQL to a dump file.

## MariaDB, SkySQL, and HTAP Platforms

According to [4], MariaDB is a fork of MySQL and the process of replacing MySQL with MariaDB can be as easy as importing a dump file to MariaDB from MySQL. When compared to MySQL, MariaDB have the capability to support a variety of storage engines. And in many cases, performance on MariaDB is better than MySQL.

SkySQL is the cloud service for MariaDB, there are four flatforms available in SkySQL. They are the Transactions Platform which is optimized for fast transaction processing (OLTP), the Analytics Platform which is optimized for running ad hoc queries on data warehouses (OLAP), and the Hybrid Transactional-Analytical Processing Platform which supports both OLTP and OLAP using different storage engines for OLTP and OLAP [3]. Here in this project we used the Hybrid Transactional-Analytical Processing Platform to simulate the real-life work environment. We can have a seemless data movement from the transaction database to analytics data warehouse.

As documented in [5], transaction database can handle OLTP queries “using row-based transactional storage engines, such as InnoDB or MyRocks,” and analytics data warehouse can handle OLAP queries “using the MariaDB ColumnStore storage engine.” The MariaDB server use MariaDB MaxScale to handle client connections. Another function of MaxScale is to differentiate between SQL queris for RowStore and SQL queries for ColumnStore. In MariaDB, regular SQL queries can be used to query from the ColumnStore databases. MaxScale can route queries to the right server [5].

## HTAP Data Warehouse Setup and Loading Data

In this project, we created a RowStore database for transactional processing and a ColumnStore database to serve as the data warehouse to store historical data for analytics.

To enable auto replication, the MariaDB server replicates “writes from InnoDB tables to the MariaDB ColumnStore tables” using MariaDB Replication. According to the MariaDB Enterprise Documentation [5], “MariaDB Replication with MariaDB MaxScale configured as a Binlog Server, MariaDB Enterprise Server can host InnoDB and ColumnStore on the same Server.”

In our practice, we specified specific the Row Store tables with certain prefix to automatically replicate to Column Store, which fed data directly from the transaction RowStore to the data warehouse ColumnStore. As demonstraded in the mariadb\_cloud\_load\_data.ipynb notebook file in the Github link under Appendix, we set up the replication filter using the set\_htap\_replication() UDF.

The next step was to create two databases with distinguishable names, one named as for menu\_rowstore, another one named as menu\_cstore. Create tables for our data model in the menu\_rowstore database and check if tables in the RowStore were auto replicated to the ColumnStore. We verified that the auto replication was working properly. Then we deleted the tables in the ColumnStore and re-created the same tables with engine specified as ColumnStore in table creation queries. Finally, we imported the dump file exported from MySQL in to the RowStore database and verified that the data was replicated to ColumnStore in nearly realtime.

## Analytics Methods 1

## Analytics Methods 2

## Analytics Methods 3

# 3 Results

## 3.1 Results 1

## 3.2 Results 2

## Results 3

# 4 Results

## 4.1 Appendices

Appendices, if present, appear online as supplemental material. In the event multiple appendices are required, they will be labeled “Appendix A,” “Appendix B, “ etc.

IEEE Computer Society Transactions accepts supplemental materials for review with regular paper submissions. These materials may be published on our Digital Library with the electronic version of the paper and are available for free to Digital Library visitors. Please see our guidelines below for file specifications and information. Any submitted materials that do not follow these specifications will not be accepted. All materials must follow US copyright guidelines and may not include material previously copyrighted by another author, organization or company. More information can be found at <http://www.computer.org/portal/web/peerreviewjournals/author#supplemental>.

## 4.2 Acknowledgments

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” Sponsor and financial support acknowledgments are included in the acknowledgment section. For example: This work was supported in part by the US Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Researchers that contributed information or assistance to the article should also be acknowledged in this section. Also, if corresponding authorship is noted in your paper it will be placed in the acknowledgment section. Note that the acknowledgment section is placed at the end of the paper before the reference section.

## 4.3 References

Unfortunately, the Computer Society document translator cannot handle automatic endnotes in Word; therefore, type the reference list at the end of the paper using the “References” style. See the IEEE Computer Society’s style for reference formatting at: <http://www.computer.org/portal/web/publications/style_refs>. The order in which the references are submitted in the manuscript is the order they will appear in the final paper, i.e., references submitted nonalphabetized will remain that way.

Please note that the references at the end of this document are in the preferred referencing style. Within the text, use “et al.” when referencing a source with more than three authors. In the reference section, give all authors’ names; do not use “et al.” Do not place a space between an authors' initials. Papers that have not been published should be cited as “unpublished” [4]. Papers that have been submitted or accepted for publication should be cited as “submitted for publication” [5]. Please give affiliations and addresses for personal communications [6].

Capitalize all the words in a paper title. For papers published in journals not published in English, please give the English citation first, followed by the original foreign-language citation [7].

## 4.4 Additional Formatting and Style Resources

Additional information on formatting and style issues can be obtained in the IEEE Computer Society Style Guide, which is posted online at: <http://www.computer.org/portal/web/publications/styleguide>. Click on the appropriate topic under the Special Sections link.

# 5 Results

Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Authors are strongly encouraged not to reference multiple figures or tables in the conclusion—these should be referenced in the body of the paper.

**Acknowledgment**

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

**References**

1. Calorie changes among food items sold in U.S. convenience stores and pizza restaurant chains from 2013 to 2017 https://www.sciencedirect.com/science/article/pii/S2211335519301068?via%3Dihub
2. http://www.menustat.org/
3. mariadb.com/docs
4. <https://kinsta.com/blog/mariadb-vs-mysql/>
5. https://mariadb.com/docs/multi-node/htap/

**First A. Author** Allbiographies should be limited to one paragraph consisting of the following: sequentially ordered list of degrees, including years achieved; sequentially ordered places of employ concluding with current employment; association with any official journals or conferences; major professional and/or academic achievements, i.e., best paper awards, research grants, etc.; any publication information (number of papers and titles of books published); current research interests; association with any professional associations. Author membership information, e.g., is a member of the IEEE and the IEEE Computer Society, if applicable, is noted at the end of the biography.

**Second B. Author Jr.** biography appears here.

**Third C. Author** biography appears here.