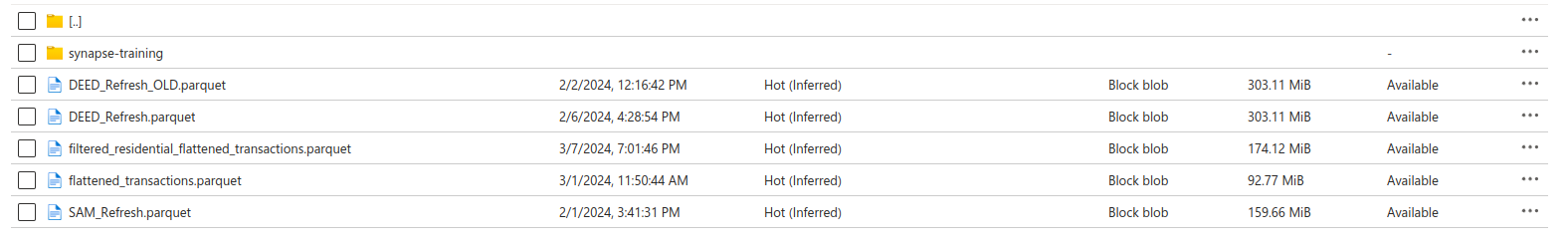
**Data Processing Workflow**

**1. Introduction:**

The following documentation outlines the data processing workflow implemented on Azure ADL Gen2 for the "deed" and "sam" datasets. The workflow involves several steps, from initial data retrieval to the final result in the "borrow\_group" table.

**2. Data Sources:**

* **Deed:** Original dataset containing deed-related information, specifically focused on buyer-related data for a property.
* **Sam:** Original dataset containing SAM information, specifically focused on Borrower-related data for a property.



**3. Workflow Steps:**

**a. Filtering:**

1. **Individual/Commercial Classification:**

**First, we created Full name by using concat before applying the condition**

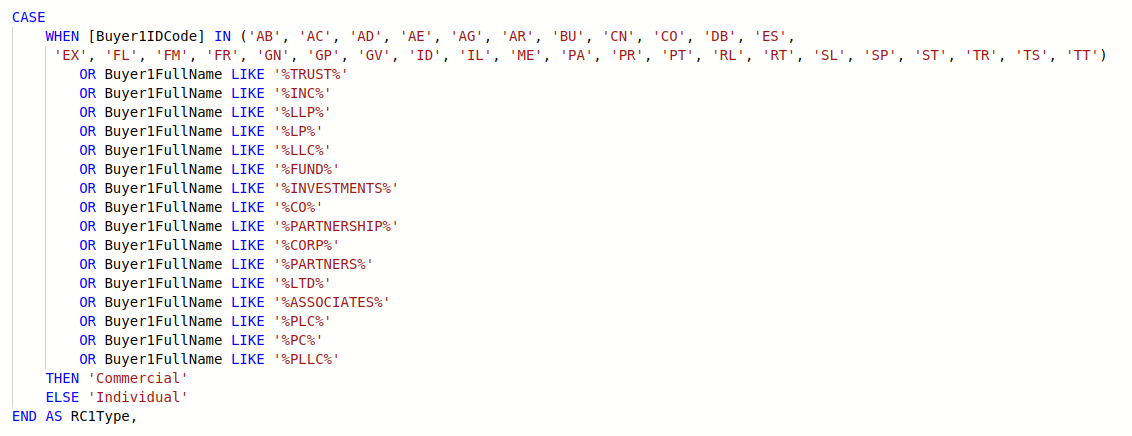


i. **Code-Based Matching:**

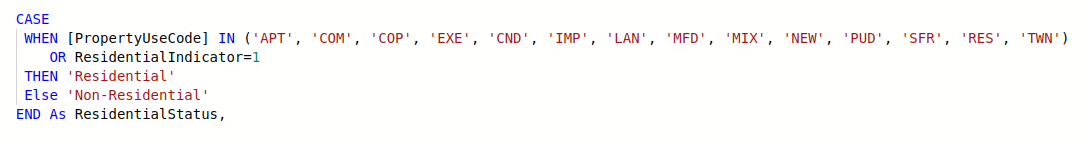
* + Applied filters based on BK codes to identify individual and commercial borrowers in both SAM and DEED datasets.
  + Used the BK code list: ["AB", "AC", "AD", "AE", "AG", "AR", "BU", "CN", "CO", "DB", "ES", "EX", "FL", "FM", "FR", "GN", "GP", "GV", "ID", "IL", "ME", "PA", "PR", "PT", "RL", "RT", "SL", "SP", "ST", "TR", "TS", "TT"].

ii. **Name-Based Matching:**

* + Checked borrower names for terms indicative of commercial status.
  + If the borrower's name incorporates any terms from the list ["TRUST", "INC", "LLP", "LP", "LLC", "FUND", "INVESTMENTS", "CO", "PARTNERSHIP", "CORP", "PARTNERS", "LTD", "ASSOCIATES", "PLC", "PC", "PLLC"] , it is classified as a commercial borrower else individual.



1. **Residential/Non-Residential Classification:**
   * Checked the "ResidentialIndicator" flag and "PropertyUseCode" to tag properties as Residential or Non-Residential.

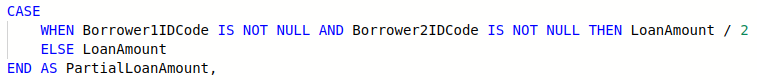


**3. Extra conditions:**

* + For individuals, transactions are picked if they occurred in the year 2019 or onwards (within the last five years). Field for that: “OriginalDateOfContract.”



* + For companies, transactions are picked only if they occurred in the year 2013 or onwards (last 10 years). Field for that: “OriginalDateOfContract.”
  + When adding transactions into "individual\_borrowers" and "company\_borrowers" tables from the original "SAM" and "DEED" tables, an extra field named "PartialLoanAmount" is added. For each transaction in both "SAM" and "DEED" tables where two borrowers are involved, the value of "PartialLoanAmount" for each borrower will be half of the original loan amount value.

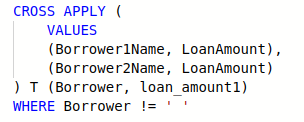
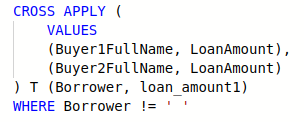


**Table Creation:**

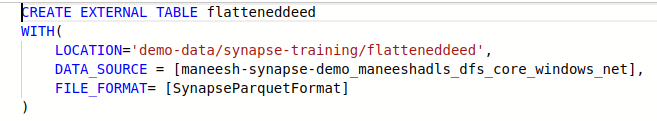
* Created 4 tables based on classification and conditions:
  + Res\_individual\_2019\_deed
  + Res\_commercial\_2013\_deed
  + Res\_individual\_2019\_sam
  + Res\_commercial\_2013\_sam

**b. Flattening and Combining:**

* Applied flattening and combining operations on both datasets individually by using Cross Apply.
* CROSS APPLY is a SQL operator used in conjunction with a table-valued function. It evaluates the function for each row of the table expression and returns the combined result set.

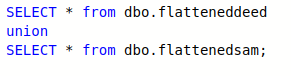


Created "flatteneddeed\_individual" and "flatteneddeed\_commercial", “flattenedsam\_individual” and “flattenedsam\_commercial” tables for deed and sam datasets respectively



**c. Union Operation:**

* Combined "flatteneddeed\_individual" and "flattenedsam\_individual" using the UNION operation, resulting in a table named individual\_borrowers and combined “flatteneddeed\_commercial” and “flattenedsam\_commercial” to create a table named commercial\_borrowers.



* The purpose of combining the tables "flatteneddeed\_individual" and "flattenedsam\_individual" using the UNION operation to create a table named "individual\_borrowers" is likely to consolidate data on individual borrowers from different sources or databases. Similarly, combining "flatteneddeed\_commercial" and "flattenedsam\_commercial" into a table named "commercial\_borrowers" serves to aggregate information about commercial borrowers.

**d. Individual and Combined Borrower Processing:**

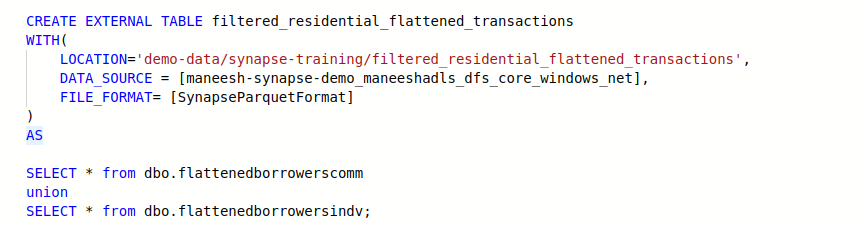
* Applied conditions to filter individual borrower information ("individual\_borrowers ") and combined borrower information ("commercial\_borrowers ").
* Conditions for individual\_borrowers:
* Exclude borrowers with missing BorrowerFullStreetAddress.
* Group the dataset by BorrowerName and Mailing Address.
* Filter transactions within the last 1 year for each group and count the number of unique DPID's.
* Filter transactions within the last 5 years for each group and count the total number of transactions.
* Select borrowers/groups where the count of distinct property transactions in the last 1 year is greater than or equal to 2, and the count of property transactions in the last 5 years is greater than or equal to 3.
* Output the selected borrowers.
* Conditions for commercial\_borrowers:
* Group the dataset by CompanyName/BorrowerName.
* Calculate the total net loan amount for each group.
* Calculate the total number of property transactions for each group.
* Apply the selection criteria:
* Exclude groups where the net loan amount is less than 1 million.
* Exclude groups where the total property transactions are less than 3.
* Output the selected borrowers.

**e. Flattening Borrower Information:**

* Flattened individual and combined borrower datasets separately, resulting in "flattenedborrowersindv" and "flattenedborrowerscomm" tables.

**f. Union Operation (Flattened Borrower Information):**

* Combined " flattenedborrowersindv " and " flattenedborrowerscomm" using the UNION operation to create table named “filtered\_residential\_flattened\_transactions”

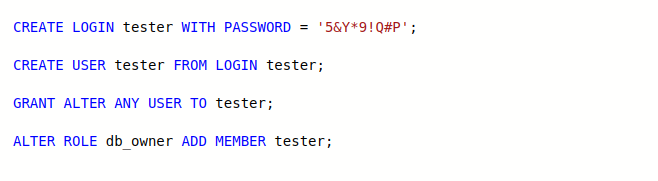


The purpose of combining the tables "flattenedborrowersindv" and "flattenedborrowerscomm" using the UNION operation to create a table named "filtered\_residential\_flattened\_transactions" is likely to consolidate borrower information from both individual and commercial sources specifically for residential transactions.

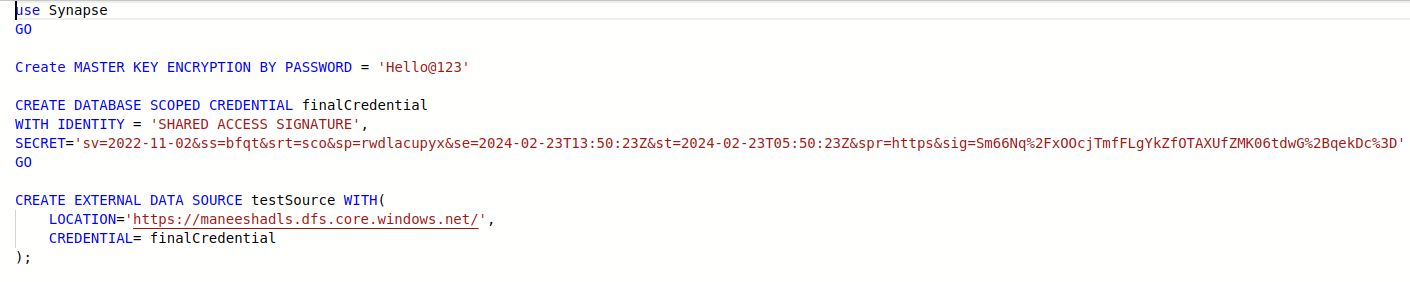
This step allows for the aggregation of borrower data from diverse sources, ensuring that all relevant information related to residential transactions is captured in a single table. By filtering out non-residential transactions,

**g. Creating Credentials/ Data Source / File Format :**

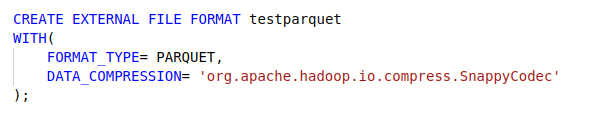
* **How to create DB Credentails:**
* Create a login with a password.
* Create a user from the created login.
* Grant the necessary permissions to the user.
* Add the user to a role.



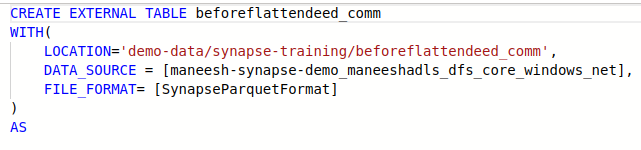
* **How to create a Data Source:**
* Identify the external data location.
* Access Synapse Studio.
* Write a CREATE EXTERNAL DATA SOURCE statement specifying the data source type and location.
* Execute the SQL script.
* Verify the data source creation.



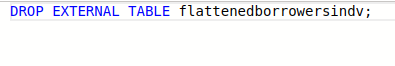
* **How to create File Format:**
* Identify the file format details such as field terminators, line terminators, encoding, etc.
* Access Synapse Studio or any SQL client connected to your Synapse Analytics instance.
* Write a CREATE EXTERNAL FILE FORMAT statement specifying the format options.
* Execute the SQL script.
* Verify the file format creation.



* **How to create External Table:**
* Identify the external data source and file format details.
* Access Synapse Studio or any SQL client connected to your Synapse Analytics instance.
* Write a CREATE EXTERNAL TABLE statement specifying the table schema, external data location, data source, and file format.
* Execute the SQL script.
* Verify the external table creation.



* **How to Drop Table:**
* Access Synapse Studio or a SQL client.
* Write a DROP TABLE statement for the table you want to delete.
* Execute the SQL script.
* Verify the table has been dropped.



**4. Conclusion:**

This documentation outlines the step-by-step workflow for processing the "deed" and "sam" datasets on Azure ADL Gen2. Each step is explained, detailing the operations performed and the resulting tables. The final output is the "filtered\_residential\_flattened\_transactions" table, reflecting the processed and analyzed borrower data.