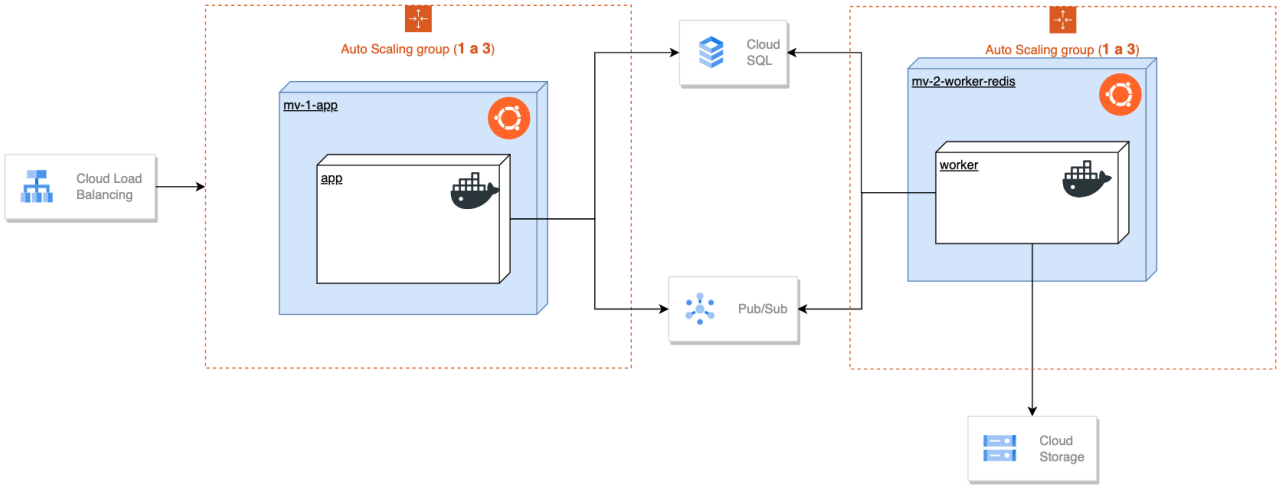
Arquitectura, conclusiones y consideraciones

Contenido

Arquitectura

Arquitectura de despliegue con servicios de GCP (Google Cloud Platform)



Consideraciones para escalar el sistema

Con esta nueva arquitectura para Cloud Conversion Tool (CCT), se puede evidenciar que la integración del servicio de Cloud Storage en nuestro sistema permitió escalar automáticamente el almacenamiento de los videos convertidos y originales; aquí se puede implementar una estrategia de retención de cada uno de los objetos para depurar estos archivos automáticamente sin ocupar espacio que no será necesario mantener.

La integración de un grupo de escalamiento de instancias sobre el aplicativo web resalta una eficiencia en la recepción de peticiones, dado que se configuro una tasa de escalamiento basado en el uso de la CPU en un 80%, de esta manera si el umbral es alcanzado por el tráfico de peticiones, automáticamente se creará una nueva instancia con todas las configuraciones necesarias para atender las solicitudes de nuestros clientes; de la misma forma como se desarrolló en la entrega pasada se incorporó un grupo de auto escalamiento para el procesamiento del Worker, pero en las pruebas de estrés identificamos que podemos tener un cuello de botella por el tipo de instancia o el número máximo de instancias Worker que podemos tener.

Finalmente, la integración del servicio Pub/Sub de GCP nos disminuyó la complegidad del sistema en configuraciones iniciales, ya que no teníamos que configurar instancias para exponer puertos para que pudiera ser visible por otra máquina. De esta manera, consideramos que el uso de este servicio mejoró en la capacidad de orquestación de los mensajes asíncronos y una mayor eficiencia en la entrega de los mensajes y el escalamiento del servicio sin afectar un componente clave de la operación.

Conclusiones

Como se menciona en las consideraciones, el sistema como aplicación cumple con las necesidades para el cliente y nos sentimos satisfechos de encontrar un MVP que puede soportar una transaccionalidad alta para ser expuesta a nuestros clientes y poder avanzar en mejoras de negocio. Sin embargo, debemos hacer un análisis más detallado al considerar cuantas maquinas deberían estar disponibles inicialmente en el grupo de escalamiento de la capa Worker y cuáles serían las unidades máximas que podríamos soportar como negocio. Con la arquitectura planteada hasta ahora, consideramos un cambio de 360º a la aplicación monolito que teníamos como prueba inicial.