|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Resource | Category | Metric | Threshold Value | Updated Threshold | Why is it important | Action to be performed as the follow-up |
| VM Instance | Agent | Memory Usage | 0.80% | No changes | Memory usage of the VM | It normally crosses the required Limit based on the usage.  Solution is to increase the memory for high process. |
|  | CPU | CPU Utilization(OS Reported) | 80% | No changes | It helps to reduce the crash of the instance | It normally reaches more than 80%  And at the maximum of 99.9% .  Solution is to increase the CPU size for high process. |
|  | Disk | Bytes read | 700000 B/s | No changes |  |  |
|  | Firewall | Dropped packets | 60 | No changes |  |  |
|  | Guest | Problem count | 0.003/s | No changes |  |  |
|  | Instance | CPU Utilization | 70% | 90% | CPU utilization of each instance | High level tensorflow instance can reach upto 99%. |
|  |  | VM Memory Total | 3221225472 B |  | Need to check the total usage for the project |  |
|  |  | VM Memory Used | 838860800 B |  | Need to check the total usage for the project |  |
|  |  | Scheduler wait time | 300 s | No changes |  |  |
| Pub/Sub | Cloud Pub/Sub Subscription | Ack latency | 200 ms | No changes | time taken when Pub/Sub sends a message to a subscriber and the Pub/Sub  topic receives an Acknowledge request for that message | If it exceeds then there is issue in on pub/sub topic |
|  |  | Delivery latency health score | 5 | 3 | Score based on ack\_latency  And ack\_deadlines.  Can find the performance of it | If score is less then we need to change the configuration of the topic or method we are using to subscribe the message from the topic |
|  |  | Oldest unacknowledged msg age | 5 |  | Subscribers not keeping up with message volume | * Need to add more subscriber threads or processes. |
|  |  | Push latency | 50 ms | 6 min | Time taken for rollup of the message from publisher to subscriber | If we are using pub/sub event based cloud function it can take more than  9 mins . Because it will respond only when the function is completed. |
|  |  | Pull latency |  |  | We don’t have any metric policy as pull latency |  |
|  |  | Subscription byte cost | 100 B | 1000 MiB | Cumulative cost of operations | It can go till GiB based on data size |
|  |  | Unacked msgs | 1 | 10 | Message unacknowledged by the subscriber. | If the subscriber is not acknowledged the message then there is issue In script where the messages are subscribed. Either the script is stopped because of issue |
|  | Cloud Pub/Sub Topic | Public msg operations | 16 | No change | Number of operation performed on a message  Like retry , ack deadline,  Dead letter |  |
|  |  | Publish msg size | 195 B | 100 MB | Message size which is published in the topic | It depends on the data size.  High volume of the data can consume more size |
|  |  | Topic byte cost | 15360 B | 100 MB |  | This is interconnected with  Publish msg size .  This is based on size of the data |
|  |  | Unacked msgs by region | 5 | 10 | Message unacknowledged by the subscriber for a region | This is interconnected  With Unacked msgs and we deploy in only one region  We no need this alert |
| Big Query Project | Query | Query execution times | 60ms | 30min | Time taken execute the a sql statement | Need to stop the job based on the query it is running.  If ML model creation we can ignore  Because it can take execution time from 30 min to 5 hours.  For select statement if it crosses threshold it needs to be stopped . |
|  |  | Statement scanned bytes billed | 20480 B | 4 GiB | Billing will be done based of bytes | It is also based on the statement  If it is a insert or delete statement we can ignore.  For select statement limit should be used in the statement |
| Big Query Dataset | Storage | Uploaded bytes billed | 1000 B | 6 GiB | Number of bytes of data gonna be inserted.  Billing will be based on this | If more records are gonna be inserted it can reach till 15GiB.  We cannot avoid that |
| Big Query BI Engine | Model | Request count | 1 | 3 | Total number of requests that were issued to a model  If it crosses 3 then those request are unwanted hit to the model | Need to stop the execution |
| Big Query DTS Config | Transfer\_config | Run latency distribution | 14400000 ms |  | We can remove this service  Because we are not gonna use bigquery data transfer |  |
| Cloud Functions | Functions | Execution time | 10000 ms | 9 min | Time taken for a single instance of complete the execution | Total timeout time for 10 min  Some script can execute for more than 9 mins. Normally when cloud function reaches 10 min it will stop the function so we can alert when it reaches 9 mins |
|  |  | Executions | 2 | 5 | The count of execution based on status  Important status  'ok', 'timeout', 'error', 'crash', 'out of memory', 'out of quota', 'load error', 'load timeout' | Need to need to function instead of retry |
|  |  | Active instances | 5 mins | 15 min | A instance which needs to be active for a time period | Maximum threshold is 23 hours  So we can alert it instead of any action. |
|  |  | Memory usage | 419430400 B | 2 GiB | Memory which takes run the function | It depends on the configuration of the specification  For 8 GiB machine we can have alert if it is more than 7 GiB |
| Consumer Quota | Quota | Allocation quota usage | 30 | Need to check based on project |  | Solution is remove the instance which is not used |
|  |  | Rate quota usage | 80 | Need to check based on project |  | Solution is to stop the instance which is not used .  Or to create a VM scheduler to start and stop the instances |
| GCS Bucket | Storage | Object count | 470 | 1000 | Number of objects per bucket | It depends on the usage of the bucket. |
|  |  | Total bytes | 3000000000 | 5 GB | Total size of the objects in the bucket | It depends on the usage of the bucket.If unwanted bucket is used for storing we can remove the bucket. |