How Managed disks are created and charged for ADF Dataflow?

Last updated by | Ranjith Katukojwala | Mar 10, 2022 at 8:59 AM PST

Contents

How Managed disks are created and charged for ADF Dataflow?

When dataflow is triggered with a cluster, it uses cluster disks to process the data. However, if the data is huge/bigger than the specified IR memory, might create additional disks as the existing disks may run out of memory.

Disks will be used to preserve the spark job result and the size of the disk gets auto-scaled depending on the customer's data flow job definition/data size. These disks would be premium disks and can not control the storage type.

The disk usage is not just aligned with the IR core count, if the customer has the data size increase for a certain set of runs, or especially when the customer has aggregation/join-like transformation (lookup/join/aggregation/) in the data flow, it will easily lead the disk usage to go higher which causes more disk charge on customer’s billing report.

Below Customer's question is one of the examples just for your reference.

Customer question: The spike occurred between the 17th and 27th of January. Apart from the spike itself, what most concerns us is that the same pattern occurred on 5 of the 6 of our Production Data Factories and for a period of 11 days.

Due to the retention date of logs, we may not have the exact report. However, based on the customer's question, let’s assume if we have a leaked compute issue, then it should be a random pattern with a small chance on all factories across the region (otherwise it would be a noticeable sev 2 issue). However, according to the customer, there are 5 out of 6 factories that have the spike during the same time period, so to be more likely it’s due to source size/structure change or data flow definition modification on their side during that time instead of service issue.

Recommendation:

We can’t guarantee the disk usage would be the same if there’s no change on activity runs or even data flow definition. Usually, the overall cost of the disk from data flow will be 20-40% of data flow vcore usage. There’s no good way to predict the disk cost given it’s controlled by internal spark core logic instead of data flow/data factory layer. The “best” indicator customers might rely on for now is to check the ratio between disk and vcore (data flow meter ID) usage. Based on previous cases experience, If the ratio is way over 40% in the customer’s monthly billing report, it might be due to some abnormal disk issue, and feel free to engage PG in AVA for further investigation.

How good have you found this content?

🙂 🙁