# 🧠 PySpark Multi-Dataset Discrepancy Categorization

### An Intelligent Framework for Detecting and Resolving Data Quality Issues in Financial & Enterprise Datasets

---

## 📘 Overview

Welcome to the \*\*PySpark Multi-Dataset Discrepancy Categorization\*\* tool — a scalable, Spark-powered framework designed to ensure data accuracy, consistency, and trustworthiness across complex, multi-source datasets. This solution focuses on identifying and resolving common discrepancies that can compromise analytical integrity and downstream modeling efforts.

The engine is optimized for datasets originating from structured sources like \*\*CSV\*\* and \*\*Excel\*\*, and it applies a comprehensive set of rules to normalize data formats, numeric precision, and text consistency.

---

## 🧪 Discrepancy Types Addressed

The engine detects and categorizes the following types of data discrepancies:

- \*\*Leading Zero Issues\*\*: Safeguards numeric identifiers by preserving leading zeros (e.g., `00755275` → `755275`)

- \*\*Decimal Precision Differences\*\*: Aligns numerical fields to a standard decimal format

- \*\*Thousands Separator Differences\*\*: Normalizes regional formatting (e.g., `5,000` → `5000`)

- \*\*Scientific Notation Differences\*\*: Converts exponential notation to full numeric values (e.g., `1.2e3` → `1200`)

- \*\*Currency Symbol Differences\*\*: Standardizes currency formats across international symbols and denominations

- \*\*Rounded Off Numbers\*\*: Identifies discrepancies caused by rounding inconsistencies

- \*\*Abbreviation vs. Full Form\*\*: Harmonizes abbreviations and full-text equivalents (e.g., `DOB` vs `Date of Birth`)

- \*\*Case Sensitivity Issues\*\*: Resolves mismatches due to inconsistent capitalization (e.g., `Credit Card` vs `CREDIT CARD`)

- \*\*Extra Space Issues\*\*: Trims unnecessary leading, trailing, or embedded whitespace

- \*\*Special Character Differences\*\*: Unifies textual fields containing symbols, punctuation, or escape sequences

---

## ✨ Key Features

- \*\*Comprehensive Discrepancy Detection\*\*

Employs advanced algorithms to identify a broad spectrum of data integrity issues

- \*\*Scalable Data Transformation\*\*

Utilizes PySpark to efficiently handle large-scale datasets with high performance

- \*\*Multi-Format Compatibility\*\*

Supports both `.csv` and `.xlsx` files, including multi-sheet Excel parsing

- \*\*Advanced Currency Handling\*\*

Enables recognition and standardization of global currency symbols and formatting

- \*\*Dual-Mode Processing\*\*

Detects discrepancies in both \*\*numeric\*\* and \*\*textual\*\* formats with equal precision

---

## 📊 Sample Input/Output Mapping

| 🔎 \*\*Source\*\* | 🎯 \*\*Target\*\* | 🏷️ \*\*Discrepancy Category\*\* |

|-------------------------------|----------------------------|------------------------------------------|

| `00755275` | `755275` | Leading Zero Issue |

| `5501` | `5501` | Decimal Precision Difference |

| `5,000` | `5000` | Thousands Separator Difference |

| `1.2e3` | `1200` | Scientific Notation Difference |

| `$1,000` | `1,000.00 USD` | Currency Symbol Difference |

| `Credit Card` | `CREDIT CARD` | Case Sensitivity Issue |

| `Savings Account` | `SAVINGS ACCOUNT` | Case Sensitivity Issue |

| `Routing Number:123456789` | `Routing Number: 123456789`| Extra Space Issue |

| `Password@123` | `Password 123` | Special Character Difference |

---

## ⚙️ Tech Stack

- \*\*Apache Spark (PySpark)\*\*

- \*\*Python 3.9+\*\*

- \*\*Pandas\*\* for local data wrangling and Excel I/O

- \*\*Jupyter Notebooks\*\* for exploratory testing

- \*\*GitHub Actions\*\* for planned CI/CD integration

---

## 🧭 Ideal Use Cases

- Financial data quality audits

- Model risk validation pipelines

- Preprocessing step in MLOps workflows

- Data migration and ETL transformation testing

---

## 🚀 Future Enhancements

- ⏱️ Airflow-based scheduled pipeline integration

- ☁️ Deployment to distributed clusters (AWS EMR, Azure Synapse)

- 🧪 Unit and integration testing with PySpark mocks

- 📦 Model/data version control with MLflow or DVC

- 📈 Visualization dashboard to monitor discrepancy trends

---

## 📩 Contact

For questions, contributions, or access to internal notebooks, please contact the Data Engineering & Risk Analytics team.

© 2025 Citi Internal – Confidential and Proprietary.