Detailed Explanation of ER Pipeline Files

# 1. prepare\_data Notebook

This notebook is responsible for creating synthetic data for testing Entity Resolution (ER) and adding controlled data errors.  
It uses the synthetic data from the test catalog, sets up the data for ER, and then writes it to the Unity Catalog for later reference.  
Key steps:  
- Sampling and defining a starting dataframe using `row\_number()` for unique indexing.  
- Preparing and writing synthetic data to Unity Catalog.  
- Defining and applying various data distortions for testing, using scenarios for phone numbers, birth dates, SSNs, names, etc.  
- Combining these distortions and writing them as a single parquet file.  
- Verifying the data integrity and unique indexes after applying the distortions.

# 2. process\_results\_lev Notebook

This notebook processes the ER matching results and writes them to Unity Catalog.  
Key steps:  
- Loading the matching results from output files.  
- Adding scenario labels to track which scenario the record was involved in.  
- Writing these results to a dedicated results table.  
- Performing a join between original and distorted data to calculate confusion matrix statistics like TP, FP, TN, and FN.  
- Running quality checks to ensure data accuracy and consistency.

# 3. Analysis Notebook

This notebook analyzes the results of ER using confusion matrix statistics, precision, recall, and F1-score.  
It compares rule-based and ML-based approaches, measuring performance metrics like:  
- Accuracy, Precision, Recall, F1 Score  
- Summarizing how many records matched by rules vs. ML  
- Explaining differences in rule vs ML precision  
- Verifying matching scenarios and performing ID splitting analysis

# 4. my\_utils.py File

This Python utility script defines helper functions for transformations, column ordering, and calculation of confusion matrix statistics.  
Key functions include:  
- `apply\_transform` - applies a transformation to a column.  
- `keep\_column\_order` - ensures the order of columns remains consistent.  
- `rename\_result\_columns` - renames columns for clarity in analysis.  
- `write\_to\_single\_directory` - writes DataFrames to single parquet file directories in S3.  
- `reset\_tmp` - cleans up temporary files in DBFS.  
- `read\_er\_results` - loads ER results from CSVs for further processing.  
- `match\_id\_join` and `row\_index\_join` - joins original and distorted data for confusion matrix calculations.  
- `quality\_control` - displays filtered data for quality checks.  
- `confusion\_stats` - calculates accuracy, precision, recall, and F1 statistics.  
- `overall\_confusion\_stats` - overall summary of confusion matrix data.