Introducción a Big Data.

En esta unidad de trabajo vamos a comenzar a conocer el fenómeno conocido como Big Data (o macrodatos).

Comenzaremos entendiendo las circunstancias que originan la aparición de las metodologías y tecnologías para Big Data y veremos qué conseguimos gracias a su uso.

A continuación, veremos qué es un clúster de computadoras, los cuales constituyen la infraestructura básica en la que se apoyan los sistemas Big Data.

Más adelante repasaremos una serie de conceptos muy importantes a la hora de terminar de comprender todo el fenómeno Big Data, tanto relacionados con almacenamiento de datos como con procesamiento de los mismos.

Por último, veremos cuál es la arquitectura en capas que generalmente se utiliza en proyectos Big Data para y terminaremos viendo lo que viene en llamarse el paisaje de Big Data.

1.- Por qué Big Data.

Las metodologías y tecnologías para Big Data aparecen como respuesta la necesidad de tratar cantidades de datos tan grandes que desbordan los sistemas convencionales mono máquina.

Debes conocer

Para complementar la información que veremos en este curso sobre Big Data, existen multitud de fuentes online.

La primera (y muy útil) referencia sería el artículo de la Wikipedia sobre lo queson los macrodatos (traducción al castellano de Big Data).
Macrodatos

Es igualmente muy importante que sepas que la literatura sobre Big Data (al igual que ocurre con las ciencias de la computación en general) es más abundante en inglés que en castellano. Por ello, si quieres profundizar por completo por lo general tendrás que acceder a la versión en inglés de la información.

Un ejemplo es el enlace equivalente al previamente indicado, pero en la versión inglesa de la Wikipedia el cual contiene más información aún.

Big data (en inglés)

1.1.- Las 5 Vs.

En esta sección hablaremos de las cinco características de Big Data que suelen emplearse para discernir si el procesamiento de datos que necesitamos realizar puede realmente considerarse comoBig Data.

A lo largo de la literatura existente acerca de Big Data, esas 5 características han venido en llamarse "las 5 Vs".

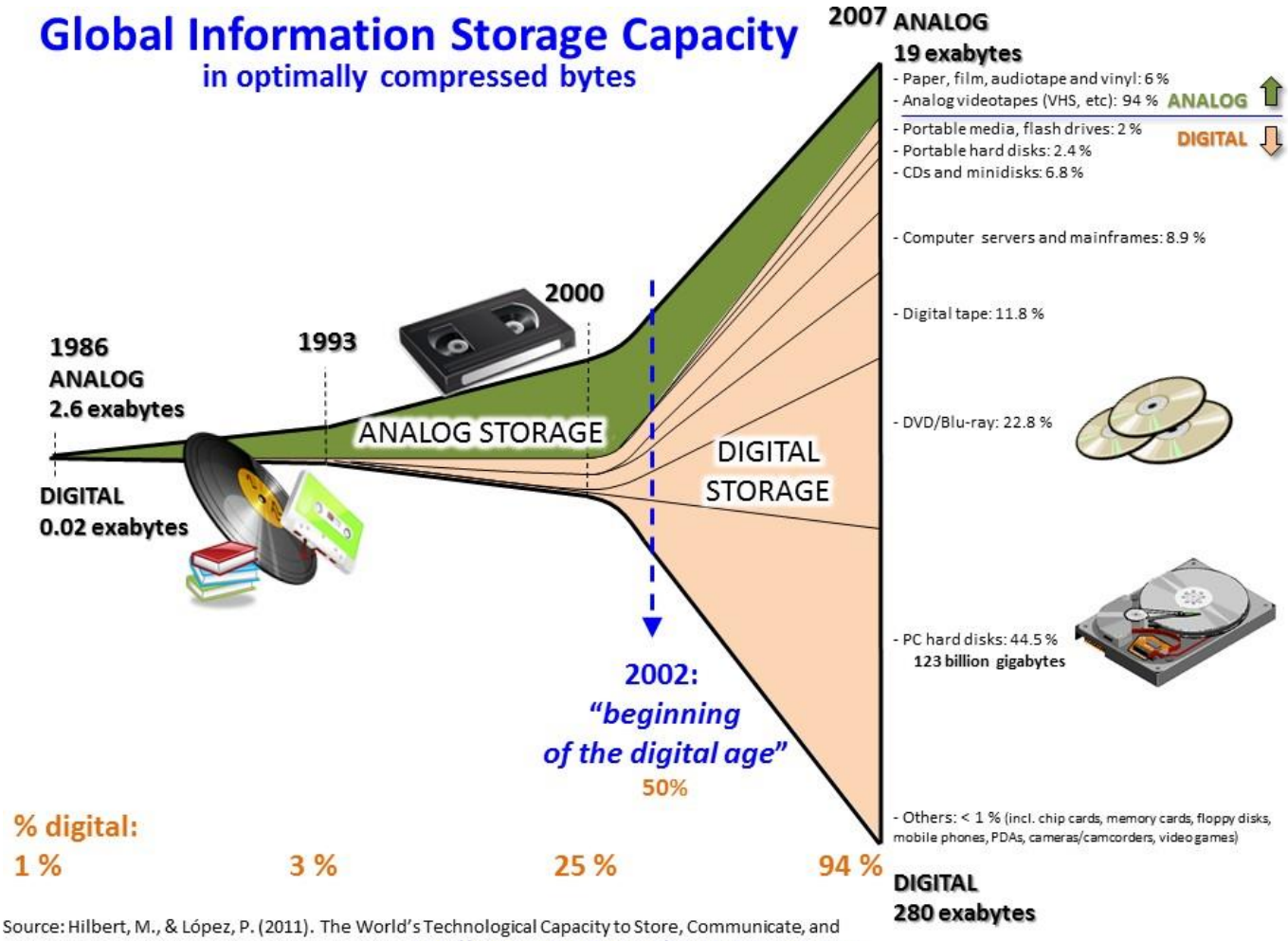
- Volumen.
- Velocidad.
- Variedad.
- Veracidad.
- Valor.

1.1.1.- Volumen.

La primera característica del reto de tratamiento de datos que ha venido en llamarse Big Data es el volumen de los mismos, es decir, la gran cantidad de bytes de información que los componen.

En la siguiente imagen podemos ver cómo ha ido creciendo en los últimos tiempos la cantidad de información almacenada por el ser humano, gracias a lo cual podemos hacernos una idea de la magnitud del reto.

Capacidad de información global



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Llegados a este punto, y para poder hacernos una idea de cuál es la magnitud de las cantidades de información con las que es necesario tratar, debemos hacer referencia al significado de las nomenclaturas que se emplean a tal efecto.

Unidades de cantidad de información digital

Unidades	Significado
Bit	Unidad mínima de información en un sistema de computación (almacena un "0" o un "1").
byte (B)	8 bits
kilobyte (kB)	1000 bytes (10^3 bytes)
megabyte (MB)	1000 kilobytes (10^6 bytes)
gigabyte (GB)	1000 megabytes (10^9 bytes)
terabyte (TB)	1000 gigabytes (10^{12} bytes)
petabyte (PB)	1000 terabytes (10^{15} bytes)
exabyte (EB)	1000 petabytes (10^{18} bytes)
zettabyte (ZB)	1000 exabytes (10^{21} bytes)
yottabyte (YB)	1000 zettabytes (10^{24} bytes)

Hay que tener en cuenta que si bien el significado de un kilobyte (kB) es 1000 bytes, dado que en ambientes de computacionales se emplea constantemente la numeración en base 2 también existe el kibibyte (KiB), el cual corresponde a 1024 bytes (2^{10}). De igual modo, también existe el mebibyte (MiB = 2^{20} bytes), el gibibyte (GiB = 2^{30} bytes), y así toda la progresión hasta llegar al yobibyte (YiB = 2^{80} bytes).

Lo que puede producir algo de confusión es que por lo general se emplean las nomenclaturas en base 10 de forma indistinta para designar tanto a la de base 10 como a la (más o menos equivalente) de base 2. Es decir, cuando vemos 1 MB, es posible quesignifique 10^6 bytes, pero también es posible que signifique 2^{20} bytes. Puede depender tanto del fabricante del dispositivo como a qué se esté refiriendo (módulos de memoria RAM, unidades de almacenamiento, hardware de red, ...).

Para hacernos una idea del volumen de datos que maneja la humanidad, según las predicciones el volumen de datos en el mundo se calculaba en unos 4.4 zettabytes en 2013, y tiene un crecimiento exponencial según el cual se espera que pueda llegar a los 163 zettabytes para el año 2025.

¿De dónde vienen todos esos datos?

- Datos de usuarios y/o clientes de instituciones y empresas. Datos generados por transacciones (compras, transferencias, ...).Datos adquiridos por sensores (de temperatura, de humedad, ...).Datos subidos a redes sociales (textos, imágenes, vídeos, ...).
- Datos relacionados con la salud (historiales y pruebas realizadas a pacientes).Datos de geolocalización (posicionamiento en cada momento según GPS).

- Datos guardados en logs (de todos los accesos que hacemos a páginas web). Datos producidos por el Internet de las cosas (de los diversos dispositivos IoT).Datos producidos por la genómica (cada vez que se secuencia un genoma).
- Datos de meteorología (información obtenida por satélites y las predicciones realizadasa partir de la misma).
- Datos producidos por cámaras (imágenes estáticas y vídeos producidos). Datos producidos por micrófonos (grabaciones de sonido producidas).
- Datos de RFID (aquellos con los que se tratan al realizar identificación porradiofrecuencia).
- Datos producidos por los sectores energético e industrial (toda la información que segenera alrededor de la energía y la industria).
- Datos Open Data (todos los datos abiertos liberados ya sea a nivel gubernamental ono gubernamental).

Debes conocer

¿A partir de qué cantidad de datos es Big Data?

No existe ninguna entidad u organismo que regule de algún modo cuál es el tamaño de datos concreto a partir de la cual se considera que estamos en un ambiente Big Data.

Simplemente nos quedaremos con que los sistemas para Big Data hoy en día trabajan con volúmenes del orden de los petabytes (PB) e incluso de los exabytes (EB).

Para saber más

En el siguiente enlace puedes ver más información los Datos Abiertos (u Open Data) sobre una de las fuentes de datos más interesantes, ya que son de dominio público.

Datos abiertos

1.1.2.- Velocidad.

No sólo tratamos con una gran cantidad de datos que hay que almacenar y procesar, sino que tales datos a su vez se siguen produciendo a una gran velocidad.

Para hacernos una idea, secalcula que en el mundo se generan cada 60 segundos:

- 350.000 tweets.

- 300 horas de vídeosubidos a YouTube (más los que se suban a otras plataformas).
- 171 millones de correos electrónicos. 330 GB's de información generados por sensores de motores de aviones comerciales.

Si volvemos a revisar la enumeración de posibles fuentes de las que provienen los datos que vimos en el apartado anterior, y tenemos en cuenta que en el mundo hay del orden de 7.870 millones de personas (datos de 2022), podremos seguir haciéndonos una idea de la gran velocidad a la que todos esos datos se siguen generando cada minuto que pasa.

El problema con respecto a la velocidad no es únicamente el hecho de que el volumen de datos continúe creciendo sin parar (ya que si hemos dimensionado el almacenamiento para el doble de lo que necesitamos entonces aún quedará mucho tiempo para que tal tamaño de almacenamiento sea un problema), sino lo rápido que es necesario obtenerlos y ser capaces de integrarlos junto con los que ya tenemos.

De la gran velocidad a la que llegan datos nuevos nacen las estrategias de procesamiento tipo *streaming*, las cuales estudiaremos más adelante.

1.1.3.- Variedad.

Además de tener que procesar una gran cantidad de datos que se generan cada vez más rápido, existe el problema añadido de la gran variedad existente en cuanto a la representación de tal información.

Datos estructurados:

Los existentes en registros (filas) de bases de datos (típicamente relacionales), los cuales existen dentro de tablas con un esquema definido que nos indica de qué tipo de datos es cada una de las columnas (entero, decimal, textual, fecha, ...).

Datos no estructurados:

Aquellos que no están regidos por un esquema. Por ejemplo:

- Vídeos.
- Imágenes.
- Audios.

Hay que tener que la proporción de datos en el mundo que son no estructurados se estima en más de un 80% del total, lo cual es fácilmente comprensible teniendo en cuenta la naturaleza de los mismos. Sólo hay que comparar el espacio de almacenamiento que ocupa un vídeo (típicamente varios megabytes) con el que ocupa un registro en una base de datos (típicamente varios bytes).

Datos semiestructurados:

Son datos definidos según una cierta estructura pero que no tienen naturaleza relacional (es decir, no son registros de una tabla con un esquema determinado).

Por lo general se almacena en ficheros de texto siguiendo un cierto formato preestablecido, de modo que se mantiene la flexibilidad que ofrece el fichero (para poder almacenar lo que sea necesario) a la vez que es posible determinar qué significa cada una de las porciones de información que se encuentran dentro del mismo.

Ejemplos de formato de fichero en los que se guardan datos semiestructurados:

CSV.
XML.
JSON.

Metadatos:

Los metadatos son datos extra (muchas veces generados de forma automática) que se guardan acerca de los propios datos para favorecer su interpretabilidad posterior.

Ejemplos de metadatos que pueden acompañar a los datos convencionales son:

- Información extra sobre su estructura.
- Fuente.
- Autor.
- Fecha de creación.
- Resolución en pixeles (si se trata de una imagen o un vídeo).Duración (si se trata de un vídeo).
- Frecuencia de muestreo (si se trata de un audio).
- Tipo de compresión.

1.1.4.- Veracidad.

Un problema extra con el que tenemos que tratar es el hecho de que los datos no siempre cuentan con la calidad deseada o no son totalmente fieles a la realidad.

Este término está muy relacionado con el concepto de relación seña/ruido en cualquier flujo de información.

- El ruido son datos que no pueden ser convertidos en información (ya sea porque no la contienen o porque ésta está corrupta y es irrecuperable).
- La señal está constituida por datos que sí pueden ser convertidos en información con sentido.

Para saber más

En el siguiente enlace de la Wikipedia puedes saber más sobre lo que significa la relación señal/ruido:

Relación señal/ruido

Por ello, por un lado, es necesario conocer en qué condiciones se adquirieron los datos (para poder así estimar su nivel de veracidad), mientras que por otro lado en muchos casos será necesario llevar a cabo un procesamiento específico de los mismos con el fin de resolver

posibles problemas y eliminar información inválida.

Por lo general los datos producidos de modo automático (como la generada cuando realizamos transacciones) contienen menos ruido que los que producen personas (como los posts de un blog).

1.1.5.- Valor.

El concepto de valor en relación con los datos tiene que ver con cómo de útiles son estos para una institución, empresa o persona.
Tiene mucho que ver con el concepto de veracidad que ya hemos visto, ya que por lo general cuanto más veraces (fieles a la realidad) sean los datos,más valor se puede obtener de ellos.

También depende en gran medida del tiempo transcurrido desde que se produjeron tales datos. Por ejemplo, si estamos operando en bolsa, el dato que nos indica el valor de una acción es mucho más valioso si corresponde a hace 1 segundo que si corresponde a hace 1 hora. En términos generales, cuanto más rápido seamos capaces de hacer llegar el dato desde donde se produce al lugar en el que se toman las decisiones, más valor podremos obtener de ellos.

Es también muy importante que los datos sean lo más completos posible para poder producir el valor deseado. Es decir, no sólo que sean veraces (que lo que viene sea correcto) sino que sean completos (que venga todo lo que necesitamos).

Por último, la propia interpretación del dato también juega un papel vital a la hora de poder obtener valor. Por ejemplo, sería absurdo estar almacenando un valor de temperatura obtenido por un sensor que está bajo tierra y querer utilizarlo para una científica sobre temperatura ambiente. En este caso el dato podría ser perfectamente veraz pero la interpretación del mismo estaría siendo errónea, lo cual disminuiría el valor producido.

1.2.- Qué conseguimos gracias a Big Data.

En esta sección veremos qué nos aporta no sólo el ser capaces de obtener y almacenar con grandes cantidades de datos sino también el poder tratarlos y analizarlos gracias a las metodologías y tecnologías de Big Data.

Aportes generales de Big Data

Las metodologías y tecnologías para Big Data nos permiten realizar diversas operaciones con grandes cantidades de datos, entre las cuales se encuentran:

- Capturarlos desde sus orígenes.
- Integrarlos para poderlos almacenar de un modo unificado.
- Almacenarlos de un modo distribuido y replicado, gracias lo cual conseguimos altos valores de disponibilidad.
- Tratarlos de forma distribuida, empleando para ello un alto número de máquinas que los procesan en paralelo.
- Aplicar técnicas de minería de datos (también llamado ciencia de datos cuando esa minería de datos se realiza en ambientes Big Data) para crear modelos predictivos.
- Usar esos modelos para realizar predicciones a utilizar en sistemas automáticos.
- Crear visualizaciones y cuadros de mando usando tanto los propios datos como los modelos creados para así dar soporte a la toma de decisiones.

El ser capaces de realizar tales operaciones con los datos, nos permiten obtener los siguientes aportes y beneficios (entre otros):

- Generar registros más detallados mediante la integración desde diversas fuentes.
- Optimizar las operaciones de instituciones y empresas.
- Poder actuar de modo inteligente basándonos en la evidencia de los datos.
- Identificar nuevos mercados.
- Realizar predicciones basándonos en modelos creados a partir de los datos.
- Detectar casos de fraude e impagos.
- Dar soporte a la toma de decisiones.
- Realizar descubrimientos científicos.
- Ayudar a los médicos a detectar enfermedades en función del historial de los pacientes y las pruebas que se les realizan.
- Crear nuevos fármacos más efectivos y con menos efectos secundarios.

1.2.1.- Desde los eventos al valor.

El tratamiento de los datos a lo largo de diversas capas de procesamiento sucesivas nos permite llegar desde los meros eventos que se producen en nuestro mundo hasta la sabiduría que necesitamos para obtener valor gracias a poder tomar las mejores decisiones.

Eventos:

En nuestro mundo se producen eventos constantemente.

- Una estación meteorológica unas mediciones de temperatura, humedad, presión atmosférica, etc.
- Una cámara toma imágenes dentro de una fábrica.
- Alguien pide un préstamo.
- Alguien realiza una llamada telefónica.
- Alguien realiza un pago con tarjeta.
- Un hospital realiza una prueba médica a un paciente.
- ...

Datos:

Los eventos son reflejados de algún modo, generándose de ese modo datos que pueden ser almacenados para su uso posterior.

- Registros de bases de datos.
- Ficheros (en diversos posibles formatos).

Información:

Cuando les damos contexto a los datos, organizándolos de algún modo lógico, tenemos información.

- Distintos registros referentes a pagos con tarjeta quedan almacenados en una misma tabla.
- Usamos jerarquías de carpetas para organizar distintos ficheros en función

Conocimiento:

Si tratamos la información dándole un significado, podemos obtener conocimiento.

- A partir de gran cantidad de datos se generan modelos mediante los cuales se representa la realidad y que pueden ser utilizados para realizar predicciones.

Sabiduría:

Si una vez tenemos conocimiento en forma de modelos predictivos añadimos el entendimiento necesario para saber de qué modo emplearlos. Como resultado obtenemos sabiduría.

Valor:

La sabiduría de por sí misma no genera ninguna acción. Sin embargo, si realizamos acciones basándonos en la sabiduría, esas acciones serán mejores que las que podamos tomar sin basarnos en los datos.

La diferencia entre el resultado que podemos obtener basándonos en la sabiduría que producen los datos y el que obtendríamos si no los hubiésemos tenido en cuenta para nada, es el valor añadido que conseguimos.

Debes conocer

A pesar de que dentro de las tecnologías de Big Data se suele englobar lo relacionado con obtener valor del dato, en la práctica son la minería de datos o la ciencia de datos las disciplinas que terminan de obtener el valor (haciendo uso de esas tecnologías).

La Minería de Datos es una rama de la Inteligencia Artificial que emplea técnicas de Aprendizaje Automático para obtener valor de los datos.

La Ciencia de Datos en el fondo es misma Minería de Datos, pero haciendo énfasis en que se realiza en entornos de Big Data.

Sin embargo, puedes encontrar los términos Minería de Datos y Ciencia de Datos siendo empleados de forma equivalente (Minería de Datos en Big Data y Ciencia de Datos fuera de Big Data).

2.- Clústeres de computadoras.

En ambientes de computación, un clúster es un conjunto de computadoras (también referenciados como servidores o como nodos) conectados entre sí mediante red para trabajar como una única unidad resolviendo cargas de trabajo de forma conjunta.

Históricamente los clústeres se construían utilizando computadoras especializadas muy caras. Sin embargo, más adelante han ido apareciendo diversos *frameworks* o plataformas de computación distribuida que emplean computadoras de uso común (el llamado *commodity* hardware), gracias al considerable aumento sus prestaciones.

Para saber más

Puedes ver más información sobre los clústeres de computadoras en el siguiente enlace:
Clúster de computadoras

El uso de clústeres nos da una serie de ventajas respecto al uso de computadoras de forma individual:

- Alto rendimiento.
- Alta disponibilidad.
- Equilibrado de carga.
- Escalabilidad.

Alto rendimiento:

Dado que cada componente del clúster es una computadora completa, con sus propios recursos (procesador, memoria y almacenamiento), las cargas de trabajo susceptibles de paralelización pueden acelerarse en gran medida dividiéndolas en subtareas y distribuyéndolas para que sean ejecutadas en los distintos nodos.

Gracias a esto se pueden resolver problemas muy complejos que no sería posibleresolver en un tiempo razonable en una máquina individual por muy potente que ésta sea.

Alta disponibilidad:

Mediante una continua monitorización entre los propios nodos del clúster, se puede detectar la no disponibilidad de un subconjunto de los mismos (ya sea por fallo eléctrico, por avería o por corte de las comunicaciones) y se pueden tomar medidas para que los servicios o datos que hay (o había) en esas máquinas siganestando disponibles.

- Rearrancando un nodo caído o arrancando un nuevo nodo para suplirlo.
- Respondiendo las peticiones desde otro nodo del clúster que también contenga una réplica de esos datos.

Equilibrado de carga:

El equilibrado de carga (o también balance o balanceo) se consigue mediante algoritmos destinados a distribuir las cargas de trabajo entre los diversos nodos del clúster para así evitar cuellos de botella. Tales cuellos de botella se producen cuando el envío de trabajos a nodos sobrecargados aumenta la latencia media con la que tales trabajos son finalizados.

Para ello, se realiza una monitorización del estado de carga de cada nodo y se decide para cada paquete de trabajo a qué nodo enviarlo, atendiendo a:

- El tamaño del trabajo.
- El estado de carga de cada nodo.
- La potencia de procesamiento de cada nodo.

Escalabilidad:

Gracias a que el clúster está formado por un número indeterminado de nodos, no sólo conseguimos una mayor potencia de cálculo al utilizarlos para una misma tarea, sino que podemos hacer crecer dicha potencia de cálculo añadiendo nuevos nodos. En otras palabras, la potencia de cálculo del clúster es ampliable.

Esta característica es muy deseable para sistemas Big Data, ya que desaparece la necesidad de realizar una estimación de potencia necesaria a priori, lo cual en por lo general siempre lleva a una sobreestimación para guardar un margen de seguridad. Con un clúster escalable podemos comenzar con un número determinado de nodos e ir añadiendo más según sea necesario.
Es interesante conocer la diferencia entre escalado horizontal y vertical:

- **Escalado vertical (scale-in):**
  - Es el que se consigue mejorando las características hardware de la computadora (individual) en el que se están ejecutando las cargas detrabajo (procesador, memoria o almacenamiento). Por lo tanto, está limitado por la mejor especificación de hardware que sea posible encontrar en el mercado.
  - Por ello, aunque reciba el nombre de "escalado" en la práctica no sirvepara conseguir la característica de escalabilidad.
- **Escalado horizontal (scale-out):**
  - Es el que se consigue añadiendo más nodos a un clúster.
  - Por ello es el tipo de escalado que realmente nos permite conseguir la característica de escalabilidad.

Para saber más

En los siguientes enlaces puedes ver más información sobre lo que significaalto rendimiento, alta disponibilidad, equilibrado de carga y escalabilidad:

Clúster de alto rendimiento Clúster de alta disponibilidad Equilibrio de carga Escalabilidad