

# [CosmosDB & JSON] Support customized schema in source

Last updated by | Zhangyi Yu | Aug 29, 2022 at 11:33 PM PDT

---

## Contents

- [Issue](#)
- [Root Cause](#)
- [Resolution](#)
- [Reference](#)
- [Additional Information](#)

## Issue

When ADF dataflow customers try to move/transfer data from CosmosDB/JSON into other data stores, some columns of source data may be missing.

## Root Cause

For schema free connectors(the column number, column name and column data type of each row can be different compared with others), today ADF uses sample rows (for example, top 100 or 1000 rows data) to infer the schema, the inferred result will be used as the schema to read data. So if customer's data store has additional columns which doesn't appear in sample rows, the data of these additional columns cannot be read, moved and transferred into sink data store.

## Resolution

In order to solve up known issue/limitation, we have enabled feature "support customized schema in schema free connectivity" in ADF dataflow, the customer can feel free to specify additional/missing columns (which could be missing in schema-infer-result) in dataflow source projection to read the data they want with one of following options to set customized schema while option-1 is more preferred:

**Option-1:** Compared with original source data that may be one big file, table or container which contains millions of rows with complexed schema, the customer can create one temp table/container with few rows where all the columns they want to read are included, then to have following operation:

1. Use dataflow source "Debug Settings" to have "Import projection" with sample file/table to get the fully schema.

1. Click on the 'Debug Settings' icon in the top toolbar.

2. Select 'Sample table' under the 'cosmosSource' section.

3. Check the 'Edit' checkbox.

4. Click the 'Save' button.

5. Click the 'Import projection' button.

2. Change "Debug Settings" back to use source dataset to have remaining data movement/transformation.

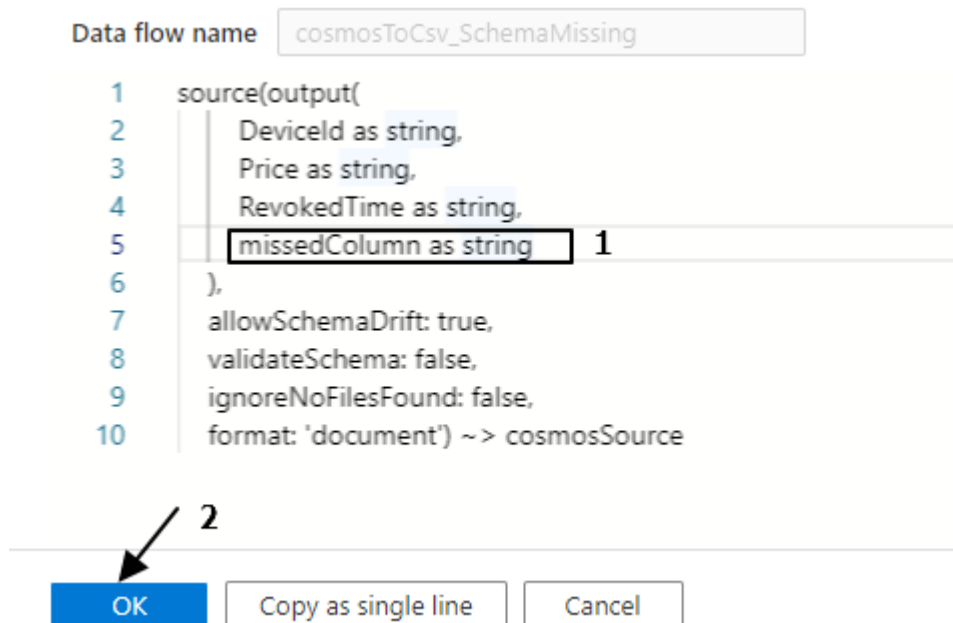
1. Click on the 'Debug Settings' icon in the top toolbar.

2. Select 'Source dataset' under the 'cosmosSource' section.

3. Click the 'Save' button.

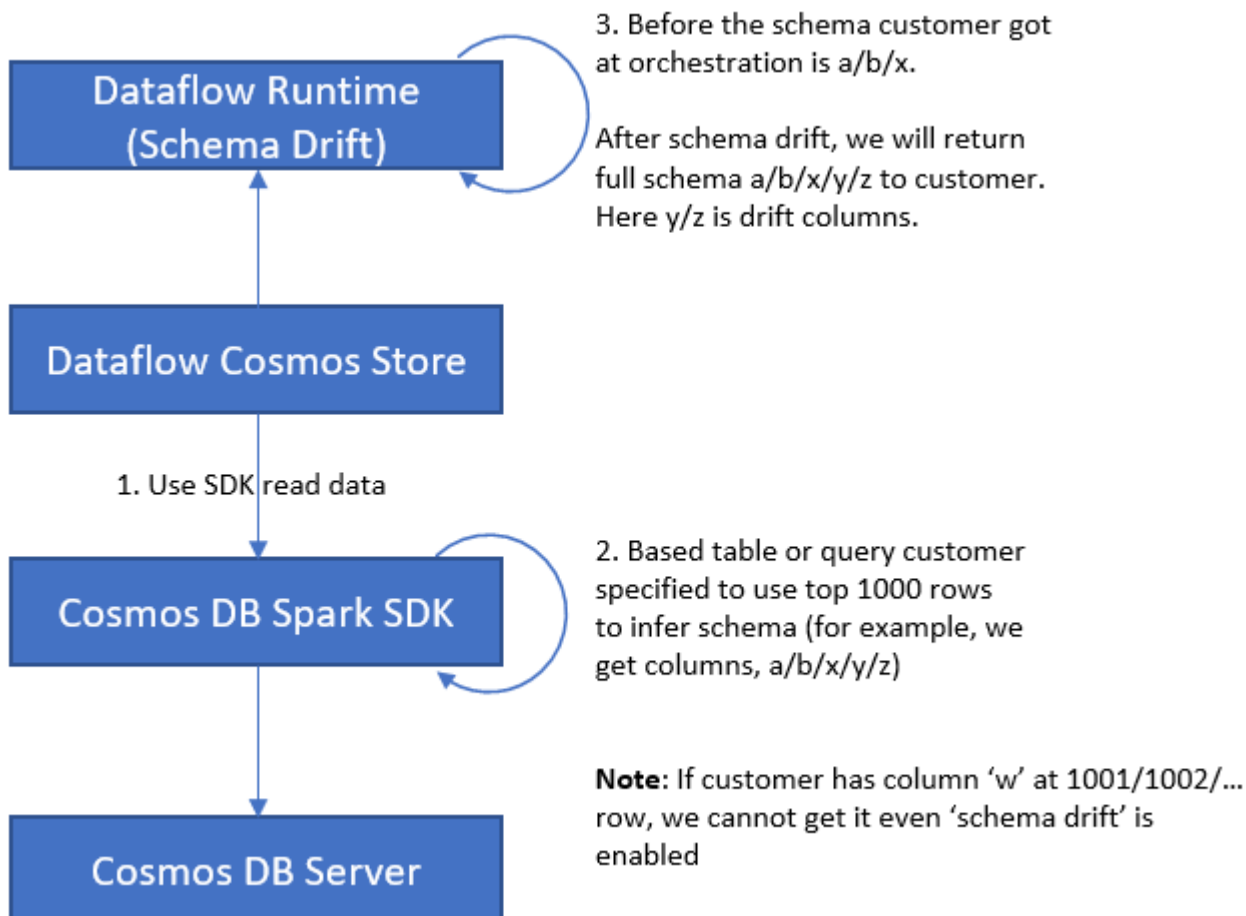
Then ADF dataflow runtime will honor and use the customized schema to read data from original data store.

**Option-2:** If customers are very familiar with source data's schema and DSL language, the customers can manually update the dataflow source script to add additional/missed columns to read data, for example:



## Reference

For most customers who are familiar with ADF dataflow, they know enable "schema drift" can get additional schema (changed or new added columns) automatically during read data in pipeline/activity execution. For cosmos DB, it always use top 1000 rows to infer schema based on customer specified container or query. If changed or new column is at 1001th, 1002th, 1003th... row, the new schema will not be included in "Schema Drift", for details, please reference following diagram:



## Additional Information

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