Simulación

UTN.BA

UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Profesores: Ing. Gladys Alfiero, Ing. Silvina Quiroga
Alumnos: Ariel Habib, Alejo Gurfein, Bibé Delfina

Estudio de una simulación para la agilización en el sistema de ventas de la distribuidora DFV utilizando un balanceador de carga

Contexto:

Durante la pandemia del 2020, la distribuidora veterinaria DFV S.R.L. se vio obligada a digitalizar sus operaciones y lanzó un servicio web para facilitar la compra de productos veterinarios. Debido a la urgencia del contexto, el sistema fue desarrollado con una arquitectura básica y diseñada para una capacidad moderada de usuarios ya que sus clientes, acostumbrados a realizar compras físicas, no utilizaban mucho la plataforma digital en sus primeros meses de operación.

Sin embargo, con el paso del tiempo, no solo aumentó la cantidad de usuarios que adoptaron la plataforma (debido a mejoras del sistema, mayor confianza en la compra online y el aumento de la demanda en el sector), sino que también se incrementó el número de clientes de la distribuidora. En consecuencia, el crecimiento sostenido en la demanda expuso las limitaciones del sistema, el cual no logra procesar eficientemente un volumen elevado de peticiones concurrentes, generando cuellos de botella y demoras considerables.

Ante este escenario, DFV ha decidido contratar nuestros servicios para diseñar una solución que permita mejorar los tiempos de respuesta y manejar una mayor cantidad de clientes de manera más eficiente. Para ello nos ha proveído un historial de servicios (tiempo y nombre de servicio) que atendió estos últimos dos años.

Descripción del Sistema Actual:

El sistema no es particularmente complejo. El equipo de sistemas nos proporcionó los servicios actuales en los que están interesados y sus métricas de uso.

- **Buscar ofertas (35%)**
- **Buscar productos veterinarios (60%)**
- **Realizar compras de productos veterinarios (5%)**

Propuesta de Escalabilidad:

DFV establece estas tres estrategias para mejorar la infraestructura y atender mejor la demanda de los usuarios:

1. **Compra de Servidores Físicos Adicionales:**
    - **Costo fijo por servidor:** USD 2.000.
    - **Tiempo de arranque:** Instantáneo (ya están encendidos permanentemente).
    - **Ventaja:** Capacidad constante de atención sin dependencia de red.
    - **Desventaja:** Riesgo de ociosidad cuando la demanda es baja, lo que podría generar gastos innecesarios.
2. **Uso de Servicios Cloud (ej., AWS EC2) con escalamiento horizontal:**
    - **Costo:** USD 0,38 por hora por instancia activa.
    - **Tiempo Inicio**: Generalmente tarda entre **30 segundos a 2 minutos**, dependiendo del tipo de instancia, AMI, configuración de red, etc.
    - **Tiempo de bajada**: Entre **10 a 30 segundos**.
    - **Ventaja:** Escalabilidad dinámica, con la posibilidad de activar o desactivar servidores según la carga en tiempo real.
    - **Desventaja:** Mayor latencia en el arranque de instancias.
3. **Uso de un WSGI Server (worker) para la atención de servicios (Ej: Gunicorn)**

    Para distribuir la carga de trabajo entre las distintas opciones, DFV desea implementar un balanceador de carga que reparta las solicitudes entrantes entre las unidades de atención disponibles, ya sean servidores físicos o instancias en la nube, en función de la carga actual. El balanceador asignará cada solicitud al servidor que tenga menor carga en ese momento.

    Cada servidor contará con una cantidad determinada de *workers*, los cuales permitirán procesar las solicitudes de manera paralela. Para calcular la cantidad de

workers por servidor, se utiliza el siguiente estándar:

Workers = 2 × Número de CPUs + 1.

Actualmente, **el servidor físico** utilizado por la distribuidora posee 10 núcleos, lo que equivale a **21 *workers***. En caso de requerir nuevos servidores físicos, planean continuar utilizando este mismo modelo de servidor.

Por otro lado, para el escalado horizontal en la **nube**, se contempla el uso de instancias AWS “t4g.medium”, las cuales disponen de 4 núcleos, resultando en **9 *workers* por instancia**. En cuanto al escalado de las instancias en la nube, se nos indicó que no se planea implementar escalado vertical; en cambio, se optará por un enfoque de escalado horizontal, generando nuevas instancias de AWS a medida que aumente la demanda de recursos.

Se nos informó también que el balanceador de carga que desea utilizar DFV realiza *una medición de la carga actual del sistema cada 10 segundos.*

Objetivos de la Simulación:

DFV desea simular distintos escenarios con el fin de determinar cuál será la mejor disposición de arquitectura para poder atender a sus clientes actuales y tener la posibilidad de seguir escalando, si llegara a ser necesario, sin necesidad de volver a hacer un análisis tan profundamente. La mejor arquitectura será la que minimice sus gastos y atenderá a todos los clientes en tiempo y forma.

Con este fin, se sentaron ciertas bases como referencia para el estudio:

- Un servidor se lo considerara ocioso si tiene 15% o menos de sus hilos totales en uso
- Un servidor se lo considerara en su capacidad máxima sí tiene 90% o más de sus hilos totales en uso
- Un servidor idealmente tiene que tener en uso entre un 65% y un 89% de sus hilos

Para hacer esta evaluación, se simulará utilizando servidores físicos comprados por la empresa y un híbrido que utilice ambas alternativas.

Para cada alternativa, será de suma importancia analizar el tiempo promedio de espera de los usuarios para los diferentes tipos de servicios y porcentaje de tiempo ocioso de los servidores. Para lograr esto se tendrá en cuenta el tiempo promedio de respuesta para cada petición.

A su vez, se nos pidió minimizar el porcentaje de tiempo en que los servidores físicos están en su capacidad máxima para alargar su vida útil. Teniendo en cuenta esto, los resultados obtenidos de los tiempos de levante y bajada para la empresa se sugerirá la cantidad óptima de servidores y su tipo correspondiente.

Con el fin de brindar una respuesta que no solo sirva para la actualidad pero también a futuro, DFV nos solicita responder las siguientes preguntas:

1. Dimensionamiento de servidores físicos

Pregunta:
¿Cuántos servidores físicos debería tener la empresa para cubrir la demanda promedio sin incurrir en tiempos de ociosidad significativos ni generar demoras innecesarias?

Objetivo:
Determinar la cantidad mínima de servidores físicos necesaria para atender eficientemente la carga promedio, evitando tanto la sobreasignación de recursos como la saturación del sistema.

Variables de resultado:

- Porcentaje de tiempo ocioso de los servidores físicos
- Promedio de peticiones por servidor físico
- Porcentaje de tiempo de workers activos
- Porcentaje de tiempo de workers ociosos
- Promedio de tiempo de demora de cada petición

---

2. Eficiencia del uso de servidores en la nube (cloud)

Preguntas clave:

- ¿Cuál es el comportamiento del sistema en términos de escalado automático?
- ¿Con qué frecuencia y efectividad se activan o desactivan instancias cloud?
- ¿Qué porcentaje del tiempo están activas las instancias y cuánto tiempo permanecen ociosas?

Objetivo:
Analizar el uso de los recursos cloud (AWS) para comprender la eficiencia del balanceo automático y estimar los costos operativos y no operativos asociados a la infraestructura dinámica.

Variables de resultado:

- Porcentaje de tiempo ocioso de servidores cloud

- Promedio de tiempo de subida de instancia
- Promedio de tiempo de bajada de instancia
- Promedio de efectividad del balanceo
- Promedio de peticiones por servidor cloud

---

3. Priorización de servicios críticos para optimización

Pregunta:
¿Cuál de los servicios principales debería ser priorizado para su optimización, en función de su frecuencia de uso y su impacto directo en la experiencia del usuario?

Servicios a evaluar:

- buscar ofertas
- buscar productos
- realizar compras

Objetivo:
Identificar cuál de los servicios representa el mayor cuello de botella y tiene mayor potencial de mejora en términos de tiempos de respuesta y percepción del usuario.

Variables de resultado:

- Promedio de demora en buscar ofertas
- Promedio de demora en buscar productos
- Promedio de demora en realizar compras

Análisis de Variables:

Tipos de Variables		Nombre	Descripción
Datos	Exógena	IAP	intervalo arribo peticion
		DBO	demora buscar_oferta
		DBP	demora buscar_productos
		DRC	demora realizar_compra
		TAC	tiempo arranque cloud
		TBC	tiempo bajada cloud
Control	Exógena	CSF	Cantidad servidores físicos
Estado	Endógena	NBP	Peticiones buscar_prod en sistema
		NBO	Peticiones bucar_oferta en sistema
		NRC	Peticiones realizar_compra en sistema
		SCA	Servidores Cloud Activos
		WOFi	workers ocupados en el servidor físico i
		WOCi	workers ocupados en el servidor cloud i
		Flag(i)	Flag para indicar el tipo de petición que se atiende; 0 para buscar producto, 1 para buscar oferta, 2 para comprar producto
Resultado	Endógena	PTOSF	Porcentaje tiempo ocioso servidores físicos
		PTOSC	Porcentaje tiempo ocioso servidores cloud
		PTSI	Porcentaje tiempo subida de instancia

		PTBI	Porcentaje tiempo bajada de instancia
		PEB	Porcentaje efectividad balanceo
		PPSF	Promedio peticiones por servidor físico
		PPSC	Promedio peticiones por servidor cloud
		PDBO	Promedio demora buscar_oferta
		PDBP	Promedio demora buscar_productos
		PDRC	Promedio demora realizar_compras
		PPR	Promedio Peticiones redirigidas

Tabla de eventos independientes:

Eventos	**EFnC**	**EFC**	**Condición**
Llegada	Llegada	Salida(i,j)	PS <= 21 * CSF + 9*SCA
Medición Carga	Medición Carga	Levantar Servidor Cloud	(CWFO+ CWCO <= CSF*21*0.65 + SCA*9*0.65) && SCA > 0
		Bajar Servidor Cloud	CWFO+ CWCO >= CSF*21*0.9 + SCA*9*0.9
Levantar Servidor Cloud			
Bajar Servidor Cloud			
Salida(i,j)	Salida(i,j)		

Tabla de eventos futuros:
TPLLP: Tiempo próxima llegada petición
TPMC: Tiempo próxima medición de carga
TPLS: Tiempo próxima levantada de servidor
TPBS: Tiempo próxima bajada de servidor
TPS(i,j): Tiempo próxima salida server i worker