Scaling MemoryDB clusters

PDF | RSS

As demand on your clusters changes, you might decide to improve performance or reduce costs by changing the number of shards in your MemoryDB cluster. We recommend using online horizontal scaling to do so, because it allows your cluster to continue serving requests during the scaling process.

Conditions under which you might decide to rescale your cluster include the following:

- **Memory pressure:**

  If the nodes in your cluster are under memory pressure, you might decide to scale out so that you have more resources to better store data and serve requests.

  You can determine whether your nodes are under memory pressure by monitoring the following metrics: *FreeableMemory*, *SwapUsage*, and *BytesUsedForMemoryDB*.
- **CPU or network bottleneck:**

  If latency/throughput issues are plaguing your cluster, you might need to scale out to resolve the issues.

  You can monitor your latency and throughput levels by monitoring the following metrics: *CPUUtilization*, *NetworkBytesIn*, *NetworkBytesOut*, *CurrConnections*, and *NewConnections*.
- **Your cluster is over-scaled:**

  Current demand on your cluster is such that scaling in doesn't hurt performance and reduces your costs.

  You can monitor your cluster's use to determine whether or not you can safely scale in using the following metrics: *FreeableMemory*, *SwapUsage*, *BytesUsedForMemoryDB*, *CPUUtilization*, *NetworkBytesIn*, *NetworkBytesOut*, *CurrConnections*, and *NewConnections*.

Performance Impact of Scaling

When you scale using the offline process, your cluster is offline for a significant portion of the process and thus unable to serve requests. When you scale using the online method, because scaling is a compute-intensive operation, there is some degradation in performance, nevertheless, your cluster continues to serve requests throughout the scaling operation. How much degradation you experience depends upon your normal CPU utilization and your data.

There are two ways to scale your MemoryDB cluster; horizontal and vertical scaling.

- Horizontal scaling allows you to change the number of shards in the cluster by adding or removing shards. The online resharding process allows scaling in/out while the cluster continues serving incoming requests.
- Vertical Scaling - Change the node type to resize the cluster. The online vertical scaling allows scaling up/down while the cluster continues serving incoming requests.

If you are reducing the size and memory capacity of the cluster, by either scaling in or scaling down, ensure that the new configuration has sufficient memory for your data and Redis overhead.