DataFlow Runtime Custom Properties

Last updated by | Anjali Sharma | Mar 15, 2023 at 12:47 PM PDT

Contents

Overview

Custom properties in dataflow runtime, is a way to control your compute requirements and helpful in setting the configurations according to the dataflow need. This customization should only be used in known scenarios otherwise it can cause unnecessary dataflow failures.

Supported Custom Properties

This customization is currently supported for the following properties:

Property	Type	Default	Description
Shuffle partitions	Integer	200	Configures the number of partitions to use when shuffling data for joins or aggregations.
Constraint propagation	Boolean	true	Constraint propagation can be computation expensive and block the driver execution for long time.

When to use

- **Shuffle partitions**

  Dataflow divides the data into partitions and transforms it using different processes. If the data size in a partition is more than the process can hold in memory then the process fails with OOM errors. If dataflow contains huge amounts of data and having joins/aggregations transformations in such a case you may want to try changing shuffle partitions in incremental way by 50 up to 2000 to avoid OOM errors.

  While increasing the shuffle partitions, make sure data is spread across well, a rough number is to have ~1.5 GB of data per partition, if data is skewed increasing the "Shuffle partitions" will not be helpful.

For example: If you have 500 GB of data then having a value between 400 to 500 should work.

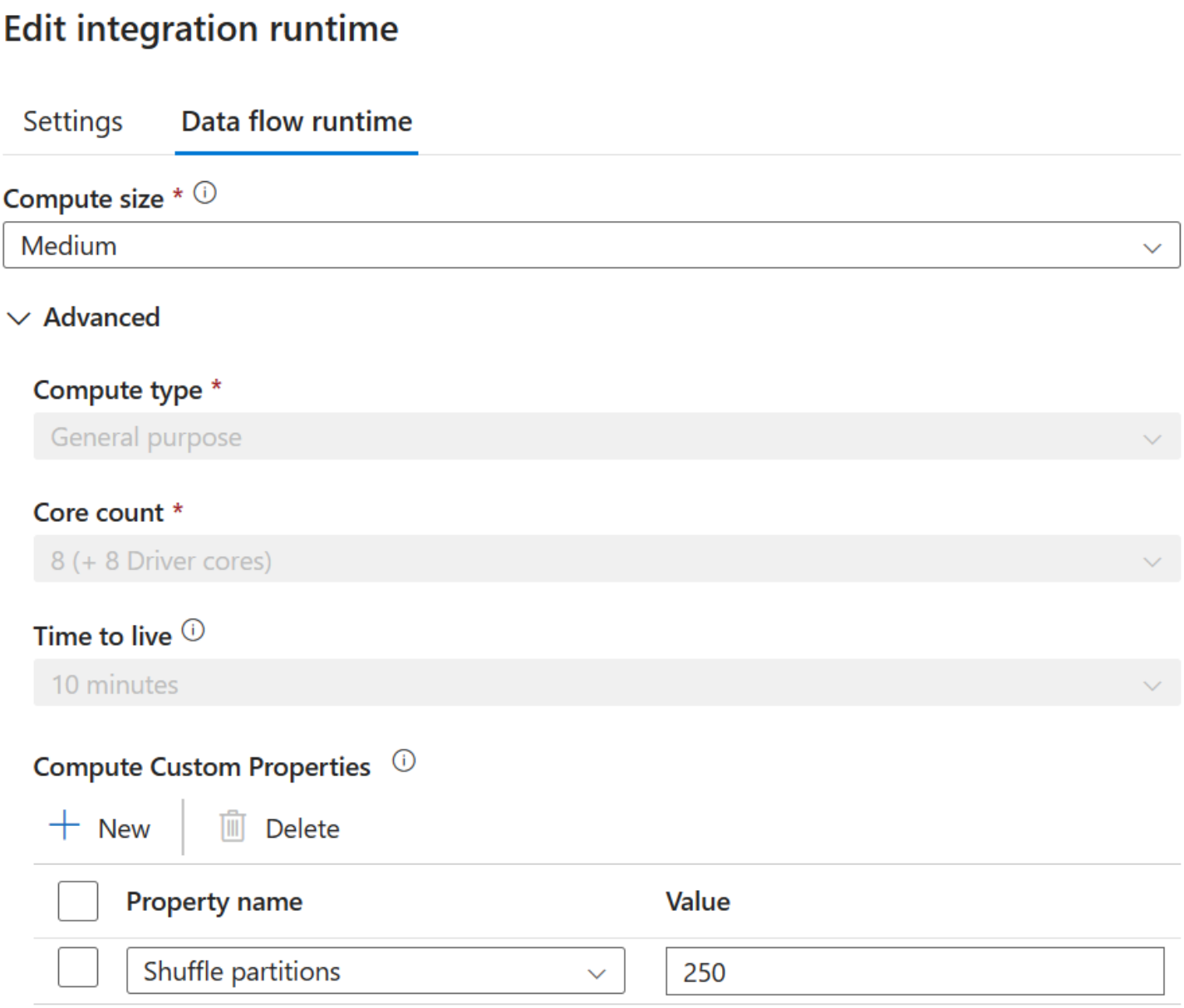
Default limit for shuffle partitions is 200 which works well for ~300 GB of data.

- **Constraint propagation**

  Query optimizer will infer and propagate data constraints in the query plan to optimize them. Constraint propagation can sometimes be computationally expensive for certain kinds of query plans such as those with a large number of predicates and aliases which might negatively impact overall runtime. In some special cases, constraint propagation will lead to OOM or job stuck. It's recommended to disable this config and set the value to "false" in those scenarios.

Update Custom Properties in Dataflow Run Time UI

Customer can go to "Edit integration runtime" and then use dataflow runtime tab to update the custom properties from advanced option such as follows,



Update Custom Properties in Integration Runtime JSON

Customer can pass the custom properties using Integration Runtime JSON editor as an array of name and value pair, where name is custom property name and value represents the value such as follows and use it to update "Shuffle partitions".

Copy to clipboard

```
{
    "name": "SparkTest-IR",
    "properties": {
        "type": "Managed",
        "typeProperties": {
            "computeProperties": {
                "location": "AutoResolve",
                "dataFlowProperties": {
                    "computeType": "General",
                    "coreCount": 16,
                    "timeToLive": 10,
                    "cleanup": false,
                    "customProperties": [
                        {
                            "name": "Shuffle partitions",
                            "value": "250"
                        }
                    ]
                }
            }
        }
    }
}
```