Remote RPC client disassociated. Likely due to containers exceeding thresholds, or network issues

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

Contents

Issue

Dataflow pipeline execution is failed with following error:

org.apache.spark.SparkException: Job aborted due to stage failure: Task 2016 in stage 547.0 failed 1 times, most recent failure: Lost task 2016.0 in stage 547.0 (TID 26149, 10.139.64.16, executor 4): ExecutorLostFailure (executor 4 exited caused by one of the running tasks) Reason: **Remote RPC client disassociated. Likely due to containers exceeding thresholds, or network issues**. Check driver logs for WARN messages.

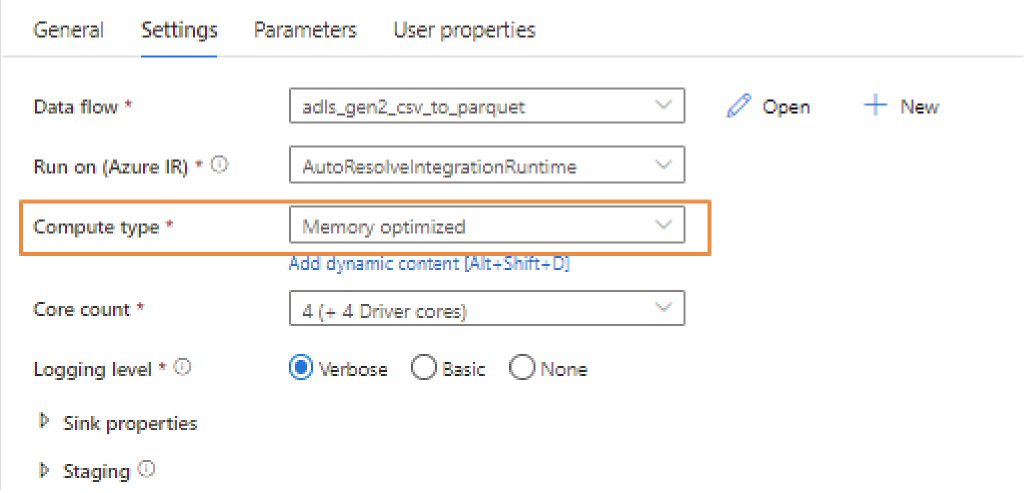
Root Cause

When customer meets up error, it could be transient network issue or one node in spark cluster runs out of memory.

Resolution

When customer meets up error, he/she can have a try to use following options to solve the problem.

Option-1: Use powerful cluster (both drive and executor nodes have enough memory to handle big data) to run dataflow pipeline by setting "compute type" as "Memory optimized"



General | Settings | Parameters | User properties

Data flow * adls_gen2_csv_to_parquet

Run on (Azure IR) * AutoResolveIntegrationRuntime

Compute type * Memory optimized

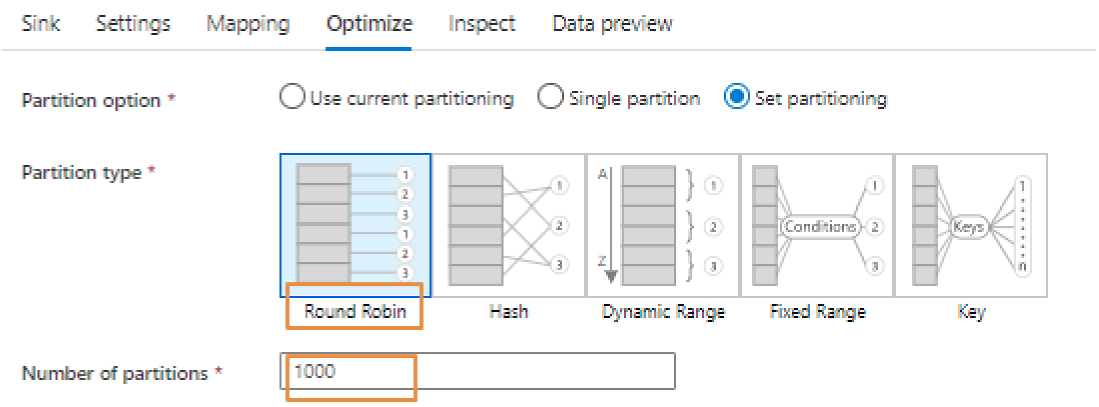
Add dynamic content [Alt+Shift+D]

Core count * 4 (+ 4 Driver cores)

Logging level * ● Verbose ○ Basic ○ None

Option-2: Use one larger cluster size (for example, 48 cores) to run dataflow pipeline. https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-performance#cluster-size

Option-3: Repartition input data. For task running on dataflow spark cluster, one partition is one task and runs on one node. If data in one partition is too big, related task running on node needs to consume more memory than the node itself, it will fail. So the customer can have repartition to avoid data skew and ensure data size in each partition is average while memory consumption is not too heavy.



Sink | Settings | Mapping | Optimize | Inspect | Data preview

Partition option * ○ Use current partitioning ○ Single partition ● Set partitioning

Partition type * Round Robin

Number of partitions * 1000

Note: The customer needs to evaluate the data size/partition number of input data, then set reasonable partition number under "Optimize". For example, the cluster customer used in dataflow pipeline execution is 8 cores and memory of each core is 20GB, but the input data is 1000GB with 10 partitions. If the customer directly run the dataflow, it will meet OOM issue because 1000GB/10 > 20GB, so it is better to set repartition number to be 100 (1000GB/100 < 20GB)

Option-4: Tune and optimize source/sink/transformation settings. For details, you can reference https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-performance For example, don't using wildcard pattern, and try to copy all the files on one container.

Additional Information

- **Icm References:** Icm Link
- **Author:** zhanyu@microsoft.com
- **Reviewer:** shawnxia@microsoft.com
- **Keywords:**

How good have you found this content?

🙂 🙁