MCO1 Query Processing Technical Report

Sean Gabriel M. Cardeno[1], Jeff Uriel V. Fonte[2], Aurelio Rodolfo F. Garcia[3], and Raine Margaux A. Siongco[4]

De La Salle University Manila

[1]sean_cardeno@dlsu.edu.ph, [2]jeff_uriel_v_fonte@dlsu.edu.ph, [3]aurelio_garcia@dlsu.edu.ph, [4]raine_siongco@dlsu.edu.ph

Abstract

With tensions of all kinds rising in almost every corner of the world, coupled with significant strides in psychology and neuroscience, the topic of Mental Health has been ubiquitous within the global public discourse. In particular, as higher education has become more competitive and rigorous, the mental health of students has received notable attention. Reports from agencies and experts worldwide consistently reach a common conclusion: students are struggling. With all this in mind, the project aimed to thoroughly examine the available data from various sources and perform various queries to gain a better understanding of this phenomenon. For the ETL script, age and gender were the only commonalities found among the datasets. The datasets were read and extracted with the Pandas library. To connect each one, data wrangling operations such as renaming columns, data type coercion, mapping, and one-hot encoding were necessary. These processes allowed the data to be aggregated as numeric values and loaded into the data warehouse for analysis and visualizations. The OLAP application uses student data to display the relationships between gender, age, and mental health. Using MySQL, queries incorporating OLAP operations such as roll-up, drill-down, slice, and dice were developed to generate analytical reports. These reports are visualized through the dashboard to provide insights. Focused on validating that the ETL process and OLAP dashboard functioned correctly. The backend queries were tested to ensure accurate results, and the dashboard was verified to update properly when filters were applied. All tests passed successfully.

Keywords

Mental Health, Students, CDC, Data Warehouse, ETL, OLAP, Query Processing, Query Optimization.

1. Introduction

The objective of this study is to identify patterns and trends between student groups and mental health indicators. To accomplish this, the following datasets were used for the study:

- **PHQ-9 Student Depression Dataset:** A dataset that contains responses from 400 students to the PHQ-9 questionnaire, a well-established tool for diagnosing depression.
- **Student Depression Dataset:** A dataset that contains data aimed at analyzing, understanding, and predicting depression levels among students.
- **Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days:** A dataset compiled by the U.S. Census Bureau, in collaboration with five federal agencies, launched the Household Pulse Survey to produce data on the social and economic impacts of COVID-19 on American households.
- **Students' Social Media Addiction:** A dataset that contains anonymized records of students' social‑media behaviors and related life outcomes.
- **Student Mental Health Crisis After COVID-19:** A dataset that explores the mental health status of students post-COVID, with a focus on depression, academic pressure, and suicidal thoughts.
- **Student Performance & Behavior Dataset:** A dataset that contains real data of 5,000 records collected from a private learning provider.

The project's data warehouse was designed using both MySQL and Jupyter Notebook, while the ETL Script was made with Jupyter Notebook. The OLAP Application was built using JavaScript and utilizes SQL and OLAP operations to perform queries. The application is designed for use by anyone, particularly researchers, educators, and students who wish to explore patterns and trends in mental health data through interactive visual reports.

2. Data Warehouse

Using all the datasets, a Data Warehouse was developed, with the following figure.

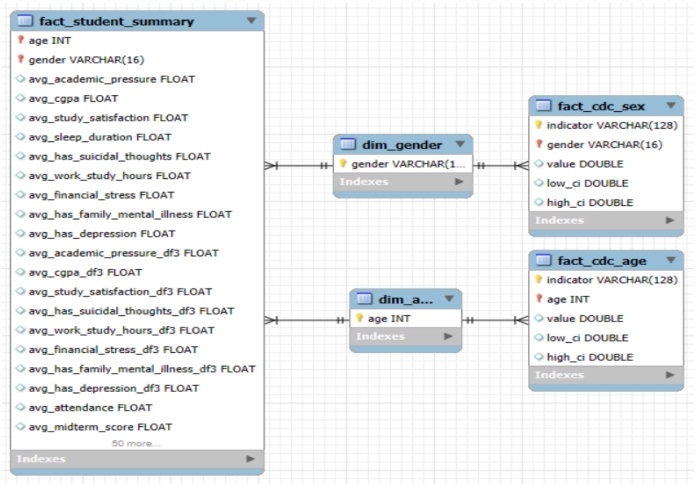

Fig 1. Data warehouse schema

The Data Warehouse follows the **Snowflake Schema Design**, with **three fact tables** and **two dimension tables**, as shown:

- **fact_student_summary:** The Student Summary Table is the primary fact table in the warehouse, containing summarized information about various student groups.
- **fact_cdc_sex:** The CDC's Statistics Based on Sex contains the CI values for different mental issues, separated by biological sex.
- **fact_cdc_age:** The CDC's Statistics Based on Age contains the CI values for different mental problems, separated by biological sex.
- **dim_gender:** The dimension table for gender, referenced by all the fact tables.
- **dim_age**: The dimension table for the different ages, referenced by all the fact tables.

To properly utilize the psychiatric information from various individuals across different sources, the data were aggregated into the Student Summary Table, the CDC Statistics by Age Table, and the CDC Statistics by Sex Table. This way, queries for the psychiatric statistics of large groups of people could still be performed. Age and Gender were converted into dimension tables due to being a common factor in the utilized datasets.

3. ETL Script

The ETL script was developed as a Python notebook. First, the datasets were loaded and extracted using the Pandas library without any issues. Initially, the extraction was problematic, as there were times when the files were corrupted when downloaded from the sites. The process was straightforward, as the Pandas library included appropriate reader functions for JSON, CSV, and Excel files. A commonality among the datasets was age and gender, which can be used to combine facts from all the datasets about those groups and their visualizations. The figure below shows the flow of the ETL process.

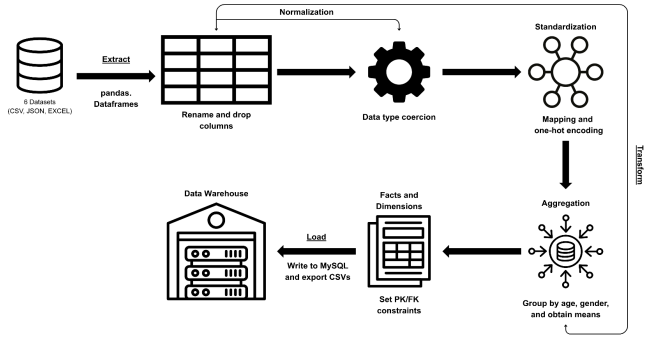

Fig 2. ETL Pipeline

After extraction, each dataset had to be cleaned and preprocessed before being aggregated into the data warehouse. For the columns, all of them had to be renamed, where the names are lowercase, underscored, and non-alphanumeric. This change summarizes survey questions and prevents the names from being misread by Python.

After this, the unnecessary columns and those with missing instances were removed from the dataframes. Normalization and standardization involve data type conversions, mapping, and one-hot encoding to convert all data into a numeric format. The purpose of this is to ensure that all the values can be aggregated for the fact table. Moreover, the categorical values were turned into columns to represent the distribution for a particular age and gender.

Once everything was converted into numeric values, the datasets were then aggregated and normalized before creating the fact and dimension tables. During the operation of merging after creating numerous one-hot columns, it experienced high memory spikes, resulting in runtime errors.

To resolve this issue, the process was broken down, where datasets were joined one at a time to reduce peak memory use. After standardization, three fact tables and two dimension tables were created. The two fact tables originate from the CDC dataset, which uses age ranges and gender groups instead of individual instances.

For the loading section, it writes the fact and dimension tables into a local MySQL Server and Workbench using the SQLAlchemy library. It also applies keys and foreign-key constraints after loading using ALTER TABLE statements.

The constraints include primary keys on dimensions, composite PKs on the main fact table, and foreign keys from facts to dimension tables. Overall, these constraints enforce referential integrity for OLAP queries. The fact and dimension tables were exported as CSV files to validate the aggregation. Lastly, once the tables were validated and written to the data warehouse, it was ready to be used for the OLAP application.

4. OLAP Application

The application is a mental-health dashboard for student data. It summarizes and visualizes relationships between sleep, stress, social-media addiction, academic performance (cGPA), and mental-health indicators across gender and age groups. The dashboard exposes slice/dice filters and renders charts (scatter, radar, bar, line) to provide in-depth visualizations.

4.1. Decision-Making / Analytical Tasks

- Identify correlations, such as sleep and stress, to inform.
- Detect risk groups in certain age groups and gender with high levels of addiction or mental health risk.
- Find relationships between social-media addiction and CGPA.
- Track trends across age groups and genders for addiction and mental health.

4.2. Analytical Reports

1. Correlation between Sleep and Stress

Examines if lower sleep hours are associated with higher stress levels in student groups (age range and gender) across different mental health scores.

OLAP Operations

- **Slice/Dice:** WHERE gender='Female' AND age BETWEEN 18 AND 20 — this is a slice (selects a specific gender and age range). Changing these filters implements dice/slice interactions from the dashboard.
- **Roll-up:** WITH ROLLUP produces subtotals and totals along the gender and age group axis.
- **Drill-Down:** The dashboard can remove the WITH ROLLUP or change grouping to show the raw age level (or fetch detail rows) to drill into individuals.

gender	age_group	avg_sleep	avg_stress_level
Female	18–20	6.52	9.67
Female	NULL	6.52	9.67
NULL	NULL	6.52	9.67

Fig 3. Sleep and Stress

Sleep duration is fairly consistent across genders (~ 6.4 - 6.5 hrs). Stress levels drop as age increases, especially among females (9.67 → 3.00). Females aged 18–20 show the highest stress, far exceeding males of any group.

There is a clear inverse relationship between sleep and stress — less stress aligns with slightly longer or more stable sleep durations. Furthermore, younger females (18–20) have the highest stress and least sleep, while older students of both genders show better stress regulation.

2. Mental Risk Indicators

Compares how different student groups (age range and gender) based on their sleep quality, addiction score, and stress index shape a risk profile.

OLAP Operations

- **Slice/Dice:** Using WHERE + CASE to focus on gender and age (slice); combining multiple dimensions (gender × age_group) is a dice operation.
- **Roll-up:** Summarizes higher-level metrics (e.g., per-gender).

gender	age_group	avg_addicted_score	avg_cgpa	avg_mental_health_score
Female	18–20	8.67	7.54	5.33
Female	NULL	8.67	7.54	5.33
NULL	NULL	8.67	7.54	5.33

Fig 4. Addicted Score, CGPA, and Mental Health Score

The average stress level (7.07) is moderately high. A mental health of 6.64 is slightly positive, but PHQ-9 = 14 suggests mild depressive symptoms. An addiction score of 5.64 indicates moderate social-media dependence. From the application's results, there

3. Impact of Social Media Addiction on Academic Performance

Compares how different student groups (age range and gender) based on their sleep quality, addiction score, and stress index shape a risk profile.

OLAP Operations

- **Slice:** Filter to a specific gender and age band.
- **Roll-up:** Aggregates at group and higher level (subtotal, total) to spot broad trends and compare group-level averages.

gender	age_group	avg_addicted_score	avg_cgpa
NULL	NULL	8.67	7.54
Female	NULL	8.67	7.54
Female	18–20	8.67	7.54

Fig 5. Addicted Score and CGPA

Younger students (18–20) have the highest addiction to social media, but their mean GPA remains decent (7.6). Addiction drops sharply in the 21–23 group (4.17), correlating with GPA slightly improves. For older students, addiction rebounds slightly (5.0) but GPA remains stable. From the output, there is a clear inverse relationship between addiction and academic performance.

4. Gender & Age Group Comparison of Stress and Mental Health

Reveals how average stress levels and mental health scores differ by students' gender and age groups. This visualization aims to identify:

- Age Trends
- Gender Gaps
- Age Groups with stable mental health but high stress scores

OLAP Operations

- **Dice:** Selecting the age_group dimension with BETWEEN and AND for age.
- **Drill-down:** The gender query is an example of drilling down to a different dimension (gender).
- **Roll-up:** For subtotals

age_group	avg_stress_level	avg_mental_health_score
18–20	8.67	5.67
NULL	8.67	5.67

Fig 6. Stress level and mental health score for age groups

gender	avg_stress	avg_mh
NULL	6.86	6.29
Female	6.86	6.29

Fig 7. Stress level and mental health score for each gender

Males display marginally higher stress and higher self-rated mental health. Females rate lower in both. For both genders, stress starkly decreases with age and mental health generally improves. The youth (18-20) are particularly more vulnerable to stress.

5. Social Media Addiction and Mental Health

Illustrates a graph showing how addiction levels vary with mental health factors—such as academic, pressure, work pressure, job satisfaction, and sleep duration—across age groups and genders.

OLAP Operations

- **Slice:** by gender + age range
- **Roll-up:** across age_group × gender; dashboard uses these aggregates to draw trend and radar charts

gender	age_group	avg_academic_pressure	avg_sleep	avg_work_study_hours	avg_mental_health_score
Female	18–20	3.39	7.83	7.24	5.33
NULL	18–20	3.39	7.83	7.24	5.33
NULL	NULL	3.39	7.83	7.24	5.33

Fig 8. Academic pressure, sleep, work hours, and mental health score

18–20-year-olds report the most academic pressure, which drops as the age increases, with mental health improving correspondingly. There is a clear positive correlation between academic pressure and negative mental health.

General Findings

Younger students, especially female students, report the highest stress and academic pressure, which has clear negative effects on their mental health and sleep patterns. As students age, both stress and social-media addiction decline, and mental well-being improves.

Advanced SQL Constructs & Purpose

- **CASE ... END** — creates the age_group hierarchy, allowing aggregates to be computed per group.
- **AVG() + ROUND()** — compute and format group averages (numeric summarization).
- **GROUP BY ... WITH ROLLUP** — produces multiple aggregation levels (detailed group rows + subtotals + grand total row) in a single query, to use roll-up OLAP operation.
- **ORDER BY gender, age_group** — ensures deterministic ordering for display.

How the Dashboard Application uses the Queries / OLAP

The application uses filters (gender, ageRange), which correspond directly to the WHERE clauses used by the queries to implement different OLAP Operations.

Charts (Filters - ALL, ALL ages):

ScatterChart → Sleep vs Stress Query (Report 1)

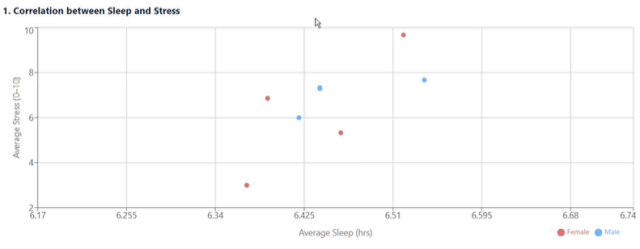

Fig 9. Scatter chart for sleep vs stress

RadarChart → Mental Risk Indicators (Report 2)

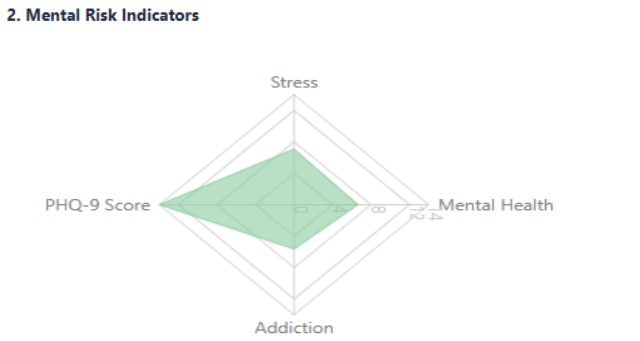

Fig 10. Radar chart for mental risk indicators

BarChart → Social Media vs GPA (Report 3)

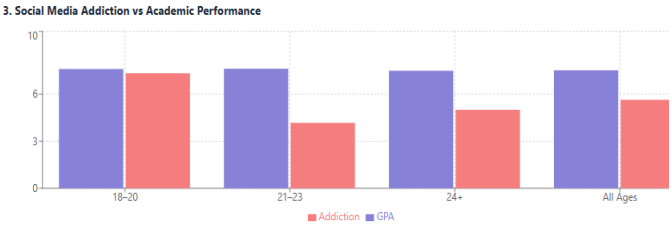

Fig 11. Radar chart for mental risk indicators

Line/Bar → Gender & Age comparisons (Report 4)

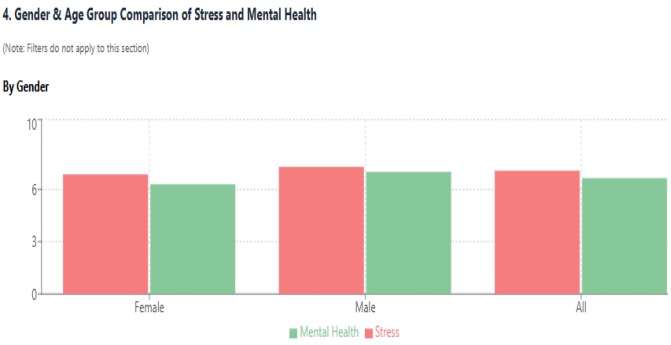

Fig 12. Bar chart for stress and mental health by gender

Fig 13. Bar chart for stress and mental health by age group

Line/Bar → Social Media Addiction and Mental Health Trends (Report 5)

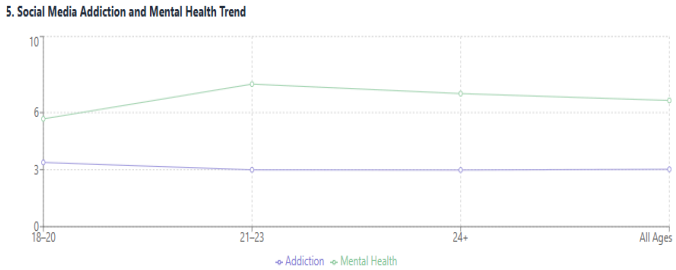

Fig 14. Line chart for social media addiction and mental health

5. Query Processing, Optimization, Results, and Analysis

5.1 Functional Testing

Test Case	Query	Expected Output	Actual Result	Status
Validate dataset extraction	Load all datasets	All loaded with no file errors	Loaded	Pass
Validate data and preprocess	Apply cleaning pipeline	Cleaned with renamed columns and no missing values	All cleaned	Pass
Validate MySQL load	Write fact and dim tables to olap_dashboard	Tables created with Primary Keys	Verified locally in Workbench	Pass
Test API endpoints	Run each endpoints	Returns JSON response with no SQL errors	Returns rows/data	Pass
Test Frontend using filters	Change gender/age filters	Dashboard dynamically updates graphs and summaries	Behavior matched manual SQL validation in Workbench	Pass

The functional testing was performed during the implementation of the ETL and OLAP pipeline: dataset loading, preprocessing, data warehouse integration, API responses, and frontend validation. The expected and actual results were consistent across all stages and passed successfully, confirming the correctness of our implementation.

5.2 Performance Testing

Query	Execution Time
Sleep-Stress	0.00059300
Mental Health Indicators	0.00040500
Social Media Impact	0.00042275
Gender-Age Comparison of Stress and Mental Health	0.00051650 0.00045200
Social Media Addiction and Mental Health	0.00053050

The performance testing was conducted in MySQL Workbench to evaluate the query performance. The tests involved the five OLAP analytical queries from the dashboard and used SET profiling = 1 to record the execution times. The resulting execution times ranged between 0.0004 - 0.0006 seconds for all queries. The results indicate that the system was already performing quite well.

Because the ETL pipeline already pre-aggregated and normalized the data into the resulting fact table "fact_student_summary", the data volume was greatly reduced before queries were executed.

Since the queries only involved simple groupings and averages, query reformulation was considered but deemed unnecessary. Index optimization was another strategy the team considered implementing but these were found to be redundant as the tables were already using age and gender as primary keys which were already defined during the ETL's loading phase.

As a result, no major restructuring and query reformulation was needed as most of the optimization goals the team had were already achieved through the ETL and schema design.

6. Conclusion

Project Summary

The project workflow follows the instructions from the machine specifications, where each member is assigned a major part to distribute the overall workload. There was one developer assigned for the schema design, ETL script, OLAP application, and query optimization.

The process of building and maintaining a data warehouse allows developers to analyze and combine datasets from different sources, at a deeper and broader level, which couldn't be achieved otherwise. It reveals significant factors, such as, in this case, mental health risk indicators amongst different student groups, such as PHQ-9 ratings, stress levels, and addiction scores.

Data Warehouses enable researchers to handle large volumes of data and perform significantly more robust queries than one could do with a regular database. The datasets used for the project, which had the psychiatric records of collectively thousands of individuals, would be extremely difficult to perform regular SQL Queries on.

The ETL script can be automated using GitHub Actions that can be scheduled weekly or monthly. This process updates the data warehouse when the datasets update or change, as the ETL script also provides support by downloading them directly from the sites.

OLAP Applications are used for performing fast, robust, and multidimensional analyses of large volumes of data, which they excel in due to their optimization for data retrieval and analysis. This is exemplified by the obtained reports from the OLAP Application, as it yielded results that would be more difficult and time-consuming to obtain if a regular database was used instead.

Query optimization is needed because it makes queries run faster, more efficient and less resource intensive as the data warehouse grows and becomes more complex.

There are many strategies with the most important being the effective and proper use of indexes, which allows databases to locate data more quickly. Other techniques include optimizing JOINs, caching, making efficient and smart ETL decisions, and avoiding the use of SELECTs to improve the systems overall performance.

The findings of the experiments offer significant insights into the relationship between the mental health of different demographic groups and various factors. In particular, the experiments confirm that extensive stress, social media addiction, and a lack of consistent sleep are all contributing to the stark decline in mental health among students.

7. References

[1] Balla, E., and Panaretou, K. 2023. A Review of Dashboard Design and Visualization for Mental Health and Public Health Data. Frontiers in Public Health, 11 (2023), Article 10192578. https://pmc.ncbi.nlm.nih.gov/articles/PMC10192578/

[2] codebasics. (2020, February 15). What is ETL | What is data warehouse | OLTP vs OLAP [Video]. YouTube. https://www.youtube.com/watch?v=oF_2uDb7DvQ

[3] DataCamp. 2023. *MySQL ROLLUP: How to Use ROLLUP for Subtotals and Grand Totals*. DataCamp Inc., New York, NY. Available at https://www.datacamp.com/doc/mysql/mysql-rollup

[4] Han, J., Kamber, M., and Pei, J. 2011. *OLAP and Data Cube Technology for Data Warehousing.* Lecture Notes for CS412 – Data Mining, University of Illinois at Urbana-Champaign. https://hanj.cs.illinois.edu/cs412/bk3_slides/04OLAP.pdf

[5] Legaspi, C. (n.d.). STADVDB Slides 03A - Data wrangling and visualization [Slide show]. Google Docs. https://docs.google.com/presentation/d/1rd6Sx0oDlnBjq0z4o_L73S6Iy-yXj8s6/

[6] Oracle.. 2024. *MySQL 8.4 Reference Manual: GROUP BY Modifiers (ROLLUP, CUBE, GROUPING SETS)*. Oracle Corporation, Austin, TX. https://dev.mysql.com/doc/refman/8.4/en/group-by-modifiers.html

[7] O&B. (n.d.). DB-201 Database Optimization [Slide show]. https://docs.google.com/presentation/d/1y9wXAcPtmvdr-2ApS773v6AASNxobc9gUBk27XIN4LI/

[8] Recharts. 2024. *Recharts: A composable charting library built on React components.* Retrieved from https://recharts.org

[9] React. 2024. *React – A JavaScript library for building user interfaces.* Meta Platforms, Inc. Retrieved from https://react.dev

[10] SmartSurvey. (n.d.). Age groups for surveys. https://www.smartsurvey.co.uk/survey-questions/demographics/age-groups

8. Declarations

Throughout the project development, the utilization of AI tools improved and quickened the implementation process for each step of the specifications.

This use has also allowed developers to expand their knowledge beyond the lecture materials, which was applied to parts such as the ETL script, OLAP application, and query optimization.

8.1. Declaration of Generative AI Usage.

In this project, the developers used the following AI tools for the following tasks:

GitHub Copilot	Assisted with the development of the ETL script by answering prompts, including: - Suggest an ETL workflow. - Show the pros and cons of different normalization and standardization techniques. - Explain one-hot encoding and mapping concepts for transforming categorical values to numerical ones.
ChatGPT	Assisted in improving grammar, phrasing, and coherence in written documentation. Provided guidance on backend development steps, including Express.js API structuring.

After utilizing the mentioned tools, the developers reviewed and edited the AI responses and took responsibility for discerning and understanding them to be explained properly during the presentation.

8.2. Record of Contribution.

Sean Cardeno: Managed functional testing, optimization, and debugging. Contributed to the "Query Processing, Optimization, Results and Analysis" and "Conclusion" chapters.

Jeff Fonte: Worked with the schematic design, then proposed and developed the ETL pipeline and script to be used for transforming and aggregating the datasets before loading into a data warehouse. Contributed to the "Introduction", and "ETL Script" chapters.

Aurelio Garcia: Worked with the schematic design and the loading of data into the data warehouse. Handled the "Introduction" and "Data Warehouse" chapters.

Margaux Siongco: Developed the OLAP Application and produced the data visualizations for each OLAP Query. Handled to the "OLAP Application" and "Analytical Reports" chapters.