**ACS Automation Document**

**Aim:**

TO improve the processing time of the loading the ACS data.

**Solution:**

Create an intermediate table with all the aggregation values and use the intermediate table to insert data into the main table.

How it differs from earlier implementation: The earlier implementation had the table being created and the values being updated by columns (for muni\_id > 352). Generally updating takes more time than insert and the update was happening for every column and not all at once. This is the reason for the performance issues earlier. Now this solution writes Insert statements for muni\_id>352.

The problem with insert solution is we need to perform aggregation for all muni\_id > 352 and the where condition for each id, so it calls for writing separate insert statements for each muni\_id. So the process of creating SQL scripts for insert has been automated using R. We have a Req\_functions. R which has functions that generates SQL scripts that could be run directly on postgres for Inserting ACS data efficiently.

**Tasks and challenges:**

* Creating the intermediate table involves creating repetitive insert queries that have large number of columns and impossible to create for so many tables
* Also create a query file to create and insert data in the final table

**Solution:**

To create an R script that can generate the intermediate

Files:

1. Req\_functions.R
2. Sample.R

**1. Req\_funtions.R**:

This file consists of all the functions required to get the previous ACS sql file and generate sql files for the new method. This file consists of four functions.

* 1. SqlFileAutomation:

This function takes the following as inputs.

Domain – The short name given to the the five to the table that we are currently dealing with. Ex: table b17001 is given domain name as ‘poverty\_by\_population’. The temporary of intermediate table created will have this AGGRE\_MUNI\_ID\_*”domain”* as the table name.

Outfilepath –output file path for the created .sql file to be stored

QueryTables – This a vector of FROM/JOIN tables except acs\_id\_lookup and acs\_muni\_geoid

QueryTableAlaises – This is a vector alias name given to the queryTables with alias name positions corresponding to positions in queryTables

Maxcols – this is the number of columns in the raw table except logrecno

1.2. LINECLEAN /ONELINEQ:

Onelineq function takes the input query file name and the input file directory as parameters and uses lineclean function to have one line queries without unnecessary indents spaces and comments. This is done to pass the query as parameters to the function query2Creator.

1.3. Query2Creator

The first three queries from the output file of ONELINEQ is the first three inputs of this function. They are DROPQ (drop query), CREATEQ (Create Query), q1 (insert Query1). It also has two other parameters queryTableAlaises (the same vector used in sqlfileAutomation function) and outFilepath (the path of the output file also the same as in sqlfileAutomation function)

1.4 Inputtooutput

The function input to output calls all the other functions internally and produces two outputfiles. Running those two output sql files in postgres will deliver the intended result. The usage of this function is shown in sample.R file.

**2. SAMPLE.R**

This file shows the parameters that needs to be given for the whole task and it calls the input-to-output function. Here we are showing sample code for educational\_attainment\_15002 educational data

The users are expected to give values for the following parameters:

Infilepath: Path where the earlier sql file used for the processing the table exists

Outfilepath: Path where the new sql file is dropped

Filename The name of the earlier sql file that was used to process the same table

Maxcols: The number of columns that the table has excluding the following(muni\_id ,municipal ,geoid logrecno,acs\_year ) . These file columns are present in all acs tables

QueryTableAlaises: The vector of alias name given for the tables (except acs\_id\_lookup, acs\_muni\_geoid) in the first create statement of the input sql file

domain: Name of the table and the output file.