

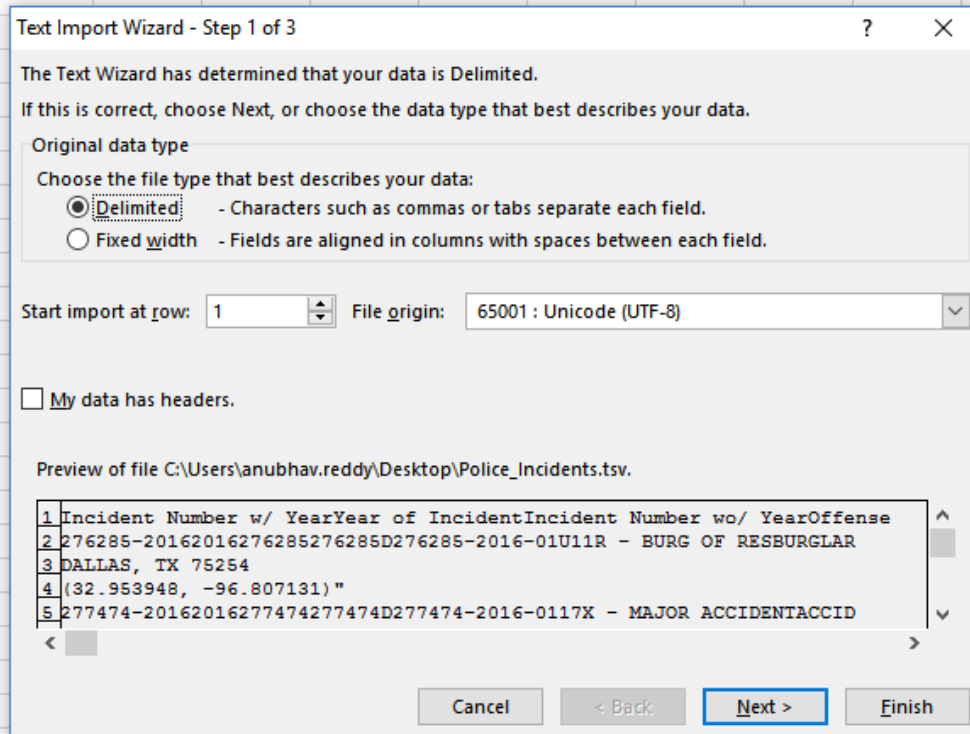
## Steps to clean the police incidents file and load it into hive

Step 1: Download police Incidents data in “.tsv” format from the following link:

<https://www.dallasopendata.com/Public-Safety/Police-Incidents/tbnj-w5hb>

- Go to Downloads and click TSV for Excel

Step 2: Open the TSV file in Excel and you should see the below screen on opening the file



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.

Start import at row: 1 File origin: 65001 : Unicode (UTF-8)

☐ My data has headers.

Preview of file C:\Users\anubhav.reddy\Desktop\Police\_Incidents.tsv.

1	Incident Number w/ YearYear of IncidentIncident Number wo/ YearOffense
2	276285-20162016276285276285D276285-2016-01U11R - BURG OF RESBURGLAR
3	DALLAS, TX 75254
4	(32.953948, -96.807131)"
5	277474-20162016277474277474D277474-2016-0117X - MAJOR ACCIDENTACCID

< >

Cancel < Back Next > Finish

Click Next

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

Incident Number w/ Year	Year of Incident	Incident Number wo/ Year	Offense
276285-2016	2016	276285	276285D
DALLAS, TX 75254			
(32.953948, -96.807131)"			
277474-2016	2016	277474	277474D

< >

Cancel < Back **Next >** Finish

Make sure **Tab** is selected in Delimiters then click **Next**.

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: MDY  
☐ Do not import column (skip)

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

Advanced...

Data preview

General	General	General	General
Incident Number w/ Year	Year of Incident	Incident Number wo/ Year	Offense
276285-2016	2016	276285	276285D
DALLAS, TX 75254			
(32.953948, -96.807131) "			
277474-2016	2016	277474	277474D

Cancel
 < Back
 Next >
 Finish

Click Finish. Once the data is loaded first couple of columns will look like below

<div><div>FILE</div><div>HOME</div><div>INSERT</div><div>PAGE LAYOUT</div><div>FORMULAS</div><div>DATA</div><div>REVIEW</div><div>VIEW</div></div>									
<div><div><div><div></div><div>Paste</div></div><div><div></div><div>Cut</div></div><div><div></div><div>Copy</div></div><div><div></div><div>Format Painter</div></div></div><div>Clipboard</div></div>		<div><div><div><div>Calibri</div><div>11</div><div>A<sup>+</sup> A<sup>-</sup></div></div><div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div></div><div>Font</div></div></div>				<div><div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div><div></div><div></div></div></div><div><div><div></div><div>Wrap Text</div></div><div><div></div><div>Merge &amp; Center</div></div></div><div>Alignment</div></div>			
<div><div>A1</div><div><div></div><div></div><div></div></div><div>Incident Number w/ Year</div></div>									
	A	B	C	D	E	F	G	H	I
1	Incident N	Year of Inc	Incident N	Offense S	Service N	Watch	Call (911)	Type of In	Penalt
2	276285-20	2016	276285	276285D	276285-20	U	11R - BUR	BURGLARY	F2
3	DALLAS, TX 75254								
4	(32.953948, -96.807131)"								
5	277474-20	2016	277474	277474D	277474-20		1 7X - MAJO	ACCIDENT	MB
6	DALLAS, TX 75211								
7	(32.749608, -96.892956)"								
8	276593-20	2016	276593	276593D	276593-20		3 55 - TRAFF	FOUND PF	NA
9	DALLAS, TX 75203								
10	(32.758751, -96.815154)"								
11	276430-20	2016	276430	276430D	276430-20		3 20 - ROBB	ROBBERY	F1
12	DALLAS, TX 75212								
13	(32.778774, -96.858248)"								
14	277049-20	2016	277049	277049D	277049-20		1 PSE/09 - T	THEFT OF	FS
15	DALLAS, TX 75201								
16	(32.794004, -96.80388)"								
17	276500-20	2015	276500	276500C	276500-20		1 24 - ABAN	ABANDON	NA
18	DALLAS, TX 75243								
19	(32.910862, -96.718216)"								
20	276005-20	2016	276005	276005D	276005-20	U	11B - BUR	CRIM MIS	MB
21	DALLAS, TX 75208								
22	(32.74511, -96.830438)"								
23	276614-20	2015	276614	276614C	276614-20		1 44 - PERSC	ABANDON	NA
24	DALLAS, TX 75224								
25	(32.695623, -96.842569)"								
26	276522-20	2016	276522	276522D	276522-20	U	PSE/09 - T	THEFT OF	MB
27	DALLAS, TX 75229								

The geo location data is also present and we need to remove it so, apply filters to all columns. **Make sure you select the entire data (ctrl + A) before applying filters.**

Excel ribbon: FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEWS

Clipboard: Paste, Cut, Copy, Format Painter

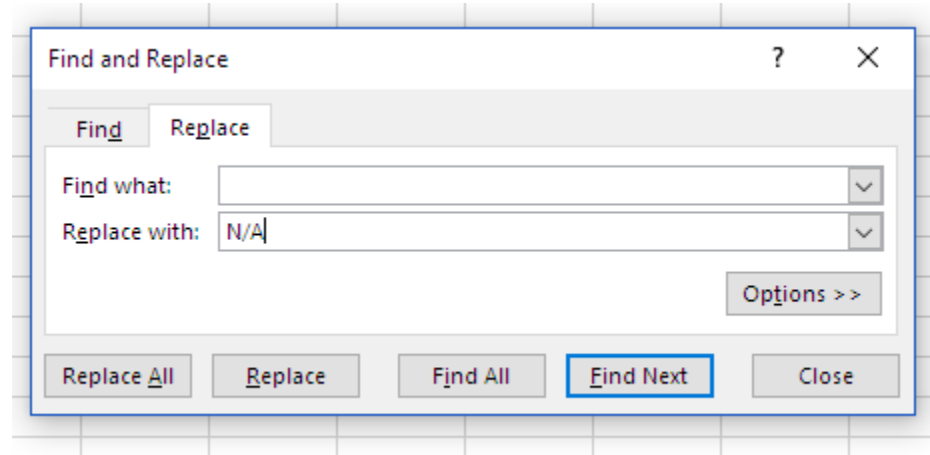
Font: Calibri, 11, Bold, Italic, Underline, Text Color, Background Color, Paragraph

Formulas: fx, 2016

	A	B	C	D	E	F	G
1	Incident	Year of	Incident	Offense	Service	Watch	Call (911)
2	Sort Smallest to Largest				276285-20	U	11R - BUR
3	Sort Largest to Smallest						
4	Sort by Color				277474-20		1 7X - MAJO
5	Clear Filter From "Year of Incident"						
6	Filter by Color						
7	Number Filters				276593-20		3 55 - TRAFF
8	Search						
9	(Select All)				276430-20		3 20 - ROBB
10	2014						
11	2015						
12	2016				277049-20		1 PSE/09 - T
13	2017						
14	(Blanks)				276500-20		1 24 - ABAN
15	OK				276005-20	U	11B - BUR
16	Cancel						
22	(32.74511, -96.830438)"						
23	276614-20	2015	276614	276614C	276614-20		1 44 - PERSC
24	DALLAS, TX 75224						
25	(32.695623, -96.842569)"						
26	276522-20	2016	276522	276522D	276522-20	U	PSE/09 - T
27	DALLAS, TX 75229						
28	(32.898695, -96.876029)"						
29	277418-20	2016	277418	277418D	277418-20		3 41/11V - B

Then select "Year of Incident" and **sort** the data from smallest to largest. Once sorting is finished filter on "blanks" and delete all blank rows. AT this point you should be left with 197191 rows of data. Now

again select entire data (ctrl +A) and find (ctrl + f) go to Replace tab. Fill the blanks as shown.



Click **Replace All**

Step 3. Upload the data to virtual machine / sandbox and select the columns using “awk” statement

Replace the number with column numbers you want to select for analysis

```
awk -F "\t" '{print $1 "\t" $2 "\t" $3 "\t" $8 "\t" $9 "\t" $10 "\t" $17 "\t" $20 "\t" $21 "\t" $22 "\t" $23 "\t" $24 "\t" $25 "\t" $26 "\t" $58 "\t" $59 "\t" $60 "\t" $64 "\t" $81 "\t" $94 "\t" $96 "\t" $97 "\t" $98 "\t" $99 "\t" $101}' Police_Incidents.tsv > Police.tsv
```

Step 4. Load the new file with selected columns into hive

SQL Query to load data into hive

```
CREATE TABLE TABLE NAME (Incident_Number_year VARCHAR(50), X INT, Y VARCHAR,.....)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' lines terminated BY '\n'  
tblproperties("skip.header.line.count"="1");
```

```
LOAD DATA LOCAL INPATH 'PATH TO FILE' OVERWRITE INTO TABLE TABLE NAME;
```