

Sebastián Valencia  
201111578

La escalabilidad es el atributo de calidad que se refiere a la habilidad de un sistema de software de crecer o manejar una mayor carga a la actualmente soportada de lo que actualmente lo soporta. La escala de un sistema de software, se refiere a lo grande que es el sistema en líneas de código, mientras la escalabilidad es un atributo de calidad que se refiere a la habilidad del sistema anteriormente mencionadas. Un ejemplo es correr en hardware más robusto (escalabilidad horizontal), o usar más unidades de hardware (escalabilidad vertical). En términos generales, la escalabilidad se refiere a la habilidad de manejar cargas de trabajo crecientes. Tal crecimiento se puede deber al aumento en procesamiento, transacciones, peticiones, mensajes, o trabajos del sistema requeridos para procesar por unidad de tiempo la complejidad creciente para administrar la carga mayor debida a éstas tareas.

En concreto, la escalabilidad horizontal se refiere a agregar más recursos a las unidades lógicas que soportan el procesamiento, un ejemplo de esto, es agregar otro servidor al clúster de servidores. La escalabilidad vertical se refiere a agregar más recursos a una unidad física, un ejemplo de esto, es agregar más memoria a un solo computador.

La elasticidad se refiere a la propiedad que le permite al usuario agregar o remover máquinas virtuales del pool de recursos. Estas máquinas virtuales están generalmente alojadas sobre una colección grande de máquinas físicas que son manejadas por el proveedor del servicio (IaaS).

En un lenguaje menos técnico si se quiere que la escalabilidad es la capacidad del sistema de acomodar cantidades grandes de carga al agregar hardware más robusto, o más unidades de hardware. En el mismo contexto, la elasticidad es la habilidad de acomodar los recursos con cargas dinámicas para escalar. Es decir, si la carga aumenta, se debe agregar más recursos, si la carga disminuye, se debe remover los recursos sub utilizados. La elasticidad en recursos, debe estar relacionada con los costos asociados al servicio según los recursos utilizados por el cliente (IaaS, SaaS).

Para medir la escalabilidad de un sistema, se debe considerar una métrica que indique que tan bien se maneja la demanda creciente. Las métricas que soportan ésta especificación, pueden ser asociadas a las siguientes medidas:

- Tasa a la cual las transacciones son procesadas por el sistema. Para medir, se asocia el nivel de concurrencia y las transacciones.
- El nivel de uso de varios recursos computacionales (memoria, CPU, disco, ancho de banda). Se mide en procesos, u otra métrica asociada al costo de computación.
- Precio por transacción. (Precio económico, o precio asociado al procesamiento).

Los nombres asociados a las medidas señaladas son throughput, utilización de recursos, costo. Dado que la elasticidad está asociada con un tipo de escalabilidad, puede medirse con métricas similares que contemplen el cambio de recursos con las curvas de demanda, es decir, tasa entre demanda y capacidad de cómputo.

Recursos.

Islam, Lee, Fekete, Liu. How a consumer can measure elasticity for cloud platforms.

Bass, Clements, Kazman. Software architecture in practice.

Fairbanks-Just enough software architecture.