Compare Strings

Contents:

Unlike other types of data, text data has very few restrictions on the kinds of values that appear in a cell. In the application, this data is typically inferred as String data type. As a result, finding string values that mean the same thing can be a challenge, as minor differences in their content or structure can invalidate a match.

This section provides some methods for comparing strings.

- Some target systems may impose limits on the lengths of imported values. For more information on managing the lengths of your strings, see *Manage String Lengths*.

Find Substrings

You can use the following functions to locate sub-strings that are part of a column's value.

Function Name	Description
LEFT Function	Matches the leftmost set of characters in a string, as specified by parameter. The string can be specified as a column reference or a string literal.
RIGHT Function	Matches the right set of characters in a string, as specified by parameter. The string can be specified as a column reference or a string literal.
FIND Function	Returns the index value in the input string where a specified matching string is located in provided column, string literal, or function returning a string. Search is conducted left-to-right.
RIGHTFIND Function	Matches the right set of characters in a string, as specified by parameter. The string can be specified as a column reference or a string literal.
SUBSTRING Function	Matches some or all of a string, based on the user-defined starting and ending index values within the string.

The following transformation checks the left five values of the lowercase version of the ProdId column to see if it matches `xxxx-`. If the value is detected, then the `ProdName` value is set to `NO_NAME`:

Transformation Name	`Edit with formula`
Parameter: Columns	`ProdName`
Parameter: Formula	`IF(LEFT(LOWER(ProdId,5))=='xxxx-','NO_NAME' ,ProdName)`

Compare String Ends by Pattern

You can use the `STARTSWITH` and `ENDSWITH` functions to determine if a string begins or ends with a specified pattern.

> **Tip:** These functions are most useful for performing pattern-based checks on strings. For string literals, you can use the `LEFT` and `RIGHT` functions. See below.

The following transformation inserts the value `error` in the `custCodeStatus` column if the `custCode` value begins with six digits in a row:

Transformation Name	`Edit with formula`
Parameter: Columns	`custCodeStatus`
Parameter: Formula	`IF(STARTSWITH(custCode,`{digit}{6}`), 'error',custCodeStatus)`

See *STARTSWITH Function*.

See *ENDSWITH Function*.

Match Strings

Exact matching

You can use the `EXACT` function to compare if two strings are exact matches. String inputs can be literals, column references, or expressions that evaluate to strings.

> **NOTE:** The `EXACT` function evaluates for exact matches. Whitespace or capitalization differences return `false`.

You can nest function expressions inside of the `EXACT` reference to eliminate common and perhaps not useful differences between strings. In the following transformation, a value of `true` is inserted into the `matches` column, if `colA` and `colB` are exact matches, after whitespace and case differences have been removed:

Transformation Name	`New formula`
Parameter: Formula	`IF(EXACT(LOWER(REMOVEWHITESPACE(colA)))==EXACT(LOWER (REMOVEWHITESPACE(colB))),'true','false')`
Parameter: New column name	`matches`

Doublemetaphone matching

The platform also supports the doublemetaphone algorithm for fuzzy matching. This algorithm provides mechanism for proximity matching; the `DOUBLEMETAPHONEEQUALS` function supports an optional second parameter to define the strength of the algorithm.

This algorithm works by generating two separate encodings for each string: a primary encoding and a secondary encoding. You can experiment with these encodings using the `DOUBLEMETAPHONE` function. See *DOUBLEMETAPHONE Function*.

This algorithm can be applied to compare two strings, as in the following transformation.

Transformation Name	`New formula`
Parameter: Formula	`DOUBLEMETAPHONEEQUALS(colA,colB,'strong')`

Parameter: New column name	`matches`

- The first two parameters of the function are the string literals, column references, or functions returning strings to compare.
- The third parameter is optional. It determines the level of matching required to return `true`. Options:

Match threshold	Description
`'strong'`	Both primary encodings must match.
`'normal'`	At least one primary encoding must match either of the other string's encodings
`'weak'`	A primary or secondary encoding from each can match.

- For more information, see *DOUBLEMETAPHONEEQUALS Function*.

Compare Strings

For string values, you can use the string comparison functions to check how strings compare using Latin collation settings.

> **Tip:** Any column can be converted to String data type to use these functions.

Collation refers to the organizing of written content into a standardized order. String comparison functions utilize collation rules for Latin. A summary of the rules:

- Comparisons are case-sensitive.
  - Uppercase letters are greater than lowercase versions of the same letter.
  - However, lowercase letters that are later in the alphabet are greater than the uppercase version of the previous letter.
- Two strings are equal if they match identically.
  - If two strings are identical except that the second string contains one additional character at the end, the second string is greater.
- A **normalized version** of a letter is the unaccented, lowercase version of the letter. In string comparison, it is the lowest value of all of its variants.
  - a is less than .
  - However, when compared to b, a = .
  - The set of Latin normalized characters contains more than 26 characters.

This table illustrates some generalized rules of Latin collation.

Order	Description	Lesser Example	Greater Example
1	whitespace	(space)	(return)
2	Punctuation	'	@
3	Digits	1	2
4	Letters	a	A
5		A	b

Resources:

> **NOTE:** In the following set of charts (linked below), the values at the top of the page are lower than the values listed lower on the page. Similarly, the charts listed in the left nav bar are listed in ascending order.

For more information on the applicable collation rules, see *http://www.unicode.org/charts/collation/*.

Available functions:

- *STRINGLESSTHAN Function*
- *STRINGLESSTHANEQUAL Function*
- *STRINGGREATERTHAN Function*
- *STRINGGREATERTHANEQUAL Function*