Dataflow taking longer at Source with fileSystemInitDuration high

Last updated by | Mithun Rajendran | Mar 3, 2023 at 10:50 AM PST

Contents

Issue

Customer's DataFlow run takes a long time to run. A lot of time is taken at the source with fileSystemInitDuration high.

Analysis

The customer might mention that the Dataflow pipeline is running more than expected time. Although increased core counts and compute type to Memory-optimized, still with less volume, the pipeline is running longer than expected.

To confirm if this is happening at Source read you can use following query

```
cluster('adfcus.kusto.windows.net').database("AzureDataFactory").DataflowClusterLogs
| union  cluster('adfneu.kusto.windows.net').database("AzureDataFactory").DataflowClusterLogs
| where Message contains "StoreContext"
| where ActivityRunId contains "<ActivityRunID>"
| where Message contains "<SourceName>"
```

Table 1 Stats

Timestamp	Message
2023-03-03 16:44:49.5940	StoreContext.timed: transformation name [AcaoExtensivaCSV]. Start execute for fileSystemInitDuration.
2023-03-03 16:44:49.6850	StoreContext.timed: transformation name [AcaoExtensivaCSV]. Start execute for pathResolutionDuration.
2023-03-03 16:44:49.7680	StoreContext.timed: transformation name [AcaoExtensivaCSV]. End execute for pathResolutionDuration. Duration is 79 ms.
2023-03-03 16:44:49.9960	StoreContext.timed: transformation name [AcaoExtensivaCSV]. Start execute for fileLoadDuration.
2023-03-03 16:45:04.1440	StoreContext.timed: transformation name [AcaoExtensivaCSV]. End execute for fileLoadDuration. Duration is 14147 ms.
2023-03-03 16:45:04.1510	StoreContext.timed: transformation name [AcaoExtensivaCSV]. End execute for fileSystemInitDuration. Duration is 14557 ms.
2023-03-03 16:45:10.4300	StoreContext.timed: transformation name [AcaoExtensivaCSV]. Start execute for moveFilesDuration.
2023-03-03 16:45:10.5860	StoreContext.timed: transformation name [AcaoExtensivaCSV]. End execute for moveFilesDuration. Duration is 156 ms.

'

Three things to look at

- pathResolutionDuration - Time taken to resolve wildCardPaths

- fileLoadDuration - Time taken to inferSchema from the source files
- fileSystemInitDuration - Total time to initialize source which includes above two durations

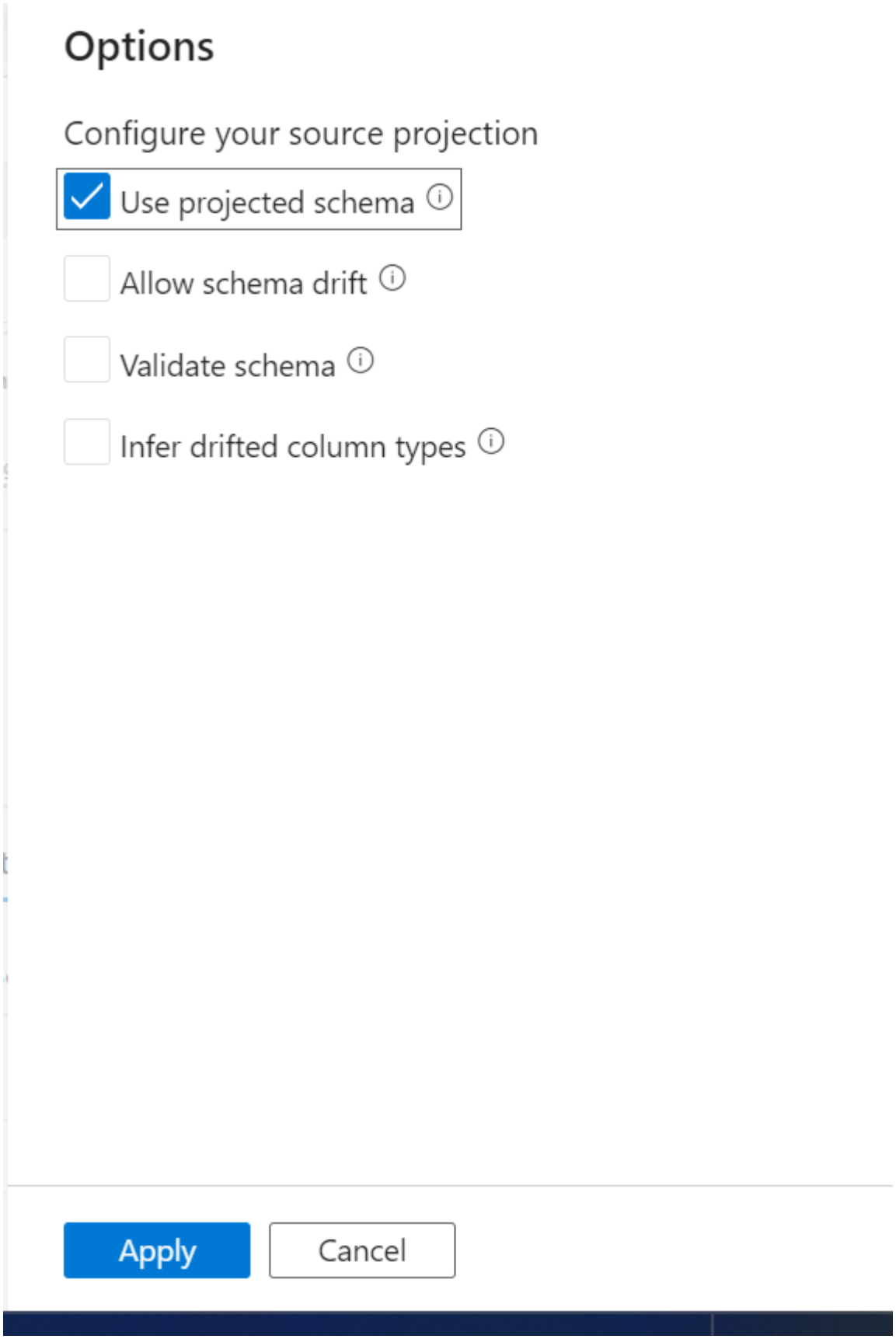
Recommendation

If ***pathResolutionDuration*** is high, it could be because of a complex wildcard path. Changing this to a less complex folder path would help reduce time.

If ***fileSystemInitDuration*** is high, this could be because inferring schema is taking longer. Actions to take :

- If *schemaDrift is not needed*, "Use Projected Schema" can be enabled for this source. This will ensure "Projection" schema for this is used and infer schema is disabled.
- Supported formats (only Inline): ***JSON, Delimited***

Ensure there is schema projected already. Go to "Projection" -> "Schema Options" -> Disable "allow schema drift" and enable "use projected schema"


Options
Configure your source projection
Use projected schema
Allow schema drift
Validate schema
Infer drifted column types
Apply
Cancel