MYSQL DATA CLEANING

Project by
Hrishikesh Helge

Layoffs 2022: Enhancing Data Integrity Through Data Cleaning

Overview

Objectives

Data Overview

Methodology

Steps

Challenges

Conclusion

Objectives

ELIMINATE DUPLICATES
Streamline the dataset by removing redundant entries.

STANDARDIZE FORMATS
Ensure uniform data formats to simplify analysis.

ADDRESS MISSING DATA
Handle incomplete entries to maintain data integrity.

IMPROVE CONSISTENCY
Align data entries with standardized categories for uniformity.

DOCUMENT PROCESSES
Clearly record the data cleaning steps for transparency and reproducibility.

ENHANCE DATA ACCURACY
Correct any errors or discrepancies to ensure the reliability of the dataset.

Data Overview

```
+---------------------------------+------+
| Field                           | Type |
+---------------------------------+------+
| company                         | text |
| location                        | text |
| industry                        | text |
| total_laid_off                  | int  |
| percentage_laid_off             | text |
| date                            | text |
| stage                           | text |
| country                         | text |
| funds_raised_millions           | int  |
+---------------------------------+------+
```

Methodology

STAGING TABLES

WINDOW FUNCTIONS

GENERATING NULL VALUES

SQL JOINS

DATA TYPE CONVERSION

ALTER AND UPDATE TABLE

Steps

STEP 01

REMOVE DUPLICATES

Identify and eliminate duplicate records from datasets.

STEP 02

STANDARDIZE THE DATA

Correct spelling errors and ensure consistent formatting across entries.

STEP 03

DEAL WITH NULL VALUES

Handle missing data by using appropriate methods like imputation or deletion.

REMOVE DUPLICATES

Missing Primary Key: The absence of a primary key in the original table made it challenging to identify and address data redundancy efficiently.

Solution: We introduced a ROW_NUMBER() window function, partitioned by all columns, to assign a unique row_num to each record. A value of 1 in the row_num column indicates a unique record, while values greater than 1 highlight redundant entries.

Data Cleanup: We removed the redundant data from the staging table, ensuring the original dataset remained unchanged.

STANDARDIZE THE DATA

Industry Column: We identified that variations such as "Crypto," "Crypto Currency," and "CryptoCurrency" referred to the same industry. To standardize this, we used the LIKE and UPDATE clauses to ensure consistency across the column.

Country Column: The "United States" was entered inconsistently, with variations such as "United States" & "United States." We standardized the country column by applying UPDATE statements to unify these entries.

Date Column: The date column was originally stored as text (VARCHAR) instead of the proper DATE format. We corrected this by using the STRING_TO_DATE function to standardize the data type.

DEAL WITH NULL VALUES

Industry Table: We found missing, null, or blank values in the industry column for some companies, while others were correctly labeled (e.g., "Airbnb: Travel," "Carvana: Transportation," "Juul: Consumer"). To ensure consistency, we used an inner join to fill in the missing industry values.

Handling Null Values: The total_laid_off and percentage_laid_off columns had over 300 null entries out of 2,300. Without employee count data, we couldn't calculate one from the other. Thus, we removed rows where both columns were null.

Trade-Off Analysis: Removing 300 rows had little impact on data quantity but greatly improved data quality, leading to more reliable analysis.

CHALLENGES

LACK OF PRIMARY KEY

IDENTIFYING DUPLICATE RECORDS

DECISION ON DATA RETENTION

HANDLING MISSING DATA

INCONSISTENT DATA FORMATTING

DOCUMENTING THE PROCESS

CONCLUSION

KEY LESSONS LEARNED

The importance of thorough data cleaning for quality and the need for clear documentation to ensure transparency and reproducibility.

AENHANCED DATA QUALITY

The cleaned dataset exhibits improved integrity, consistency, and accuracy, making it a robust foundation for further analysis.

IMPROVED ANALYTICAL READINESS

With the dataset now carefully structured and ready for comprehensive analysis, it supports enhanced decision-making and more effective strategic planning.

EFFICIENCY GAINS

By streamlining the data, the project has reduced potential bottlenecks in the analysis phase, leading to more efficient data processing.

FUTURE RECOMMENDATIONS

Regular data audits and cleaning should be integrated into the data management process to maintain high standards of data integrity.

SUCCESSFUL DATA CLEANING

The data cleaning process removed duplicates, standardized formats, addressed missing values, and eliminated unnecessary data, resulting in a more reliable dataset.