# ¿Historia, data centers tradicionales y evolución?

No se puede concebir el crecimiento en los negocios sin el advenimiento de las tecnologías en la nube. Tradicionalmente tener un servidor en un propio centro de datos no solo requería de infraestructura (de electricidad, control de temperatura y humedad, energía de respaldo, y un largo etc) también se requería personal altamente capacitado en ámbitos como redes y servidores haciendo prácticamente inalcanzable para organizaciones medianas o pequeñas contar con soluciones de software que impulsaran sus negocios.

Aunque suene extraño, el concepto de compartir tiempo y recursos de un sistema de cómputo no es algo que se haya dado en las últimas décadas, ni siquiera en este siglo, [John McCarthy](https://es.wikipedia.org/wiki/John_McCarthy) a finales de la década de los 50 ya se había planteado e implementado un sistema donde múltiples personas podían compartir CPU, memoria y tiempo de ejecución en una IBM 704. El concepto se hizo muy popular en décadas siguientes, pero ya para la década de los 80 comienza a perder popularidad debido al auge de los microprocesadores cada vez más rápidos y potentes.

El lanzamiento de la WWW con el protocolo de transferencia de hipertexto el 30 de Diciembre del 1990 por [Tim Berners-Lee](https://en.wikipedia.org/wiki/Tim_Berners-Lee) marca un hito en la historia que haría que la demanda de sistemas de cómputo creciera en las décadas venideras, ya que la web se idealiza como algo abierto, libre para todos, editable y libre de cargos o patentes haría que cualquier persona pudiera hacer uso de esta nueva tecnología para compartir sus ideas al mundo entero, ¿cuál fue el problema?, los servidores. Para hacer funcionar una página web en ese momento se requería un ‘web server’ (hoy día también además de otras tecnologías), siendo httpd el primero escrito por el propio Tim. No era muy pesado para ser ejecutado, hay que recordar el contexto temporal en que los recursos de cómputo eran prácticamente limitados a empresas con gran capital o instituciones educativas y de investigación, así que solo unos privilegiados podían generar sus propias páginas y compartirlas al mundo. No fue hasta 1994 cuando se lanzan los primeros servidores montables en un rack por parte de Compaq, los [ProLian](https://iweb.com/wp-content/uploads/2018/09/5.jpg), ese fue el primero paso hacia centros llenos de computadoras potentes para el momento.

Uno de los precursores de renta de computadoras con conexión a internet para tener páginas web propias fue Richard Yoo, él en 1996 ya servía como un ISP pequeño y para 1997 ofrecía servicio de web hosting, después de algunos negocios en 1998 funda ‘RackSpace’ con él mismo como CEO. Además de ofrecer servicios como web hosting se enfocaba a servir soporte de nivel “fanático”, es decir un servicio con la primicia en el soporte y servicio al cliente. Recordemos que a este punto la burbuja de [las punto com](https://blog.r4.com/burbuja-de-las-puntocom/) estaba en plena gestión, todo mundo quería una página web no importando en realidad si era negocio en el corto plazo, el punto era hacer una página para después monetizarla.

Como ellos muchos otros nacieron en el mercado con la promesa de brindar un espacio para alojar páginas web pero después con la explosión del lenguaje PHP y más tarde de CMS como Wordpress los hosting ya ofrecían también soporte para este lenguaje e incluso soporte para base de datos, todo en uno, una real ganga. Esto le dio otro impulso a la demanda centrada en productos de software de tipo web, pero la alta demanda de algunos sitios en especial de noticias comienza a demandar arquitecturas y recursos más allá de los que podía brindar un hosting de un solo servidor, por lo regular eran servidores que no podías modificar mucho, la base de datos ya venía definida, el lenguaje para trabajar también, algunos ofrecen servicio de email integrado que también ya viene definido, otro problema que se suscitaba también es el ¿en quién confiar?, un día el web hosting que daba servicio a la organización para podría desaparecer sin dejar rastro, en cuanto a los costos, se cobra normalmente una tarifa plana con ciertos recursos que no se deben sobrepasar al mes sobretodo en ancho de banda, si son sobrepasados es cobrado un cargo extra, y si están subutilizados no es posible aminorar los costos extra con el excedente en la mayoría de los proveedores, ante picos de operación son poco flexibles.

La evolución de la web a la llamada [Web 2.0](https://whatis.techtarget.com/definition/Web-20-or-Web-2) aceleró indudablemente la necesidad por más servicios, más rápidos y más flexibles, la Web 2.0 se caracteriza por su dinamismo, la interactividad que brindan y la capacidad de ver contenido más rico visualmente. Es en esta etapa ya se puede hablar de cómputo móvil, software como servicio, web apps, generación de contenido en audio y vídeo, comunicaciones unificadas y el ya conocido social media.

El 14 de marzo de 2006 el tablero de juego iba a cambiar, Amazon lanza su servicio de cómputo en la nube llamado Amazon Web Services, lanzando aparte de S3 (servicio de storage) y SQS (Servicio de colas) el servicio ‘Elastic Compute Cloud’ mejor conocido como EC2, la peculiaridad y novedoso del servicio es el cobro de tarifas por la cantidad de recursos que uses, tanto en CPU y memoria RAM como en storage, se basa también en la modalidad de ‘auto servicio’ con lo cual uno mismo puede provisionar tantos servidores como sea necesario y configurarlos con el lenguaje de programación de la preferencia del cliente. Estos puntos de flexibilidad dieron a empresas pequeñas y medianas la capacidad de ser más competitivas sin incurrir en los costos excesivos y el personal que requiere un servidor en una oficina. Aunque, es cierto, para operar AWS también se requiere personal capacitado, con experiencia, que sepa y reconozca que servicios deben ser usados ante la necesidad de los diferentes negocios que requieren una solución tecnológica a su operación. Aquí es donde entras tú, capacitarte para las tecnologías de nube es una inversión que seguro recuperarás rápidamente.

# ¿Opciones comunes de cómputo?

Ya se ha hablado brevemente de las ventajas de tener los servidores en algún proveedor de servicios en la nube, pero, ¿es la mejor opción?.

Hoy día en el mercado se pueden encontrar diversas ofertas de cómputo, por un lado es común encontrar los servidores instalados en sitio. Este tipo de servicio de cómputo se requiere a la hora de cumplir con normativas, ej. en industrias como las de seguros no es posible alojar los datos de los clientes fuera de México, por lo que un servicio de cómputo en la nube no es posible para la mayoría de servicios. Tener servicios en sitio implica un gran esfuerzo para mantenerlos operando constantemente, los servidores prácticamente son computadoras con muchos gigas de memoria RAM, múltiples procesadores con múltiples núcleos cada uno, múltiples discos duros formado arreglos [RAID](https://searchstorage.techtarget.com/definition/RAID) que dependiendo de la configuración brindan respaldo a los datos, brindan velocidad o ambos, todo esto hace que los servidores desprendan mucho calor el cual al tener varios servidores apilados en [racks](https://www.capitolinetraining.com/getting-your-data-centre-ready-for-open-compute-and-open19-rack-layouts/) hacen que el lugar físico donde se encuentran tranquilamente pueda llegar a temperaturas de 70 °C, la temperatura es un enemigo a vencer siempre que se usen servidores en sitio, para lo cual se debe contar con una solución confiable de control de temperatura con equipos especiales de enfriamiento que deben estar encendidos por meses continuamente y que además garanticen parámetros de humedad en el aire muy específicos, demasiada humedad causará condensación en los servidores, poca humedad generará un aire en el centro de datos propenso a tener carga estática, imaginar la memoria RAM de un procesador dañada por la descarga de cargas estáticas acumuladas. Para cerrar el tema del enfriamiento también se debe contar con una instalación física adecuada que garantice el correcto flujo tanto de aire caliente despedido por los servidores como de aire frío que debe entrar a ellos, [aquí](https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-680202.html) se puede profundizar más al respecto.

El siguiente enemigo a vencer es la electricidad, los servidores normalmente deben estar encendidos 24/7 por años inclusive, por lo que se debe garantizar un flujo continuo de electricidad, para lo cual se usan tradicionalmente equipos [UPS](https://www.energystar.gov/products/data_center_equipment/uninterruptible_power_supplies) que mantienen los servidores encendidos por algunos minutos tiempo suficiente para el arranque y regularización de voltaje de una [planta de energía](https://www.generatorsource.com/Supplying_Backup_Power_to_Data_Centers.aspx), tanto como el o los UPS y la o las plantas de energía requieren de mantenimiento continuo a fin de evitar anomalías ante un corte impredecible de electricidad de la red principal.

La comunicación entre [servidores](https://www.sdxcentral.com/data-center/definitions/data-center-networking-explained/) es crucial para tener servicios operando, sin problema un call center debe contar con las grabaciones del servidor PBX, las grabaciones se llevan a un servidor de storage para recolectar las grabaciones de llamadas. Para mantener una sana comunicación entre ambos se requieren equipos de alta calidad, un buen cablead instalado por profesionales y una impecable administración y configuración de las redes, eso sin contar con la seguridad que conlleva la correcta segmentación y dimensionamiento de las mismas, ah, la redundancia es también un factor importante. Algunos de los equipos usados en la industria se presentan [aquí](https://www.router-switch.com/Price-cisco-switches-cisco-switch-catalyst-6500_c18).

Hasta este punto se presentan los principales problemas para arrancar un centro de datos, eventualmente los equipos fallan, se deterioran, se debe estar dando constante mantenimiento tanto al espacio físico, al propio hardware de los servidores también hay que hacerlos, los discos duros fallan y a pesar de en arreglo RAID eso no exime de reemplazarlos si se presenta una falla en alguno de ellos, en cuanto a los sistemas operativos se deben aplicar parches de seguridad y coordinar ventanas de mantenimiento necesarias para mantener la usabilidad y seguridad. Aquí entran en juego temas de [soporte y garantías](https://marketing.dell.com/Global/FileLib/hp_microsite/dell-support_services.pdf), ¿si se daña un disco quien que va a dar el soporte y garantía para reemplazarlo?, ¿si un ventilador del servidor deja de funcionar?, ¿si alguna de las fuentes redundantes de un servidor deja de funcionar?, ¿si un puerto en una tarjeta de red deja de operar?, todos esos escenarios son plausibles que sucedan.

## Sub utilización y virtualización.

La complejidad en los centros de datos aunque solo cuenten con un rack en él (a un rack le pueden caber 24 servidores físicos, 48 Unidades de rack, cada servidor puede ocupar 2 Unidades dependiendo del modelo) crece rápidamente, cada servicio como telefonía o aplicación web requiere tener servidores para manejar storage, servidores web, sftp, telefonía, bases de datos, donde dependiendo de la criticidad que se le asigne a cada servicio será determinante pasa saber en que servidor físico se configurarán los servicios. Imagina por un momento a un proveedor de hosting, aloja cientos de páginas y aplicaciones de sus clientes, realmente se vuelve inviable económicamente tener un servidor para cada cliente a un precio accesible, lo mismo pasa en empresas y centros de investigación, no se vuelve viable tener servidores físicos para cada servicio. En los albores de la computación múltiples personas deseaban utilizarlos pero no era posible, al inicio se optaron por estrategias de tiempo compartido donde el servidor podía procesar los datos de una persona mientras otra tomaba su tiempo para decidir cuál sería el siguiente set de instrucciones que el servidor debía cumplir por medio de terminales tontas, lo cual era ineficiente, no era raro que un usuario sobrecargara al servidor afectando en rendimiento a los demás usuarios conectados impactando sus tareas. Fue hasta 1988 con el lanzamiento de [SoftPC](https://www.nytimes.com/1988/06/19/business/the-executive-computer-choosing-a-link-from-mac-to-dos.html) que la idea de virtualización cobró vida.

La virtualización es en concepto, tomar un servidor físico y poder montar sobre él múltiples sistemas operativos que incluso pueden ser diferentes entre sí, brindando aislamiento y seguridad, así los procesos de un sistema operativo se mantienen aislados de otros sistemas operativos, ¿se tienen cuatro aplicaciones web con sus respectivas bases de datos?, no hay problema, es posible ejecutar cada tupla web server y base de datos en un sistema operativo en el mismo servidor físico. Esta es una estrategia hoy día ampliamente usada para reducir costos y utilizar mejor los recursos de un servidor, adquieres un servidor con 4 TB de disco duro, 2 procesadores con 8 núcleos cada uno y 32 GB de RAM, una forma de disponer de esos recursos es asignar a 4 sistemas operativos en su propia maquina virtual 2 núcleos de los 8 disponibles, 8GB de RAM y 1 TB a cada una, en el caso de la red con una conexión de 1 Gbps es suficiente para la mayoría de aplicaciones aunque si se requiere más se pueden poner tarjetas extra de 10 Gbps ya sea de fibra o cobre con escudo a menudo conocido como [Cat 7](https://www.amazon.com/Shielded-Ethernet-Cable-Pack-10GB/dp/B06XWN3S5B).

Si se requiere tener más servicios funcionando simplemente se provisionan más servidores físicos con un factor de 4 a 1 para esta configuración, si una organización tiene 42 servicios en lugar de usar 42 servidores físicos solo serían 11 y aún sobra capacidad para un par más.

## El Cloud

Pero, ¿cómo puedo tener más de un sistema operativo ejecutándose al mismo tiempo?, si has tenido una PC o laptop con Windows y alguna distribución GNU/Linux instalada se debe apagar un sistema operativo para encender el otro. Para instalar un sistema operativo dentro del sistema operativo principal es requerido un Hypervisor como [VirtualBox](https://www.virtualbox.org/wiki/Downloads) o VMWare Workstation (hosted hypervisor o de tipo 2), para soluciones de servidores (native hypervisor o de tipo 1) se tiene [Linux KVM](https://www.linux-kvm.org/page/Main_Page), [VMWare vSphere](https://www.vmware.com/products/vsphere.html) o [Hyper-V](https://docs.microsoft.com/en-us/windows-server/virtualization/hyper-v/hyper-v-technology-overview). Cualquiera sea la solución todas permiten la generación u orquestación de recursos de las máquinas virtuales, desde ellos es posible agregar más memoria, más disco, más núcleos, apagar y encenderlas prácticamente sin afectar a otras máquinas virtuales aun hospedadas en el mismo servidor físico.

De tener los suficientes recursos tener 100 servidores físicos en unos 5 o 6 racks con la tasa de 4 a 1 se contaría con 400 servidores listos para montar ahí cualquier solución de software, bases de datos, servidores web por ejemplo. ¿Y si el excedente de recursos/servicios se vende? en la empresa de esos 400 solo serán usados 300, los otros 100 se pueden rentar a otras personas u organizaciones que tengan necesidad de procesamiento y capacidad de cómputo. Amazon y eventualmente otros grandes tecnológicos vieron esta posibilidad de rentar espacio e infraestructura a terceros, espacio e infraestructura que ocupan en sus negocios principales. Mismo Amazon ha dicho como usa la infraestructura de Amazon Web Services para soportar sus operaciones de [black friday](https://aws.amazon.com/es/blogs/aws/how-aws-powered-amazons-biggest-day-ever/).

## ¿Otras alternativas? ¿Cuál es mejor para mi?

* Colocation: Hay lugares o centros de datos donde es posible rentar un espacio físico (espacio en un rack medido normalmente en unidades de Rack) donde instalar y configurar de acuerdo a las necesidades un servidor físico propio. Dependerá de proveedor el costo y terminaos y condiciones del servicio. Es una opción interesante si se requiere cumplir con regulaciones nacionales, por ejemplo, es común escuchar la restricción de los datos no pueden salir del país o encontrarse operando en una zona de alto riesgo de desastres naturales donde el centro de datos se puede ver amenazado, otra opción es contar con esos servidores para un [Disaster Recovery Plan](https://searchdisasterrecovery.techtarget.com/definition/disaster-recovery-plan) y un [BCP](https://www.ready.gov/business-continuity-plan). Un posible [proveedor](https://www.digitalserver.com.mx/colocacion-de-servidores-en-mexico.shtml) de este servicio.
* Baremetal o Servidores dedicados: Son servidores que no están compartidos por otros clientes, es un servidor dedicado para un solo cliente, puede estar en una nube comercial, en un pequeño centro de datos o dentro de un centro de datos propietario, la característica principal es que no es compartido. Hay aplicaciones altamente demandantes que requieren bastos recursos como algoritmos de Inteligencia artificial y Data Lakes por ejemplo. [AWS](https://aws.amazon.com/es/blogs/aws/now-available-bare-metal-arm-based-ec2-instances/) provee de este tipo de servidores. [IBM](https://www.ibm.com/mx-es/cloud/bare-metal-servers) también lo hace.
* Nube privada, pública e híbrida: En general todas las soluciones de nube como AWS, GCP, IBM y Azure son consideradas nubes públicas, ya que cualquier persona del planeta las puede utilizar y puede compartir el mismo servidor físico, las mismas redes y los mismos equipos físicos de storage ya que se rentan en un modelo compartido. Por otro lado la nube privada se caracteriza por que solo el cliente tiene acceso y para hacer uso de los servicios computacionales que en ella se hospeda, es normal que los servicios hospedados en esa nube privada solo sean accesibles a una empresa muy específica, se usan normalmente por gobierno, bancos, fintech, aseguradoras, etc. Los esquemas híbridos hacen una fusión de la nube privada y nube pública, en los casos en los que los datos sensibles de la empresa no puedan salir del país por regulaciones es posible dejar esa infraestructura en una nube privada de la empresa y otros servicios como páginas web, aplicativos web y bases de datos en la nube pública. De hecho AWS permite por ejemplo, conectar una solución local de backups como Backup Exec para depositar los backups en la nube por medio de [AWS Storage Gateway](https://aws.amazon.com/es/storagegateway/) operando así en ambos esquemas. Bajo el esquema híbrido las cosas se pueden complicar aún mas, a veces las soluciones demandadas al departamento de IT se vuelven complejas y un solo proveedor de nube no tiene las herramientas con la calidad y necesidades que la organización requiere, por ejemplo, si se requiere hacer text to speech desde un proveedor diferente a Amazon por que al director del proyecto no le gustaron las voces de Amazon, se puede hacer el procesamiento text to speech en una nube distinta terminado el procesado los datos pueden ser regresados a AWS para seguir el flujo de trabajo destinado para ellos, el esquema es conocido como [Hybrid MultiCloud](https://www.netapp.com/us/info/what-is-hybrid-multicloud-experience.aspx).

## Impulsando negocios

Muchas veces los servidores dedicados pueden llegar a quedar grandes para muchas empresas, recordemos que en México la gran mayoría de empresas son medianas y pequeñas. Los baremetal también pueden ser algo muy grande, con el advenimiento de las distintas ofertas de nube se pueden contratar servicios de cómputo por precios muy baratos, en el caso de Amazon Web Services un servicio interesante para comenzar con AWS es [Lightsail](https://aws.amazon.com/es/lightsail/), sin muchos conocimientos se puede tener un servidor de Wordpress en pocos clicks funcionando. Hay otros servicios especializados en temas de: - IoT (internet of things): Comunicar objetos por medios de protocolos ligeros como WebSockets o MQTT, [AWS IoT Core](https://aws.amazon.com/es/iot-core/) es la opción.

* Inteligencia artifical con servicios de [texto a voz](https://aws.amazon.com/es/polly/?c=ml&sec=srv), [análisis de texto](https://aws.amazon.com/es/comprehend/?c=ml&sec=srv), [análisis de imágenes](https://aws.amazon.com/es/rekognition/?c=ml&sec=srv).
* Block Chain: [Servicios como Amazon QUantum Ledger Database](https://aws.amazon.com/es/qldb/?c=bl&sec=srv) y [Amazon Managed Blockchain](https://aws.amazon.com/es/managed-blockchain/?c=bl&sec=srv)
* Analytics: ¿Trabajos de ETL? no hay problema [AWS Glue](https://aws.amazon.com/es/glue/) hace el trabajo, [AWS Lake Formation](https://aws.amazon.com/es/lake-formation/) es la solución si se requiere guardar información en múltiples formatos antes de ser procesada y analizada.

# Modelos de Servicios

Los servicios en general de los proveedores de Cloud en la nube se dividen en tres grandes categorías.

* IaaS (Infrastructure as a Service) : Caracterizado por tener acceso a la configuración del servidor a nivel de sistema operativo con todo y las ventajas y desventajas que pueda acarrear, también se caracteriza por poder manejar la configuración de Storage y las redes para los servidores. En este nivel el proveedor de nube absorbe los gastos operativos de electricidad, hardware, garantías, refrigeración, acceso físico y gestión del hypervisor, aquí la responsabilidad del cliente comienza en el sistema operativo, pasando por los componentes del aplicativo como bases de datos o web servers hasta el código fuente de la aplicación.
* PaaS (Platform as a Service): Muchas veces al instalar servicios y soluciones hay que preocuparse por el sistema operativo, parches, storage, redes, balanceo de carga, los clusters de base de datos,etc. En el esquema de plataforma como servicio el administrador olvida prácticamente el provisionamiento, solo hay que configurar el servicio, ya en este nivel el administrador se debe hacer cargo del código y los datos que serán procesados.
* SaaS (Software as a Service): Literal paga y disfruta. No hay que provisionar código, no hay que preocuparse por redes ni sistemas operativos, storage, disponibilidad, o planes de recuperación de desastres.

En general hay aplicaciones que se siguen haciendo en el modo tradicional, se requiere una base de datos, se provisiona una instancia de máquina virtual en la nube y se instala un gestor de base de datos, las nuevas tendencias de aplicaciones en pro de ser mas flexibles en su evolución y en la demanda por parte de los usuarios se tiende a hablar de Cloud-Native Services. ¿se requiere reconocimiento de texto cuando un cliente sube una factura?, no necesariamente debe ser implementado desde cero (eso demandaría semanas de trabajo y personal altamente capacitado), es posible usar algún servicio ya existente. Hay un trio de formas de proveer servicios que han tenido auge en los últimos años:

* Serverless: Es un paradigma en el cual ya no se tienen servidores que administrar, prácticamente un servicio serverles es una instancia muy pequeña (poca RAM, poco CPU y poco Storage) de un servidor que es creada al vuelo, este servidor es preconfigurado con código que ejecutará así como las entradas de datos que dispararán el arranque de esta unidad de procesamiento y qué datos además de dónde los debe ingerir y a donde almacenará la salida. Estas unidades de procesamiento son efímeras, es decir una vez completada su tarea que debe durar algunos minutos (máximo 15 en el caso de [AWS Lambda](https://aws.amazon.com/es/lambda/)) la unidad de procesamiento desaparece, los datos en memoria RAM ya no pueden ser accesados, en general la Lambda es destruida. Las Lambdas se vuelven especialmente útiles para tareas repetitivas de baja duración, por ejemplo, el envío de email al registrarse un usuario, no es una tarea pesada de procesar y a veces solo provisionar un servidor para esta tarea puede ser demasiado. Además este esquema es bajo la filosofía paga lo que usas, así si la aplicación tuvo un mal mes y nadie la usó, los costos asociados serán muy pequeños.
* Contenedores: Los contenedores han venido a dar flexibilidad a las operaciones muy grandes y complejas con esquemas de microservicios y arquitecturas de software como CQRS orquestando todo con la visión de la cultura DevOps. Un contenedor es un paquetes de software que contiene todas las dependencias de software necesarias para ser ejecutado, con la particularidad que este paquete esta desacoplado del sistema operativo, no importa se genera un contenedor en Windows, dará lo mismo si se ejecuta en Windows o en Linux, esa flexibilidad se hereda y resiente en las operaciones mas complejas.
* IaaC ó IaC (Infrastructure as a Code): Significa tener la capacidad de definir la infraestructura por medio de un lenguaje de programación normalmente [declarativo](https://en.wikipedia.org/wiki/Declarative_programming) y dejar que el intérprete comience a provisionar el storage, la red, instancias y todas las configuraciones asociadas, todo a partir de un script. Esta es la promesa de [AWS CloudFormation](aws.amazon.com/es/cloudformation/), generar un script y toda la infraestructura necesaria es provisionada.