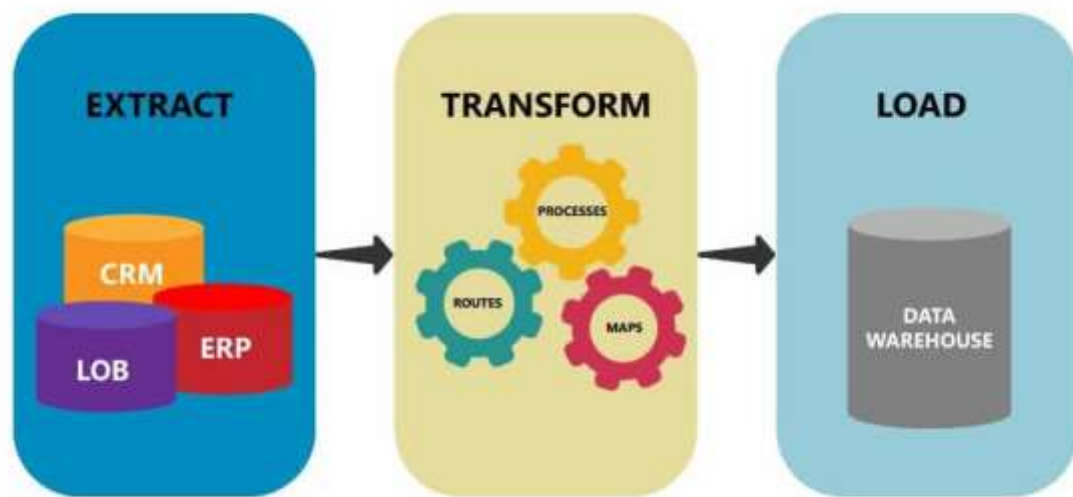


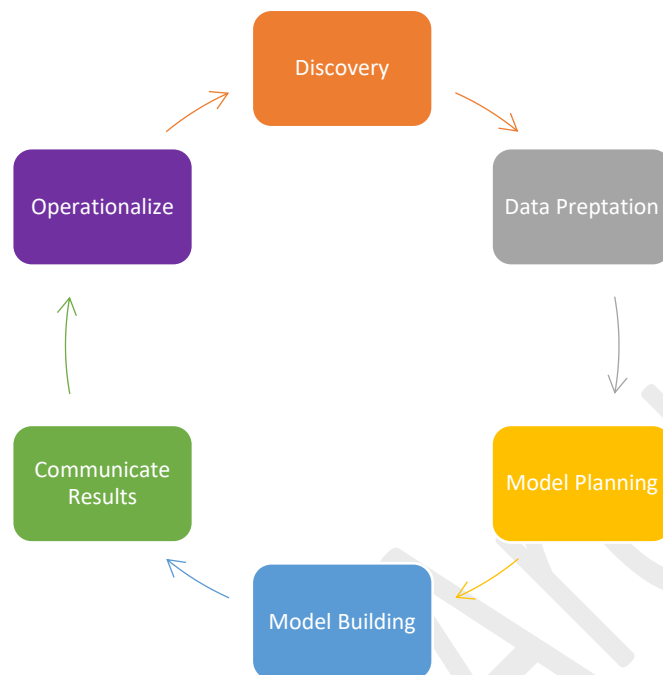
Data Analytics Lifecycle and Demonstration of the process of Extract Transform and Load using Excel and SSIS



ETL - Extract, Transform, Load

Made By
Dhawal Arora
MBA Business Analytics

Data Analytics lifecycle



Phase 1 – Discovery

1. Learning the business domain
2. Resources – Technology, tools, system, data & people.
3. Framing the problem – Identifying main objective of the project.
4. Identifying the key Stakeholders.
5. Interviewing the Analytics Sponsor.
6. Developing initial hypothesis.
7. Identifying potential Data sources.
 - Identifying data sources
 - Capture aggregate data sources
 - Review the raw data
 - Evaluate the data structure & tools needed.
 - Scope the sort of data infrastructure needed for this type of problem.

Phase 2 – Data Preparation

It includes steps to explore, per-process & condition data prior to modelling & analysis.

1. Preparing the analytics sandbox / workspace sand box should 5 – 10 times the size of original dataset.
2. Preforming ETLT (Extract Transform Load Transform) – Analytics sandbox / workspace should have reliable network connections & bandwidth to underlying data sources for uninterrupted read & write.

For Big ETL Hadoop or MapReduce is used advisable to make on inventory.

3. Learning about the data – Making relationship / understand the data some outputs could be surprising.

Some Points –

- Clarifying the data that team has before starting.
- Highlight gaps by identifying datasets which could be inaccessible but is useful in an organisation or data owner.
- Identifying dataset outtrade the organisation that may be useful to obtain through API's, data sharing or purchasing data to supplementing already existing datasets.

4. Data Conditioning

Refers to the process of cleaning data normalizing datasets and performing transformation on the data, it is also known as pre-processing.

5. Survey & Visualize

After collecting the dataset, it needs a subsequent analysis, a useful step to leverage data visualisation tools to gain an overview of the data. Seeing high level patterns in the data enables one to understand characteristics about data very quickly.

6. Common tools for data preparation phase preparation phase.

Hadoop – Can perform massively parallel ingest and custom analysis for web traffic parasing, GPS location analytics, genomic analysis & combining massive unstructured data feeds from multiple source.

Alpine Miner – Provides graphical user interface for creating analytics workflows, including data manipulation & a series of analytics event such as staged data mining technique on Postgres SQL and other Big Data Sources.

Open Refine – A free open source, powerful tool for working with messy data. 'IT is a popular GUI based tool for performing data transformation and it's one of the most robust free tools.

Data Wrangler- Tool for cleaning & transferring subset of the data can be manipulated in wrangler via its GUI and then same operations can be written in Java & Python code to be executed offline.

Phase 3 – Model Planning

In this phase data science team identifies candidate model to apply to the data for clustering, classifying or finding relationship. It is during this phase that the team refers to the hypothesis developed in Phase 1.

Research and model planning in Industry Verticals

Consumer Packaged Goods- Multiple linear regression automatic relevance determination (ARD) and decision tree.

Retail Banking – Multiple Regression

Retail Business – Logistic Regression, ARD, decision tree

Wireless Telecom – Neural Network, Decision Tree, Hierarchical neuro fuzzy system, rule evolver, logistic regression.

- 1) Data Exploration – Objective is to understand the relationship among the variable and method to understand the problem domain.
- 2) Model Selection – Choose an analytical technique. Big data involves determining If the team will be using technique best suited for structured data or a hybrid approach. Create the initial models using a statistical software package such as R, SAS or Matlab.
- 3) Common tools for Planning Phase
 - R has complete set of modelling capabilities and provides a good environment for building interpretive models with high quality code. R contains nearly 5000 packages for data analysis and graphical representation. It has ability to interface with database via ODBC (open Database connectivity) connection and execute statistical test & analysis against Big Data via open source connection.

- **SQL Analysis service** – Can perform in database analytics of common data mining functions, involved aggregation and basic predictive modelling.
- **SAS/ACCESS** – Provide integration between SAS and the analytics sandbox via multiple data connection such as ODBC, JDBC and OLE DB.

Phase 4 – Model Building

Data Scientist team needs to develop datasets for training, testing and production purpose. These data sets enable the data scientist to develop the analytics model and train it & use other side for testing model.

1) Common tools for the model Building Phase. There are many tools available.

SAS Enterprise Miner – Allows to run predictive and descriptive models based on large volumes of data.

Gretl and SPSS – Offers method to explore & analyze data through a GUI and Command-line interface.

Matlab – Provide high language for performing a variety of data analytics, algorithm and data exploration.

Alpine Miner – Provide GUI front end for users to develop analytics workflow and interact with Big Data tools and performs on the back end.

STATISTICA and Mathematica are also popular and well regarded mining and analytics tools.

R & PL (Procedural Language) – Procedural language for Post Gre SQL.

Octave – A free software programming language for computational modelling has some feature of matlab used in major university for teaching machine learning.

WEKA – Free data mining software package with analytics work bench, functions can be also be executed within JAVA code.

SQL – In database implettation such a s MATLAB alternative to in-memory desktop analytical tools. MATLAB provides on open source machine learninglibrary of algorithm that can be executed in database for Postgre SQL or Greenplum.

Python – PL that provides toolkits for Machine Learning and analysis, such as Scikit – Learn, numpy, scipy, pandas and related data visualization using matplotlib.

Phase 5 – Communicate Results

The team needs to determine if it succeeded or failed in its objective. The team must be rigorous with the data to determine whether it will prove or disprove the hypothesis outlined in phase 1 discovery.

Before deploying models on large scale on production environment team can manage risk more effectively and the team can learn by undertaking a small scope pilot deployment before a wide scale roll out.

Phase 6 - Operationalize

This approach enables the team to learn about the performance and related constraint on small scale and make some adjustment before deployment.

Main stake holder of project and their expectation.

Business User – Benefits and implications of the finding to the business.

Project Sponsor – Asks questions related to the business impact of the project, the risk and returns on investment and the way project can be evangelized in organization.

Project Manager – Whether project was completed on time and within budget and how will the goods be met.

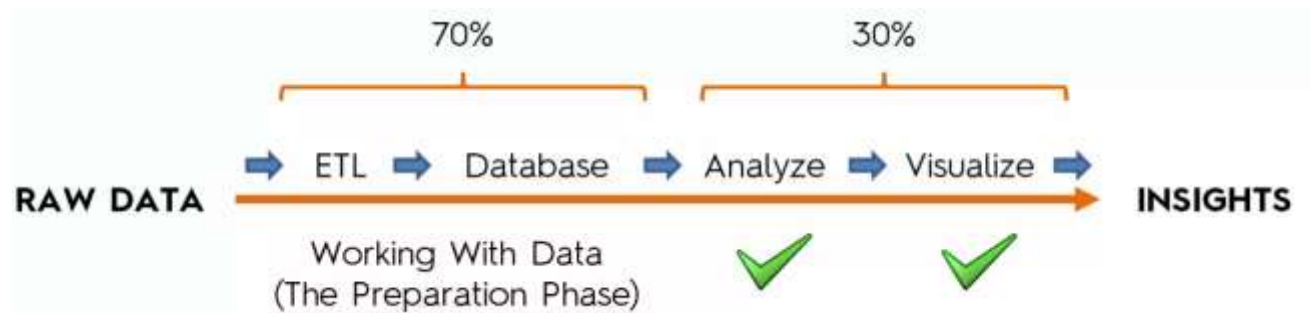
Business Intelligence Analyst – Reports and dashboard he manages will be impacted and need to change.

Data Engineer and Data Base Administrator – Needs to share their code from the analytics project and create a technical document on how to implement it.

Data Scientist – Needs to share the code and explain the model to her peers, managers and other stakeholders.

Working with Data

In order to draw insights from a raw data, the data has to go through a long journey and this journey has four major steps.



ETL – Extract Transform Load

Data we want to analyse can be stored in number of locations.

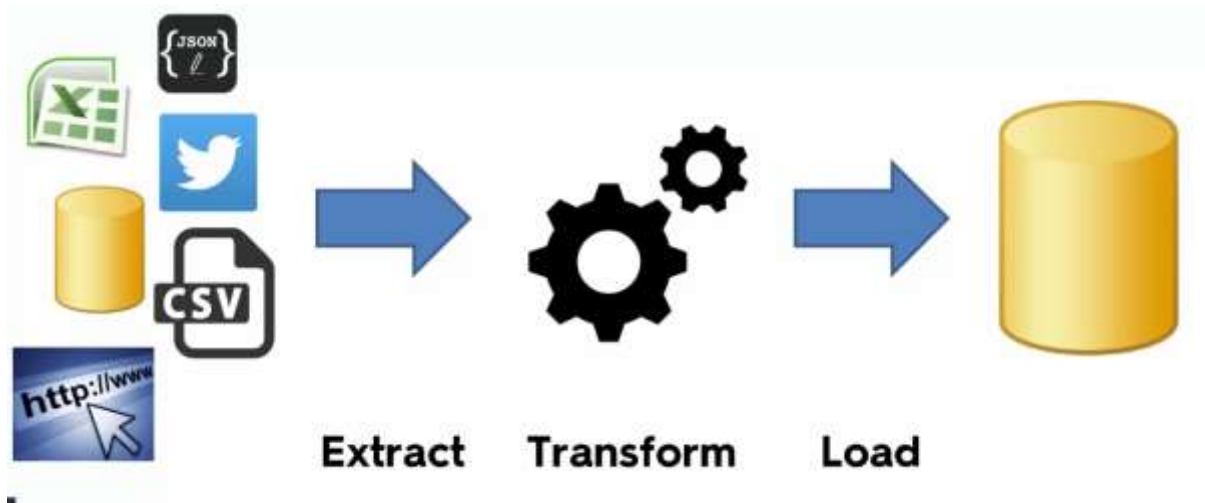
For Instance, A Data Base, Excel Spread Sheet, Website, Twitter, JSON file and CSV file.

Working on data on Source file could be risky, one can modify a row of data and jeopardise the work. In the worst case one can severely impact the critical business process or can crash internal system.



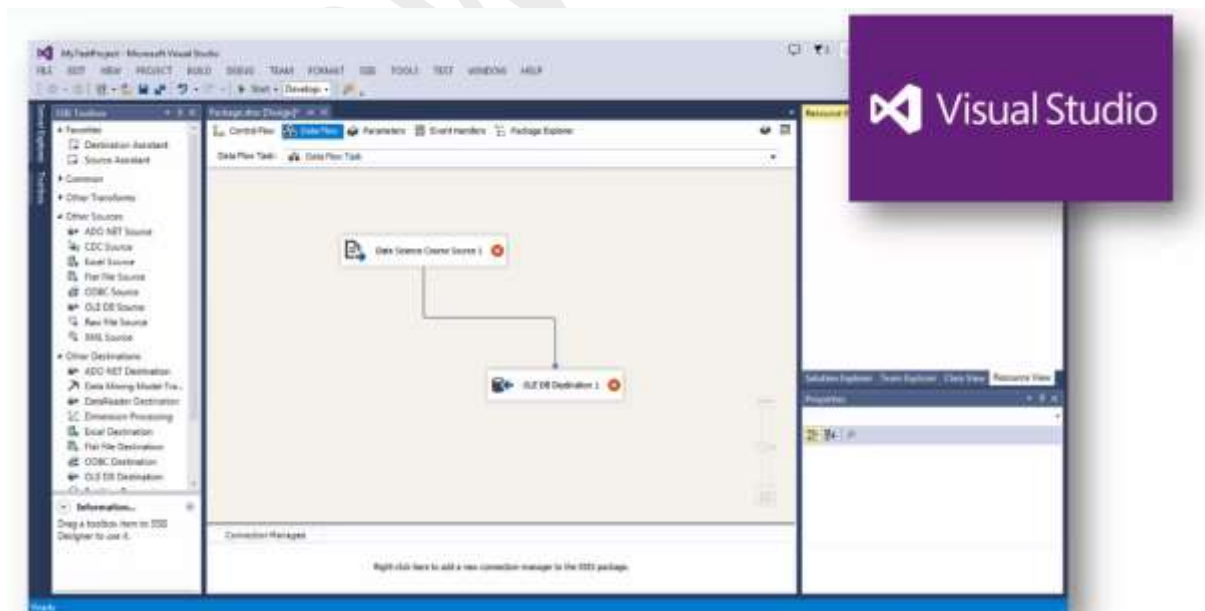
Business Intelligence Tools

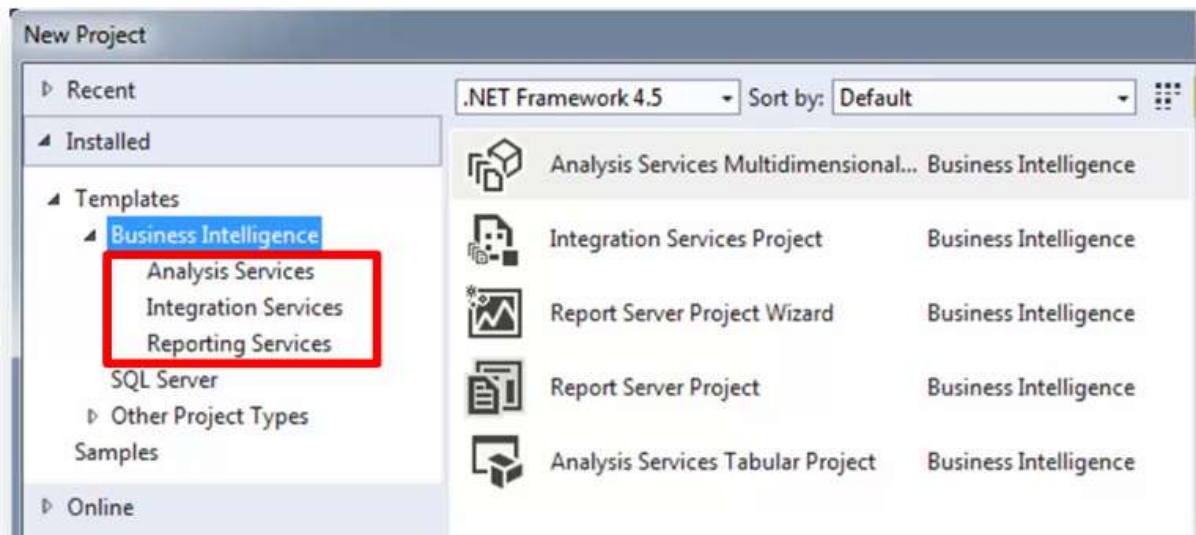
In our process of transformation we are going to use Microsoft Visual Studio - MSVS



Visual Studio – We can work in C++, C sharp, create apps and Build Software. MSVS will be used to manage our ETL process.

We will be working on SSDT- BI Domain of MS Visual Studio. Previously known as Business Intelligence Development Studio.





SSAS (Analysis Services) = **Analyse** (We use R/ Python)

SSIS (Integration Services) = **ETL**

SSRS (Reporting Services) = **Visualize** (Tableau / etc.)

Performing Phase 1 – Folder Structure

In our overall process we will organise our work into different folders Structure. Using smart folder structure increases efficiency, we can keep track of our work in different phases and able to audit the data errors with ease.

We have added a date which will sort our folders chronologically when we order them by name.

20190617 Bank Demo	27-06-2019 02:47 ...	File folder
20190628 Fake Names	04-07-2019 01:29 ...	File folder
20190717 Fake NamesUK	31-05-2020 08:47 ...	File folder
20190718 Office Supplies Store	18-07-2019 10:52 ...	File folder
20190718 SQL Practice Dataset	18-07-2019 02:03 ...	File folder
20190725 FakeNamesCanada	25-07-2019 04:24 ...	File folder
YYYYMMDD Project Name	27-06-2019 02:46 ...	File folder
YYYYMMDD Project Name - Copy	26-07-2019 05:45 ...	File folder

<input type="checkbox"/> Name	Date modified	Type
1. Original Data	17-07-2019 05:55 ...	File folder
2. Prepared Data	17-07-2019 06:14 ...	File folder
3. Uploaded Data	17-07-2019 06:18 ...	File folder
4. Analysis	17-07-2019 07:58 ...	File folder
5. Insights	27-06-2019 02:43 ...	File folder
6. Final	27-06-2019 02:44 ...	File folder

Original Data - We will begin with saving/adding our raw data into Original Data folder which we will not temper with in any circumstances.

Prepared Data - We will save our original data into Prepared Data folder which for any modification done to original data. This include cleaning up of data.

Uploading date folder - It's a temporary stop for data, when we are ready to upload our file we store our data into this folder in a subfolder with a date in YYYYMMDD format and name.

Analysis Folder -It is for storing our script, codes, errors etc that we create in our course of analysis

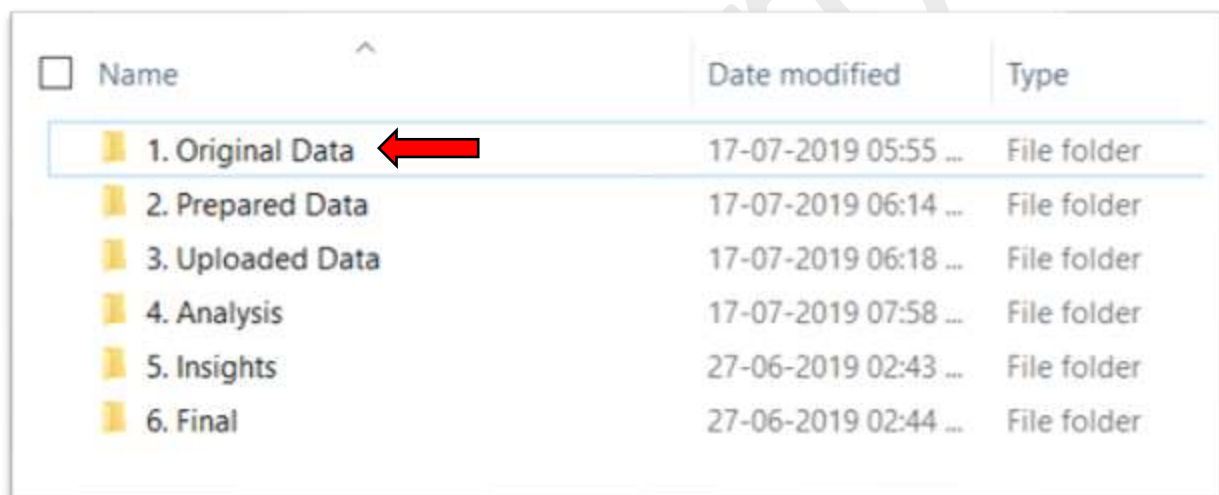
Insight Folder -It is for any preliminary results.

Final Folder – For drafts and final reports

Working with Raw Data

Original Data

We will begin with saving/adding our raw data into Original Data folder.



<input type="checkbox"/> Name	Date modified	Type
1. Original Data	17-07-2019 05:55 ...	File folder
2. Prepared Data	17-07-2019 06:14 ...	File folder
3. Uploaded Data	17-07-2019 06:18 ...	File folder
4. Analysis	17-07-2019 07:58 ...	File folder
5. Insights	27-06-2019 02:43 ...	File folder
6. Final	27-06-2019 02:44 ...	File folder

Preparation of Data, Handling Raw Data

Taking an Overview of our raw data before transforming it into excel. We are using Notepad++ here.

Excel file: Data Science Exercise\Data Science Project\20190717 Data\Homes\K05_Original Data\PreparedData\K05 - Homes.xlsx

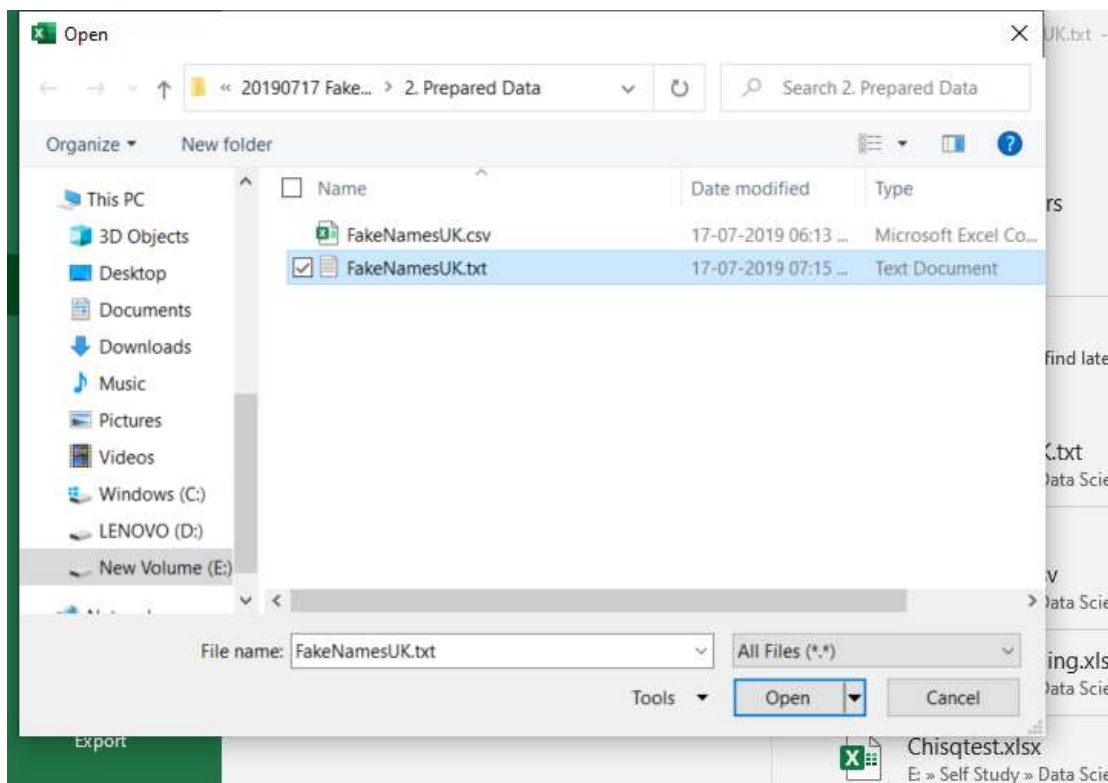
	Number	Gender	Title	Name	Name	Address	ZipCode	EmailAddress	Username	Password	BirthDay	CCType	CCNumber	CVV2	CCExpires	BloodType	K11
1	male	Mr.	Dewei	Tuan	"42 Main Rd,	FOKI, UK	"SY21 3FF	DeweiTuan@fleckens.hu	Piten1973	ohSoo22d	7/31/1973	MasterCard	5371472825372				
2	female	Ms.	Natasha	Watts	"46 Coast Rd,	KIRKPATRICK-FLEMING, UK	"DG11 2HN	NatashaWatts@jourrapide.com	Fultank	Weey0ko8	10/30/193						
3	female	Mrs.	Qing	Yuan	Ch'ien	"3 Friar Street,	CLAKUD-NEMYDD, UK	"LL15 8SQ	QingYuanChien@rhyta.com	Diater	IemahStoo	6/8/1981	Visa				
4	male	Mr.	Msthan	Bykov	"79 Abbey Row,	NORTON BRIDGE, UK	"ST15 3JF	NathanBykov@einrot.com	Coogge	cieMaSchai81	10/23/1979	Visa	4453				
5	female	Mrs.	Mua	Han	"37 Shire Oak Road,	SAXTHORPE, UK	"NN11 4NK	Muallan@superrito.com	Prisam	ad6irGleach	12/10/1943	Visa	4485599				
6	male	Mr.	Chan	K'ung	"57 Sigone Weil Avenue,	WEATHEROAK HILL, UK	"B48 1AN	ChanKung@superrito.com	Cercs1939	iePneeOph	1/10/1939	M					
7	male	Mr.	Innocent	Golubov	"97 Winchester Rd,	METHERINGHAM, UK	"LN4 8SN	InnocentGolubov@rhyta.com	Heramntooped	Aeth6ewoh4	2/24/						
8	female	Ms.	Valentine	Ilyina	"4 Earls Avenue,	WHITESOUSE, UK	"PA29 6FY	ValentineIlyina@jourrapide.com	Vareat	dae1Ahp6	3/4/1947						
9	male	Mr.	Nicholas	Lees	"89 Botley Road,	HIDELETON-ON-THE-MOLDS, UK	"YO25 8QU	NicholasLees@armyspy.com	Sularoat	oleifaim6Eis	12/						
10	female	Ms.	Ling	Bocharova	"51 London Road,	COMPTON, UK	"PO18 7AL	LingBocharova@dayrep.com	Therip	ohSiende	12/30/1949	Visa	4556				
11	male	Mr.	Shing	Ting	"9 St Denys Road,	POVEY CROSS, UK	"RM6 9JF	ShingTing@armyspy.com	Thibust	aboh3BaStho	7/11/1936	Visa	471633				
12	female	Dr.	Praskovya	Isayeva	"53 Thornton St,	HUNTERSTON, UK	"KA23 9XB	PraskovyaIsayeva@armyspy.com	Ebothe	iePooOhoo	5/16/1974						
13	female	Ms.	Tao	Shih	"80 Ramsgate Rd,	WILMINGTON, UK	"BN26 0BX	TaoShih@fleckens.hu	Distert1960	aduaV9Cee001	10/16/1960	MasterCa					
14	female	Mrs.	Jade	Marsh	"18 Redcliffe Way,	WOODSTON, UK	"PE2 6RD	JadeMarsh@gustr.com	Scoul984	aeko2An7	8/13/1984	Visa	453975239				
15	male	Mr.	Ewan	Lowe	"0 Well Lane,	PAVENHAM, UK	"MK43 4DE	EwanLowe@fleckens.hu	Dingdowas	ELRe2eeobus	4/9/1940	Visa	4916024504531				
16	female	Mrs.	Park	Chuang	"9 Glodiseth Street,	KINNSROOK, UK	"LN3 0AD	ParkChuang@einrot.com	Thattly	Xah3Eelie	6/7/1974	MasterCard					
17	female	Ms.	Jasmine	Thornton	"55 Thornton St,	HUNTLEY, UK	"GL19 5GG	JasmineThornton@teleworm.us	Bentarast	ahkuria7Ee	10/15/1954						
18	female	Ms.	Ling	Feng	"79 Walwyn Rd,	CHARSFIELD, UK	"IP13 8EP	LingFeng@armyspy.com	Fightfil	Ishaip9B	4/10/1977	Visa	49169876734				
19	male	Mr.	Evan	Ward	"1 Stone Cellar Road,	KINGSTON RUSSELL, UK	"DT2 8LE	EvanWard@rhyta.com	Symbeptere	yoh1KaCu6	11/14/1983	Mas					
20	female	Mrs.	Daniella	Mills	"91 Ploughley Rd,	TODHILLS, UK	"CA6 3QN	DaniellaMills@dayrep.com	Morchad	jiiNgeSealah	12/18/1955	Ma					
21	female	Ms.	Isabella	Ross	"62 Stamford Road,	ANGLE, UK	"SA71 0QU	IsabellaRoss@dayrep.com	Gholeake	chailshk9ph	12/23/1953	Visa	4				
22	male	Mr.	Gabriel	Kodryashov	"7 Hertingfordbury Rd,	NEWNUM, UK	"NN11 6RD	GabrielKodryashov@fleckens.hu	Nhavruid	shahioiOdyah	8/						
23	female	Ms.	Hannah	Reeves	"21 Ash Lane,	SEALS, UK	"BA12 7BH	HannahReeves@einrot.com	Sonters	shuThu2chl	7/13/1959	MasterCard	514				
24	female	Mrs.	Greta	Sayceva	"82 Bridge Street,	GONACHAN, UK	"G63 2GB	GretaSayceva@einrot.com	Artswr	uThTheR0ph	3/22/1959	Master					
25	female	Mrs.	Lan	Tu	"73 Holburn Lane,	HEATHTON, UK	"WV5 4PB	LanTu@gustr.com	Reyes1962	Yalquesas2e	2/27/1962	MasterCard	5337606				
26	male	Mr.	Stepan	Kovalev	"40 Tonbridge Rd,	COOMBE, UK	"TQ9 6QN	StepanKovalev@dayrep.com	Iday1974	Albelgie90Fo	1/16/1974	Visa	49				
27	male	Mr.	Shaining	Chien	"97 Stamford Road,	APPLETON, UK	"CX13 9XS	ShainingChien@einrot.com	Behiliess	ohNoVie9eja1	12/15/1986	V					
28	male	Mr.	Bo	Wei	"92 Bury Rd,	HAM, UK	"SNE 7QR	BoWei@dayrep.com	Grased	Bengiab2ph	4/17/1975	MasterCard	5533084196036220	111	3/2		
29	male	Mr.	Sun	Chao	"6 Slice Lane,	CROWFIELD, UK	"IP6 1WD	SunChao@cuvov.de	Selamudder	ieFiej4Ae	8/10/1952	MasterCard	545197033376				
30	female	Ms.	Elizabeth	Srnakova	"38 Wrexham Rd,	EYPE, UK	"DT6 1QG	ElizabethSrnakova@superrito.com	Othapprocy	Calliaf7	10/12/1960						
31	female	Ms.	Pompy	Baxter	"56 East Street,	MAKE CROSS, UK	"TN6 5JH	PompyBaxter@armyspy.com	Compter	cheili7Aiaul	1/16/1981	Visa	4				

We will save our original data into Prepared Data folder which for any modification done to original data. This include cleaning up of data.

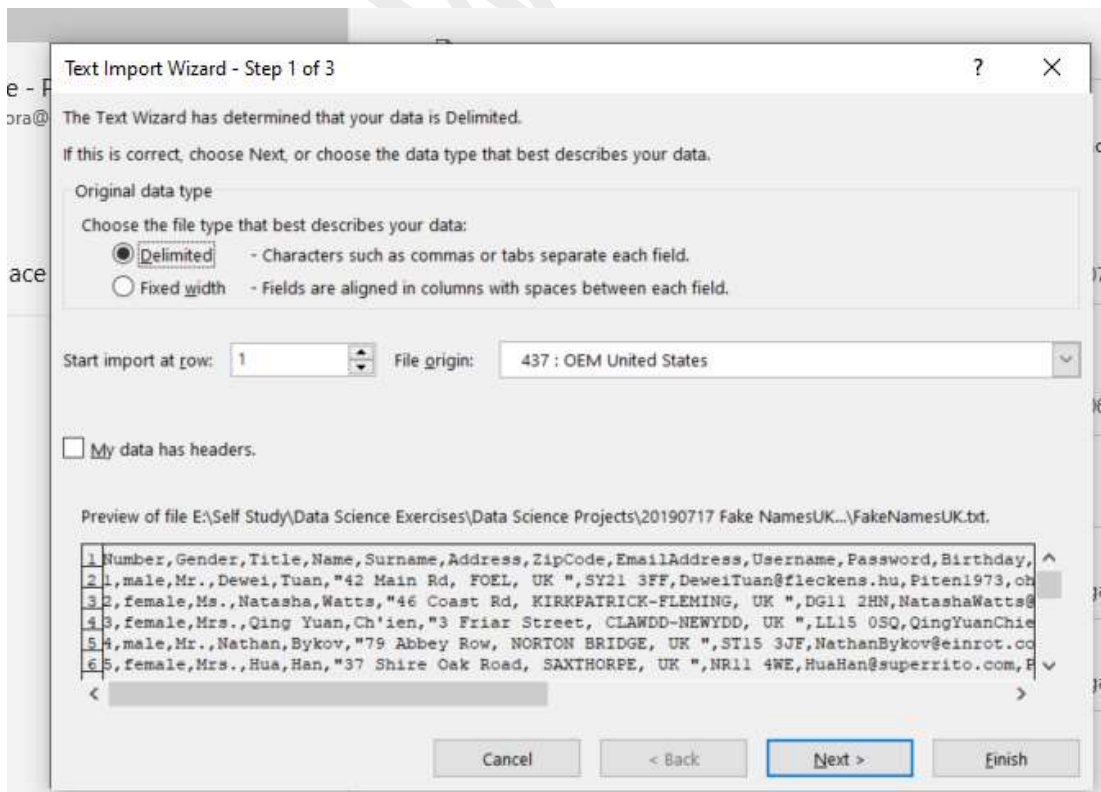
<input type="checkbox"/> Name	Date modified	Type
1. Original Data	17-07-2019 05:55 ...	File folder
2. Prepared Data	17-07-2019 06:14 ...	File folder
3. Uploaded Data	17-07-2019 06:18 ...	File folder
4. Analysis	17-07-2019 07:58 ...	File folder
5. Insights	27-06-2019 02:43 ...	File folder
6. Final	27-06-2019 02:44 ...	File folder

We will prepare it to transform it into excel sheet. We will change the extension of the file with .txt so it can be readable by excel.

Data Wrangling Phase 1 - Using Exel to transform raw data



On opening of our file, we get option of Text import wizard where we will Delimit our file using comma “,” separator.



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☐ Tab

☐ Semicolon

☒ Comma

☐ Space

☐ Other:

☐ Treat consecutive delimiters as one

Text qualifier:

Data preview

Number	Gender	Title	Name	Surname	Address	ZipCode	Email
1	male	Mr.	Dewei	Tuan	42 Main Rd, FOEL, UK	SY21 3FF	DeweIT
2	female	Ms.	Natasha	Watts	46 Coast Rd, KIRKPATRICK-FLEMING, UK	DG11 2HN	Natash
3	female	Mrs.	Qing Yuan	Ch'ien	3 Friar Street, CLAWDD-NEWYDD, UK	LL15 0SQ	QingYu
4	male	Mr.	Nathan	Bykov	79 Abbey Row, NORTON BRIDGE, UK	ST15 3JF	Nathan
5	female	Mrs.	Hua	Han	37 Shire Oak Road, SAXTHORPE, UK	NR11 4WE	HuaHan

Cancel < Back Next > Finish

We will select all the columns and format it into Text

Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

☐ General

☒ Text

☐ Date: DMY

☐ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Data preview

Text	Text	Text	Text	Text	Text	Text	Text
Number	Gender	Title	Name	Surname	Address	ZipCode	Email
1	male	Mr.	Dewei	Tuan	42 Main Rd, FOEL, UK	SY21 3FF	DeweIT
2	female	Ms.	Natasha	Watts	46 Coast Rd, KIRKPATRICK-FLEMING, UK	DG11 2HN	Natash
3	female	Mrs.	Qing Yuan	Ch'ien	3 Friar Street, CLAWDD-NEWYDD, UK	LL15 0SQ	QingYu
4	male	Mr.	Nathan	Bykov	79 Abbey Row, NORTON BRIDGE, UK	ST15 3JF	Nathan
5	female	Mrs.	Hua	Han	37 Shire Oak Road, SAXTHORPE, UK	NR11 4WE	HuaHan

Cancel < Back Next > Finish

We will change the Birth date column which is in text to **MDY – Month Date Year**. Date format using Text to Columns Wizard.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Number	Gender	Title	Name	Surname	Address	ZipCode	EmailAddr	Username	Password	Birthday	CCType	CCNumber	CVV2	CCExpires	BloodType	Kilograms	Centimeters
2	1	male	Mr.	Dewei	Tuan	42 Maie RrSY21 3FF		DeweiTuan	Pten1973	ohSooZ2d	7/31/1973	MasterCar	53714728	919	9/2019	B+	89.6	183
3	2	female	Ms.	Natasha	Watts	46 Coast RlD011 2HN		NatashaW	Futbank	Nenrykoc8	10/30/1973	Visa	47167531	577	3/2017	B+	52.4	170
4	3	female	Mrs.	Qing Yuan	Ch'en	3 Friar StreLL15 0SQ		QingYuanC	Diater	IemahRlov	6/8/1981	Visa	44851220	895	5/2016	AB+	91.0	158
5	4	male	Mr.	Nathan	Bykov	29 Abbey 1ST15 3IF		NathanByl	Cougge	ieMa8ch	10/23/1973	Visa	45126443	295	5/2018	O+	110.4	187
6	5	female	Mrs.	Hua	Han	37 Shine Q.NR11 4Wl		Hualland	Prine	edbr01eic	12/19/1943	Visa	44855993	711	3/2017	B+	84.1	155
7	6	male	Mr.	Chan	K'ung	97 Simone B48 1AW		ChanKung	Cers1939	ieP1neeOp	1/10/1939	MasterCar	52480959	308	7/2020	A+	74.5	162
8	7	male	Mr.	Innocent	Golubov	97 WiescheLN4 8SH		InnocentG	Henventou	AathGeoov	2/24/1962	Visa	45194966	143	2/2019	B+	86.7	168
9	8	female	Ms.	Valentine	Ilyina	4 Earls AvePA29 0FY		Valentinell	Varenet	dae1Aphc	3/4/1947	MasterCar	51921958	189	4/2020	B+	96.6	156
10	9	male	Mr.	Nicholas	Lees	89 Botley 1YQ25 0QU		NicholasL	Cularopat	oleifalM6	12/26/1953	Visa	49166430	769	8/2018	O+	117.1	179
11	10	female	Ms.	Inga	Bocharova	51 LondonPD18 7AL		IngaBocha	Therip	nhBim4e	12/30/1943	Visa	45561409	715	8/2018	O+	60.7	164
12	11	male	Mr.	Shing	Ting	9 St Denys RH46 9IF		ShingTing	Thibust	abch3ba5	7/11/1930	Visa	47163385	156	5/2018	A+	83.7	163
13	12	female	Dr.	Praskovya	Isayeva	53 Thorns KA23 9XB		Praskovya	Ebothe	iePoo0hod	5/14/1974	MasterCar	51456749	743	1/2016	B+	62.8	162
14	13	female	Ms.	Tao	Shih	80 RamagelBN26 0BK		TaoShihg	Diater196	aduaV9Gw	10/16/196	MasterCar	51186247	162	11/2017	A+	60.9	165
15	14	female	Mrs.	Jodie	Marsh	18 RedcliffPE2 6RD		JodieMarsh	Scout1984	ieko2Au7	8/13/1984	Visa	45197523	888	7/2017	O+	83.3	169
16	15	male	Mr.	Ewan	Lowie	8 Well Lan MK43 4DE		EwanLowie	Dingdown	ERReZeobc	4/9/1940	Visa	49160245	138	8/2017	B+	101.5	166
17	16	female	Mrs.	Park	Chuang	9 Gloddaw LN3 0AU		ParkChuan	Thattly	Kah3Eelo	6/7/1974	MasterCar	52535892	740	8/2020	B+	81.1	166
18	17	female	Ms.	Jasmine	Thornton	55 Thorns GL19 5GG		JasmineTh	Bentast	ahxuria7E	10/15/197	MasterCar	54489945	567	2/2019	A+	51.2	159
19	18	female	Ms.	Ling	Feng	29 Waiway IP13 8EP		LingFeng	Fightfl	ishuap8b	4/10/1973	Visa	49169876	911	4/2020	B+	52.5	155
20	19	male	Mr.	Evan	Ward	1 Stone CeDT2 8LB		EvanWard	Symbentou	yoh1KaCu	11/14/194	MasterCar	54759885	670	1/2018	B+	99.7	177
21	20	female	Mrs.	Danielle	Mills	91 Ploughl CA6 3QN		DanielleM	Morchad	jiiNgeSeal	12/18/195	MasterCar	55242947	484	9/2018	O+	65.5	163
22	21	female	Ms.	Isabella	Ross	62 Stamford SA71 0QU		IsabellaRo	Sholeake	chaiLahk9	12/23/195	Visa	45398018	938	2/2017	O+	96.6	168
23	22	male	Mr.	Gabriel	Kudryasho	7 Hertingh NN11 6RD		GabrielKur	Whawuld	shohjo1O	9/20/1963	Visa	45360526	539	4/2019	O+	89.1	178
24	23	female	Ms.	Hannah	Reeves	23 Ash Lan BA12 7BH		HannahRe	Sonters	shuThro2c	7/13/1953	MasterCar	51432140	725	5/2017	O+	95.2	152
25	24	female	Ms.	Greta	Zaytseva	82 Bridge 1 GS3 2GB		GretaZayt	Artmer	uThTheRD	3/22/1958	MasterCar	52242569	079	2/2020	A+	74.4	163
26	25	female	Mrs.	Lan	Tu	39 Holbur WVS 4PR		LanTu	Bgr Heyes196	Yaiquesa	2/27/1962	MasterCar	53376064	601	12/2019	O+	52.5	165
27	26	male	Mr.	Stepan	Kovalev	40 Tonbriol TQ9 6QN		StepanKov	Iday1974	Abetgiev1	1/16/1974	Visa	49169913	035	2/2018	O+	125.4	182
28	27	male	Mr.	Shaiming	Chien	87 Stamford OX13 5XS		ShaimingC	Bethless	uHNoVier9	12/15/198	Visa	49295559	608	11/2019	A-	89.9	189
29	28	male	Mr.	Bo	Wei	92 Bury RdSN8 7QR		BoWei	@ Grased	Eenglab2p	4/13/1979	MasterCar	55330841	111	3/2017	A+	102.6	179

Connections
Data Types
Sort & Filter
Data T

ay

E

urname

uan

/atts

h'ien

ykov

an

'ung

olubov

yina

ees

ochard

ing

ayeva

hih

larsh

owe

huang

hornto

eng

/ard

tills

oss

91 Ploughl

CA6 3QN

DanielleM

Morchad

jiiNgeSeal

12/18/195

MasterCar

55242947

484

9/2018

62 Stamfo

SA71 0QU

IsabellaRo

Sholeake

chaiLahk9

12/23/195

Visa

45398018

938

2/2017

Convert Text to Columns Wizard - Step 1 of 3

?

×

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

☒ Delimited - Characters such as commas or tabs separate each field.
☐ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

1 Birthday

2 7/31/1973

3 10/30/1934

4 6/8/1981

5 10/23/1979

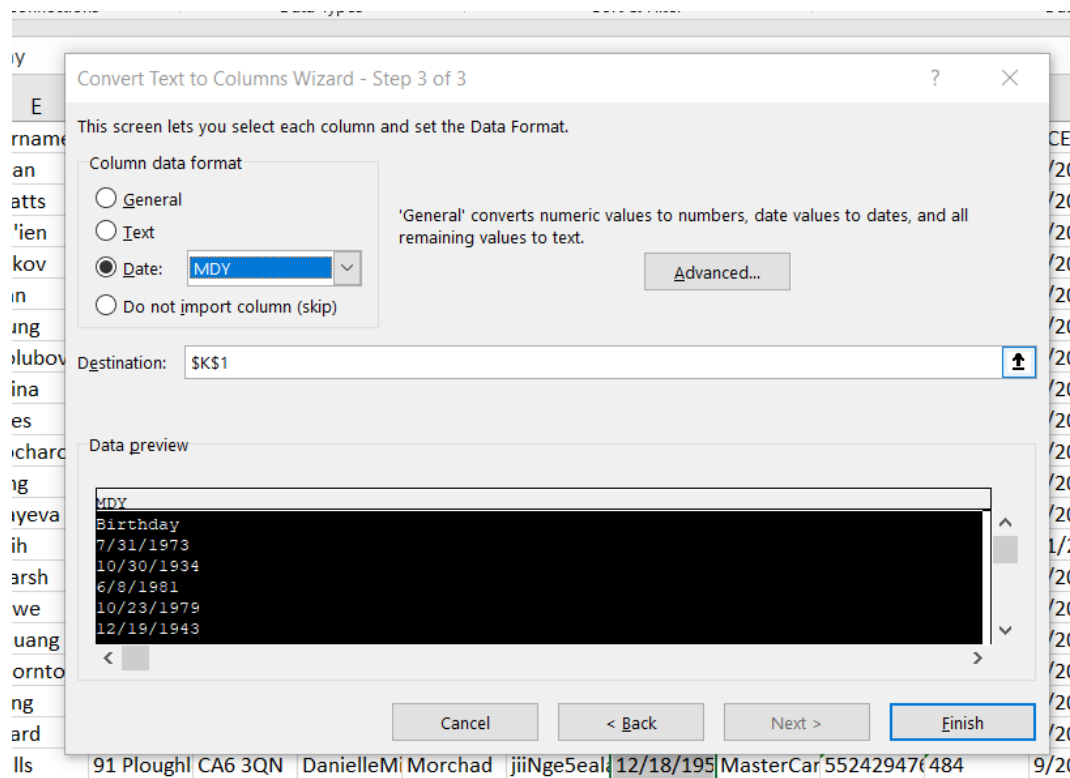
6 12/19/1943

Cancel

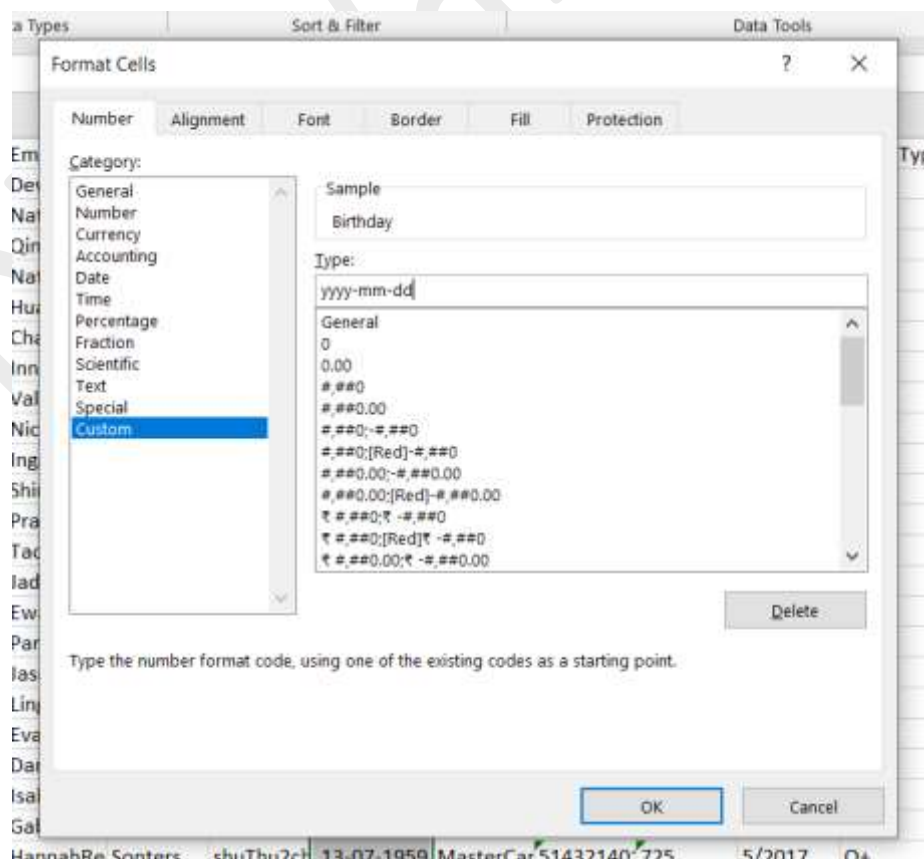
< Back

Next >

Finish



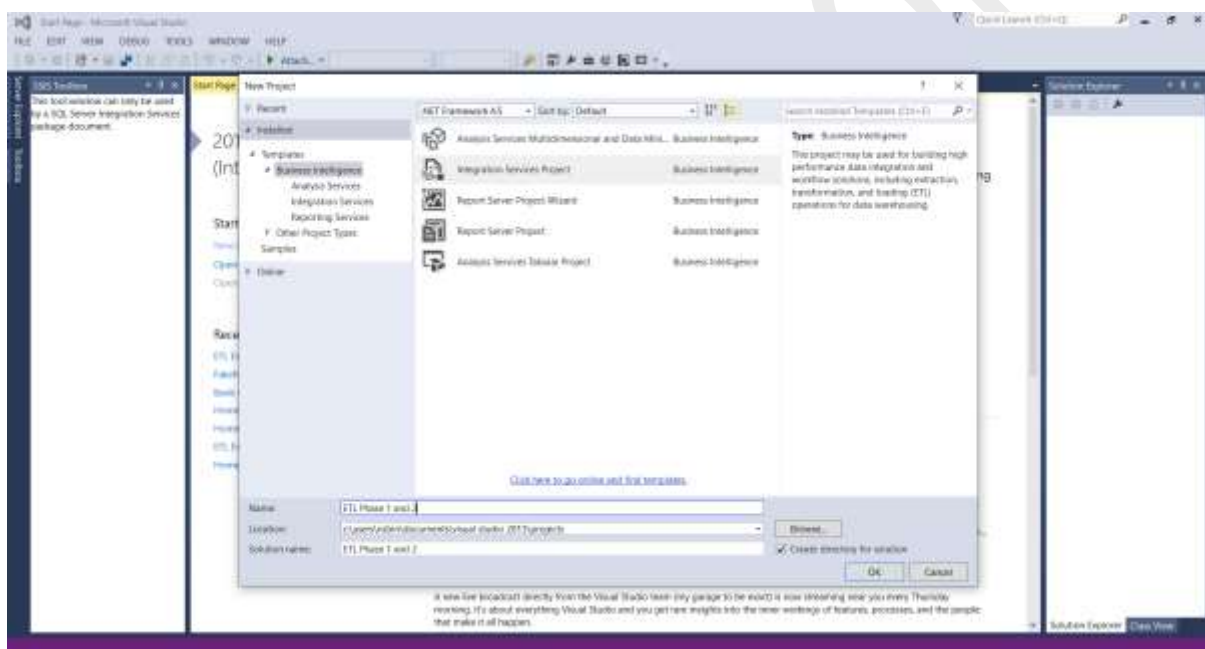
After converting our dates in MYD format, we will change the format of MYD to YYYY-MM-DD as it is international date format.



Similarly we have to change other column such as currency format into Numbers if it is present in our data.

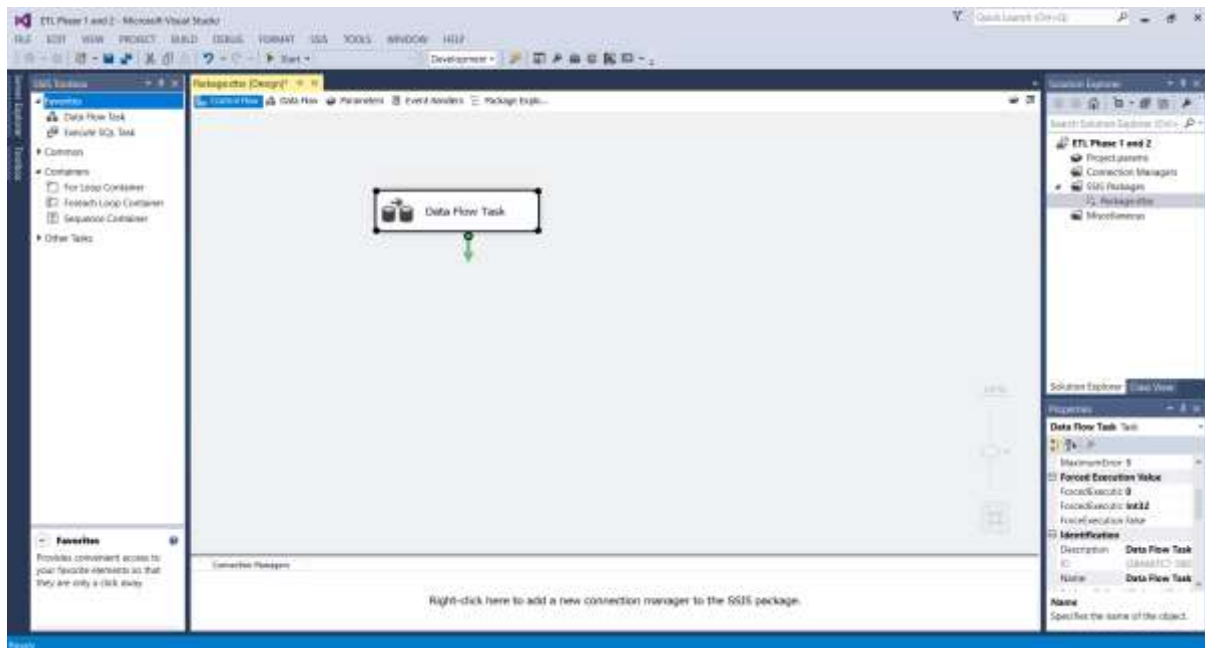
Phase 2 -Working in Microsoft Visual Studio for Transforming and Mapping up of Data using SSIS

After converting our data to a readable format, we will begin with next step of loading our data in Microsoft Visual Studio. We will begin with selecting Integration Service Project and naming our file as shown in the picture below.



MS Visual Studio have drag and drop option, using which we will drag **Data Flow Task** into the **Control Flow** pane. Double clicking it will take us into Data Flow.

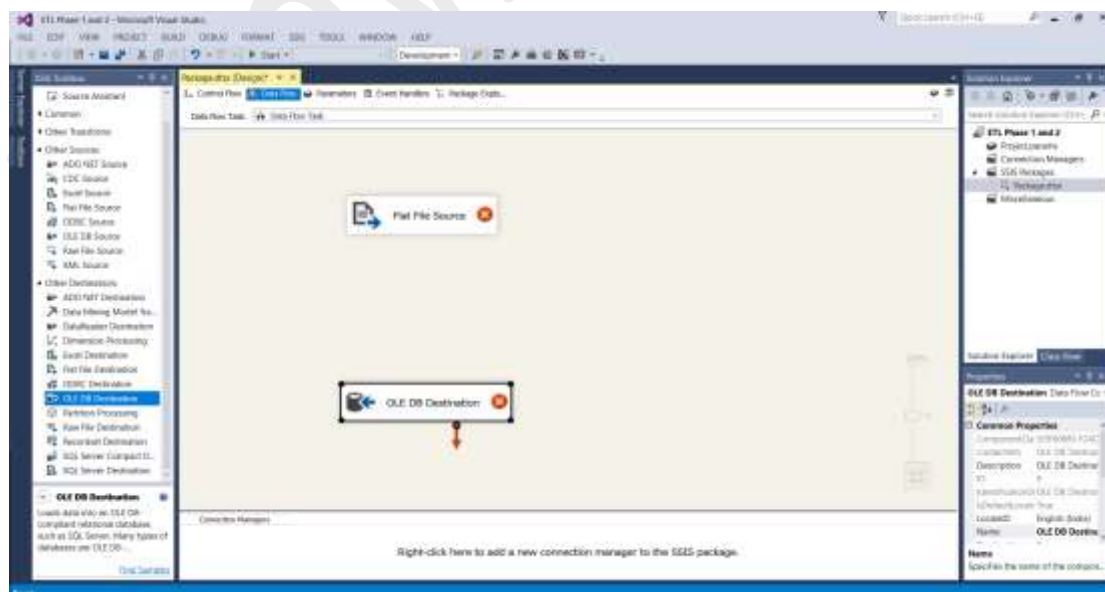
The **Data Flow task** encapsulates the **data flow** engine that moves **data** between sources and destinations, and lets the user transform, clean, and modify **data** as it is moved. Addition of a **Data Flow task** to a package control **flow** makes it possible for the package to extract, transform, and load **data**



Here we will drag **Flat File Source** and **OLE DB Destination** from Source Assistant on the left.

A **Flat File Source** in SSIS is used to extract or reads data from text files. **Flat File Source** uses the **Flat File Connection Manager** to connect with the text files.

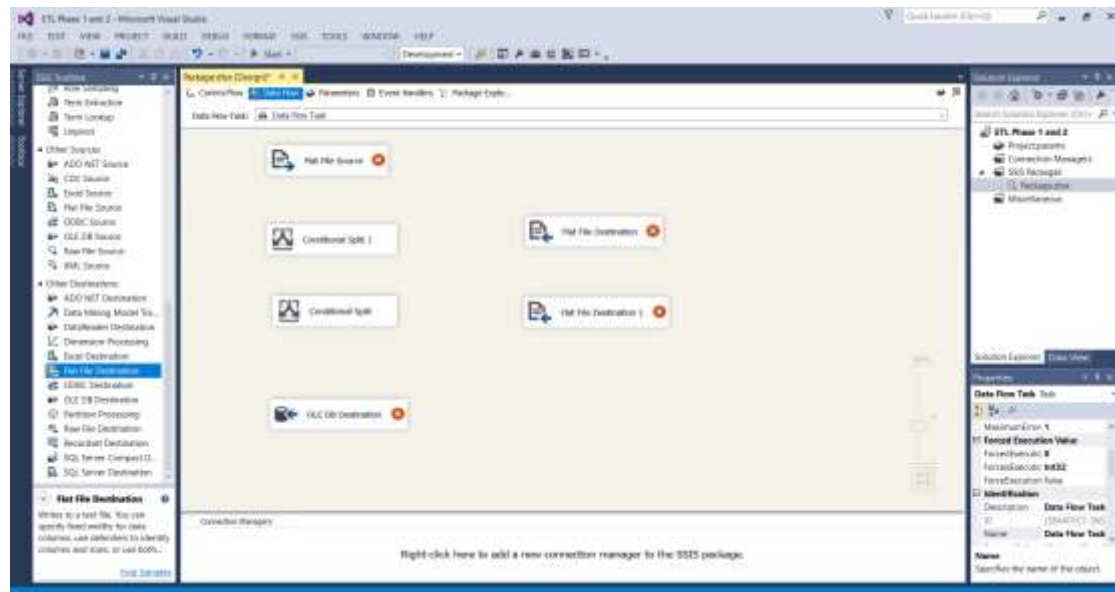
The **SSIS OLE DB Destination** is used to load data into a variety of database tables or views or SQL Commands. **OLE DB destination** editor provides us the choice to select the existing table(s), View(s), or you can create a new table.



In next step we will drag Conditional Split and Flat File Destination twice and arrange in the format shown in the picture.

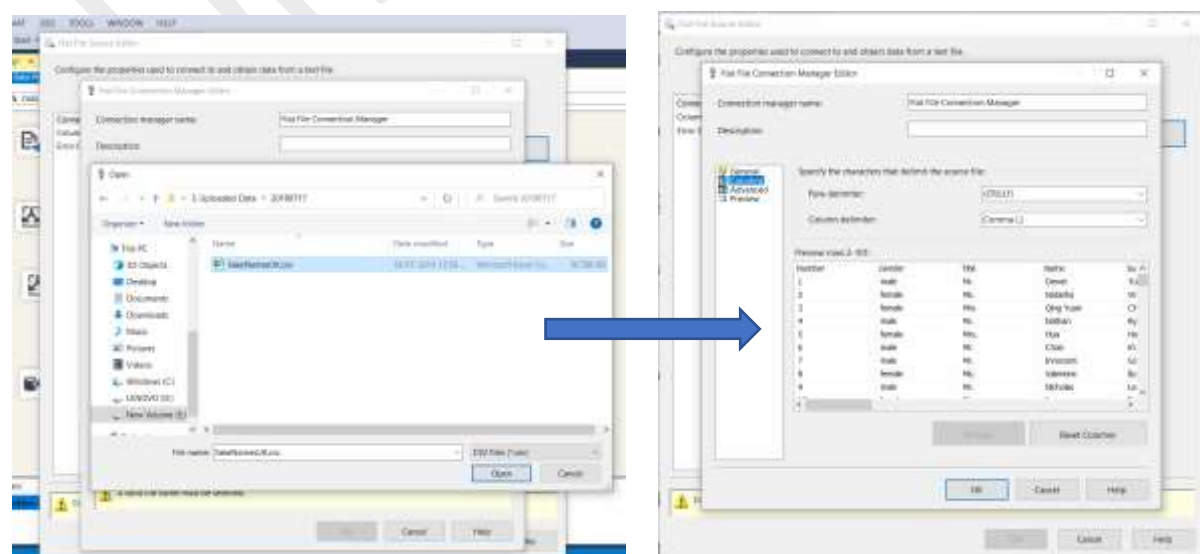
The **Conditional Split** can route data rows to different outputs depending on whatever criteria of the data that you wish. ... The transformation lets you route your data flow to different outputs, based on criteria defined within the transformation's editor

he **Flat File destination** writes data to a text **file**. The text **file** can be in delimited, fixed width, fixed width with row delimiter, or ragged right format. You can configure the **Flat File destination** in the following ways: Provide a block of text that is inserted in the **file** before any data is written.

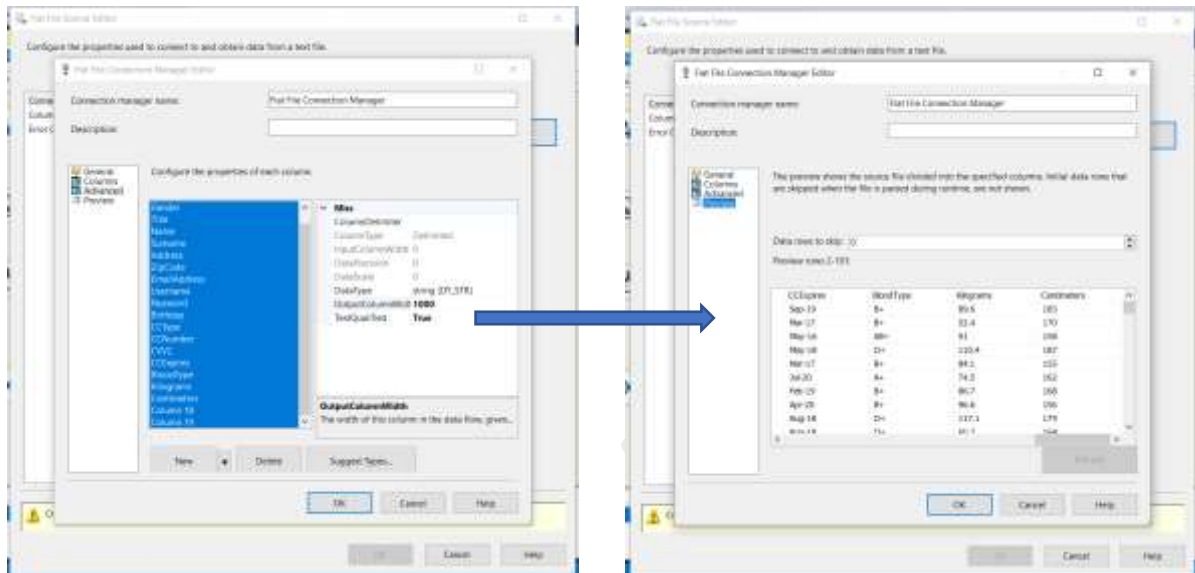


Loading Data into **Flat File Source**, it includes following steps shown in the picture below.

Double Click Flat File Source -> New Connection Manager -> Browse File Destination -> Open



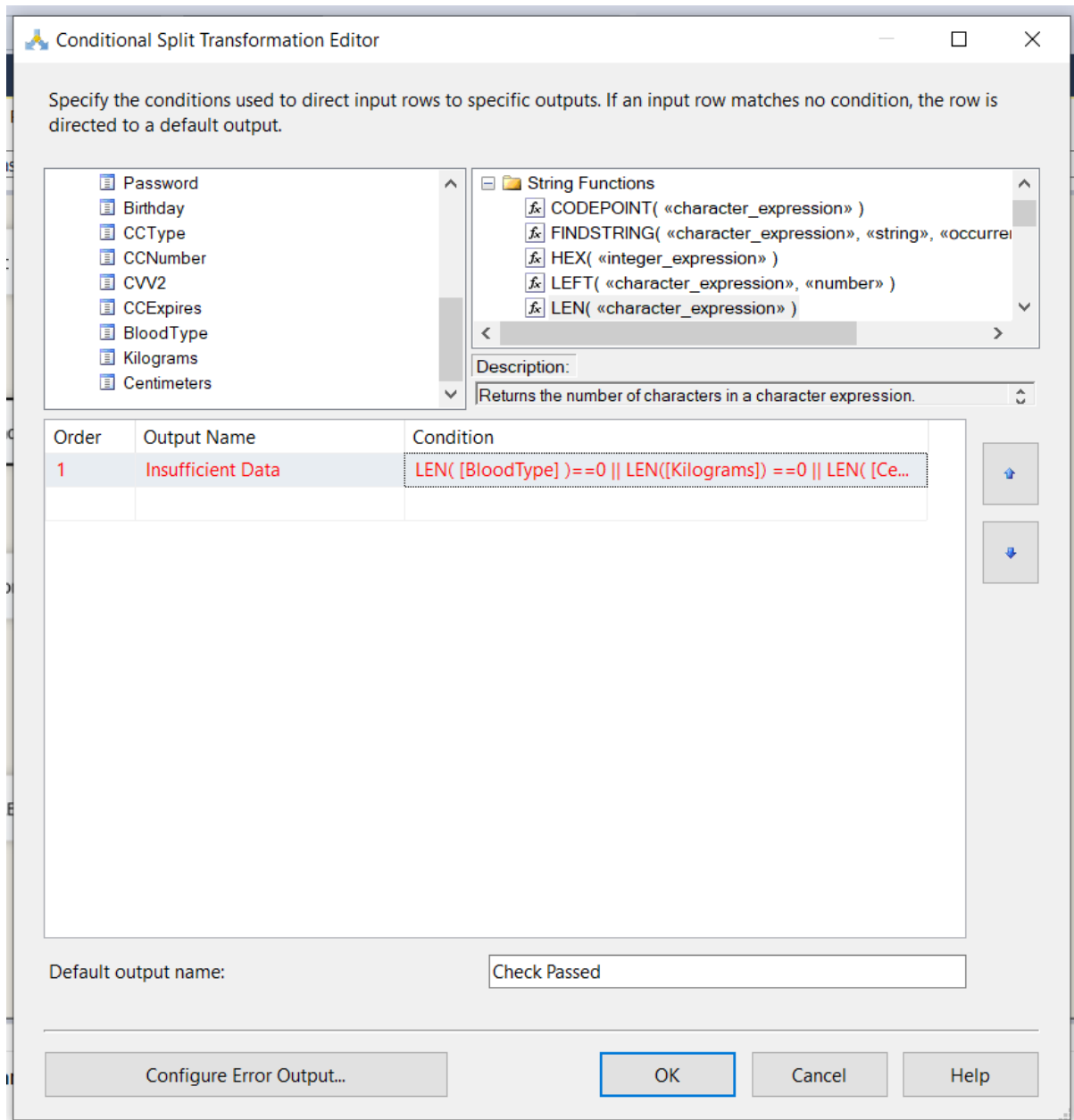
Take over view of data to check if its correct then go to advance column and change the Output Column Width of all column from 50 to 1000. We increase the limit because it increases the character limit to store our data in SQL Database. There are high chance that data may contain character of default 50 word limit.



Click Okay after taking a preview. In preview section.

Double click on **conditional split**, here we will set limitations to filter our data into bad records and insufficient data.

Our data have two extra columns in the end because there is a chance that some our data rows might have shifted to right due to errors.

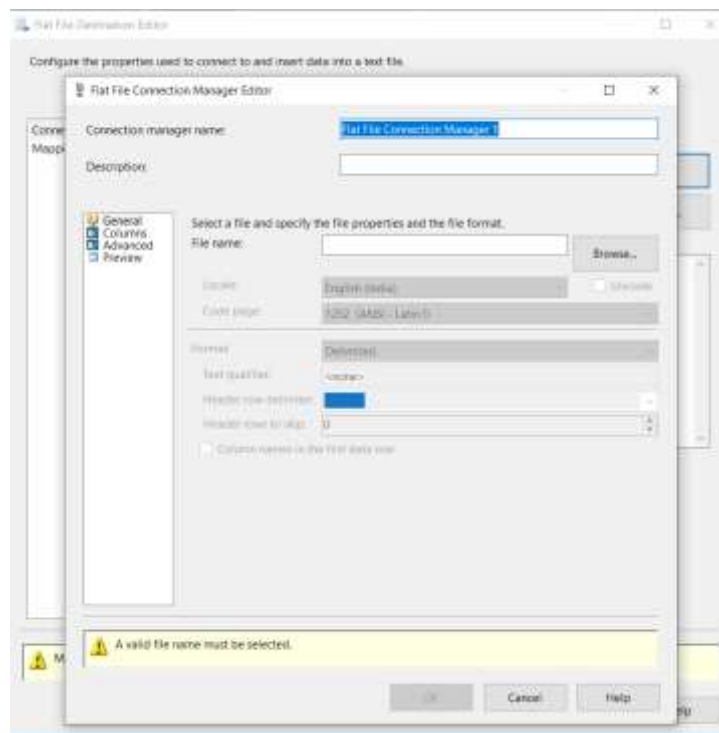


We have given output name as Insufficient Data

So here we have put a condition that if Blood Type or Kilogram or Height column is equal to Zero then those rows will filter out to insufficient data file in Flat File Destination.

Configuring Flat File Destination

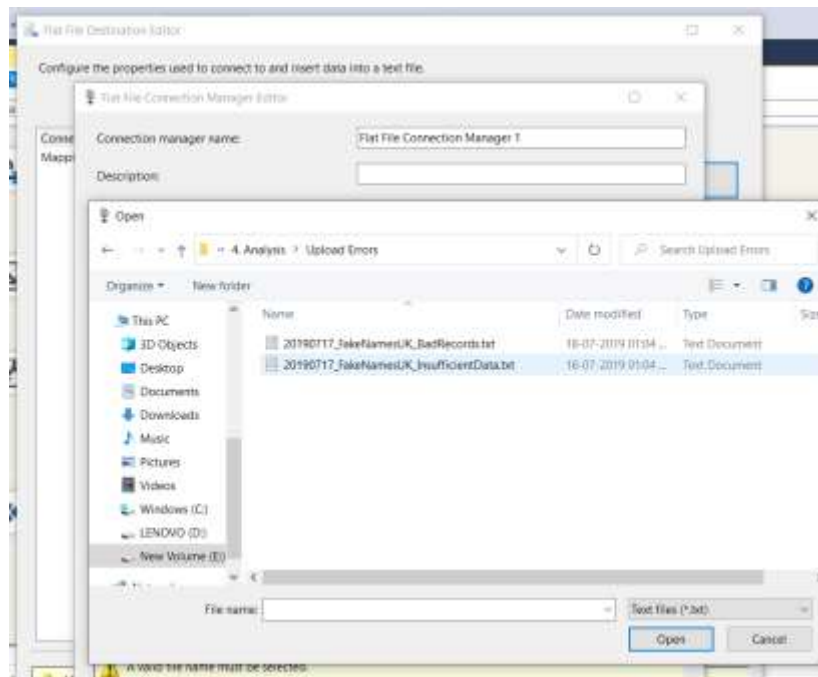
Double click Flat File Destination -> Browse -> Go to File Destination Folder which is Analysis



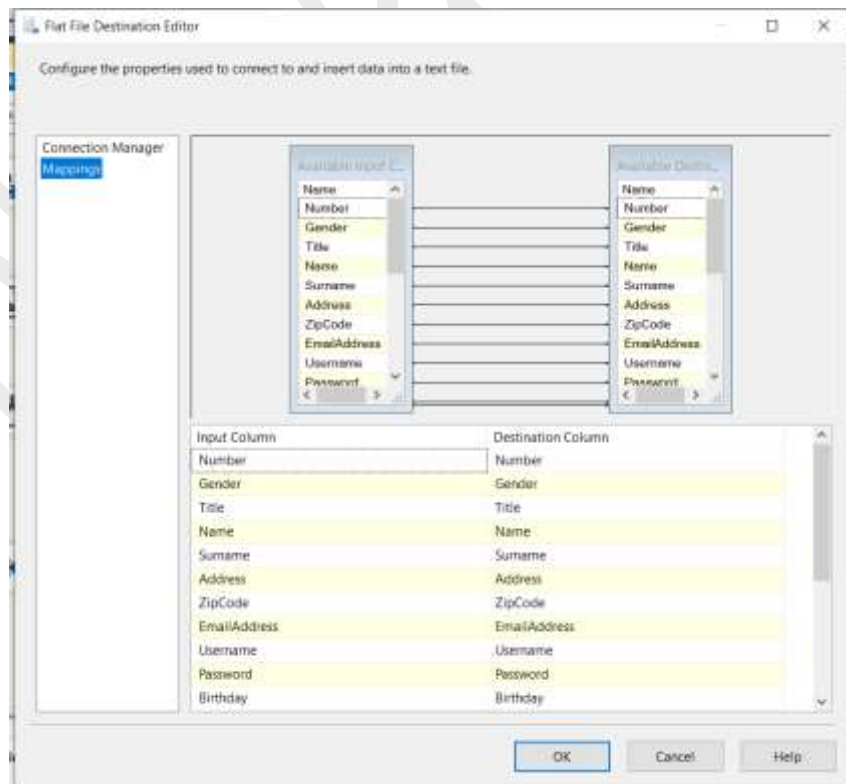
Analysis Folder -It is for storing our script, codes etc that we create in our course of analysis

<input type="checkbox"/> Name	Date modified	Type
1. Original Data	17-07-2019 05:55 ...	File folder
2. Prepared Data	17-07-2019 06:14 ...	File folder
3. Uploaded Data	17-07-2019 06:18 ...	File folder
4. Analysis ←	17-07-2019 07:58 ...	File folder
5. Insights	27-06-2019 02:43 ...	File folder
6. Final	27-06-2019 02:44 ...	File folder

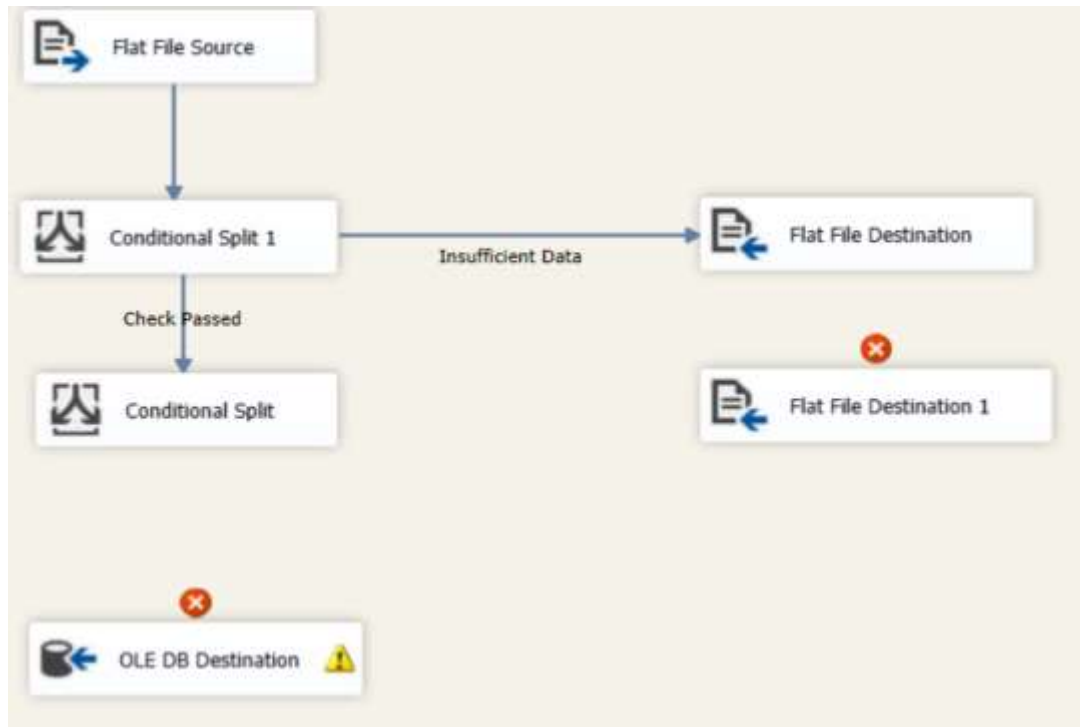
We will create an empty text files in international date format + File name + InsufficientData and Open it.



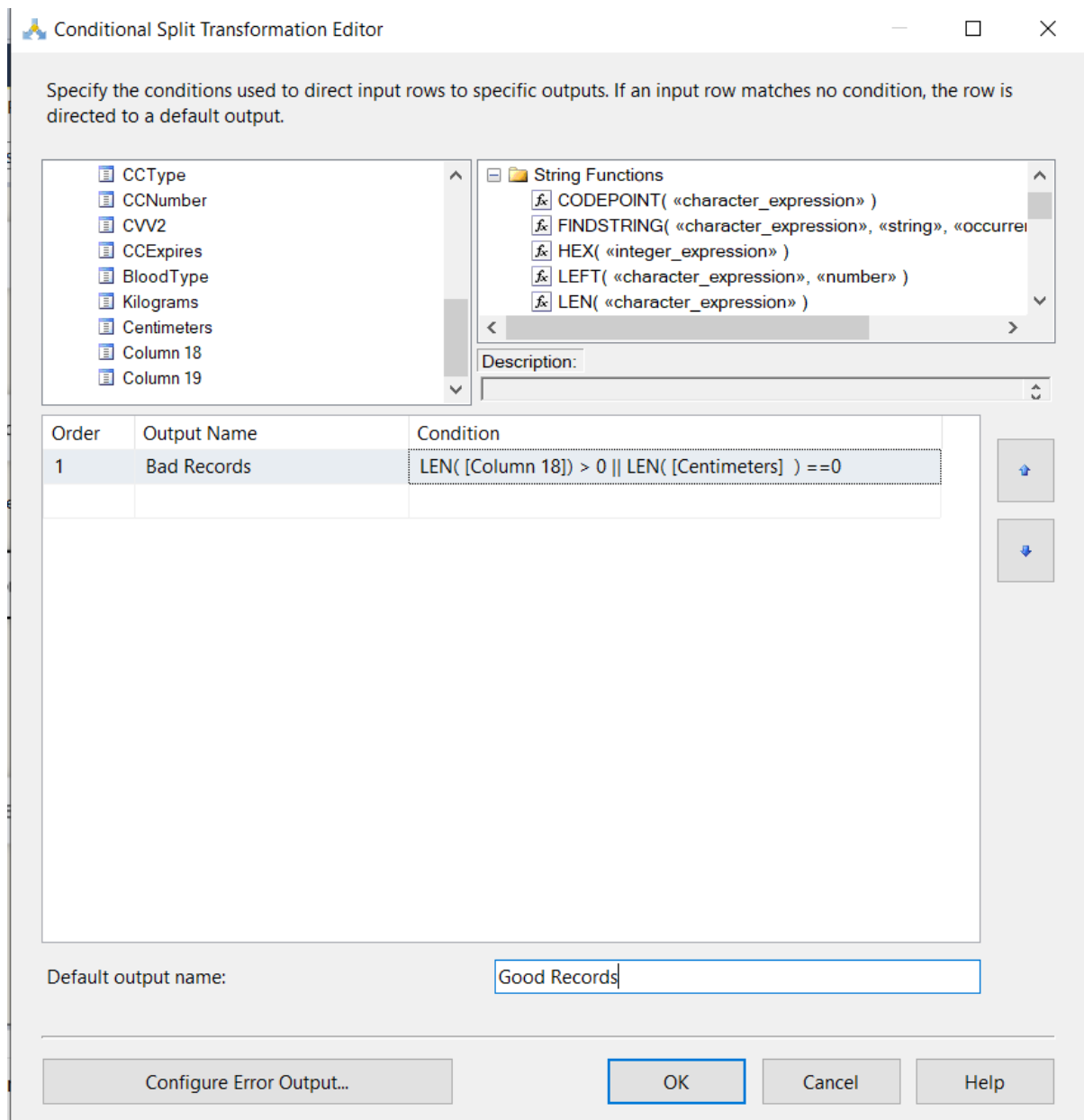
We will check the mapping so see if it matches to the columns of our original Flat File Source File. Click Okay



Two arrows extend below the component. These are called **data paths**. In this case, there is one blue and one red. The blue data path marks the flow of data that has no errors. The red data path redirects row whose values are truncated or that generate an error. Together these data paths enable the developer to specifically control the flow of data, even if errors are present.



We will follow same steps with our other **Conditional Split** and then drag blue arrows to Flat File Destination 1 and OLE DB Destination.



We have put a condition that if length of the characters of **column 18** is more than 0 or **centimetres** = 0, those rows will filter out as a Bad Records in **Flat File Destination 1**

We will repeat same procedure with Flat File Destination 1 of Creating an empty text file in a Analysis Folder\Upload Errors and Name that file in YYYYMMDD+File Name+ BadRecords.txt

In this case for instance we named it as **20190717_FakeNamesUK_BadRecords**

The screenshot shows the 'Flat File Connection Manager Editor' window. It has a sidebar with 'General', 'Columns', 'Advanced', and 'Preview' tabs. The 'General' tab is active. The 'Connection manager name' is 'Flat File Connection Manager 2'. The 'Description' field is empty. Below the sidebar, there's a section titled 'Select a file and specify the file properties and the file format.' The 'File name' field contains 'Errors\20200601_FakeNamesUK_BadRecords.txt' and has a 'Browse...' button next to it. The 'Locale' is set to 'English (India)' and 'Unicode' is unchecked. The 'Code page' is '1252 (ANSI - Latin I)'. The 'Format' is 'Delimited'. The 'Text qualifier' is '"'. The 'Header row delimiter' is '{CR}{LF}'. The 'Header rows to skip' is '0'. There is an unchecked checkbox for 'Column names in the first data row'. At the bottom are 'OK', 'Cancel', and 'Help' buttons.

Flat File Connection Manager Editor

Connection manager name: Flat File Connection Manager 2

Description:

General
Columns
Advanced
Preview

Select a file and specify the file properties and the file format.

File name: Errors\20200601_FakeNamesUK_BadRecords.txt Browse...

Locale: English (India) ☐ Unicode

Code page: 1252 (ANSI - Latin I)

Format: Delimited

Text qualifier: "

Header row delimiter: {CR}{LF}

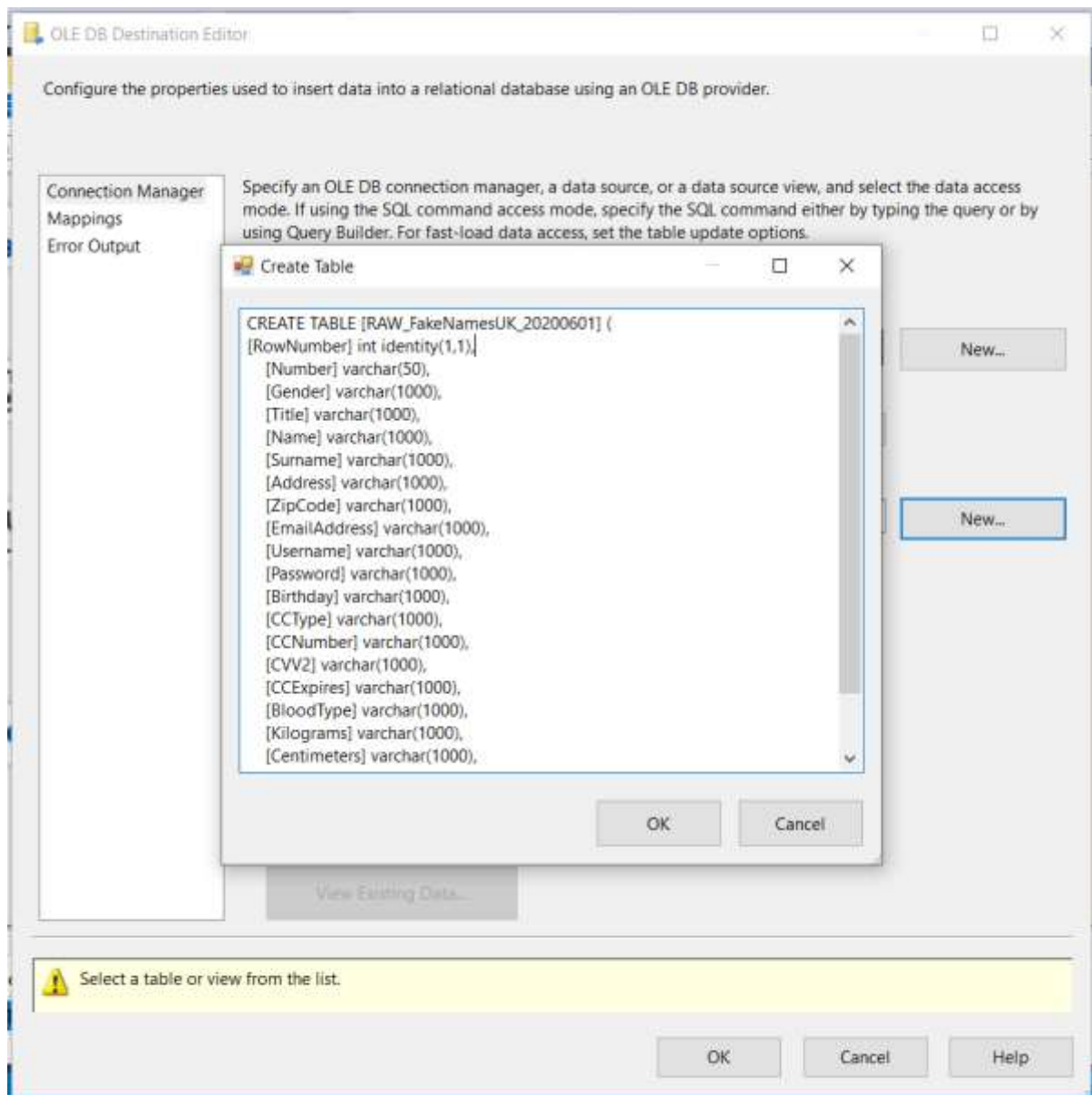
Header rows to skip: 0

☐ Column names in the first data row

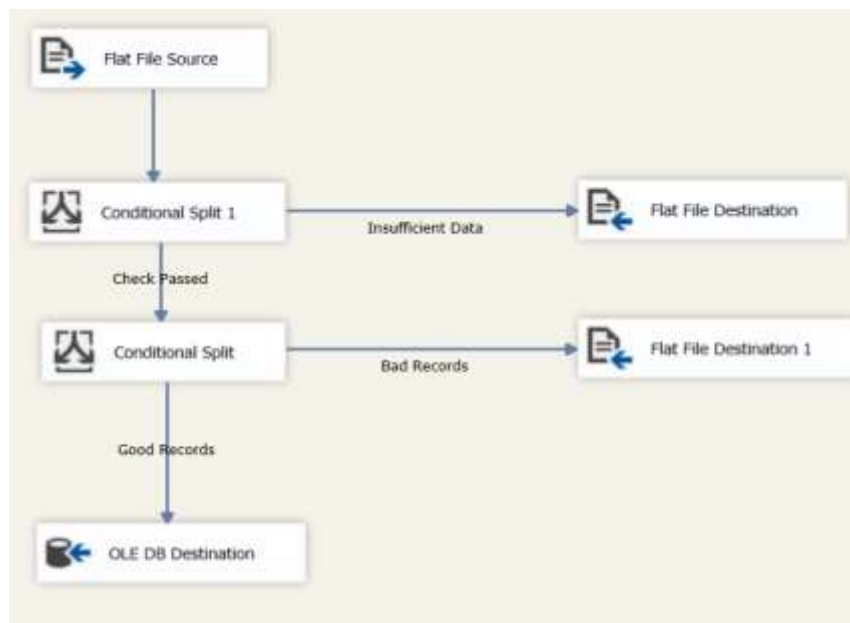
OK Cancel Help

The next step is to work with OLE DB Destination

The **SSIS OLE DB Destination** is used to load data into a variety of database tables or views or SQL Commands. **OLE DB destination** editor provides us the choice to select the existing table(s), View(s), or you can create a new table.



This is how our pane should look in the end after mapping.



Now all we need to do is execute the package and see if it works. To do this, click the **Execute** button. It's the green arrow on the toolbar.

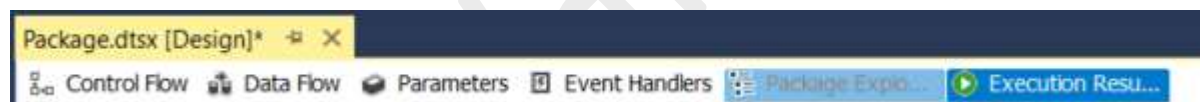
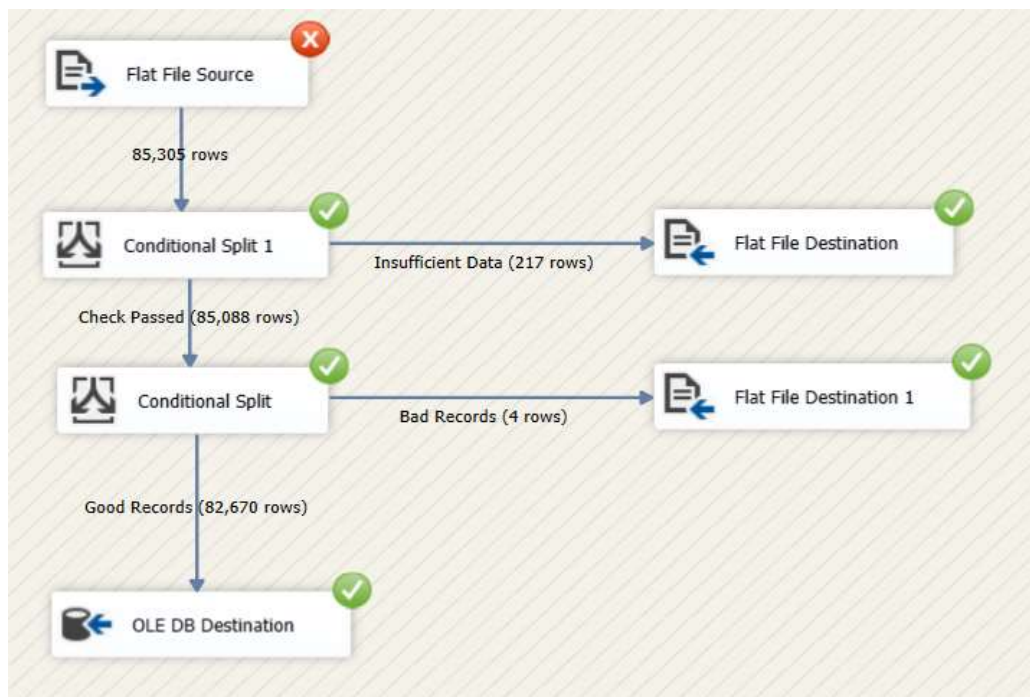


As the package progresses through the data flow components, each one will change color. The component will turn yellow while it is running, then turn green or red on completion. If it turns green, it has run successfully, and if it turns red, it has failed.

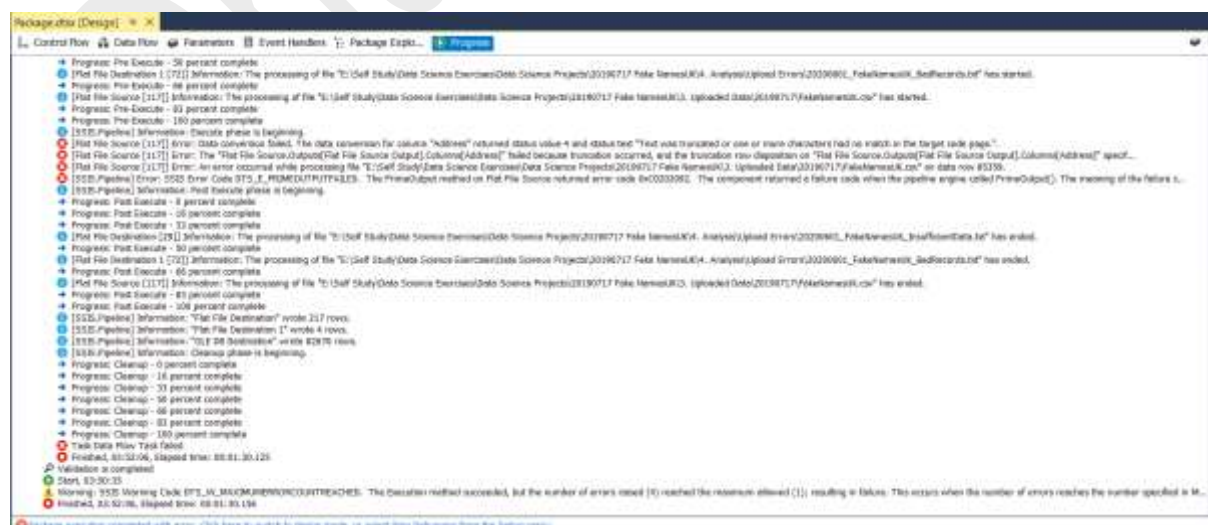


Handling Errors

We have encountered an error while running our package this is how process stops, to look into it we will check the execution result section.



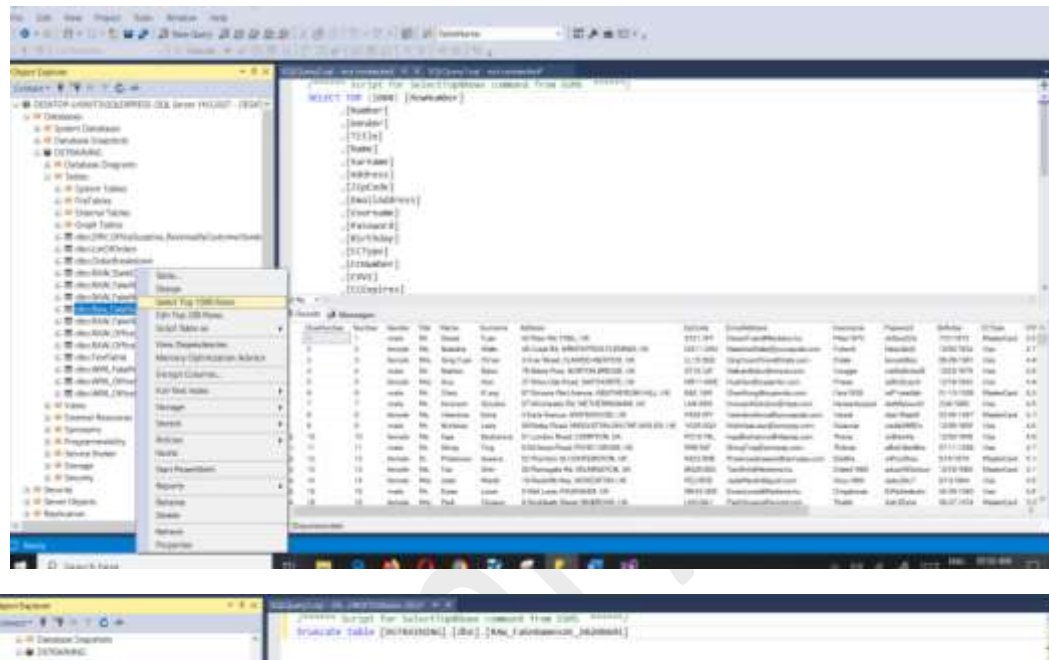
In execution result red circles with white cross is the step where we have faced error. Here it says from **Flat File Source** Data conversion has failed as Address column have one or more character had no match in the target code page.



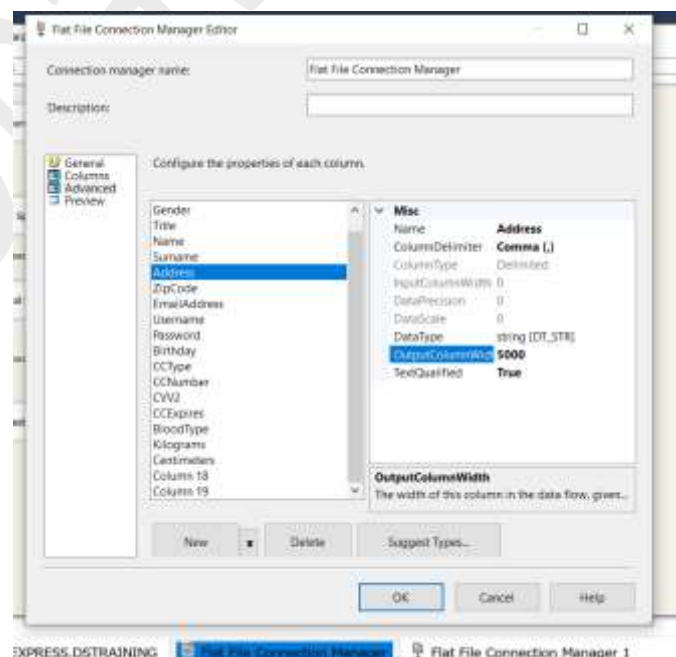
Second step after encountering error in our package, we have to delete this data from data base i.e. **Microsoft SQL Server Management Studio** as we will reupload our data after removing error, to do this go to data base, find this file in data base.

Right Click on it -> Select top 1000 rows -> Remove all the text leaving the last row which will be `"FROM [DSTRAINING].[dbo].[Raw_FakeNamesUK_20190717]"`.

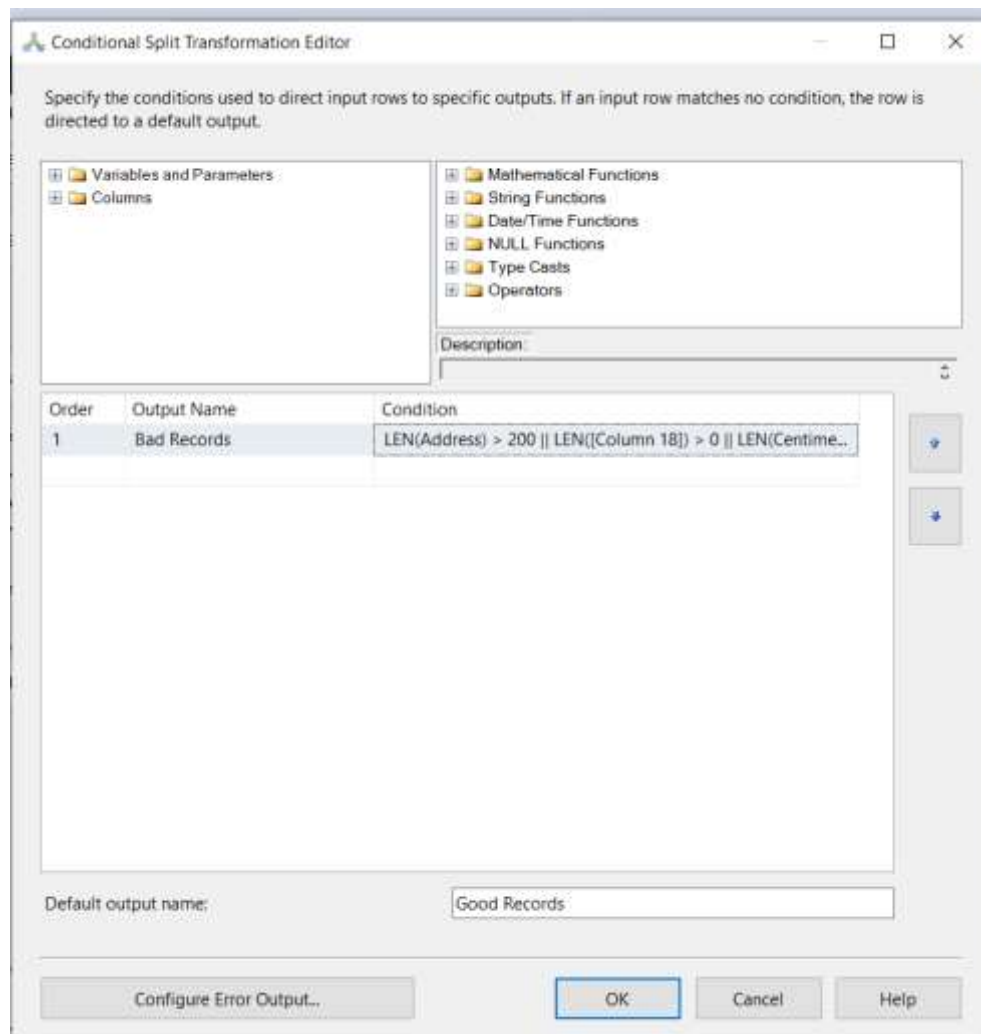
Replace **FROM** with **truncate table** then execute the command.



Third step is to remove the error we were facing, we will open **Flat File Connection Manager 1**, open **Advanced**, select Address column then change its Output Column Width to 5000. Click OK



We know that address can't be of 5000 characters, not even of 100 characters. So we will add a new condition in **Conditional Split** that any address which is too long will also be considered as bad records which we want to exclude it.



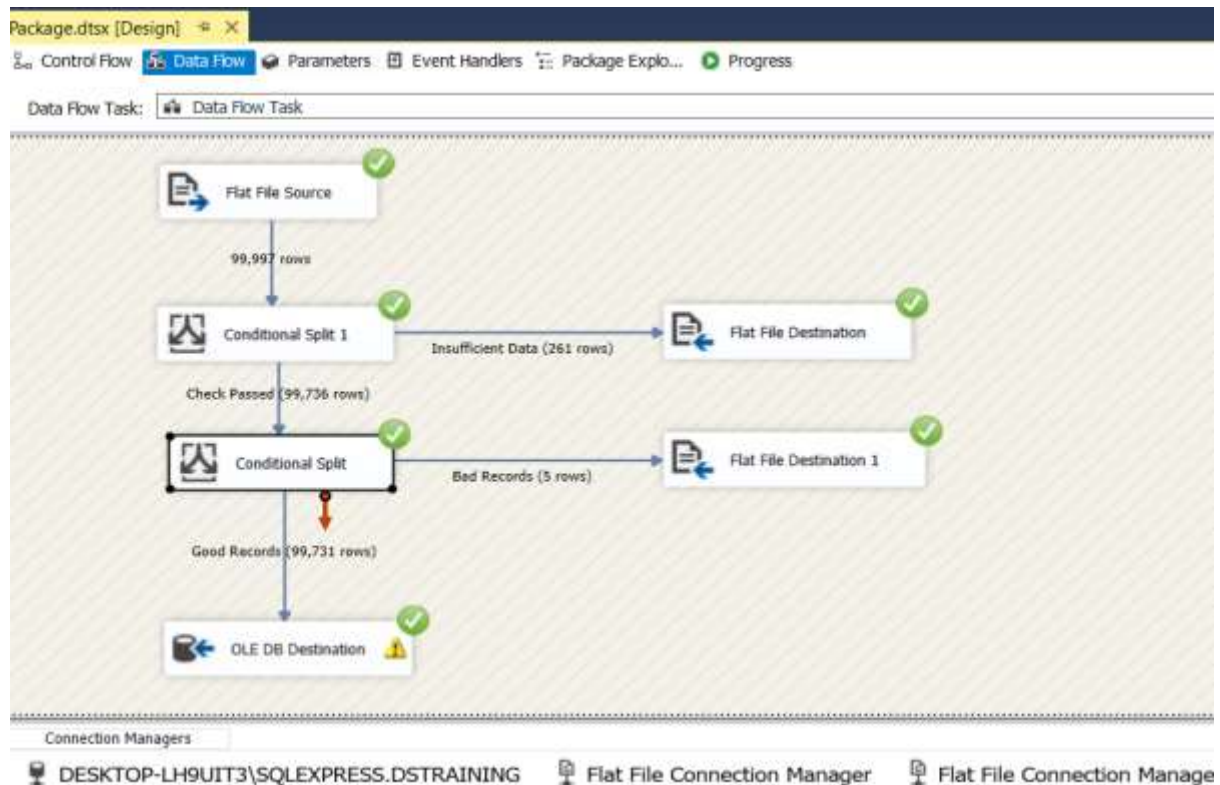
Update the Source and Connection Manager – Double click Flat File Source -> Click Yes -> Re open Flat File Source -> Review columns -> Click OK

Now we are ready to run our package again, save the file before running, now the error we were facing will be filtered out in Bad Record text File.

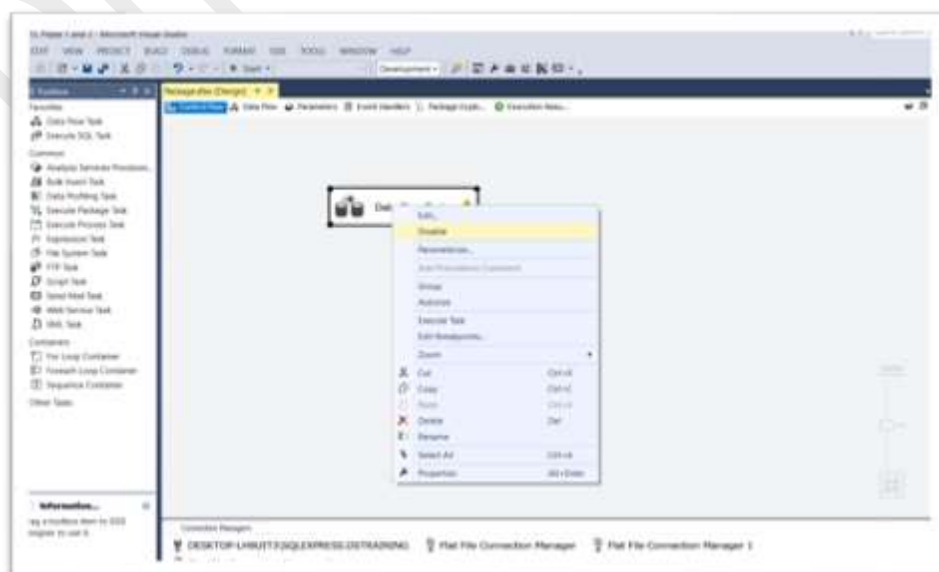
We can also filter that file into some other text file by adding another conditional split and Flat File Destination which is optional.

We can see that 99,731 Rows has been uploaded to Data Base as Good Records.

Now stop the debugging the process by clicking the stop or Ctrl + F5.

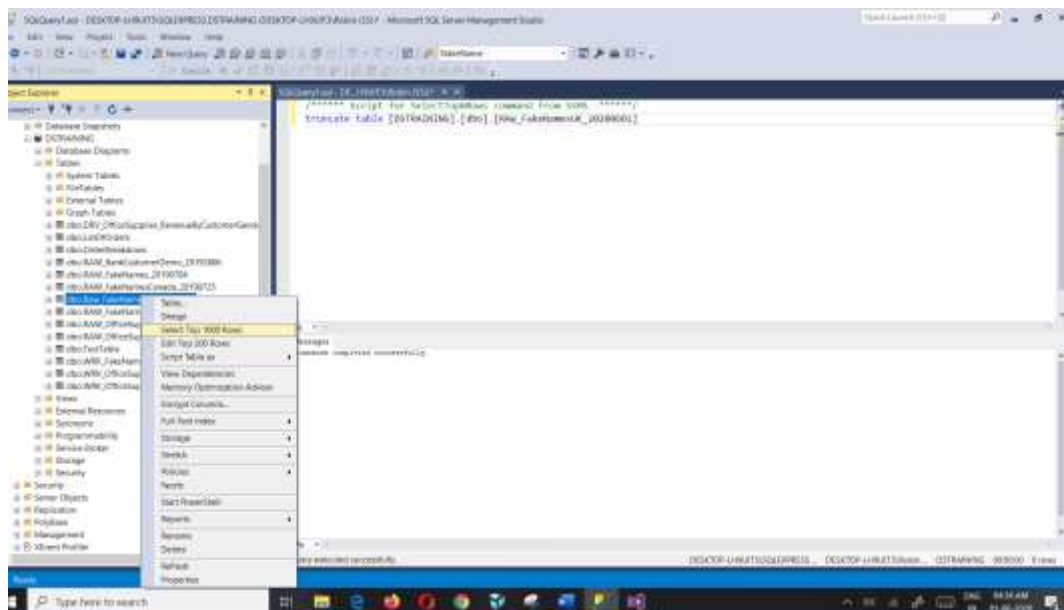


Go to **Control Flow** and Disable the **Data Flow** before exiting MS Visual Studio Program.

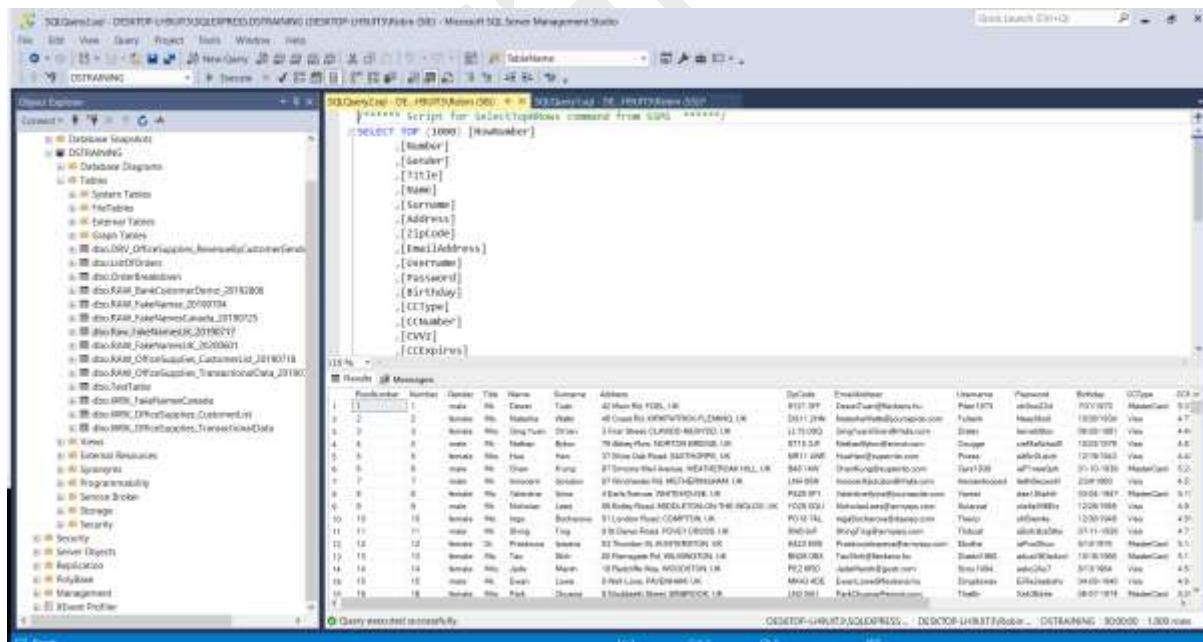


Now to Data Base in **MS SQL Server Management Studio**, refresh tables.

Select our file, Right click and click Select Top 1000 rows.



This is our data in our Data Base which looks good.



We will review our other two text files Bad Records and Insufficient Data, which are in our Analysis folder - > Upload Errors.

In Bad Records file there are few rows which shifted toward right and have lots of blank spaces.



```
1 "484","female","Mr.", "Wei", "Shou", "", "", " UK ", "", "", "", "", "", "", "", "06", "Apr-17", "AB+", "03.4", "168"  
2 "32960","male","Mr.", "Jack", "Humphreys", "14 Crown Street, LONDON, UK ", "SW1P 4BQ", "Jack", "Humphreys@armyspy.com", "Wered76", "cyoxch  
3 "33201","male","Mr.", "Albert", "Bezrukov", "32 Inga Lane, DEAM, UK ", "CA14 0QB", "AlbertBezrukov@teleworn.us", "Unterequan", "zahChai", "  
4 "76920","male","Mr.", "Ethan", "Gould", "73 Thornton St", " HUNTINGFORD", " UK ", "SP0 7BZ", "EthanGould@guatr.com", "Faidi946", "miN6quasT  
5 "85361","female","Mrs.", "Yun", "Wei", "45 Fulford Road, FENFORD, UK "
```

The row which stopped our processing of package have a space of character count 1119 which was more then Address Output Column Width.



```
1  
2  
3  
4  
5 ", "BA43 5XD", "YunWei@armyspy.com", "Tel1993", "iuY2tha
```



Count characters Reset 1119

Summary

In this article, I have shown you the 6 different phases of the Data Analytics Life cycle and its associated processes and tools. Along with it demonstrated the process of Extract Transform and Load, beginning with Data Wrangling where we worked on raw data, transforming it into accurate format using excel, then loading it into MS Visual Studio and Mapping it to get data into MS SQL Data Base without errors along with data anomaly in separate text file.

Dhawal Arora