Dataflow job failed with 'Requested array size exceeds VM limit' error

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

Contents

Issue

Dataflow job failed with the following error:

{"message":"Job failed due to reason: **Requested array size exceeds VM limit**. Details:Requested array size exceeds VM limit","failureType":"UserError","target":"ExcludeHistoryDflow","errorCode":"DFExecutorUserError"}

It seems customer is trying to perform a transformation which is consuming too much memory.

We could see the following error in the backend cluster logs:

20/09/11 17:09:43 WARN TaskSetManager: Lost task 0.0 in stage 9.0 (TID 124, 10.139.64.29, executor 0): java.lang.**OutOfMemoryError**: Requested array size exceeds VM limit at java.util.Arrays.copyOf(Arrays.java:3236)

Resolution

There are two common way to resolve this:

1. Use "MemoryOptmized" instead of general purpose type. Refer **Check compute used by customer** section below to find out compute used by the customer for Dataflow run.

   OR

2. Try to increase the partition count from source/transformation in order to reduce the memory usage for each partition. Refer **Check Partition Counts** used by the customer at source

---

Check compute used by customer

Query the following table with the ActivityRunId to check the compute used by the customer:

```
cluster('azuredmprod.kusto.windows.net').database('AzureDataMovement').TelemetryEvent
| union cluster('azuredm.kusto.windows.net').database('AzureDataMovement').TelemetryEvent
| where ActivityId == "<activityRunId>"
| project PreciseTimeStamp, RoleInstance, Payload, ActivityId, SubscriptionId, Tenant
```

Check the ComputeType in the Payload properties, to find out the compute being used.

Check Partition Counts

Run the following query to find out the partition counts used by the customer:

```
cluster('adfcus.kusto.windows.net').database('AzureDataFactory').DatabrickLoggingTest
| union cluster('adfneu.kusto.windows.net').database('AzureDataFactory').DatabrickLoggingTest
| where ActivityRunId == "<ActivityRunId>"
| where caller in ("com.microsoft.datafactory.dataflow.AdmsClient.processedEvent")
```

Additional Information:

- **Icm References:** Icm Link
- **Author:** vimals
- **Reviewer:** vimals
- **Keywords:**

How good have you found this content?