**SSIS ETL-Processing**

1. **I used 3 Sequence Containers to process 3 different data source**

**files;**

1. **Brief Explanation about logic for each data source file:**

* **Round 1: Structural Data Cleansing:**

**It refers to standardize, correct, and unify data at the structural**

**level to ensure conformity in format, type, and basic integrity :**

**- Missing Value Handling**

**- Duplicate Data Handling -> It is considered as Logical checks in Round 2 below**

**- Data Type Standardization**

**- Field Format Standardization**

**- Data Length Regulation**

**- Illegal Character Cleaning**

**- Field Value Unification**

* **Round 2: Logical Check**

**It refers to an attempt to verify whether the logical relationships among fields align with theory or common sense in real life.**

**- Identify and group duplicated data rows based on values of multiple fields such as state\_city\_suburb etc.**

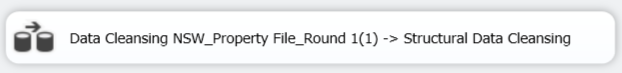
**- In each group of these duplicated values, only 1 row was retained, while the rest were removed.**

**- Check whether these retained values have correct corresponding field values that conform to real-life scenario information.**

**- Because of the sheer volume of data in each data source file, the logical checks were only conducted on these duplicated data, while it is assumed that all field values within the non-duplicate records adhere to real-world conventions and logical consistency.**

**3.2 NSW\_Property\_File**

**In SSIS,**



- When attempting to import data from the **NSW\_Property** file into the

temporary database (**Property Data Engineering**), I encountered an issue:

the file contains a field named **'Property\_Median\_Value'**, which includes

missing values. As a result, the data import task to the temporary table

- **NSW\_Property\_original\_data(Do Not Truncate)**’consistently failed;

- Therefore, **Round 1(1)** is initiated to address this issue by filling

the missing values with **0**.

- Then, all the data was successfully imported into the table -

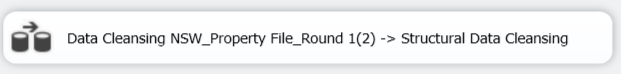
‘**NSW\_Property\_original\_data(Do Not Truncate)**’in the temporary

database (**Property Data Engineering**);

The purpose of this table is to retain the original data extracted from

the step above for downstream ETL operations.It is critical that

under no circumstances should this table be truncated.

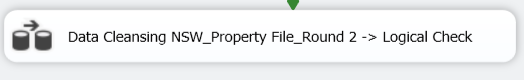


- Round 1(2) is initiated to conduct a structural data cleansing on the **NSW\_Property** dataset.

The cleansing strategy was developed based on an exploratory analysis of various data blocks extracted through SQL queries(Refer to ***NSW Property\_Round 1(2).sql***) tailored to specific patterns identified in the source data.

Detailed descriptions of the cleansing procedures are documented in the corresponding SSIS packages.

Upon completion of this round, all cleansed data were stored in the temporary table named '**NSW\_Property\_data\_cleansing\_Round\_1(2)\_end**'.



- Round 2 is initiated to conduct a logical check on the **NSW\_Property** dataset.

- Round 2 involves a logical consistency check to assess whether the

relationships among multiple fields within each record align with

common sense or theoretical expectations - This task has been escaped for lack of reference data and it is assumed that source data are correct;

- A reference table (***duplicated\_data\_records.csv***) was created by identifying all duplicate records and their frequencies in the source data using SQL queries (***NSW Property\_Round 2.sql***).

- Using the combination of state + city + suburb from the reference table, duplicate entries in the source data were located and grouped accordingly. Within each group, only the first occurrence was retained while all subsequent duplicates were removed.

-The cleansed data were then loaded into the table-(***load\_NSW\_PropertyMedainValue***) in the database-(***DataWarehouse***) for storage.