Google Cloud

Best Practices- FCT Portugal

Account Types

Billing Ids + Safeguarding

Creating a billing id account requires entering credit card details. Billing ids with safeguarding in place allows you to have optimal use of your award without getting your credit card charged.

To avoid charges you would need to avoid accessing your project during the interim period and would need to set a budget cap to automatically disable billing.

Instrumentless Accounts

Using instrutmentless accounts does not require inputting credit card details. These accounts have limitations including the following:

- Once promotional credits are consumed or expired the **account is depleted**. This means you can no longer access your work.
- GPUs are not available in these types of accounts.
- Should you require additional credits, it is not possible to top off these accounts.

Making the most of your GCP spend

Shut off nodes when not in use

- Many of the interactive nodes don't need to operate while your researchers or students are asleep. Shutting off the nodes does not delete the root volumes, it just suspends the VM until you start it up again. When it comes up, it may have a different IP, but if you need a consistent IP address, you can assign one. This can easily save ⅓ to ½ of your credit allocation.
- Scale your cluster sizes dynamically as needed. Slurm scheduler has the capability to do queue-based scaling. With other schedulers, you can use managed instance group.

Preemptible machines save up to 80% of cost

- For batch workloads, consider using preemptible virtual machines that provide a cost reduction of up to 80% as compared with regular vms. The tradeoff is that these nodes may be reclaimed without warning. Traditionally, customers have experienced a 5-10% reclamation rate. By using a workload manager such as HTCondor or Slurm, the jobs on the reclaimed nodes will be rescheduled and run again. Alternatively, you can set up a managed instance group such that the instance group manager will automatically provision new nodes when the others are reclaimed.
- To minimize preemption, use a larger number of small core machines
- There are preemptible GPUs too!

Use custom machine types

- If you know your workload characteristics, customize your machine-type with the specific number of cores and RAM needed to avoid paying for unutilized resources

Set budget notifications

- You can apply a budget to either a billing account or a project, and you can set the budget at a specific amount or match it to the previous month's spend. You can also create alerts to notify billing administrators and billing account users when spending exceeds a percentage of your budget.

Match your storage requirements with appropriate storage tiers

- Google Cloud Storage has four storage tiers with different pricing. Don't default to multi-regional storage.

Best Practices to control costs for BigQuery

- Query only the columns that you need.
- Don't run queries to explore or preview table data.
- Before running queries, preview them to estimate costs.

Rightsizing Recommendations

- Compute Engine provides machine type recommendations to help you optimize the resource utilization of your virtual machine instances. These recommendations are generated automatically based on system metrics gathered by the Google Stackdriver Monitoring service over the previous 8 days. Use these recommendations to resize your instance's machine type to more efficiently use the instance's resources.