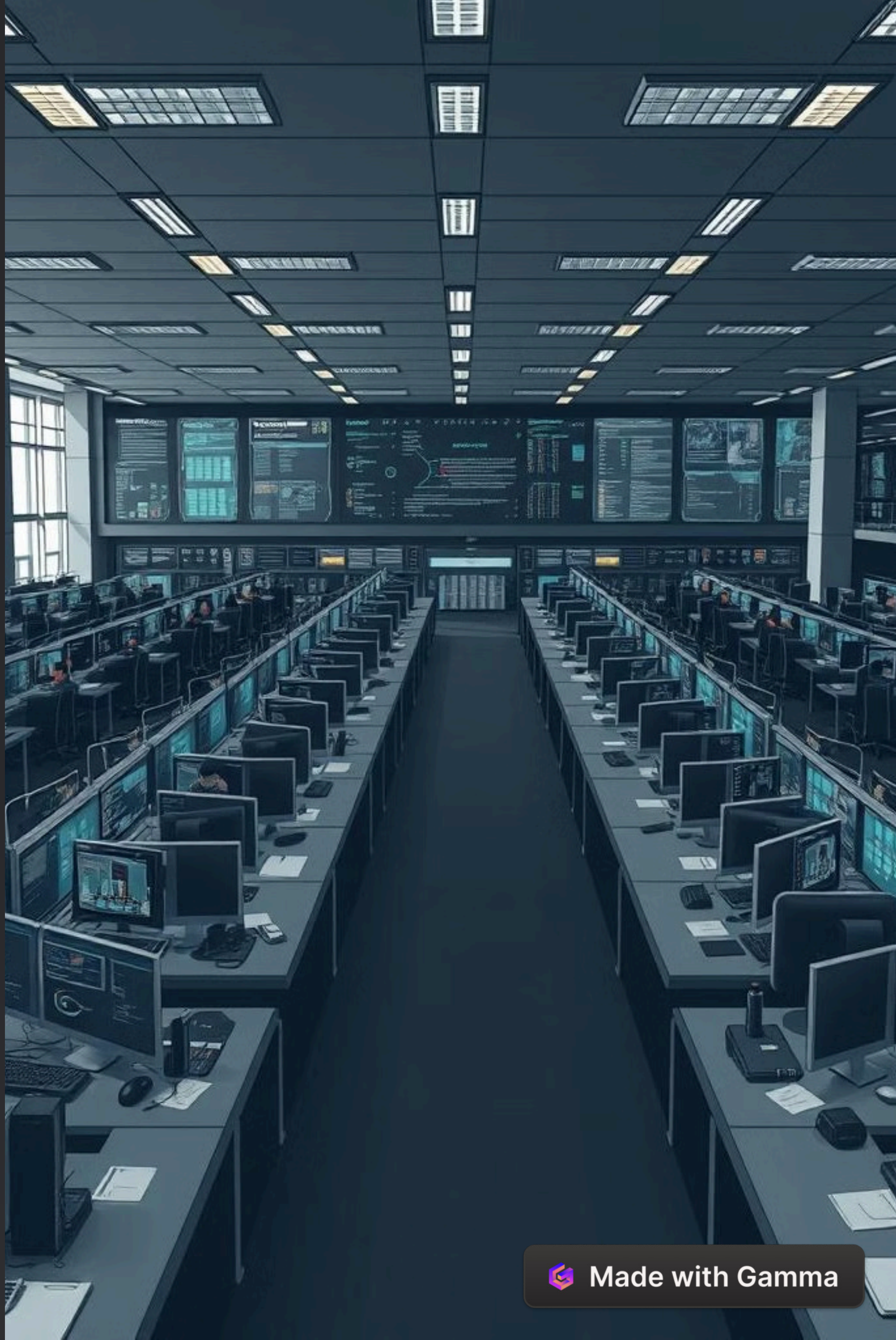
# SQL Project - Data Cleaning

This presentation will guide you through the process of cleaning and analyzing a dataset of layoffs from 2022. We will use SQL to identify and address data inconsistencies, explore trends, and gain insights from the data.

# Introduction to the Dataset

Tech firms around the globe are fighting the economic slowdown. The slow consumer spending, higher interest rates by central banks and strong dollars overseas are hinting towards possible recession and tech firms have started laying employees off. This economic slowdown has made Meta recently fire 13% of its workforce, which amounts to more than 11,000 employees.



# Data Cleaning: Removing Duplicates and Standardizing Data

## Removing Duplicates

We start by identifying and removing duplicate entries in the dataset. This involves using SQL queries to identify rows with identical values across key columns. We then use a DELETE statement to remove these duplicates, ensuring data integrity.

## Standardizing Data

Next, we standardize data by addressing inconsistencies in values. This includes converting empty strings to NULL values, updating industry names to a consistent format, and correcting inconsistencies in country names. We use UPDATE statements to modify the data in the staging table.

# Data Cleaning: Removing Null Values

When working with a dataset in SQL, it's crucial to handle null values appropriately to ensure the accuracy and reliability of your analysis. Here are some common ways to deal with null values:

- **Identifying Null Values:** Use SQL queries to identify columns with null values by running queries like `SELECT * FROM table_name WHERE column_name IS NULL`.
- **Counting Null Values:** Determine the number of null values in each column using SQL functions like `COUNT(*)` or `SUM(CASE WHEN column_name IS NULL THEN 1 ELSE 0 END)`.
- **Handling Null Values:** Decide on the best approach to handle null values, which may include filling them with a default value, removing rows with null values, or imputing missing values based on certain criteria.

# Data Cleaning: Removing any columns and rows that are not necessary

In SQL, it's essential to streamline your dataset by removing any columns or rows that are not necessary for your analysis. Here are a few ways to achieve this:

## Removing Unnecessary Columns:

- **Using `SELECT` Statement:** Specify only the required columns in your `SELECT` statement to exclude unnecessary columns.

```
SELECT column1, column2, column3

FROM table_name
```

- **Dropping Columns:** If certain columns are entirely unnecessary, you can drop them from the table using the `ALTER TABLE` statement.

```
ALTER TABLE table_name

DROP COLUMN column_name
```

## Removing Unnecessary Rows:

- **Filtering Rows:** Use `WHERE` clause to filter out rows that are not needed for your analysis.

```
SELECT *

FROM table_name

WHERE condition
```

By looking at null values and removing unnecessary columns and rows, you can ensure that your SQL analysis is focused, accurate, and efficient.

# Exploratory Data Analysis (EDA) with SQL

### Identifying Trends

We use SQL queries to explore the cleaned data and identify trends and patterns. This includes finding the companies with the most layoffs, the locations with the highest number of layoffs, and the industries most affected by layoffs.

### Analyzing Layoffs Over Time

We analyze the data to understand how layoffs have changed over time. We use SQL queries to calculate the total number of layoffs per year and per month, and to create a rolling total of layoffs over time.

# Identifying Trends

1 Find the companies with the most layoffs.

2 Identify the locations with the highest number of layoffs.

3 Discover the industries most affected by layoffs.

4 If we order by funds raised millions we can see how big some of these companies were.

5 Find Companies with the most Layoffs. Now let's look at that per year.

6 Let's see the rolling Total of Layoffs Per Month.

# Identifying Trends

- Find the companies with the most layoffs.

```
SELECT TOP(10)company, SUM(total_laid_off) AS [Sum_of_total_laid_off]

FROM layoffs_staging

GROUP BY company

ORDER BY 2 DESC
```

# Identifying Trends

Identify the locations with the highest number of layoffs.

```
SELECT TOP(10)location, SUM(total_laid_off) AS [Sum_of_total_laid_off]

FROM layoffs_staging

GROUP BY location

 ORDER BY 2 DESC
```

# Identifying Trends

Discover the industries most affected by layoffs.

```
SELECT TOP(10)industry, SUM(total_laid_off) AS [Sum_of_total_laid_off]

FROM layoffs_staging

GROUP BY industry

ORDER BY 2 DESC
```

# Identifying Trends

If we order by funds raised millions we can see how big some of these companies were.

```
SELECT *

FROM layoffs_staging

WHERE percentage_laid_off = 1

ORDER BY funds_raised_millions DESC
```

# Identifying Trends

- Find Companies with the most Layoffs. Now let's look at that per year.

```
WITH Company_Year AS (
  SELECT company, YEAR(date) AS years, SUM(total_laid_off) AS Sum_of_total_laid_off
  FROM layoffs_staging
  GROUP BY company, YEAR(date)
), Company_Year_Rank AS (
  SELECT company, years, Sum_of_total_laid_off,
          DENSE_RANK() OVER(PARTITION BY years ORDER BY Sum_of_total_laid_off DESC) AS ranking
  FROM Company_Year
)
SELECT *
FROM Company_Year_Rank
WHERE ranking <= 5 AND years IS NOT NULL
ORDER BY years ASC, Sum_of_total_laid_off DESC
```

# Identifying Trends

Let's see the rolling Total of Layoffs Per Month.

```
WITH DATE_CTE AS (
    SELECT SUBSTRING(CONVERT(VARCHAR, DATE, 120), 1, 7) AS MonthYear, SUM(total_laid_off) AS TotalLaidOff
    FROM layoffs_staging
    GROUP BY SUBSTRING(CONVERT(VARCHAR, DATE, 120), 1, 7)
)
SELECT MonthYear, (SELECT SUM(TotalLaidOff) FROM DATE_CTE t2 WHERE t2.MonthYear <= t1.MonthYear) AS RollingTotalLayoffs
FROM DATE_CTE t1
ORDER BY MonthYear
```

# Insights From Dataset

- The highest company that laid off employees was **Amazon**, with a total number of **18150** employees.
- The largest city in which the largest number of employees were laid off was **SF Bay Area**, with a total number of employees amounting to **125631** employees.
- The largest industry in which the largest number of employees are laid off is in **Consumer**, with the total number of employees reaching **45182** employees.
- **BritishVolt** sounds like an electric car company, I know this company raised almost **$2 billion** and then collapsed.
- It appears that **January 2023** was the month in which the most employees were laid off, reaching **84,714** employees.

# Thank You

Thank you for taking the time to view my project.

I hope you learned something new from this.

Connect with me on LinkedIn: https://www.linkedin.com/in/muhammed-imam/