**Data Engineering Case Study**

* **Data Sets Used**

-  Certificate Holders Statement

-  Enhanced Loan-Level Data

-  Loan-Level Data

* **Requirements**

After downloading the relevant files from the webpage, please perform the following for each of the months of the deal:

* 1. Findthesectioninthe“CertificateHolderStatement”PDFfilethatcontains the “Principal Funds Available” title and extract the value for “Total Principal Funds Available”. This represents the total principal payments that were collected from the pool of assets for that month.
  2. Load the data from the loan-level data files into a database or a file format that you and a data analyst can run queries against.
  3. Queryyourdatabasetoverifythatthetotalprincipalreportedintheloan- level data files matches the value you extracted from the PDF file.
  4. Summarize your findings and note any issues you ran into while downloading and working with the data.

**Challenges:**

1. Using Loan-level data: It was very challenging as you have to identify the correct formula on how to get the “Total Principal Funds Available” as its corresponding header breakdown won’t be queried from the csv file in reference to the pdf file that was downloaded from Certificate Holder Statement.
   * + Loan-Level Data 2006 for the month of Oct, Nov, and Dec
2. Using Enhanced Loan-Level Data: With this one, you will still use a and identify the formula in order to get the “Total Principal Funds Available” but it was easier to figure out as some of the headers are the same or related to the breakdown presented in the pdf file.
   * + Enhanced Loan-Level Data 2007 for the month of Oct

**Recommendations:**

1. Use uniform or consistent terminologies so it will be easier to query data.
2. Be mindful on the data types.

To expand our work at Citi, we can enhance our data pipeline by:

1. Automated Data Quality Checks: Add automated checks for data integrity and consistency.
2. Data Profiling: Develop tools to analyze data characteristics and detect issues like missing values or format inconsistencies.
3. Data Cleansing: Create routines for automatic data cleansing, fixing common issues.
4. Scalability: Design the pipeline to handle multiple data fields and pipelines concurrently using distributed computing.
5. Centralized Repository: Establish a central hub to manage and prioritize data issues.
6. Documentation: Document processes and best practices for consistent data quality improvements.

These changes will support scaling data validations and corrections across numerous pipelines at Citi, ensuring data reliability.