



Datasets Cleaning Automation

Sponsored by





Delora Parks
Scrum Master



George Zoto
Product Owner



Arsadur Rahman
Developer



Vincentt Gomes
Developer

Team No Kid Hungry



Outline

1. **Introduction**
2. **Data Overview**
3. **Approach**
4. **Performance Evaluation**
5. **Tools Used**
6. **Future Prospect**



1

Introduction

Problem Definition - Objectives - Complexity



“

The focus of this project is to automate data cleaning from different datasets based on some provided criteria(Data Dictionary) to have quick and time saving approach than the current current solution.



**Why need a different
Solution ?**

A diagram illustrating a data cleaning process. It features three document icons at the bottom, with a large purple arrow pointing upwards from them to a single document icon at the top. This central graphic is surrounded by a circular arrow, with a purple segment on the left and a grey segment on the right. The entire diagram is set against a background of concentric white circles on a dark grey field.

**Current solution takes a lot of time to
clean raw data files**



Objectives

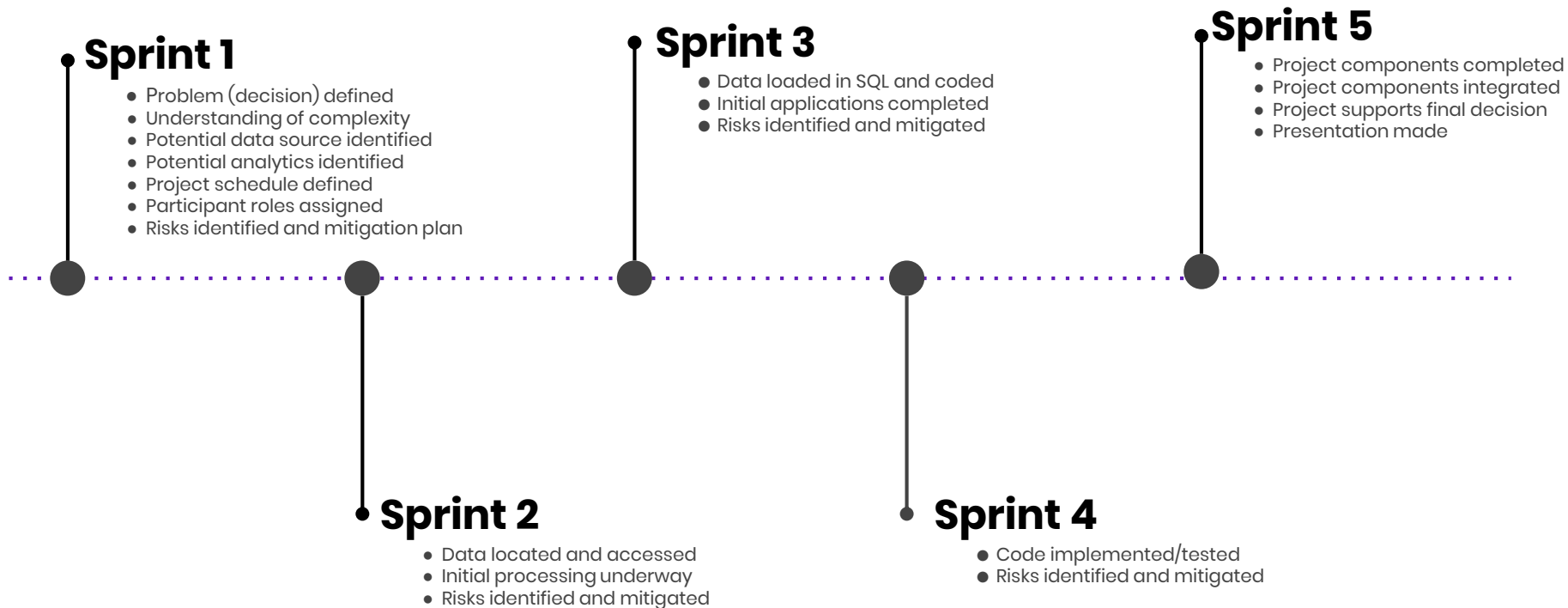
- Identifying the raw data files and clean them based on Data Dictionary autonomously.
- Identify different cleaning steps for a particular state and apply them on SQL
- Revealing the implications for the automatic cleaning..

Risks and Complexities

- Identifying scenarios of potential data loading in SQL
- Combining across different datasets of lunch and breakfast data with data dictionary
- Calculating more than 20 new fields and add them to the clean data set



Project Sprint History





2

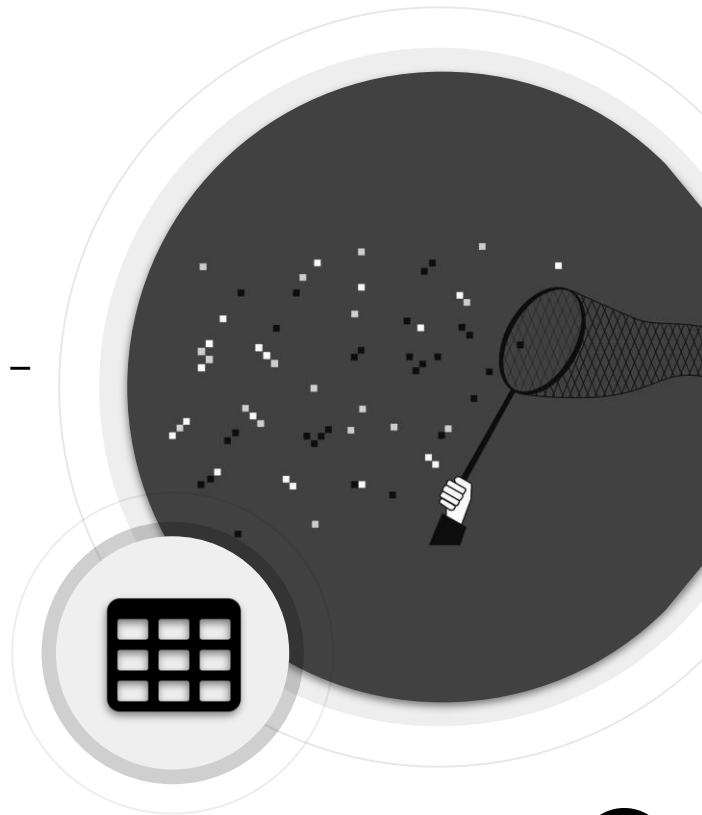
Data Overview

Data Description – Pre Processing – Distribution Graphs



Datasets

- Wisconsin Data sets
 - Raw data
 1. 2017_2018_SBP_MEAL_PARTICIPATION
 2. SEP_2018_JUNE_2019_NSLP_MEAL_PARTICIPATION – Strength
 3. WI NCES School Data, SY15-16
 - Data Dictionary
 1. data_dictionary_template_1
 2. data_dictionary_template_2



Data dictionary Sample

AutoSave OFF Data Dictionary_CA Breakfast SY18-19 - Protected View - Repaired - Excel arahma9

File Home Insert Draw Page Layout Formulas Data Review View Help Search Share Comments

PROTECTED VIEW Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing

A1 Clean Data No.

	A	B	C	D	E	F	G	H	I	J	K
	Clean Data No.	Clean Data Column Names	Clean Column in Raw Data	Critical Clean Column	Check						
1	n/a	NOT USED	n/a	n/a	OK						
2	n/a	SEE NOTES	n/a	n/a	OK						
3	41	BAB Implementation Date from Campaign Tracking Data	no	may not have	Omission						
4	42	BAB Implementation Date from State Agency Tracking	no	may not have	Omission						
5	43	BAB Implementation Date from Other Source	no	may not have	Omission						
6	44	BAB Implementation Date Other Source, Note Source Here	no	may not have	Omission						
7	37	Breakfast Delivery Model from Campaign Tracking Data-Original	no	may not have	Omission						
8	38	Breakfast Delivery Model from State Agency Tracking-Original	no	may not have	Omission						
9	39	Breakfast Delivery Model from Other Source-Original	no	may not have	Omission						
10	40	Breakfast Delivery Model from Other Source, Note Source Here	no	may not have	Omission						
11	30	Breakfast Meals-Free	yes	Absolute need	OK						
12	33	Breakfast Meals-Free and Reduced	no	Medium need	OK						
13	32	Breakfast Meals-Paid	yes	Medium need	OK						
14	31	Breakfast Meals-Reduced	yes	Absolute need	OK						
15	13	CEP (Y/N)	yes	Medium need	OK						
16	7	City	yes	Medium need	OK						
17	18	Claim Date	yes	Absolute need	OK						
18	19	Claim Month	no	Medium need	Priority omission						
19	20	Claim Year	no	Medium need	Priority omission						
20	6	County	yes	Medium need	OK						
21	3	District ID	yes	Absolute need	OK						
22	4	District Name	yes	Absolute need	OK						
23	23	Enrollment-Free	yes	Absolute need	OK						
24	25	Enrollment-Free and Reduced	no	Medium need	OK						
25	22	Enrollment-Paid	yes	Medium need	OK						
26	24	Enrollment-Reduced	yes	Absolute need	OK						
27	21	Enrollment-Total	yes	Absolute need	OK						
28	64	ISP	no	may not have	Omission						
29	30	Lunch Meals-Free	yes	Absolute need	OK						

Clean Data Name Lookup Instructions Template1 Template2 Template3 Template4 Template5 Template6 Template7 Template8

Ready

Search the web and your PC

8:23 PM 2/17/2020

Raw Data Sample

AutoSave On fpm1819 - Protected View - Excel arahma9

File Home Insert Draw Page Layout Formulas Data Review View Help Search Share Comments

PROTECTED VIEW Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing

A1 fx Unduplicated Student Poverty – Free or Reduced Price Meals Data 2018–19

	A	B	C	D	E	F	G	H
1	Unduplicated Student Poverty – Free or Reduced Price Meals Data 2018–19							
2	Academic Year	County Code	District Code	School Code	County Name	District Name	School Name	District Type
3	2018-2019	01	10017	0112607	Alameda	Alameda County Office of Education	Envision Academy for Arts & Technology	County Office
4	2018-2019	01	10017	0123968	Alameda	Alameda County Office of Education	Community School for Creative Education	County Office
5	2018-2019	01	10017	0124172	Alameda	Alameda County Office of Education	Yu Ming Charter	County Office
6	2018-2019	01	10017	0125567	Alameda	Alameda County Office of Education	Urban Montessori Charter	County Office
7	2018-2019	01	10017	0130401	Alameda	Alameda County Office of Education	Alameda County Juvenile Hall/Court	County Office
8	2018-2019	01	10017	0130419	Alameda	Alameda County Office of Education	Alameda County Community	County Office
9	2018-2019	01	10017	0131581	Alameda	Alameda County Office of Education	Oakland Unity Middle	County Office
10	2018-2019	01	10017	0136101	Alameda	Alameda County Office of Education	Connecting Waters Charter - East Bay	County Office
11	2018-2019	01	10017	0136226	Alameda	Alameda County Office of Education	Opportunity Academy	County Office
12	2018-2019	01	10017	0137448	Alameda	Alameda County Office of Education	Aurum Preparatory Academy	County Office
13	2018-2019	01	10017	6001788	Alameda	Alameda County Office of Education	Cox Academy	County Office
14	2018-2019	01	10017	6002000	Alameda	Alameda County Office of Education	Lazear Charter Academy	County Office
15	2018-2019	01	31609	0131755	Alameda	California School for the Blind (State Special)	California School for the Blind	State Special
16	2018-2019	01	31617	0131763	Alameda	California School for the Deaf-Fremont (State Special)	California School for the Deaf-Fremont	State Special
17	2018-2019	01	61119	0000001	Alameda	Alameda Unified	Nonpublic, Nonsectarian Schools	Unified School
18	2018-2019	01	61119	0106401	Alameda	Alameda Unified	Alameda Science and Technology Institute	Unified School
19	2018-2019	01	61119	0111765	Alameda	Alameda Unified	Ruby Bridges Elementary	Unified School
20	2018-2019	01	61119	0119222	Alameda	Alameda Unified	Nea Community Learning Center	Unified School
21	2018-2019	01	61119	0122085	Alameda	Alameda Unified	The Academy of Alameda	Unified School
22	2018-2019	01	61119	0126656	Alameda	Alameda Unified	Maya Lin	Unified School
23	2018-2019	01	61119	0130229	Alameda	Alameda Unified	Alameda High	Unified School
24	2018-2019	01	61119	0130609	Alameda	Alameda Unified	Alameda Community Learning Center	Unified School
25	2018-2019	01	61119	0130625	Alameda	Alameda Unified	Alternatives in Action	Unified School
26	2018-2019	01	61119	0131805	Alameda	Alameda Unified	The Academy of Alameda Elementary	Unified School
27	2018-2019	01	61119	0132142	Alameda	Alameda Unified	Encinal Junior/Senior High	Unified School
28	2018-2019	01	61119	0134304	Alameda	Alameda Unified	Island High (Continuation)	Unified School

Title Page FRPM School-Level Data Data Field Descriptions

8:21 PM 2/17/2020

The image features a large white circle centered on a black background. To the left of the white circle, there is a series of three concentric circles in shades of gray, with the number '3' in white at the center. To the right of the white circle, there is a series of five concentric white circles. The word 'Approach' is written in bold black text across the middle of the white circle.

3

Approach


Autonomous Pre Processing(1)

**Load in Dictionary
Templates
In
SQL**

**Load in Raw State file
(Dynamically)**

This is done for each file

Autonomous Pre Processing(2)



**Join Dictionary Templates
to Raw State Files to get a
clean data lookup for
column names**

Autonomous Pre Processing (3)



**Use lookup from previous
step to determine which
columns need to be deleted
and delete them
dynamically**

Autonomous Pre Processing



**Use lookup from previous
step to determine which
columns need to be deleted
and delete them
dynamically**

Autonomous Pre Processing



**Use lookup from above to
determine which columns
need to be renamed and
rename them dynamically**

Autonomous Pre Processing



Create conditional columns

Autonomous Pre Processing



**Join tables from above into
one table**

Autonomous Pre Processing



Create calculated columns

Sponsor given Rules

“

Will have to match with the Clean Data files.

”



5

Performance Evaluation

Comparing the generated clean data file
in T-SQL with the given clean data file by
our sponsor



6

Tools Used

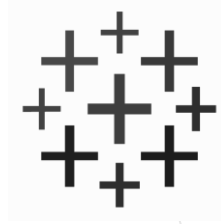
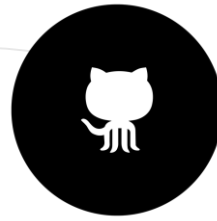
Tools - Commands

Tools

- MS T-SQL
- Skype
- YouTrack
- Hangouts
- Github



T-SQL
Commands



Commands

- CREATE TABLE
- BULK INSERT
- DECLARE
- SELECT *
- INNER JOIN/FULL JOIN
- CASE WHEN
- WHILE LOOP
- AGGREGATION
- SET FUNCTIONS





Results

The SQL code joined all the datasets, type of Join to perform

Join Scenario #1

After performing the SQL code run we achieved to gather 614columns.



7

Future prospect

Results – Conclusion



- The code we established can be used for all 40 states But need some changes based on the raw data structure
- Some other possible cases it can be run/rerun on stored procedure



**Thank
You** 

YouTrack Story

YT Issues Dashboards Agile Boards Reports Projects Create

DAEN690005NHKS2020... Filter cards on the board

Sprint 2 11 — 24 Feb S XL TV

Open 0 In Progress 0 To Verify 0 Done 0

Uncategorized Cards 2 cards

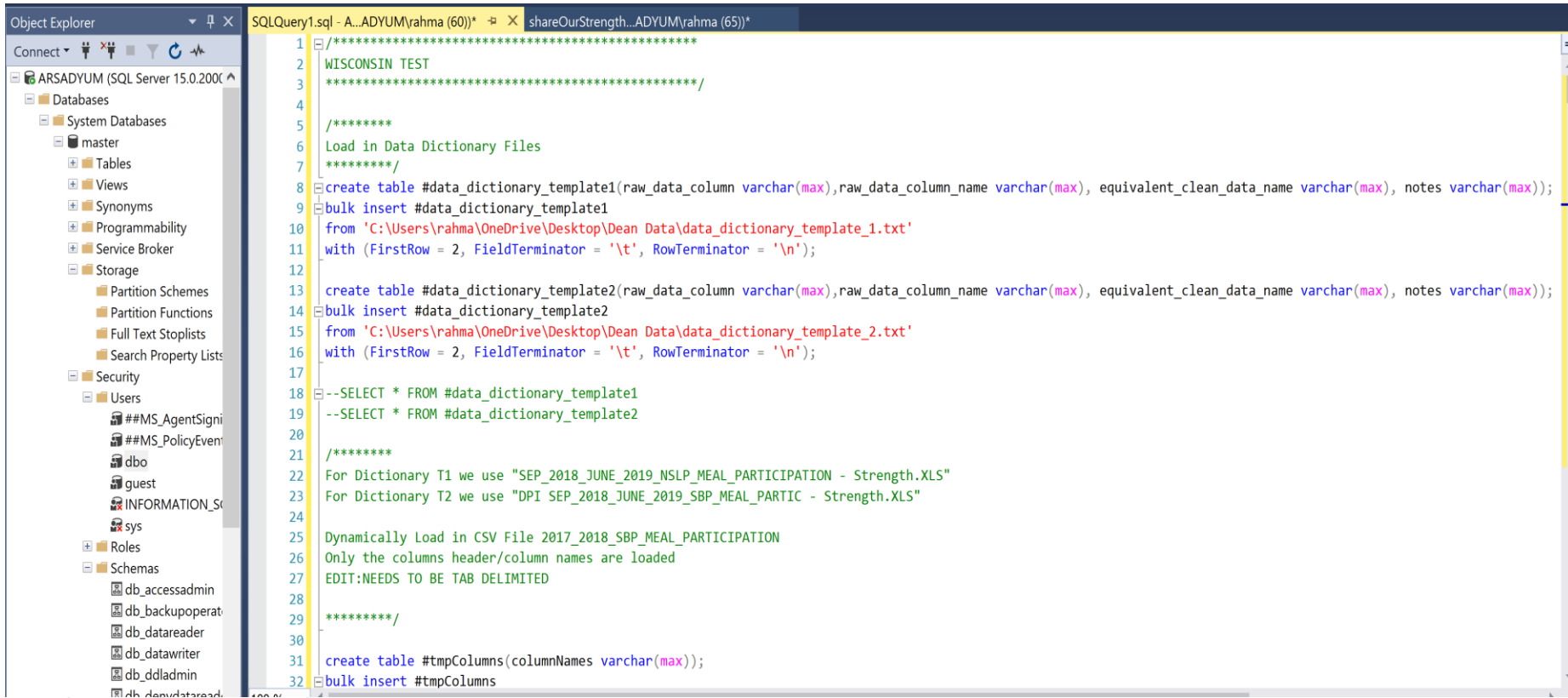
- DAEN690005NHKS2020-14 Weekly sponsor meeting
Normal Task Not estimated
- DAEN690005NHKS2020-9 Create Data Cleaning Process Doc
Normal Task Not estimated

▼ Solution Selection- DAEN690005NHKS2020-4 3 cards

- DAEN690005NHKS2020-8 Select States to Model
Normal Task Not estimated
- DAEN690005NHKS2020-2 Identify Software or Cloud Solutions to Investigate
Normal Task Not estimated
- DAEN690005NHKS2020-6 Get Demo of Current

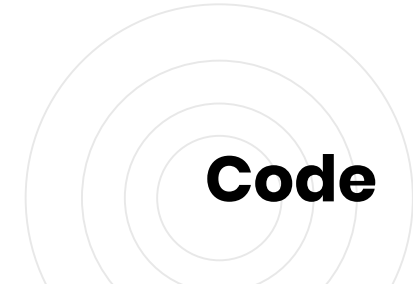
Board owner: F. Brett Berlin

New swimlane ...



```
1  /*****
2  WISCONSIN TEST
3  *****/
4
5  /*****
6  Load in Data Dictionary Files
7  *****/
8  create table #data_dictionary_template1(raw_data_column varchar(max),raw_data_column_name varchar(max), equivalent_clean_data_name varchar(max), notes varchar(max));
9  bulk insert #data_dictionary_template1
10 from 'C:\Users\rahma\OneDrive\Desktop\Dean Data\data_dictionary_template_1.txt'
11 with (FirstRow = 2, FieldTerminator = '\t', RowTerminator = '\n');
12
13 create table #data_dictionary_template2(raw_data_column varchar(max),raw_data_column_name varchar(max), equivalent_clean_data_name varchar(max), notes varchar(max));
14 bulk insert #data_dictionary_template2
15 from 'C:\Users\rahma\OneDrive\Desktop\Dean Data\data_dictionary_template_2.txt'
16 with (FirstRow = 2, FieldTerminator = '\t', RowTerminator = '\n');
17
18 --SELECT * FROM #data_dictionary_template1
19 --SELECT * FROM #data_dictionary_template2
20
21 /*****
22 For Dictionary T1 we use "SEP_2018_JUNE_2019_NSLP_MEAL_PARTICIPATION - Strength.XLS"
23 For Dictionary T2 we use "DPI SEP_2018_JUNE_2019_SBP_MEAL_PARTIC - Strength.XLS"
24
25 Dynamically Load in CSV File 2017_2018_SBP_MEAL_PARTICIPATION
26 Only the columns header/column names are loaded
27 EDIT:NEEDS TO BE TAB DELIMITED
28
29 *****/
30
31 create table #tmpColumns(columnNames varchar(max));
32 bulk insert #tmpColumns
```

Code

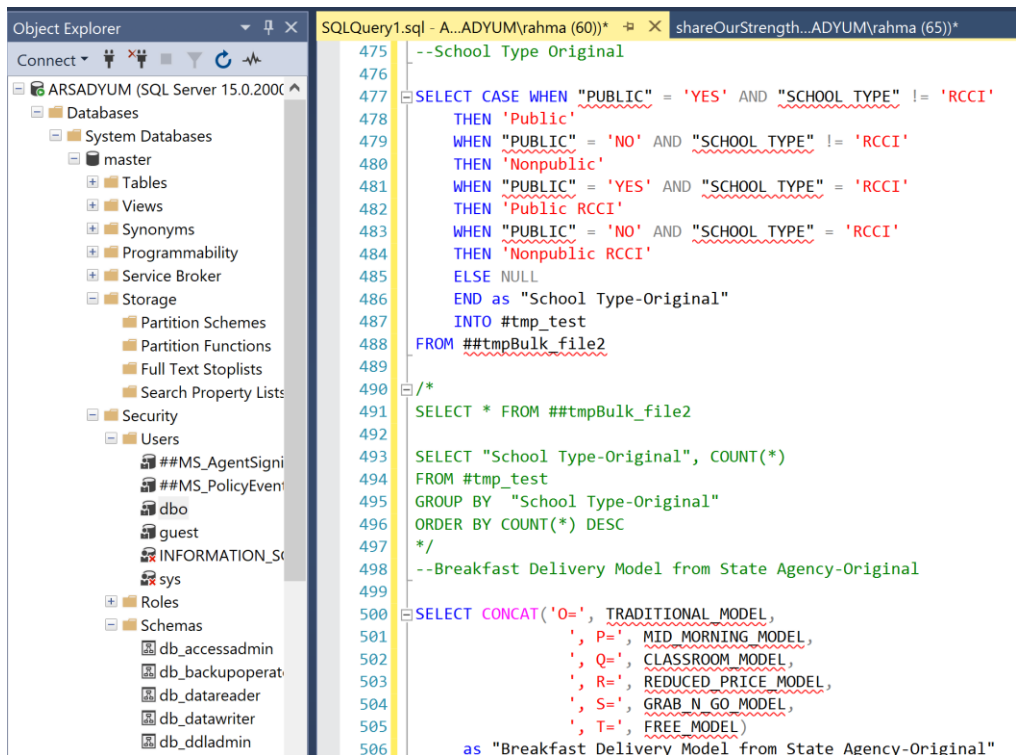


Code

```

47 -- SELECT * FROM #tmpColumns2
48 /****
49 The following loop dynamically creates a table based on the column names from above
50
51 For standardization purposes all columns are varchar(1000) and should be changed later in the script
52
53
54 *****/
55 drop table ##tmpBulk
56 declare @sql as varchar(max)
57 select @sql = 'create table ##tmpBulk (' + replace(columnNames,',',' ' varchar(max),') + ' varchar(max));
58             bulk insert ##tmpBulk
59             from 'C:\Users\rahma\OneDrive\Desktop\Dean Data\2017_2018_SBP_MEAL_PARTICIPATION.txt'
60             with (FirstRow = 2, FieldTerminator = '\t', RowTerminator = '\n');
61
62
63             create table ##tmpBulkColumns (' + replace(columnNames,',',' ' varchar(max),') + ' varchar(max));
64             bulk insert ##tmpBulkColumns
65             from 'C:\Users\rahma\OneDrive\Desktop\Dean Data\2017_2018_SBP_MEAL_PARTICIPATION.txt'
66             with (FirstRow = 1, LastRow = 1, FieldTerminator = '\t', RowTerminator = '\n');
67
68             '
69 from #tmpColumns3
70
71 exec(@sql)
72
73 -- SELECT * FROM ##tmpBulk
74 --Drop column tables
75 drop table #tmpColumns
76 drop table #tmpColumns2
77

```



```
475 --School Type Original
476
477 SELECT CASE WHEN "PUBLIC" = 'YES' AND "SCHOOL TYPE" != 'RCCI'
478 THEN 'Public'
479 WHEN "PUBLIC" = 'NO' AND "SCHOOL TYPE" != 'RCCI'
480 THEN 'Nonpublic'
481 WHEN "PUBLIC" = 'YES' AND "SCHOOL TYPE" = 'RCCI'
482 THEN 'Public RCCI'
483 WHEN "PUBLIC" = 'NO' AND "SCHOOL TYPE" = 'RCCI'
484 THEN 'Nonpublic RCCI'
485 ELSE NULL
486 END as "School Type-Original"
487 INTO #tmp_test
488 FROM ##tmpBulk_file2
489
490 /*
491 SELECT * FROM ##tmpBulk_file2
492
493 SELECT "School Type-Original", COUNT(*)
494 FROM #tmp_test
495 GROUP BY "School Type-Original"
496 ORDER BY COUNT(*) DESC
497 */
498 --Breakfast Delivery Model from State Agency-Original
499
500 SELECT CONCAT('O=', TRADITIONAL MODEL,
501 'P=', MID MORNING MODEL,
502 'Q=', CLASSROOM MODEL,
503 'R=', REDUCED PRICE MODEL,
504 'S=', GRAB N GO MODEL,
505 'T=', FREE MODEL)
506 as "Breakfast Delivery Model from State Agency-Original"
```

Code