Project Report

Bronze Layer

Introduction

The Bronze Layer project focuses on integrating and analyzing data from various tables related to health camps. The project aims to provide insights into patient attendance, health scores, and donation patterns across different health camps.

Data Loading

Health_Camp_Detail: This table contains information about health camps, including the camp ID, start and end dates, and camp categories.

Patient_Profile: This table provides details about patients, such as patient ID, online presence (followers on various platforms), income, education score, age, and other demographic information.

First_Health_Camp_Attended: This table includes details of patients who attended the first health camp, including their patient ID, health camp ID, donation amount, and health score.

Second_Health_Camp_Attended: This table contains similar information as the first health camp table but for the second health camp. To align the columns for both tables, the columns in the second_health_camp_Attended table were reordered to match the sequence of the First_Health_Camp_Attended table.

Data Integration

The tables were integrated using a union operation to create a single table containing data from both the first and second health camps. This integration allows for comprehensive analysis across all health camps.

Conclusion

In the Bronze Layer of our project, we successfully loaded and prepared the data from multiple tables related to health camps. This foundational step is crucial for ensuring that the data is clean, standardized, and ready for further analysis. By loading and organizing the data from tables such as Health_Camp_Detail, Patient_Profile, First_Health_Camp_Attended, and Second_Health_Camp_Attended,

Silver Layer

Introduction

The Silver Layer of the project builds upon the Bronze Layer by further refining the data integration process and focusing on patient-centric analysis. This layer aims to identify and analyze patient data across different health camps, providing deeper insights into patient behavior and engagement.

Data Refinement

Check for Common Health Camp IDs: Using an inner join between the First_Health_Camp_Attended and Second_Health_Camp_Attended tables, common health camp IDs were identified. This step ensures that there are no duplicate health camp IDs between the two tables, as they will be unioned later.

Duplicate Record Removal: Duplicate records were removed from all tables to ensure data cleanliness and accuracy.

Adding Camp Identifier: A new column named camp was added to both the First_Health_Camp_Attended and Second_Health_Camp_Attended tables. The value "A" was assigned to the camp column in the First_Health_Camp_Attended, representing the First Camp, and "B" was assigned to the Second_Health_Camp_Attended tables, representing the second camp.

Data Integration and Joining

Union of Health Camp Attendance Tables: The First_Health_Camp_Attended and Second_Health_Camp_Attended tables were unioned to create a single table containing attendance data from both camps.

Joining with Health Camp Details: The unioned attendance table was left-joined with the Health_Camp_Detail table using the Health_Camp_ID as the key. This step adds detailed information about each health camp to the attendance data.

Distinct Patient IDs: Distinct patient IDs were identified in the left_joined_data just for checking how many Distinct patient IDs are present in the left_joined_data

Joining with Patient Profile: The joined data was then right-joined with the Patient_Profile table to include detailed patient information. This step ensures that all patients, including those who did not attend any health camp, are included in the analysis.

Conclusion

The Silver Layer enhances the data integration process by refining patient-centric analysis and providing a more comprehensive view of patient behavior across different health camps. This layer sets the foundation for further analysis and insights in the project. At the conclusion of our data processing, we have obtained the "sample_data" table, which serves as the foundational dataset resulting from the integration of all relevant tables.

Gold Layer: Data Transformation to Abacus Data Model Format

In the Gold Layer of our project, we focused on transforming the sample_data table into the Abacus Data Model format. This involved selecting specific columns required for the model and applying various transformations to ensure consistency and usability of the data. Below are the transformations applied to the selected columns:

Patient_ID: Retained as is.

Health_Camp_ID: Retained as is.

Camp_Start_Date: Converted to "dd/MM/yyyy" format. If the data is null, it is left unchanged.

Camp_End_Date: Converted to "dd/MM/yyyy" format. If the data is null, it is left unchanged.

Category1, **Category2**, **Category3**: Retained as is.

Donation: Retained as is.

Health_Score: Converted to "%f" format with up to 6 decimal places and cast to float format.

Online_Follower, **LinkedIn_Shared, Twitter_Shared, Facebook_Shared:** Retained as is.

Income: If the value is "None", it is replaced with "Unemployed"; otherwise, it is kept as is.

Education_Score: If the value is "None", it is replaced with "Uneducated"; otherwise, it is kept as is.

First_Interaction: Converted to "dd/MM/yyyy" format. If the data is null, it is left unchanged.

City_Type: Retained as is.

Employer_Category: Retained as is.

camp: Retained as is.

By applying these transformations, we have prepared the data from the sample_data table to conform to the Abacus Data Model format, ensuring consistency and compatibility for further analysis and integration into the Abacus system . At the conclusion of our data processing, we have obtained the "final_table" table, which serves as the foundational dataset resulting from the integration of all relevant tables.

SAM Layer: Data Transformation for Client Requirements

In the SAM Layer, we focused on fulfilling the client's specific requirements by creating a table that meets their format and includes only the fields they need. The transformations were applied to the final_table to create a new table called sam_attendance, which includes information about the patient's attendance at different health camps. We also created a field called Attendance_info to categorize the attendance status of each patient.

Attendance_info Creation:

- Selected columns "Patient_ID", "Health_Camp_ID", and "Camp" from final_table.
- Used group by and aggregate functions to create a new field called grouped_camp, which defines how many and which camps each patient attended.
- Created the Attendance_info table based on the grouped_camp field to categorize the attendance status of each patient as "Attended both camps", "Attended camp A", "Attended camp B", or "Not attended any camp".
- Dropped the grouped_camp field to obtain the Attendance_info table, which includes "Patient_ID" and "Attendance_Status".

Patient Details Selection :

sam1

from final_table.

- Converted "Camp_Start_Date", "Camp_End_Date", and "First_Interaction" to to_date ("dd/MM/yyyy") format for doing operations on date ,if the data is not null.
- Calculated the "Health_Score_Difference" based on the difference between the last and first health scores for each patient . after arranging “Camp_Start_Date” in ascending format

And other fields as it is.

sam2

using sam1 I created new table sam2

- Added a row number to each row partitioned by "Patient_ID" and ordered by "Camp_Start_Date" in descending order.
- Filtered only the rows where the row number is 1 to get the “health_score_difference” for each patient because it sam1 giving running difference and I only want last “health_score_difference”.

Patient Details Selection :

patient_details

Selected specific fields required for patient details from sam2.

- "Patient_ID",
- "Online_Follower",
- "LinkedIn_Shared",
- "Twitter_Shared",
- "Facebook_Shared",
- "Income",
- "Education_Score",
- "Age",
- "First_Interaction",
- "City_Type",
- "Employer_Category",
- Formatted the "Health_Score_Difference" field to %.6f format.

Camp Info Aggregation :

camp_info

- Grouped sam1 by "Patient_ID" and "camp" (or "none" if "camp" is null).
- Aggregated camp details including "Health_Camp_ID", "Camp_Start_Date", "Camp_End_Date", "Health_Score", "Category1", "Category2", "Category3", and "Donation" into a struct.
- Aggregated the camp details into a map with "camp" as the key and the list of camp details as the value.

Output_table

- After completing the transformations in the SAM Layer, the final step involved joining the camp_info, patient_details, and Attendance_info tables to create the output_table.
- This table combines all the necessary information about patient attendance, health scores, and camp details in the required format.
- The join was performed using inner joins on the "Patient_ID" column to ensure that only patients with complete information in all tables are included in the final output.

Output_table2

- To include only those patients who attended health camps, we filtered the rows to retain only those with a non-null health score.

Uploading And Downloading file (AWS s3)

- Using the boto3 library, we successfully uploaded and downloaded our DataFrame in CSV format.

Based on the final analysis:

- There are 29,363 patients who did not attend any health camp, which is 78.02% of the total patients.
- There are 2,147 patients who attended the first health camp, accounting for 5.71% of the total patients.
- There are 4,722 patients who attended the second health camp, representing 12.55% of the total patients.
- There are 1,401 patients who attended both health camps, making up 3.72% of the total patients.

In conclusion, approximately 78% of patients did not attend any health camp, while only 22% attended at least one health camp.