## MINT ETL running steps with code

select @@servername

--SQLBI-1\OLAPG3

--TODO:CREATE DATABASE Jim\_Zhu

IF NOT EXISTS (SELECT \* FROM sys.databases WHERE name = N'KoreAssignment\_Jim\_Zhu')

BEGIN

CREATE DATABASE[KoreAssignment\_Jim\_Zhu];

END

GO

USE [KoreAssignment\_Jim\_Zhu]

GO

--Check and create stg schema if it dose not exist

IF NOT EXISTS (SELECT \* FROM sys.schemas WHERE name = N'stg')

BEGIN

EXEC('CREATE SCHEMA stg')

END

--Check and create prod schema if it does not exist

IF NOT EXISTS (SELECT \* FROM sys.schemas WHERE name = N'prod')

BEGIN

EXEC('CREATE SCHEMA prod')

END

SELECT \* FROM INFORMATION\_SCHEMA.SCHEMATA WHERE SCHEMA\_NAME in ('stg','prod') --list all schemas

/\*

CATALOG\_NAME SCHEMA\_NAME SCHEMA\_OWNER DEFAULT\_CHARACTER\_SET\_CATALOG DEFAULT\_CHARACTER\_SET\_SCHEMA DEFAULT\_CHARACTER\_SET\_NAME

KoreAssignment\_Jim\_Zhu prod dbo NULL NULL iso\_1

KoreAssignment\_Jim\_Zhu stg dbo NULL NULL iso\_1

\*/

--Check and create stg.Users table if it does not exist

IF NOT EXISTS (SELECT \* FROM sys.objects WHERE object\_id = OBJECT\_ID(N'stg.Users1') AND type in (N'U'))

BEGIN

CREATE TABLE stg.Users1 (

StgID INT IDENTITY(11,1) PRIMARY KEY,

UserID INT,

FullName NVARCHAR(255),

Age INT,

Email NVARCHAR(255),

RegistrationDate DATE,

LastLoginDate DATE,

PurchaseTotal FLOAT

);

END

GO

--select \* from stg.Users1

--DROP TABLE stg.Users1

--Check and create prod.Users table if it does not exist

IF NOT EXISTS (SELECT \* FROM sys.objects WHERE object\_id = OBJECT\_ID(N'prod.Users') AND type in (N'U'))

BEGIN

CREATE TABLE prod.Users (

ID INT IDENTITY(1,1) PRIMARY KEY,

UserID INT,

FullName NVARCHAR(255),

Age INT,

Email NVARCHAR(255),

RegistrationDate DATE,

LastLoginDate DATE,

PurchaseTotal FLOAT,

ProdStgID INT,

RecordLastUpdated DATETIME

);

END

GO

--DROP TABLE prod.Users

--select \* from stg.Users1

/\*

SELECT fullname, COUNT(\*) as cnt

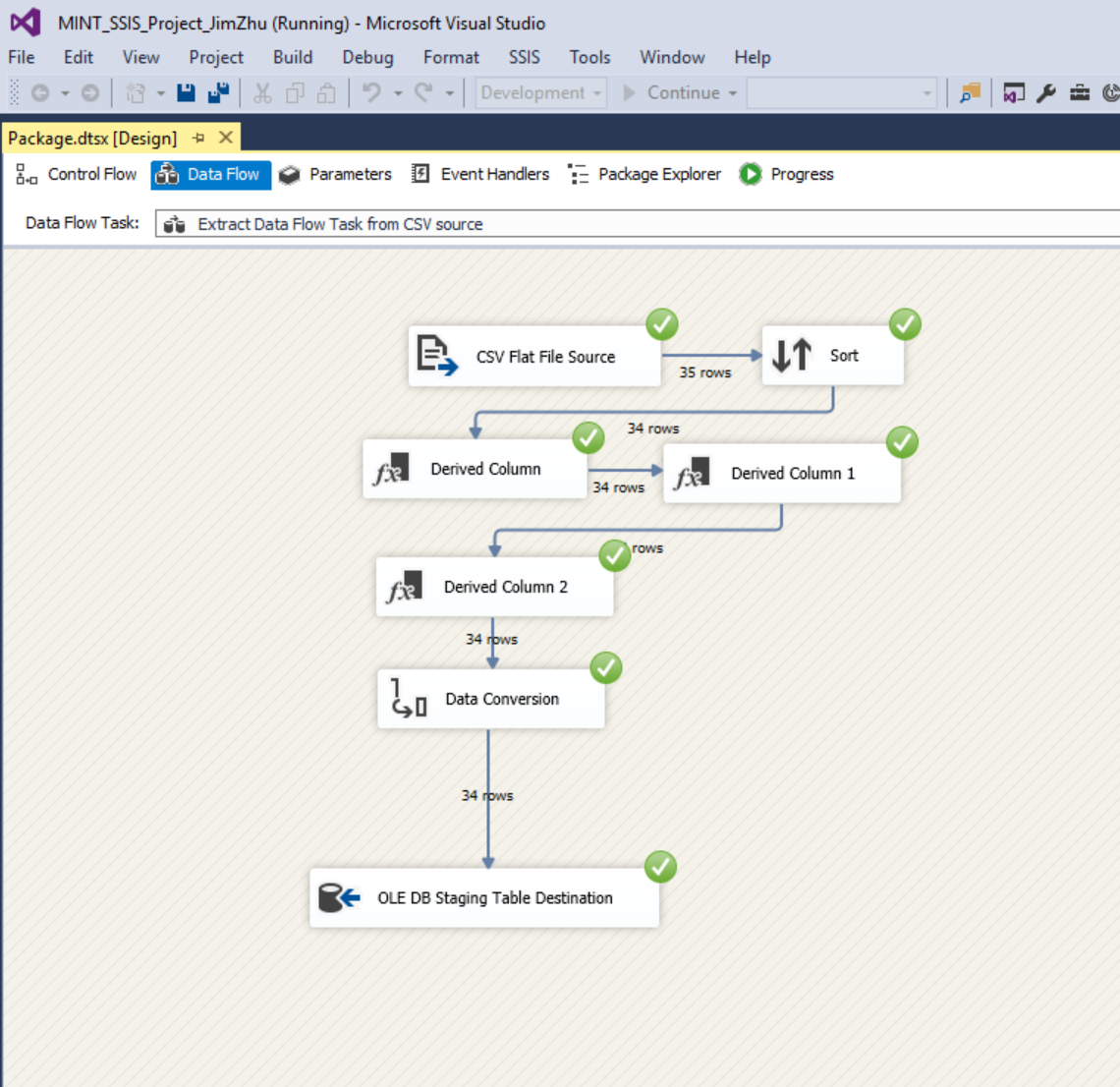
FROM stg.users

GROUP BY fullname

HAVING COUNT(\*) > 1;

\*/

--Run SSIS package 'Extract Data flow task' to load source csv flat file to stg.User1 table



--Then perform stg.User1 table cleaning after data is loaded in stg.User1 table:

UPDATE stg.Users1 SET Age = ISNULL(Age, -1) WHERE Age IS NULL --cleanse null age value to -1

UPDATE stg.Users1 SET Age = -1 WHERE Age = 0 --cleanse 0 age value to -1

UPDATE stg.Users1 SET Email = '-1' WHERE Email='' --cleanse empty email value to '-1'

UPDATE stg.Users1 SET Email = '-1' where CHARINDEX('@', Email) = 0 --verify email format and set value '-1' to email address without '@'

UPDATE stg.Users1 SET RegistrationDate = ISNULL(RegistrationDate, '1900-01-01') --cleanse null date to '1900-01-01'

UPDATE stg.Users1 SET LastLoginDate = ISNULL(LastLoginDate, '1900-01-01') --cleanse null date to '1900-01-01'

UPDATE stg.Users1 SET PurchaseTotal = ISNULL(PurchaseTotal, -1) --cleanse null PruchaseTotal value to float value -1

GO

--Remove duplicate records if any by table self join

DELETE t2

FROM stg.users1 t1

INNER JOIN stg.users1 t2

ON t1.FullName = t2.FullName AND t1.LastLoginDate = t2.LastLoginDate AND t1.Email = t2.Email

AND t1.PurchaseTotal = t2.PurchaseTotal AND t1.Age = t2.Age AND t1.RegistrationDate = t2.RegistrationDate

AND t1.LastLoginDate = t2.LastLoginDate

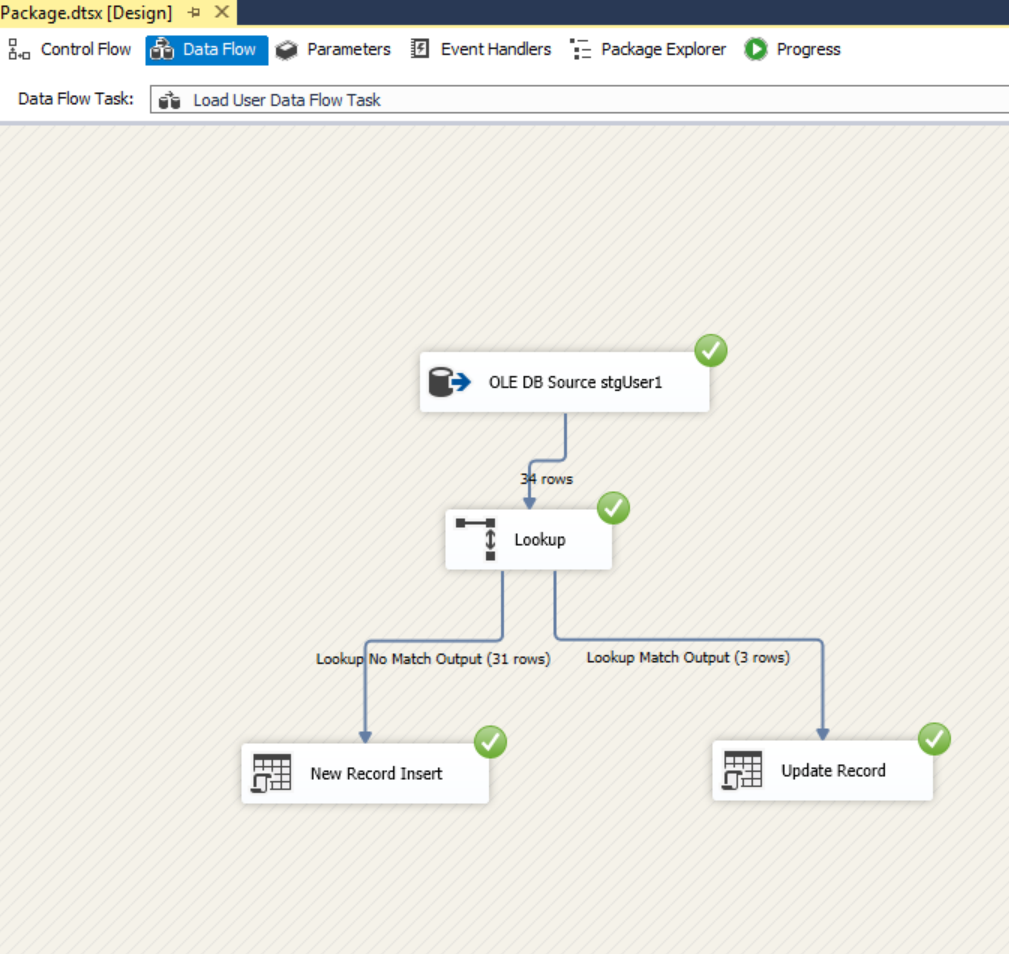
AND t1.stgid < t2.stgid

GO

-- Load to Production Table prod.Users

--Now run SSIS data flow--'Load User Data flow task' of project package, it is to load stg.Users1 table to prod.Users table

--Very important: please disable “Extract Data flow task” in Control Flow tab, when running “Load User Data Flow”, otherwise, cvs flat source file will be running each time “Load User Data Flow” running. This will introduce redundant extract.

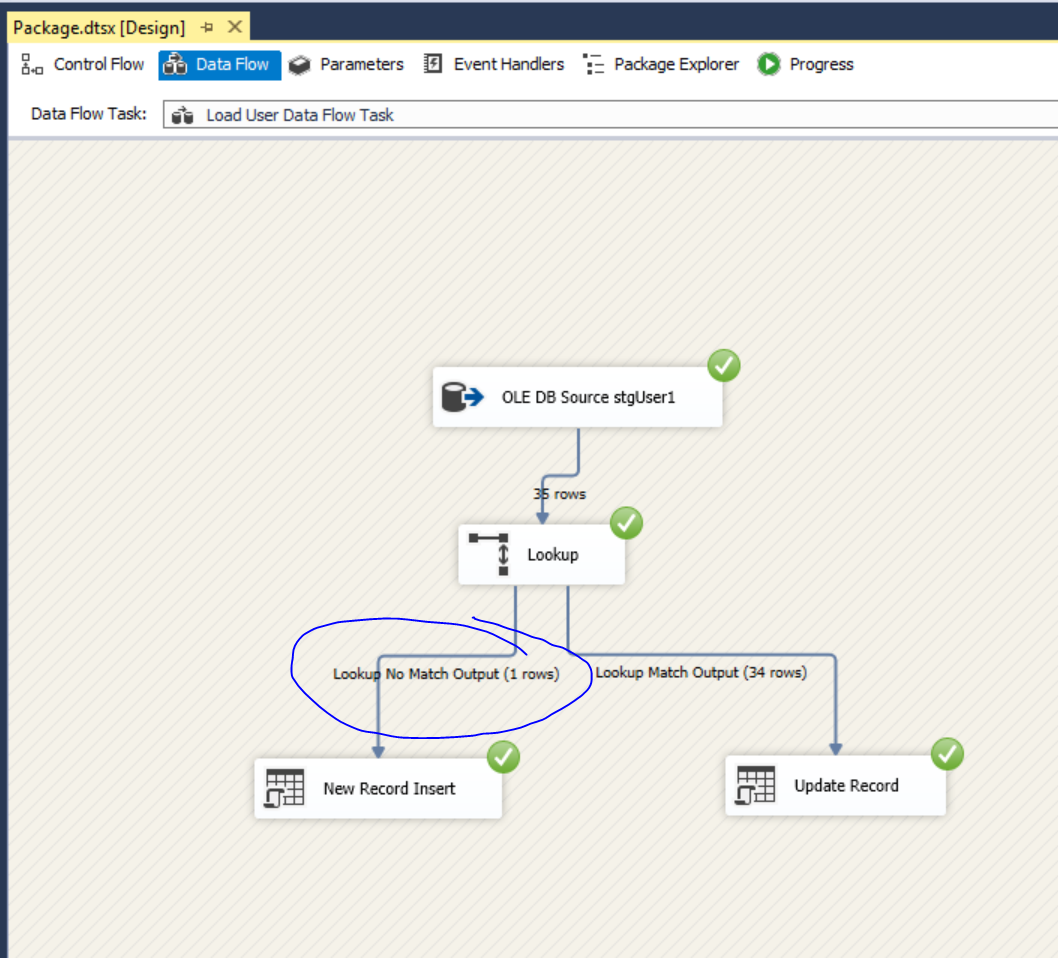


--Incremental Load to Production Table

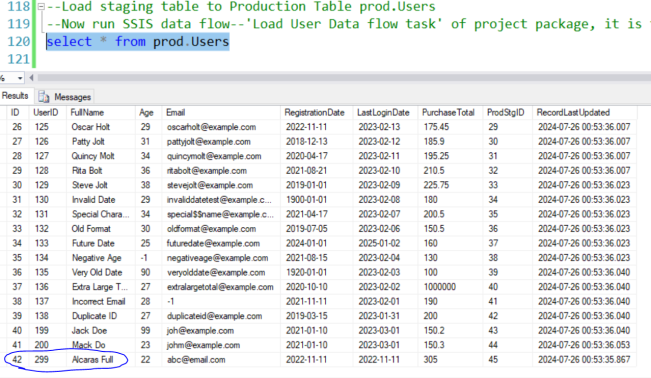
--a: insert a new record to staging table

Insert into stg.Users1(UserID,FullName,Age,Email,RegistrationDate,LastLoginDate,PurchaseTotal)

values(299,'Alcaras Full',22,'abc@email.com','2022-11-11','2022-11-11',305)



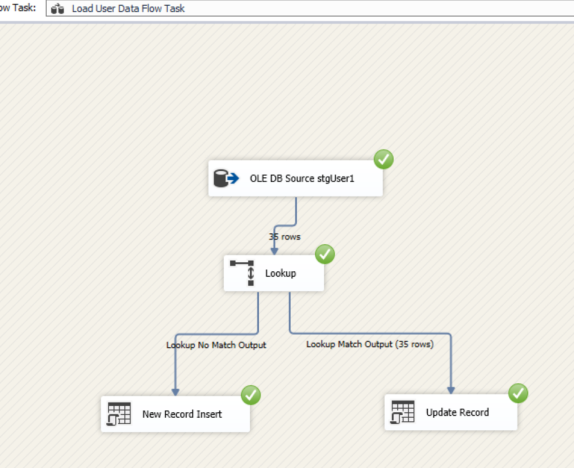
--New record was inserted to production table via SSIS data flow

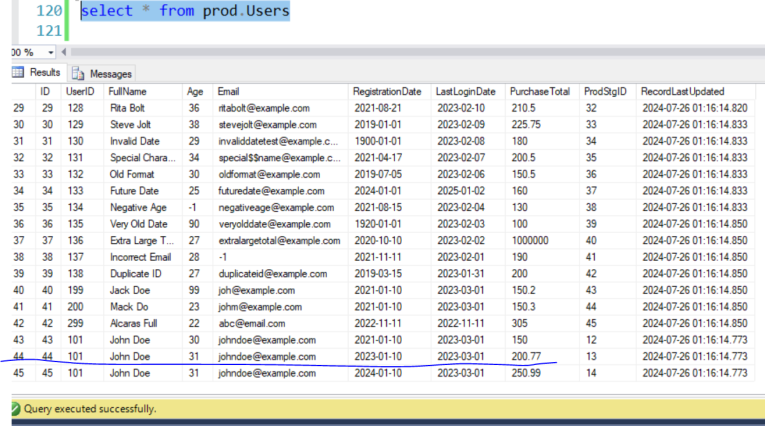


--b: update existing record testing

update stg.users1 set Age= 31, PurchaseTotal= 200.77 where StgID = 13

--Re-run Load User Data Flow task in SSIS, existing record is updated in production





Test done.

---handling error records, isolate error records to a seperate table stg.ErrorTable as below

IF EXISTS (SELECT \* FROM sys.objects WHERE object\_id=OBJECT\_ID('stg.ErrorTable') AND type in ('U'))

DROP TABLE stg.ErrorTable

GO

IF NOT EXISTS (SELECT \* FROM sys.objects WHERE object\_id = OBJECT\_ID(N'stg.ErrorTable') AND type in (N'U'))

BEGIN

CREATE TABLE stg.ErrorTable (

StgID INT PRIMARY KEY,

UserID INT,

FullName NVARCHAR(255),

Age INT,

Email NVARCHAR(255),

RegistrationDate DATE,

LastLoginDate DATE,

PurchaseTotal FLOAT

);

END

GO

/\*SELECT \*

INTO stg.ErrorTable FROM stg.Users

WHERE UserId=0 OR Age = -1 OR Email = '-1' OR RegistrationDate = '1900-01-01'

OR LastLoginDate = '1900-01-01' OR PurchaseTotal = -1

--(9 row(s) affected)

GO

\*/

-- Move error records to stg.ErrorTable

INSERT INTO stg.ErrorTable

SELECT \*

FROM stg.Users1

WHERE UserId=0 OR Age = -1 OR Email = '-1' OR RegistrationDate = '1900-01-01'

OR LastLoginDate = '1900-01-01' OR PurchaseTotal = -1

SELECT \* FROM stg.ErrorTable

GO

/\*

StgID UserID FullName Age Email RegistrationDate LastLoginDate PurchaseTotal

2 111 Alice Johnson -1 alicejohnson@example.com 2022-07-15 1900-01-01 -1

4 113 Cathy Smith -1 cathysmith@example.com 2019-10-12 2023-02-27 210.5

5 0 Null User 25 -1 2020-12-01 2023-02-25 100

7 115 Eva Green -1 evagreen@example.com 2018-08-22 2023-02-23 -1

14 121 Kevin Yolt -1 kevinyolt@example.com 2021-03-14 2023-02-17 -1

22 112 Bob Marley 27 bobmarley@example.com 2021-05-20 2023-02-28 -1

24 130 Invalid Date 29 invaliddatetest@example.com 1900-01-01 2023-02-08 180

28 134 Negative Age -1 negativeage@example.com 2021-08-15 2023-02-04 130

31 137 Incorrect Email 28 -1 2021-11-11 2023-02-01 190

\*/

DELETE FROM stg.Users1 --keep stg.Users table healthy from error records

WHERE UserId=0 OR Age = -1 OR Email = '-1' OR RegistrationDate = '1900-01-01'

OR LastLoginDate = '1900-01-01' OR PurchaseTotal = -1

-- Now stg.Users1 table are all good with records. These records are awaiting happily to load to prod.

----END

By Jim Zhu