[External] Data Analysis Cost

THIS DOCUMENT IS NOW DEPRECATED, please refer to the Help Center article.

Summary

We are autoscaling Data Analysis in Glean Assistant up and down based on usage, for AWS and GCP customers. Up to 50 concurrent users will be able to use Data Analysis within a few minutes, and we will spin down their sandboxes within 30 minutes of inactivity.

We expect it to cost between $30-45/month at a base level, and scale efficiently. For example, with 10 users concurrently using the sandbox for an hour a day, total cost will be $45-60/month.

Technical Details

Hosting and autoscaling primer

We use shared core machines for running sandbox and orchestrator pods. Shared core machines are ideal for our use case because sandboxes are mostly idle and require short bursts of high CPU and memory when executing code as part of an analytical glean chat query.

One pod runs on one machine. In steady state, we run 2 sandboxes and the orchestrator which sums to a base monthly cost of 3 * $(cost per machine) per month. We will autoscale to more than 2 sandboxes when data analysis sees higher usage and then downscale back to 2 sandboxes when the usage goes back down. The autoscaling algorithm roughly tries to maintain enough sandboxes so that only 75% of the sandboxes are being used. When upscaling, we'll

incur additional cost for the extra machines running (see examples below) but these will roughly be for the amount of time sandboxes are being used and should add negligibly to the base cost.

Example

10 users ask data analysis questions to hold one sandbox for 1 hour every day. We will assume that all the usage happens at the same time to get a worst case calculation. Then, for 1 hour, we run 14 machines (to ensure that utilization is <= 75%). After the hour of usage, we mark these sandboxes for deletion after 15 mins of inactivity. The machines are then downscaled 15 more minutes later. So, 14 - 2 = 12 extra sandboxes run for 1.5 hours everyday. This equals 12 * 1.5 * 30 = 540 hours of extra machine uptime per month over the base cost.

Cost in GCP

We use the e2-small machine type. Each machine costs $0.016751 per hour or $12.06 per month.
Monthly base cost is $36.18. With the example case, 540 hours * $0.016751 = $9.04554 extra per month.

Total cost = 45$ per month.

Cost in AWS

We use the t3.small machine type. Each machine costs $0.0208 per hour or $14.97 per month.
Monthly base cost is $44.91. With the example case, 540 hours * $0.0208 = $11.232 extra per month.

Total cost = 56$ per month.

Notes

The price of machines can vary slightly over regions. The machines prices are for us-east regions of GCP and AWS.

We (configurably) limit upscaling to a maximum of 50 sandboxes. This ensures an upper bound of cost incurred to $603/month for GCP and $748/month on AWS.

If you anticipate your needs to be higher, these sandbox limits can be increased to 450 on AWS and 1000 on GCP.

If you have any questions, please reach out to your Glean contact.