DWI LAB – Mini-Project stage II - RAPORT

Piotr Stróż and Piotr Kontek

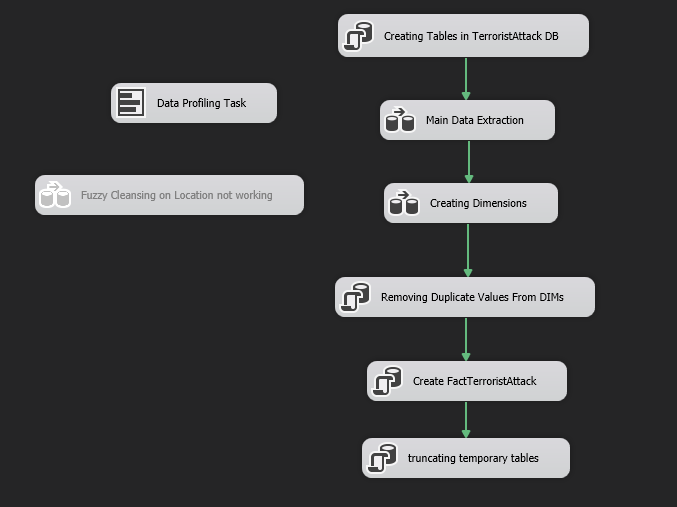
# 1.Data profiling

We used data profiling, to find out what types should we use to store our data. We use precise size of varchars, small ints, binary values, so no unnecessary storage is reserved for our database.

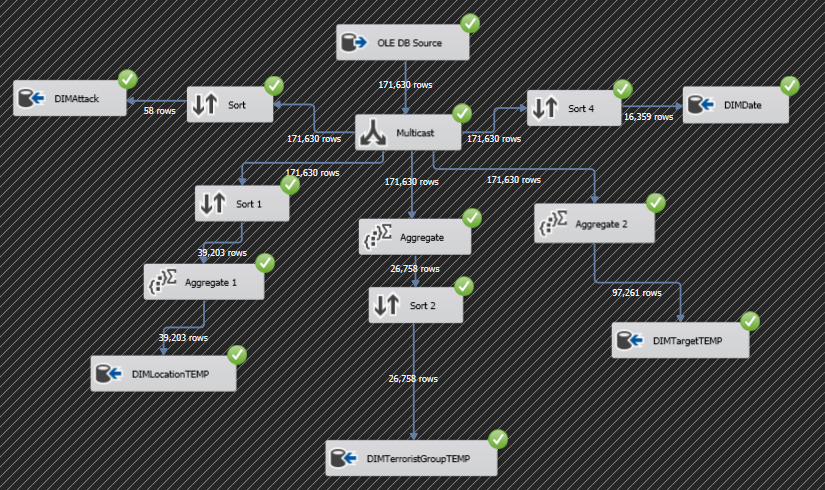
Moreover we found that some columns happen to have a small number of null values, which we consider as unknown value.

# 2. Implementation

Our data source is terroristDB.csv. It contains 181692 rows. First SQL TASK creates database with all necessary tables, then all data is extracted to MainData table, and it is presorted and fuzzy grouped, which shortens our database to 171630 rows(10000 rows were removed as effect of duplicate values removal and fuzzy grouping).



After Main Data Extraction we create dimensions from it, again data is sorted and duplicates are removed once again, as now with smaller number of columns it is still possible to find dimension specific duplicate. DIMDate and DIMAttack contain numeric values, so fuzzy grouping or so is not possible on them, only duplicates removal, therefore they are instantly placed in final tables. Dimensions: Location, TerroristGroup, Target, still needs some processing and are inserted inside temporary tables.



Then some final filtering of three previously mentioned dimensions, by executing SQL TASK, which selects only distinct values for our tables. This solution is necessary, as Visual Studio SSIS string matching is different from SQL Server Management Studio, and produces different results.

With proper values inside our dimensions we reference their keys inside our FactTerroristAttack fact table. We used this SQL query to achieve desired goal:

insert into FactTerroristAttack

(numberOfKilledCivilians, numberOfKilledTerrorists,

numberOfWoundedCivilians, numberOfWoundedTerrorists,

valueOfDestroyedProperty, dateID, locationID, attackID,

terroristGroupID, targetID)

select nkill, nkillter, nwound, nwoundte, propvalue,

dd.id, dl.id, da.id, dtr.id, dt.id

from MainData md

join DIMDate dd on dd.day = md.iday and dd.month = md.imonth and dd.year = md.iyear

join DIMAttack da on da.attackType = md.attacktype1\_txt and da.wasMultiple = md.multiple and da.wasSuccesfull = md.success and da.wasSuicide = md.suicide

join DIMTarget dt on dt.nationality = md.natlty1\_txt and dt.targetName = md.target1 and dt.targetSubtype = md.targsubtype1\_txt and dt.targetType = md.targtype1\_txt

join DIMLocation dl on dl.City = md.city and dl.Region = md.region\_txt and dl.State = md.provstate and dl.Country = md.country\_txt

join DIMTerroristGroup dtr on dtr.motive = md.motive and dtr.name = md.gname and dtr.numberOfMembersParticipating = md.nperps and dtr.usedWeaponType = md.weaptype1\_txt

Last step is to remove records from temporary tables from our database.

In the end we visually compared data of two parallel queries, first from MainData table, second from join of all dimensions to fact table. They provide identical results, which proves that process of creating and normalizing our Datawarehouse went correct.

