Definición de servicios

Philipp Maier
Desarrollador de cursos, Google Cloud

En este módulo, nos enfocaremos en la definición de los servicios.

Los desarrollos nuevos comienzan por planificar y diseñar las fases. Esto requiere recopilar información y se debe comenzar con los requisitos empresariales. Una vez que se definan los requisitos, es importante medir que proporcionen valor empresarial. En este módulo, hablaremos sobre la recopilación de requisitos y, luego, sobre las técnicas para medir el impacto de las soluciones.

Analicemos con más detalle los temas del módulo.

Los requisitos cualitativos definen los sistemas desde el punto de vista del usuario

Quiénes	¿Quiénes son los usuarios? ¿Quiénes son los desarrolladores? ¿Quiénes son las partes interesadas?
Qué	¿Qué hace el sistema? ¿Qué funciones principales tiene?
Por qué	¿Por qué es necesario el sistema?
Cuándo	¿Para cuándo los usuarios necesitan o quieren la solución? ¿Cuándo pueden terminar el trabajo los desarrolladores?
Cómo	¿Cómo funcionará el sistema? ¿Cómo se definirá la cantidad de usuarios? ¿Cómo se definirá la cantidad de datos?

Estas son preguntas útiles que los arquitectos de la nube deben hacerse para definir los requisitos: ¿Quiénes? ¿Qué? ¿Por qué? ¿Cuándo? ¿Cómo?

- “Quiénes” sirve para determinar no solo a los usuarios del sistema, sino también a los desarrolladores y las partes interesadas. El objetivo es obtener un panorama completo de las personas sobre las que influirá el sistema, tanto directa como indirectamente.
- Responder “qué” es fácil y difícil al mismo tiempo. Debemos establecer las principales áreas de funcionalidad requeridas, pero de una manera clara y sin ambigüedades.
- “Por qué” es necesario el sistema es una pregunta muy importante. ¿Cuál es el problema que busca abordar o solucionar el sistema propuesto? Si no se entiende bien la necesidad, es probable que se deban incluir requisitos adicionales. El *porqué* también podrá ayudar a definir los KPI, los SLO, los ANS, etcétera.
- “Cuándo” determina un cronograma realista y puede ayudar a limitar el alcance.
- “Cómo” permite determinar muchos de los requisitos no funcionales, que pueden ser, por ejemplo, la cantidad de usuarios que el sistema debe admitir de forma simultánea, cuál es el tamaño promedio de la carga útil de las solicitudes de servicio, si hay requisitos de latencia, etc. También pueden ser que los usuarios se ubicarán en todo el mundo o solo en una región en

Los roles representan el objetivo de un usuario en un momento determinado

Los roles no son personas ni cargos	Los roles deben describir los objetivos de los usuarios	Ejemplos de roles
• Las personas pueden tener varios roles. • Varias personas pueden desempeñar un mismo rol.	• ¿Qué desea hacer el usuario? • "Usuario" no es un rol recomendable porque *todos son usuarios*	• Comprador • Titular de la cuenta • Cliente • Administrador • Gerente

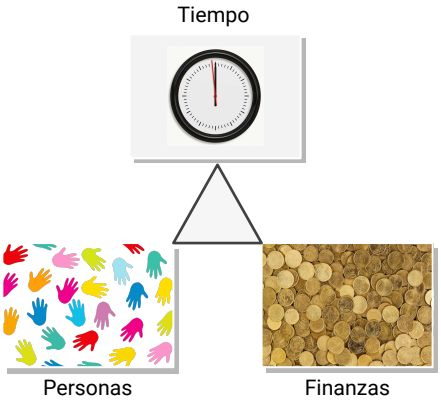
En la actividad de diseño anterior, definió los roles del usuario para su aplicación. Los roles representan el objetivo de un usuario en un momento determinado y permiten analizar un requisito en un contexto particular. Es importante mencionar que un rol no es necesariamente una persona, sino que un actor en el sistema y puede ser otro sistema, como un cliente de un microservicio que accede a otro microservicio.

El rol debe describir el objetivo del usuario cuando utiliza el sistema. Por ejemplo, el rol de un comprador en una aplicación de comercio electrónico define claramente lo que el usuario desea hacer. Existen varias formas de determinar los roles de los requisitos en los que está trabajando. Aquí encontrará un proceso que funciona particularmente:

- Primero, intercambie ideas para un conjunto inicial de roles. Escriba todos los que se le ocurran (cada uno debe ser de un solo usuario).
- Ahora organice el conjunto inicial: puede identificar los roles que se superponen y se relacionan, y agruparlos.
- Con el conjunto de roles agrupado, consolide los roles. El objetivo de este paso es consolidar y resumir los roles para quitar los duplicados.
- Por último, defina mejor los roles, incluidos los internos y externos, y los distintos patrones de uso. En este punto, puede proporcionar información adicional, como el nivel de experiencia del usuario en el campo o la frecuencia de uso del software propuesto.

Seguir un proceso sencillo como este proporciona estructura y enfoque a la tarea.

Los requisitos cuantitativos son elementos medibles



Dadas estas restricciones:

- Tiempo
- Finanzas
- Personas

Qué se puede lograr:

- ¿Cuántos usuarios hay?
- ¿Cuántos datos hay?
- ¿Cuáles son los beneficios y los riesgos?
- ¿Qué funciones se pueden lanzar?

Para administrar bien un servicio, es importante comprender qué comportamientos importan y cómo medirlos y evaluarlos. Siempre deben considerarse en el contexto de las restricciones, que suelen ser el tiempo, las finanzas y las personas. Luego, debemos considerar lo que se puede lograr. El tipo de sistema que se evalúa determina los datos que se pueden medir.

Por ejemplo, en el caso de los sistemas orientados a los usuarios, ¿se respondió una solicitud? (disponibilidad); ¿cuánto tiempo tardó la respuesta? (latencia) y ¿cuántas solicitudes se pueden manejar? (capacidad de procesamiento).

En el caso de los sistemas de almacenamiento de datos, ¿cuánto tarda la lectura y escritura de datos? (latencia); ¿hay datos cuando se necesitan? (disponibilidad) y, si ocurre una falla, ¿se perderán datos? (durabilidad).

La clave para todos estos elementos es que las preguntas pueden responderse con los datos recopilados de los servicios.

Los indicadores clave de rendimiento (KPI) son métricas que pueden usarse para medir el éxito

En los negocios, los KPI más comunes incluyen los siguientes:

- Retorno de la inversión (ROI)
- Ganancias antes de los intereses y los impuestos (EBIT)
- Reemplazo de los empleados
- Deserción de los clientes

En el ámbito del software, los KPI más comunes incluyen los siguientes:

- Páginas vistas
- Registros de usuarios
- Proporción de clics
- Confirmaciones de compra

Los encargados de las decisiones de negocios quieren medir el valor de los proyectos. Esto les permite dar más apoyo a los proyectos más valiosos y no desperdiciar recursos en los que no son beneficiosos. Una forma habitual de medir el éxito es usar KPI. Los KPI se pueden categorizar como comerciales y técnicos.

Los KPI comerciales son una manera formal de medir lo que valora el negocio, como el ROI, en relación con un proyecto o servicio. Otros incluyen las ganancias antes de los intereses y los impuestos, o el impacto en los usuarios, como la deserción de clientes o la rotación de personal.

En los KPI técnicos o de software, se pueden considerar aspectos como la eficacia del software a través de las páginas vistas, los registros de usuarios y la cantidad de confirmaciones de compra. Estos KPI también deben estar alineados estrechamente con los objetivos comerciales.

Como arquitecto, es importante que comprenda cómo el negocio mide el éxito de los sistemas que diseña.

Los KPI indican si está bien encaminado para cumplir el objetivo

Objetivo: Aumentar las ventas de una tienda en línea
KPI: El porcentaje de conversiones en el sitio web

Un KPI no es lo mismo que un objetivo. El objetivo es el resultado que quiere lograr. El KPI es una métrica que indica su progreso para alcanzar el objetivo.

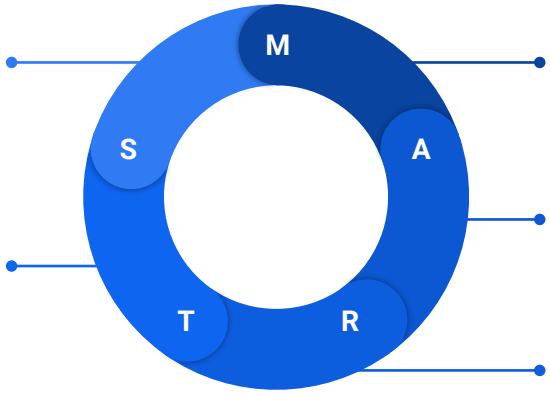
Para aprovechar al máximo los KPI, estos requieren de un objetivo complementario. Ese debe ser el punto de partida para definir los KPI. Luego, para cada objetivo, debe definir los KPI que le permitirán supervisar y medir el progreso. Para cada KPI, debe establecer cómo se define el éxito. Supervisar los KPI en función de los objetivos es importante a fin de lograr el éxito y permite realizar ajustes según los comentarios.

Por ejemplo, un objetivo puede ser aumentar las ventas de una tienda en línea y un KPI asociado puede ser el porcentaje de conversiones en el sitio web.

Para que los KPI sean eficaces, deben cumplir con los criterios SMART

Específico
"Fácil de usar" no es tan específico como "accesible en virtud del artículo 508".

Limitado en el tiempo
99% disponible: ¿Al año? ¿Al mes? ¿Al día? Si no sabemos, ¿cómo podemos medirlo?



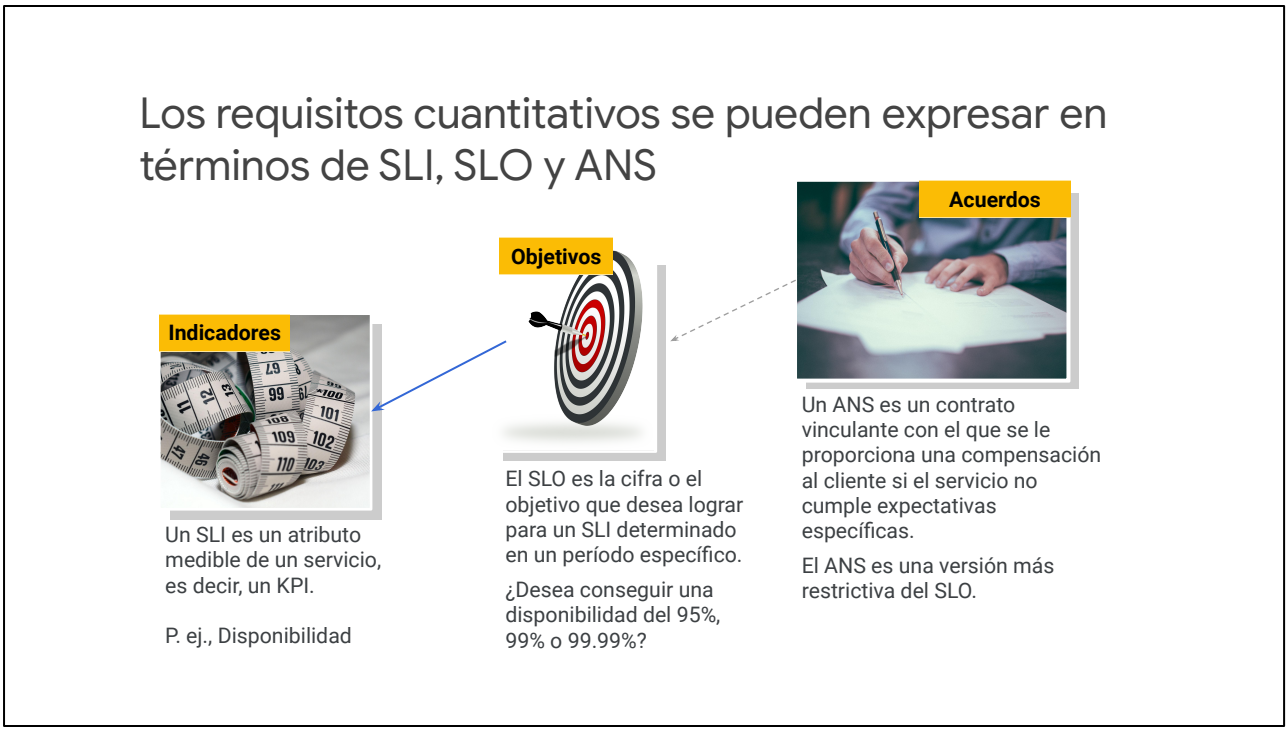
Medible
Debe buscar una manera objetiva de comprobar si está alcanzando los KPI.

Alcanzable
"100% de disponibilidad" puede sonar bien, pero no es posible.

Relevante
¿De verdad es importante para el usuario? ¿Ayudará a lograr los objetivos de la aplicación?

Que sea limitado en el tiempo ayuda a medir el KPI. Algunos KPI están más limitados por el tiempo. Por ejemplo, ¿la "disponibilidad" es por día, mes o año?

En resumen, los KPI se usan para medir el éxito o el progreso con respecto a un objetivo.



Hablemos sobre la terminología del nivel de servicio. Para ofrecer un determinado nivel de servicio a los clientes, es importante definir los indicadores (SLI), los objetivos (SLO) y los acuerdos (ANS) de nivel de servicio. Son mediciones que describen propiedades básicas de las métricas que se evaluarán, los valores que las métricas deben analizar y cómo reaccionar si no se pueden alcanzar las métricas.

El indicador de nivel de servicio es una medida cuantitativa sobre algún aspecto del nivel de servicio que se proporciona. Entre los ejemplos se incluyen la capacidad de procesamiento, la latencia y la tasa de error.

El objetivo de nivel de servicio es un objetivo o un rango de valores acordado para un nivel de servicio que mide un SLI. Por lo general, se expresa en el siguiente formato: SLI ≤ objetivo O límite inferior ≤ límite superior. Un ejemplo de un SLO es que la latencia promedio de las solicitudes HTTP de nuestro servicio debe ser menor que 100 milisegundos.

El Acuerdo de Nivel de Servicio es un acuerdo entre un proveedor de servicios y un consumidor. Define las responsabilidades en cuanto a la entrega del servicio y las consecuencias cuando estas no se cumplen. El ANS es una versión más restrictiva del SLO. Lo ideal es diseñar una solución y mantener un SLO acordado para tener capacidad libre con respecto al ANS.

Los SLO deben ser alcanzables y relevantes

SLI	SLO	
Las cargas de fotos HTTP POST se completan dentro de 100 ms agregados por minuto	**99%**	✖ Si los usuarios utilizan teléfonos celulares, quizás es excesivo.
	80%	✔ Este valor puede ser suficiente.
Disponible según la medición con una verificación de tiempo de actividad cada 10 segundos, agregados por minuto	**100%**	✖ Parece una buena idea, pero no es práctica.
	99.999%	✖ Es posible, pero tal vez sea demasiado costoso.
	99%	✔ Quizás es suficiente, más fácil y más rentable.

La relevancia de los SLO es fundamental. Le recomendamos tener objetivos que ayuden o mejoren la experiencia del usuario. Es fácil definir SLO en función de elementos que son fáciles de medir en lugar de útiles. Para mayor claridad, los SLO deben especificar cómo se miden y las condiciones en las que son válidos.

Suponga que la disponibilidad se mide con una verificación de tiempo de actividad de más de 10 segundos, agregados por minuto. No es realista ni deseable tener SLO con un objetivo del 100%. Ese objetivo genera soluciones costosas y demasiado conservadoras que, de todos modos, tienen pocas probabilidades de alcanzar el SLO. Es mejor hacer un seguimiento del porcentaje con el que los SLO no se alcanzan y trabajar para mejorarlo. En muchos casos, un 99% de disponibilidad puede ser suficiente y mucho más fácil de lograr y diseñar. También es muy probable que sea más rentable de ejecutar.

También se debe tener en cuenta el caso de uso. Por ejemplo, si un servicio HTTP para cargas de fotos requiere que un 99% de las cargas se completen dentro de 100 ms agregados por minuto, puede ser una cifra poco realista o excesiva si la mayoría de los usuarios usan teléfonos celulares. En ese caso, un SLO del 80% es mucho más alcanzable y es suficiente.

A menudo, es buena idea especificar varios SLO. Tenga en cuenta lo siguiente:

Un 99% de las llamadas GET de HTTP se completarán en menos de 100 ms

Sugerencias para determinar los SLO

- El objetivo no es que los SLO sean lo más altos posible, sino que sean tan bajos como se pueda y que, al mismo tiempo, satisfagan las necesidades de los usuarios. Por eso es importante comprenderlos.
- Mientras más alto establezca el SLO, mayor será el costo en recursos de procesamiento (redundancia) y tareas de operaciones (tiempo-persona).
- Las aplicaciones no deben superar de forma significativa sus SLO, ya que los usuarios esperarán el nivel de confiabilidad que generalmente reciben.

Seleccionar SLO tiene consecuencias en el producto y el negocio. A menudo, se deben compensar las limitaciones, como el personal, el tiempo de salida al mercado y el financiamiento. Como se indica en la diapositiva, el objetivo es mantener a los usuarios satisfechos, no tener un SLO que requiera de esfuerzos sobrehumanos para lograrlo. Le daré algunas sugerencias para seleccionar los SLO:

- No apunte demasiado alto: Es mejor tener SLO más bajos para comenzar y ajustarlos con el tiempo a medida que conozca el sistema, en lugar de definir SLO que no sean alcanzables y que requieran de grandes esfuerzos y costos para lograrlos.
- Asegúrese de que sean simples: Los SLI más complejos pueden ocultar los cambios importantes en el rendimiento.
- Evite los valores absolutos: Tener un SLO que indica un 100% de disponibilidad no es realista. Un SLO de ese tipo aumenta el tiempo de compilación, la complejidad y el costo de operación y, en la mayoría de los casos, es poco probable que sea necesario.
- Minimice los SLO: Un error común es tener demasiados SLO. Es recomendable que tenga solo los suficientes para abarcar los atributos clave del sistema.

En resumen, los buenos SLO deben reflejar lo que les importa a los usuarios. Funcionan como un estímulo para los equipos de desarrollo. Un SLO deficiente generará mucho trabajo perdido (si es demasiado ambicioso) o un producto de mala

Un ANS es un contrato comercial entre el proveedor y el cliente

El ANS estipula lo siguiente:

- La penalización que se aplicará al proveedor si el servicio no cumple con ciertos umbrales de disponibilidad o rendimiento.
- Si el ANS no se cumple, el cliente recibirá una compensación por parte del proveedor.

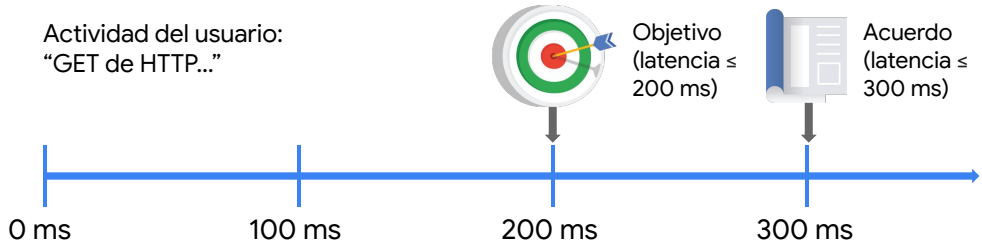
No todos los servicios tienen ANS, pero todos los servicios deben tener un SLO.

Sus umbrales de SLO deben ser más estrictos que el ANS.

Un ANS es un contrato comercial entre el proveedor de servicios y el cliente. Se aplicará una penalización si el proveedor no mantiene los niveles acordados. No todos los servicios tienen ANS, pero todos los servicios deben tener SLO.

Al igual que con los SLO, es mejor ser conservador con los ANS, ya que es muy difícil cambiar o quitar aquellos que ofrecen poco valor o generan un exceso de trabajo. Además, debido a que pueden tener consecuencias financieras a través de la compensación a los clientes, establecer ANS demasiado altos puede causar que se deban pagar compensaciones innecesarias. Para proporcionar protección y cierto nivel de seguridad, un ANS debe tener un umbral más bajo que el SLO. Siempre debe ser así.

Ejemplo: SLI, SLO y ANS



Latencia de extremo a extremo de las respuestas HTTP exitosas (HTTP-200) (con un promedio de más de 1 minuto)

- SLI: La latencia de las respuestas HTTP exitosas (HTTP-200)
- SLO: La latencia del 99% de las respuestas debe ser ≤ 200 ms
- ANS: El usuario recibe una compensación si la latencia del percentil 99 excede los 300 ms.

Consideremos un ejemplo de servicio y sus SLI, SLO y ANS. El servicio es un extremo de HTTP al que se accede mediante un GET de HTTP.

El SLI es la latencia de extremo a extremo de las respuestas HTTP exitosas, es decir, las HTTP 200. Se calcula el promedio de ellas durante un minuto. En el SLO, se acordó que la latencia del 99% de las respuestas debe ser menor o igual que doscientos milisegundos.

El ANS indica que el usuario recibirá una compensación si la latencia del percentil 99 excede los 300 ms. El ANS compiló claramente un búfer sobre el SLO, lo que significa que, incluso si se excede el SLO, hay algo de capacidad disponible antes de que se incumpla el ANS. Esta es la posición deseada en la relación entre el SLO y el ANS.

Revisión de la actividad 3: Defina los SLI y SLO

- Escriba los SLI y SLO para las funciones de su caso de éxito.

En la tercera actividad, se le solicitó escribir los SLI y SLO para su caso de éxito.

Historia de usuario	SLO	SLI
Búsqueda de hoteles y vuelos	Disponibilidad de un 99.95%	Fracción de respuestas HTTP 200 frente a HTTP 500 desde el extremo de la API (medición mensual)
Búsqueda de hoteles y vuelos	El 95% de las solicitudes se completará en menos de 200 ms	Tiempo hasta el último byte de solicitudes GET medido cada 15 segundos, agregado por cada 5 minutos
Inventario de suministros de hotel	Tasa de errores menor que el 0.00001%	Errores de carga medidos como un porcentaje de las cargas masivas que se subieron al día según una métrica personalizada
Inventario de suministros de hotel	Disponibilidad de un 99.9%	Fracción de respuestas HTTP 200 frente a HTTP 500 desde el extremo de la API (medición mensual)
Analizar el rendimiento de ventas	El 95% de las consultas se completará en menos de 10 s	Tiempo hasta el último byte de solicitudes GET medido cada 60 segundos, agregado por cada 10 minutos

Estos son algunos ejemplos de los SLO y SLI de nuestra aplicación del portal de viajes. Tenga en cuenta que el SLI describe lo que vamos a medir y cómo lo haremos; por ejemplo, “Fracción de respuestas HTTP 200 frente a HTTP 500 desde el extremo de la API (medición mensual)”. Este ejemplo es una forma de medir la disponibilidad.

El SLO representa el objetivo que intentamos alcanzar para un SLI determinado. Por ejemplo, “Disponibilidad de un 99.95% del tiempo”.

No dude en pausar el video para leer los otros SLO y SLI de cada historia de usuario.