Deepgram

February 2024 On-Prem Release: Update Recommendation

In Deepgram's February on-prem release (`release-240228`), we add new functionality for token-based billing. This is relevant for our new audio intelligence features (sentiment analysis, intent recognition, topic detection), as well as our text intelligence endpoint (submitting text rather than audio as input to Deepgram). Usage of these features is measured by number of input and output tokens, rather than ASR's measurement of seconds of audio transcribed.

Due to this addition, the February on-prem release is not backwards-compatible with previous, but rather it is a breaking change in the API and License Proxy. This dependency also means that the updated version of the License Proxy node (`1.6.0`) must be running in advance of the updated version of the API node (`1.113.2`) serving requests. Note that the new version of the License Proxy (`1.6.0`) is compatible with previous versions of the API, so the License Proxy container can be deployed before the API container; the blue-green deployment strategy discussed in this document is one possible deployment strategy, but there are others that satisfy the requirement that the License Proxy is deployed first.

The Engine node is not impacted by breaking changes, but in the context of a complete Deepgram on-prem deployment, it is most cohesive to also include the update to the Engine node (`3.64.7`) in the blue-green deployment.

Blue-Green Deployment Overview

Blue-green deployments are an effective strategy for managing on-prem updates, especially in cases when backwards compatibility is not maintained and there are multiple container dependencies. The deployment model illustrated below involves maintaining two production environments during a deployment, Blue and Green. The Blue deployment is the current production environment, and the Green deployment is the update that is being rolled out. When deploying new changes that may not be compatible with the existing system, the new changes are initially released in the currently-unused Green environment. This allows you to deploy and test in a mirror of the live environment without impacting end-users. When ready, you can route your requests to the new deployment, requiring no downtime. If an issue surfaces in the new Green environment upon routing initial traffic to it, then a rollback to the previous Blue environment is instantaneous, ensuring system stability. Once the Green deployment has been completed, verified, and all production traffic shifted over, then the Blue deployment may be decommissioned, and the Green deployment is promoted to the new Blue deployment. This method helps reduce operational risk in updates, offering seamless transition and contingency plans.