# 

# Data Science Project

# Submitted By:

# Danish Sarwar (SP23-BCS-030)

# Usman Ahmad (SP23-BCS-031)

# Ahmad Ashraf (SP23-BCS-047)

# Submitted To: Mr. Fakhar Mustafa

# Course Title: Introduction to Data Science

# Department of Computer Science

# COMSATS University Islamabad, Sahiwal Campus

# Data Importing, Cleaning and Transformation

## Dataset Description:

The dataset used in this project focuses on cybersecurity breach incidents primarily within different public and private sectors of the United States. Each record represents a distinct breach event and provides critical details such as the name of the affected organization, type of breach, date of occurrence, number of individuals affected, and the location of breached information.

Key attributes include:

* **Name of Covered Entity:** Organization involved in the data breach.
* **State:** U.S. state where the breach occurred.
* **Individuals Affected:** Number of individuals whose personal information was exposed.
* **Date of Breach:** The reported date when the breach took place.
* **Type of Breach:** Describes the nature of the security incident (e.g., Theft, Hacking/IT Incident, Loss, Unauthorized Access).
* **Location of Breached Information:** Indicates where the information was stored (e.g., Laptop, Network Server, Paper Records).
* **Date Posted or Updated:** The date the breach was officially listed or last updated in the records.

## Dataset Preparation Challenges:

* High Percentage of Missing Values: Several columns had significant missing values, particularly the “breach\_end” column, which was missing in more than 90% of cases. Such columns were removed to maintain dataset quality.
* Irrelevant and Redundant Columns: The dataset contained an “Unnamed” index column and another column “Business\_Associate\_Involved,” which had mostly “Unknown” values and were dropped.
* Duplicate Records: Duplicate entries likely due to repeated reports of the same breach were removed.
* Inconsistent Date Formats: Several date-related columns used inconsistent formats and were standardized.

## Data Cleaning:

The cleaning phase ensured removal of unnecessary or corrupt data. Key actions included:

* **Dropping Columns:** Removed non-informative and redundant columns such as Unnamed: 0, Business\_Associate\_Involved, and breach\_end.
* **Handling Missing Values:** Columns with over 80% missing data were dropped. Remaining missing values in categorical fields were replaced with “Unknown,” and numeric columns were filled using median values.
* **Duplicate Removal:** Identified and removed all duplicate rows.
* **Data Type Conversion:** Converted date-related fields into datetime format and categorical variables into category types.

## Data transformation:

To enhance analytical usefulness, several transformation steps were applied:

* Date Adjustment: “Date\_of\_Breach” was used as the primary date attribute, with missing dates filled using “breach\_start” or “year.”
* Derivation of New Features: Two new time-based attributes were created:  
   Breach\_Month(1–12)  
   - Breach\_Quarter (1–4)
* Categorical Encoding: Categorical variables like organization name, state, and breach type were label-encoded.
* Aggregation: Computed average number of individuals affected per breach type for quick insights.

## Conclusion:

After cleaning and transformation:

* All irrelevant and redundant fields were removed.
* Dates were standardized and formatted correctly.
* Categorical variables were properly encoded.
* Derived variables for temporal analysis were added.
* Dataset is ready for visualization and further analysis.