**Fuzzy Grouping Transformation in SSIS**

# Fuzzy Lookup Transformation in SSIS

# Term Lookup Transformation in SSIS

# Term Extraction in SSIS

**Fuzzy Grouping Transformation in SSIS**

The Fuzzy Grouping Transformation in SSIS is used to replace the wrongly typed words with correct words.

SSIS Fuzzy Grouping Transformation does not require any reference table to correct the data. It will use the grouping technique to check for the wrongly typed words (type mistakes) and correct them.

For example, if 98 people typed the country name as India and two people types as Indi, then SSIS Fuzzy Grouping Transformation will replace the Indi with India.

Before we start configuring the SSIS Fuzzy Grouping Transformation, let us see some important properties of this:

* This transformation uses the Grouping technique to replace the wrong word in source data with the correct word.
* This transformation allows us to use only DT\_WSTR and DT\_STR Data type column for Fuzzy Matching, and Exact matching can apply to any data type except DT\_TEXT, DT\_NTEXT, and DT\_IMAGE.
* It creates temporary tables and indexes in the SQL Server database at runtime.
* This transformation will only use the OLE DB Connection Manager to establish a connection to store the temporary tables and indexes.
* To configure the transformation, you must select the Match Type (Fuzzy or Exact) for an input columns

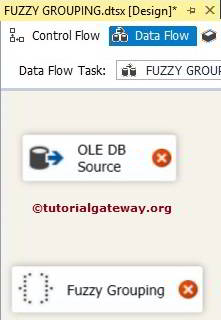


**STEP 1:** Open BIDS and Drag and drop the data flow task from the toolbox to control flow and rename it as Fuzzy Grouping Transformation in SSIS.

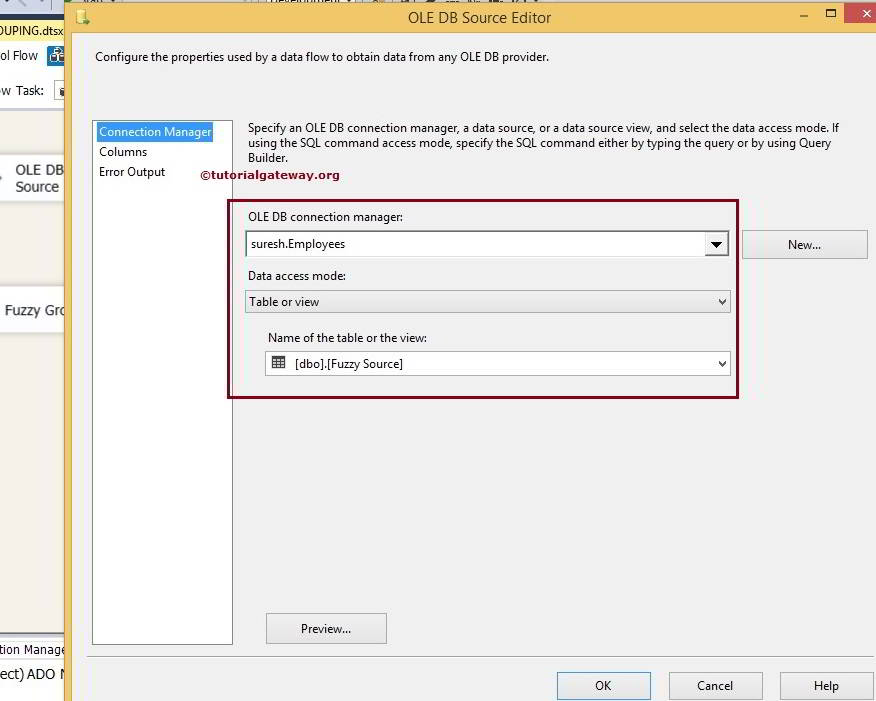


Double click on it, and it will open the data flow tab.

**STEP 2:** Drag and drop OLE DB Source, Fuzzy Grouping transformation from the SSIS toolbox to the data flow region

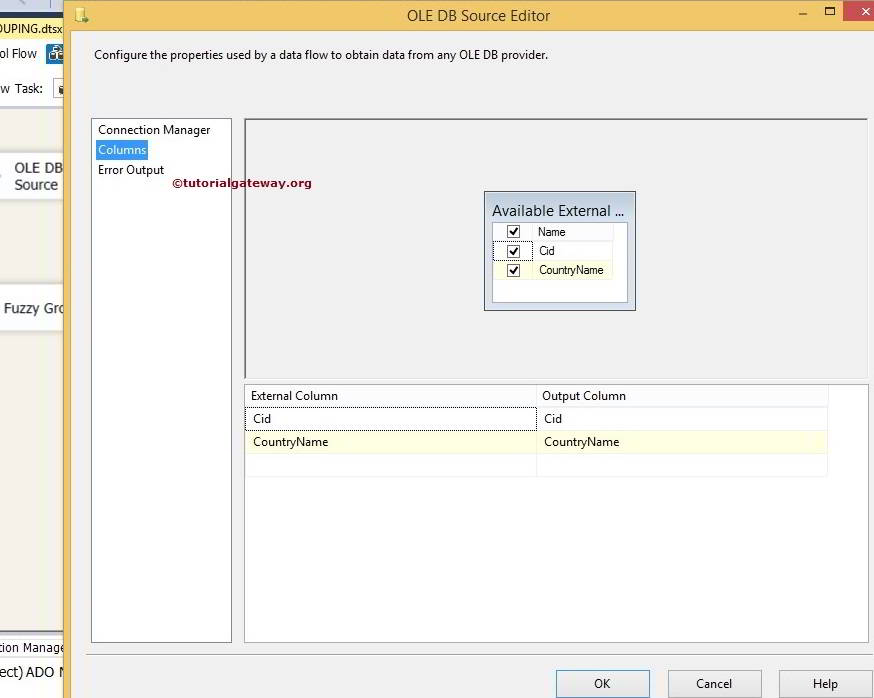


**STEP 3:** Double click on OLE DB source in the data flow region will open the connection manager settings and provides an option to select SQL Table or space to write our SQL statement.



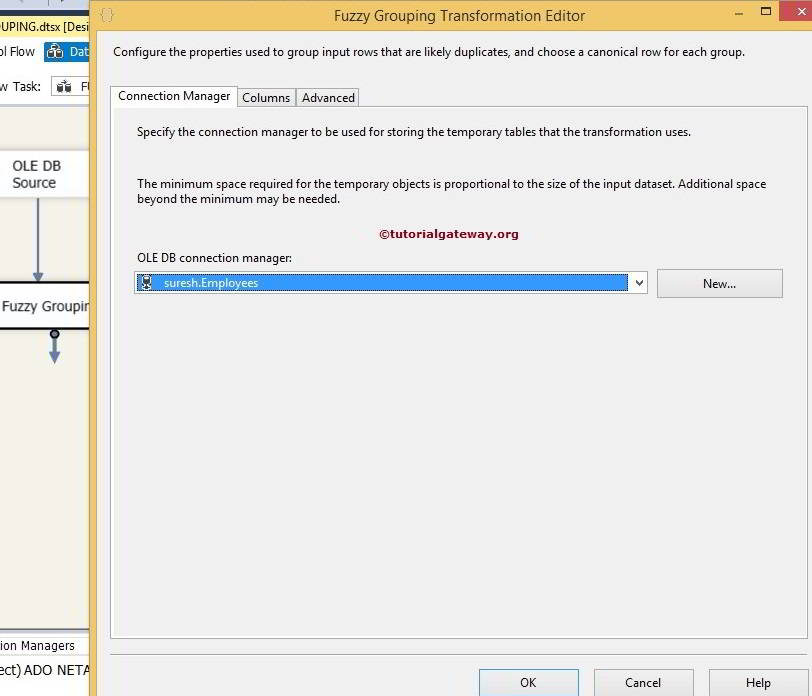
Here we selected the Employees database as our source database and [Fuzzy Source] as SQL table.

**STEP 4:** Click on the columns tab to verify the columns. In this tab, we can uncheck the unwanted columns.

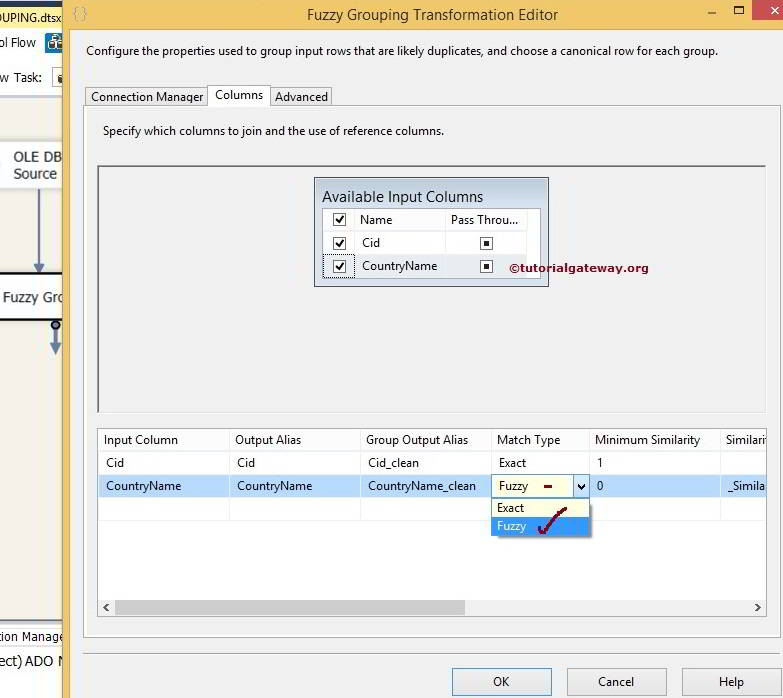


**STEP 5:** Right-click on the SSIS Fuzzy Grouping Transformation will open the Fuzzy Grouping Transformation Editor window to configure it.

Within the Connection Manager tab, we have to configure the OLE DB connection Manager setting. Within this source, SSIS Fuzzy Grouping Transformation will create a temporary table and indexes to perform the Fuzzy Grouping operation.



**STEP 6:** Within the Columns Tab, We have to configure the match Type. You will need to select the columns that you want to group as Fuzzy Match Type and other columns as Exact match. For this example, the Country name is a string, and we want to find the wrong values in this column. So, change the match type to Fuzzy, and Cid is the Int value, so the match type is Exact.

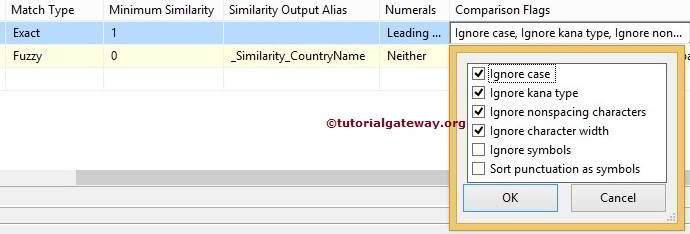


Two parameters are important when you are Configuring a fuzzy grouping Transformation. They are Numerals and Comparison Flags.

#### Comparison Flags

From the below screenshot, you can see the available options in this.

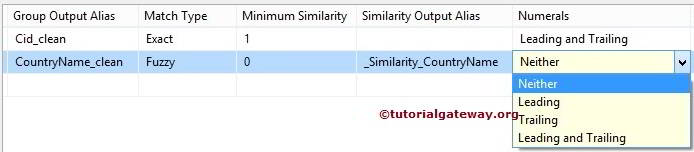
* **Ignore case:**If we check mark this option, Fuzzy Grouping will ignore the case. Both XYZ and xyz will be the same.
* **Ignore kana type:** This option of SSIS Fuzzy Grouping ignores the difference between the Japanese hiragana and katakana.
* **Ignore non spacing characters:**If we check mark this option then Fuzzy Grouping will ignore the difference between the diacritics and character
* **Ignore character width:** If we checkmark this option, SSIS Fuzzy Grouping ignores the difference between single-byte character and double-byte character.
* **Ignore symbols:**SSIS Fuzzy Grouping will ignore the difference between the letters and symbols (white spaces, punctuations, currency symbols, and mathematical symbols). For example, \*xy treated the same as xy
* **Sort punctuation as symbols:** Fuzzy Grouping will all the punctuation symbols (except apostrophe and hyphen) write before the letters. For example, .xyz will sort before the xyz



#### ****Numerals****

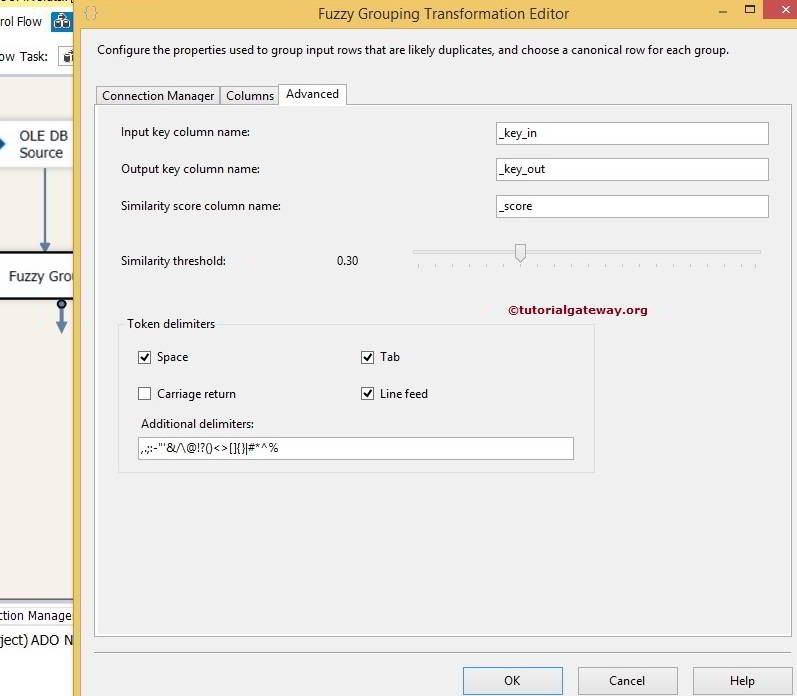
In this option, We have to specify the significance of starting and ending numerals while comparing the column data. For example, if leading numerals are significant, “93 New lands Street” will not be grouped with “99 New lands Street”

|  |  |
| --- | --- |
| **VALUE** | **DESCRIPTION** |
| Neither | Numerals at the starting and ending positions ignored while grouping. |
| Leading | Numerals at the starting position will ignore while grouping. |
| Trailing | Numerals at the ending positions ignored while grouping. |
| LeadingAndTrailing | Numerals at the starting and ending positions will not ignore while grouping. |



**STEP 7:**Within the Advanced Tab, we have to configure the Similarity Threshold. The similarity threshold ranges between 0 and 1, where 1 is an exact match. The SSIS Fuzzy Grouping Transformation Editor provides a slider to adjust the similarity between 0 and 1. If the similarity threshold is closer to 1, the source column should match more accurately to reference data.

In this example, We are doing Fuzzy Grouping on Country Name and find the fuzzy match. If we give a Similarity threshold as 0.76, the string column values should match more than 76%, then only it will treat as a valid record.

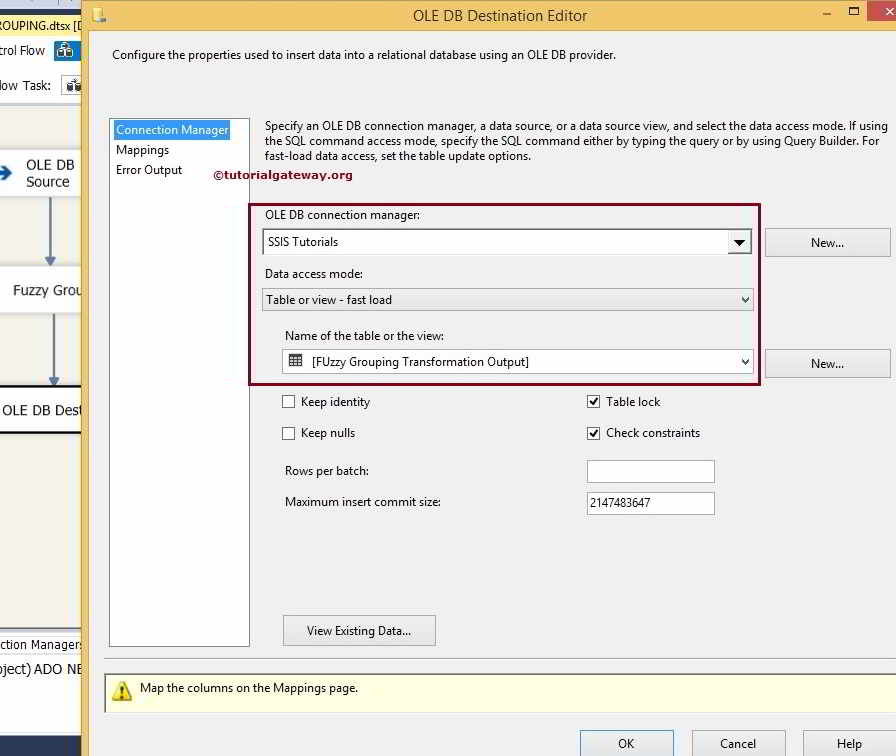


The SSIS Fuzzy Grouping Transformation produces additional columns along with existing columns, and they are:

* **Input key column name:**Provide the Unique name for Input Key Column. **key\_in**is the default name.
* **Output key column name:** Provide the Unique name for Output Key Column. **key\_Out**is the default name.
* **Similarity score column name:**Provide the Unique name for the similarity score column. **\_score** is a value between 0 and 1. It will indicate the similarity of the input row to the canonical row.

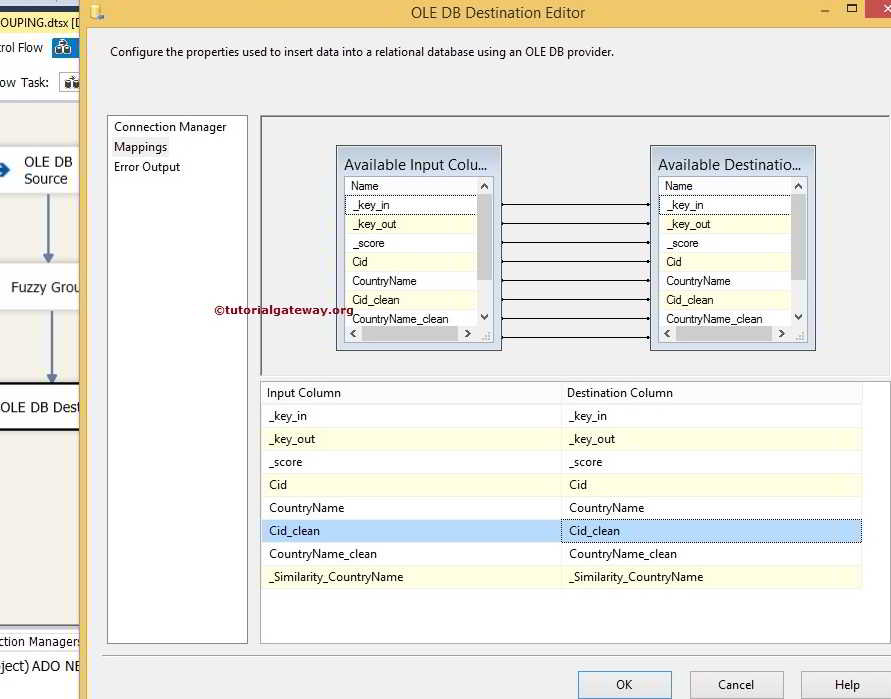
For the time being, we left them to default values and Click ok to finish configuring the SSIS Fuzzy Grouping Transformation. Now drag and drop the OLE DB destination into the data flow region.

**STEP 8:** Now, we have to provide the Server, database, and table details of the destination. So double-click on the OLE DB Destination and provide the required information



Here we selected the [SSIS Tutorials] database from SQL Server as our destination database and [Fuzzy Grouping Transformation Output] table as our destination table.

**STEP 9:** Click on the Mappings tab to check whether the source columns exactly mapped to the destination columns.



**STEP 10:** Click ok to finish designing the SSIS Fuzzy Grouping Transformation package. Let us run the package



# Fuzzy Lookup Transformation in SSIS

The Fuzzy Lookup Transformation in SSIS is used to replace the wrongly typed words with correct words.

Fuzzy Lookup transformation in SSIS uses fuzzy matching to find one or more close matches in the reference table and replace the source data with reference data.

The Fuzzy Lookup Transformation in SSIS is an important transformation in real-time. For example, while entering the product information, sometimes, we may enter the data with spelling mistakes. While doing Lookup Transformation, due to these wrongly typed words, we can’t match the source data with a lookup table. In these situations, SSIS Fuzzy Lookup Transformation will look for the nearest matching right word and replace the wrong value with the correct word. Please remember the following points before working with SSIS Fuzzy lookup

* SSIS Fuzzy Lookup Transformation only use OLE DB Connection Manager to establish a connection with the Reference table
* This transformation uses the Reference table to compare the wrong word in source data with the correct word in the reference table.
* SSIS Fuzzy Lookup allows us to use only DT\_WSTR and DT\_STR Data type column for Fuzzy Matching.
* This transformation creates temporary tables and indexes in the SQL Server database at runtime

## Fuzzy Lookup Transformation in SSIS Example

In this example, we are going to show how this SSIS Fuzzy lookup transformation works on the source data. The below screenshot is our source data



Fuzzy Transformation Lookup reference table is

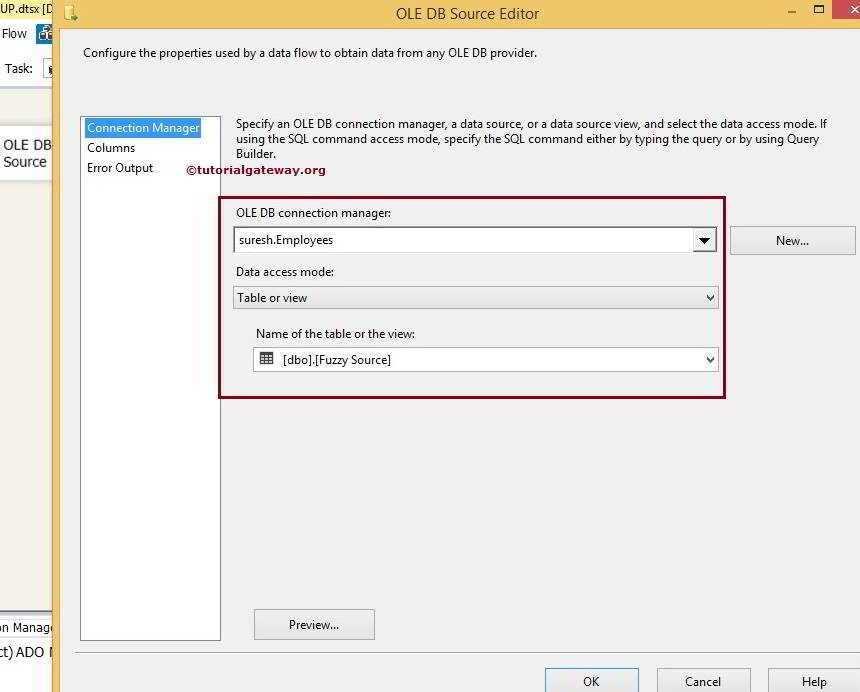


**STEP 1:** Open BIDS and Drag and drop the data flow task from the toolbox to control flow. Next, rename it as Fuzzy Lookup Transformation in SSIS.



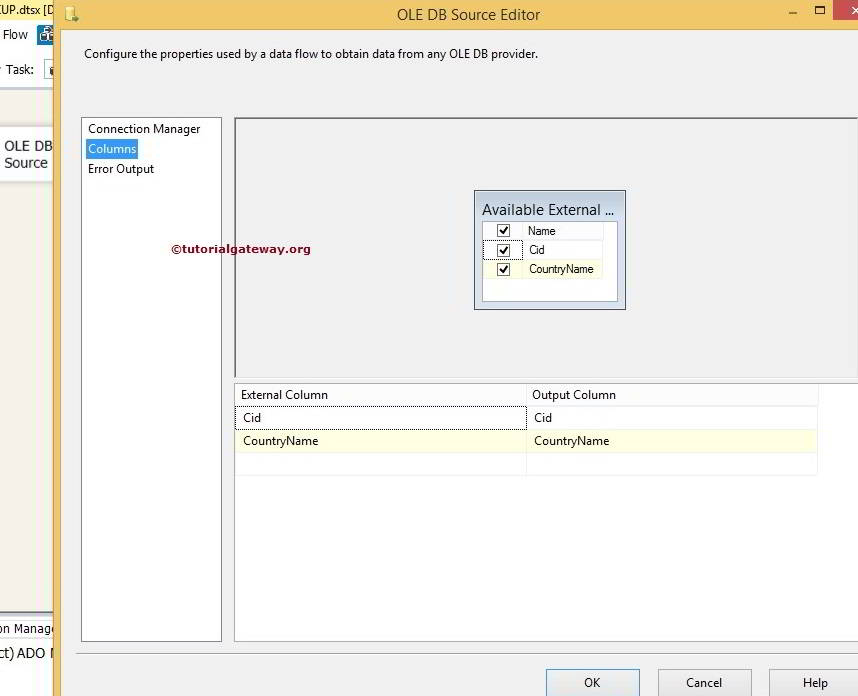
Double click on it and it will open the data flow tab.

**STEP 2:** Drag and drop OLE DB Source from toolbox to data flow region. Double click on OLE DB source in the data flow region will open the connection manager settings and provides an option to select SQL Table or SQL statement.



Here we selected the Employees database and previously shown Fuzzy Source table as our OLE DB source database and table.

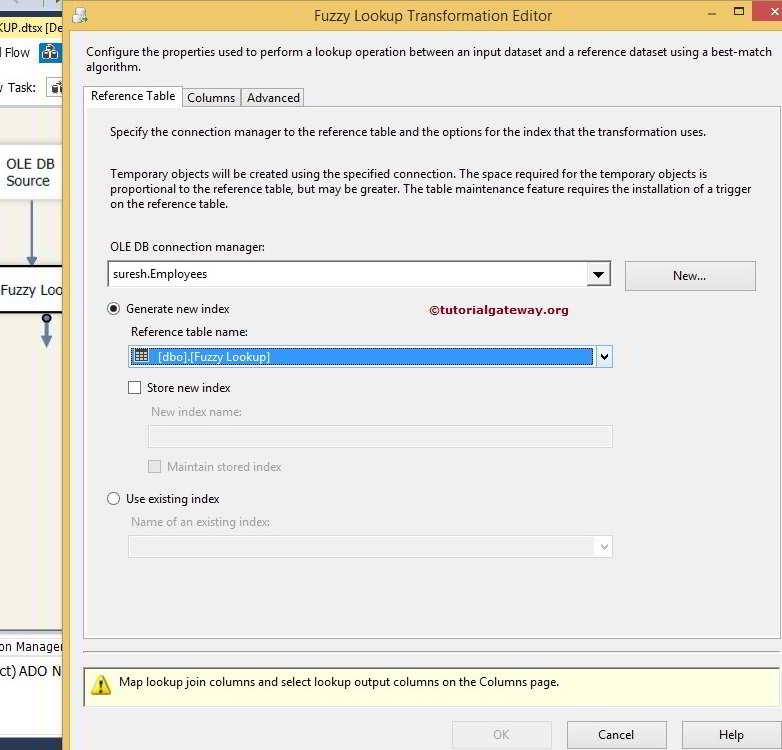
**STEP 3:** Click on the columns tab to verify the columns. In this tab, we can uncheck the unwanted columns also.



Drag and drop the Fuzzy Lookup Transformation in the SSIS toolbox to the data flow region and connect the OLE DB source output arrow to it.

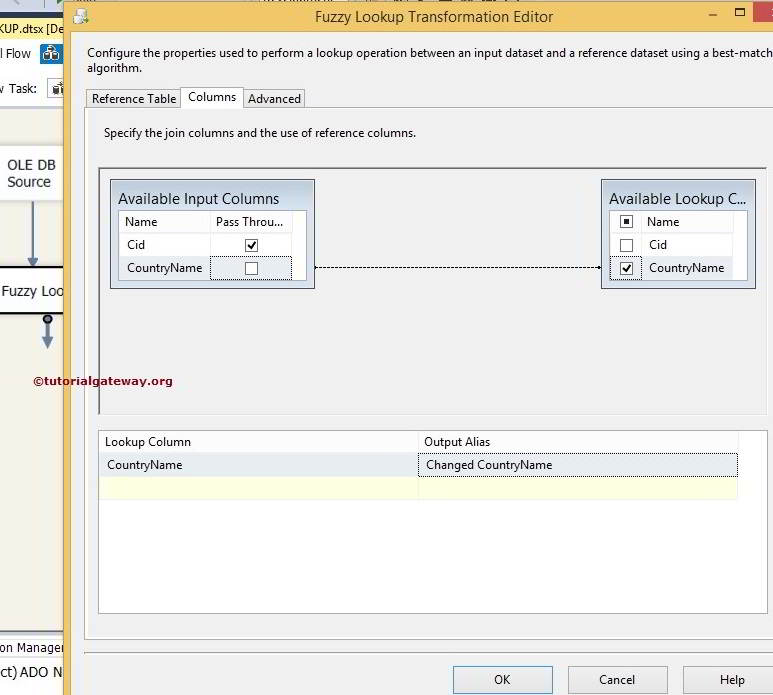
**STEP 4:** Double click on the SSIS Fuzzy Lookup Transformation will open the Fuzzy Lookup Transformation Editor. In the reference table tab, we have to configure the connection manager and also select the fuzzy lookup table. For this, this tab has multiple options.

* **OLE DB Connection Manager:** SSIS Fuzzy Lookup Transformation only supports OLE DB Connection Manager to establish a connection with the Reference table. So, create a connection using the OLE DB Connection Manager.
* **Generate new index:**SSIS Fuzzy Lookup transformation will generate a new index on the referenced table to use for the Fuzzy lookup.
* **Reference table name:**From the drop-down list select the reference table from the OLE DB connection
* **Store new index:**If you want to save the new index on the Fuzzy lookup table, select this option.
* **New index name:**If you have chosen the **Store New index**option, write the index name here.
* **Maintain stored index:** If you want the SQL Server to maintain the newly created index using the **Store New index**option, checkmark this option.
* **Use existing index:**If you want to use the existing (already created) index on the fuzzy lookup table then select this option
* **Name of an existing index:**Please select the existing index from the drop-down list



From the above screenshot, you can observe that we configured the OLE DB connection manager to Employees database and then selected the Fuzzy Lookup Table as a reference table.

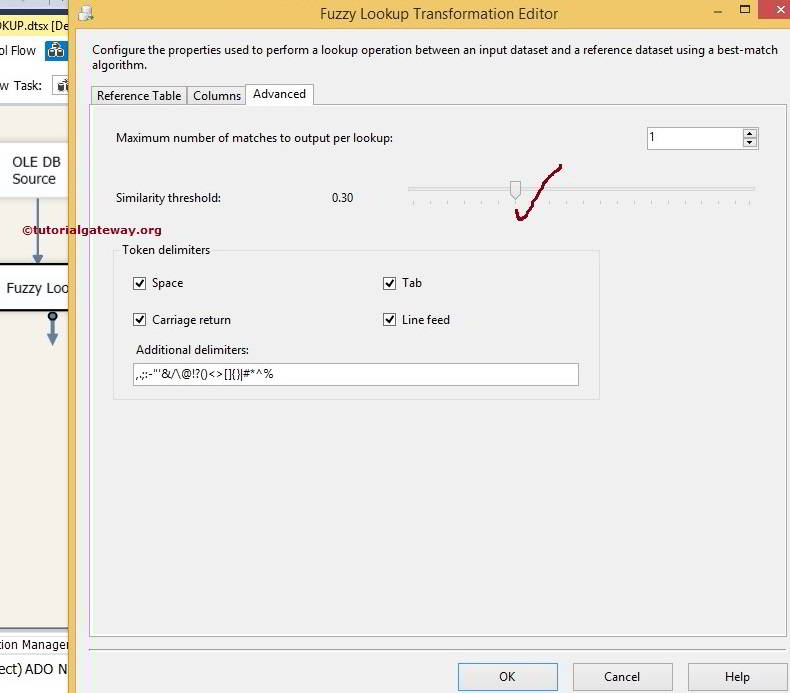
**STEP 5:** Checkmark the columns if you wish to pass through and Join the source column (Country Name) to the reference table column name (again Country Name in this example). If you want to change the output column, then change the alias name as we turned in the below screenshot



If you required both changed and unchanged country name columns in the destination folder, checkmark the pass-through option country name column

**STEP 6:** In the advanced tab, we can configure the similarity threshold by dragging the bar pointer to the required limit. Available options in the Advanced tab are as follows:

* **Maximum number of matches to output per lookup:**Please specify the Maximum number of matches the SSIS Fuzzy Lookup Transformation may return. The default is 1.
* **Similarity threshold:**The similarity threshold ranges between 0 and 1, where 1 is an exact match. The SSIS Fuzzy Lookup Transformation Editor provides a slider to adjust the similarity between 0 and 1. If the similarity threshold is closer to 1, the source column should match more accurately to reference data.

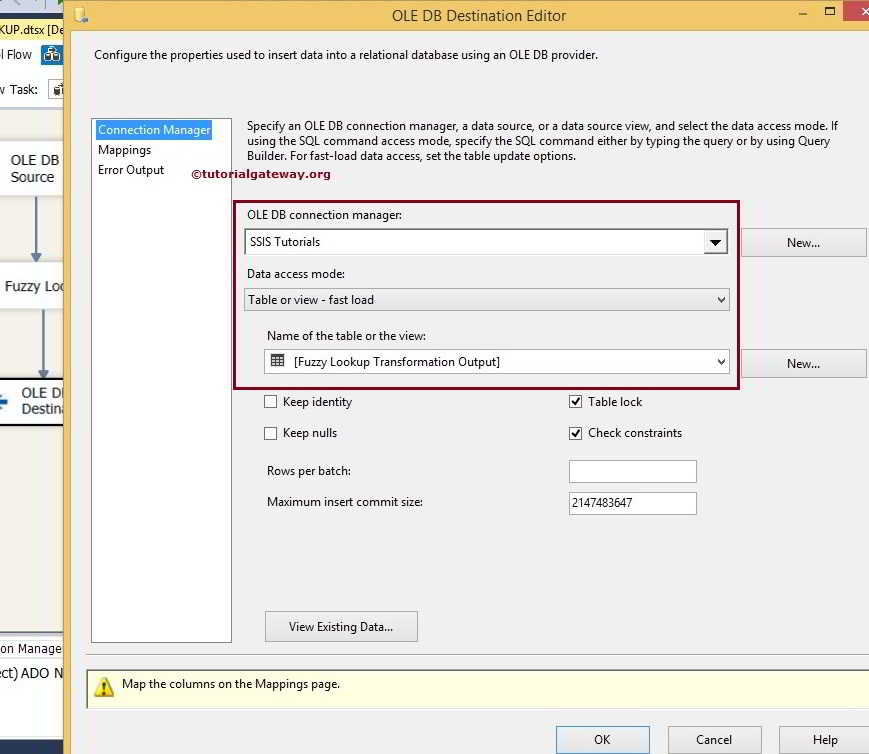


Click ok to finish configuring the Fuzzy lookup Transformation in SSIS. Now drag and drop the OLE DB destination into the data flow region.

The SSIS Fuzzy lookup Transformation Editor generated two extra columns along with the existing columns

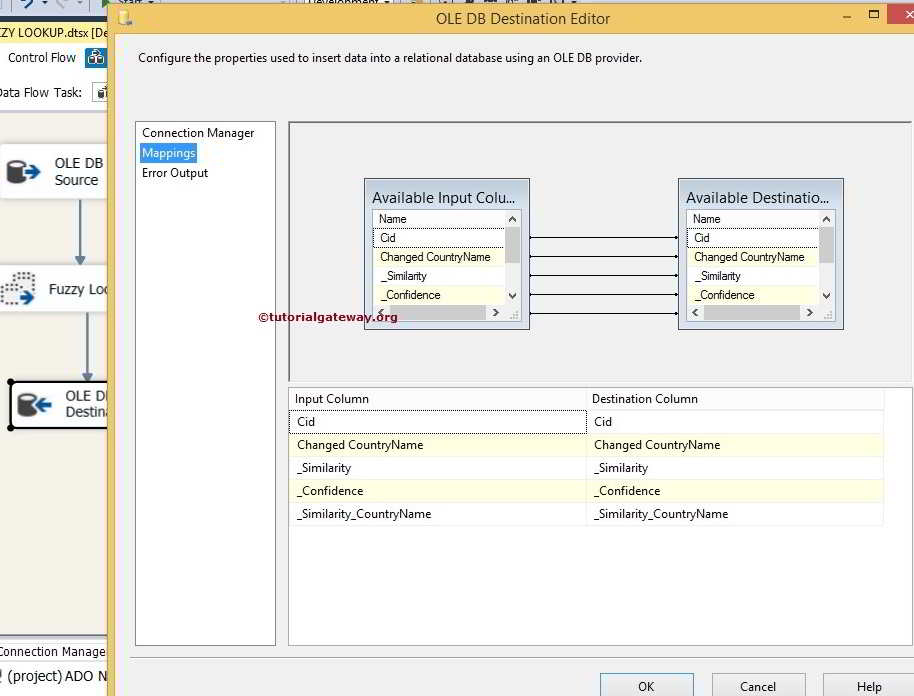
* **\_Similarity:**The similarity score is between 0 and 1. It will show how accurate the source data matched with the fuzzy lookup data. For instance, 0.80 means source data is 80% matched with a destination.
* **\_Confidence:**Confidence Score shows how much confident Fuzzy lookup Transformation is about the best match in the fuzzy lookup table. For instance, 0.50 means Fuzzy lookup Transformation is 50% sure that Australia is the best match for Australi.

**STEP 7:** Now, we have to provide the Server, database, and table details of the destination. So double click on the OLE DB Destination and provide the required information

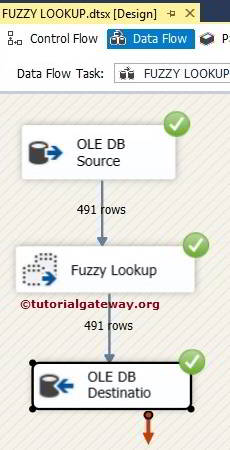


Here we selected the [SSIS Tutorials] database as our destination database and selected the destination table.

**STEP 8:** Click on the Mappings tab to check whether the SSIS Fuzzy lookup source columns exactly mapped to the destination columns.



Click ok to finish designing the Fuzzy Lookup transformation in SSIS package. Let us run the package



Let’s open the SQL Server Management Studio and select the required table to check the results

SELECT [Cid]

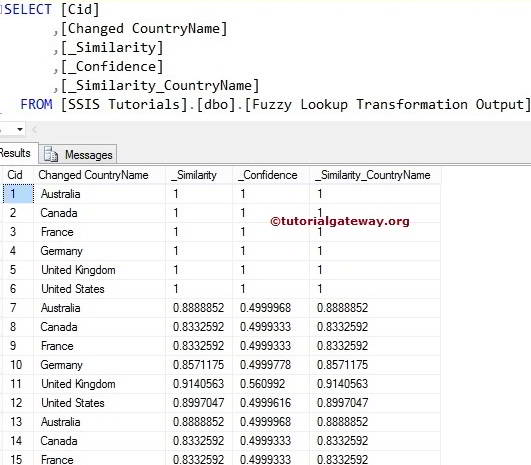
,[Changed CountryName]

,[\_Similarity]

,[\_Confidence]

,[\_Similarity\_CountryName]

FROM [SSIS Tutorials].[dbo].[Fuzzy Lookup Transformation Output]



# Term Lookup Transformation in SSIS

The Term Lookup Transformation in SSIS is used to find terms within the string or text. It uses the reference table (contains terms) to find the matched sentences from the source data and also counts the number of times a term repeated in the source data row.

The SSIS Term Lookup transformation adds two extra columns to the transformation output called **Term** and **Frequency** by default. SSIS Term contains a term from the lookup table, and Frequency includes the number of times the phrase is repeated or occurred in the input data set.

For example, if your company has 100 products and it is allowing users to write reviews on your products. Imagine you are getting 2000 reviews per day then, how will you find which review belongs to which product? To resolve these types of issues, SSIS introduced the Term Lookup Transformation. Using this, you can find which row has product 1 review and which row has product 90 review.

### Term Lookup Transformation Matching Techniques

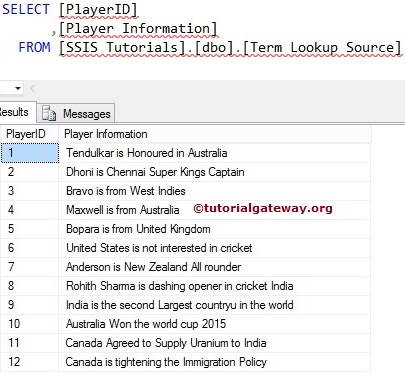
The SSIS Term Lookup Transformation uses the following rules to perform Term lookup on the source Data:

* If the reference table contains a singular form of the word, SSIS Term Lookup Transformation considers both the singular and plural forms of the word in the source data as a singular form. For example, if the term lookup table contains Employee and we have Employee and Employees terms in source data, SSIS Term Lookup Transformation would count both of them as a match for the lookup term
* If the reference table contains a plural form of the word, Term Lookup Transformation considers both the singular and plural forms of the word in the source data as separate Terms. For example, if the term lookup table contains Employees and we have Employee and Employees terms in source data, Term Lookup Transformation would count Employee separately and Employees separately.
* If we configured the SSIS Term Lookup Transformation to perform case-sensitive matches, the Employee is treated separately from the EMPLOYEE. If we have Employee as a first word in the sentence, then Term Lookup Transformation will count Employee and Employee as the same term.

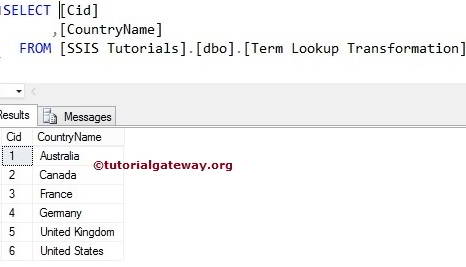
## Term Lookup Transformation in SSIS Example

In this example, we are going to show you how to configure Term Lookup Transformation in SSIS with sample data.

Here is our Lookup Source table



Here is our lookup table from the [SSIS Tutorials] Database.

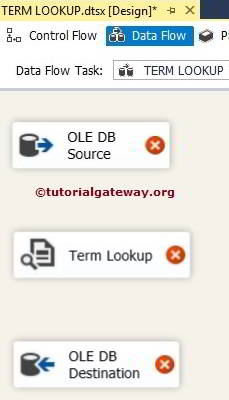


**STEP 1:** Create a new project in BIDS with the Integration Service Project Template. Drag the Data flow task to the Control flow region and rename it as per your requirements.

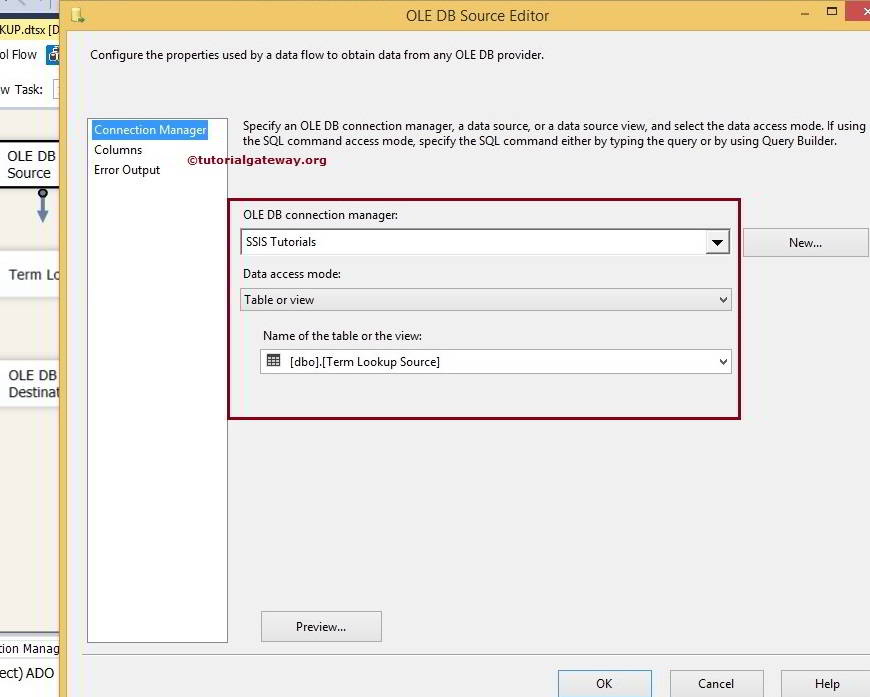


Double click on it, and it will open the data flow tab.

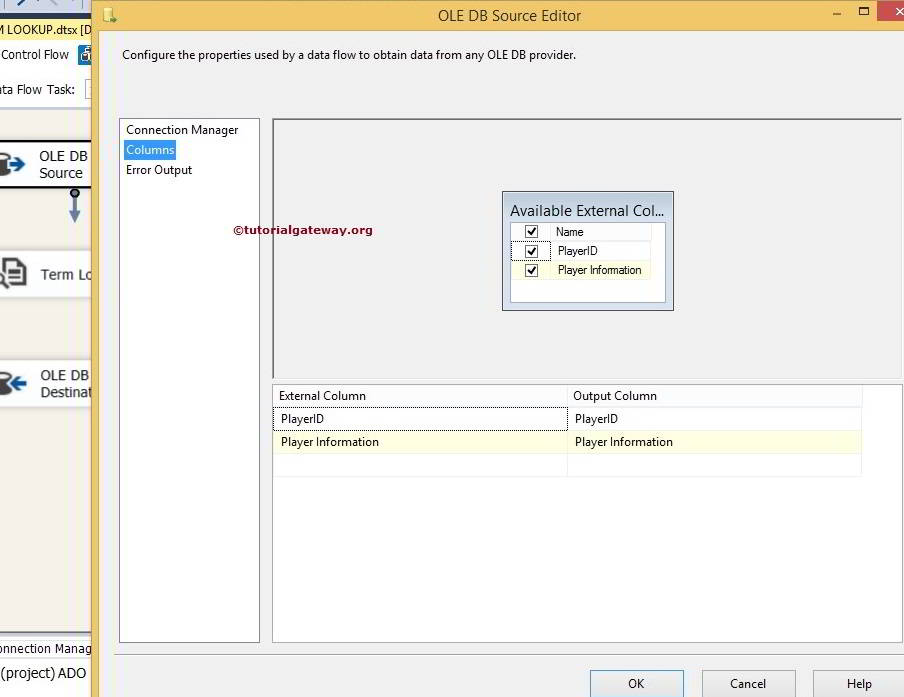
**STEP 2:** Drag and drop OLE DB Source, Term Lookup Transformation and OLE DB destination to data flow region



**STEP 3:** Double click on OLE DB source in the data flow region will open the connection manager settings. It provides space to write our SQL statement. Or we can select the table from the Dropdown list.



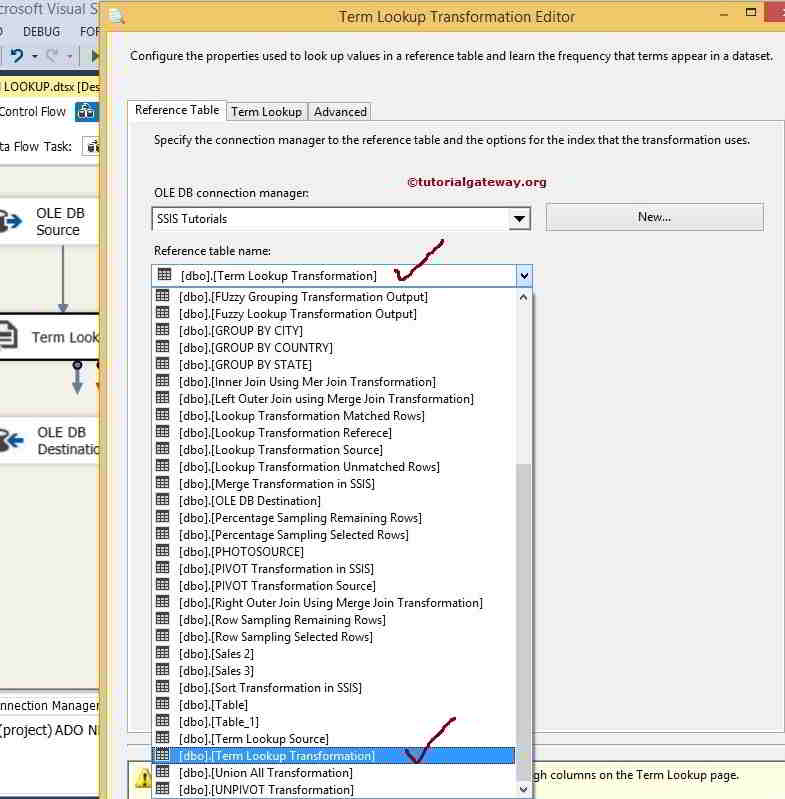
**STEP 4:** Click on the columns tab to verify the columns. In this tab, we can uncheck the unwanted columns.



**STEP 5:** Click ok and edit the SSIS Term Lookup Transformation

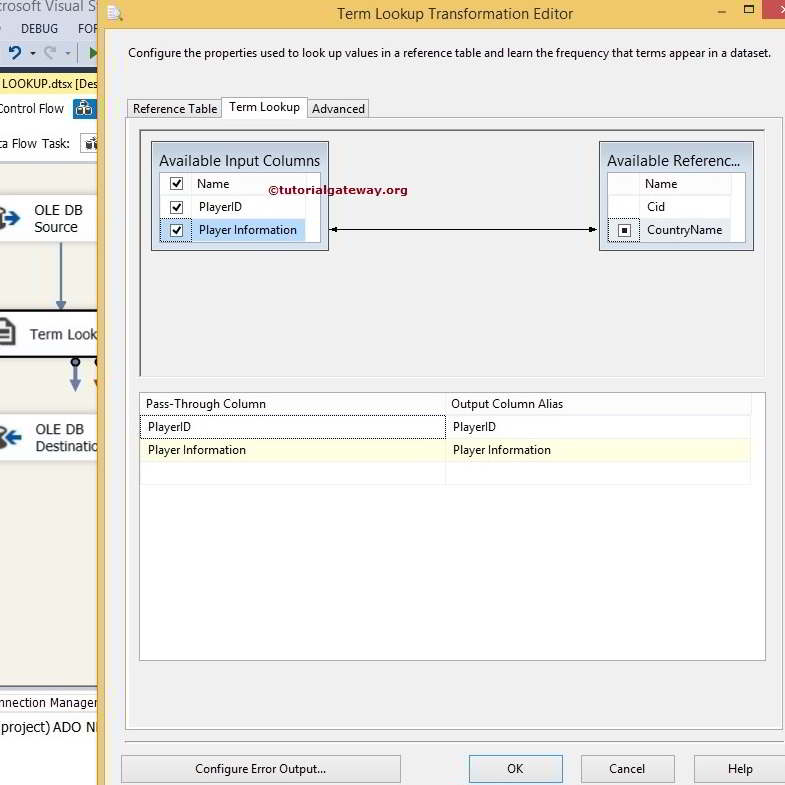
In the reference table tab, Configure the OLE DB connection manager to reference database (Here it is [SSIS Tutorials]) and then select the table you wish to add as reference table (Here it is Term Lookup Transformation table).

**TIP:** This SSIS Term Lookup Transformation only supports OLE DB Connection Manager to connect with the reference table.

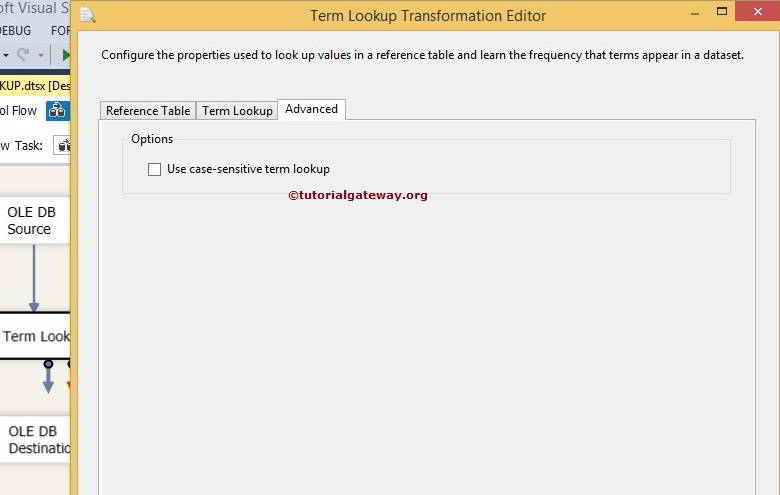


**NOTE:** This Transformation will accept only Unicode String [DT\_WSTR] or Unicode text stream [DT\_NTEXT] as its data type.

**Step 6:** Use the Term Lookup tab to map an input column to a lookup column in a reference table and to provide an alias for each output column.

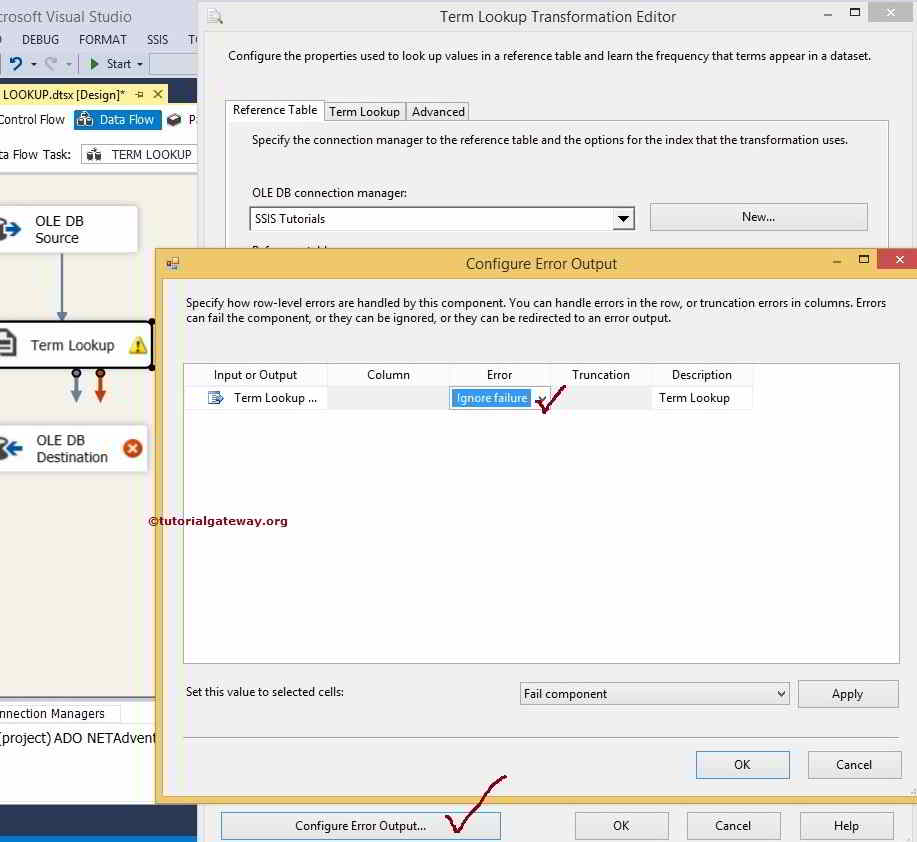


Use the Advanced tab of the Transformation Editor Dialog box to specify whether the lookup should be case-sensitive or not.



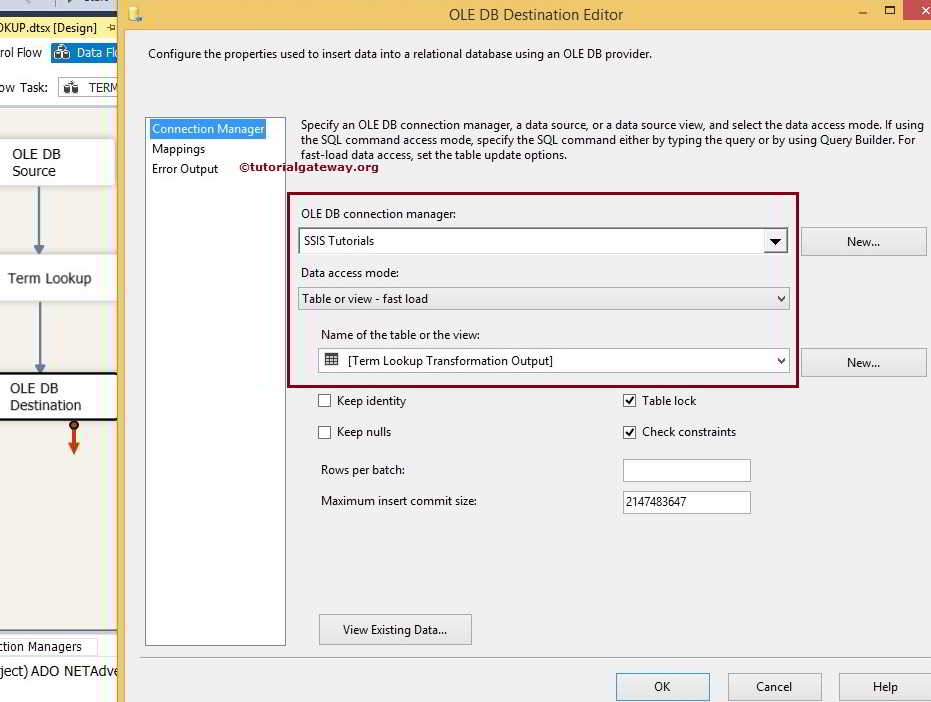
From the below screenshot, you can see there is a warning symbol on the Transformation. And it is telling that error output is not connected. You can remove the warning symbol by configuring the error output. So double click on the *Configure Error Output*button will open a new window to set the error output.

The default configuration of this Transformation is to redirect error rows. You can get rid of this warning by connecting the error output, or by changing the default behavior to Ignore Failure or Fail Component. Let me change to Ignore Failure



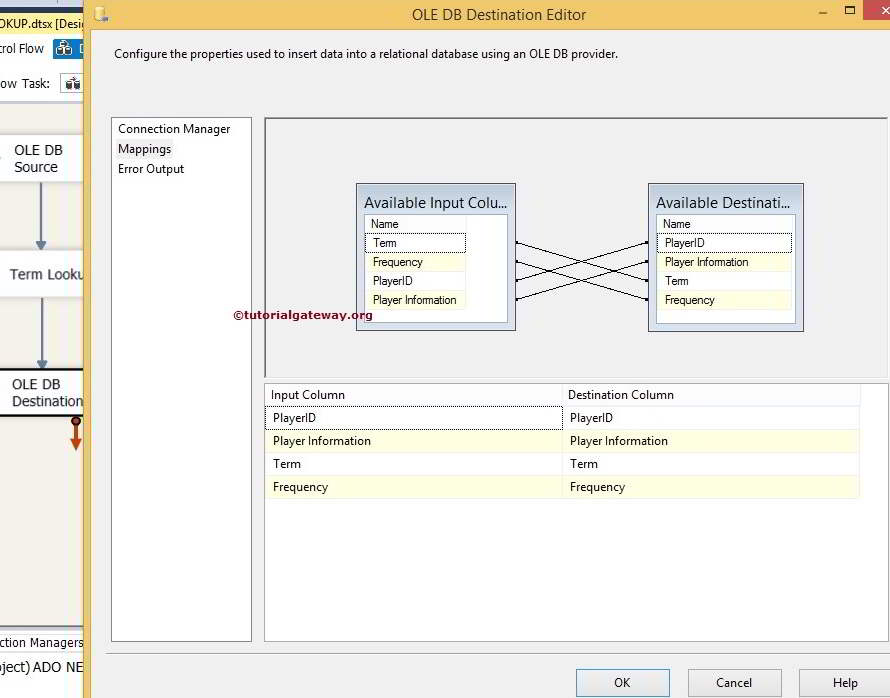
Click ok to finish configuring the Transformation.

**STEP 7:** Now, we have to provide the Server, database, and table details of the destination. So double click on the OLE DB Destination and provide the required information

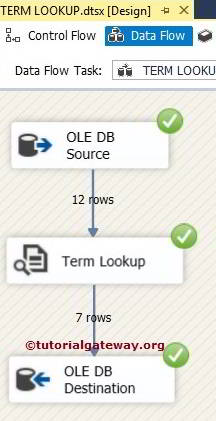


From the above screenshot, see that we selected the [SSIS Tutorials] database as our destination database and [Term Lookup Transformation Output] table as destination table.

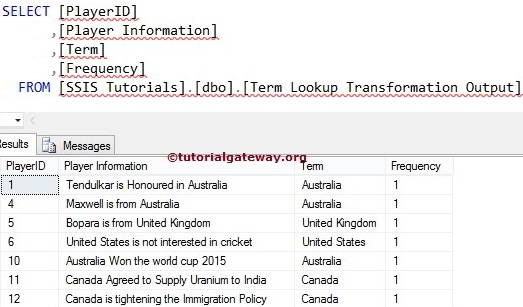
**STEP 8:** Click on the Mappings tab to check whether the source columns exactly mapped to the destination columns.



Click ok to finish designing our SSIS Term Lookup Transformation package. Let us run the package



Let’s open the SQL Server Management Studio and check the results



From the above screenshot, you may see, this transformation is displaying the Term and Frequency (number of times term is repeating) of the Term in a sentence.

# Term Extraction in SSIS

The Term Extraction transformation in SSIS first extracts terms from the text present in the source data and then writes the extracted terms to a Transformation output column.

For example, people are writing reviews on your products, and you want to contact them for further assistance. In these situations, you can use SSIS Term Extraction transformation to extract the email address and name of the user from the reviews.

**NOTE:** The SSIS Term Extraction transformation uses its own English dictionary and linguistic setting to extract the Term from the source data.

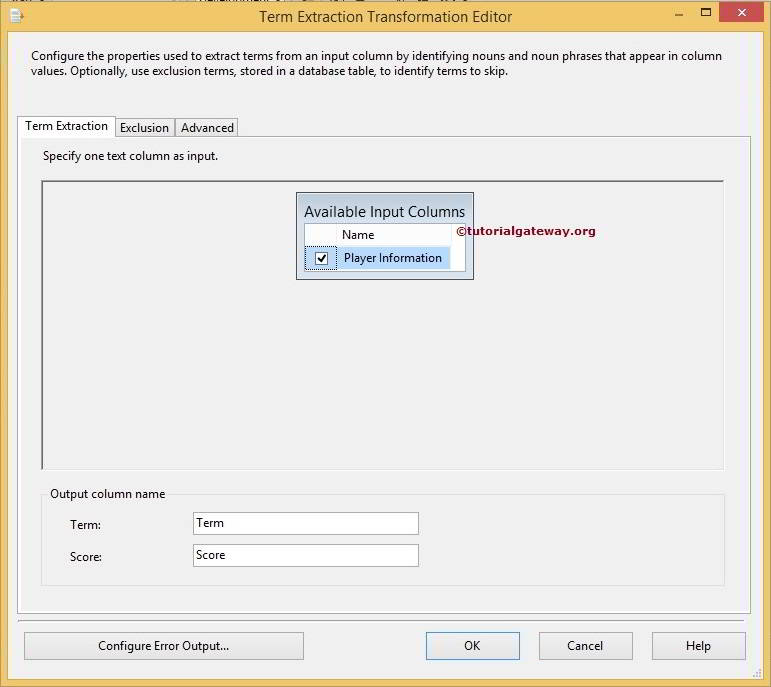
In SSIS, We can perform Term Extraction Transformations only on the column with the DT\_WSTR and DT\_NTEXT data type. If your input column is different from these two then Please use the SSIS Data Conversion to convert other Data Types to DT\_WSTR and DT\_NTEXT data type

## Steps involved in configuring Term Extraction in SSIS

When you double-click on this transformation, an SSIS Term Extraction Transformation Editor window will open to configure it. It contains three Tabs, such as **Term Extraction**, **Exclusion**, and **Advanced** Tab.

### Term Extraction Tab

Within the Term Extraction tab, We have to select the column name of the Source data from the **Available Input Columns** option.



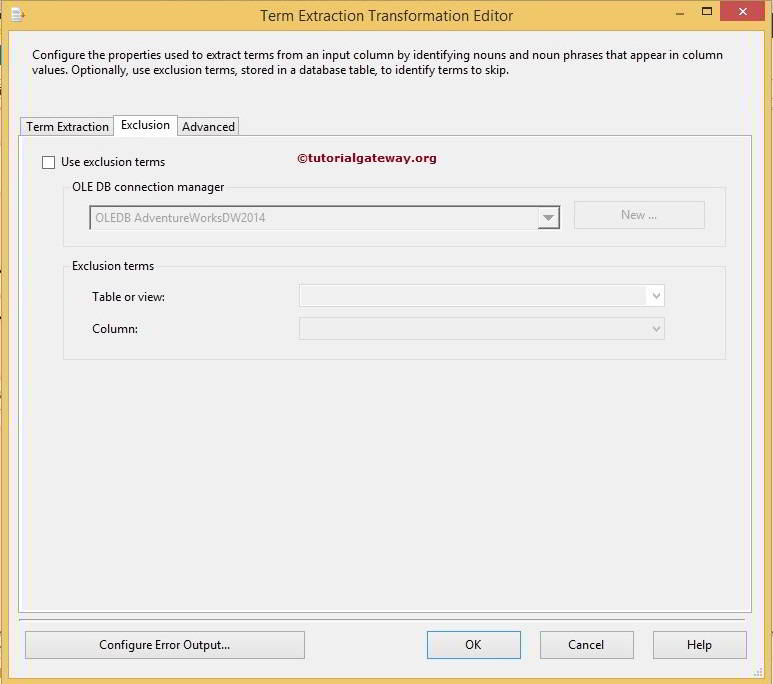
Term Extraction transformation produces only two output columns. The default names of the columns are **Term** and **Score,**but you can change them as per your requirement.

* **Term:**This column contains the extracted terms from the text. For example, if we are extracting the Nouns, then all the nouns will be stored in this column.
* **Score:** This column contains the number of times a term repeated in the input column. For example, India is the first term extracted from the text. Term Extraction Transformation will check all the rows and counts the number of times Term India repeated in all the rows available in that input column.

### Exclusion Tab in Term Extraction Transformation

This tab is used to exclude unwanted terms from the extraction. For example, when we are extracting terms from source data that contains product reviews about all our company products, we don’t need to extract Product name from the input text. To add the Exclusion Terms to the Term Extraction Transformation, please checkmark the **Use Exclusion Terms** option

**TIP:** Please refer to Exclusion Tab in Term article to understand the configuration of the Exclusion tab.

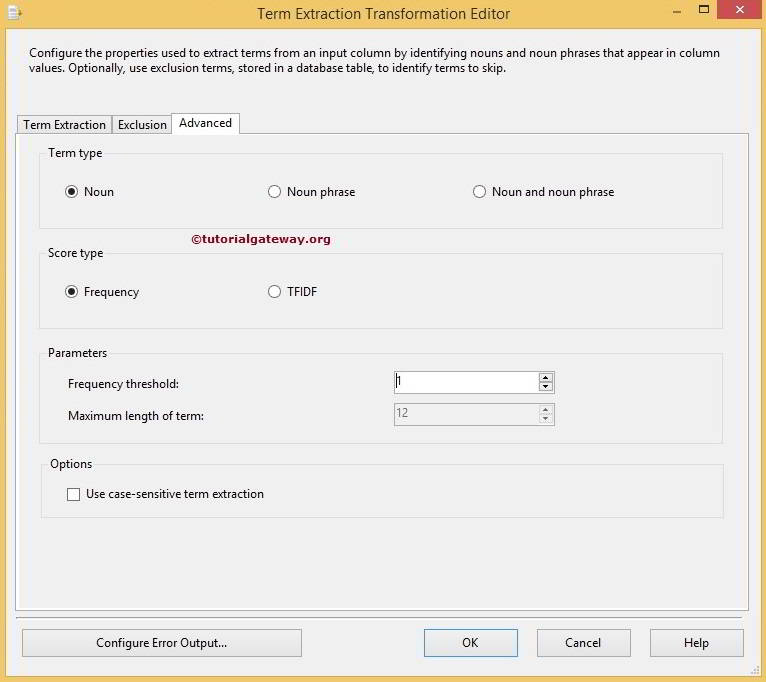


List of options available in Exclusion Tab to configure the exclusion list is:

* **OLE DB connection manager:**SSIS Term Extraction Transformation only supports the OLE DB connection manager to connect with the server holding the exclusion list. So, select an existing one from the drop-down list if you already created it. Or, if you want to create a new connection, then click on the New button.
* **New:**Create New connection to a database using the OLE DB Connection Manager dialog box.
* **Table or view:**Select the table or view from the drop-down list, which contains the exclusion terms.
* **Column:**Select the column name from the table or view, which contains the exclusion terms.
* **Configure Error Output:** Click on this button to configure the errors.

### Advanced Tab

Use the Advanced tab in the SSIS Term Extraction Transformation Editor to configure the extraction properties.



From the above screenshot, you can observe that the following are the list of options available in the Advanced Tab

* **Noun:** If you select this option, SSIS Term Extraction Transformation will extract only Nouns from the input text. Please refer to Term Extraction Transformation in SSIS for further reference.
* **Noun phrase:**If you select this option, the transformation will extract only Noun Phrases from the input text. Please refer to Extract Noun Phrases using Term Extraction Transformation in SSIS for further reference.
* **Noun and noun phrase: T**he SSIS Term Extraction transformation will extract both Nouns and Noun Phrases from the input text. Please refer to Extract Nouns and Noun Phrases using Term Extraction Transformation in SSIS for further reference.
* **Frequency:**If you select this option, the **Score**column stores the information of, Frequency of the Term repeated in the input column.
* **TFIDF:** The **Score**column will store the information of, TFIDF value of the Term.
* **Frequency threshold:**If we specify 3, the transformation extracts the Terms if they are repeated at least three times in the column, and it ignores the terms repeated less than three times.
* **Maximum length of term:**Please provide the maximum length of a word or phrase. This option is available if we selected the Noun Phrases only option.
* **Use case-sensitive term extraction:**Please check mark this option if you want to perform the Case-Sensitive extraction.