ETL CUSTOMER DATA PROJECT

Author: Hien Tran, github:

**Contents**

[**Summary 1**](#_8hqqgbwzvbd6)

[**Extract 1**](#_vjsgldkd2xd1)

[**Transform 2**](#_rxc9mitmoun)

[Processing Log Content Data 2](#_d63fwt5xn4ej)

[Processing Log Search Data 2](#_wxr2lzkltrb)

[Unionizing Data 2](#_49x00fvm5g34)

[**Load 3**](#_1nl94c5ak5yr)

[Visualize 3](#_2osxuwu2hp2k)

# Summary

This project involves the development of the Customer360 model, an analytical CRM system designed to deliver a comprehensive view of customer interactions and behaviors. The model integrates two key datasets: **log\_content** (CSV format containing interaction data) and **log\_search** (Parquet format containing behavioral data). The ETL (Extract, Transform, Load) process is implemented using PySpark to cleanse, transform, and enhance the data, followed by consolidation into a CSV file and storage in an Azure MySQL database. The data is then visualized using Power BI, providing stakeholders with interactive dashboards that facilitate data-driven decision-making.

# Extract

In the extraction phase, data is retrieved from two primary sources:

* **Log Content Data (CSV)**: Contains records of customer interactions.
* **Log Search Data (Parquet)**: Captures customer behavioral data, such as search queries.

Using PySpark, data from each source is extracted within a user-defined date range. The extracted datasets (**log\_content** and **log\_search**) are preserved as distinct entities, setting the stage for further processing and eventual union during the transformation phase.

# Transform

The transformation phase utilizes a PySpark script to perform several key operations aimed at cleansing, transforming, and enriching the extracted data. Below is a summary of the transformation logic:

(For more information about the columns, go to data\_dictionary: [data\_dictionary](https://docs.google.com/spreadsheets/d/1DYHyV6LdsOXG8MlbxhJC1crQRA9xBm4RZ84mGTEhM14/edit?gid=0#gid=0))

## Processing Log Content Data

**Calculating Total Devices**: This function calculates the number of unique devices associated with each customer.

**Calculating Activeness**: An "Activeness" metric is computed, representing the number of days a customer has interacted with the system.

**Transforming Category**: The data is transformed to categorize different types of content.

**Calculating Statistics**: Various statistics are computed, summarizing customer total watch values for each **Category**.

**Calculating MostWatch and CustomerTaste**: These functions determine the most-watched content type and derive a "CustomerTaste" metric, which reflects customers’ preferences.

**Adding Activeness and TotalDevices**: The DataFrame is enriched by integrating the calculated "Activeness" and "TotalDevices" columns.

**Calculating Customer Type**: Customers are categorized into tiers (e.g., Bronze, Silver, Gold, Diamond) based on their Activeness patterns and total watch values.

**Renaming Columns**: Specific columns are renamed to indicate that they represent total values (e.g., **Giai\_tri** renamed to **Total\_Giai\_tri**).

## Processing Log Search Data

**Filtering Rows**: Rows with null values, duplicates, and irrelevant months are filtered out.

**Finding Most Searched Keywords**: The script identifies the most searched keywords for each user, with a focus on specific months (e.g., June and July).

**Reading Mapping File**: A mapping file is loaded, which maps search keywords to specific categories.

**Categorizing Searches**: Searches are categorized based on the mapping file, enriching the dataset with contextual information.

**Determining Trending Type and Previous Search**: Additional columns are added to capture trending search types and previous search behavior.

## Unionizing Data

Finally, the transformed **log\_content** and **log\_search** datasets are unionized based on **user\_id** and **contract**, resulting in a comprehensive table that integrates both interaction and search data.

# Load

The final dataset is loaded into an Azure MySQL database and also saved as a CSV file, ensuring accessibility for further analysis and reporting.

# Visualize

Power BI is utilized to create interactive dashboards that track and visualize key customer insights. These dashboards empower stakeholders to make informed, data-driven decisions.

# 