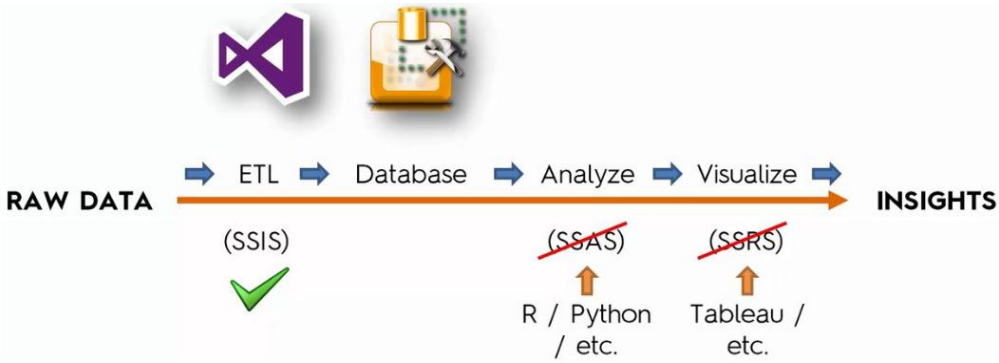
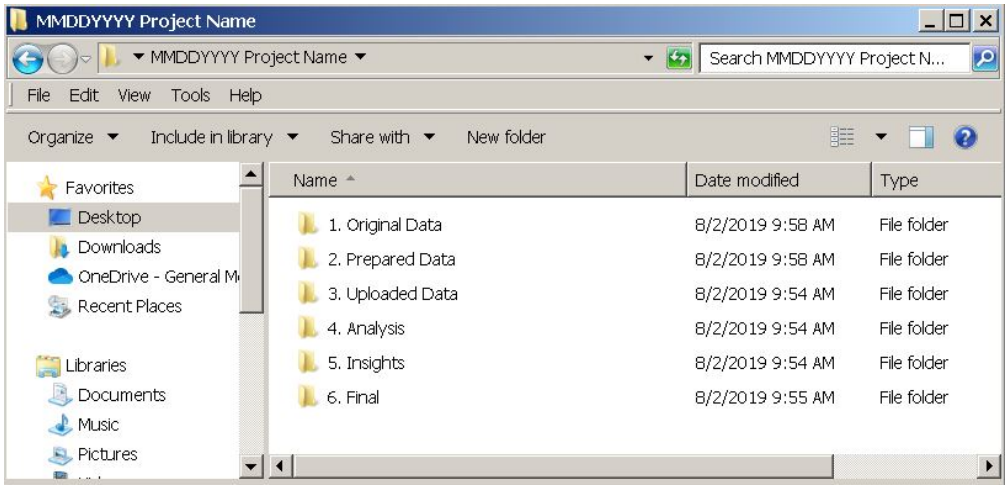
Data Preparation

Tools for Data Science



- **ETL(Extract Transform Load):** Use Microsoft Visual Studio (shell), which is only the business intelligence tools in Visual Studio.
  - SSDT-BI(SQL Server Data Tools - Business Intelligence): Building SSIS/SSAS/SSRS solutions
    - SSIS(SQL Server Integration Service): Use Microsoft Visual Studio Shell to Run SSDT-BI, then need SSIS part for ETL
    - SSAS(SQL Server Analysis Service): Analyze. But we use Python or R
    - SSRS(SQL Server Reporting Service): Visualize. But we use Tableau
- **Database:** Microsoft SQL Server
- **Analyze:** Python / R
- **Visualize:** Tableau

Suggested Data Science Project Folder Structure



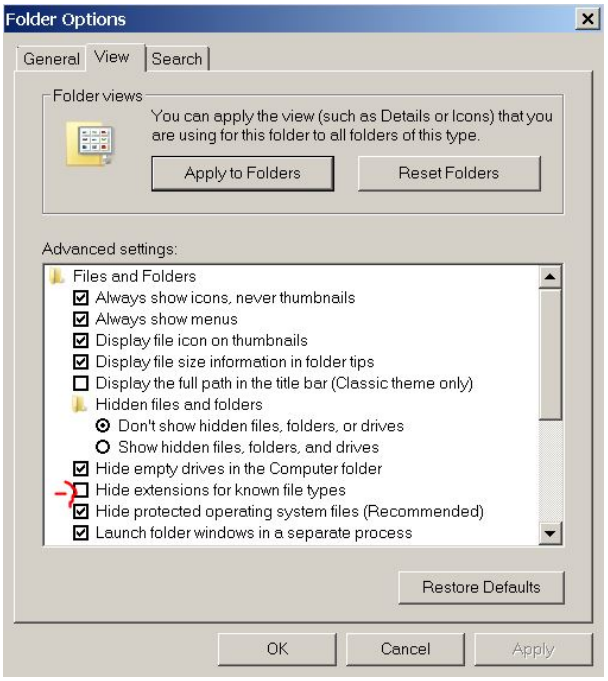
- **MMDDYYYY Project Name:** Use "Date" + "Project Name" for the Data Science Project
  - **Original Data:** Store all the raw data extracted from other systems, and no modification.
  - **Prepared Data:** Any modification we made to the raw data (cleaning the data).
  - **Uploaded Data:** Temporary Stop. Need subfolder with just date(MMDDYYYY). Once the data is ready to upload, then put in these subfolders.
  - **Analysis:** Any codes, scripts been created during the anaylsis.
  - **Insights:** Any preliminary(初步准备) results.
  - **Final:** Store the draft and final reports.

ETL (Phase 1: in EXCEL)

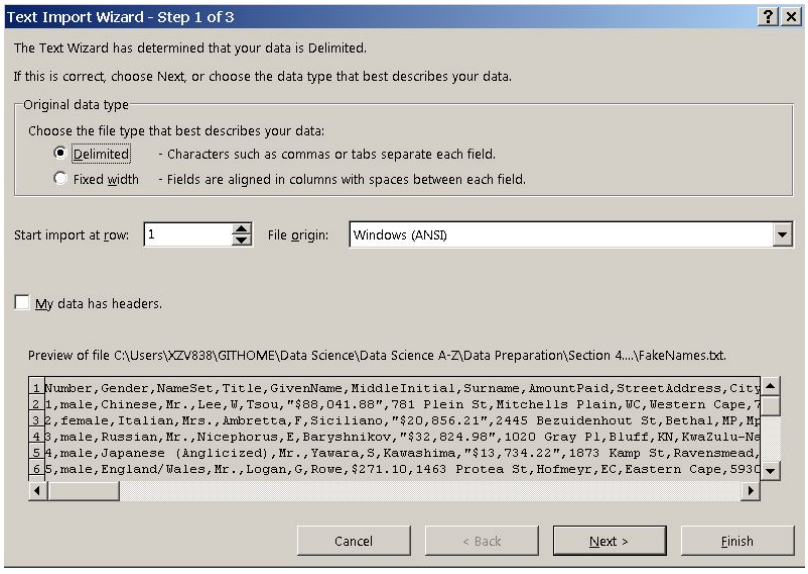
Do NOT open and save data by EXCEL directly. It will mess up the original data format (ie. Date and long int).

Deal with Large CSV

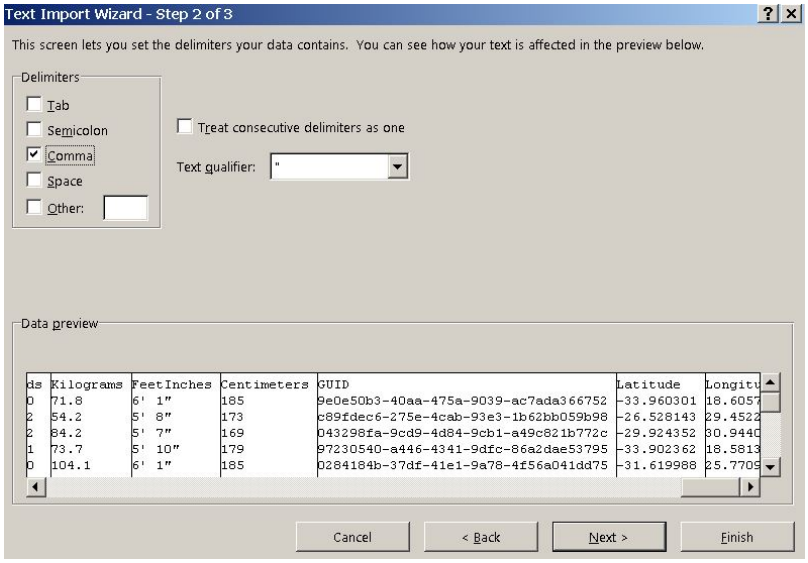
STEP 1: Rename the extension to be ".txt"
STEP 2: In the same directory, Go to "Tools" → "Folder options..." → "View" Tab → Uncheck box 'Hide extensions for known file type'



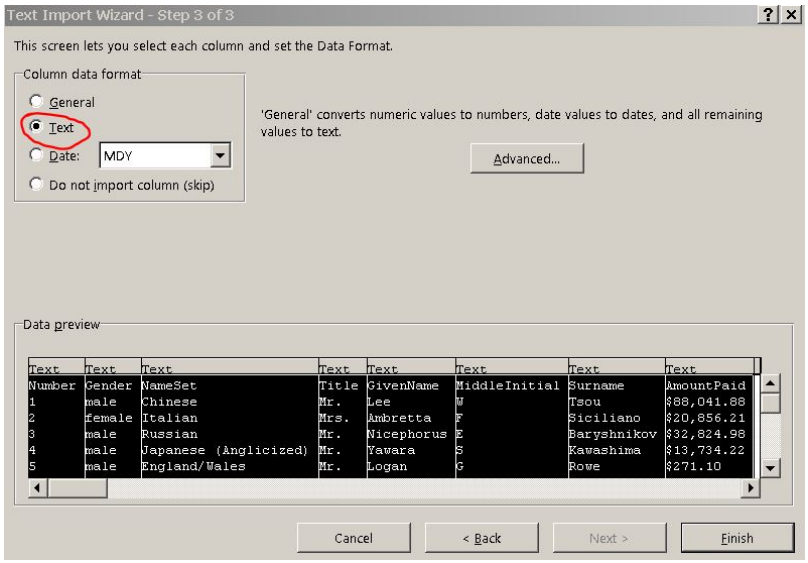
STEP 3: Use normal "File" → "Open" to open ".txt" data in EXCEL. Use "Text Import Wizard" to force EXCEL open the file correctly.
STEP 4.1: Make sure 'Delimited' has been selected



STEP 4.2: Change 'Delimiter': 'Comma'(only) and 'Text qualifier': ' " ', Check all the imported columns are correct. 根据导入的 .txt 文件的类型，可以选择不同的 Delimiter。如 'Tab'。。。

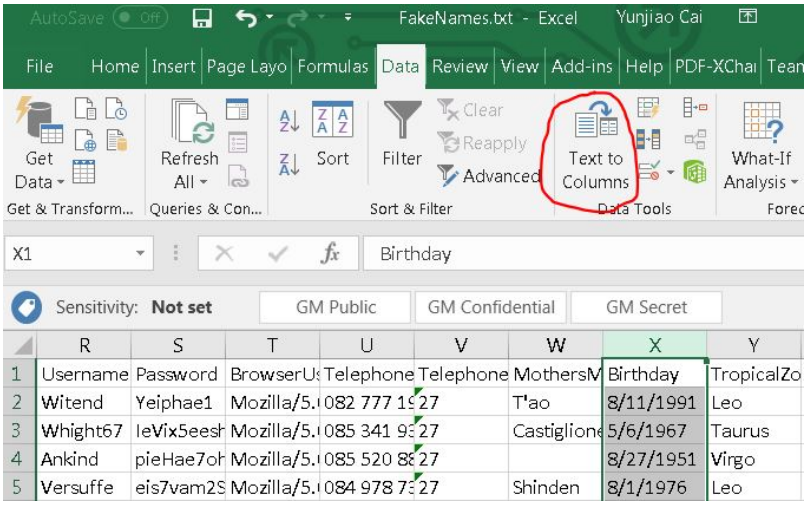


STEP 4.3: Select from the 1st column to the last column. Then check 'Text' in the 'Column data format' section.

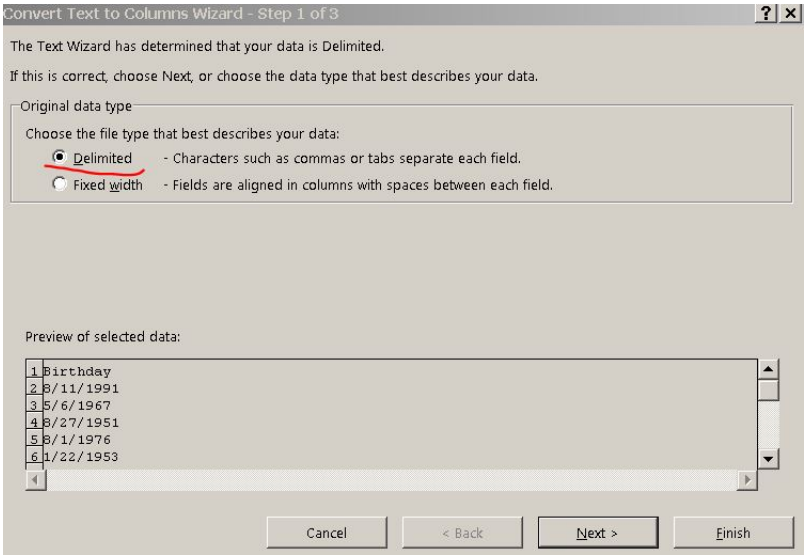


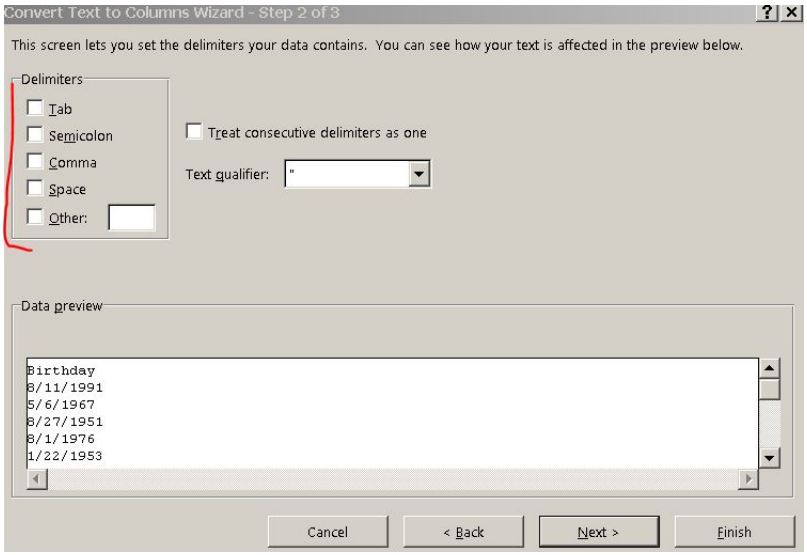
STEP 5: In general, check and fix up the "Date" in the data (convert txt to date)

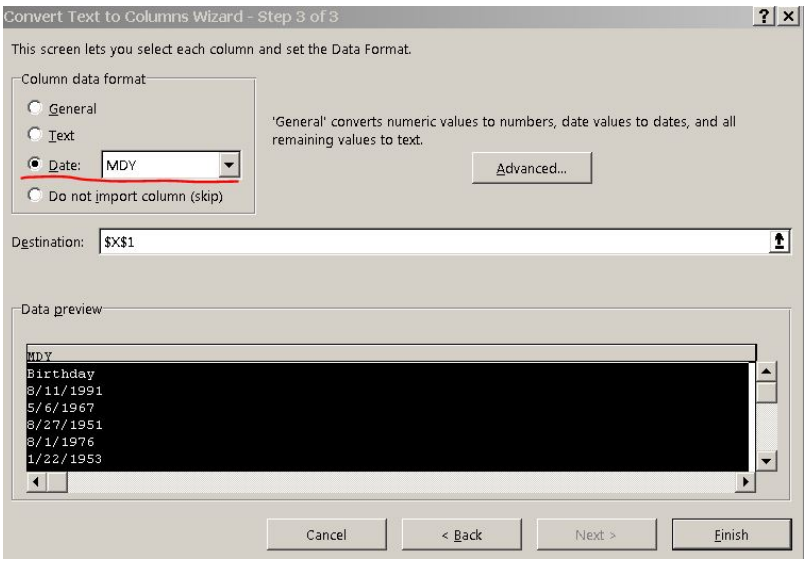
- Select whole "Date" column, then select "Data" → "Text to Columns"



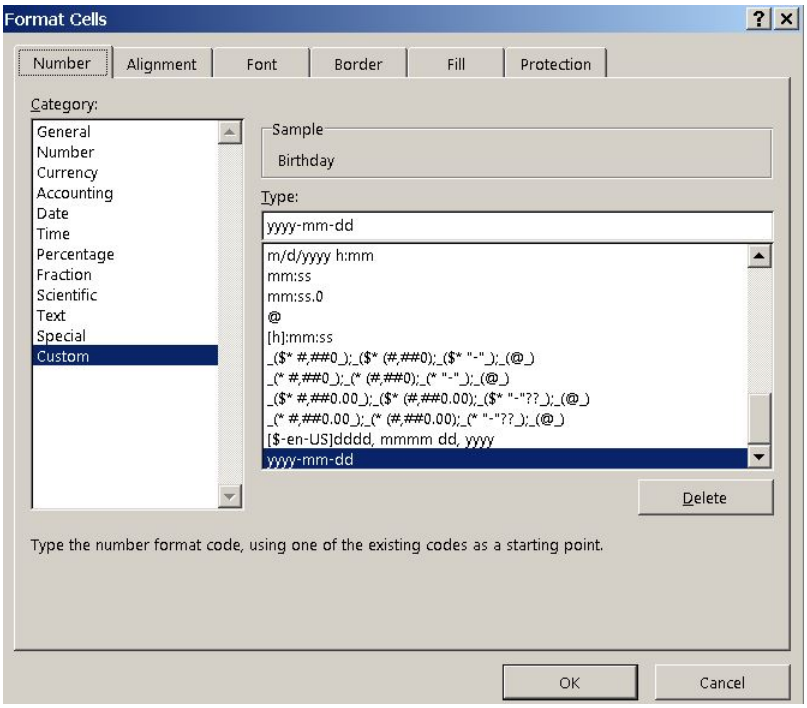
- Check 'Delimited', Uncheck any 'Delimiters' since only dealing with 1 column, Select 'Date' in 'Column data format' section to be "MDY" (Month, Day, Year)







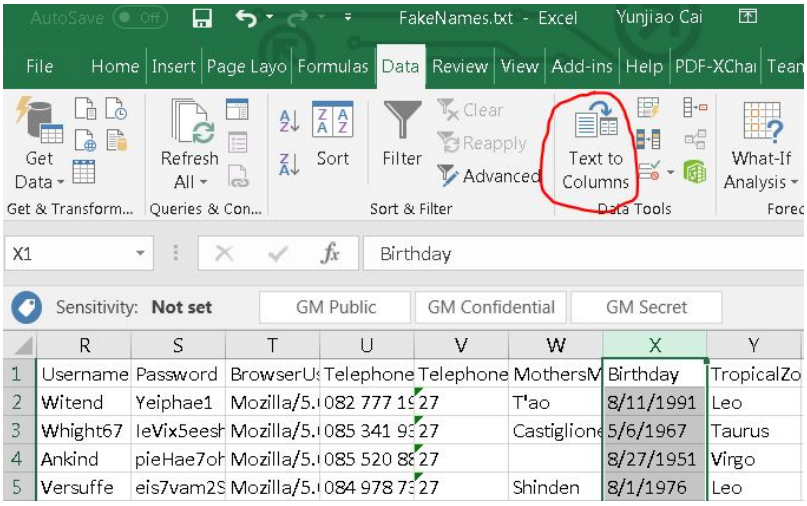
- Now we can change Date type by using EXCEL "Format Cells..." (we will use yyyy-mm-dd in the example)



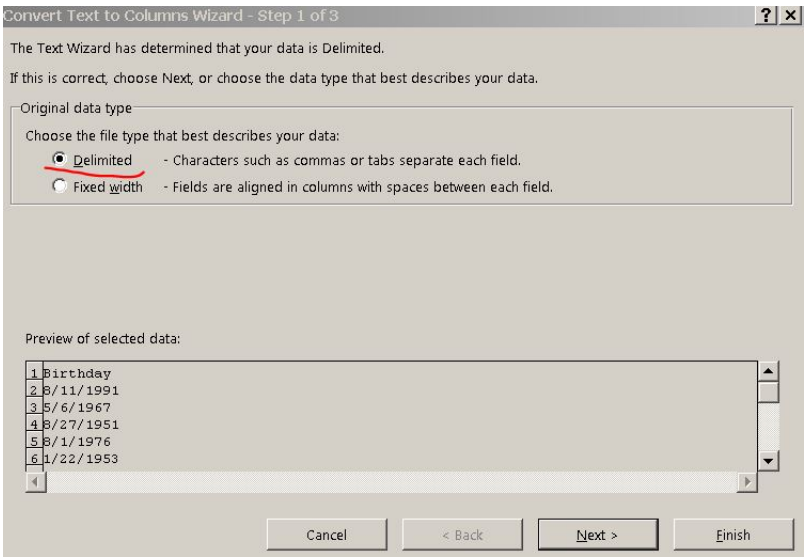
- Repeat the above steps for every column has "Date".

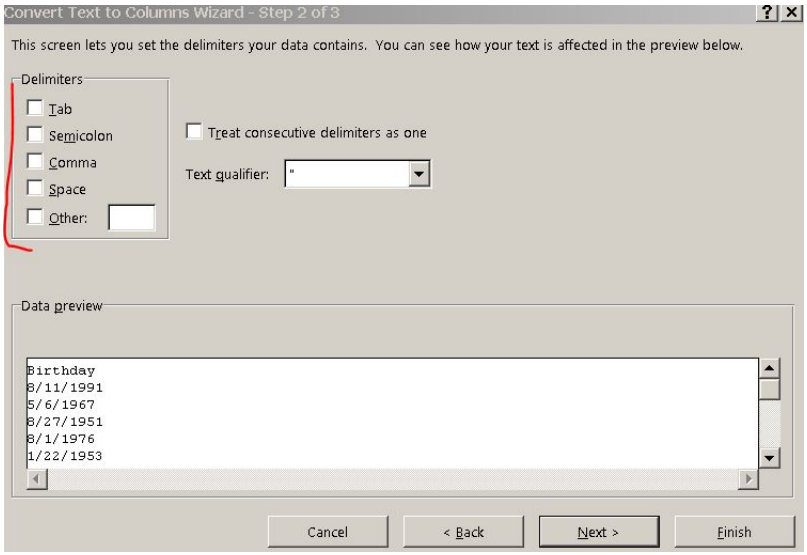
STEP 6: Same as STEP 5 to fix up the "Dollar Amounts" (convert txt to number) 此方法同样用于 Float 和 百分数%。

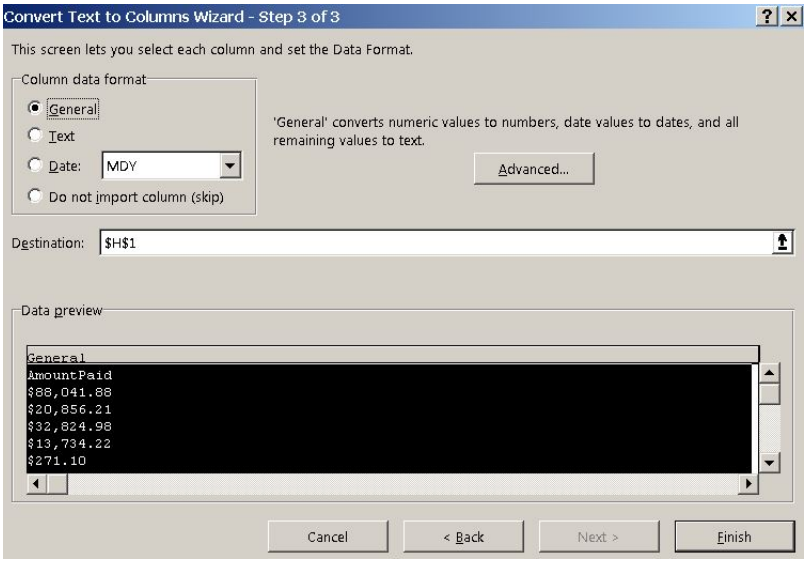
- Select whole "Dollar" column, then select "Data" → "Text to Columns"



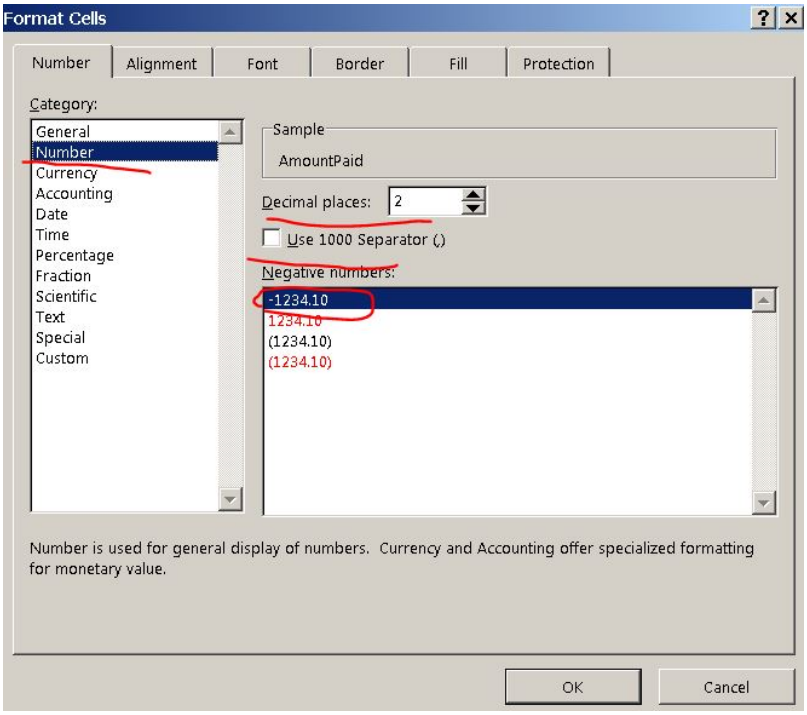
- Check 'Delimited', Uncheck any 'Delimiters' since only dealing with 1 column, Select 'General' in 'Column data format' section







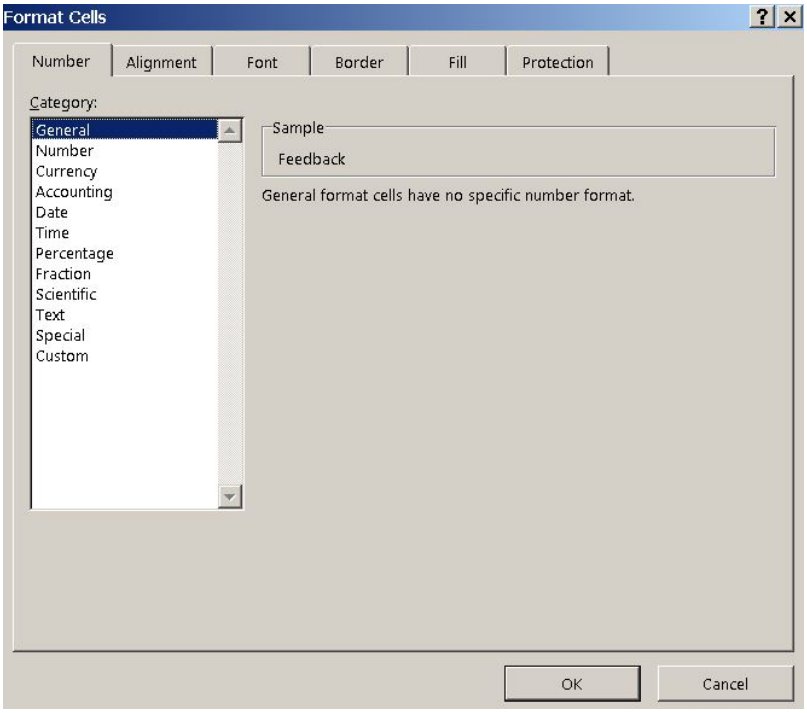
- Now we can change Date type by using EXCEL "Format Cells...". Choose 'Number', 'Decimal places', no 'Use 1000 Separtor()', 'Negative numbers'



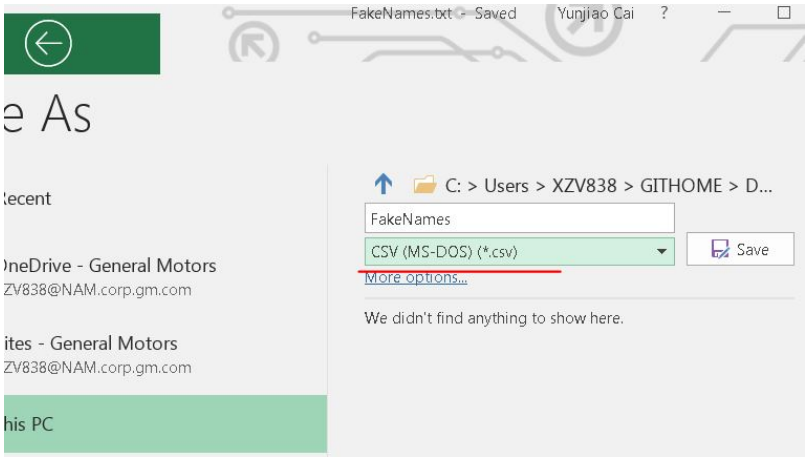
- Repeat the above steps for every column has "Dollar".

STEP 7: Fix the column which contain more than 256 characters

- Select whole column, right click "Format Cells..." → Select 'General' in "Category" section



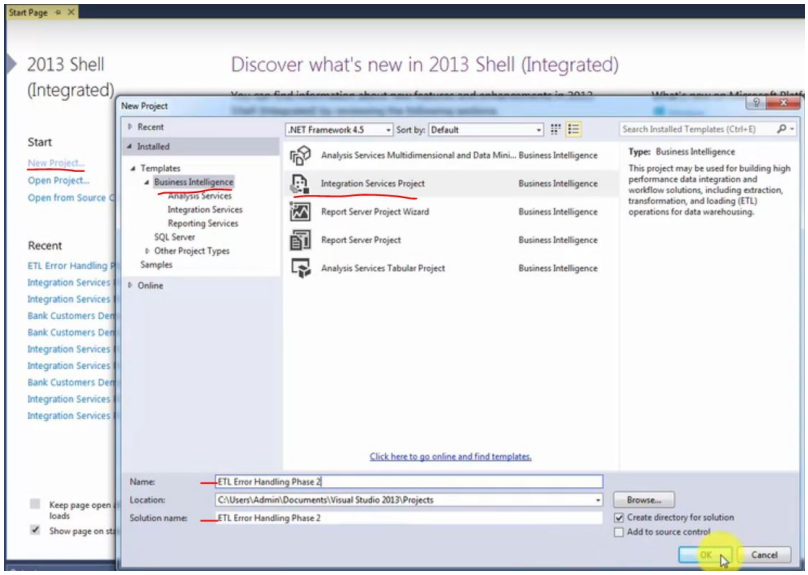
STEP 8: Now save data as ".csv" format



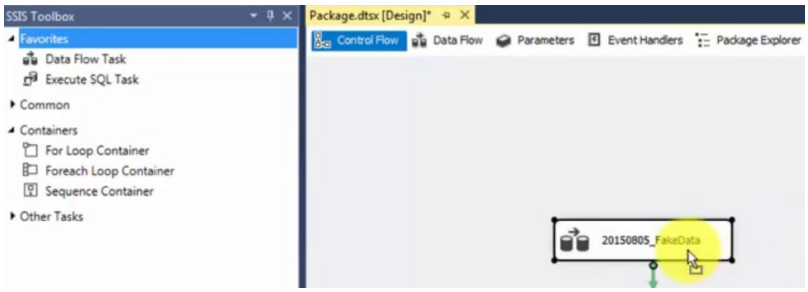
ETL (Phase 2: in SSIS)

How to upload a raw file where everything is in text into SQL

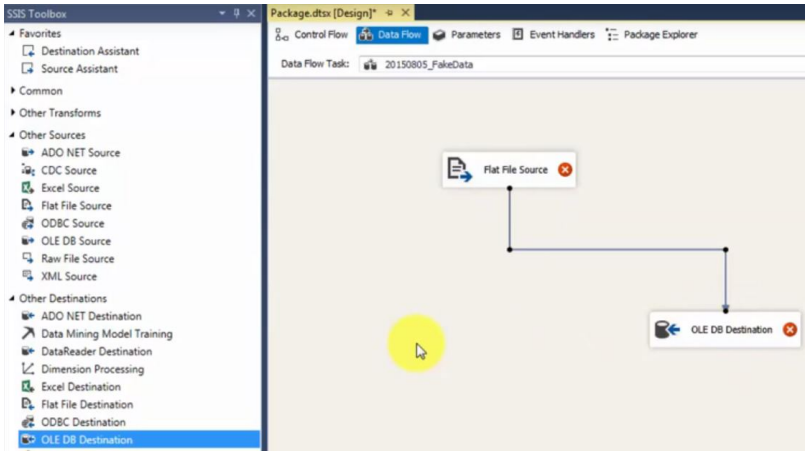
STEP 1: Open Microsoft Visual Studio → "New Project" → "Business Intelligence" → "Integration Services Project" → Enter "Name" and "Solution Name"



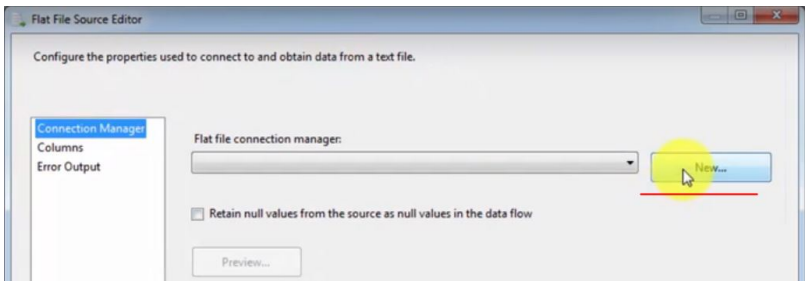
STEP 2: In the "Control Flow" section, drag 'Data Flow Task' from "SSIS Toolbox"/"Favorites". Then give it a name



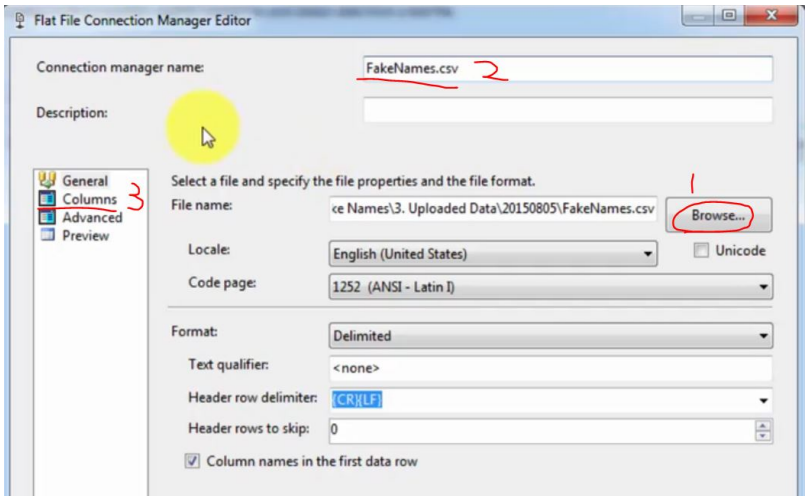
STEP 3: Double click dragged "Data Flow Task" block to open "Data Flow" section. Drag "Flat File Source" from "Other Sources" and "OLE DB Destination" from "Other Destinations". Then connect those blocks.



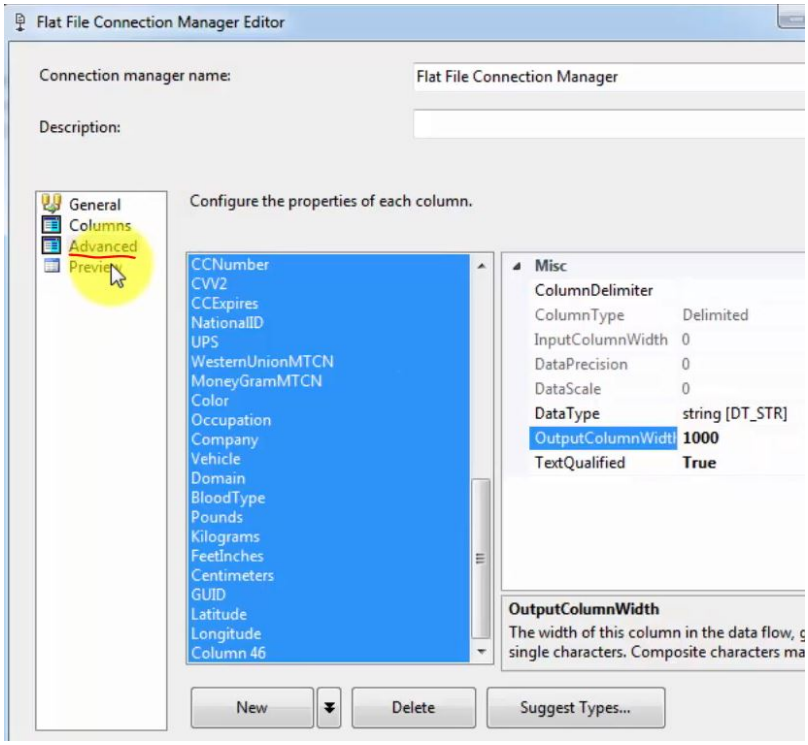
STEP 4.1: Double click "Flat File Source" to set up.



STEP 4.2: Use "Browse..." to select the data, and rename the 'Connection manager name:'
STEP 4.3: Go to the "Columns" section on the left side to check the table information (ALL the time). To find which column do errors start



STEP 4.4: Select "Advanced" → Select all the columns → 设置 'OutputColumnWidth' 为1000或2000 (For long comment cell)

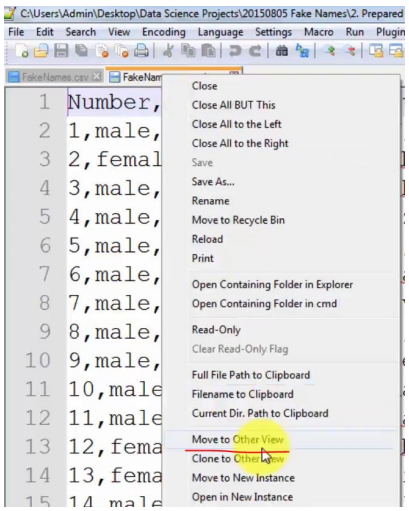


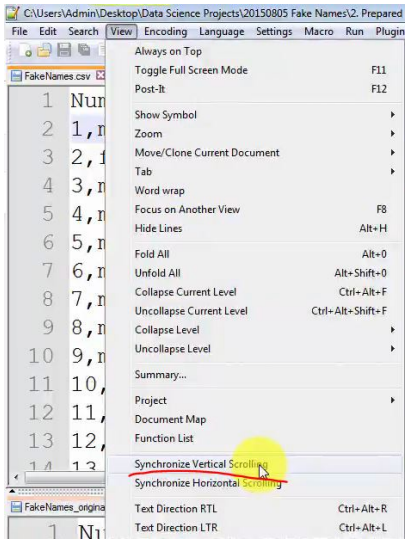
STEP 4.5: 用 Automating Error Handling in SSIS 来分流合格/不合格数据 (见 "Automating Error Handling in SSIS" 章节)

STEP 5: Open both prepared .csv and original .csv file on the same view site in Notepad++. Compare both files to find where does error start and what causes it incorrect.

(注：EXCEL会自动添加双引号"来弥补缺少的引号"，会造成问题)

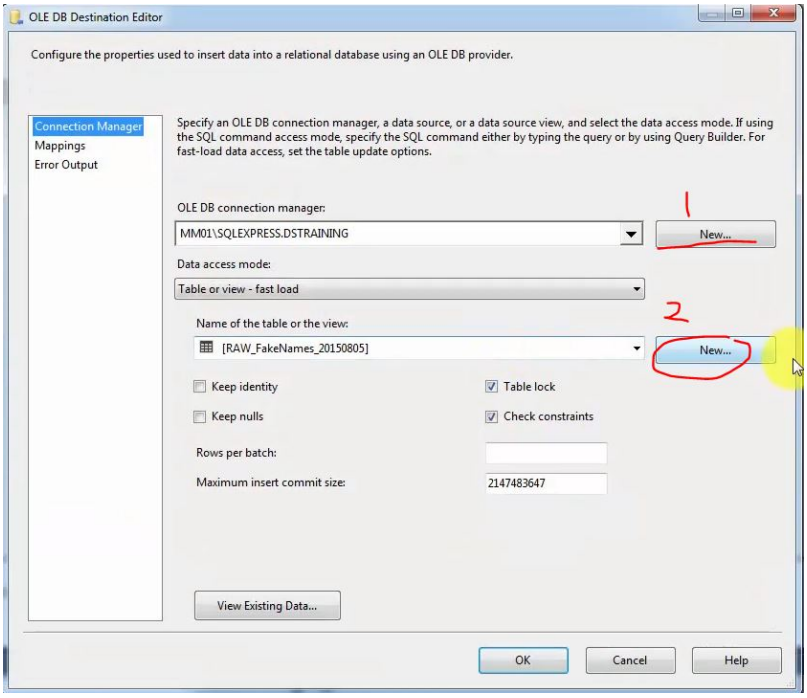
- Right click one of the data file → "Move to Other View" (To rotate the views to be either side-by-side or top-and-bottom simply right-click the divider line)
- Synchronize the scrolling: "View" → "Synchronize Vertical Scrolling"



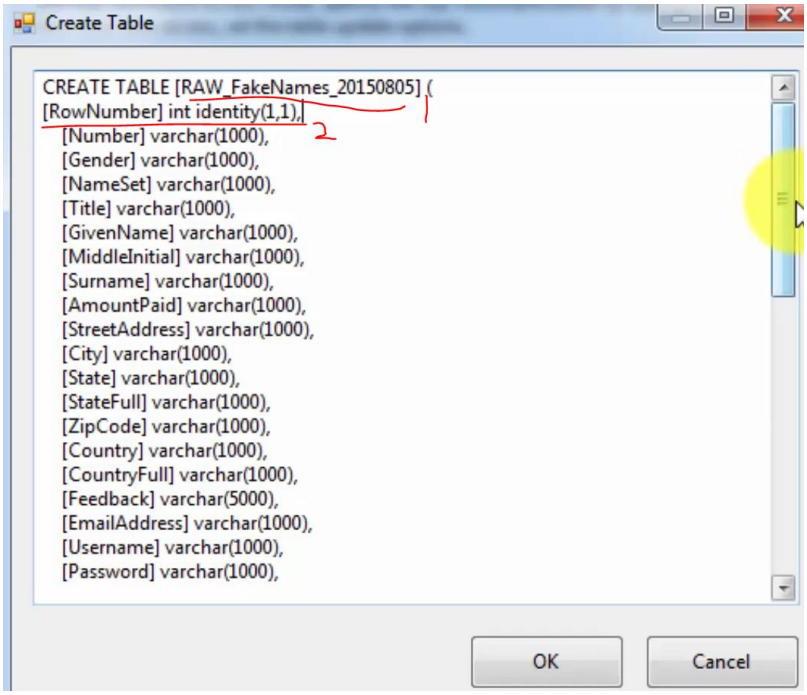


STEP 6: 见"常见问题"章节。如果 Data 结构应 EXCEL 自动排版导致问题的话，Modify/fix the prepared data in Notepad++。如果是信息缺失，则将信息缺失行消息反馈给相应部门，让他们处理。

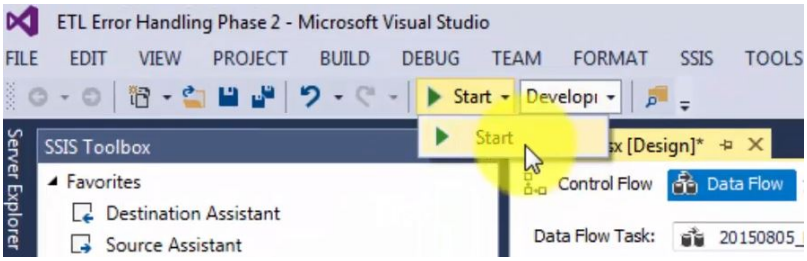
STEP 7: Double Click "OLE DB Destination" to CREATE TABLE in SQL server. Select the Database, then create new table with columns



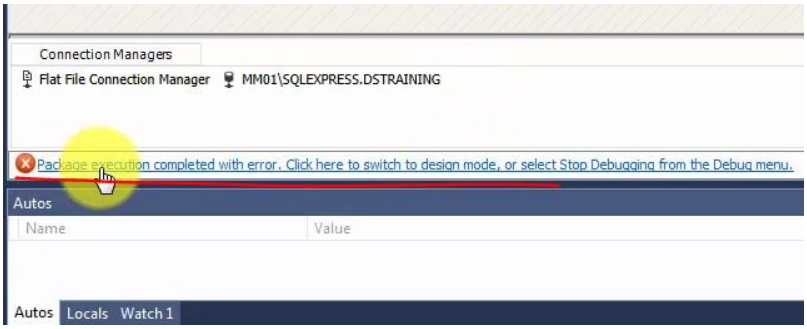
STEP 8: Rename the table name, and add one column (RowNumber) with identity(1,1) manually (index and improve data search speed)



STEP 7: 点击 "Start" 开始将 Data 导入 SQL Server.



STEP 8: 出现 error 的话，点击底部 link，再点击新页面上的 "Excution Results"。



STEP 9: 任何上传 Data 失败后，需要删除上传失败的 Database。回到 STEP 7 重新上传。

***SQL** 语句：*

```
DROP TABLE TableName # Delete Entire table
```
或者
```
TRUNCATE TABLE TableName # Delete the data in the table, but keep the header
```

STEP 10: 完成该 **block** 的 **Control Flow** 后，必须将这个 **block** 给 **Disable** 掉。不然下次执行时，这个模块的数据会再次上传。 执行 Control Flow 时，Control Flow 中所有模块都会同时上传数据。

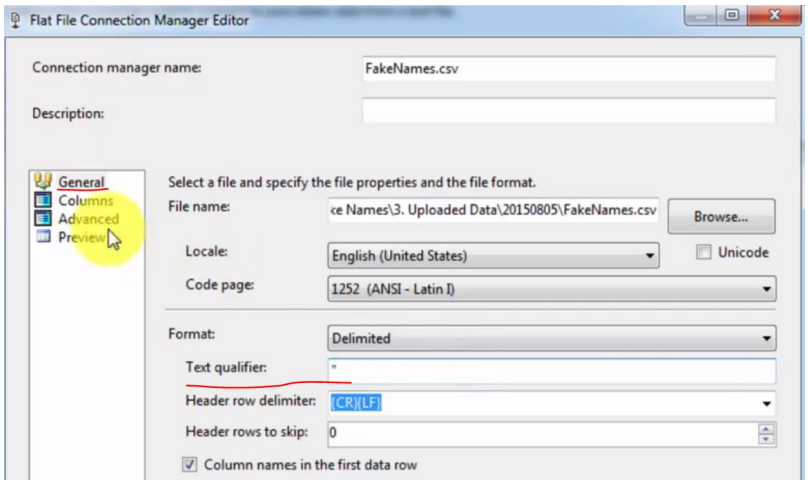
常见问题

如果 Data 结构应 EXCEL 自动排版导致问题的话，Modify/fix the prepared data in Notepad++。

Text Qualifier

"Text Qualifier": 让 SSIS 将 "Text Qualifier" 设定的符号(如 ")中，囊括的任何内容，视作一个整体(1个 cell)。 "Text Qualifier" 允许我们用设定的符号来分割内容。

比如, "I am ok, how are you". 如果 "Text Qualifier" 设置为空的话。将会产生两个 cell: "I am ok" 和 "how are you"。如果设置了 "Text Qualifier": " 。则整个 "I am ok, how are you" 会在同一个 cell 里。



Data 信息缺失

如果 Data 信息缺失，某些 Columns 没有数据，判断是否重要，并且反馈给相关部门。

Data Truncation

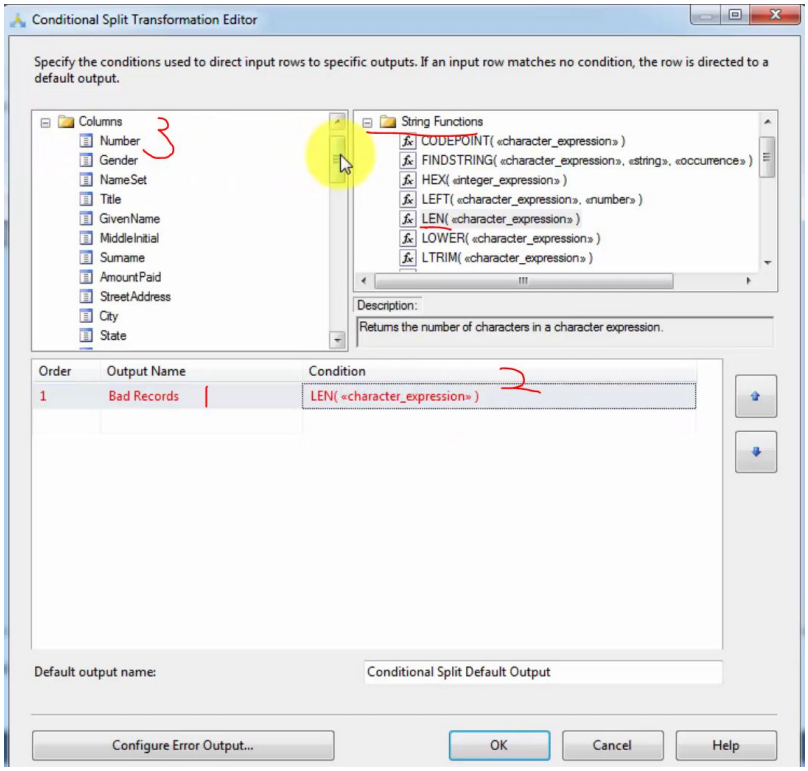
回到 "Flat File Source"，重复 STEP 4.4。增加相应 Truncation 行的 Output Column Width (如增加到5000)

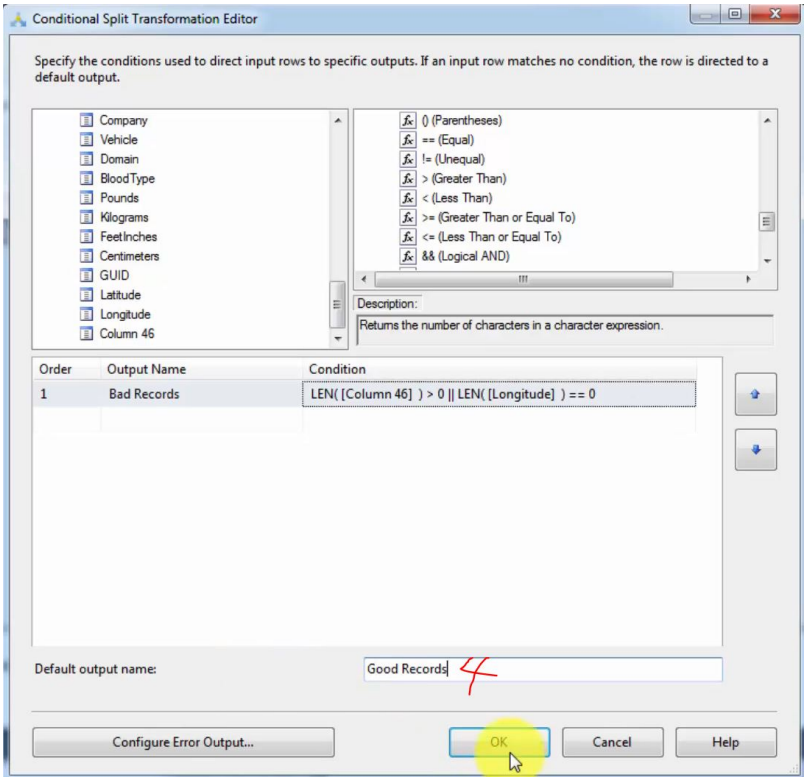
Automating Error Handling in SSIS: Conditional Split

通过使用 SSIS 中，"Conditional Split" block，设置对应过滤函数，来自动分流好/坏数据，并导入相应的 .txt 文件。

STEP 1: Delete/Truncate the created data (table) by using SQL commands in SQL Server
STEP 2: In SSIS, delete the connection line between "Flat File Source" and "OLE DB Destination"
STEP 3: In the "SSIS Toolbox / Common / Conditional Split". Use "Conditional Split" for transform. Then connect "Flat File Source" and "Conditional Split" blocks
STEP 4: Set up "Conditional Split" block. (1) Give "Output Name"; (2) Drag Functions from "String Functions" to "Condition"; (3) Use "Columns" to complete dragged function in the "Condition"; (4) Give the Default output name.

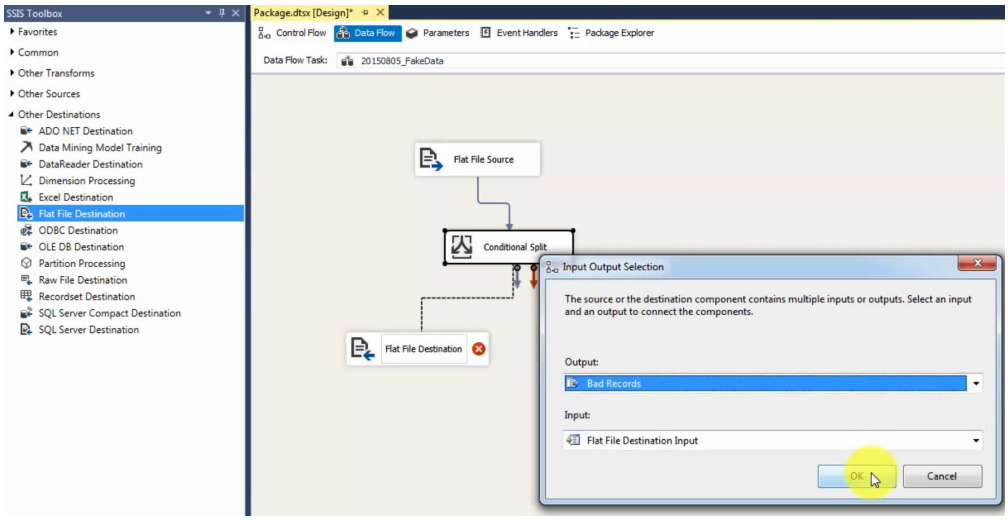
注：通过在 **Condition** 里设置含有 **Columns** 变量的 **Function**，来区分 **Data** 是否合格。将不合格的数据分流至 **"Output Name",** 合格数据分流至 **"Default output name"**。

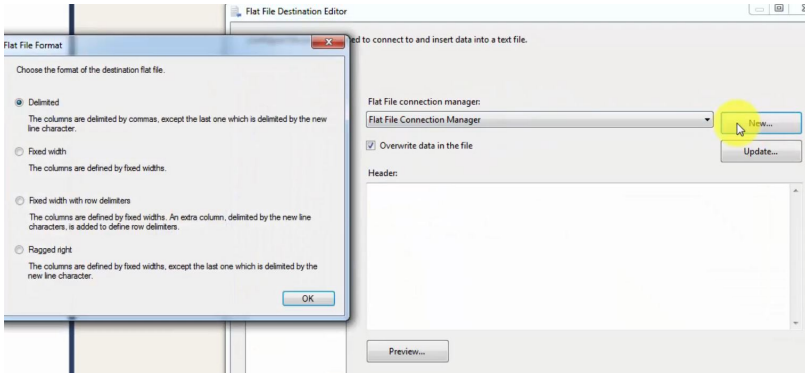


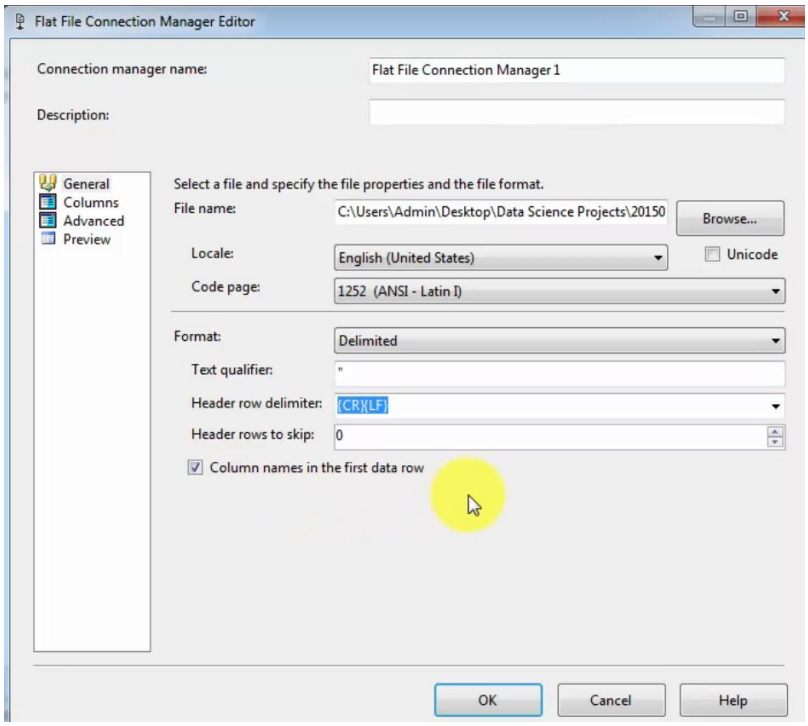


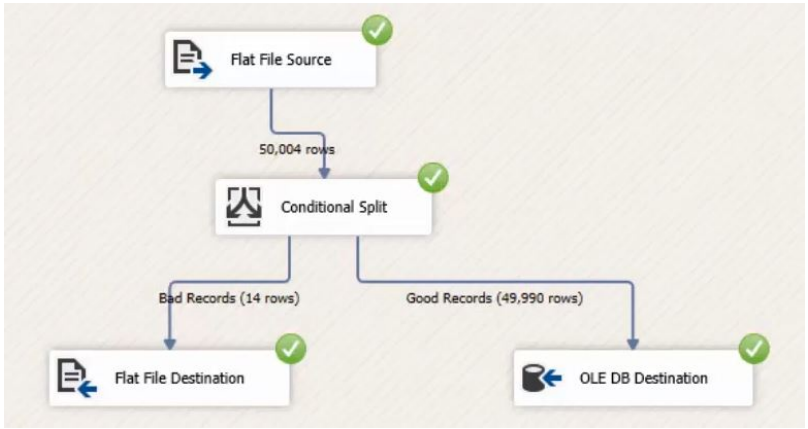
STEP 5: Set up "Flat File Destination" to store 不合格数据。(1) 当与 "Conditional Split" block 连线时，"Input Output Selection" window 自动弹出，选择对应的 Output。(2) 随后双击，在 "Flat File Connection Manager" 选择 "New...",选择 "Delimiter", 在弹出的新窗口，点击 "Browse..."。(3) 建议在 "Analysis" Folder 里新建文件夹和 .txt 文件 "Automatically Excluded Results/YYYYMMDD_ProjectName_xxxRecords.txt" 来储存分流后的数据。并选择该新建的 .txt 文件。(4) 设置 "Text qualifier" 为 "。(5) Check the box "Column names in the first data row"

。



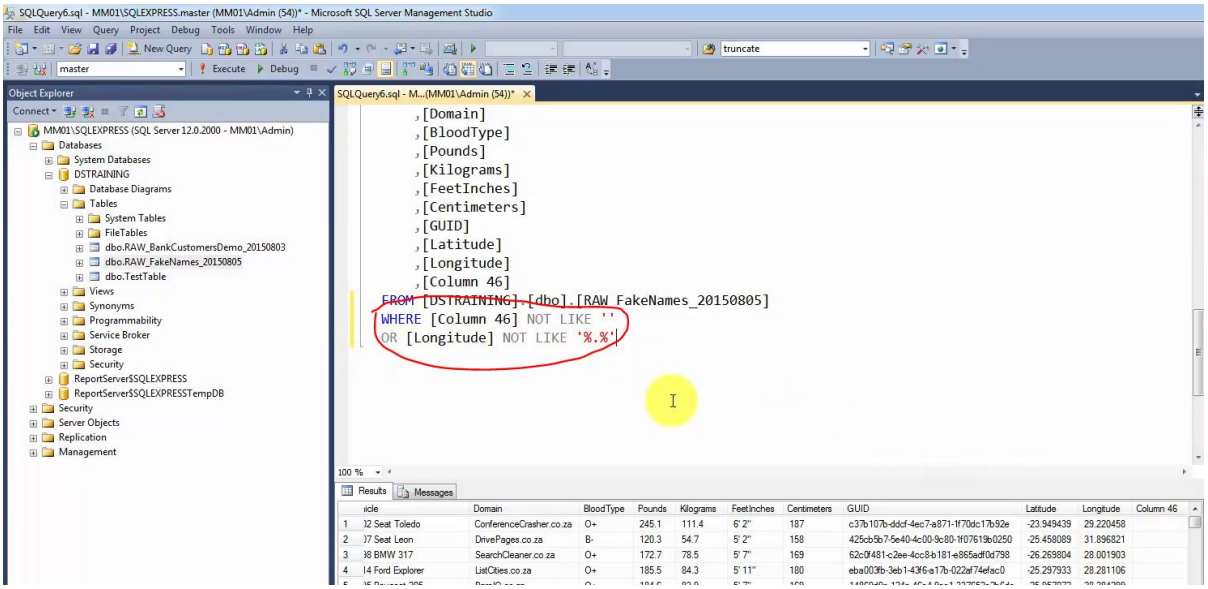

Flat File Destination Editor
Flat File Format
Choose the format of the destination flat file.
Delimited
The columns are delimited by commas, except the last one which is delimited by the new line character.
Fixed width
The columns are defined by fixed widths.
Fixed width with row delimiters
The columns are defined by fixed widths. An extra column, delimited by the new line characters, is added to define row delimiters.
Ragged right
The columns are defined by fixed widths, except the last one which is delimited by the new line character.
OK
Flat File connection manager:
Flat File Connection Manager
New...
Overwrite data in the file
Update...
Header:
Preview...


Flat File Connection Manager Editor
Connection manager name:
Flat File Connection Manager 1
Description:
General
Columns
Advanced
Preview
Select a file and specify the file properties and the file format.
File name:
C:\Users\Admin\Desktop\Data Science Projects\20150
Browse...
Locale:
English (United States)
Unicode
Code page:
1252 (ANSI - Latin I)
Format:
Delimited
Text qualifier:
"
Header row delimiter:
{CR}{LF}
Header rows to skip:
0
Column names in the first data row
OK
Cancel
Help


Flat File Source
50,004 rows
Conditional Split
Bad Records (14 rows)
Good Records (49,990 rows)
Flat File Destination
OLE DB Destination

Finding Anomalies in SQL

Open SQL Server and find the uploaded table in the correct DataBase. To add a filter to check where are the anomalies.



Only 2 types of rows will be output: (1) not empty in Column 46 (2) no period in the Longitude column. Since the source corruption always shifts the row to right or left. So this way helps to check the issue.

***SQL** 语句：*

```
SELECT *
FROM TableName
WHERE [last_column] NOT LIKE ''
OR [the_one_before_last_column] NOT LIKE '%.%'
```

SQL Programming for Data Science

SELECT *

选择 DataBase: 这样就不用每次在 table name 前，加上 DataBase path 前缀了。
USE DataBase_name
GO

从 table 中选择所有 columns:
SELECT *
FROM [table_name]

Using WHERE clause to filter data

SELECT [col_1], [col_2]...
FROM [table_name]
WHERE condition1 AND condition2 OR ...

注意：因为上传的 table 里面的值全是 'text'。所以要比较数字的话，需先 CONVERT(FLOAT, [col_x])
如：CONVERT(FLOAT,[Sales]) > 100

Use Regular Expressions in SQL

只能在 WHERE 子句中使用 LIKE
% : 顶替任意数量的字符。
_ : 只顶替 1 个字符。

如：选择所有 'IT' 开头的 'Order_ID'
SELECT *
FROM ListOfOrders
WHERE [Order_ID] LIKE 'IT%'

如：All customers whose 2nd letter in their name is 'e'
SELECT *
FROM ListOfOrders
WHERE [Customer_Name] LIKE '_e%'

Comments in SQL

/ *comments* / 或者 --

ORDER BY

通常不会在 SQL 中使用排序。因为没有必要在数据库中展示顺序，通常在应用中排序。
按照[col_name]来排序：DESC 降序，升序是默认的
SELECT *
FROM [TableName]

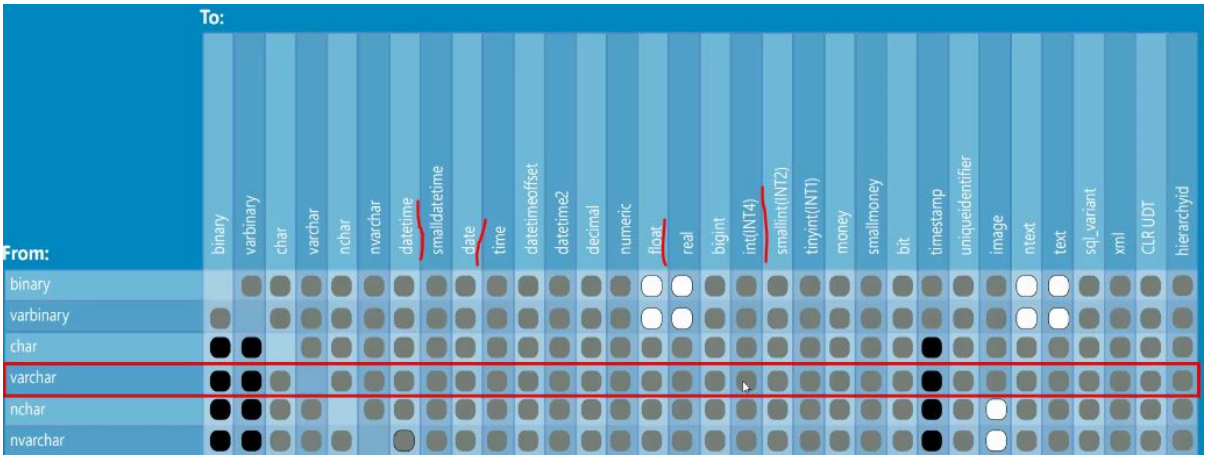
ORDER BY [col_name] DESC
注意：因为上传的 Table 是 text，所以按数字列来排序，将会出现问题。需要先 convert text table into a working table 后再排序。或者 CONVERT(FLOAT,[col_name]) 或者 CAST([col_name] as FLOAT)。

Data Types in SQL

VARCHAR(n): n 字节的可变长度
INT: $-2^31 \sim 2^31 - 1$整数
FLOAT: 浮点数，1040.53
DATE: 日期，2015-08-11

Implicit Data Conversion in SQL

因为导入SQL的表格数据是 text 类型 (varchar)，所以下表为 varchar 能隐式转换的类型(灰色为可以，黑色不可以)



Using CAST() vs CONVERT() 显式转换

CONVERT(FLOAT,[col_name]) 或者 CAST([col_name] as FLOAT)
转换日期 Dates，只能用 CONVERT()

Working with NULLs

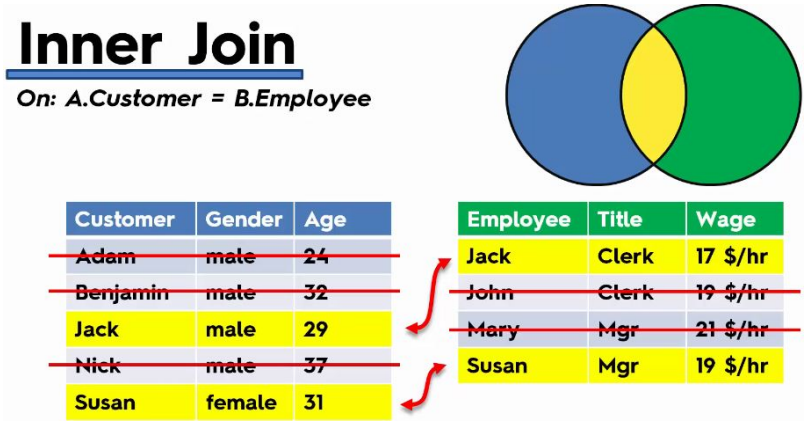
In SQL Server, alter the table value to be NULL: 当 Col_x 等于某值 y 时，替换 y 为 NULL
UPDATE [Table_Name]
SET [Col_x] = NULL
WHERE [Col_x] = value

找出哪些 value 是(不是) NULL:
SELECT [col_1],[col_2],[col_3]...
FROM [TableName]
WHERE [col_x] IS (NOT) NULL

Types of JOINs

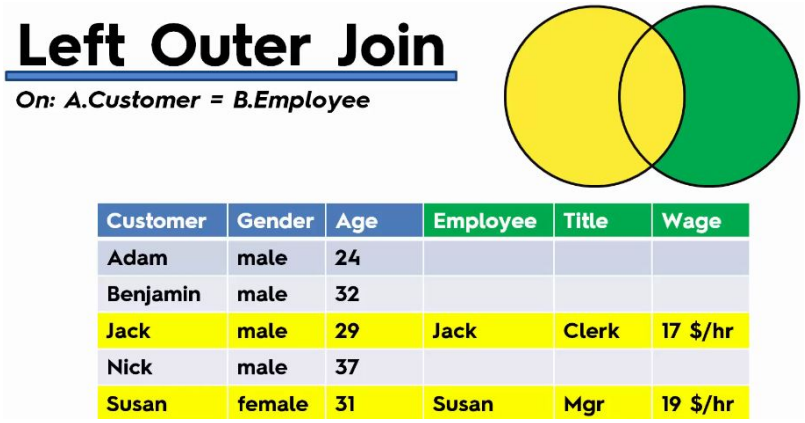
INNER JOIN

通过表格中指定 column 的相同元素，剔除不同，进行合并。



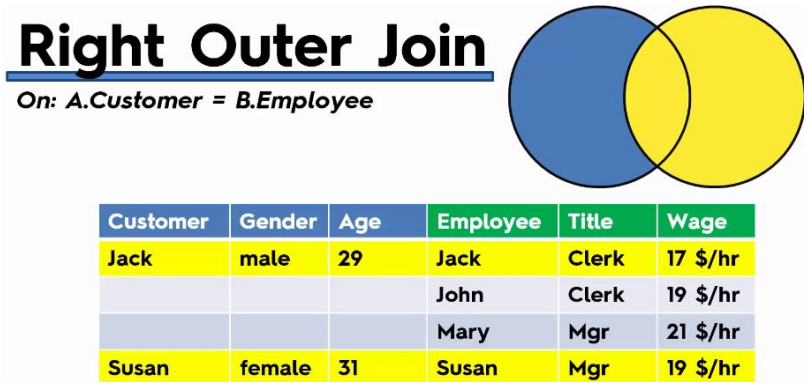
LEFT (OUTER) JOIN

左侧的表格为主表，左侧表中指定的合并 column 为主列，合并后，左侧表内容都在，右侧表指定 column 中与左表不同的行，全部删除。合并后，左表缺失数据为 NULL。



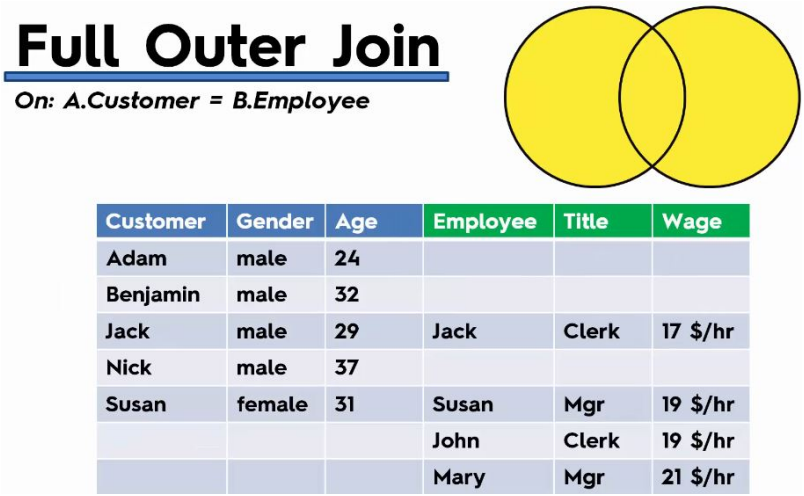
RIGHT (OUTER) JOIN

和 LEFT JOIN 相似。右表为主表，保留右表指定列的内容，合并后，删除左表不同行。



FULL (OUTER) JOIN

合并表格内容全部保留，指定列中相同元素对应合并。其他不同处，缺失元素为 NULL。



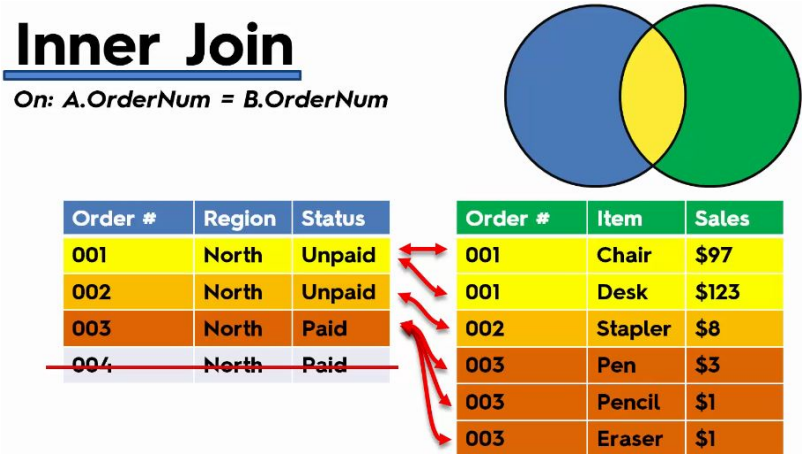
Customer	Gender	Age	Employee	Title	Wage
Adam	male	24			
Benjamin	male	32			
Jack	male	29	Jack	Clerk	17 $/hr
Nick	male	37			
Susan	female	31	Susan	Mgr	19 $/hr
			John	Clerk	19 $/hr
			Mary	Mgr	21 $/hr

最常用的是 INNER JOIN 和 LEFT JOIN。

Duplicates in JOINs

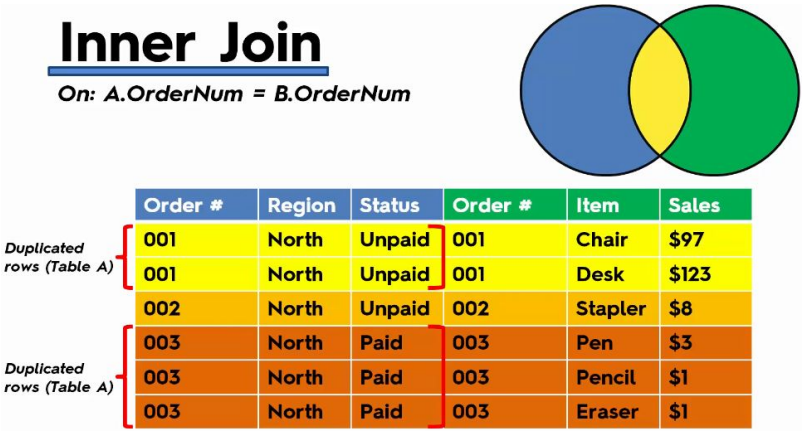
INNER JOIN

比对指定 column 中的相同元素，如果表中相同指定列的行数不相等，则复制行直至与对应表中行数相同。



Order #	Region	Status
001	North	Unpaid
002	North	Unpaid
003	North	Paid
~~004~~	~~North~~	~~Paid~~

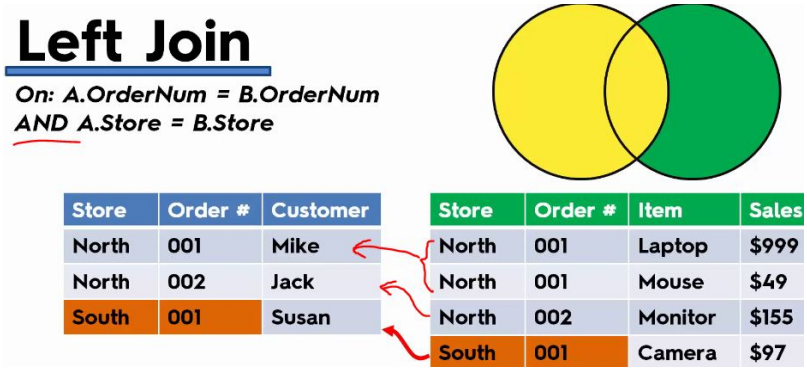
Order #	Item	Sales
001	Chair	$97
001	Desk	$123
002	Stapler	$8
003	Pen	$3
003	Pencil	$1
003	Eraser	$1

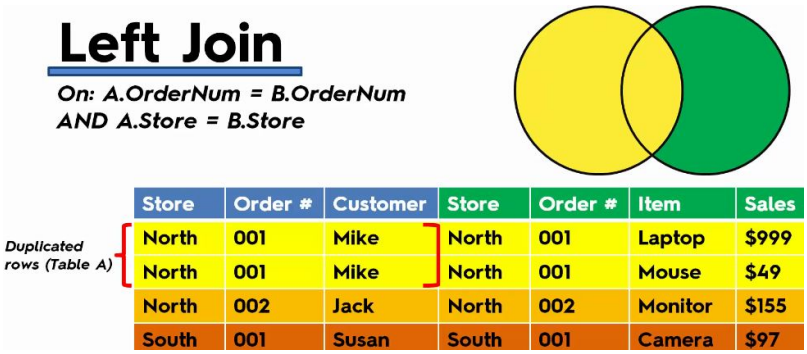


	Order #	Region	Status	Order #	Item	Sales
Duplicated rows (Table A)	001	North	Unpaid	001	Chair	$97
	001	North	Unpaid	001	Desk	$123
	002	North	Unpaid	002	Stapler	$8
Duplicated rows (Table A)	003	North	Paid	003	Pen	$3
	003	North	Paid	003	Pencil	$1
	003	North	Paid	003	Eraser	$1

JOIN on Multiple Fields

合并 Key 由多列组成。如下表中由 "Store" 和 "Order #" 组成合并 Key。单由 "Order #" 作为合并 Key，将有歧义。
下例中，想找出某一单和某个客户总共的消费





Practicing JOINs

```
SELECT *
FROM [TableA] as A
LEFT JOIN [TableB] as B
ON A.[Col_x] = B.[Col_y]
```

ETL (Phase 3: in SQL)

- RAW Table: Formatted as Text. We need format it properly (numbers with decimal to be float, dates to be dates...)
- Working Table: We will use "Stored Procedure"(script) in SQL to build Working Table from RAW Table. Normally, ETL is end once we got the Working Table, we can connect the analytical tools (Tableau, Python...) to the server and access the Working Table from there.
- Derived Table: 使用 VIEW 从 Working Tables 中生成 the Derived Table, and then analysis it.

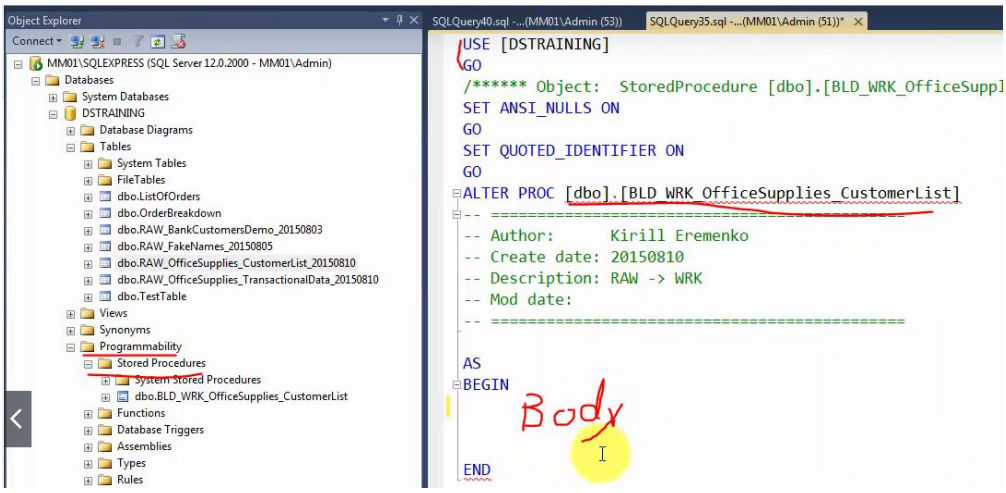
Stored Procedures

Advantages of using Procs:

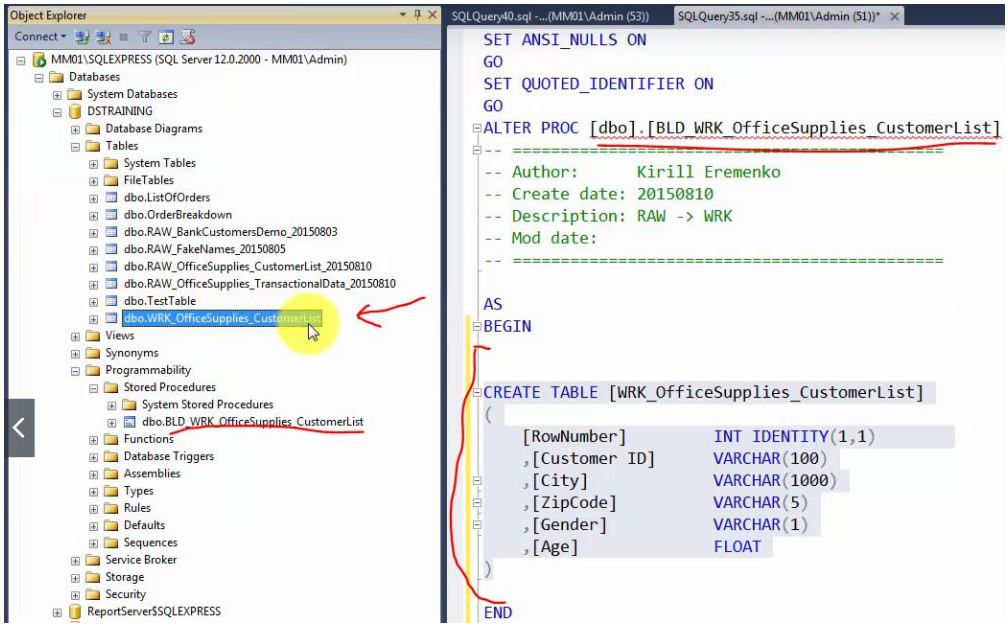
- Save your script in SQL Server MS
- Easily modify your BLD process (will see later in this section)
- Auditability, Reliability, Repeatability

Type the code between BEGIN to END and it's the only way to save the SQL code.
不推荐用 save 键，运行整个 Proc，使用 "!Execute" 键 (只运行 PROC body BEGIN...END 不会 save)，就会自动 save。

STEP 1: 按下图修改基本格式和信息(**注意**：第一次新建 PROC 时，在 Script 中使用 "CREAT" PROC, 一旦点击 "!Execute" 运行后，需要把 Script 中的 "CREATE" PROC 改为 "ALTER" PROC。防止重复生成相同 PROC。)



STEP 2: 基于原 RAW Table, 生成(CREATE) Working Table。在 CREATE TABLE 时，设置 Columns 的 Data Type。

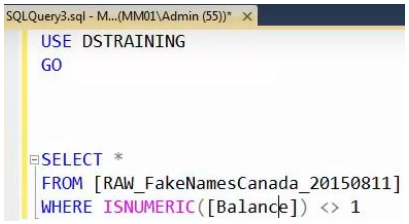


STEP 3.1: 为防止之前已经生成过 Table，避免重复生产相同 Table。在 ***CREATE TABLE*** 前使用 ***DROP TABLE***
STEP 3.2: 为防止之前已经在 Table 中输入过数据，建议先使用 ***TRUNCATE TABLE*** 用 ***INSERT INTO...SELECT*** 从 RAW Table 中复制插入新的 Working Table。SQL会通过隐式转换，自动转换数据类型。

```
AS
BEGIN

-- ============================================
-- DROP TABLE
-- ============================================
IF OBJECT_ID('WRK_OfficeSupplies_CustomerList') IS NOT NULL
DROP TABLE [WRK_OfficeSupplies_CustomerList]

-- ============================================
-- CREATE TABLE
-- ============================================
CREATE TABLE [WRK_OfficeSupplies_CustomerList]
(
    [RowNumber]         INT IDENTITY(1,1)
    ,[Customer ID]      VARCHAR(100)
    ,[City]             VARCHAR(1000)
    ,[ZipCode]          VARCHAR(5)
    ,[Gender]           VARCHAR(1)
    ,[Age]              FLOAT
)

-- ============================================
-- TRUNCATE TABLE
-- ============================================
TRUNCATE TABLE [WRK_OfficeSupplies_CustomerList]

-- ============================================
-- INSERT INTO
-- ============================================
INSERT INTO [WRK_OfficeSupplies_CustomerList]
(
    [Customer ID]
    ,[City]
    ,[ZipCode]
    ,[Gender]
    ,[Age]
)
SELECT
      [Customer ID]
      ,[City]
      ,[ZipCode]
      ,[Gender]
      ,[Age]
FROM [RAW_OfficeSupplies_CustomerList_20150810]
--(43 row(s) affected)



END
```

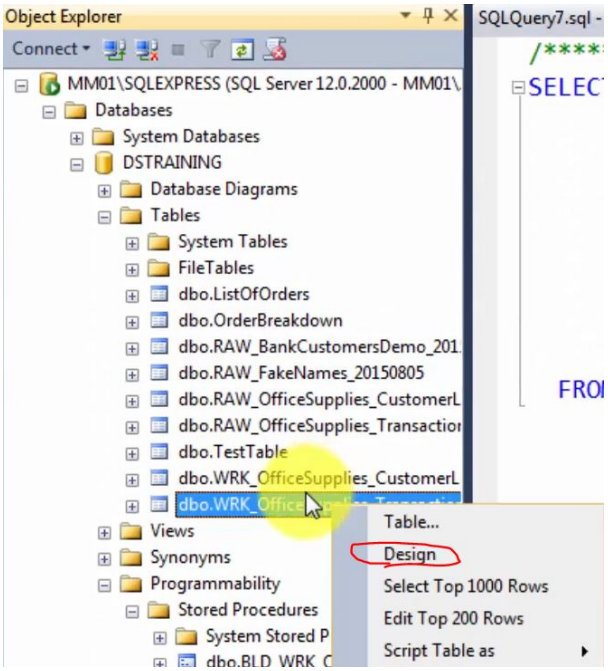
STEP 4: 如果复制转换时，出现错误。可以 comment 掉可以某些 column，逐步排除，最终发现哪个 column 导致问题。同时可以新开一个 script， 使用各种 function 来识别出问题所在。

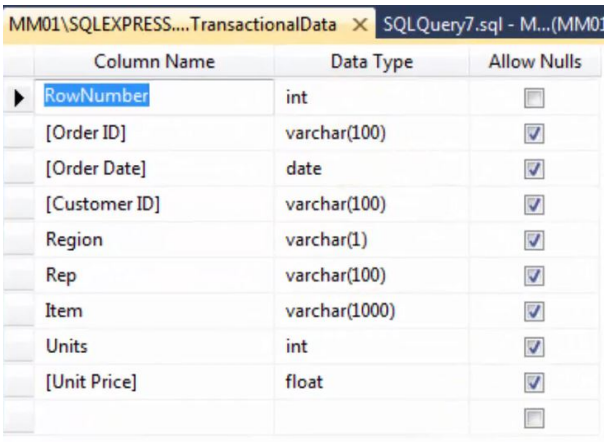
- Error converting data type varchar to float. 可以用 ***ISNUMERIC()*** 来查看哪些应该是 float 的 column 出了问题。如下哪些[Balance]的行不是 numeric (ISNUMERIC == False)



- String or binary data would be truncated. 可以在 ***CREATE TABLE*** 中逐个增大最小的 varchar 值后，运行程序，逐个排除嫌疑，一旦确定哪列后，可以用 LEN([Col]) > n 来收集不合格数据。
- Conversion failed when converting date and/or time from character string. 用 ***ISDATE([Col])*** 来检查 Date 列。

快速查看并设置 Table 中的 NULL 值: 右键点击相应 Table，选择 "Design"。

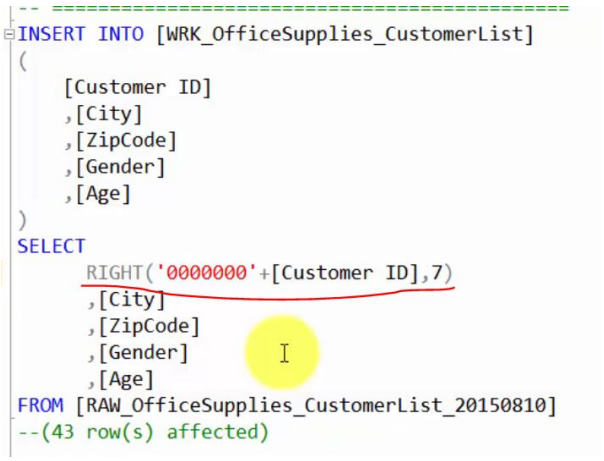




Lost Leading Zeros Issue

Normally, Lost leading Zeros issue occurs in the RAW table. 通过修改 PROC 来纠正。

原理：在需要修改的 column 前，添加需要的'0' (string/varchar/text)，生成'00000...xyz' (string/varchar/text)。然后再用 RIGHT('0000...xyz',从右数起需保留的位数)。如下图中，从右数起，保留7位。

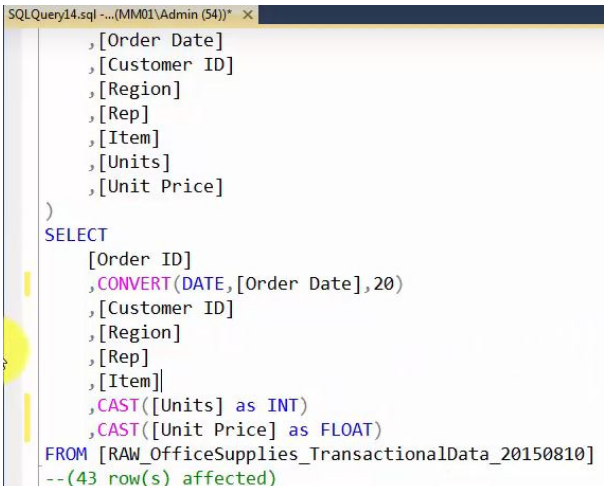


```
-- ============================================
INSERT INTO [WRK_OfficeSupplies_CustomerList]
(
    [Customer ID]
    ,[City]
    ,[ZipCode]
    ,[Gender]
    ,[Age]
)
SELECT
      RIGHT('0000000'+[Customer ID],7)
      ,[City]
      ,[ZipCode]
      ,[Gender]
      ,[Age]
FROM [RAW_OfficeSupplies_CustomerList_20150810]
--(43 row(s) affected)
```

显式转换

虽然可以在复制表格时，使用SQL的隐式转换，但是也可以用显式转换，强制转换成想要(但必须是允许转换)的类型。

如下图中，'20' 为 DATE 的表示种类 yyyy-mm-dd (日期转换一般用 ***CONVERT()***)。根据 RAW Table 中的日期，只能转换成特定种类。同理，'Unit Price' 不能转成 INT。



```
SQLQuery14.sql -...(MM01\Admin (54))*
      ,[Order Date]
      ,[Customer ID]
      ,[Region]
      ,[Rep]
      ,[Item]
      ,[Units]
      ,[Unit Price]
)
SELECT
      [Order ID]
      ,CONVERT(DATE,[Order Date],20)
      ,[Customer ID]
      ,[Region]
      ,[Rep]
      ,[Item]
      ,CAST([Units] as INT)
      ,CAST([Unit Price] as FLOAT)
FROM [RAW_OfficeSupplies_TransactionalData_20150810]
--(43 row(s) affected)
```

PROC Template (SQL Server)

In [ ]:

```
USE <DataBase Name>
GO

SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

# Could have Schema, in SQL Server, the schema is 'dbo'
CREATE PROC [dbo].[__tmpl__BLD_WRK_TableName]

# ================================================
# Author:
# Create date:
# Description: RAW -> WRK
# Mod date:
# ================================================

AS
BEGIN

# DROP TABLE
IF OBJECT_ID('WRK_TableName') IS NOT NULL
DROP TABLE [WRK_TableName]

# CREATE TABLE
CREATE TABLE [WRK_TableName]
(
    [RowNumber]    INT IDENTITY(1,1),
    [AAA]          VARCHAR(100),
    [BBB]          VARCHAR(1000),
    [DDD]          DATE,
    [EEE]          INT,
    [FFF]          FLOAT
)

# TRUNCATE TABLE
TRUNCATE TABLE [WRK_TableName]

# INSERT INTO
INSERT INTO [WRK_TableName]
(
    [AAA],
    [BBB],
    [DDD],
    [EEE],
    [FFF]
)
SELECT
    [AAA],
    [BBB],
    [DDD],
    [EEE],
    [FFF]
FROM [RAW_TableName_YYYYMMDD]
```

```
/* # Quick Check
SELECT *
FROM [WRK_TableName]
*/
```

使用 **VIEW** 生成 **Derived Table**

推荐使用 ***CREATE VIEW*** 来生成可以保存的 Script。根据要求，使用 ***SELECT*** 语句来建立相应的虚拟表，供 Analysis Tool 使用。
下图中，通过合并不同表格，将需要的列从各自表格中，汇总到一张表。

```
-- =============================================
-- Author:
-- Create date:
-- Description: WRK -> DRV
--              COMBINING TABLES:   [WRK_OfficeSupplies_CustomerList]
--                                  [WRK_OfficeSupplies_TransactionalData]
--              TO UNDERSTAND REVENUE SPLIT BY GENDER
-- Mod date:
-- =============================================

AS
BEGIN


SELECT
    A.[Customer ID]
    ,A.Gender
    ,B.[Units] * B.[Unit Price] AS Revenue
FROM [dbo].[WRK_OfficeSupplies_CustomerList] A
LEFT JOIN [dbo].[WRK_OfficeSupplies_TransactionalData] B
ON A.[Customer ID] = B.[Customer ID]
```