ENTREGA 4 - ESCALABILIDAD
DISEÑO E IMPLEMENTACIÓN DE UNA APLICACIÓN WEB ESCALABLE EN NUBE PÚBLICA

OBJETIVOS

- Conocer las consideraciones técnicas que deben ser tenidas en cuenta para escalar la capa web de una aplicación sobre la infraestructura de un proveedor IaaS público.

- Implementar políticas de autoscaling que permitan que los servidores web (API REST) de la aplicación puedan escalar automáticamente (autoscaling + servicio de monitoreo).

- Definir una estrategia de despliegue de código y de inicio de los servidores (web y workers) para que los servidores de la aplicación puedan ser desplegados y escalados bajo demanda (sin que el administrador despliegue manualmente).

- Utilizar un balanceador de carga que distribuya la carga entre los diferentes servidores web de la aplicación.

- Implementar políticas de autoscaling que permitan que los procesos en batch (workers) puedan escalar automáticamente (autoscaling + monitorización).

- Utilizar un sistema escalable para el envío de emails administrado por un proveedor de nube pública.

- Configurar, implementar y utilizar el sistema de almacenamiento de objetos para cumplir con los requerimientos de almacenamiento de una aplicación Web escalable.

- Utilizar un sistema escalable de paso de mensajes entre los servidores Web y los workers como es el caso de las colas.

- Ejecutar pruebas de estrés de la aplicación en los servidores web.

- Ejecutar pruebas de estrés de la aplicación tanto para los servidores web como para los procesos que ejecutan trabajos en batch (workers).

TIEMPO DE DEDICACIÓN

Esta entrega “Escalabilidad” del proyecto está planeada para tres semanas, en las que cada participante deberá invertir las horas definidas en la planeación semanal.

FECHA DE ENTREGA

La entrega “Escalabilidad” deberá ser realizada de acuerdo con las fechas publicadas en la plataforma.

La socialización del proyecto estará a cargo de uno de los miembros del equipo de trabajo seleccionado aleatoriamente por ello es importante que los miembros del equipo conozcan en detalle el funcionamiento de toda la aplicación. Para el día de la sustentación **del modelo de despliegue ya debe estar funcionando en GCP**.

LECTURAS PREVIAS

Material de lectura entregado durante las semanas correspondientes y documentación disponible por el proveedor GCP para los servicios de Compute Engine, Cloud SQL, servicios de autoscaling y balanceadores de carga.

ESQUEMA DE EVALUACIÓN

La distribución de la calificación del proyecto está distribuida de la siguiente manera:

- Actividades requeridas para la migración inicial de la aplicación: 80%
- Documento de escenarios y resultados de las pruebas de estrés: 15%
- Documento de la arquitectura y consideraciones de la aplicación: 5%

LUGAR Y FORMATO DE ENTREGA

La entrega deberá ser realizada de la siguiente manera:

- Aplicación desplegada y en ejecución sobre GCP.
- Documentación de la aplicación, con el conjunto de instrucciones necesarias para la ejecución de esta.
- Crear un release del código fuente en el repositorio del grupo en GitHub/GitLab.
- Entregar toda la documentación vía GitHub/GitLab.

DOCUMENTACIÓN – ARQUITECTURA DE LA APLICACIÓN

Se deberá entregar un documento donde se describa la arquitectura de la aplicación, las conclusiones identificadas con las pruebas de estrés ejecutadas y las consideraciones que deben ser tenidas en cuenta para que la aplicación pueda escalar a cientos de usuarios finales que van a estar utilizando la aplicación web de manera concurrente. En este documento se deben describir las limitaciones del desarrollo realizado.

El nombre de este documento deberá ser: "Arquitectura, conclusiones y consideraciones"

RECOMENDACIONES Y CONSIDERACIONES

Para este proyecto es necesario que los **servidores web** (API REST) y los **workers** escalen automáticamente. Se mantiene el almacenamiento escalable basado en contenedores de objetos.

Se recomienda desplegar las instancias sólo cuando vayan a hacer pruebas sobre GCP y detener las instancias (acción stop) cuando no las vayan a utilizar, con el fin de evitar que se consuman la totalidad de los créditos asignados.

TAMBIÉN SE RECOMIENDA TENER CUIDADO AL HABILITAR POLÍTICAS DE AUTOSCALING, ya que una mala configuración puede hacer que se desplieguen silenciosamente nuevas instancias, **LO CUAL PUEDE LLEGAR A GENERAR A CONSUMIR TODOS LOS CRÉDITOS ASIGNADOS**. A**ctive las alarmas de consumo y los presupuestos para monitorear de forma continua la utilización de su cuenta de GCP.**

Deberán hacer los ajustes necesarios en el entorno de desarrollo para que puedan probar que toda la aplicación funciona adecuadamente. Servicios como el balanceador de cargas y autoscaling son difíciles de simular en el entorno local, por lo cual ese tipo de servicios los pueden probar directamente en el entorno de producción desplegado en el proveedor de nube.

Se recomienda hacer la configuración de los nuevos servicios y la modificación de las aplicaciones (web y worker) de manera incremental con el fin de que el trabajo se facilite. Tenga en cuenta que las condiciones para escalar los servidores web pueden ser muy diferentes a las condiciones para escalar los servidores workers (que serán escalados en la siguiente entrega).

Dado que durante el proyecto van a tener que trabajar con credenciales de acceso a GCP, se recomienda NO copiar las credenciales de GCP en texto plano en el código fuente de la aplicación, sino utilizar variables de entorno.

Nota Importante: Para la socialización de la entrega deben tener disponibles archivos de prueba en los diferentes formatos soportados por la aplicación y que permitan mostrar las funcionalidades de esta.

PARTE A - MODELO DE DESPLIEGUE - ESCALABILIDAD EN LA CAPA WEB

Con el fin de lograr que la aplicación web esté diseñada para escalar, se deberá modificar la aplicación desarrollada en la entrega anterior con el fin de que esta pueda responder de acuerdo con la demanda de usuarios. La compañía ya ha identificado que la aplicación deberá utilizar los siguientes servicios:

- **Compute Engine**: para la ejecución de los servidores Web y Worker. Por decisión de negocio se seleccionaron instancias de cómputo Serie N1 - F1 Micro, con 1 vCPU, 614 MiB en RAM y 10 GiB en almacenamiento.

- **Cloud SQL**: para almacenar los datos de la aplicación en una base de datos relacional. Seleccione una base de datos de desarrollo (Development) para minimizar los costos. En las fases iniciales de desarrollo puede usar una instancia de Compute Engine y reemplazarla solo para las pruebas de estrés a Cloud SQL.

- **Cloud Storage**: para el almacenamiento de los archivos originales y los procesados.

- **Cloud Monitoring**: para monitorear las instancias y los demás servicios utilizados.

- **Autoscaling**: Servicio para escalar la capa web de acuerdo con la demanda de usuarios.

- **Load Balancers**: para distribuir la carga entre los diferentes servidores Web.

Para **implementar el modelo propuesto de despliegue de la aplicación**, es necesario desarrollar las siguientes actividades:

1. **(20%) Este modelo tiene como objetivo realizar el despliegue de la aplicación bajo una arquitectura recomendada para aplicaciones escalables sobre GCP. Este despliegue incluye el uso de un balanceador de carga, hasta 3 servidores web, políticas de auto-scaling.** Configurar el servicio de balanceo de carga para poder desplegar varios servidores web**.**

2. **(5%)** Definir e implementar una estrategia para que los servidores Web escalan de manera automática. Se deberá definir en qué condiciones los servidores Web deberán escalar utilizando los servicios de monitoreo, autoscaling y de balanceo de cargas.

3. **(10%)** Configurar un bucket en el servicio de almacenamiento de objetos para almacenar todos los archivos subidos por los usuarios, tanto los originales como los procesados. Hacer los cambios en las aplicaciones **web** y **worker** para que los archivos (originales y procesados) se almacenen en un contenedor de objetos.

4. **(5%)** La aplicación web debe cumplir con todos los requerimientos funcionales que ya fueron definidos (que están mencionados completamente en el enunciado del proyecto). Eso implica que debe desplegar igualmente la instancia que soporta la capa worker, el sistema de base de datos y el sistema de envió de correos masivos, y validar el correcto funcionamiento de cada uno de los endpoints definidos en la entrega #1.

PARTE B - MODELO DE DESPLIEGUE - ESCALABILIDAD EN EL BACKEND

Con el fin de lograr que la aplicación Web esté diseñada para escalar, se deberá modificar la aplicación desarrollada en la entrega 3 con el fin de que ésta pueda escalar de acuerdo con la demanda de usuarios. La compañía ya ha identificado que la aplicación deberá utilizar los siguientes servicios:

- **Cloud Pub/Sub**: permite que las aplicaciones intercambien mensajes de forma confiable, rápida y asíncrona
- **Además de los servicios presentados en la parte A**.

El despliegue parte B, incluye las características y requerimientos del despliegue parte A y deben ser realizados los ajustes solicitados a continuación. Además, el despliegue incluye el uso de un balanceador de carga, hasta 3 servidores web, hasta 3 servidores workers, políticas de autoscaling, sistema de mensajes.

Para **implementar el modelo propuesto de despliegue de la aplicación**, es necesario desarrollar las siguientes actividades:

1. **(15%)** Diseñar e implementar una estrategia para que los servidores de procesamiento de archivos (workers) puedan escalar de manera automática. Se deberá definir en qué condiciones los workers deberán escalar utilizando los servicios de monitoreo, autoscaling y cola de mensajes.

2. **(15%)** Configurar el servicio de mensajes Cloud Pub/Sub, que será el sistema de comunicación a través del cual los servidores web y los procesos workers se comunicarán. Se deberán hacer los cambios necesarios para que ahora los servidores web coloquen las solicitudes para procesar nuevos archivos en una Cloud Pub/Sub y los workers saquen los archivos que tienen que procesar de dicha cola.

3. **(10%)** Configurar la capa web (API REST) de la aplicación en alta disponibilidad. Para esto configure el balanceador de carga y los nodos para que se desplieguen sobre dos zonas de disponibilidad de GCP.

ANÁLISIS DE CAPACIDAD

Con base a su documento de la “Entrega - Análisis de Capacidad” ejecute sus pruebas de desempeño de la aplicación.

Considere los dos escenarios propuestos

- **(7.5%)** Escenario 1 – Pruebas de Escalabilidad Parte A, es decir, pruebas de carga solo con los ajustes de escalabilidad a la capa web.

- **(7.5%)** Escenario 2 – Pruebas de Escalabilidad Parte B, es decir, pruebas de carga a la aplicación considerando la escalabilidad de la capa web y la capa de procesamiento asíncrono.