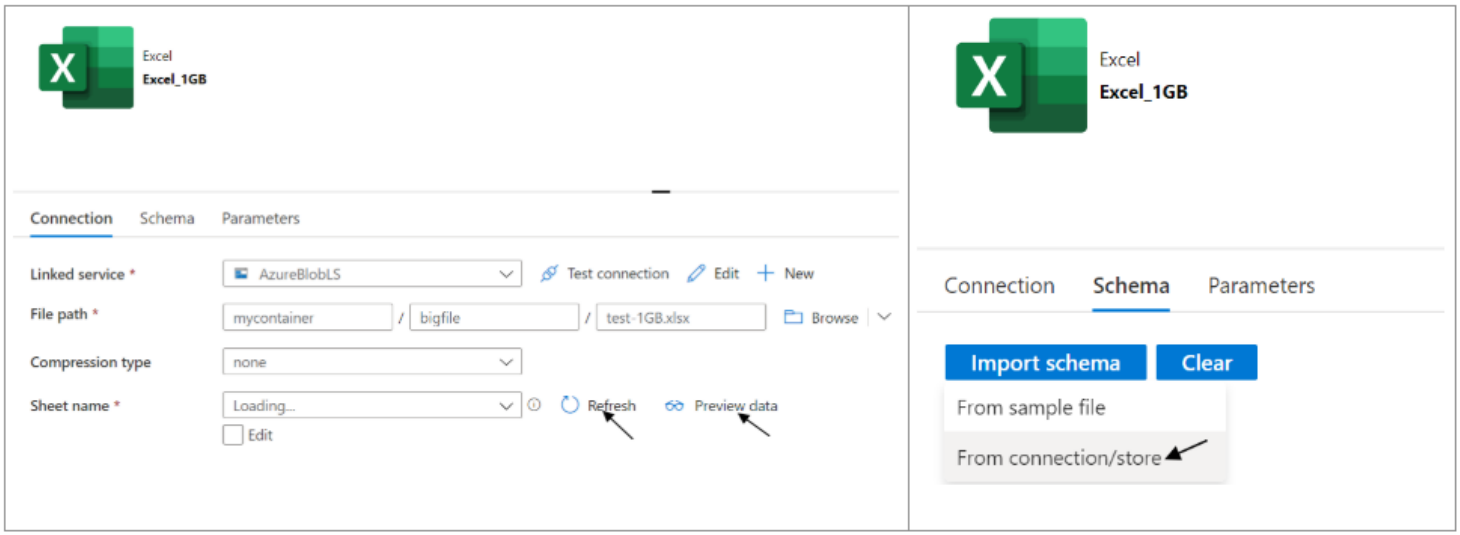
Excel - It is timeout or has low performance with big file

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

Contents

Issue

1, When customer creates excel dataset, or have "import schema", "preview data" and "refresh" on dataset, he may meet timeout issue if related excel file is big.



2, When customer uses copy activity to move big excel file (>= 100MB) into other data store, he may meet low performance, low network or OOM issue.

Root Cause

Today ADF excel in copy does not support streaming read which is limited by third part library/SDK (NPOI) we used, it must load whole excel file into memory, then to find customer specified worksheet to read data row by row, cell by cell.

- For all the operation ("import schema", "preview data", "refresh") on excel dataset, it is interactive operation, the data must be returned by ADMS runtime immediately before http request is timeout (100s). For big excel file, these operation cannot be finished in 100s, so timeout issue is raised.

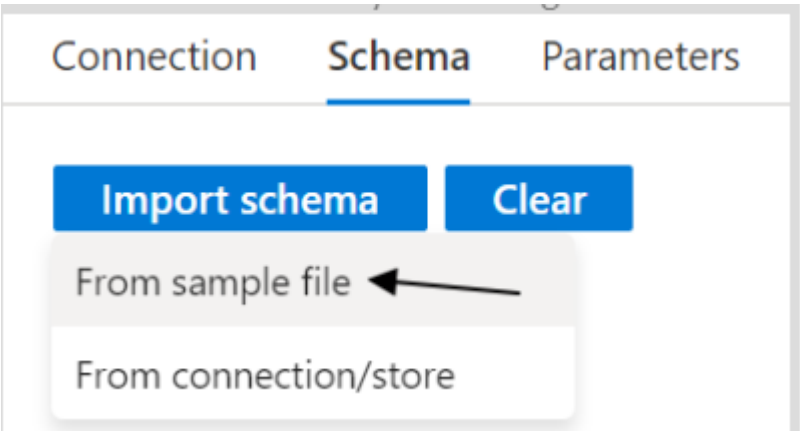
- For copy big excel file, it has low performance, network or OOM, it is the same root cause because of streaming read is not supported.

Resolution

If customer wants to move/transfer big excel file (>100MB) into other data store in ADF, he/she can use one of following option to work around it:

- Option-1: Register and online self-hosted integration runtime (SHIR) with powerful machine, then use copy activity to move big excel file into other data store with that IR.
- Option-2: Split big excel file into several small ones (for example, split one 100MB excel file into ten 10MB ones), then use copy activity to move the excel folder.
- Option-3: Use dataflow activity to move big excel file into other data store. In dataflow, the excel has supported streaming read with few CPU/Memory consumption. Based on perf test, it can move/transfer 1GB excel file in dataflow within 5 minutes while using Azure Integration runtime with 32 cores.
- Option-4: The customer can manually convert/save big excel file as csv format, then use copy activity to move it.

For interactive operation ("import schema", "preview data", "refresh") on excel dataset which has big file, the customer can create one sample excel file which is subset of original file, then use sample file to finish "import schema", "preview data" and "refresh".



Note: If customer already know to read data from which worksheet, he can directly check "Edit" of "Sheet name" to input related value.

Later changing the file name back to be original file to deploy this dataset, then this excel dataset can be reused by different activity (copy, lookup, get-metadata...).

Additional Information:

Icm Reference: N/A

Author: Zhangyi Yu

Reviewer: Zhangyi Yu; Shawn Xiao

Keywords: