[CSV, Excel] Read files with different schema error

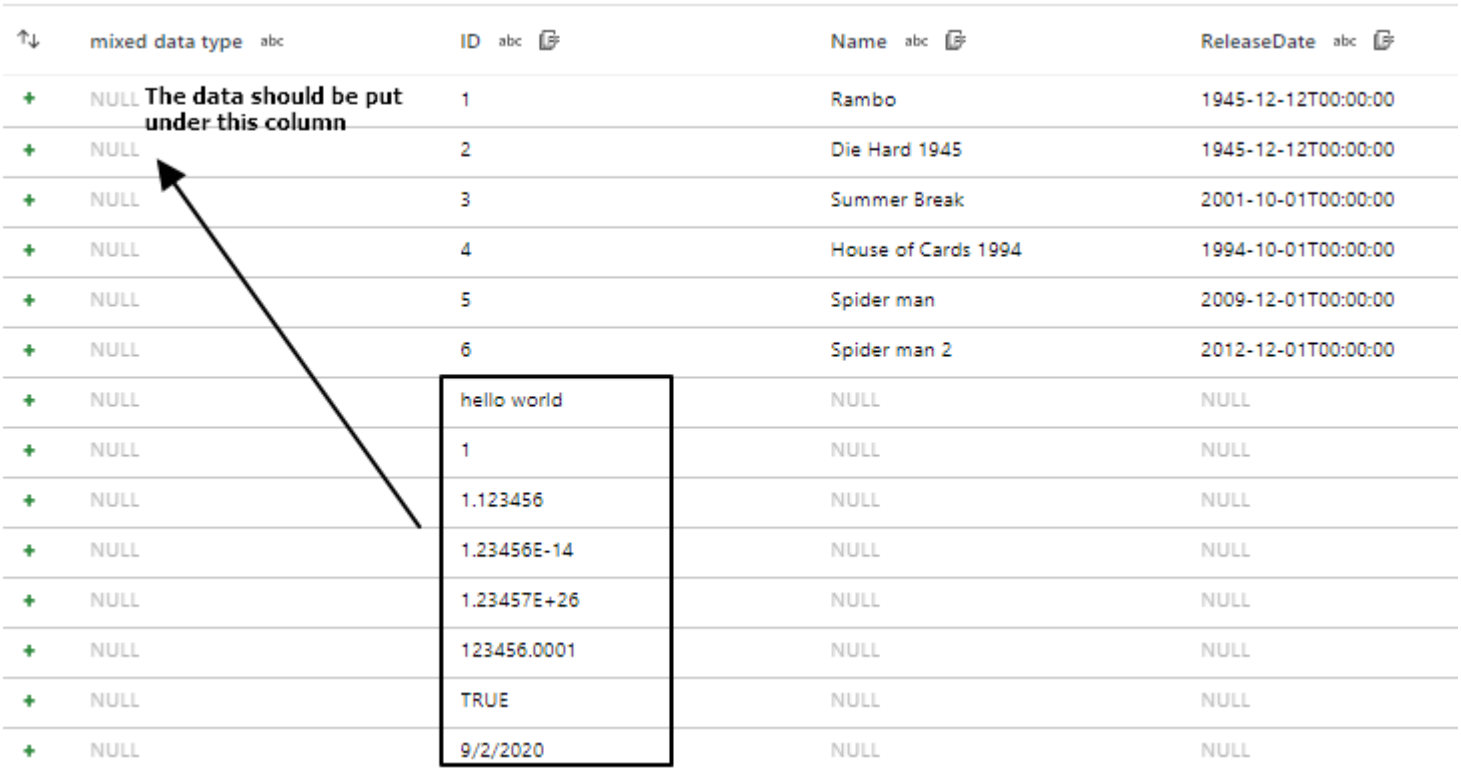
Last updated by | Jackie Huang | Jan 4, 2022 at 12:24 AM PST

Contents

Issue

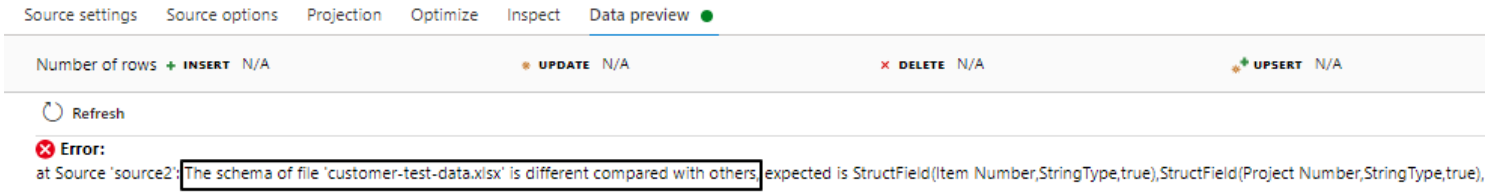
When customer uses dataflow to read files (csv, excel...) while the schema is different, the dataflow debug, sandbox or activity run will fail.

CSV: it has data misalignment when the schema of files is different.



mixed data type	ID	Name	ReleaseDate
NULL	1	Rambo	1945-12-12T00:00:00
NULL	2	Die Hard 1945	1945-12-12T00:00:00
NULL	3	Summer Break	2001-10-01T00:00:00
NULL	4	House of Cards 1994	1994-10-01T00:00:00
NULL	5	Spider man	2009-12-01T00:00:00
NULL	6	Spider man 2	2012-12-01T00:00:00
NULL	hello world	NULL	NULL
NULL	1	NULL	NULL
NULL	1.123456	NULL	NULL
NULL	1.23456E-14	NULL	NULL
NULL	1.23457E+26	NULL	NULL
NULL	123456.0001	NULL	NULL
NULL	TRUE	NULL	NULL
NULL	9/2/2020	NULL	NULL

Excel: it throws error when the schema of files is different.



Root Cause

For reading files with different schema in dataflow, it is not supported.

Resolution

If customer still wants to transfer files (csv, excel...) with different schema in dataflow, he can use following ways to work around:

CSV: The customer needs to manually merge the schema of different files to get the full schema. For example, file_1 has columns c_1, c_2, c_3 while file_2 has columns c_3, c_4,... c_10, so the merged and full schema is c_1, c_2... c_10. Then make other files also have full same schema even though it does not have data, for example, file_x only has columns (c_1, c_2, c_3, c_4), please add additional columns (c_5, c_6, ... c_10) in the file, then it can work.

Excel:

- Option-1: The customer needs to manually merge the schema of different files to get the full schema. For exa
- Option-2: Use range (for example, A1:G100) + firstRowAsHeader=false, then it can load data from all excel fi

Additional Information:

- Icm Reference: N/A
- Author: Zhangyi Yu
- Reviewer: Zhangyi Yu; Shawn Xiao
- Keywords:

How good have you found this content?