**Name: Hieu Nghia Huynh**

**Student Number: n10315071**

**Assignment 1 Report**

**Task 1:**

* The source data consists of 2 files: North\_America.csv and Other\_regions.xls. Each file contains information about a person and his/her respond to the survey, in the respective locations.
* When I did data profiling, I found out that in these columns: *blood\_type*, *income, smoking, working* contains “?”. During ETL, I will remove those data rows from the database.

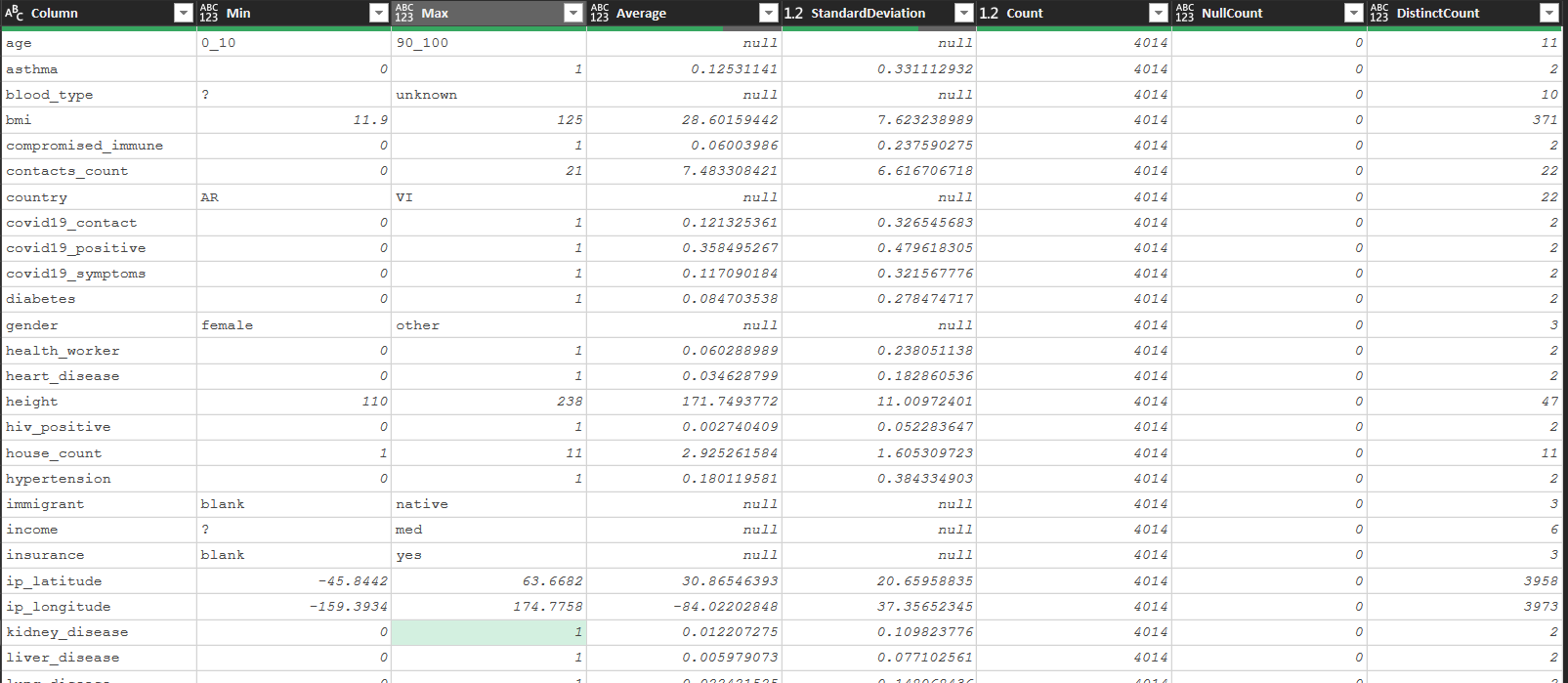


Fig 1. Data Profiling using Excel (North\_America.csv)

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Name | Data Type SQL | Data Type SSIS | Max length |
| survey\_id | varchar | string | 50 |
| survey\_date | varchar | string | 10 |
| region | varchar | string | 2 |
| country | varchar | string | 2 |
| ip\_latitude | float | double-precision float |  |
| ip\_longtitude | float | double-precision float |  |
| participant\_id | varchar | string | 50 |
| gender | varchar | string | 50 |
| age | varchar | string | 50 |
| height | int | four-byte signed integer |  |
| weight | int | four-byte signed integer |  |
| bmi | float | double-precision float |  |
| blood\_type | varchar | string | 50 |
| insurance | varchar | string | 50 |
| income | varchar | string | 50 |
| race | varchar | string | 50 |
| immigrant | varchar | string | 50 |
| response\_id | varchar | string | 50 |
| contacts\_count | int | four-byte signed integer |  |
| house\_count | int | four-byte signed integer |  |
| public\_transport\_count | int | four-byte signed integer |  |
| working | varchar |  |  |
| covid19\_positive | int | four-byte signed integer |  |
| covid19\_symptoms | int | four-byte signed integer |  |
| covid19\_contact | int | four-byte signed integer |  |
| asthma | int | four-byte signed integer |  |
| kidney\_disease | int | four-byte signed integer |  |
| liver\_disease | int | four-byte signed integer |  |
| compromised\_immune | int | four-byte signed integer |  |
| heart\_disease | int | four-byte signed integer |  |
| lung\_disease | int | four-byte signed integer |  |
| diabetes | int | four-byte signed integer |  |
| hiv\_positive | int | four-byte signed integer |  |
| hypertension | int | four-byte signed integer |  |
| other\_chronic | int | four-byte signed integer |  |
| nursing\_home | int | four-byte signed integer |  |
| health\_worker | int | four-byte signed integer |  |
| risk\_infection | int | four-byte signed integer |  |
| risk\_mortality | float | double-precision float |  |
| fact\_id | int | four-byte signed integer |  |
| location\_id | int | four-byte signed integer |  |
| day | int | four-byte signed integer |  |
| month | int | four-byte signed integer |  |
| year | int | four-byte signed integer |  |

* With data profiling, I can determine variables’ types, as well as where it should be in the tables. In the end, this is how I designed my schema:

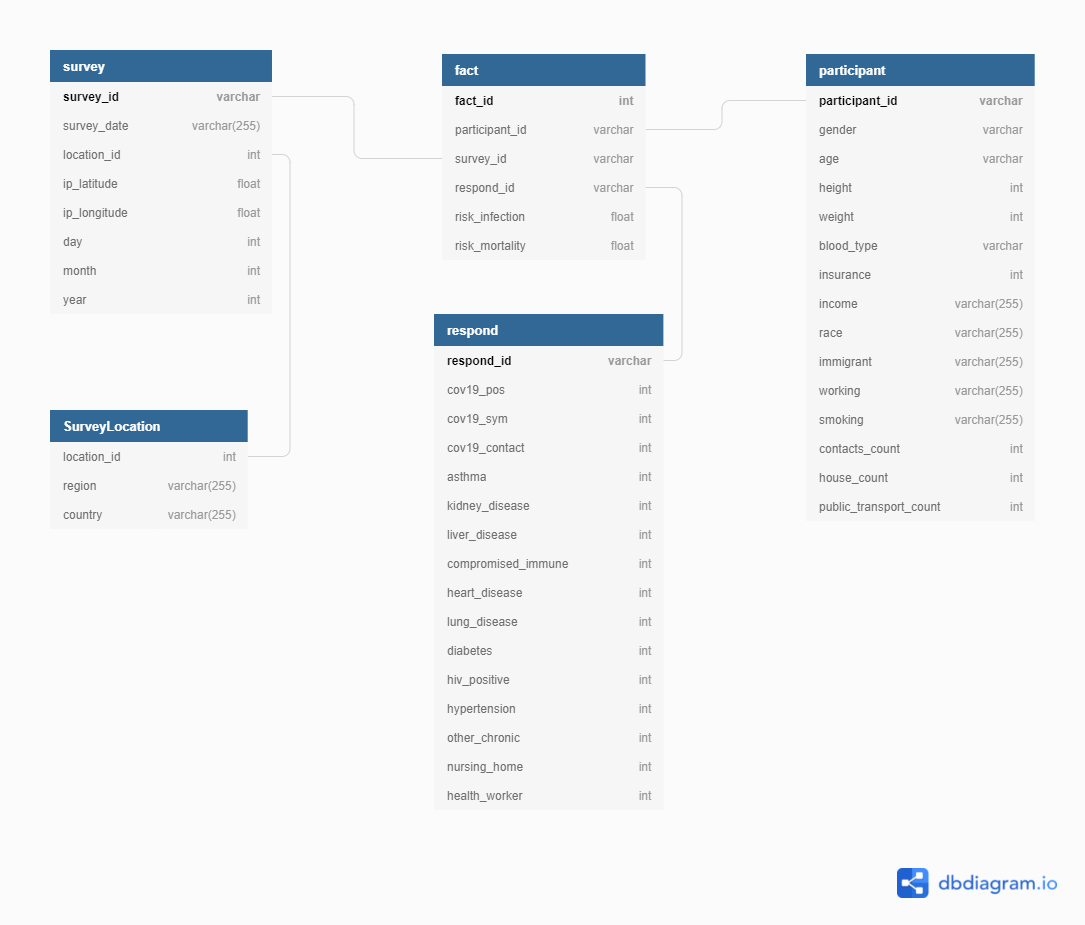


Fig 2. Star Model Diagram.

**Task 2:**

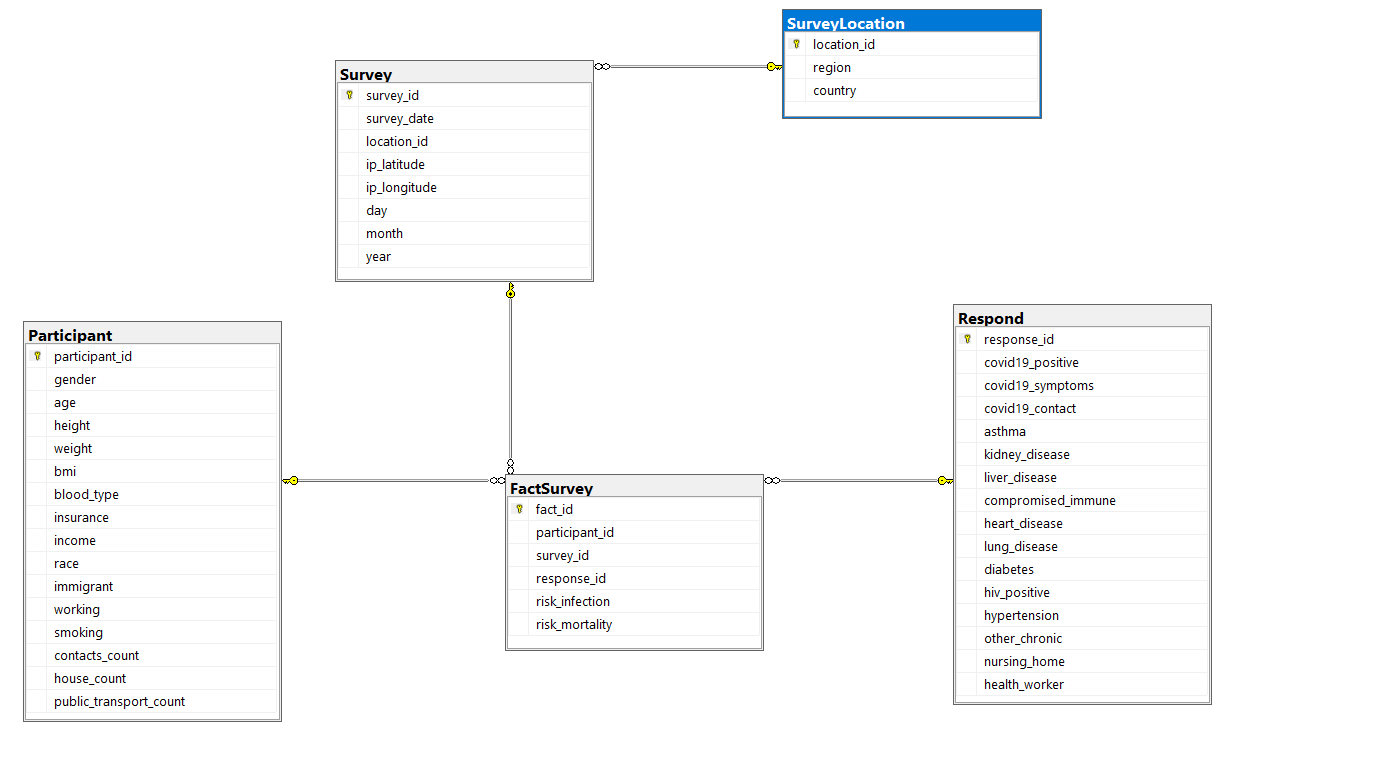
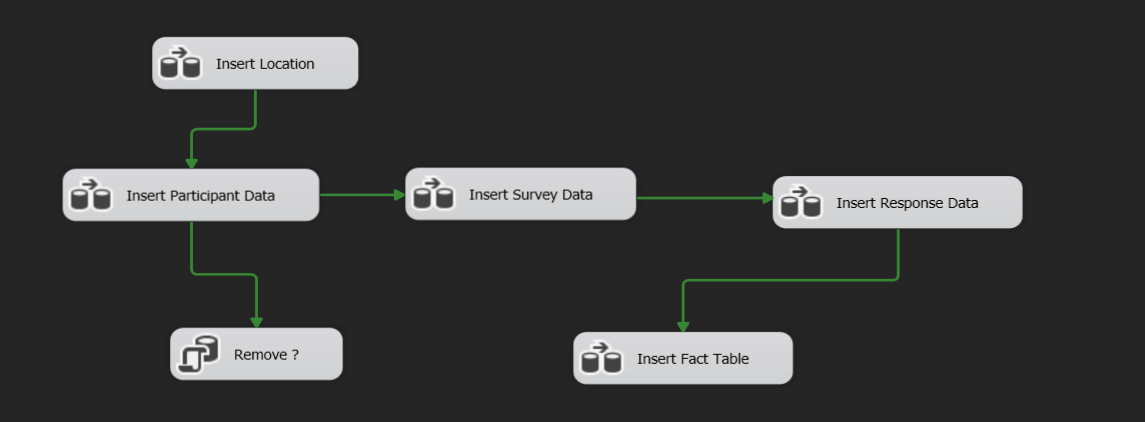


Fig 3. ER Diagram in SQL Server Management Studio

* The above figure is created using SQL. With 4 dimensions and 1 FactTable.

**Task 3:**

**a) Overview**



* Due to my constraint (foreign keys in Fact Table), I have to insert the data sequentially.

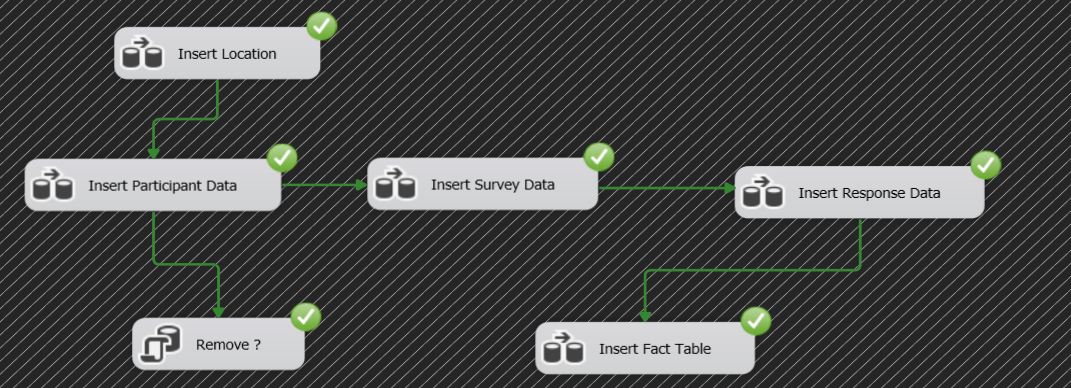
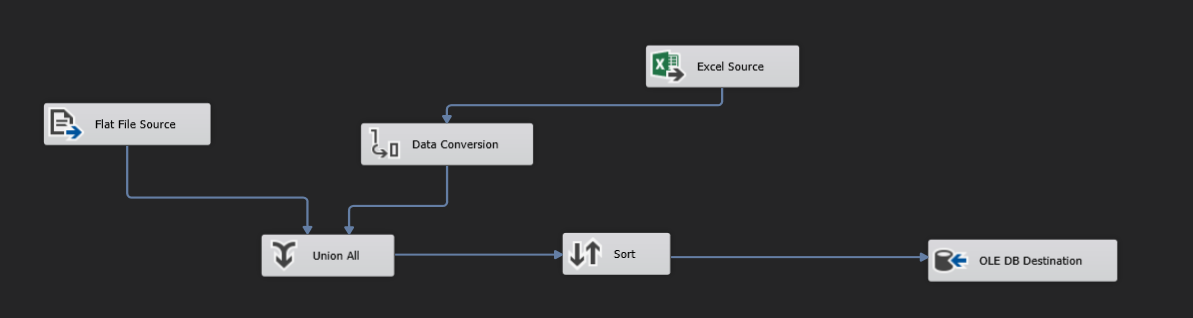
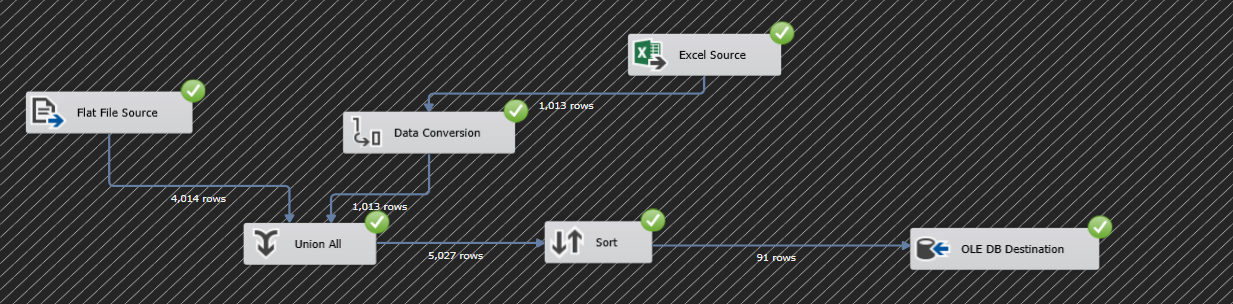


Fig 6. ETL Project Control Flow

* My Control Flow consists of: 5 different Data Flows to insert into my tables, and 1 SQL Task to delete the “?” records.

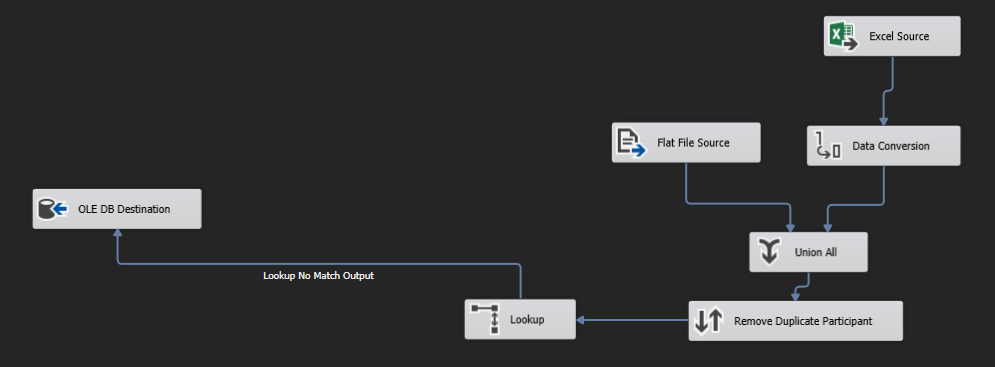
1. **Insert Location**

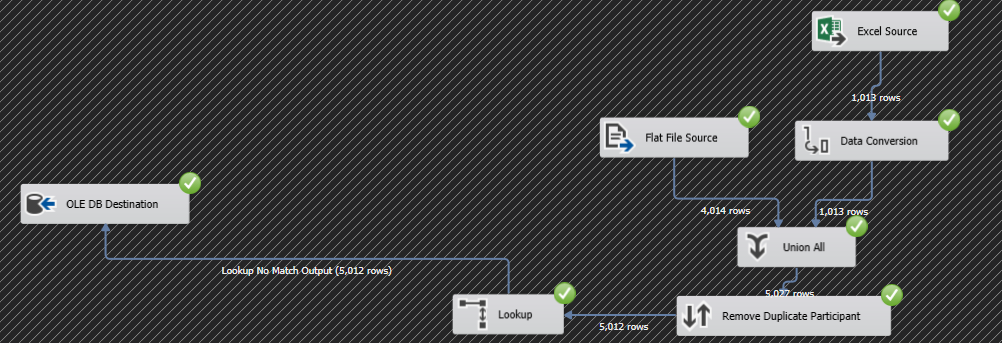




* From the excel and csv file, combine them, then sort + remove duplicates values in *region* and *country* and then insert into the database
* Data Conversion: convert *region* and *country* from Unicode String to String ( to match up with CSV datatype)
* Union All: Combine region and country from csv file and excel file.
* Sort: Sort *Region/ country* then remove duplicates

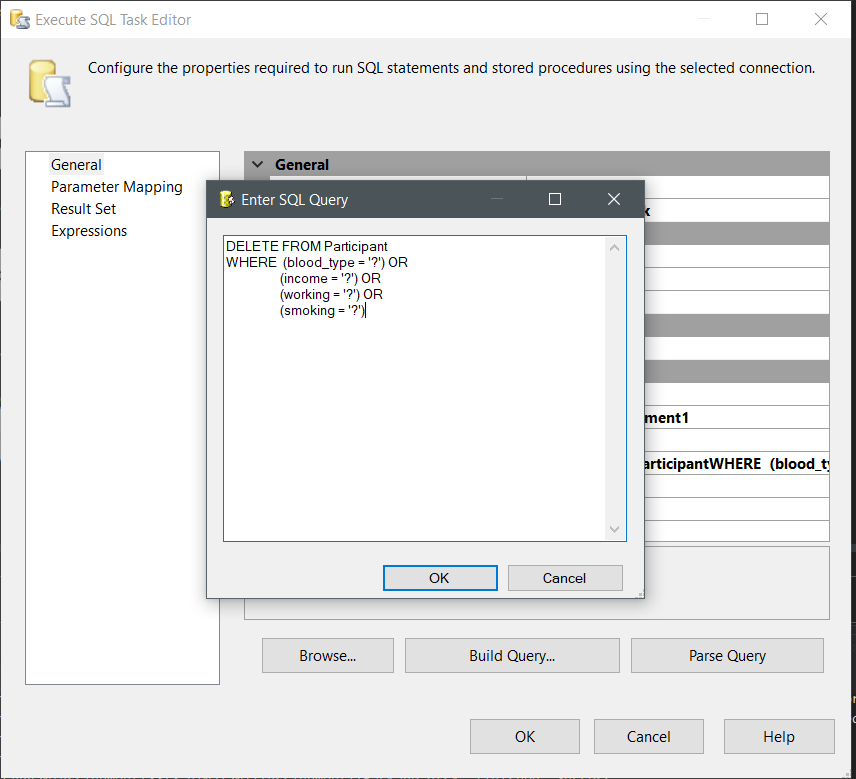
1. **Insert Participant Data**





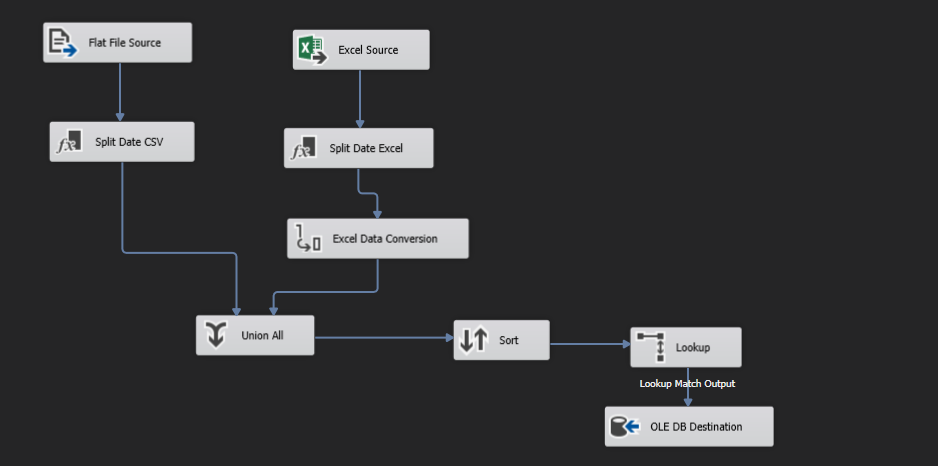
* From the excel and csv file, combine them, sort + remove all the duplicates records and insert into the database
* Data Conversion: convert Excel’s data type to match with CSV’s data type
* Union All: Combine all data from 2 files.
* Remove Duplicate Participant: sort base on *participant\_id* and remove all duplicate records.
* Lookup: Look up for non-existing data in the database before insert

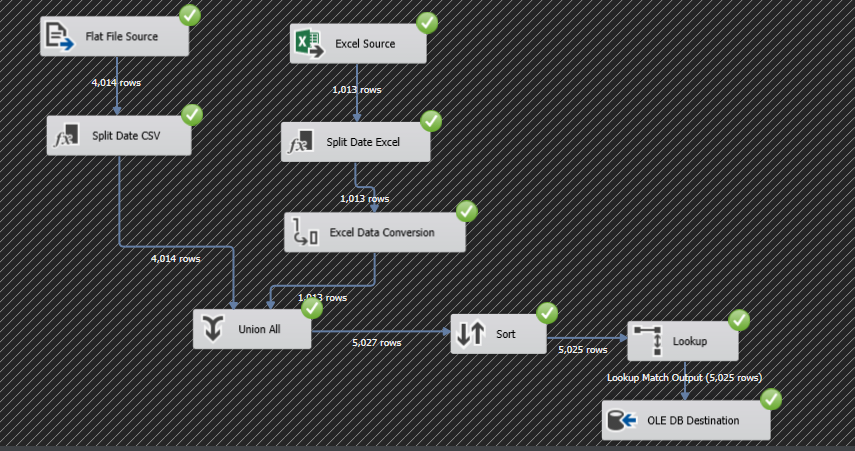
1. **Remove ?**



* After insert the participant info into database, delete the records with “?” values.

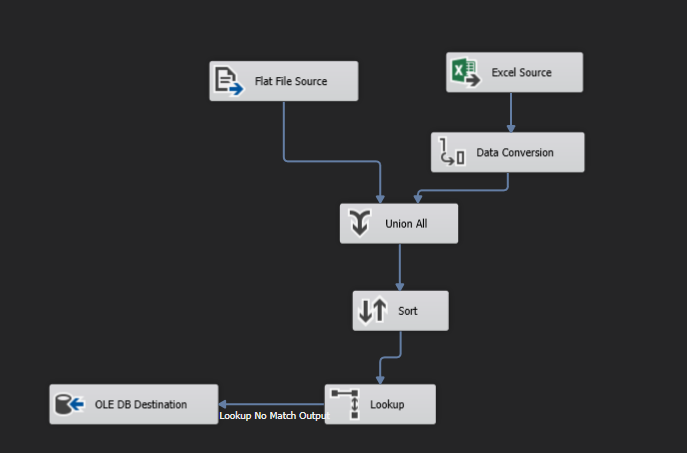
1. **Insert Survey Data**

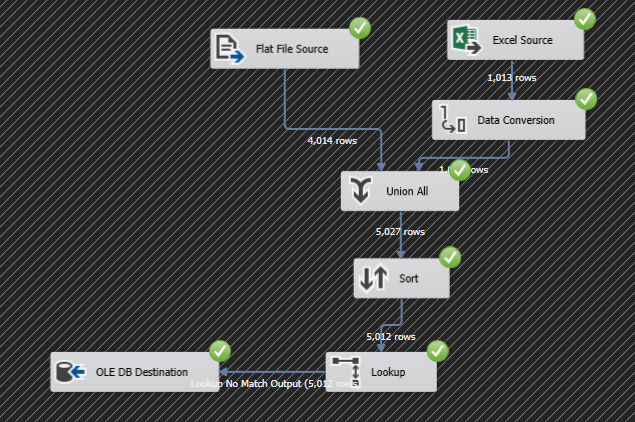




* From the excel and csv file, split the date into 3 columns, combine data from 2 files, sort + remove all duplicate records and then insert into database.
* Data conversion: Convert Excel’s data type to match with CSV’s data type
* Split Date: I first read the date as string, then split them into Day-Month-Year, then insert them into the data as 3 new columns
* Union All: Combine 2 files (with new 3 columns)
* Sort: Sort the data base on *survey\_id* and remove all duplicate records
* Lookup: Look up for location\_id to add into each survey record, then insert to the database.

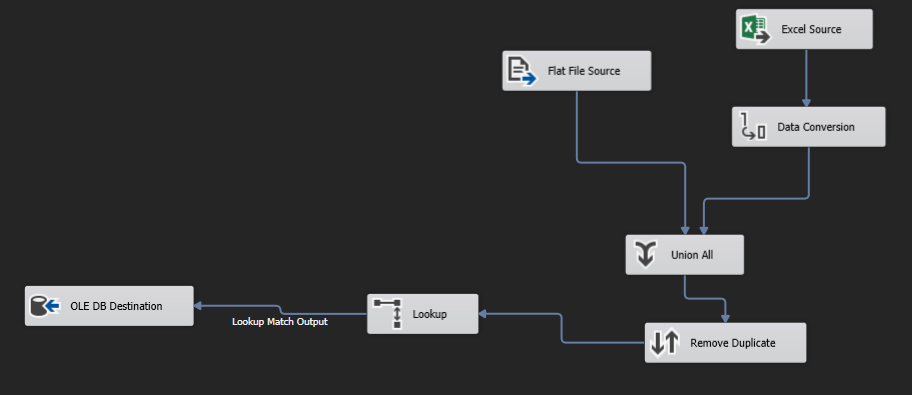
1. **Insert Response Data**

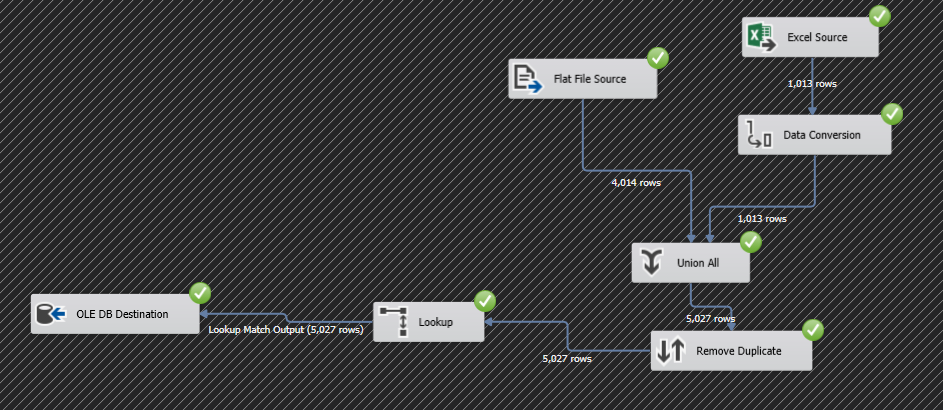




* From the excel and csv file, combine them, remove duplicate records and insert into database.
* Data Conversion: Convert Excel’s data type to match with CSV’s data type
* Union All: combine data from 2 files
* Sort: Sort the data base on *response\_id* and remove duplicate records.
* Lookup: Look up for non-existing data before insert into the database

1. **Insert Fact**





* From the excel and csv file, combine them, then sort + remove duplicates values and then insert into the database
* Data Conversion: Convert Excel’s data type to match with CSV’s data type
* Union All: combine data from 2 files
* Sort: Sort the data base on *survey\_id, participant\_id* and *response\_id* and remove duplicate records.
* Lookup: Look up for non-existing data before insert into the database

**Task 4:**

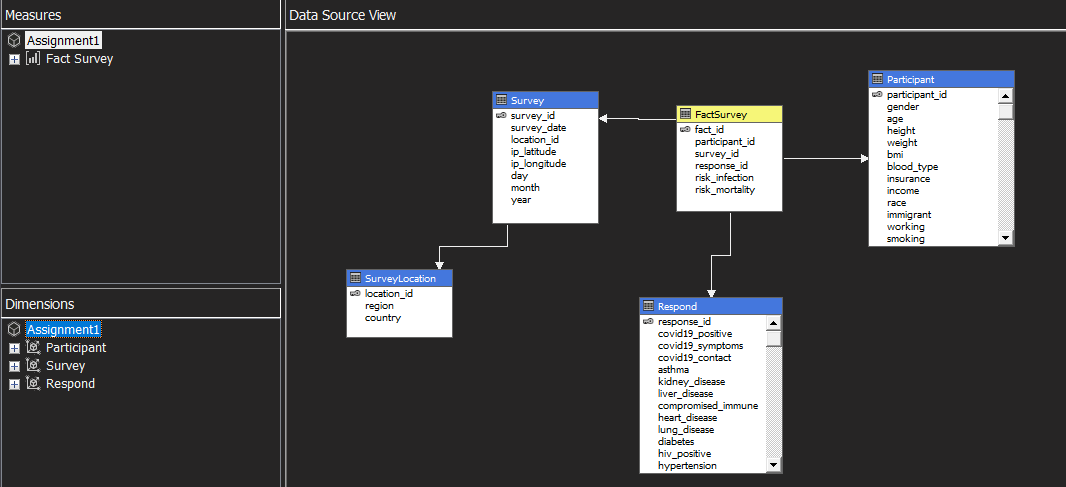
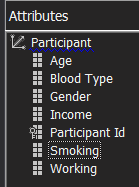


Fig 4. SSAS Cube Overview.

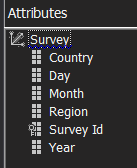
* The cube includes:
  + 1 fact table FactSurvey and 3 dimensions: Participant, Survey and Respond.
  + 3 measures: Risk Infection (Sum), Risk Mortality (max) and Fact Survey Count (Count).
  + 3 Dimensions: Participant,Survey and Respond with:

1. Participant

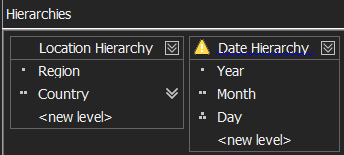


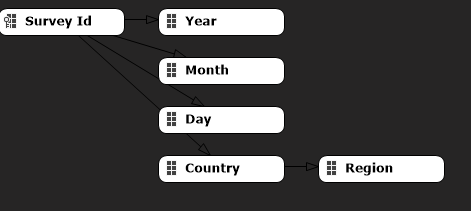
* + - Allow user to search by Age, Blood Type, Gender, etc..
    - No Hierachies in this dimension

1. Survey

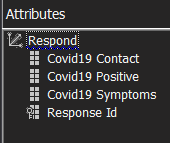


* + - Allow user to search by location(by region or country or both) or time( date month year separately or combined)
    - Contains 2 hierarchies





1. Respond



* + - Allow user to search by Covid19 Contact, Positive or Symptoms
    - No Hierachies in this dimension
* Deployed cube:

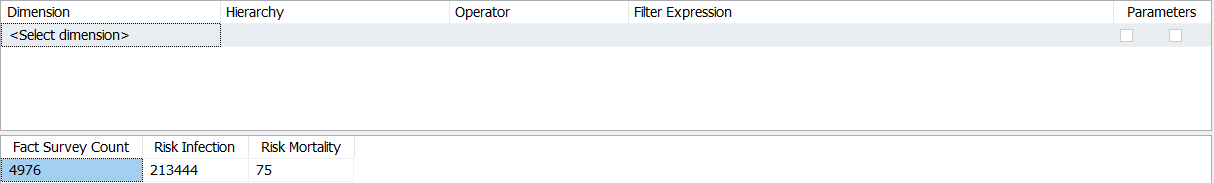


Fig 5. Deployed Cube Overview.

* Search using one dimension:

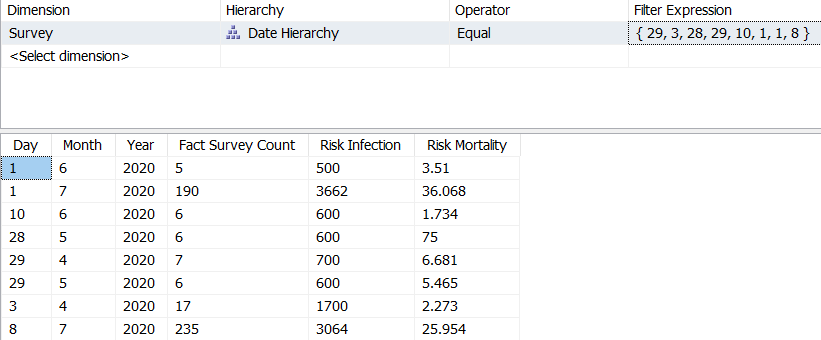


Fig 6. Search using Date Hierarchy in Survey Dimension

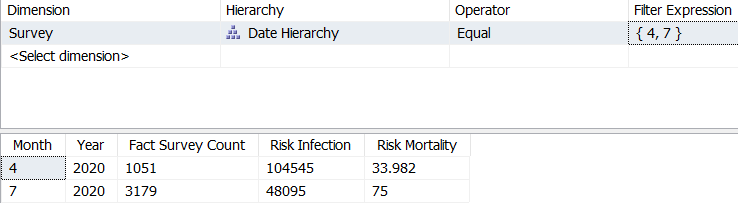


Fig 7. Search by Month in Date Hierarchy

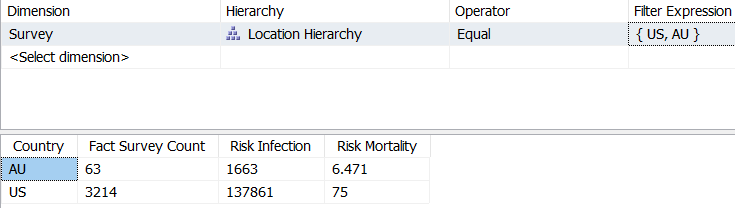


Fig 8. Search by Country in Location Hierarchy

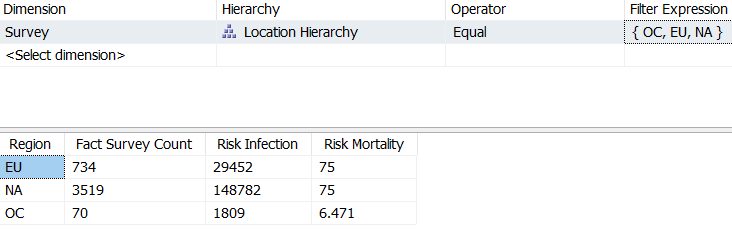


Fig 9. Search by Region in Location Hierarchy

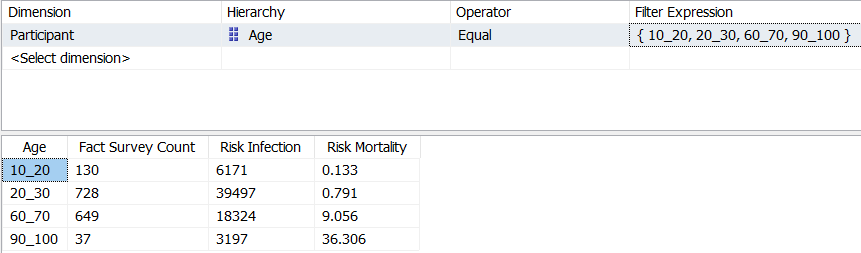


Fig 10. Search by Participant’s Age

* Search using multiple dimensions:

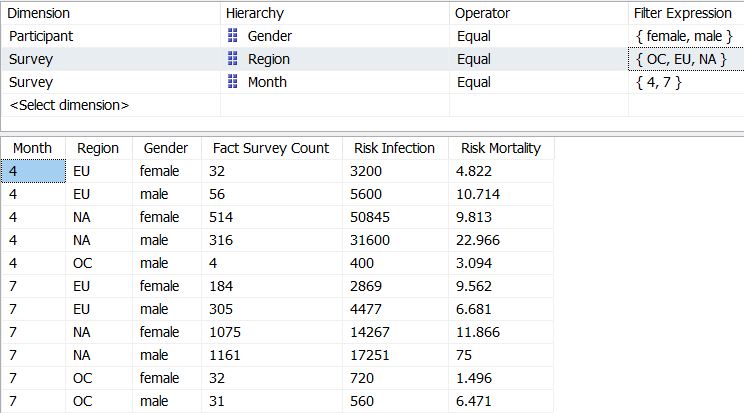


Fig 11. Search using 2 dimensions Survey and Participant