## MINT ETL task documentation

By Jim Zhu

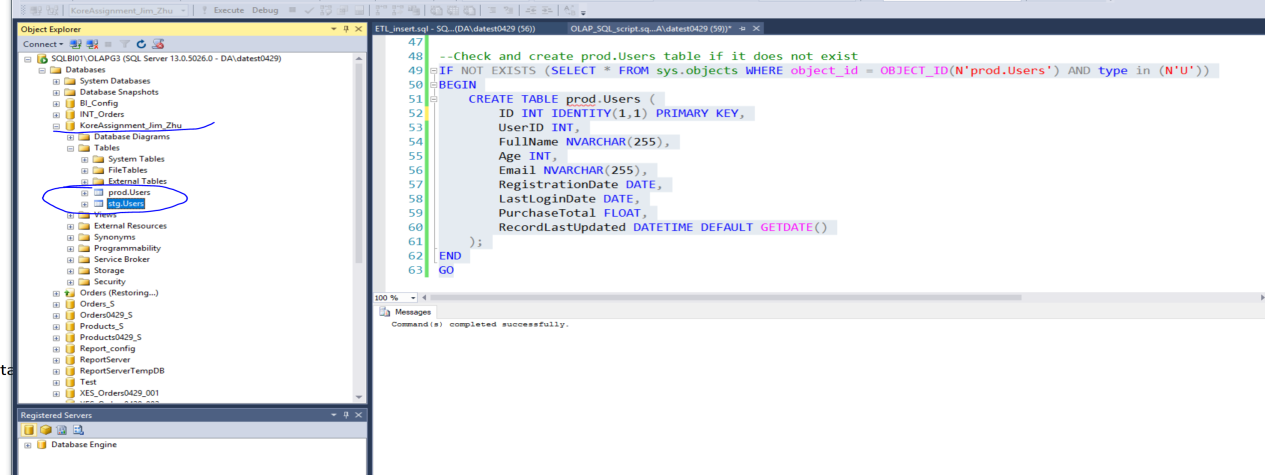
Overview of the project:

Extract csv flat file to staging SQL server as table [stg.Users], load stg.Users to SQL server production as [prod.Users].

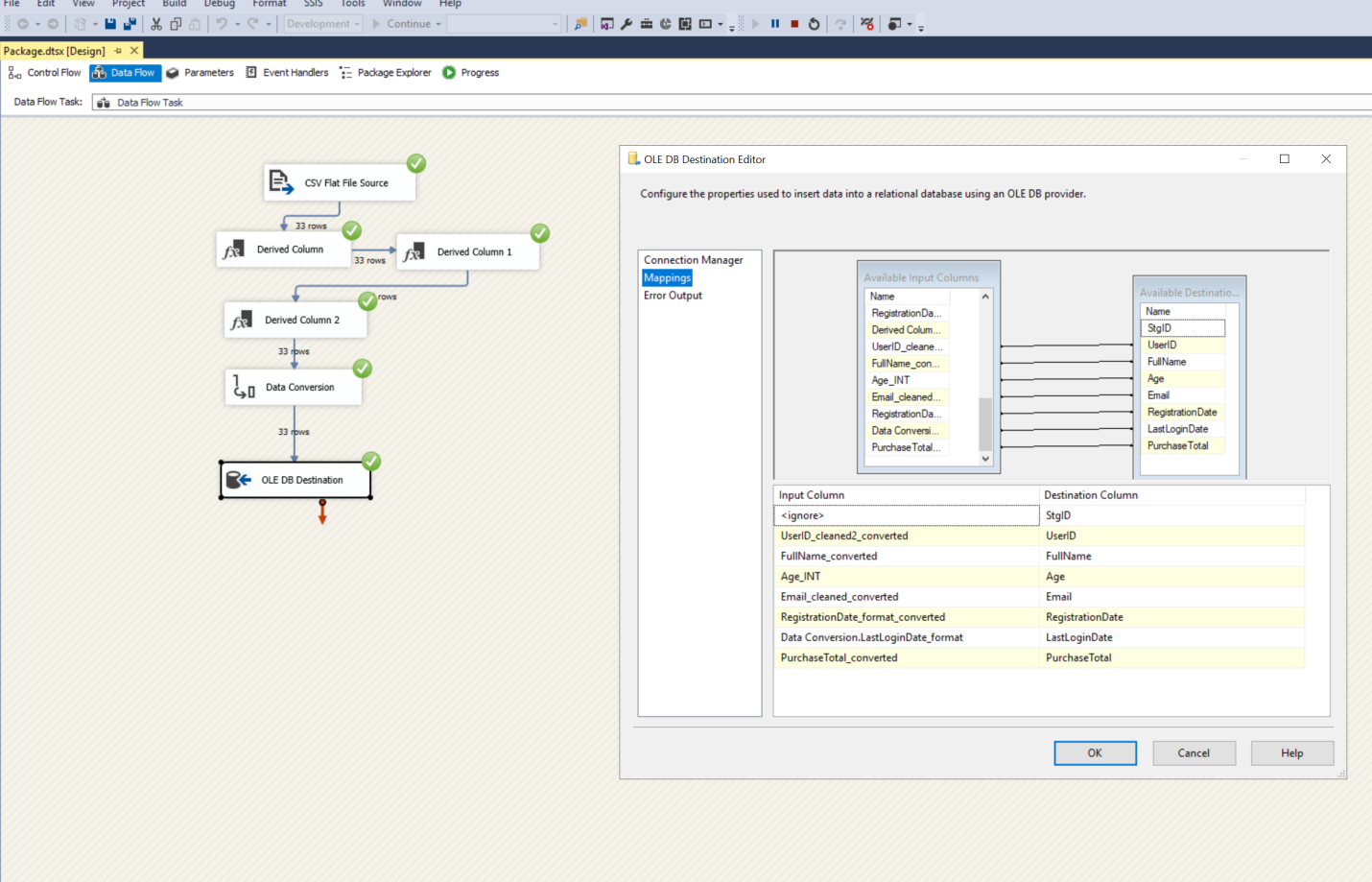
**ETL-SSIS**

**ETL-SSIS**

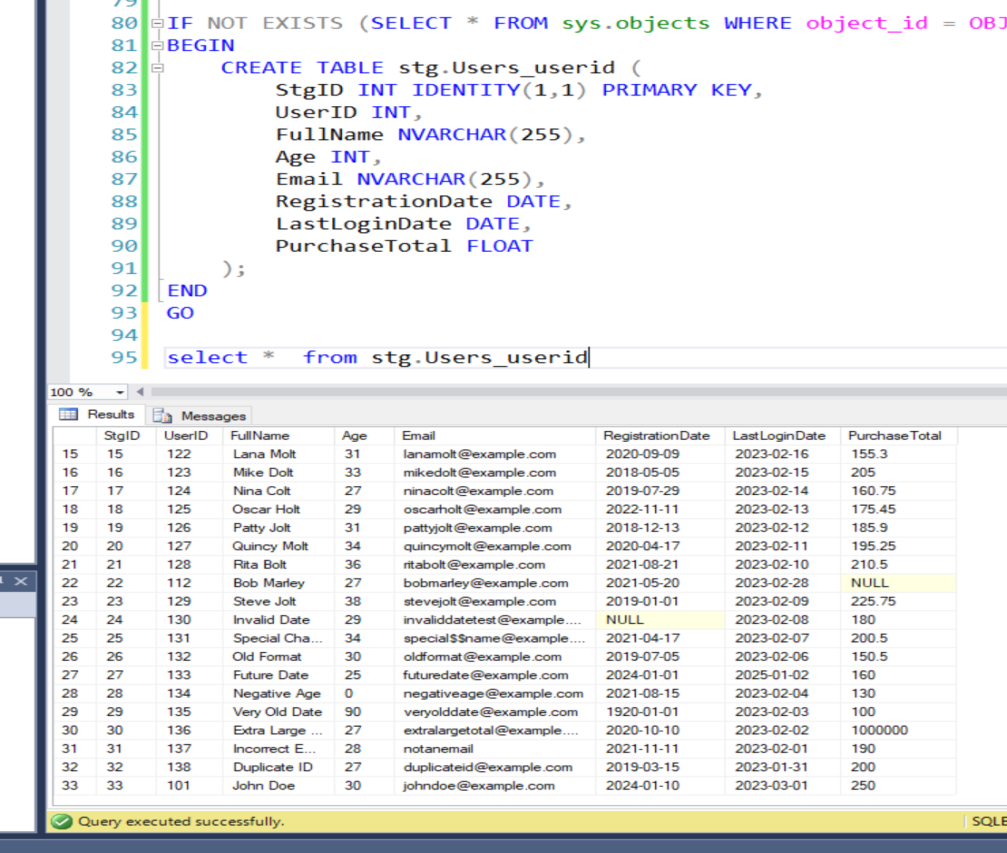
Detailed instructions on how to set up database and SSIS package  
 -- Run OLAP\_SQL\_script.sql to create database KoreAssignment\_Jim\_Zhu in both SQL server staging and production environment.  
 -- Create stg.Users and prod.Users table structure as shown in below screen shot.



--Create SSIS project and setup control flow and data flow in SSIS package:



---Run SSIS package, extract and transform source file to staging table in SQL server: [stg.Users]

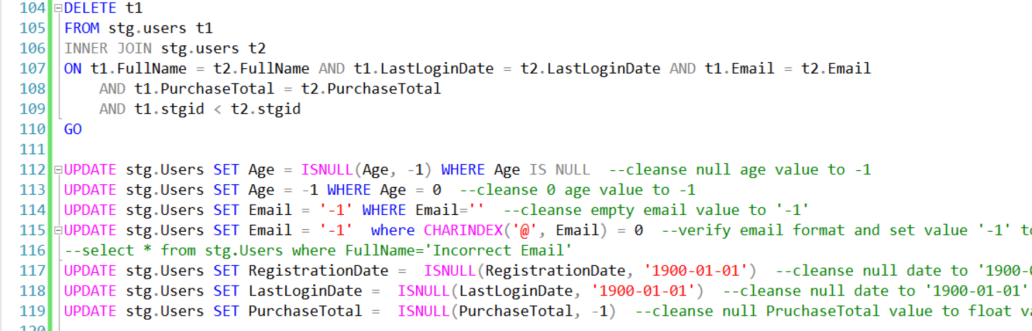


2. Data Cleaning in Staging Table: stg.Users

2.1 Removing duplicates records in stg.Users. Using “Sort Transformation” in SSIS to remove duplicates for small file. For big file using T-SQL script in SQL staging table.

2.2 Conduct more data cleaning for staging table, using T-SQL scripts as below screen shot:   
 Such as cleanse NULL value, Empty string, email format verification etc.

2.3 Staging table are good enough for production



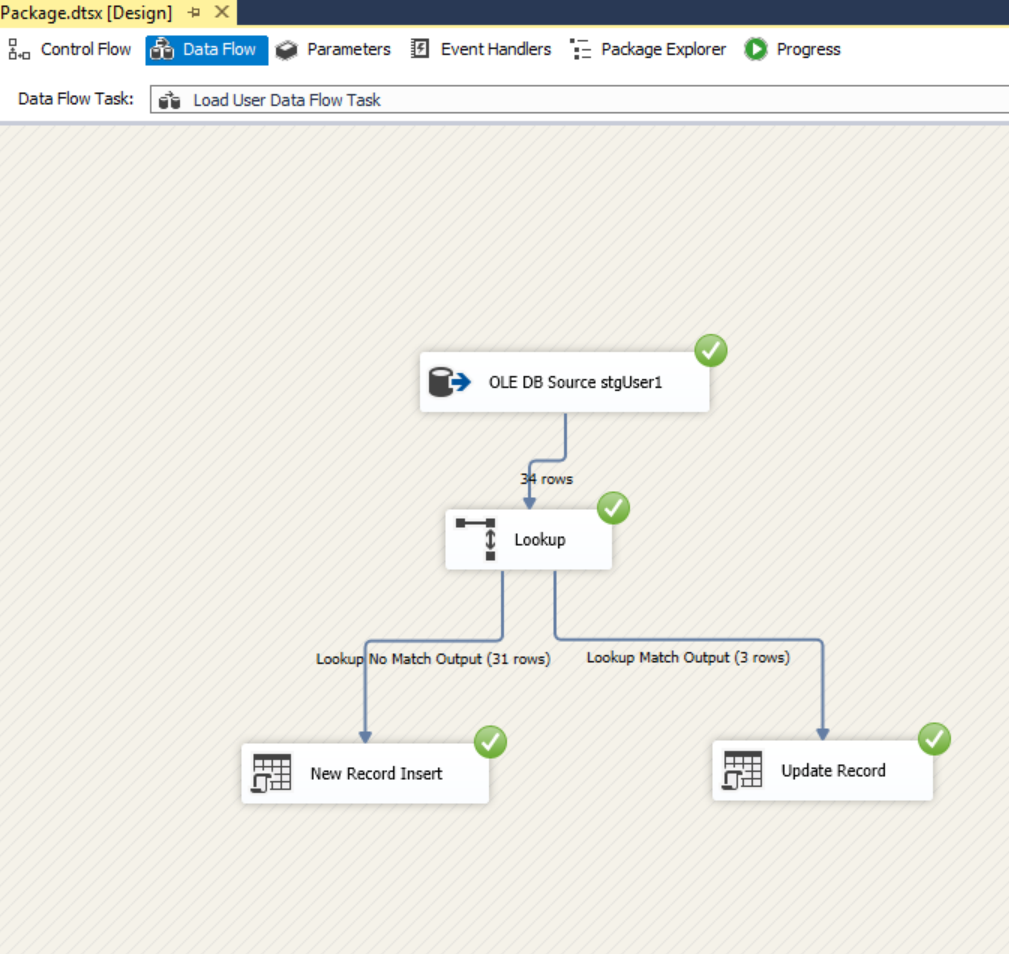
3. Incremental Load to Production Table

3.1 Create data flow from staging to production in SSIS.

If there is new record in staging table then insert new records.

Or update existing records if there is any change.

Below snap shot showed successful insert and update respectively for prod.



Approach summary:

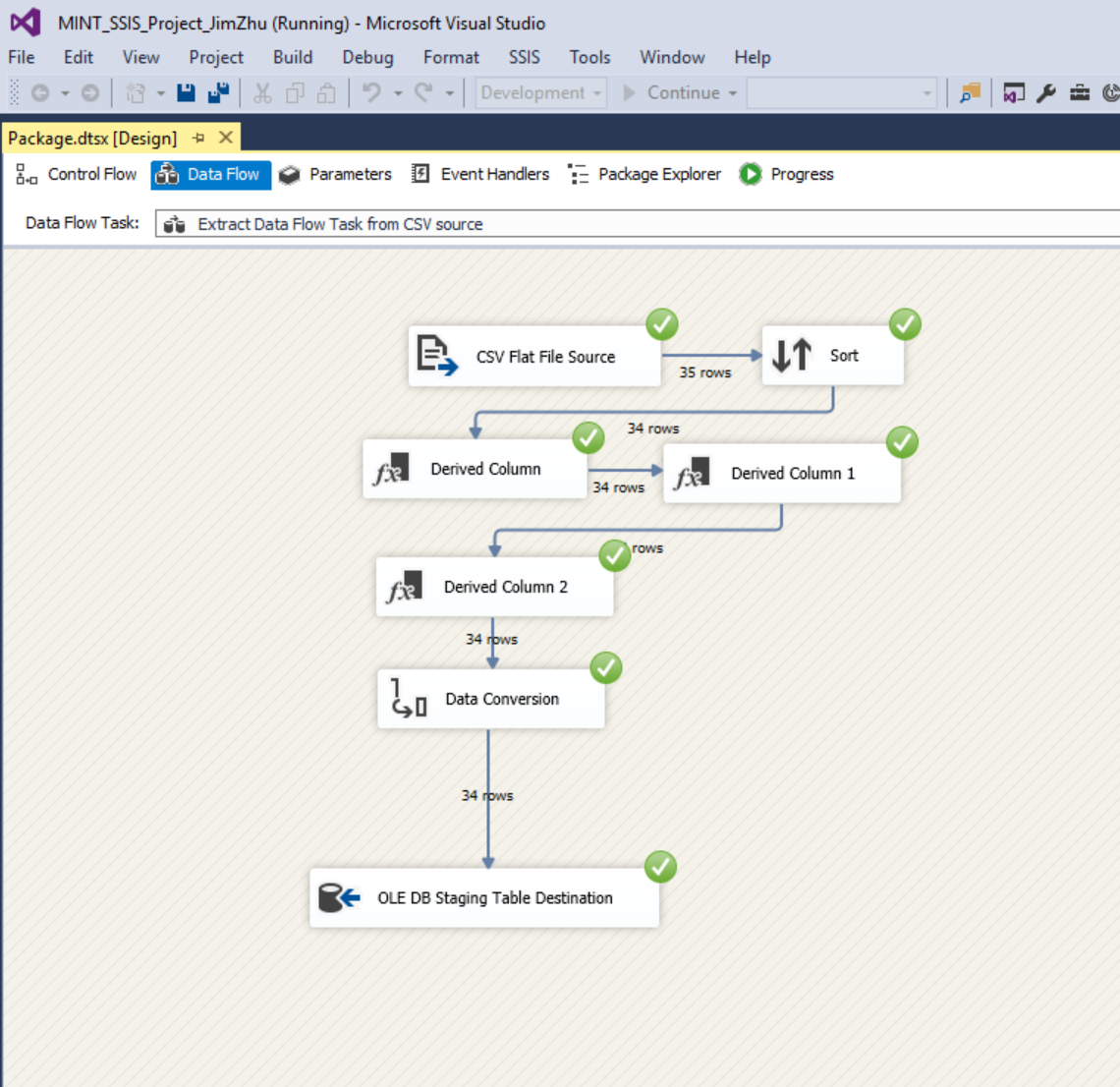
USE SSIS data flow to extract csv source file, load csv source file into staging table.

Then conduct data cleaning, removing duplicates and standardization for staging table. Error records can be isolated and be moved to a separate Error table for further modification.

After that, staging table is good enough to be loaded to production table by SSIS as part of ETL process.

Execution Report: MINT ETL running instruction (uploaded on web)

The results of the execution 34 records are loaded with success



There are more than 9 records have null/invalid values. These records have been identified and cleansed.

They can be moved to stg.ErrorTable for further investigation.

This can be done using SQL script or SSIS dataflow. SQL scripts are included in ‘ETL running instruction’ document.