**DATA PROFILING TOOL**

**Overview:**

The **Data Profiling Tool** simplifies data quality analysis within a Snowflake database. Users provide the *database, schema, and table* name as parameters, and the tool runs a pre-defined procedure to profile the data. It generates insightful visualizations to evaluate the structure, content, and quality of the data, helping identify issues like missing values or inconsistencies for improved data management.

*Note: A sample “CUSTOMER” table is pre-populated with the package in the DATAPROFILER.CORE schema to understand the workings of the application.*

A close up of a screen

Description automatically generated

**Components:**

1. **Choose Database**: Select the database to be analysed.

2. **Choose Schema**: Specify the schema within the selected database.

3. **Choose Table**: Select the specific table for data profiling.

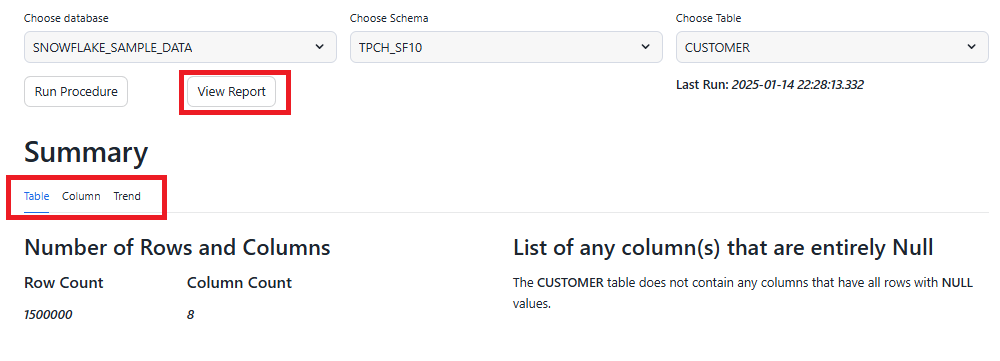
4. **Run Procedure**: Initiate the data profiling procedure for the chosen table.

5. **View Report**: Access the generated report for the data profiling results.

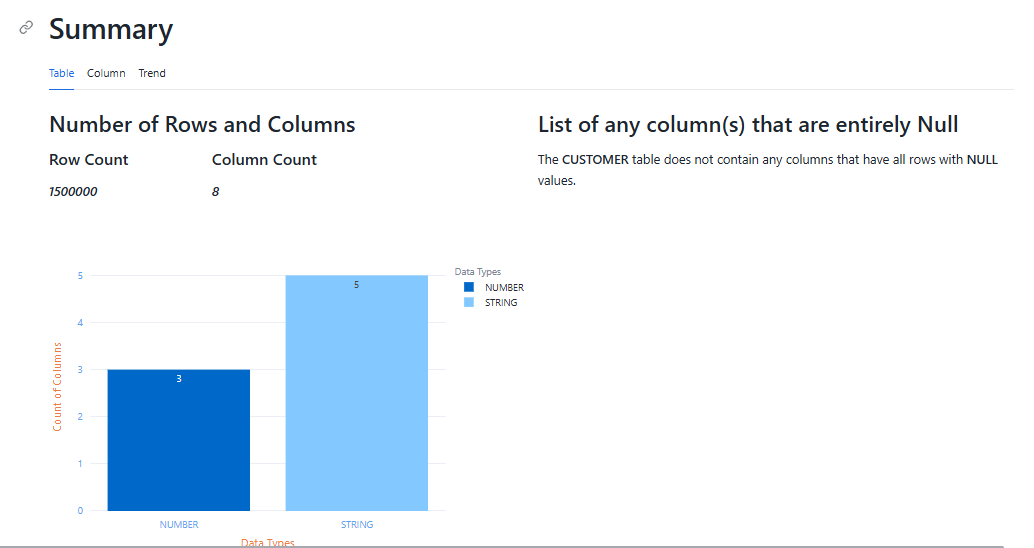
**Last Run Status:**

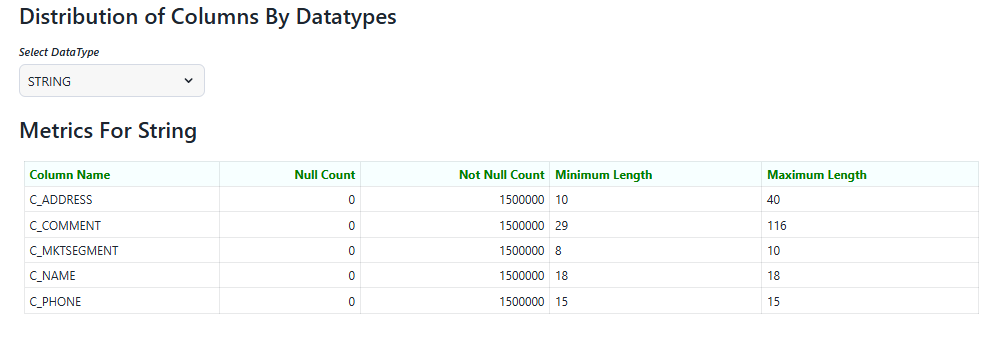
It displays the last run status for a specific table, indicating whether any data profiling procedures have been executed for that table.

**Table Summarization:**



If the table has already been processed using the procedure, clicking **View Report** will display a summary. This summary is organized into three sections: *Table, Column, and Trend*. Each section provides detailed insights, which are explained briefly below.





The **Table** section offers a comprehensive overview of the data, starting with the total count of rows and columns. It includes a bar chart that visualizes the Distribution of Columns by Datatypes, providing a breakdown of columns based on their data types. The report also highlights Columns That Are Completely Null, ensuring the table does not contain columns entirely filled with NULL values. This visualization provides a clear understanding of the table's structure and data distribution.

Additionally, the table visualization includes an option to filter columns by data type. Based on the selected data type, corresponding columns and their attributes are displayed for further exploration.

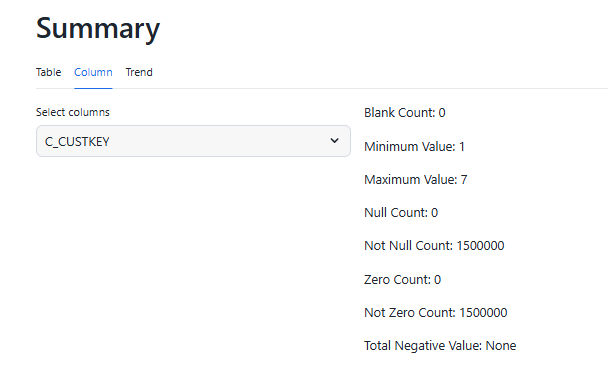
A screen shot of a computer

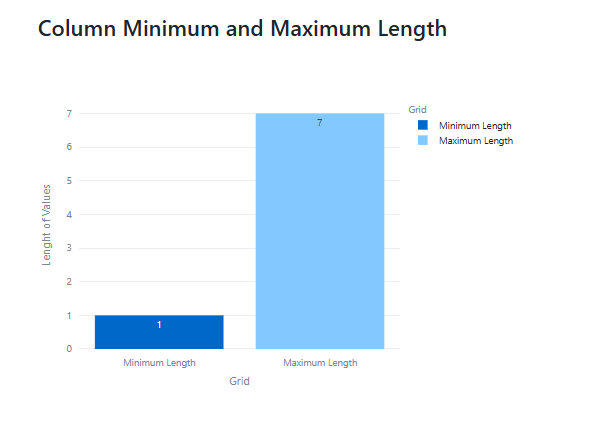
Description automatically generated

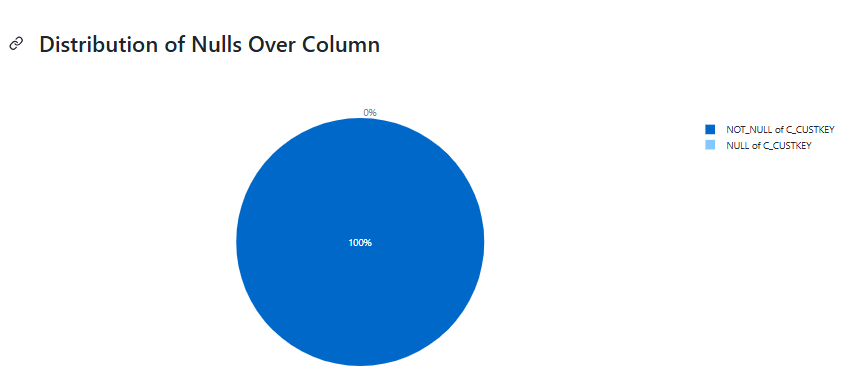
The **Recent Record Count Trend** visualization presents the number of rows affected during the last 5 procedure executions for the selected table. It provides a clear view of the row count over these recent runs, highlighting any changes or consistency. In this case, it shows a stable record count with no changes, reflecting the data’s stability across the executed procedures. This trend is useful for monitoring data consistency and detecting any unexpected variations in row counts.

**Column Summarization:**

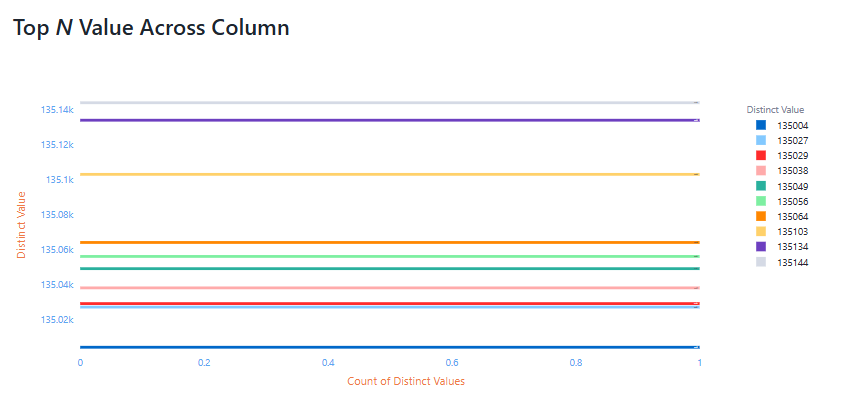
The **Column** section provides an overview of the selected column, summarizing key data characteristics such as the absence of *null, blank*, or *special characters*. It highlights the range of text lengths and ensures all entries are consistent, containing only valid alphabetic data. A bar graph visually represents the variation in *text length*, confirming the column is well-structured for further analysis.







The **Distribution of Null Over Columns** pie chart illustrates the proportion of *NULL and NOT NULL* values in the selected column.



Additionally, the **Top N Values Across Columns** visualization highlights the most frequently occurring values in the column.

**Trend Summarization:**

The **Trend** section highlights changes in selected metrics for a chosen column over recent profile runs. It allows users to track metrics like null counts, distinct values, or length variations, visualizing trends over time. This helps monitor data quality and consistency across multiple profiling sessions.

A screenshot of a computer

Description automatically generated