Cloud Vision and AutoML Vision Service Level Agreement (SLA)

Last modified: January 9, 2020 | Previous Versions (#previous_versions)

During the Term of the agreement under which Google has agreed to provide Google Cloud Platform to Customer (the "Agreement"), the Covered Service will provide a Monthly Uptime Percentage to Customer as follows (the "Service Level Objective" or "SLO"):

Covered service	Monthly uptime percentage
Cloud Vision	>= 99.9%
AutoML Vision Deployment and Online Prediction for models deployed on 2 or more nodes	>= 99.9%
AutoML Vision Batch Prediction	>= 99.9%

If Google does not meet the SLO, and if Customer meets its obligations under this SLA, Customer will be eligible to receive the Financial Credits described below. This SLA states Customer's sole and exclusive remedy for any failure by Google to meet the SLO. Capitalized terms used in this SLA, but not defined in this SLA, have the meaning stated in the Agreement. If the Agreement authorizes the resale or supply of Google Cloud Platform under a Google Cloud partner or reseller program, then all references to Customer in this SLA mean Partner or Reseller (as applicable), and any Financial Credit(s) will apply only for impacted Partner or Reseller order(s) under the Agreement.

Definitions

The following definitions apply to the SLA:

- **"Back-off Requirements"** means, when an error occurs, the Customer is responsible for waiting for a period of time before issuing another request. This means that after the first

error, there is a minimum back-off interval of 1 second and for each consecutive error, the back-off interval increases exponentially up to 32 seconds.

- "**Covered Service**" means Cloud Vision, AutoML Vision Deployment and Online Prediction for models deployed on 2 or more nodes, or AutoML Vision Batch Prediction.
- "**Downtime**" means more than a 5% Error Rate. Downtime is measured based on server side Error Rate.
- "**Downtime Period**" means a period of one or more consecutive minutes of Downtime. Partial minutes or Intermittent Downtime for a period of less than one minute will not be counted towards any Downtime Periods.
- "**Error Rate**" means (i) with respect to Cloud Vision, the number of Valid Requests that result in a response with HTTP Status 500 and Code "Internal Error" divided by the total number of Valid Requests during that period; (ii) with respect to AutoML Vision Deployment and Online Prediction, the number of predict API calls that result in a response with HTTP Status 500 or 503 divided by the total number of predict API calls; and (iii) with respect to AutoML Vision Batch Prediction, the number of batchPredict API calls that result in a response with HTTP Status 500 or 503 divided by the total number of batchPredict API calls. Repeated identical requests do not count towards the Error Rate unless they conform to the Back-off Requirements.
- "**Financial Credit**" means the following:

Monthly uptime percentage	Percentage of monthly bill for the respective Covered Service which does not meet SLO that will be credited to future monthly bills of Customer
99% – < 99.9%	10%
95% – < 99.0%	25%
< 95%	50%

- "**Monthly Uptime Percentage**" means total number of minutes in a month, minus the number of minutes of Downtime suffered from all Downtime Periods in a month, divided by the total number of minutes in a month.
- "**Valid Requests**" are requests that conform to the Documentation, and that would normally result in a non-error response.

Customer must request financial credit

In order to receive any of the Financial Credits described above, Customer must notify Google technical support (https://support.google.com/cloud/contact/cloud_platform_sla) within 30 days from the time Customer becomes eligible to receive a Financial Credit. Customer must also provide Google with identifying information (e.g. project ID) and the date and time those errors occurred. If Customer does not comply with these requirements, Customer will forfeit its right to receive a Financial Credit. If a dispute arises with respect to this SLA, Google will make a determination in good faith based on its system logs, monitoring reports, configuration records, and other available information.

Maximum financial credit

The total maximum number of Financial Credits to be issued by Google to Customer for any and all Downtime Periods that occur in a single billing month will not exceed 50% of the amount due by Customer for the Covered Service for the applicable month. Financial Credits will be made in the form of a monetary credit applied to future use of the Service and will be applied within 60 days after the Financial Credit was requested.

SLA exclusions

The SLA does not apply to any: (a) features designated Alpha or Beta (unless otherwise stated in the associated Documentation), (b) features excluded from the SLA (in the associated Documentation), or (c) errors: (i) caused by factors outside of Google's reasonable control; (ii) that resulted from Customer's software or hardware or third party software or hardware, or both; (iii) that resulted from abuses or other behaviors that violate the Agreement; (iv) that resulted from quotas applied by the system or listed in the Admin Console; or (v) that resulted from Customer use of the Covered Service inconsistent with the Documentation, including but not limited to invalid request fields, unauthorized users, inaccessible data, or, with respect to AutoML Vision, use of a model that is beyond the recommended model lifespan described in the applicable Documentation for Image Classification (https://cloud.google.com/vision/automl/docs/models) and Object Detection (https://cloud.google.com/vision/automl/object-detection/docs/models).

Previous versions

- November 20, 2019 (https://cloud.google.com/vision/sla-20191120)
- November 19, 2018 (https://cloud.google.com/vision/sla-20181119)
