[JSON] Consume json file generated by Copy Activity

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

Contents

Issue

If customer use copy activity to generate some JSON files, then try to read these files in dataflow, it may fail with following error message:
JSON parsing error, unsupported encoding or multiline

Root Cause

There are following known limitations on JSON for copy & dataflow respectively:

- For Unicode encodings (utf-8, utf-16, utf-32) JSON files, copy activity always generate the JSON files with BOM
- Dataflow JSON source with "single document" or "Array of documents" selected as document form doesn't support Unicode encoding with BOM

dataflow2 ● ☷ csv_test_1 ●

✓ Validate ⬤ Data flow debug ⬤

☷ source1

| Source settings | **Source options** | Projection | Optimize | Inspect | Data preview |

◢ **File settings**

File mode ⓘ ⦿ File ◯ Wildcard

File path * [Container] / [Folder path] / [File name] 🗁 Browse

Allow no files found ⓘ ☐

Compression type [None ⌄]

Encoding [Default(UTF-8) ⌄]

◢ **Other settings**

Document form ⓘ ◯ Single document ◯ Document per line ⦿ Array of documents

Unquoted column names ⓘ ☐

Has comments ⓘ ☐

Single quoted ⓘ ☐

Backslash escaped ⓘ ☐

Partition root path ⓘ [   ]

List of files ⓘ ☐

Column to store file name ⓘ [   ]

After completion * ⦿ No action ◯ Delete source files ◯ Move

Filter by last modified Start time (UTC) [   ] End time (UTC) [   ] ⓘ
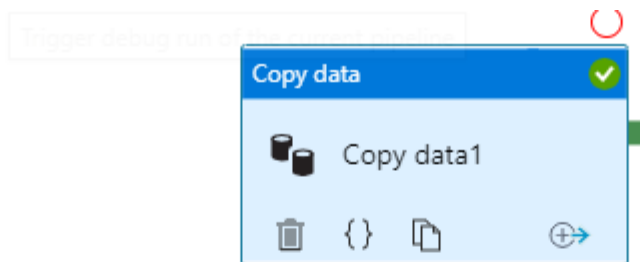
So if it meets following criterial, customer will experience issue:

- The sink dataset used by copy activity is set to Unicode encoding (utf-8, utf-16, utf-16be, utf-32, utf-32be) or default is used
- Copy Sink is set to use "Array Of Objects" file pattern (no matter whether "single document" is enabled or not in dataflow JSON source)

Resolution #1

- Set encoding to "UTF-8 without BOM" in copy sink dataset, then the json files generated won't have BOM.

Resolution #2

- Always use default file pattern or explicit "Set of Objects" pattern in copy sink if the generated files will be used in dataflow
- Select "Document per line" as document form in dataflow JSON source

Note: Using "Set of Objects" is also the recommended practice from performance perspective, as the "Single document" JSON in dataflow can't enable parallel read for single large files, so this resolution won't have any negative impact.