Escuela de Negocios
Tipo de documento: Tesis de maestría

EMBA | Executive MBA

Ventajas operativas y mecanismos de optimización en infraestructura de nube pública

Autoría: Gross, Germán
Año: 2025

UNIVERSIDAD
TORCUATO DI TELLA

MBA 2017

VENTAJAS OPERTIVAS Y MECANISMOS DE OPTIMIZACIÓN EN INFRAESTRUCTURA DE NUBE PÚBLICA

Alumno: Gross, German

Tutor: Eisbruch, Gabriel Andrés

Buenos Aires, 2024

AGRADECIMIENTOS

A mi hijo, Nacho, fuente inagotable de motivación. A mi esposa, Silvina, por el apoyo, la paciencia, y el soporte infinito durante todos estos (largos) meses.

A MercadoLibre y Wildlife Studios, por la oportunidad de desarrollarme en una disciplina novedosa y apasionante, en un ambiente inmejorable desde lo humano y lo profesional.

Por último, al “Colo”, amigo y tutor durante esta tesis de posgrado, por su tiempo, apoyo, consejo, y ayuda desinteresada. ¡Gracias!

RESUMEN

Desde hace varias décadas, la evolución tecnológica ha transformado profundamente la vida de las personas y los entornos de trabajo laborales, convirtiendo a los departamentos de TI en una capacidad esencial para el correcto funcionamiento y la competitividad de las empresas. Como parte de esta transformación, la gestión de la infraestructura de tecnología ha evolucionado de “modelos tradicionales”, apalancados sobre centros de cómputo propios, hacia soluciones de servicio basadas en “nube pública” que ofrecieron operaciones a escala con menor inversión, y mayor velocidad y flexibilidad. Sin embargo, esta transformación ha traído nuevos desafíos asociados a la gobernabilidad de los recursos, la gestión de los costos, la eficiencia de las operaciones, y la capacidad de las empresas para adaptarse a esta transformación; con especial énfasis en América Latina, donde aún persisten dudas acerca de la viabilidad y la sostenibilidad económica/financiera de este modelo.

Para analizar este problema, se utilizó un enfoque metodológico descriptivo, combinando análisis cuantitativo y cualitativo. Se realizaron encuestas a profesionales de áreas de tecnología, y entrevistas a expertos en infraestructura de tradicional, y de nube pública. Se complementó el análisis con casos de estudio reales de empresas líderes de tecnología, como Wildlife Studios. El principal objetivo de este trabajo consistió en analizar las ventajas operativas y los mecanismos de optimización en el uso de servicios de infraestructura de nube pública, identificando prácticas clave, herramientas, mecanismos y métricas para mejorar la eficiencia y reducir los costos, sin comprometer la operación del negocio.

Los resultados de este análisis revelaron que, si bien la adopción de nube pública ofrece beneficios como mayor simplicidad, velocidad, flexibilidad y escalabilidad, la falta de una estrategia adecuada para gestionar y optimizar estos servicios usualmente deriva en un aumento descontrolado de los costos, y en una reducción significativa de la capacidad de identificar y asociar esos mismos costos a sus correspondientes unidades de negocio. Se identificaron métricas clave, como los *ratios* de utilización de recursos, y los costos totales por transacción; junto con mecanismos de optimización efectivos, tales como la reserva de recursos y la

utilización de recursos efímeros. Estas métricas y herramientas (entre muchas otras) permitieron a las organizaciones lograr una operación significativamente más eficiente, demostrando que una adopción estratégica de la nube pública no es solo absolutamente viable, sino un diferenciador competitivo crucial en los mercados actuales.

PALABRAS CLAVE

Nube Pública; Eficiencia; Optimización; *FinOps; Cloud Economics*.

ÍNDICE

INTRODUCCIÓN

Tradicionalmente, las empresas siempre fueron dueñas de sus propios centros de cómputos, necesarios para soportar los servicios de las áreas de IT. El hardware y, en general, todo el equipamiento requerido no eran la excepción, significando grandes inversiones y erogaciones de dinero en activos (y soporte), que debían ser gestionados y mantenidos de forma regular y constante.

Durante la década del 2000, nació una nueva alternativa a este modelo: Cloud Computing (público) como servicio, que cambiaría radicalmente la forma en que las empresas (tecnológicas o no) gestionarían su infraestructura IT. En contraposición al modelo tradicional, este permitió prescindir de poseer Hardware propio, dándole la posibilidad a las empresas de reducir el esfuerzo operativo de IT, eliminar el costo que significa adquirir (e instalar) equipamiento propio (CAPEX), y transformar los costos de operación de fijos a variables, teniendo que abonar únicamente por lo que se usa, en el momento en que se usa.

Sin embargo, esta transformación vino acompañada de nuevos desafíos y problemas de diversa índole: desde problemas técnicos, que requirieron nuevos conocimientos en los equipos de ingeniería; problemas operacionales, que obligaron a las organizaciones a adoptar nuevas prácticas, metodologías y herramientas; hasta problemas financieros, asociados al rápido, y muchas veces inesperado, aumento de los costos operativos asociados a recursos de tecnología.

El presente trabajo se focalizó principalmente en explorar y describir las ventajas operativas de la utilización de servicios de nube pública, en comparación con los modelos tradicionales de infraestructura *On-Prem*; y los principales desafíos que esta transformación generó en materia de gobierno operativo, y gestión de costos. En pocas palabras, se describió uno de los principales problemas actuales de la utilización de nube pública: cómo lograr que las organizaciones tengan visibilidad sobre lo que se provisiona y consume como servicio de nube pública, y más

importante aún, cómo asegurar que los costos de la operación tengan sentido para el negocio.

Existen innumerables casos de estudios, incluyendo a grandes empresas como Netflix, Spotify, Instagram, Etsy, entre otras, que resaltan las ventajas de migraciones exitosas hacia la nube pública, explicando como “antes de la pandemia (COVID-19), [la migración hacia la nube pública] era una estrategia popular, pero hoy en día esta migración se convirtió en mandatoria para todas las organizaciones, sin importar que tan grande o pequeñas sean” (Gaca, 2023). Sin embargo, estos procesos estuvieron plagados de riesgos y desafíos, y su éxito no estuvo garantizado en lo absoluto.

Existen decenas de casos donde organizaciones de tecnología deciden migrar hacia la nube pública, y fallaron en el intento, solo para arrepentirse algunos años más tarde, pagando un costo enorme en recursos y tiempo (Metz, 2016). De la misma forma, existen muchos otros casos donde si bien la migración resultó exitosa en el largo plazo, las mismas fueron mucho menos eficientes de lo que podrían haber sido, a causa de una mala planificación y/o ejecución. En muchos de los mismos, el costo de migración terminó por ser exponencialmente más alto de lo estrictamente necesario, usualmente debido a la falta de conciencia y nula visibilidad respecto del costo económico de las decisiones de los equipos de ingeniería, escalando rápidamente a los millones de dólares en ineficiencias (J. R. Storment, 2020).

A lo largo de esta tesis se describieron y exploraron en profundidad estos problemas, y aquellos mecanismos identificados para mitigarlos, con el objetivo de dar una respuesta a aquellas dudas y mitos respecto del Cloud que persisten en la actualidad, especialmente en aquellas empresas que siguen operando de la forma tradicional: ¿Conviene pasarse a un modelo de Cloud Computing? ¿Cuáles son sus ventajas (y desventajas)? A su vez, en este trabajo se analizó un campo para el cual aún no existe estudio extensivo, y que comienza a tomar cada vez más fuerza a medida que más y más empresas se suman a esta tendencia: Economía de la Nube. ¿Es realmente más barato operar en Cloud? Y más importante aún, una vez que se comienza a operar en la Nube, ¿cómo se puede lograr ser más eficiente en la utilización de los recursos? ¿Qué mecanismos existen para reducir los costos operativos?

Con un enfoque especialmente dirigido hacia empresas (pequeñas, medianas y grandes) radicadas en Latinoamérica, el presente trabajo tuvo como objetivo analizar la conveniencia de operar en Cloud, y definir las condiciones bajo las cuales resulta beneficioso adoptar este modelo. Asimismo, se analizó y describió el escenario óptimo para operaciones en ambientes de nube pública: para aquellas empresas que aún operen con infraestructura propia, migrar hacia una infraestructura Cloud es solo el primer paso; existe un esfuerzo posterior (y mucho más grande) si se quiere lograr verdadera eficiencia, que consiste en desarrollar mecanismos efectivos para detectar desvíos, identificar responsables, sugerir y/o automatizar soluciones, y finalmente medir la efectividad de los mismos. Por ende, resultó de vital importancia analizar y presentar herramientas y procesos de diversa naturaleza:

- **Técnicos**, enfocados en recomendaciones y buenas prácticas respecto de decisiones y diseño de arquitectura.
- **Financieros**, asociados a las distintas modalidades de contratación disponibles en los principales proveedores de Cloud.
- **De Negocio**, que apuntan a aquellas decisiones disponibles que permitan lograr descuentos, asumiendo compromisos y/o riesgos operacionales.
- **Procedurales**, que exponen la forma de asegurar que el nivel de eficiencia económica sea sustentable y perdurable en el tiempo.

En última instancia, se intentó medir el beneficio potencial de aplicar los mecanismos y recomendaciones presentados a lo largo del trabajo, para comprobar que, para empresas medianas y grandes, utilizar servicios de Infraestructura Cloud de manera eficiente otorgó un beneficio diferencial, que superó con creces el esfuerzo requerido para lograrlo (incluso considerando el costo de una eventual migración inicial).

Para realizar este estudio se eligió un método de trabajo descriptivo, con un análisis que combina tanto recursos cuantitativos (encuestas a actores de la industria), como cualitativos (entrevistas a expertos). Esta metodología permitió obtener el conocimiento requerido para habilitar un análisis profundo y abarcativo del tema en cuestión, proveyendo información cuantitativa para identificar patrones y correlaciones; y data cualitativa, para comprender y analizar en profundidad las

razones, requerimientos y desafíos del camino hacia la adopción de tecnologías de nube.

El presente trabajo se estructuró en un *Cuerpo Teórico*, compuesto por los capítulos que proporcionan el marco conceptual que permiten al lector comprender los conceptos fundamentales del modelo de infraestructura tradicional, para luego describir el nuevo paradigma de Cómputo en la Nube Pública. A continuación, en el *Cuerpo Empírico* se describió la metodología de investigación, el caso de referencia de *Wildlife Studios* y su experiencia en la utilización de recursos de nube pública; y aquellos mecanismos, métricas y prácticas identificadas como esenciales a la hora de garantizar el éxito en la adopción de *Cloud Computing*. Por último, se procedió a enunciar las conclusiones del presente trabajo, seguido de la bibliografía y los anexos, que contienen la información complementaria utilizada para elaborar la presente tesis.

CUERPO TEÓRICO

El cuerpo teórico del presente trabajo está compuesto por los capítulos que proporcionan el marco conceptual, teórico, e histórico, que permiten al lector comprender los conceptos fundamentales del Cómputo en la Nube.

En esta sección, se intentó describir los conceptos claves y centrales de esta tesis, y definir los términos técnicos fundamentales para entender la presente investigación, dotando al lector de los conocimientos y herramientas requeridos para poder abordar los distintos mecanismos de optimización, necesarios para garantizar una operación eficiente y rentable de las nuevas tecnologías de *Cloud Computing*.

Se comenzó por describir la infraestructura tecnológica tradicional, fuertemente asociada a la construcción y operación de Centros de Cómputos propios, para luego analizar y contrastar con las nuevas tendencias tecnológicas, apalancadas sobre la tercerización de la operación y la propiedad de los Datacenters, la infraestructura de IT, y el consumo de dichos recursos a través de “servicios” comúnmente conocidos como “Servicios Cloud”.

CAPÍTULO I: EL MODELO TRADICIONAL CON DATACENTERS

1.1. Definiendo On-Premise Computing

Es generalmente aceptado que Alan Turing se considera el padre de la Computación Moderna. En 1936, Turing presentó el concepto de “Máquina Universal”, el cual, en la teoría, permitía computar cualquier algoritmo que fuera efectivamente computable. Esta *Máquina Universal* sería luego conocida como “Máquina de Turing”, y sería la piedra fundamental para diseñar las computadoras modernas como las que se usaron para escribir esta tesis, o la que seguramente esté utilizando el lector para leerla.

Es incuestionable la enorme evolución que ha tenido la informática en general desde 1936 a la fecha, principalmente apalancada por el enorme valor que estas computadoras agregaron a la sociedad desde su mismísima invención: Alan Turing utilizó los conceptos de su máquina universal para diseñar y construir una computadora capaz de descifrar los mensajes encriptados del ejército alemán durante la segunda guerra mundial (Agar, 2017). Hoy en día, resulta imposible imaginar una vida sin computadoras.

Mucho antes de ser un bien que en la actualidad se encuentra, en alguna de sus diferentes formas, en casi todos los hogares, a medida que las computadoras modernas fueron evolucionando y su valor de producción se tornó cada vez más accesible, las mismas dejaron de ser bienes reservados para algunas pocas entidades gubernamentales, y empezaron en la segunda mitad del siglo XX, a ser utilizadas por el sector privado. Muchos años después, luego de la invención de los transistores y los medios de almacenamiento portátiles, surgieron los estándares de comunicación que permitirían conectar múltiples computadoras entre sí, y empezaron a surgir lo que hoy conocemos comúnmente como *Data Centers*.

Los *Data Center*s (o Centros de Datos) son los espacios utilizados para alojar las computadoras, las cuales llamaremos “servidores”, para desambiguarlas de lo que hoy en día conocemos como “computadoras personales”. Con una necesidad creciente de contar cada día con más poder de cómputo para realizar cada vez más (y más complejas) operaciones, surgió la necesidad de poder contar con un espacio

capaz de albergar una gran cantidad de servidores, que permitiera operarlos, interconectarlos, proveer la suficiente capacidad energética, seguridad y todos los múltiples requerimientos necesarios para lograr que estos servidores funcionen de forma adecuada.

Un Datacenter está compuesto por múltiples partes, que se agrupan principalmente en cuatro categorías, que se enumeran a continuación (Telecommunications Industry Association, 2024):

- **Instalaciones:** Es la edificación de las distintas partes de un *Datacenter*, incluyendo las soluciones para garantizar la seguridad física, y los sistemas de prevención de incendios y otras catástrofes.

- **Salas de IT:** Son las salas específicas para instalar y operar los servidores, incluyendo los racks donde se monta el *hardware*, las unidades de distribución de energía (PDUs) para alimentar los dispositivos, los distintos sensores que miden las condiciones ambientales y energéticas, el piso técnico, y otros componentes.

- **Energía y Refrigeración:** Son todos los componentes destinados a proveer y garantizar el suministro energético de forma segura e ininterrumpida. Incluye los transformadores, las baterías, el cableado, y los generadores de emergencia, entre otros. Se incluye además el equipamiento requerido para garantizar los controles ambientales óptimos.

- **Infraestructura:** Se compone por todo el equipamiento de IT desplegado para correr las aplicaciones, y proveer servicio al negocio y sus usuarios. Comúnmente, los servidores y otros componentes de *hardware* necesarios para ejecutar aplicaciones informáticas, y transmitir y almacenar información.

1.2. Instalaciones

Las instalaciones de un Datacenter son, en su expresión más básica, almacenes donde se instala y opera equipamiento de IT. Sin embargo, estas instalaciones se

construyen con el objetivo de garantizar las siguientes características: i) Garantizar la seguridad física; ii) Asegurar la continuidad ante eventuales catástrofes; iii) Maximizar las prestaciones de conectividad; y iv) Proveer suficiente capacidad (medida en m^2/m^3) y potencial para expansión futura.

Es usual que los Datacenters modernos sean un activo crítico para el sector privado y el sector público. Hoy en día, las naciones y la mayoría de las empresas requieren, en menor o mayor medida, de sistemas informáticos para funcionar. Es por esto que protegerlos, para garantizar la continuidad del negocio, la seguridad nacional, o la correcta operación de una organización, resulta de vital importancia.

Garantizar la seguridad física es un objetivo primordial en todo Datacenter. Los Datacenters modernos garantizan la seguridad mediante un diseño en capas, que generalmente comienza por segurizar un **perímetro exterior** mediante muros o rejas. Dicho perímetro exterior debe contar con puntos de acceso limitados y controlados, y estar preparados para minimizar intrusiones por choques con vehículos motorizados, o a través de escaladas.

La siguiente capa de seguridad consiste en el **perímetro interior**, constituido por las edificaciones cerradas. Dichas edificaciones deberán estar lo suficientemente alejadas, de modo de garantizar que un evento en el perímetro exterior (por ejemplo, una explosión) no afecte la seguridad del perímetro interior. Es en este perímetro donde generalmente ocurren los chequeos de acceso más rigurosos, como validar la identidad de las personas, validar sus niveles de acceso autorizados, que no se ingresen sustancias o elementos peligrosos, etc.

Una vez superado el perímetro interior, el siguiente estrato son los **cuartos (o jaulas) de cómputo**, lugar donde se encontrará el equipo de IT o hardware. Dichos cuartos generalmente cuentan con control de acceso independiente, y automatizado por llaves físicas o digitales.

Por último, la siguiente y última capa se trata del acceso al **Rack**. Los Racks son estructuras metálicas (gabinetes), que se utilizan para montar múltiples equipos, de

una forma espacio-eficiente. Es usual y recomendable que cada Rack cuente con acceso mediante una llave física independiente.

Sin embargo, la seguridad física de un Datacenter no solo se asegura a través de prevenir ataques humanos de terrorismo, vandalismo, u otros; sino también ante eventuales catástrofes naturales. Es por esto que los Datacenters son construidos usualmente en áreas que minimicen la probabilidad de ocurrencia de actos de Dios, como, por ejemplo, inundaciones, terremotos, huracanas, tsunamis, etc.

Otra característica fundamental de los Datacenters es que rara vez operan de forma independiente. Usualmente, las aplicaciones alojadas en dichos Centros de Cómputo tienen dependencias con otras aplicaciones o componentes alojados en otros Datacenters, que pueden estar alejados por algunos pocos kilómetros, o en otros continentes. Para garantizar que esta dependencia se satisfaga, es necesario que los Datacenters cuenten con excelente conectividad, tanto a través de internet, como de redes o vínculos privados. Es por esto que, generalmente, se busca emplazar los Centros de Cómputos cercanos a centros neurálgicos de conectividad internacional. Ejemplos de estas locaciones son ciudades como Ashburn, Sterling (Virginia del Norte, EE. UU.), Londres (Reino Unido), Tokio (Japón), Frankfurt (Alemania), entre muchas otras (Pilz, 2023).

Por último, la construcción de un Datacenter moderno significa una inversión financiera significativa. Es por esto que a la hora de diseñar un Datacenter, se debe contemplar una capacidad suficiente para cubrir las necesidades del negocio, pero también ser diseñado con capacidad de ampliación, para soportar potencial demanda futura.

1.3. Salas de IT

El objetivo principal de las salas de IT es albergar los equipos que conforman la infraestructura de IT de un Datacenter. Las Salas de IT son diseñadas con el objetivo de garantizar un entorno óptimo para la correcta operación de la infraestructura, y generalmente se dividen en dos tipos de Salas:

- **Salas de Cómputo:** Son las salas de mayor tamaño, aquellas que albergan los dispositivos de infraestructura que actúan como los "cerebros" de la operación: procesan, almacenan y transmiten datos.

- **Salas de Comunicación:** Son salas (generalmente) de menor tamaño, que albergan los dispositivos de comunicación, aquellos encargados de transmitir y dirigir los datos dentro y hacia fuera del centro de cómputos. Dichas salas son igualmente criticas para la operación, pero generalmente requieren de menor espacio, menor capacidad energética y menor capacidad de refrigeración.

Las Salas de IT son el corazón de un centro de cómputos, y por ende, son los espacios con mayor seguridad y protección, ubicándose lejos de fuentes de contaminación y vibraciones. Generalmente, se busca que sus paredes y techos sean resistentes al fuego y a las interferencias electromagnéticas. Para ello, se utilizan materiales como concreto, paneles de yeso y acero para garantizar la seguridad y estabilidad estructural.

Si bien existen distintos paradigmas para construir una Sala de IT, es muy común contar con techos con altura suficiente para para permitir la adecuada circulación del aire, la instalación de sistemas de enfriamiento, de extinción de incendios, y requerimientos de cableado.

Las principales características que distinguen a las Salas de IT de otros edificios, o secciones del Centro de Datos, son:

- **Sistemas de Control de Incendios:** Al estar densamente poblados por componentes eléctricos, uno de los mayores riesgos dentro de los Centro de Datos son los incendios, que representan una amenaza significativa que puede destruir (en un tiempo muy corto) el Datacenter de forma permanente, ya sea parcialmente, o en su totalidad. Los centros de datos modernos confían en diferentes sistemas de supresión de incendios, cada uno con ventajas y aplicaciones específicas, para proteger estos espacios críticos, siendo el más

común el sistema de supresión por gases (*Clean Agent Fire Supression*). En cuanto detectan humo, o al ser accionados de forma manual, dichos sistemas inundan las Salas de IT con agentes gaseosos que desplazan el oxígeno o interrumpen químicamente la reacción de combustión, extinguiendo el fuego sin utilizar agua. La principal razón para evitar el agua se centra en evitar el daño que los líquidos ocasionan en equipos eléctricos, no solamente destruyéndolos, sino también, en muchos casos, empeorando el problema. Los tipos de gases más comúnmente utilizados por estos sistemas son el *FM-200*, *Novec 1230*, o *Inergen*, siendo el primero el más utilizado. Como principal ventaja, los sistemas de supresión por gases no dejan residuos, no causan daños a equipos electrónicos, y extinguen el fuego rápidamente. Sin embargo, cabe aclarar que dichos sistemas son significativamente más costosos (en comparación con los sistemas de agua), y podrían representar riesgo de asfixia para el personal si no se evacúa a tiempo (Nilsson, 2014).

Cabe aclarar que estos sistemas de supresión de incendios no reemplazan la necesidad de contar con matafuegos (especialmente de tipo C, para extinguir fuegos de origen eléctrico), los cuales son requerimientos de seguridad básicos a la hora de operar un Centro de Datos.

- **Gestión de Cables:** Las Salas de IT que alojan cientos, miles, o incluso decenas de miles de dispositivos electrónicos, deben contar también con cientos o miles de cables para alimentar individualmente estos dispositivos, como así también conectarlos entre sí para permitir la transferencia de datos. En un Datacenter promedio, se podrá encontrar con una gran cantidad de cables, que deben ser organizados y gestionados de manera estructurada, para garantizar una operación eficiente, contemplando aspectos como mantenibilidad, extensibilidad, confiabilidad, y solución de problemas.

  La arquitectura de cableado estándar incluye tres componentes principales: i) el **Cableado Horizontal**, que conecta los equipos de comunicación con los servidores; ii) el **Cableado Vertical (o Troncal)**, que conecta los distintos espacios o secciones de los Datacenters, y por ende deben ser conexiones de alta capacidad (ancho de banda); y iii) los **Canales y Espacios**, que definen la estrategia de cableado, que podrá ser aérea (a través de bandejas elevadas) o bajo-piso (colocando los cables por debajo del piso técnico del Datacenter).

En cualquier caso, la estrategia de cableado debe garantizar la correcta capacidad para albergar los cables requeridos, garantizar el fácil acceso a los mismos, facilitar el flujo de aire, y mantener una separación física de cables de energía y datos, para minimizar la interferencia electromagnética.

Por último, dentro de un Datacenter, es de vital importancia el etiquetado o rotulado de los cables, para facilitar la gestión de estos. Es de práctica común etiquetar cada cable, en ambos extremos, con la información correspondiente que detalle i) el **Tipo y Propósito** del cable (por ejemplo, fibra o cobre, y categoría del cable); ii) el **Origen y Destino** del cable, usualmente incluyendo un identificador único de equipo y numero de puerto; y iii) la **Fecha de Instalación**, para identificar antigüedad del cable, a fines de mantenimiento y reemplazo preventivo (IEC, 2023).

- **Iluminación:** En Salas de IT, se recomienda instalar luz uniforme y suficiente en todos los espacios, siendo recomendable contar con ~750 Lux de intensidad y una temperatura de color entre 4000K y 5000K. En Datacenters donde la presencia humana es frecuente pero escasa, suele recomendarse sistemas de control automatizado para apagado de luces (por falta de movimiento) para maximizar la eficiencia energética.

  Adicionalmente, los Datacenters también deben respetar los códigos de construcción locales, específicamente aquellos asociados a requerimientos de iluminación de emergencia en caso de fallas en el suministro energético (Formation, 2024).

CAPÍTULO II: SERVICIOS E INFRAESTRUCTURA DE UN DATACENTER

2.1. Energía y Refrigeración

La provisión energética es un aspecto fundamental en el diseño y operación de un Centro de Datos. La gran mayoría de los sistemas y equipos que se alojan dentro de los mismos requiere de energía eléctrica para funcionar, y en su conjunto, el consumo de energía es muy elevado (especialmente en grandes centros). Se calcula que, durante 2022, alrededor del 1.0% a 1.3% del consumo de electricidad global se debió a la operación de Data Centers (excluyendo la minería de criptomonedas). Algunos Centros de Datos pueden incluso tener consumos superiores a los 100 mega watts, energía suficiente para alimentar más de 80.000 mil hogares (Andrae, 2015).

En promedio, más del 80% de la energía necesaria para operar un Datacenter es utilizada por servidores, y sistemas de refrigeración y provisión de energía. El 20% restante es utilizado por los sistemas de almacenamiento, sistemas de comunicación y otros (iluminación, controles de acceso, CCTV, etc.).

Debido a que la provisión energética es un requisito excluyente para la correcta operación de un Centro de Datos, y que a su vez la continuidad en la operación de los mismos es crítica para la operación de uno o más negocios, los Datacenters son diseñados con alta toleración a fallos, de manera de garantizar redundancia ante incidentes, asegurando la provisión ininterrumpida del servicio. Los principales medios para lograr estos niveles de servicio se basan en la implementación de las siguientes estrategias (Uptime Institute, 2017):

- **Múltiples fuentes de Alimentación Externa:** La provisión eléctrica es generalmente un servicio contratado a terceros. Es por esto que se busca tener redundancia y contar con, al menos, dos proveedores independientes. Si uno cualquiera fallara, el segundo proveedor deberá ser capaz de proveer la energía requerida para soportar la totalidad de la operación.

- **Múltiples fuentes de Distribución Interna:** Los proveedores de energía eléctrica generalmente llegan hasta las inmediaciones de los Centros de Cómputo. El Datacenter deberá contar con transformadores y líneas de alimentación duplicados e independientes, de manera de eliminar puntos de fallos únicos en la distribución energética desde la sub-estación del proveedor de energía, hasta los equipos del centro de cómputos.

- **Sistemas de Alimentación Ininterrumpida (UPS):** Incluso con múltiples proveedores externos de energía, existe la posibilidad de interrupción en la provisión energética, por ejemplo, porque los múltiples proveedores fallan de forma simultánea, o porque los sistemas de provisión energética son pasivo-activo y necesitan de unos instantes para conmutar entre proveedores (o generadores de emergencia). Es por esto por lo que la mayoría de los Datacenters cuentan con UPS: grandes baterías que son capaces de sostener la operación del Datacenter por tiempos breves (desde varios minutos, a algunas pocas horas).

- **Generadores de Emergencia:** Ante fallas en el suministro de energía eléctrica externa, es una práctica común contar con generadores de emergencia que puedan abastecer las necesidades energéticas del Datacenter. Estos deben activarse de manera automática ante la detección de una interrupción en el servicio de suministro energético externo (o ante la detección de una degradación del servicio externo, tal como una caída en el voltaje requerido). Los generadores de emergencia funcionan comúnmente a diésel o gas natural, y dependiendo del tamaño y criticidad del Datacenter, pueden requerirse múltiples generadores, tanto para garantizar la provisión de energía suficiente, como para garantizar la redundancia en caso de fallas en los mismos generadores de emergencia. Adicionalmente, los Datacenters que cuentan con generadores de emergencia, deben también contar con i) reservas de combustible (gasoil o gas natural) para operar el generador; y ii) acuerdos de provisión de combustible con terceros para garantizar el suministro ininterrumpido de gasoil o gas natural, en caso de problemas prolongados con el suministro externo de energía.

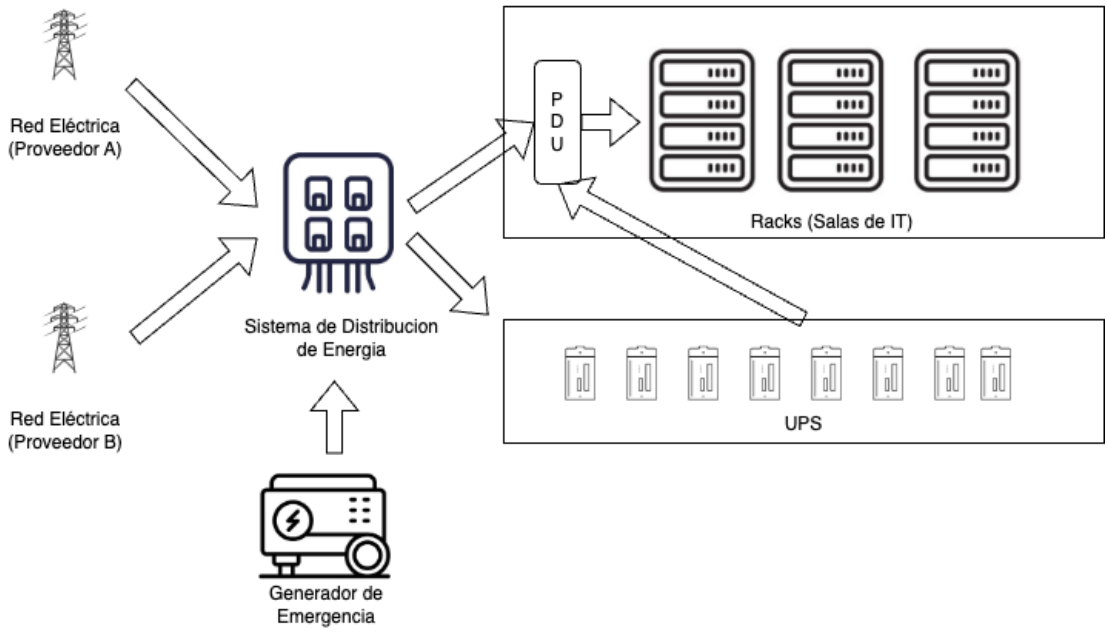


Ilustración 1 - Diagrama de Provisión de Energía (Fuente: Elaboración propia)

Vale mencionar que los PDUs (*Power Distribution Unit*) son los dispositivos (usualmente, dos por Rack) encargados de distribuir y monitorear la energía eléctrica a los servidores, de forma organizada, redundante y segura.

La refrigeración es otro aspecto critico en el diseño y operación del centro de datos, ya que los servidores (y otros equipos) generan una gran cantidad de calor que debe ser disipado, para no alterar la correcta operación, rendimiento y confiabilidad de los equipos de infraestructura que funcionan dentro de los Datacenters.

Los Centros de Datos deben operar bajo ciertas condiciones ambientales que en general oscilan entre los 15 y 25 grados Celsius, y una humedad relativa entre 40% y 60%. Para lograr estas condiciones, existen diversas técnicas de refrigeración, cada una con sus propias ventajas y desventajas, que se adaptan a diferentes necesidades y presupuestos. A continuación, se enumeran las principales, y por ende, más comunes (Henshaw, 2015):

- **Pasillos fríos y calientes:** Puede considerarse la técnica más común y efectiva para la refrigeración de Centros de Datos. Su funcionamiento se basa en la separación física de los pasillos donde se ubica el frente de los servidores

(pasillos fríos, por donde se absorbe el aire), y los pasillos por donde se ubica el contrafrente de los servidores (pasillos calientes, por donde se expulsa el aire recalentado).

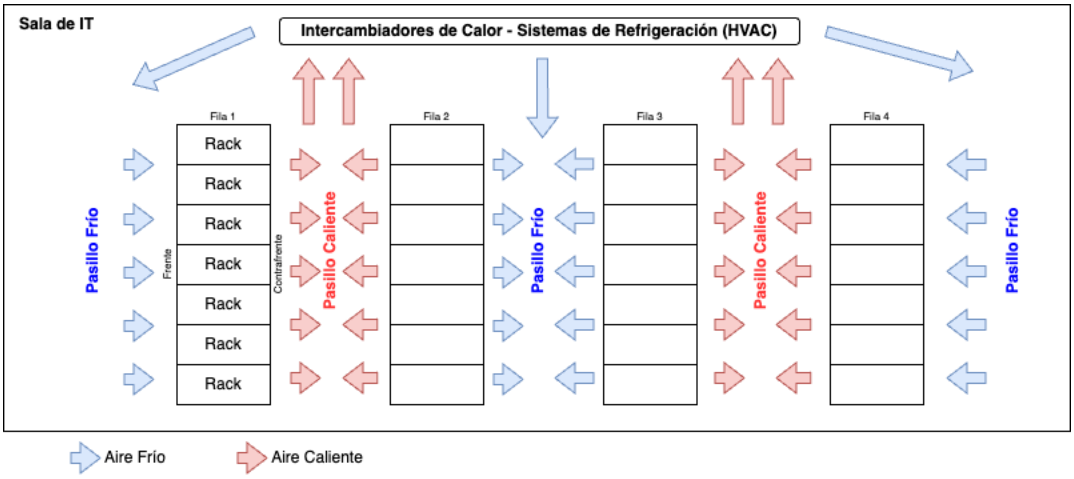


Ilustración 2 - Diagrama de Refrigeración: Pasillo Frío/Caliente (Fuente: Elaboración propia)

Esta técnica se caracteriza por ser relativamente simple de implementar y escalar a distintos tamaños y configuraciones de Datacenters. Es, además, una técnica de bajo costo de implementación, en comparación con otras.
En contraposición, esta estrategia no sobresale por su eficiencia, y necesita de mayor espacio adicional para crear pasillos fríos y calientes, lo que puede ser un problema en centros de datos con limitaciones de tamaño. Adicionalmente, esta configuración puede crear puntos más calientes que otros, dependiendo de la densidad y el tipo de equipos, lo que puede representar un problema difícil de gestionar y/o solucionar (Cho & Kim, 2011).

- **Contención de aire frío/caliente:** Esta técnica es una evolución de la anterior que se logra aislando los pasillos fríos de los pasillos calientes, usualmente mediante paneles o cortinas de plástico. De esta forma, se minimiza la mezcla de aire frío con aire caliente. La principal ventaja de este método es la mejor en la eficiencia, y por ende, la disminución de la energía requerida para refrigeración. Asimismo, la contención de aire frío/caliente disminuye la formación de puntos calientes. En contraposición, esta técnica tiene una complejidad mayor, y un costo de implementación más elevado. Al igual que el método anterior, implica contar con espacio adicional no aprovechable.

- **Extracción de Calor Intra-Rack:** Este método consiste en extraer el calor generado por cada Rack, dentro del mismo Rack, y redirigiéndolo hacia el exterior, de modo que el aire recalentado por los servidores nunca llegue a la Sala de IT. Otra alternativa para implementar este método consiste en tener unidades de refrigeración (compresores y enfriadores) incorporadas en cada Rack, absorbiendo el calor generado por los equipos, y expulsando aire enfriado a la Sala de IT.

  Las ventajas de esta técnica es que son muy eficientes en el uso de espacio. En contraparte, son generalmente más costosas de implementar, mucho más complejas, y la densidad máxima por Rack suele ser menor.

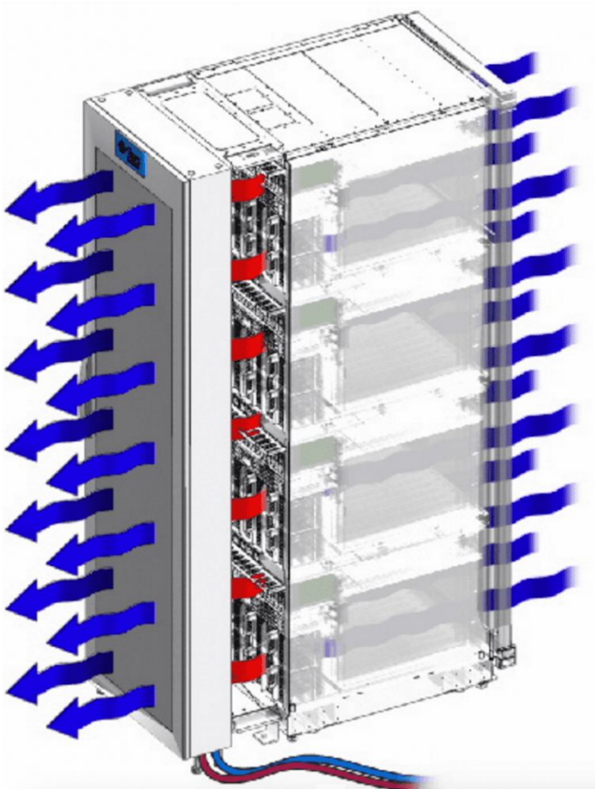


Ilustración 3 - Diagrama de Refrigeración: Extracción de Calor Intra-Rack (Fuente: (Henshaw, 2015))

- **Otras:** Existen otras técnicas de refrigeración más modernas y experimentales, como por ejemplo *Racks Refrigerados por Agua*, o *Enfriamiento por Inmersión Líquida* que son mucho más eficientes que los métodos listados anteriormente. Sin embargo, aún son poco usados en la actualidad, principalmente por su complejidad y costo, y sobre todo por el riesgo implícito de utilizar líquidos en cercanías de componentes eléctricos críticos (Haghshenas, 2023) (ante fallas,

el potencial de desastre es alto). Dichos mecanismos no serán analizados en el presente trabajo.

2.2. Infraestructura

El término “infraestructura” refiere a aquellos equipos y componentes que se encuentran dentro de los Racks, en las Salas de IT. Si bien pueden encontrarse diversos tipos de equipos, de un sinfín de fabricantes distintos, generalmente, la mayoría de la infraestructura encontrada en un Centro de Datos puede clasificarse dentro las siguientes categorías:

- **Equipos de Cómputo (*Servers*):** Tal como fue mencionado con anterioridad, los servidores son el “cerebro” del Centro de Datos. Son los equipos que se encargan de procesar los datos y las operaciones ejecutadas dentro del Datacenter. Generalmente, se mide su capacidad en cantidad de CPU, velocidad de procesamiento (GHz), y memoria disponible. Pueden encontrarse en distintos formatos, siendo los más comunes los i) *Servidores en Rack*, diseñados para ser apilados dentro del Rack, ofreciendo flexibilidad de configuración; y ii) *Servidores Blade*, modulares y de alta densidad que se instalan dentro de un chasis, permitiendo una gestión simplificada.

- **Equipos de Almacenamiento (*Storage*):** Son los equipos que se utilizan para almacenar y recuperar datos de forma masiva y eficiente. Son equipos dedicados que cuentan con una (o más) controladoras (el cerebro del equipo), y cajones de dispositivos de almacenamiento (discos). Los discos de almacenamiento pueden ser de distintas clases, desde los tradicionales discos mecánicos magnéticos (lentos, pero baratos) hasta los más modernos discos de estado sólido (más rápidos, pero significativamente más caros). Es muy común encontrar Equipos de Almacenamiento con configuraciones hibridas, contando con discos magnéticos, que se complementan con otros de estado sólido. Comúnmente, los equipos de Storage se dividen en dos tipos de arreglos (Mohammed, 2024): i) SAN (*Storage Area Network*), que son sistemas de almacenamiento de alta velocidad con bloques de datos, ideales para sistemas de alto tráfico como bases de datos y aplicaciones; y ii) NAS (*Network*

Attached Storage), que son sistemas de almacenamiento basados en archivos, fáciles de acceder y gestionar, utilizados principalmente para almacenar y compartir archivos.

- **Equipos de Comunicación (*Networking*):** Son los equipos que se encargan de distribuir, dirigir y filtrar los paquetes de datos a través del Centro de Datos, tanto internamente, como aquellos que se reciben o envían desde el exterior del Datacenter. Son los equipos que se encargan de definir y habilitar las distintas redes del Centro de Datos. Los Equipos de Comunicación se dividen en i) Pasivos, aquellos que no necesitan electricidad, como las *patcheras*, encargadas de gestionar las distintas conexiones de cables dentro del Rack; y ii) Activos, aquellos que si necesitan de energía eléctrica para funcionar.
  Dentro de la categoría de componentes Activos, los equipos que más comúnmente se encuentras en los Centros de Datos tradicionales son: i) *Switches*, aquellos dispositivos que conectan servidores, dispositivos de almacenamiento y otros componentes dentro del Data Center. Los *switches* de un Data Center, generalmente se dividen en *swicthes* de *core* (la capa superior que conecta con los principales enlaces de datos del Centro de Datos), de *distribución* (la capa intermedia que conecta los *switches de core* con los *switches de acceso*) y de *acceso* (la capa inferior, que conecta con los dispositivos). En segundo lugar, encontramos los ii) *Enrutadores* (o *Routers*), que dirigen el tráfico entre las diferentes redes, incluyendo internet y WANs (*Wide Area Network*); también iii) *Balanceadores de Carga* (*o Load Balancers*), cuya función es distribuir el tráfico entre varios servidores para optimizar el rendimiento y evitar sobrecargas; y por último iv) *Cortafuegos* (o *Firewalls*), que filtran el tráfico entrante y saliente para proteger la red contra amenazas de seguridad (Mauricio Arregoces, 2003).

- **Equipos de Seguridad (Security):** Los equipos de seguridad son aquellos que están exclusivamente dedicados a detectar, y en algunos casos prevenir o mitigar, problemas o amenazas que puedan comprometer la seguridad lógica de los sistemas alojados en el Centro de Datos. Los dispositivos más comunes en esta categoría son i) Sistemas de detección de intrusiones (IDS), pensados

para detectar actividades maliciosas dentro de la red del centro de datos; y ii) Sistemas de prevención de intrusiones (IPS), que detienen activamente los ataques detectados, idealmente, antes de que puedan causar daños.

CAPÍTULO III: UN NUEVO PARADIGMA CON CLOUD COMPUTING

3.1. Definiendo Cloud Computing

Siendo “Cloud Computing” un concepto principalmente abstracto, si se consultara a distintos especialistas sobre una definición universal que describa la naturaleza de del mismo, probablemente se obtenga una definición distinta por cada persona a la que se le consulte. Las siguientes se tratan de las definiciones provistas por los tres proveedores más importantes de Nube Pública en la actualidad, con un 63% del *Market Share* en Q3 2024 (Richter, 2024):

- **Amazon Web Services (AWS):** “El Cómputo en la nube es la entrega bajo demanda de recursos de TI, a través de Internet, con un modelo de precios *pago-por-uso*. En lugar de comprar, poseer y mantener centros de datos físicos y servidores, se puede acceder a servicios de tecnología, como capacidad de cómputo, almacenamiento y bases de datos a demanda a través de un proveedor de nube...” (AWS, What is cloud computing?, 2013)

- **Microsoft Azure (Azure):** “... la entrega de servicios de cómputo —incluyendo servidores, almacenamiento, bases de datos, conectividad, analítica e inteligencia— a través de internet (‘la nube’) para ofrecer innovación acelerada, recursos flexibles y economías de escala.” (Azure, 2022)

- **Google Cloud Services (GCP):** “[en la Nube] el capital invertido en construir y mantener centros de datos se reemplaza por el consumo de recursos elásticos de TI en forma servicio de utilidad, entregados por un ‘proveedor’ (incluyendo almacenamiento, cómputo, conectividad, procesamiento de datos y analítica, desarrollo de aplicaciones, aprendizaje automático, e incluso servicios completamente gestionados).” (Google, 2016)

Las tres definiciones anteriores fallan en explicar las características básicas de un Servicio de Nube Pública, y apuntan más bien a definir un modelo de servicio donde la característica fundamental consiste en la tercerización de la propiedad del capital de trabajo asociado a los centros de datos y el hardware necesario para montar una

operación de TI. Sin embargo, si este fuera el caso, cabe preguntarse por qué el mercado de nube pública es ampliamente dominado por AWS, GCP y Azure, y no por proveedores como IBM, HPE, Dell, Lenovo, que dominan desde hace décadas el mercado de fabricación y venta de Hardware para centros de datos, con un *market share* del 47% durante 2019 (IDC, 2020). Las definiciones anteriores fallan en resaltar y explicar la diferencia fundamental entre el modelo de Infraestructura Tradicional, en comparación con los nóveles modelos basados en *Cloud*: La Nube Pública es mucho más que un servicio de *Hardware as a Service*, donde se disponibilizan dispositivos para que sean utilizados y administrados por el cliente. El valor de la Nube Pública radica en los servicios y ecosistemas montados sobre ese *hardware*, que hoy en día es considerado un *commodity*.

En base a lo anterior, se podría definir *Cloud Computing* como la provisión dinámica de capacidades de TI a través de una red, incluyendo *hardware*, *software* y/o servicios gestionados. Adicionalmente, para ser considerado tal, un servicio de *Cloud Computing* debe cumplir, al menos, con las siguientes características:

1. **Servicios básicos convergentes y elásticos:** Los servicios ofrecidos deben incluir las capacidades fundacionales mínimas y necesarias de TI, incluyendo pero no limitado a, cómputo, almacenamiento y conectividad; y las mismas deben tener la capacidad de crecer y aumentar tanto como el usuario lo requiera[1].

2. **Acceso a los recursos a través de la Red:** Los servicios deben estar disponibles para ser consumidos a través de la red pública (Internet), y/o privada (enlaces dedicados). Dichos accesos deben establecerse utilizando mecanismos estandarizados.

3. **On-Demand y Self-Service:** Los usuarios deben ser capaces de provisionar (y des-provisionar) recursos y servicios unilateralmente, y de forma automática.

[1] Sujeto a ciertos límites técnicos, específicos a la tecnología y/o el proveedor de nube utilizado

Dichos recursos y servicios deben estar disponibles y expuestos a través de algún catálogo.

4. ***Accountability* por la provisión y uso de los recursos:** Deben existir mecanismos (automáticos) de trazabilidad para entender qué se consume, quién lo consume, cuándo, y por cuánto tiempo.

5. **Monitoreo y auditoría:** Debe ofrecerse capacidades mínimas de monitoreo, de forma de poder visualizar y evaluar el estado y uso de los servicios y recursos contratados, en tiempo real. Adicionalmente, todas las acciones realizadas sobre sobre dichos recursos y servicios deben ser registradas de forma segura. Dichos registros de auditoria deben ser puestos a disposición del cliente y/o usuario del servicio.

3.2. Los Distintos Tipos de Cloud Computing

Si bien existen múltiples formas de catalogar los distintos tipos de *Cloud Computing*, predominan dos clasificaciones que son las más utilizadas para describir las distintas ofertas: A) a partir de la **Propiedad**; y B) a partir del **Nivel de Acceso**.

Desde un punto de vista de la *Propiedad*, existen cuatro tipos principales de Nube:

- **Privada:** Son Nubes administradas y en posesión de una única empresa, para uso privado (*Single Tenant*), la cual generalmente, no se vende como servicio a terceros. Las Nubes Privadas están asociadas con altos costos de TCO (*Total Cost of Ownership*) y son altamente compatibles con operaciones que involucran grandes volúmenes estables de trabajo.

- **Pública:** Al contrario de la anterior, las Nubes Públicas son *Multi Tenant*. Esto significa que son utilizadas y consumidas por más de una entidad o cliente, o en su defecto, están disponibles para serlo. Generalmente, el dueño de este tipo de Nubes vende el derecho de uso de la misma, cobrando de acuerdo con los recursos y servicios utilizados. Ser consumidor de servicios de Nube Pública tiene bajas barreras de entrada, con costos de TCO y mantenimiento

relativamente bajos. Es altamente compatible con flujos de baja demanda de recursos, o alternativamente, con volúmenes de trabajo altamente variables. Ejemplos de proveedores de Nube Pública son los mencionado anteriormente: Amazon (AWS), Microsoft (Azure), y Google (GCP).

- **Híbrida:** Las nubes híbridas son aquellas nubes que combinan componentes tanto de Nube(s) Privada(s) como de Nube(s) Pública(s). Un ejemplo teórico de este caso es una empresa que monta su propia Nube Privada en sus Centros de Datos (utilizando alguna tecnología de las muchas disponibles, como *OpenStack*), pero decide alojar una parte de sus flujos de trabajo en una Nube Pública (por ejemplo, AWS). En general, los flujos alojados en una y otra Nube tendrán algún tipo de conexión, de forma de poder comunicarse en caso de ser necesario, y poder ser administrados de forma centralizada por el equipo de TI correspondiente.

- **Comunitaria:** Las Nubes Comunitarias son esencialmente Nubes Privadas, pero en posesión de varios individuos u organizaciones que, generalmente, comparten algún interés o actividad. Si bien el uso de dicha Nube es compartido entre todos los participantes, la administración puede ser compartida entre los miembros de la Comunidad o delegada en una tercera parte.

Por otro lado, cuando la clasificación corresponde al *Nivel de Acceso*, existen tres grandes tipos de *Cloud Computing* (Jackson, 2018):

- **Infraestructura como Servicio (IaaS):** Es la capacidad de provisionar procesamiento, almacenamiento, conectividad, entre otros recursos fundamentales de IT. El consumidor es capaz de instalar y ejecutar software de forma arbitraria, incluyendo sistemas operativos y aplicaciones propias o de terceros. Ejemplos de servicios de IaaS son: AWS Elastic Cloud Computing (EC2), AWS Simple Storage Service (S3), Google Compute Engine (GCE), entre otros.

- **Plataforma como Servicio (PaaS):** Se provee una plataforma tecnológica que permite a sus usuarios desarrollar, ejecutar y gestionar aplicaciones. A diferencia del modelo de IaaS, el usuario no debe ocuparse de las complejidades propias de construir y gestionar la infraestructura subyacente a los entornos de ejecución. Ejemplos de servicios de PaaS son: AWS Elastic Beanstalk, Heroku, Google App Engine, entre otros.

- **Software como Servicio (SaaS):** Se trata de un modelo de licenciamiento y entrega de Software, usualmente basado en la *web* o móvil nativo, accedido a través de Internet y alojado en servidores centrales, donde los usuarios abonan una suscripción que les otorga el derecho a acceder y a utilizar dicho Software por un período de tiempo determinado. Ejemplos de servicios de SaaS son: Microsoft Office 365, Google Workspace, Mercado Pago, entre otros.

Las tres categorías anteriores no se distinguen entre sí por el *stack* tecnológico que utilizan, sino por el nivel de abstracción y grado de control que otorgan al usuario del servicio. En la siguiente imagen se ilustran los principales componentes tecnológicos que conforman un servicio de *Cloud Computing* genérico, discriminando de acuerdo con el color, quién es el actor responsable por administrar y gestionar cada uno de ellos:

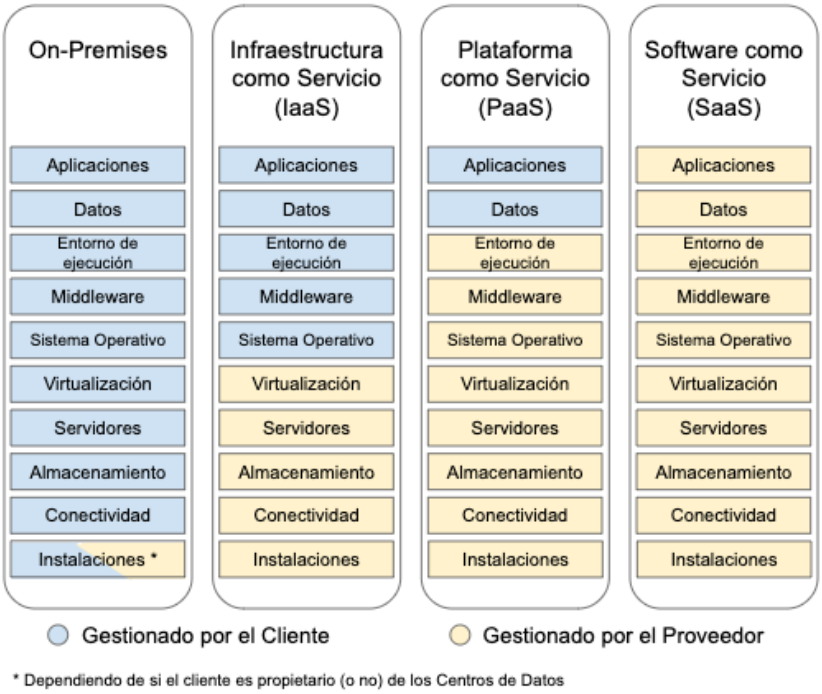


Ilustración 4 - Tipos de Servicios de Nube (Fuente: Elaboración propia)

Cabe mencionar que existen fuentes que describen clasificaciones adicionales. Sin embargo, las mismas suelen ser algún tipo de variación menor, o combinación de uno o más tipos de los anteriormente citados.

CAPÍTULO IV: GESTIÓN DE LA CAPACIDAD

4.1. Gestionando la Capacidad de Datacenters On-Premise

En la operación de Datacenters físicos, la práctica de gestión de la capacidad emerge como una herramienta indispensable para asegurar el dimensionamiento y la utilización óptima de los recursos. La gestión de la capacidad tiene como finalidad garantizar que los recursos tecnológicos posean la potencia y capacidad adecuadas para asegurar la demanda actual y futura del centro de datos, optimizando los costos de capital y operación (D. Gmach, 2007).

Para lograr este objetivo, la gestión de la capacidad debe enfocarse en cumplir con los siguientes objetivos:

- **Asegurar la capacidad del servicio:** manteniendo los recursos mínimos necesarios para cumplir con los niveles de servicio comprometidos para con el negocio.

- **Mitigar riesgos operativos:** identificando y eliminando cuellos de botella que puedan comprometer la operación.

- **Optimizar los costos:** garantizando que no se posea más recursos de los necesarios, minimizando la subutilización de las instalaciones, y por ende, evitando costos de capital y de operación innecesarios.

Siempre que se decida ampliar la capacidad instalada de un centro de cómputos, se deberá realizar una inversión de capital significativa, para adquirir el *hardware* requerido. Adicionalmente, se deberán contemplar los costos logísticos para trasladar dicho equipamiento hacia su locación final, y los costos por única vez de instalación y configuración de dicho equipamiento, que consistirá principalmente en mano de obra, pero también podría significar costos de licenciamiento de software, en aquellos casos donde el modelo sea dependiente de la capacidad del *hardware* subyacente donde se ejecuta dicho software.

La gestión de la capacidad es la práctica responsable de decidir el momento y la cantidad en la que se expandirá la capacidad de un centro de cómputos, buscando siempre minimizar la capacidad ociosa. Sabiendo que la capacidad ociosa corresponde a recursos sin utilizar, estos significan un costo de oportunidad que podría ser invertido con otros fines, como por ejemplo, otros proyectos que otorguen una rentabilidad positiva (Allspaw, 2008).

En la Ilustración 5 - Demanda vs Capacidad On-Prem, se muestra en el área verde el costo de oportunidad de los recursos ociosos de un centro de cómputos.

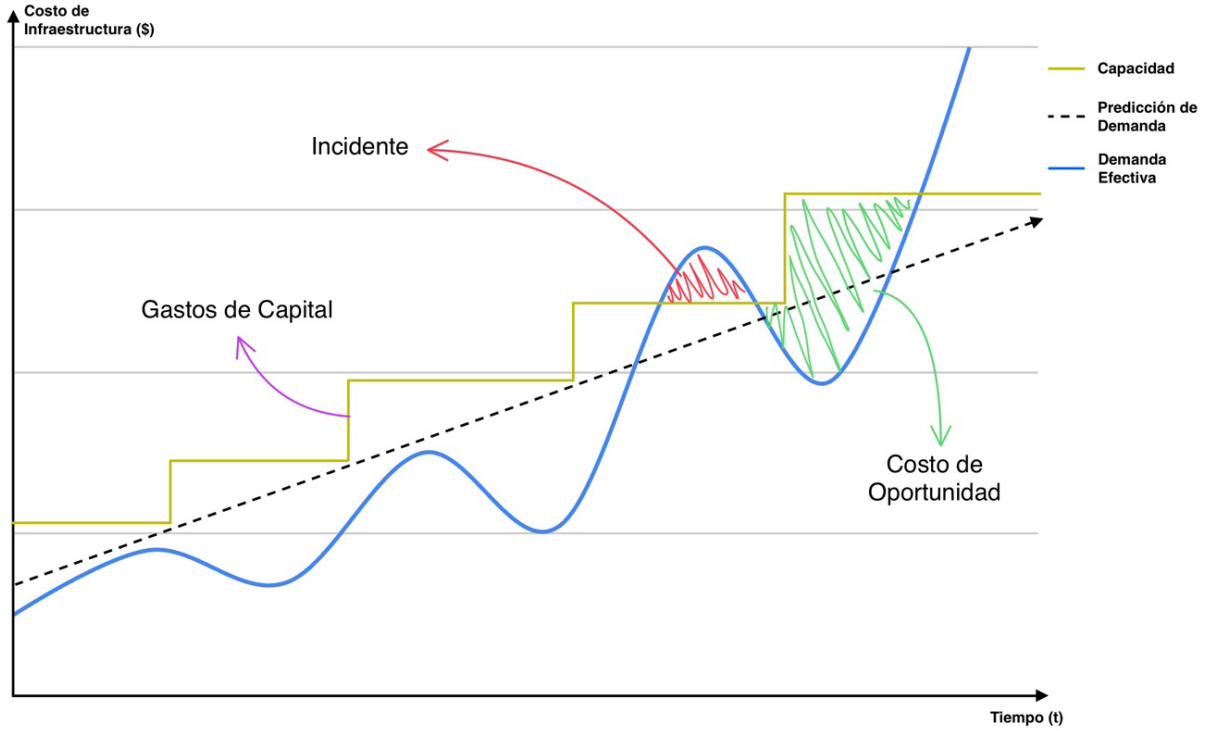


Ilustración 5 - Demanda vs Capacidad On-Prem (Fuente: Elaboración propia)

Basados en lo anterior, se podría concluir que la mejor estrategia es contar con la capacidad justa y necesaria para atender la demanda, logrando que la curva de capacidad (en amarillo) “imite” a la perfección la curva de la demanda efectiva (en azul). Sin embargo, esto es solo posible en la teoría, y resulta impracticable en un caso real. La demanda, en la gran mayoría de los casos es altamente variable y difícil de predecir con exactitud. Esto se debe a que la demanda usualmente depende de innumerables factores, como por ejemplo: la estacionalidad del negocio, el éxito de

un producto o servicio, el impacto de una campaña de *marketing*, la necesidad operativa de realizar una operación intensiva, etc.

Por otro lado, habrá que tener en consideración otros factores que imposibilitan tener una capacidad flexible en un centro de cómputos físico. En primer lugar, resulta altamente ineficiente reducir la capacidad: generalmente, el equipo de *hardware* se compra, o se alquila por períodos de tiempo extenso (meses o años). Es por esto que una vez adquirido el equipamiento tecnológico, resulta altamente ineficiente deshacerse del mismo. Las ventas de equipamiento usado a escala, y a nivel empresarial, es una práctica extremadamente infrecuente, y en muchos casos resulta incluso impracticable ante la imposibilidad legal de vender activos, el riesgo de seguridad que ello implica, o la imposibilidad de transferir los activos junto con la garantía o las licencias de uso del fabricante de los mismos (Suwan, 2024).

En segundo lugar, hay que considerar que ampliar la capacidad instalada de un centro de cómputos es una tarea que requiere de extensa planificación, mano de obra intensiva, y son usualmente proyectos que llevan semanas o incluso meses, dependiendo del volumen o las obras de infraestructura necesarias. Realizar una ampliación implica planificar y dimensionar la misma, seleccionar el equipamiento que cumpla con las especificaciones requeridas, encargar dicho equipamiento al fabricante, aguardar que dicho equipamiento arribe al centro de cómputo, instalar el nuevo *hardware*, configurar y probar el mismo, y solamente luego, poner en producción y hacer efectiva la expansión.

Es por lo anterior, que las expansiones de capacidad no son operaciones que se realicen a diario, resultando en la imposibilidad de ajustar la curva de capacidad para “seguir” a la curva de demanda efectiva, e incluso, de predicción de demanda.

Por último, tener una utilización de recursos muy cercana a la capacidad máxima instalada aumenta el riesgo de que la demanda supere a la capacidad. Como se explicó con anterioridad, en la mayoría de los casos resulta imposible predecir la demanda con extrema exactitud, por lo que si la demanda llegara a superar a la capacidad del centro de cómputos, significa que durante ese periodo, no se contarán con los recursos necesarios para atender las necesidades del negocio. Usualmente,

durante estos eventos ocurren una de dos cosas: 1) el negocio sufre un incidente o, 2) el negocio pierde una oportunidad. La diferencia de ambos consistirá en entender si la falta de capacidad resulta en limitar la operación propia o de los clientes servidos. Para los casos donde la falta de capacidad tenga un impacto en la operación, se catalogará como un incidente.

Para ilustrar lo anterior, se puede pensar en un centro de cómputos destinado a gestionar documentos de usuarios (por ejemplo, Google Drive). Si en algún momento dicho centro de cómputo tuviera un faltante de capacidad de almacenamiento, el servicio se vería significativamente afectado, ya que los usuarios no podrían crear, copiar o modificar sus documentos existentes. Esto resultaría en un incidente, donde el usuario estaría impedido de hacer uso del servicio contratado.

Por otro lado, existe la posibilidad de que la falta de capacidad implique un problema para el negocio, sin interrumpir la operación, y en este caso no sería considerado un incidente. Esto ocurre cuando el negocio solicita recursos adicionales, que en este caso no se encuentran disponibles, para realizar una operación nueva o de ampliación. Un ejemplo de lo anterior seria adquirir nuevos clientes, o vender servicios adicionales a usuarios existentes. En estos casos, la falta de capacidad resultará una limitante para expandir el negocio, resultando en un cuello de botella en el camino crítico hacia el éxito del mismo (Espin, 2024).

4.2. Gestionando la Capacidad en la Nube

La gestión de la capacidad en la nube es significativamente más simple que en la modalidad On-Prem, dado que la misma es responsabilidad del proveedor de nube, quien deberá asegurarse que sus centros de cómputo siempre cuenten con la capacidad suficiente de hardware, para proveer los recursos requeridos por sus clientes.

La gestión de la capacidad en un servicio de nube pública se limitará a asegurarse que los sistemas y aplicaciones propios soliciten en tiempo y forma los recursos necesarios al proveedor: esto significa, identificar cuándo se requiere recursos adicionales, y enviar la petición de estos al proveedor de nube (Menascé, 2009).

Deberá tenerse en cuenta que la provisión de recursos nuevos no es instantánea, sino que, dependiendo del tipo de recurso, demorará desde algunos segundos, hasta varios minutos, por lo que se deberá contar con capacidad ociosa suficiente, en todo momento, para poder atender la demanda actual, hasta tanto el proveedor disponibilice los recursos solicitados.

En la Ilustración 6 (Demanda vs Capacidad Cloud) se muestra en el área verde la capacidad ociosa de recursos de nube pública, requerida para evitar incidentes.

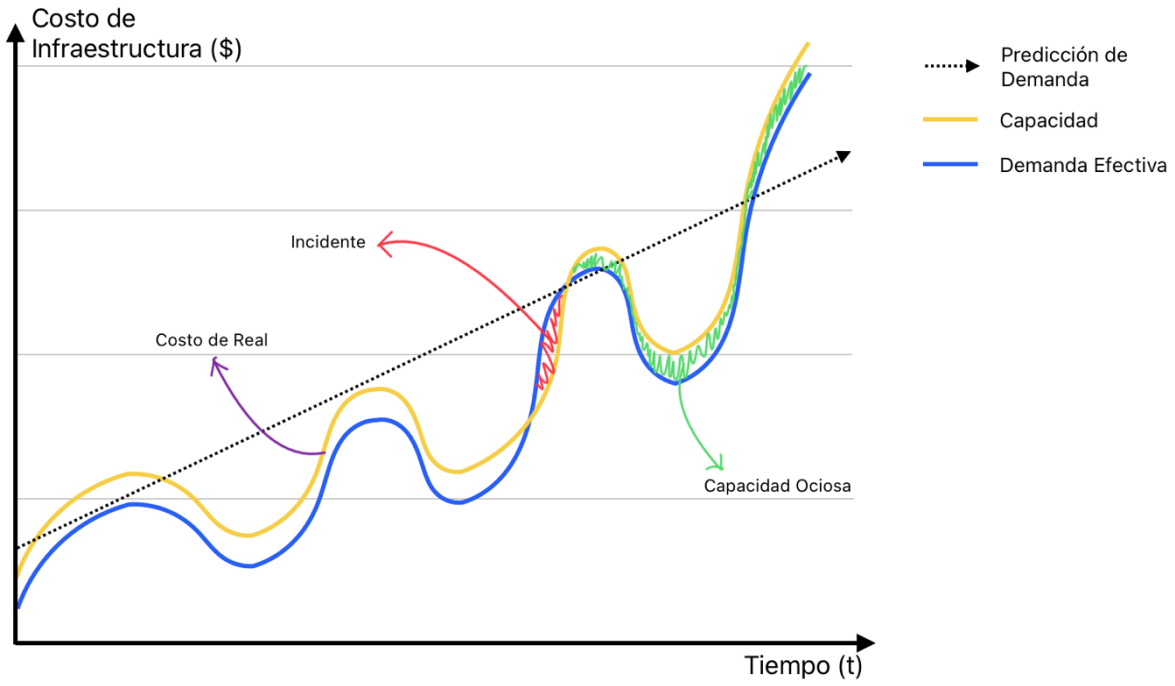


Ilustración 6 - Demanda vs Capacidad Cloud (Fuente: Elaboración propia)

De la misma forma, cada vez que la capacidad adicional ya no sea necesaria, deberá "devolverse" al proveedor, evitando de esta forma que dichos recursos sigan siendo cobrados al cliente.

Resulta evidente que la flexibilidad para demandar y devolver recursos en un ambiente de nube pública permite minimizar la capacidad ociosa significativamente cuando se lo compara con el esquema tradicional de infraestructura On-Prem (y, por ende, minimizar los costos fijos). Dicha capacidad ociosa puede minimizarse hasta un

punto que el negocio considere óptimo entre el costo adicional abonado y la probabilidad del riesgo de quedarse sin recursos en un momento de crecimiento rápido en la demanda.

Por último, vale aclarar que si bien el riesgo es bajo (especialmente para consumo bajo demanda), puede ocurrir que el proveedor de servicios no cuente con recursos adicionales para disponibilizar, debido a la alta demanda consolidada de todos sus clientes en una zona específica. Para mitigar estos riesgos, el cliente puede optar por reservar recursos de forma anticipada, lo que garantizará la demanda futura, pero incrementará el costo unitario de provisionar dichos recursos (AWS, Amazon Elastic Compute Cloud User Guide, 2024).

CUERPO EMPÍRICO

El cuerpo empírico ha sido del tipo descriptivo, apuntando a exponer hechos, herramientas y procesos aplicables a la mayoría de las empresas que se hayan encontrado, desde considerando migrar hacia la nube, hasta comenzando a lograr la madurez con su operación en la misma.

Esta sección incluyó los diferentes capítulos que describen el trabajo de campo realizado, y las conclusiones y lecciones que se desprenden del mismo. En la misma, se intentó proporcionar al lector con herramientas y una guía conceptual sobre aquellos factores que se detectó que tuvieron mayor influencia a la hora de determinar una adopción de Nube pública eficiente y efectiva.

Se comenzó por describir la operación y el contexto de Wildlife Studios, empresa en la cual fueron explorados e implementados los mecanismos descriptos en el presente trabajo; con la firme convicción de que los resultados podrán ser extrapolados hacia otras empresas comparables, incluso en industrias diferentes. Luego, se describieron aquellas métricas e indicadores que resultaron altamente significativas a la hora de evaluar y medir los resultados de aquellas iniciativas que fueron ejecutadas con el objetivo de mejorar la eficiencia en la utilización de recursos de nube. Por último, se describió los mecanismos que fueron identificados como aquellos con mayor impacto a la hora de lograr resultados positivos, alineados con el objetivo perseguido de reducir los costos *cloud*, sin comprometer la operación del negocio.

CAPÍTULO V: METODOLOGÍA DE INVESTIGACIÓN

En el presente capítulo se describió la metodología de investigación utilizada para el análisis de la utilización de *Cloud Computing*, y los mecanismos de eficiencia disponibles para asegurar una adopción sustentable de la misma. El objetivo buscado fue proveer una visión clara y estructurada del proceso ejecutado durante la investigación, asegurando su validez, confiabilidad y rigor de las conclusiones derivadas de la misma.

5.1. Diseño de la Investigación

Para lograr una comprensión integral y suficiente de las estrategias de adopción y configuración de la utilización de recursos de nube pública, se eligió un enfoque descriptivo, con un análisis que combinó recursos cuantitativos y cualitativos. Este diseño permitió obtener el conocimiento requerido para habilitar un análisis profundo y abarcativo del tema en cuestión, proveyendo tanto información cuantitativa para identificar patrones y correlaciones, como data cualitativa, para comprender y analizar en profundidad las razones, requerimientos y desafíos del camino hacia la adopción de tecnologías de nube.

Para el **análisis cuantitativo**, se optó por realizar una encuesta anónima (Ver Anexo I) a directivos y gerentes de una amplia gama de empresas tecnológicas que están actualmente utilizando proveedores de nube, con distintos niveles de adopción; o que están considerando hacerlo en el corto plazo. Este mecanismo permitió obtener una cantidad considerable de respuestas, que permitió identificar patrones y/o tendencias comunes significativos. El mecanismo de recolección de información consistió en un cuestionario estructurado, distribuido de forma *online* mediante *Google Forms*. La encuesta incluyó una mayoría de preguntas cerradas para medir variables como el grado de adopción de nube pública, el grado de madurez en la misma, y la efectividad percibida en los esfuerzos de eficientización del consumo *cloud*. La encuesta fue distribuida principalmente a líderes, gerentes y directivos de IT, CTOs, CIOs y profesionales con influencia para tomar decisiones tecnológicas en empresas chicas, medianas y grandes.

En cuanto al **análisis cualitativo**, se realizaron entrevistas (Ver Anexo II) a profesionales expertos y experimentados en infraestructura y tecnologías de nube, con extensas trayectorias comprobables en el área de tecnología e infraestructura *cloud*. Estas personas, con roles clave dentro de sus organizaciones, y alta capacidad de influencia para decidir sobre cuestiones de tecnología, brindaron información precisa y profunda acerca de los problemas que enfrentaron en sus distintas experiencias a la hora de utilizar recursos de nube, los mecanismos y procesos utilizados para encontrar soluciones pragmáticas a esos desafíos, y finalmente, aquellas dificultades y riesgos que aún tienen por resolver o mitigar. Adicionalmente, se investigó sobre las nuevas capacidades que habilitó la adopción de nube pública en sus organizaciones, y sus expectativas de cara al futuro respecto de estos servicios. Este enfoque permitió profundizar sobre aspectos cualitativos y de contexto, y complementar de manera satisfactoria la información más relevante recabada durante el análisis cuantitativo.

El análisis cualitativo consistió en unas pocas entrevistas con preguntas semi-estructuradas, en su mayoría abiertas. Como foco principal, se recabó información referida a la utilización de nube pública, los procesos de toma de decisiones respecto de su utilización en distintos casos de uso, la implementación de estrategias de adopción y/o migración, los mecanismos de eficiencia utilizados, los desafíos enfrentados en el proceso, y aquellos que aún quedan por delante. Todas las entrevistas fueron realizadas de forma asincrónica.

Adicionalmente a los análisis cuantitativo y cualitativo, el presente cuerpo se apalancó significativamente sobre reportes y publicaciones disponibles en línea, y publicaciones de los principales proveedores de nube. Asimismo, se utilizó extensivamente la experiencia personal del autor, quien se ha desempeñado en roles técnicos y de gestión en la industria de tecnología. Por más de 15 años, el autor ha trabajado en distintas empresas de diversos tamaños, incluyendo MercadoLibre y Wildlife Studios, las cuales utilizan Nube Pública de forma extensiva. Es de especial interés la experiencia armando, liderando y desarrollando equipos de *Cloud Economics* (o *FinOps*), los cuales tuvieron como principal objetivo diseñar, desarrollar e implementar muchas de las iniciativas expuestas y descriptas en el presente trabajo.

5.2. Consideraciones Éticas

Para cumplir con un estándar ético suficiente, toda la investigación fue realizada respetando los principios de confidencialidad e integridad. Los participantes fueron informados acerca del propósito de los cuestionarios y las entrevistas, y la naturaleza de su participación. Asimismo, la información sensible recolectada fue tratada como confidencial, y siempre fue almacenada de forma segura.

CAPÍTULO VI: LA NUBE PÚBLICA EN LA INDUSTRIA DE LA TECNOLOGÍA

En el presente capítulo se describieron los principales resultados y conclusiones obtenidos a partir de las encuestas (Ver Anexo I) y las entrevistas (Ver Anexo II), realizadas como parte de la investigación.

6.1. La Industria

El 100% de los participantes de la encuesta resultó trabajar en empresas que ya han adoptado la nube pública como un servicio productivo dentro de sus operaciones. Esta cifra resultó esperable, teniendo en cuenta que es realmente difícil encontrar empresas que no hayan experimentado con la nube pública, migrando procesos y aplicaciones desacopladas, que resultan fácilmente trasladables a estos ambientes. Esto se evidenció a través del hecho que menos de la mitad de los encuestados (45%) trabaja en empresas con operaciones 100% basadas en cloud pública, aquellas que usualmente corresponden a lo que se conoce como empresas *cloud-natives*. Como contraparte, más de la mitad de los encuestados (55%) admitió trabajar con modelos híbridos, apalancados sobre nube pública, pero también sobre infraestructura tradicional, con *hardware* propio alojado en d*atacenters* convencionales.

La mayoría de los encuestados (87%) refirió una utilización de recursos de nube pública madura (madurez “Definida”, o superior), lo que se corresponde con el porcentaje de empresas que vienen utilizando este tipo de servicios hace más de 3 años (81%).

Como es de esperarse, el nivel de madurez alcanzada tiene una correlación directa con la antigüedad en el uso de servicios cloud: a mayor antigüedad, mayor madurez alcanzada, significando que aquellos usuarios con más tiempo y experiencia en la utilización de servicios cloud, lograron una adopción más medida, controlada, optimizada y eficiente, que aquellos con menor experiencia.

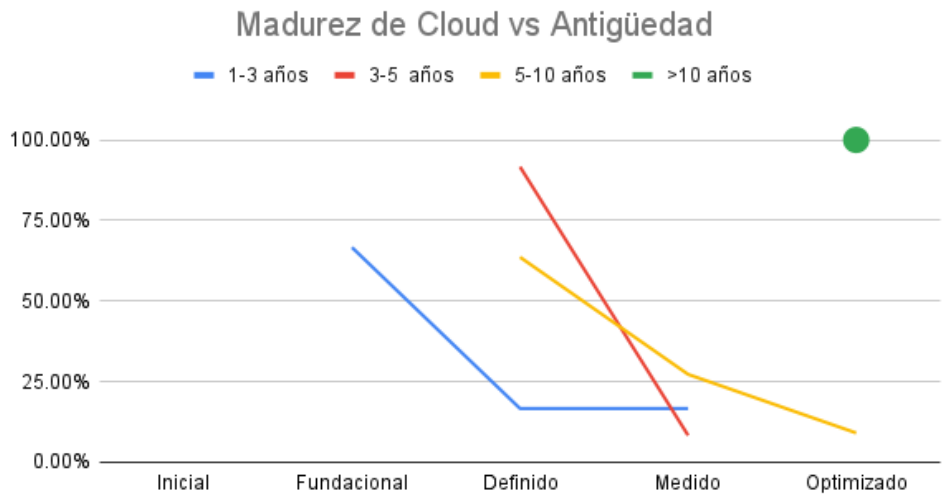


Ilustración 7 - Madurez en la adopción de Cloud vs Antigüedad de Uso (Fuente: Elaboración propia)

El *market share* de los principales jugadores (AWS, GCP y Azure) es significativamente superior a la estadística global citada en el Capítulo III, resultando en un valor cercano a 89%. Esta diferencia se interpreta a partir del hecho que todos los encuestados trabajan en empresas de origen occidental, con fuerte mayoría en empresas latinoamericanas. Por motivos de geografía, los *vendors* de nube pública asiáticos (Huawei, Ali Cloud, Tencent, entre otros) resultan poco elegidos por empresas que tienen su base de clientes fuera de Asia (especialmente, China).

AWS resulta un claro ganador entre las empresas encuestadas, con una utilización del 77%. Sin embargo, el segundo lugar lo ocupa GCP (58%), en lugar de Azure (39%). La principal hipótesis para explicar esta diferencia resultó ser el hecho de que GCP tiene una base de clientes más grande que Azure, pero donde la mayoría de los mismos corresponden a clientes chicos (*startups*) con bajo gasto en servicios de nube, mientras que gran parte de los clientes de Azure corresponde al segmento corporativo, con gastos promedio significativamente superiores (HGInsights, 2024).

Sin embargo, a partir de los datos de la encuesta, se observó que AWS, y en menor medida Azure, son más adoptados por aquellos usuarios que comenzaron a utilizar la nube hace mayor tiempo. Por el contrario, la nube de Google (GCP) es la más adoptada por aquellos usuarios con menos de 5 años de antigüedad. Por últimos, se observó que los proveedores de nube “secundarios” fueron adoptados, en su

mayoría, por aquellos usuarios con más de 5 años de experiencia en la utilización de servicios de nube pública.

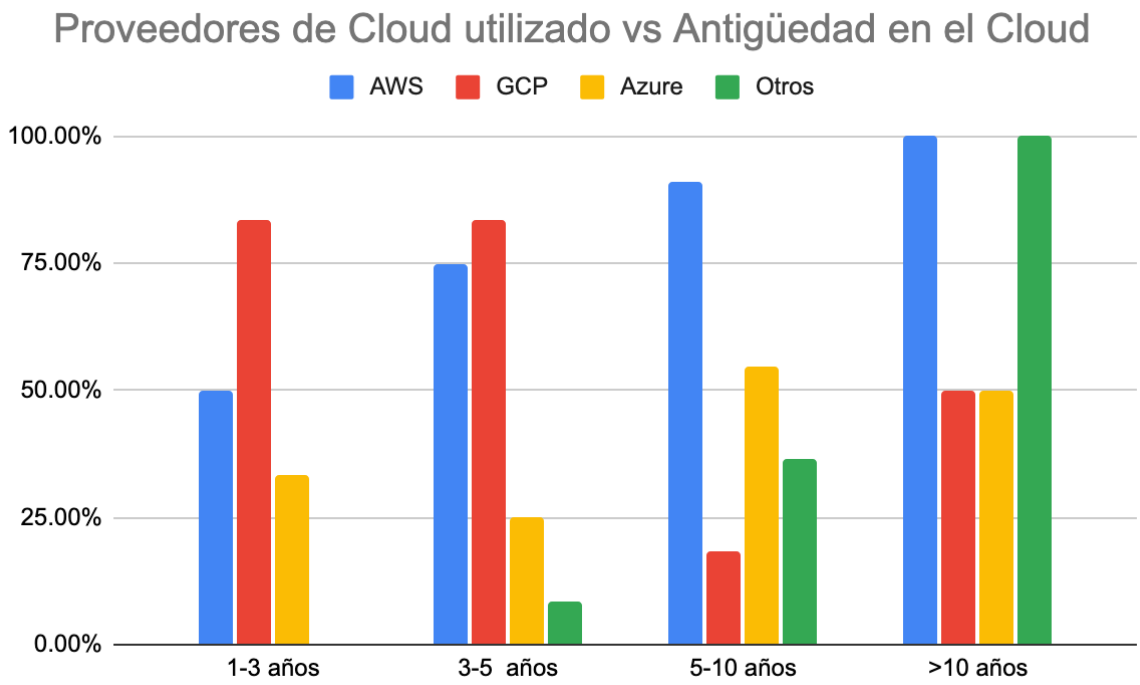


Ilustración 8 - Proveedores de Cloud Utilizados vs Antigüedad en el Cloud (Fuente: Elaboración propia)

Resultó interesante descubrir que el 32% de los encuestados utilizan un único proveedor de cloud, mientras que el 39% utiliza dos proveedores. El 29% restante utiliza tres o más proveedores diferentes de nube pública. Lamentablemente, no se contó con información suficiente para inferir si aquellas empresas con ambientes *multi-cloud* (que operan con más de un proveedor) lo hacen de forma balanceada, distribuyendo sus aplicaciones de forma pareja entre los distintos proveedores, o si por el contrario, concentran la mayoría de sus aplicaciones en un solo jugador, y utilizan otros proveedores para casos de nicho.

Sí se observó una correlación directa entre cantidad promedio de proveedores utilizados y la antigüedad en el uso de Cloud: a mayor tiempo utilizando servicios de nube pública, mayor la cantidad de proveedores utilizados. Se explica dicha correlación a partir del incremento en la madurez de la adopción: en la medida que los usuarios ganan madurez, deciden utilizar distintos proveedores para aprovechar las ventajas propias de cada uno.

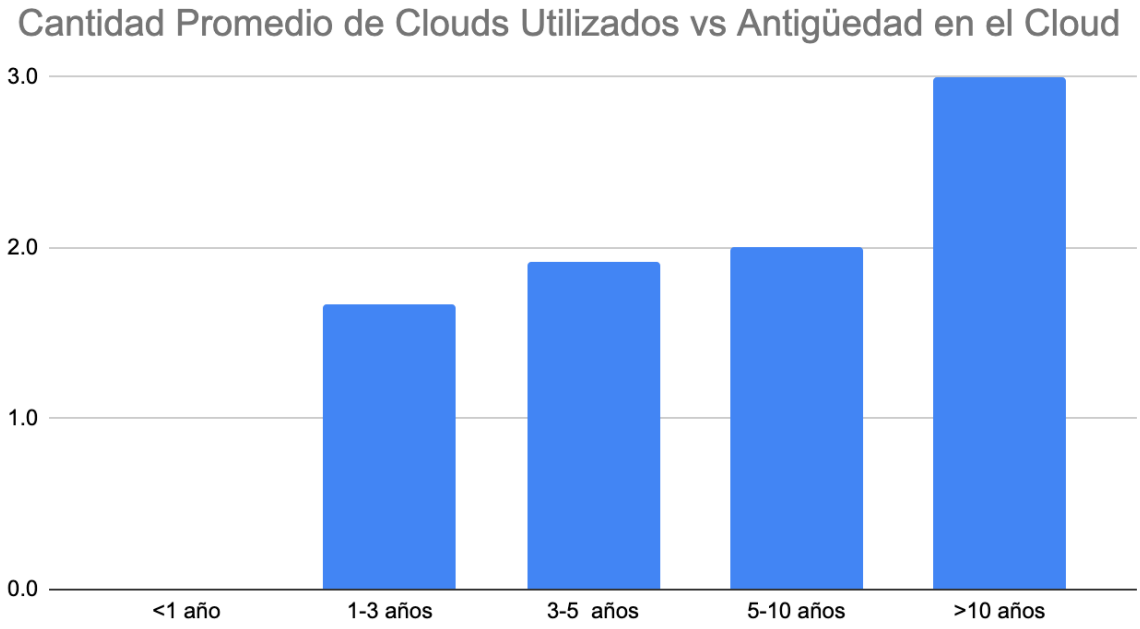


Ilustración 9 - Cantidad Promedio de Clouds Utilizados vs Antigüedad en el Cloud (Fuente: Elaboración Propia)

En relación con lo anterior, 87% de los encuestados reportó que sus empresas tienen consumos anuales de nube pública significativos (mayores a USD 100.000), con un 42% que consumió más de USD 500.000 en el último año, y un 19% con consumos anuales superiores al millón de dólares. Estos datos resultaron particularmente de interés cuando se los cruzó con los resultados de la pregunta sobre contratos de largo plazo (pregunta 5): resultó evidente que un gran número de empresas (39%) opera en la nube pública sin contratos. Esto significa que no poseen ningún compromiso, y por ende, mantienen un nivel de riesgo bajo. Sin embargo, al no poseer contratos de mediano y largo plazo, estas empresas no acceden a potenciales descuentos comerciales que pueden influir positiva y significativamente sobre el gasto total de *cloud*, con descuentos comerciales para nada despreciables.

Se observó una correlación inversamente proporcional entre la existencia de contratos y el gasto total anual en servicios cloud. Esto se explica teniendo en cuenta que los proveedores de nube poseen valores mínimos de gasto anual para permitir la firma de acuerdos comerciales específicos (Tony Chan, 2024), y adicionalmente en el hecho de que los ahorros potenciales derivados de un contrato no justifican el riesgo de los compromisos de gasto mínimos, para aquellos casos que el costo total de nube pública sea considerado “bajo” (menor a USD 500.000 / año).

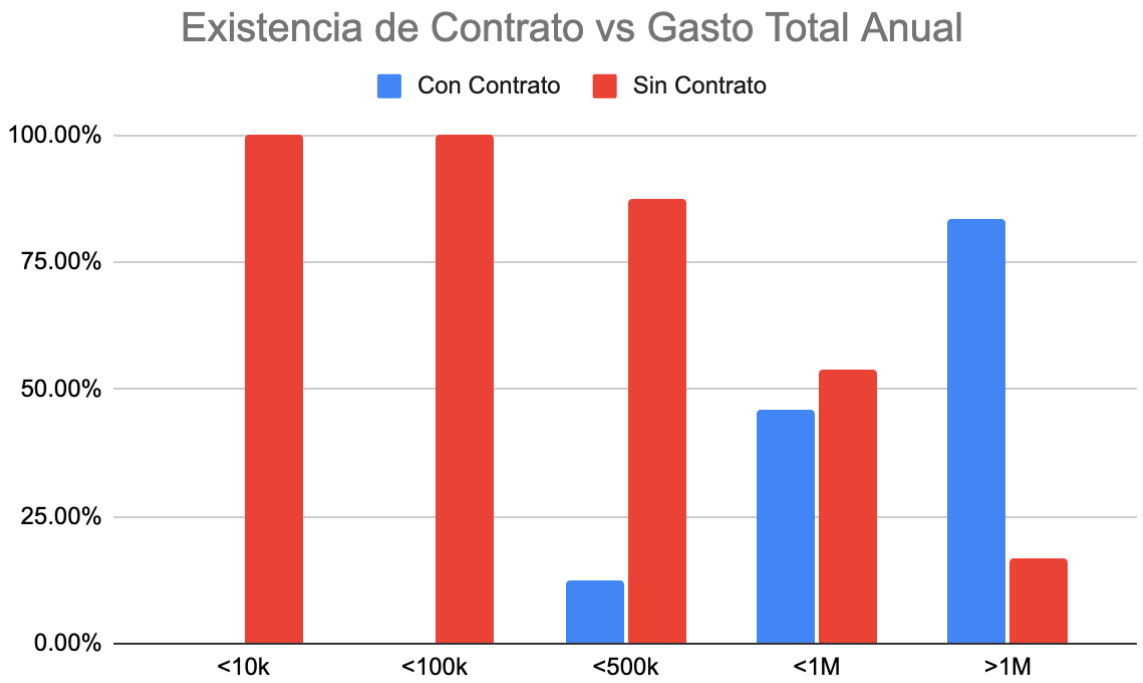


Ilustración 10 - Existencia de Contrato vs Gasto Total Anual (Fuente: Elaboración Propia)

No sorprendió descubrir que un 67% de las personas encuestadas demostró niveles de preocupación alto (“Preocupado”, o superior) respecto de los costos efectivos en los que incurrieron sus organizaciones, asociados al consumo de servicios de nube pública. Consecuentemente, esta preocupación se ve reflejada bajo el hecho que el 81% de las empresas tienen iniciativas activas para optimizar estos mismos costos.

Sin embargo, se concluyó que estas iniciativas de optimización demostraron tener poco foco e inversión organizacional, dado que solo el 45% de las empresas dedicó recursos exclusivos a abordar el problema, lo que demuestra la poca madurez que aún tienen estas prácticas en el mercado actual. En contraposición, solamente el 19% de los encuestados reportó tener equipos significativos (cinco personas, o más) dedicados a esta tarea.

Por otro lado, el nivel observado de preocupación por los costos incurridos, a diferencia de lo esperado, no se incrementa con el nivel de gasto total. No se encontró ninguna correlación evidente ni significativa, entre el gasto total y el nivel de preocupación. Esto pudo deberse a varias razones, incluyendo i) a medida que los costos se incrementan, el nivel de esfuerzo por optimizarlos también lo hace, lo que

disminuye el nivel de preocupación percibido; o ii) el gasto total debe analizarse en relación con el tamaño de la empresa, en niveles relativos, para entender que tan significativos son los costos de cloud en relación con sus costos totales de operación[2].

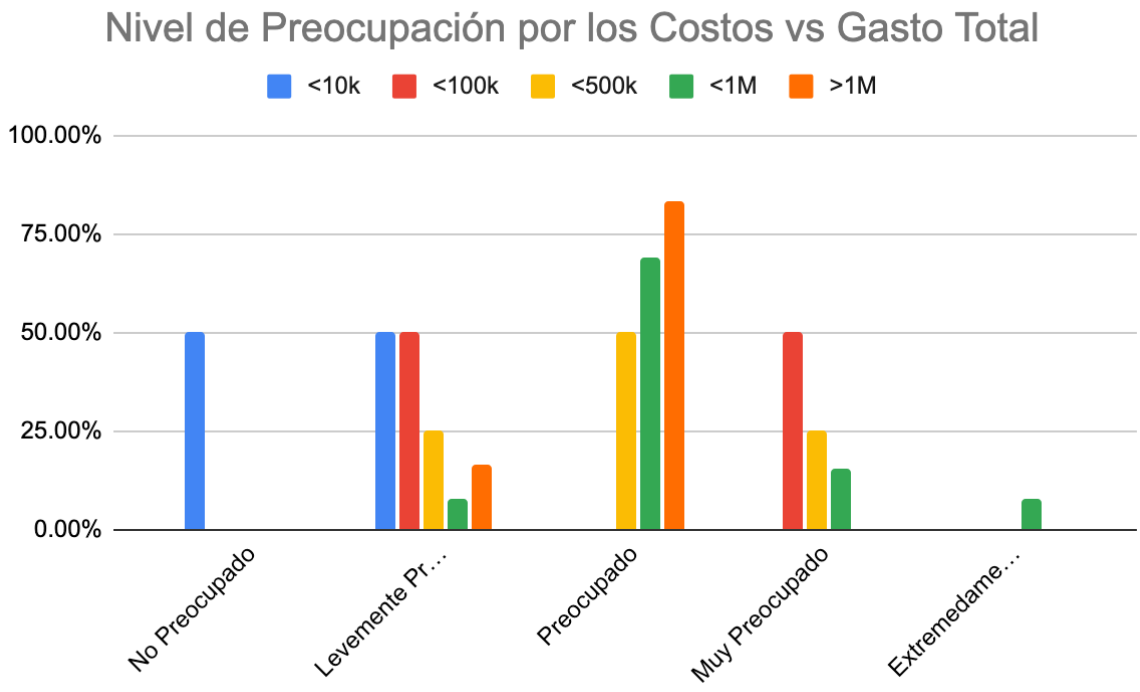


Ilustración 11 - Nivel de Preocupación por los Costos vs Gasto Total (Fuente: Elaboración Propia)

En línea con el bajo nivel de compromiso mostrado por las respuestas a la pregunta apuntada a dimensionar el foco organizacional en iniciativas de optimización (pregunta 9), se evidenciaron niveles de eficacia percibida correlacionados a este último. El 61% de los encuestados admitieron niveles de eficacia considerados bajos ("Algo efectivo", o inferior). Esto demuestra la gran oportunidad que aún resta por capturar en la mayoría de las empresas que deciden operar utilizando servicios *cloud*.

Cabe destacar que el 77% de los encuestados refieren ahorros de entre 10% y 50% respecto del consumo de nube, gracias a los esfuerzos de optimización realizados dentro de sus empresas; pero, sin embargo, los mismos son percibidos como insuficientes o escasos. Por ende, se observó que gran parte de los encuestados es perfectamente consciente del enorme potencial que tienen los esfuerzos de *FinOps*,

[2] No se incluyó esta pregunta en la encuesta, debido a que hubiera comprometido la confidencialidad de la misma.

a la hora de optimizar y reducir el gasto en nube pública, sin comprometer las capacidades del negocio.

Surgió como evidencia de los anterior, cruzar los datos de ahorro percibido contra el tamaño de los equipos dedicados a la optimización de los mismos (FinOps). Cuanto más grandes resultan los equipos de *Cloud Economics*, mayor el porcentaje de ahorros percibido en comparación con la falta total de iniciativas de eficiencia.

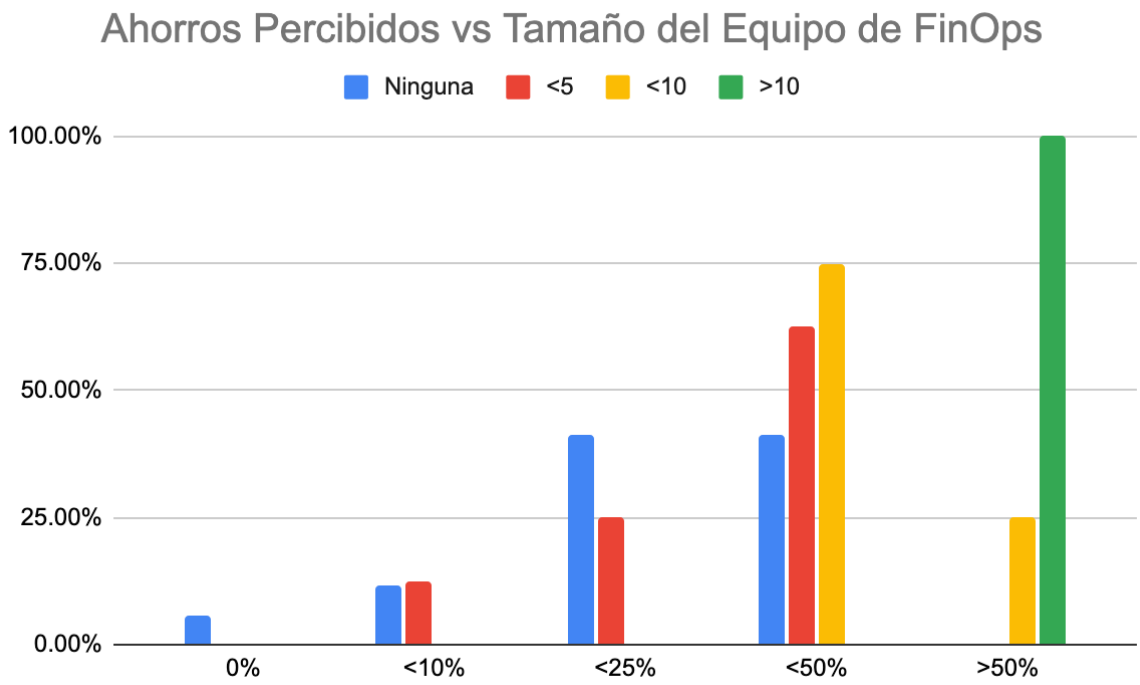


Ilustración 12 - Ahorros Percibidos vs Tamaño del Equipo de FinOps (Fuente: Elaboración Propia)

Por último, resultó interesante entender que la gran mayoría de las empresas identifica a la nube pública como un diferencial de valor, con el 84% de las mismas convencidas que los servicios de cloud son valiosos, y volver al esquema tradicional de infraestructura no es una opción que esté sobre la mesa, al momento de realizada la encuesta. Del 17% restante, la absoluta mayoría plantea un escenario donde potencialmente considerarían utilizar más infraestructura tradicional en el futuro, probablemente asociado al hecho de que con más volumen y más consumo, contar con un centro de cómputos propio se vuelve cada vez más atractivo desde un punto de vista de eficiencia financiero.

6.2. La Opinión de los Expertos

Los expertos entrevistados coinciden al destacar las múltiples ventajas de los servicios de nube pública, en comparación con el modelo tradicional. Los pros más destacados siendo i) la velocidad y la reducción del *time-to-market*; ii) la flexibilidad para provisionar y de-provisionar recursos rápidamente; y iii) las bajas barreras de entrada, debido a su simplicidad y bajo *overhead* operacional.

Todos coinciden en que la nube pública es la opción ideal para empresas y emprendimientos chicos y medianos que prioricen la agilidad, y los bajos costos operativos. Por otro lado, la infraestructura tradicional suele ser una opción más adecuada para aquellas empresas que operen bajo regulaciones muy estrictas, o que necesiten mantener un control muy granular del *hardware* que utilizan para montar y alojar sus servicios tecnológicos; siendo banca tradicional un ejemplo que surgió de forma repetida.

Si bien se identifica al *cloud* como un servicio con bajas barreras de entrada, se reconocen algunos factores comunes clave a la hora de incrementar las probabilidades de una adopción exitosa: contar con una estrategia de adopción/migración adecuada, que garantice el conocimiento y la utilización de buenas prácticas; y la correcta adecuación cultural, que facilite el cambio y garantice la confianza organizacional en el nuevo paradigma. La importancia de una estrategia adecuada es fuertemente reconocida como un factor clave, también en el ámbito global (Orban, 2017).

Todos los especialistas coinciden en la predicción acerca de que los servicios de nube pública seguirán ganando terreno sobre el modelo de infraestructura tradicional, el cual quedará relegado para dos casos principales: a) los sectores obligados por regulaciones particulares; y b) los grandes jugadores que consigan el volumen necesario, llegando al punto donde las ventajas de contar con datacenters propios haga sentido financiero.

La correcta gestión de los costos, y la adecuada optimización de los recursos resulta un factor recurrente entre los expertos a la hora de identificar dificultades y riesgos en

la adopción de nube. Dicha práctica se reconoce como un factor crucial y mandatorio a la hora de asegurar la sostenibilidad de los servicios de nube en el largo plazo, una vez alcanzados volúmenes operacionales significativos. Adicionalmente, se mencionan como riesgos la falta de mecanismos para contar con una correcta gobernabilidad de los recursos, y el *vendor lock-in* en el largo plazo.

Para garantizar una implementación eficiente en el *cloud*, los expertos recomiendan contar con equipos dedicados a la optimización y al aseguramiento de buenas prácticas. Dichos equipos deberán ser los responsables por la implementación de mecanismos de optimizaciones centralizados, y por la generación de visibilidad acerca de los costos incurridos por el negocio respecto de la utilización de recursos tecnológicos de nube pública; lo que permitirá a las distintas áreas tomar decisiones apalancadas sobre la viabilidad financiera de los productos y funcionalidades desarrollados.

Al igual que en las encuestas, las entrevistas a expertos arrojaron algunas de las principales herramientas a la hora de mantener costos optimizados dentro de los distintos proveedores de *cloud*: recursos reservados, utilización de recursos efímeros, negociación de contratos, presupuestos, métricas en tiempo real, generación de una cultura de *accountability*, entre muchas otras. Todas estas herramientas, y más, se analizaron en capítulos posteriores dentro de la presente tesis.

Por último, todos los especialistas identifican a la práctica de *FinOps*, como una disciplina que, si bien no es necesaria ni conveniente en etapas tempranas para empresas con gastos bajos, resulta absolutamente fundamental una vez que los costos de operación de tecnología alcanzan volúmenes significativos, entregando retornos considerablemente más altos que la inversión dedicada a estos esfuerzos.

CAPÍTULO VII: WILDLIFE STUDIOS Y LA INDUSTRIA DE LOS JUEGOS MOVILES

El presente capítulo refiere principalmente a la industria de los videojuegos para dispositivos móviles, entendiéndose como tal los celulares y *tablets*, con especial énfasis en aquellos que usan los sistemas operativos iOS (Apple) y Android (Google), dado que representan un *market-share* superior al 99% de la totalidad (statcounter, 2024). Quedaron excluidos del presente análisis la industria de los videojuegos para PC y consolas (tanto de escritorio, como portátiles).

7.1. La Industria de los Juegos Móviles

La industria de los juegos móviles sufrió una profunda transformación, desde sus comienzos a la actualidad, impulsada por avances tecnológicos, cambios en el modelo de negocios, la masificación de internet en los celulares, y los cambios en la demanda de los usuarios.

Los primeros videojuegos para dispositivos móviles surgieron en la década de 1990, con la aparición de celulares con pantallas LCD monocromáticas. Estos juegos venían pre-instalados en los celulares, y en la mayoría de los casos eran desarrollados por la compañía fabricante del celular, o licenciados por ellos para ser adaptados (*porteados*) para sus dispositivos. Estos juegos, de naturaleza sencilla y para un solo jugador, incluyen juegos clásicos como el *Tetris* (Wikipedia, Tetris, 2024) o el *Snake* (Wikipedia, Snake (1998 video game), 2024).

Todo cambió radicalmente en el año 2008, con la llegada de las tiendas de aplicaciones: Apple Store (para iOS) y Google Play (para Android). Estas tiendas facilitaron enormemente la distribución de aplicaciones móviles (incluyendo juegos, entre otros), para cualquier persona o empresa que deseara desarrollar un juego para estos sistemas. El modelo de negocio de estas tiendas ha sido relativamente sencillo: Las *stores* son desarrolladas y mantenidas por los desarrolladores del sistema operativo (Apple en iOS, Google en el caso de Android), y forman parte (y vienen instaladas por defecto) de todos los dispositivos que usen estos sistemas operativos. Cualquier desarrollador puede crear y subir su(s) aplicación(es) a estas tiendas de

forma muy sencilla, como un medio/plataforma para distribuirlas hacia los usuarios finales. El creador de estas aplicaciones es quien define el precio de las mismas (y de las compras adicionales que pueden realizarse *in-app*), desde $0, hasta los cientos o miles de dólares.

Los usuarios, por otro lado, pueden ingresar libremente a las tiendas de sus teléfonos, explorar el catálogo de aplicaciones disponibles, y bajar e instalar la(s) aplicación(es) que desee. En el caso de que las aplicaciones sean pagas, el usuario deberá abonar (a Google y/o Apple) el precio correspondiente antes de poder proceder con la descarga.

Los desarrolladores de las tiendas (Apple y Google) proveen el mecanismo de distribución, las herramientas de desarrollo e integración, procesan los pagos, y garantizan la seguridad del sistema. Generalmente, cobran una tarifa del 30% de todos los pagos procesados por sus plataformas (Ling, 2021).

Con la aparición de estas tiendas, cualquier persona con los medios y conocimientos para desarrollar una aplicación móvil, se encontró habilitada para publicar y comercializar su aplicación en cuestión de horas. Esta tecnología revolucionó el mercado de los juegos móviles, y dio lugar a la aparición de incontables nuevos actores, en un mercado que vio crecimientos exponenciales por los siguientes años.

7.2. El Modelo de Negocio

Si bien se encuentran algunas excepciones, generalmente existe una motivación económica detrás de cada juego móvil: sus creadores buscan generar dinero a partir de sus productos. Existen diversos mecanismos que los creadores o estudios de juegos utilizan para monetizar sus aplicaciones, siendo las siguientes las más comunes en la industria:

- **Juegos gratuitos con publicidad:** Son juegos completamente gratis de jugar, pero que muestran anuncios de forma compulsiva (el jugador debe mirar el anuncio para poder jugar), u optativa (el jugador elige mirar un anuncio para obtener una recompensa dentro del juego). Los creadores de este tipo de

juegos ofrecen espacio dentro sus juegos para que terceros puedan publicitar sus propios servicios o productos, y así obtener nuevos clientes. La forma más utilizada para lograr esto es utilizando servicios de mediación de publicidad, donde el juego se integra con un servicio que busca, provee y despliega los anuncios, y paga un porcentaje de los ingresos generados por ese anuncio (IronSource from Unity, 2021).

- **Juegos "*freemium*":** Son juegos gratis de jugar, pero ofrecen capacidades dentro del juego que deben ser adquiridas con dinero real. Ejemplos muy comunes de esta práctica, son aquellos juegos que ofrecen al jugador la posibilidad de jugar y disfrutar del juego sin pagar, pero venden artículos cosméticos (por ejemplo), que amplían la experiencia del jugador, y que solo pueden ser desbloqueados mediante compras con dinero.

- **Juegos de apuestas:** Como su nombre lo indica, son juegos donde el jugador apuesta dinero real, con la posibilidad de ganar (o perder) dinero. Estos juegos emulan la experiencia de un casino físico y generalmente ofrecen juegos tradicionales como ruleta, blackjack, póker, o *slot machines*.

- **Juegos pagos:** Consiste en cobrar una suma de dinero fija, y de única vez, para poder descargar el juego. Es uno de los modelos más tradicionales dentro de la industria de juegos para PC o consolas. Existe también en el universo móvil, donde los jugadores deben pagar una suma *one-time* o abono mensual para poder jugar el juego.

- **Juegos de Micro-financiación (*Crowdfunding*):** Son aquellos juegos que se desarrollan con los aportes de múltiples personas o entidades, generalmente a través de plataformas como *Indiegogo* o *Kickstarter*. En la mayoría de los casos, el creador del juego declara su intención de desarrollar un juego en particular, y la cantidad de dinero que necesita para lograrlo. Si y solo si la comunidad aporta el capital requerido, entonces el juego se desarrollará.

7.3. Wildlife Studios

Wildlife Studios, inicialmente llamada TFG (Top Free Games), es una empresa latinoamericana fundada en Brasil, en el año 2011, por dos hermanos: Víctor y Arthur Lazarte, cuando la industria móvil era realmente chica (en comparación con la actualidad). La empresa fue fundada sin capital externo, y con una mínima inversión inicial por parte de inversores ángeles. En los años siguientes, y con la rápida expansión del mercado de juegos móviles, Wildlife Studios vio un crecimiento exponencial.

Desde sus orígenes, Wildlife Studios se dedicó principalmente a crear y monetizar juegos *freemium*, algo realmente innovador por el año 2011, cuando el estándar de la industria eran los juegos pagos.

El primer juego lanzado por la empresa fue "*Racing Penguin*", que obtuvo un éxito inesperado, alcanzado las primeras posiciones de los rankings de las tiendas de Apple a nivel mundial, y particularmente en EE. UU. *Racing Penguin* le permitió a Wildlife Studios ser una empresa rentable desde prácticamente sus comienzos, y facilitó enormemente la tarea de buscar capitales externos y expandir la empresa.

La primera inversión externa llegó en el año 2012, de la mano de *Bessemer Venture Partners*, habilitando el crecimiento y lanzamientos de juegos exitosos como "*Bike Race*", uno de los primeros juegos competitivos multi-jugador, llegando a tener más de 100.000 jugadores en simultaneo. Otros ejemplos de juegos exitosos de Wildlife Studio son: *Sniper 3D*, *War Machines*, *Zooba*, *Tennis Clash*, entre otros.

En 2019, luego de años de crecimiento ininterrumpido, Wildlife Studios volvió a recibir inversiones externas, obteniendo una valuación de 2.900 millones de dólares americanos (USD). A finales de ese mismo año, con el surgimiento de la pandemia de COVID-19, la industria de los juegos móviles solo aceleró su crecimiento, dejando excelentes perspectivas de crecimiento futuro para la empresa.

A partir de 2020, Wildlife Studios cambió su estrategia de negocio, en un intento de transformar las dinámicas de la industria de los videojuegos. La empresa dejó de

poner foco en su estudio interno (responsable por la creación de juegos nuevos), y en su lugar comenzó a invertir fuertemente en la creación de pequeños estudios distribuidos, alrededor de unos pocos talentos pre-existentes en la industria. A partir de ese momento, Wildlife comenzó a centrar sus esfuerzos no en crear nuevos juegos, sino en desarrollar una plataforma que simplificara y acelerara la creación de los mismos, para que sus estudios afiliados pudiesen idear, crear y operar juegos a gran escala, con una cantidad reducida de personas.

Al día de hoy (2024), Wildlife no tuvo nuevas rondas de inversión, y actualmente cuenta con un portafolio de más de 60 juegos lanzados, que fueron descargados más de 3 mil millones de veces. La empresa cuenta con aproximadamente 800 empleados, distribuidos alrededor del mundo, con una fuerte concentración en San Pablo (Brasil), Buenos Aires (Argentina), Seattle (Estados Unidos), y Tel Aviv (Israel), entre otros (Wildlife, 2020).

CAPÍTULO VIII: DRIVERS DE NEGOCIO Y MÉTRICAS

8.1. La Importancia de Contar con Observabilidad en Eficiencia

A medida que las empresas adoptan tecnologías de nube pública, es una constante, en la mayoría de los casos, que los costos de su utilización sean más altos de lo inicialmente previsto. Ya sea porque muchos recursos son significativamente más caros en comparación con su comparable en Centros de Datos propios; porque el conocimiento para provisionar dichos recursos de forma eficiente no es el adecuado, porque el esfuerzo dedicado a optimizar los recursos provisionados es insuficiente; o una combinación de varios de estos factores mencionados (Martens, 2012). Adicionalmente, es muy común encontrarse ante escenarios donde los costos de nube pública crecen a una velocidad mayor a la del negocio y, por ende, significan un riesgo para la rentabilidad futura del mismo. Durante este capítulo veremos métricas e indicadores que ayudarán a detectar y conseguir un balance óptimo entre costos de nube pública, e inversión destinada a su correcta configuración y optimización.

La facilidad (y velocidad) que ofrecen las plataformas de Nube Pública para provisionar nuevos recursos es asombrosa: lo que antes llevaba días, o incluso semanas (como, por ejemplo, crear una nueva máquina virtual), en el nuevo paradigma de *Cloud Computing* puede lograrse en minutos, mediante algunos pocos *clics*, con reducciones de tiempo de hasta 76% (Perry, 2009).

Sin dudas, esta mejora en la velocidad trajo como consecuencia una aceleración en los negocios que supieron aprovecharla, que permitió hacer más en menos tiempo, iterando más rápido, y acortando los tiempos de experimentación, creación, y *time-to-market*. Sin embargo, esta facilidad trajo consigo nuevos desafíos, principalmente orientados hacia el gobierno, control y eficiencia de la utilización de los servicios de Nube Pública.

A partir de 2017, Wildlife Studios, como muchas otras empresas apalancadas sobre *Cloud*, y que vieron un rápido crecimiento de su negocio, empezó a encontrarse con un escenario donde sus costos de Nube Pública crecían más rápido que su facturación. Esta tendencia de crecimiento “desmedido” se fue consolidando con los

meses, e hizo evidente que Wildlife se encontraba ante un problema que podría comprometer las finanzas del negocio, de perdurar en el tiempo.

Tan solo 6 años luego de su concepción, Wildlife Studios ya se encontraba con una preocupación significativa respecto de sus costos de nube pública. Esto coincide con el análisis de las encuestas realizado en el sub-capitulo 6.1, donde se describió como la totalidad de las empresas con más de 3 años, y consumos significativos en cloud, mostraron niveles de preocupación alto respecto de sus gastos. Demostrado, de esta manera, que el caso particular de Wildlife se encuadra a la perfección con un caso promedio del grupo observado.

En la Ilustración 7 puede apreciarse el rápido crecimiento en los consumos de Nube Pública a partir del año 2017. Cabe aclarar que, por razones de confidencialidad, los valores de la ordenada (costo) no se muestran; pero sepa el lector que, para los últimos años de la serie, Wildlife Studios gastaba en el orden de las decenas de millones de dólares americanos por año. Sin dudas, un número más que significativo para el balance de la gran mayoría de las empresas. Sin embargo, lo realmente importante que se desea ilustrar no es el valor absoluto del gasto, sino la aceleración que el mismo mostró durante el período comprendido entre los años 2017 y 2021.

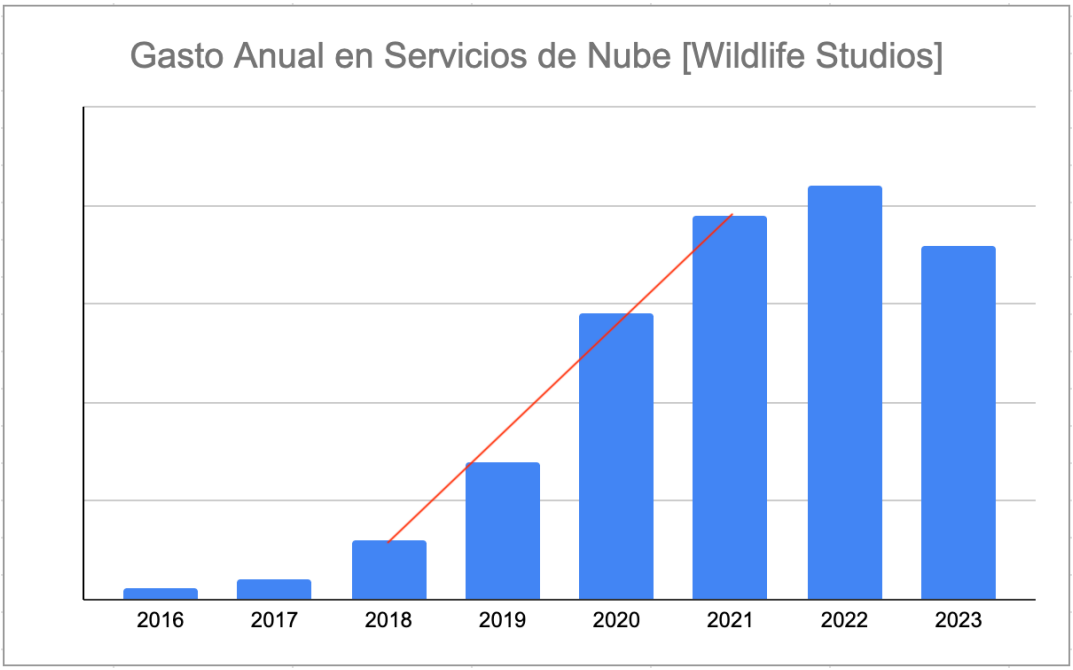


Ilustración 13 - Gasto Anual en Servicios de Nube pública de Wildlife Studios (Fuente: Elaboración propia)

La aceleración de los costos a partir de 2017 resulta un comportamiento específicamente descripto por los distintos expertos entrevistados, donde se identifica un patrón común: para aquellas empresas que no dediquen suficiente esfuerzo en construir un marco de gobierno y observabilidad, los costos de cloud crecerán fuera de control. Esto fue exactamente lo observado en el caso de referencia, donde en los años 2018, 2019 y 2020 se observaron crecimientos YoY[3] en el gasto de cloud superiores al 100%, y muy por encima al crecimiento del negocio, o del *revenue*. De tal forma, dicho gasto resultó un riesgo operativo identificado por el equipo de finanzas que, de no mitigarse, se constituiría indefectiblemente en un problema que atacaría la rentabilidad del negocio y pondría en jaque la sostenibilidad del mismo, en el mediano plazo.

A partir de principios de 2019, Wildlife Studios comenzó con sus primeros esfuerzos para primero entender, y luego buscar una solución que hiciera sustentable la utilización de nube pública en el largo plazo. El primer paso fue identificar los principales servicios que representaban alrededor del 80% del costo total de Nube: Cómputo (EC2), Almacenamiento (S3), Bases de datos (RDS, MongoDB, DynamoDB, etc.), y Networking.

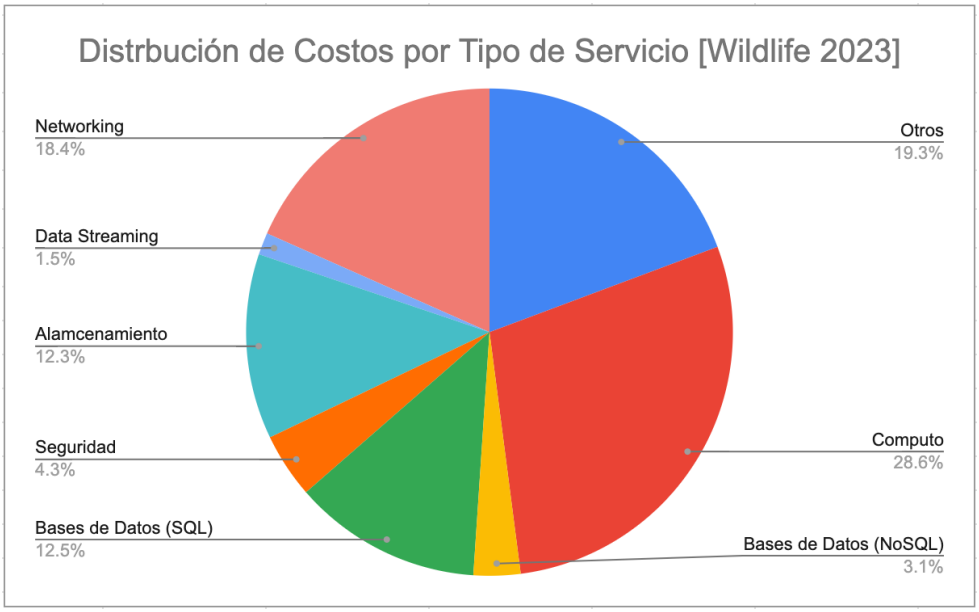


Ilustración 14 - Distribución de Costos por Tipo de Servicio (Fuente: Elaboración propia)

[3] *Year-over-Year*

Luego, se procedió a identificar y comenzar a medir aquellas métricas que midieran la eficiencia en el provisionamiento y utilización de los principales servicios; para luego poder a) realizar *benchmarks* con pares en la industria, y así identificar aquellas oportunidades con mayor posibilidad de retorno; y b) definir objetivos y medir el progreso respecto de los avances realizados por el equipo responsable.

Cabe destacar que Wildlife tomó la decisión de conformar un equipo dedicado de FinOps para atacar este problema, contratando especialistas con experiencia en la industria. Esta solución, que demostró ser efectiva, coincide con la recomendación de los especialistas entrevistados, y los comentarios observados por varios encuestados del sector. Se comprueba a través de esta, que dicho mecanismo resulta particularmente efectivo para resolver el problema antes descripto.

8.2. Métricas de Eficiencia Significativas

A continuación, se listan algunas de las métricas utilizadas por Wildlife (y MercadoLibre, entre otros) para evaluar y monitorear la eficiencia *Cloud*. Cabe destacar que el siguiente listado comprende aquellas métricas lo suficientemente genéricas para poder ser utilizadas por la gran mayoría de empresas que hacen uso de servicios de Nube Pública, en cualquiera de los principales proveedores (X. Guerron, 2020):

- **Porcentaje de utilización de CPU:** métrica que mide cuán usada es la capacidad de cómputo disponible durante un periodo de tiempo fijo. Generalmente, se busca que esta métrica sea lo más alta posible, para asegurar un aprovechamiento máximo del recurso. Resulta relativamente sencillo lograr porcentajes de utilización alto cuando la aplicación tiene un consumo estable y constante del procesador. Cuanto más grande sea la diferencia entre el máximo de utilización y el promedio de utilización del CPU, más difícil será lograr valores de eficiencia alta para esta métrica.

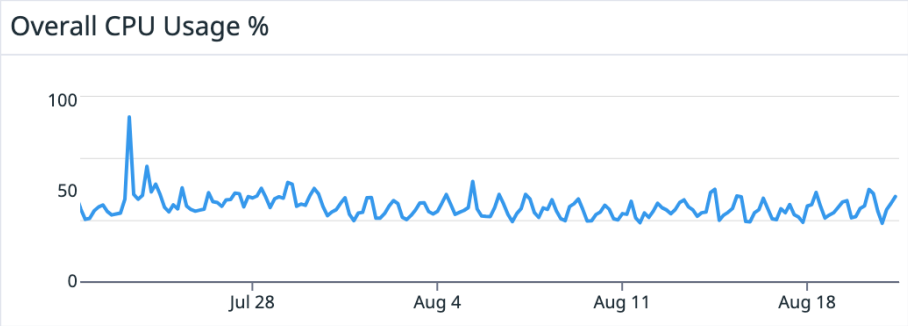


Ilustración 15 - Utilización de CPU (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de dos expertos (Pampliega y Bustos), relevada durante las entrevistas a especialistas del sector.

- **Porcentaje de utilización de Memoria:** Al igual que la anterior, esta métrica apunta a medir cuan eficiente es el uso del recurso de memoria RAM de un servidor. Generalmente, se busca que esta métrica sea lo más alta posible, para asegurar un aprovechamiento máximo del recurso, y así evitar contar con recursos ociosos, que igualmente están siendo pagados. A diferencia del CPU, la memoria es un recurso no compresible, lo que significa que no existe manera de "limitar" el uso de la memoria, por lo que lograr valores de eficiencia en la utilización de memoria puede resultar más complejo respecto del procesamiento.

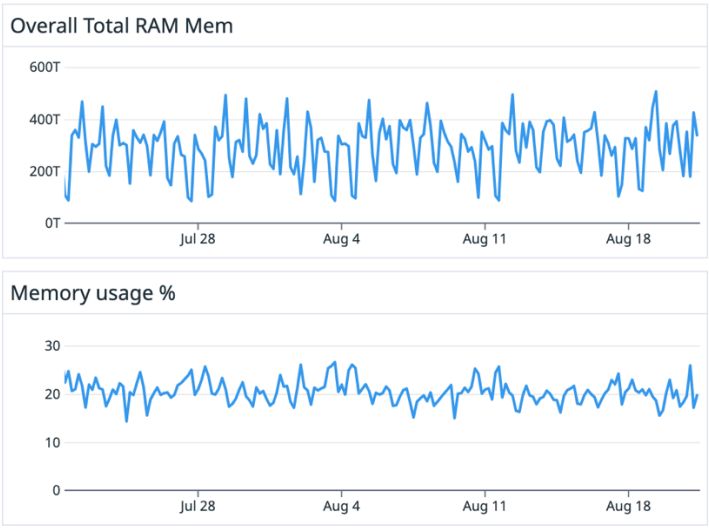


Ilustración 16 - Utilización de Memoria (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de dos expertos (Pampliega y Bustos), relevada durante las entrevistas a especialistas del sector.

- **Costo por GB de información almacenada:** Mide el costo promedio por unidad de almacenamiento. Como se describió en capítulos siguientes, existen diversas formas de almacenar la misma información, por lo que calcular el costo promedio por GB será un indicador de la eficiencia de almacenaje. Se busca que esta métrica sea lo más baja posible.

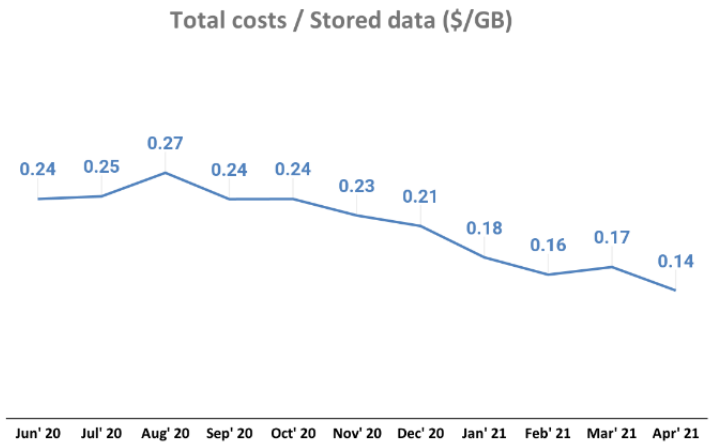


Ilustración 17 - Costo por Unidad de Almacenamiento (Fuente: Elaboración propia)

- **Costo por GB de información transferida:** Es una métrica análoga a la anterior. Existen diversos mecanismos para transferir información que serán detallados en este trabajo. Elegir el mecanismo, o la combinación de mecanismos correctos dará como resultado un costo por unidad de información transferida más eficiente. Al igual que la anterior, se busca que esta métrica sea lo más baja posible.

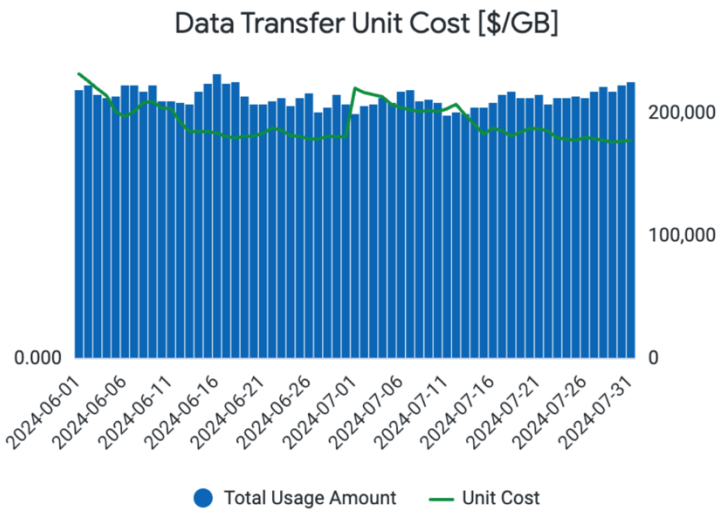


Ilustración 18 - Costo Unitario por GB Transferido (Fuente: Elaboración propia)

- **Costo por transacción:** Mide el costo total agregado por transacción procesada, siendo la definición de transacción variable de acuerdo con el negocio o al alcance del área analizada. Ejemplos de transacciones podrían ser i) comprar un producto (por ejemplo, en MercadoLibre); ii) reproducir un tema musical (por ejemplo, en Spotify); o iii) programar un viaje (por ejemplo, en Uber), entre muchas otras. Lo que busca esta métrica es sumar todos los costos de nube asociados a cada una de estas transacciones para poder tener un punto fijo de referencia que permitirá entender el costo de escalar el negocio. Esta métrica siempre contará con un componente fijo, que no dependerá de la cantidad de transacciones, y una componente variable, que tendrá una dependencia directa y proporcional con la cantidad total de transacciones. Se busca que esta métrica sea lo más baja posible.

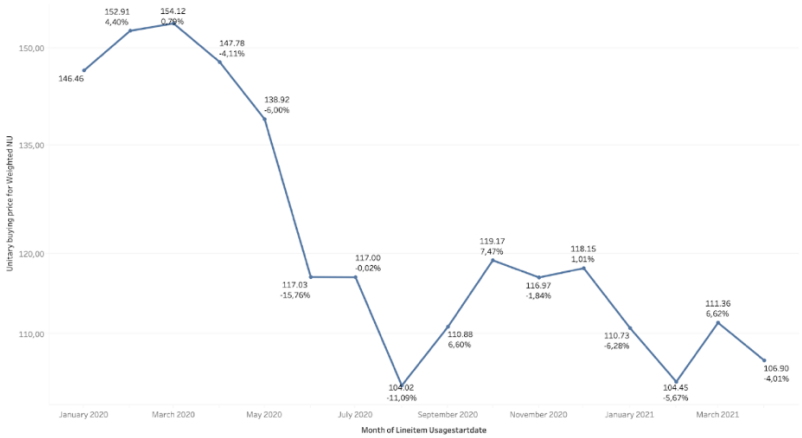


Ilustración 19 - Costo Unitario por Transacción (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de un experto (Simonassi), relevada durante las entrevistas a especialistas del sector.

- **Costo por usuario:** Mide el costo total agregado por usuario. Al igual que la anterior, esta métrica suma todos los costos de nube asociados a cada usuario del servicio para poder tener un punto fijo de referencia que permitirá entender el costo de escalar el negocio. Un ejemplo de esta métrica podría ser calcular el costo de recursos de nube por cada jugador conectado a una partida en línea (por ejemplo, del juego Fortnite, de Epic Games). Esta métrica siempre contará con un componente fijo, que no dependerá de la cantidad de usuarios, y una componente variable, que tendrá una dependencia directa y proporcional con la cantidad total de usuarios. Se busca que esta métrica sea lo más baja posible.

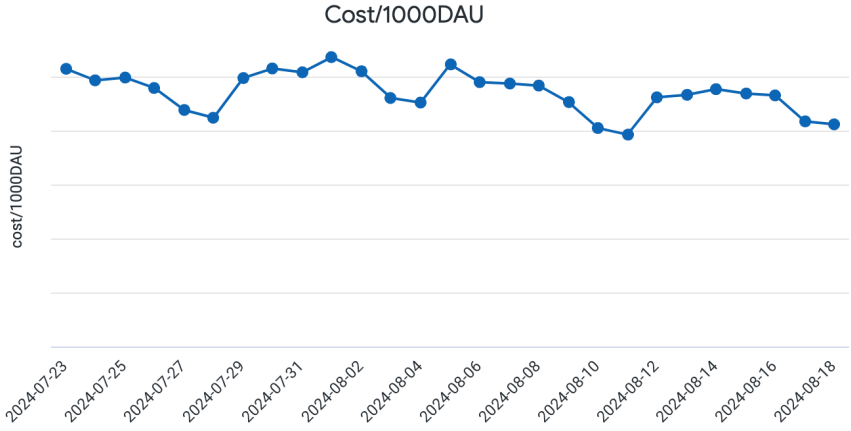


Ilustración 20 - Costo cada mil usuarios (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de un experto (Simonassi), relevada durante las entrevistas a especialistas del sector.

- **Costo por empleado**[4]**:** Esta métrica mide el costo total agregado por desarrollador. Suma todos los costos de nube, asociados a un área específica, en un periodo de tiempo fijo, y se los divide por la cantidad de desarrolladores que mantienen el o los productos de la empresa. Se calcula para tener un punto fijo de referencia que permitirá entender la eficiencia general de las prácticas de desarrollo. Se busca que esta métrica sea lo más baja posible.

- **Eficiencia de *upscaling* y *downscaling*:** Se calcula midiendo la diferencia entre la curva de capacidad alocada y la demanda requerida en un periodo de tiempo fijo. Se busca que la infraestructura escale (aumente la capacidad) en la medida que la demanda se incrementa, y de la misma forma, des-escale (disminuya la capacidad) lo más rápidamente posible, una vez que la demanda disminuye. Se busca que la diferencia entre curvas sea lo más baja/chica posible.

- **Porcentaje de utilización de recursos efímeros:** Se calcula midiendo la utilización de recursos de cómputo efímeros, en comparación con la utilización total de recursos de cómputo. Dado que los recursos efímeros es usualmente la modalidad de provisionamiento más económica, un alto uso de recursos efímeros indicará una alta eficiencia en la modalidad de contratación de recursos de cómputo. Por ende, se busca que esta métrica sea lo más alta posible.

[4] En general, se utiliza con perfiles de desarrolladores de software, analistas de negocio, perfiles de BI, científicos de datos, ingenieros de datos, y otros.

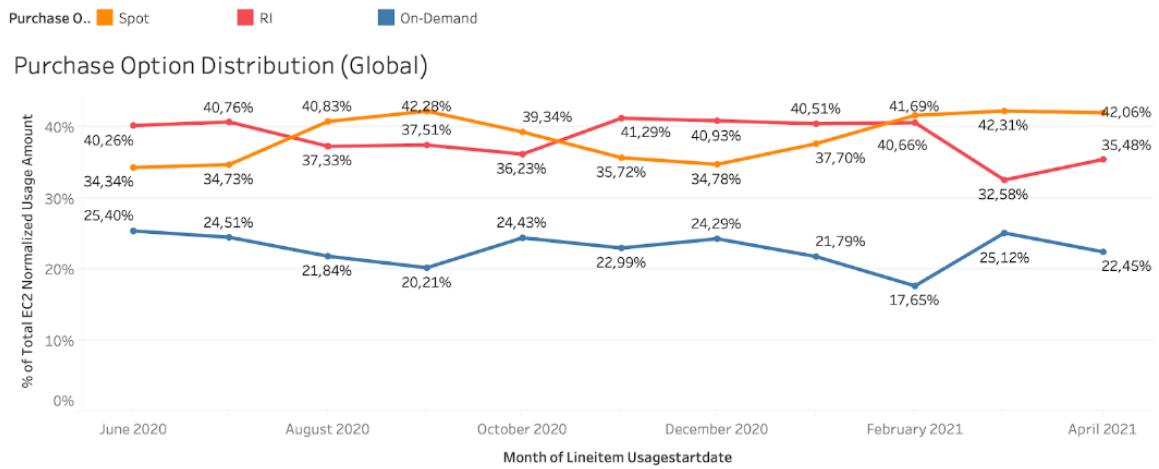


Ilustración 21 - Distribución de Modos de Provisionamiento (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de un experto (Pampliega), relevada durante las entrevistas a especialistas del sector. A su vez, esta métrica resultó identificada por una porción significativa de los especialistas que respondieron la encuesta, identificando la misma como uno de los mecanismos más efectivos a la hora de eficientizar la operación de cloud y reducir el costo total de la misma, sin comprometer la operación del negocio.

- **Cobertura de recursos reservados:** Métrica que indica la cantidad de recursos reservados, en comparación con la cantidad de recursos totales (excluyendo los recursos efímeros). Se utiliza para detectar oportunidades de reservas, con el potencial de disminuir los costos totales de recursos de nube. Se busca que esta métrica sea lo más alta posible (Storment, 2015).

- **Utilización de recursos reservados:** Se mide calculando la utilización de los recursos reservados, en comparación con la totalidad de recursos reservados. Es una métrica que ayuda a detectar excesos (o faltantes) de reservas, que pueden ser convertidas a otros tipos de recurso (que sea actualmente demandado por la organización), o vendidas en mercados secundarios. Se busca que esta métrica sea lo más alta posible. En contraposición, se busca también que la subutilización de recursos reservados sea cercana a 0% (Dar Juan, 2024).

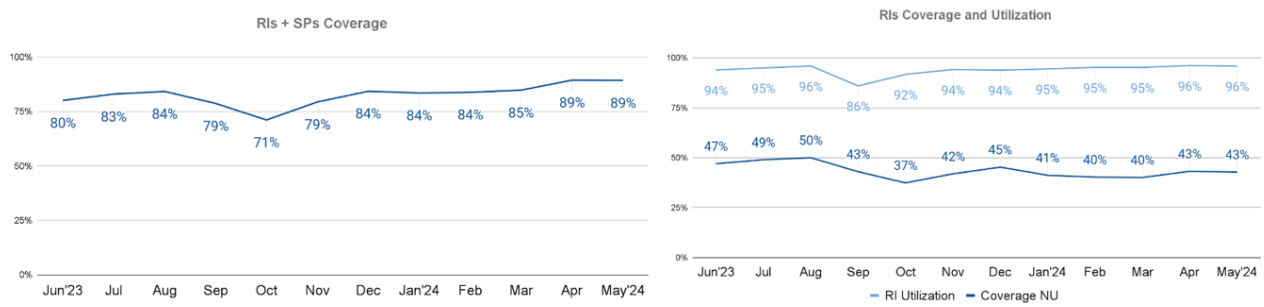


Ilustración 22 - Cobertura y Utilización de Recursos Reservados (Fuente: Elaboración propia)

Se destaca que dicha métrica coincide con la recomendación de dos expertos (Simonassi y Pampliega), relevada durante las entrevistas a especialistas del sector. A su vez, esta métrica resultó identificada por una porción significativa de los especialistas que respondieron la encuesta, identificando la misma como uno de los mecanismos más efectivos a la hora de eficientizar la operación de cloud y reducir el costo total de la misma, sin comprometer la operación del negocio.

- ***Cache hit ratio*:** Para aquellas aplicaciones que utilicen mecanismo de caché, esta métrica muestra el porcentaje de pedidos que son efectivamente servidos por la capa de *caching*, en comparación con la totalidad de pedidos. Servir información a través de mecanismos de caché resulta siempre más rápido, más barato y eficiente. Es por esto por lo que se busca que esta métrica sea lo más alta posible. Adicionalmente, una métrica de *cache hit ratio* baja indicará una pobre utilización de los servicios de caché, lo que puede ser indicador de i) una mala implementación; o ii) una capa innecesaria que puede ser eliminada (Novotný, 2022).

Vale aclarar, que para que la mayoría de las métricas descriptas métricas sean comparables, se debe tener en cuenta procesos o aplicaciones que sean de naturaleza similar.

CAPÍTULO IX: MECANISMOS DE EFICIENCIA OPERACIONALES

En el presente capítulo se describieron aquellos mecanismos que permitieron influenciar las métricas descriptas en el Capítulo VIII, con particular foco sobre el potencial de cada uno, las dependencias entre los mismos, y las estrategias empíricamente probadas para maximizar el éxito en su ejecución.

9.1. Acuerdos Comerciales

Negociar acuerdos comerciales es una estrategia que, según el tamaño de la organización compradora, puede derivar en descuentos realmente significativos en comparación con los precios de lista de los principales proveedores de servicios *cloud*. Los acuerdos comerciales privados son una herramienta que cualquier consumidor mediano[5] puede aspirar a negociar. Vale aclarar que los montos de consumo mínimo para acceder a estos acuerdos, y los potenciales descuentos por conseguir, tendrán una fuerte dependencia en el proveedor de *cloud* que se elija, el mercado o la industria donde la compañía opere, y con la región desde la cual se contraten dichos servicios (Lee, 2019).

Asegurar un acuerdo comercial de precios requiere planeamiento estratégico, conocimiento extensivo de la demanda de recursos de la empresa contratante, de los modelos de precio del proveedor de servicios, y habilidades de negociación. Sin embargo, una negociación efectiva puede brindar descuentos que van desde pequeños porcentajes (~5%), hasta descuentos agregados superiores al 50% (Krishnakumar, 2024).

Los acuerdos comerciales privados implican, en la mayoría de los casos, un compromiso de consumo por parte del cliente, por un período mínimo de 1 (un) año. Por supuesto, compromisos de consumo mayores y períodos de contrato más largos contribuirán significativamente a obtener descuentos más altos. Es por esto que el primer paso para encarar una negociación siempre será analizar y predecir la necesidad de recursos de la empresa contratante. De esta forma, la empresa

[5] Se considera consumidor mediano a todo aquel cliente que tenga un gasto mayor a USD 100.000 por año.

contratante deberá tener una predicción de la necesidad futura, que permita asumir un compromiso de consumo, minimizando el riesgo de incumplimiento.

En segundo lugar, contar con un caso de negocios que ilustre el crecimiento potencial de la empresa en los siguientes años puede contribuir enormemente a la efectividad en la negociación. Los grandes proveedores de nube suelen mostrarse más abiertos a negociar descuentos cuando el cliente muestra signos de crecimiento potencial exponencial para los años subsiguientes.

El próximo paso consistirá en auditar el uso actual de infraestructura y servicios de nube, entendiendo su estructura de costos y métricas de utilización. Esto permitirá identificar potenciales oportunidades de eficiencia, asegurando que la base utilizada para proyectar el uso futuro esté correctamente dimensionada para la necesidad actual. A su vez, se recomienda identificar aquellos flujos que son fácilmente migrables entre distintos proveedores. De esta forma, se podrá contar con una herramienta adicional a la hora de presionar por descuentos mayores, apalancados sobre la competencia entre los distintos proveedores. Es importante para esto, tener un profundo entendimiento del ofrecimiento y los modelos de precios de los distintos proveedores de nube.

Adicionalmente a los aspectos propiamente técnicos, antes de encarar una negociación, será aconsejable identificar requerimientos propios del negocio, como por ejemplo certificaciones de cumplimiento sobre ciertas normas y políticas (PCI, ISO, SOX, entre otras), y/o requerimientos de disponibilidad mínimos, usualmente especificado en los que se conoce como SLAs (*Service Level Agreements*). En caso de requerir condiciones especiales al respecto, será increíblemente más fácil hacerlo a la hora de negociar grandes contratos, especialmente porque lo principales proveedores se muestran generalmente reticentes a aplicar condiciones particulares para clientes específicos, en especial si los mismos no son clientes realmente grandes (y atractivos para los proveedores de nube pública).

Otra dimensión que será importante considerar a la hora de negociar un acuerdo de precios privado será las condiciones de pago, específicamente el periodo disponible entre la facturación y el pago efectivo mensual o anual (normalmente NET30, pero

resulta posible extenderlo), y el momento en que se aplicará el otorgamiento de créditos, que podrán afectar considerablemente el flujo de caja de ambas partes.

Por último, será posible en algunos casos (y aconsejable), introducir cláusulas de salida, que permitan a la empresa contratante terminar de forma anticipada el contrato a negociar, de manera de reducir el periodo de efectivo de contratación, y por ende el riesgo asumido con la rúbrica de este.

Negociar contratos agresivamente con los proveedores resultó una de las mayores oportunidades de optimización relevadas a partir del análisis de las respuestas a la encuesta a especialistas (ver sub-capitulo 6.1). De dicho análisis se desprendió que un gran número de empresas (39%) opera en la nube pública sin contratos. Esto significa que no poseen ningún compromiso, y por ende, mantienen un nivel de riesgo bajo. Sin embargo, al no poseer contratos de mediano y largo plazo, estas empresas dejaron de acceder a potenciales descuentos comerciales que pueden influir positiva y significativamente sobre el gasto total de cloud, con descuentos comerciales para nada despreciables.

Adicionalmente, este mecanismo fue identificado como de gran importancia por al menos un experto, a partir de las entrevistas analizadas en el sub-capítulos 6.2. Allí puede observarse como Pampliega destaca la importancia de “negociar contratos comerciales de forma agresiva con el/los proveedores de cloud elegidos” para asegurar consumos medidos.

9.2. Reservas de Recursos

Tal como se describió en capítulos anteriores, los servicios de cómputo en la nube comprenden la capacidad de provisionar procesamiento, almacenamiento, conectividad, entre otros recursos fundamentales de IT, a través de la red, evitando que los usuarios deban invertir, mantener y operar la infraestructura física. Estas responsabilidades son de dominio exclusivo de los proveedores de servicio, quienes deben gestionar centros de cómputos masivos, a escala.

Una de las dificultadas descriptas anteriormente, asociadas a la operación de Centros de Cómputos, es la *gestión de la demanda*, que se refiere al proceso de asegurar la capacidad óptima para satisfacer la demanda actual y futura de servicios. Básicamente, asegurar que los Centros de Computo cuenten con los recursos necesarios para prestar servicios: si no se cuenta con los recursos suficientes, habrá nuevos clientes sin atender, y/o o peor aún, clientes actuales que sufrirán incidentes en sus negocios, por no poder escalar su operación. Por el contrario, si se contara con recursos en exceso, se incurrirá en costos adicionales relacionados al capital de trabajo, que atentarán contra la rentabilidad del negocio del proveedor de servicios. Es por esto por lo que contar con la capacidad justa y necesaria es de vital importancia para un prestador de servicios de cómputo en la nube.

Para realizar una correcta estimación de los recursos necesarios, es de gran importancia poder predecir la utilización futura de los mismos. Sin embargo, los principales proveedores de servicios de nube afrontan la complejidad de contar con miles de clientes, en diversos mercados, con dinámicas heterogéneas, y necesidades dispares. Es por esto que la demanda de recursos es usualmente altamente volátil, y estimar la misma con precisión resulta un desafío mayúsculo.

Para mitigar esta situación, los proveedores de nube generalmente cuentan con modelos de “Reservas de Recursos”, donde los clientes asumen un compromiso de utilización a futuro (que simplifica enormemente la capacidad de predecir la demanda), a cambio de descuentos significativos en los precios de esos recursos. Al final de cuentas, un Recurso Reservado representa un acuerdo de pago. Usualmente, los clientes se comprometen a usar un recurso por un periodo prolongado de tiempo, que usualmente va desde 1 a 3 años. A cambio, el proveedor otorga un descuento significativo, que usualmente ronda el ~20%, y puede alcanzar valores superiores a 70%, dependiendo del tipo de recurso, la forma de pago, y la duración del compromiso. Por supuesto, una vez asumido el compromiso, el mismo no puede cancelarse. Por ende, los clientes que elijan esta modalidad para provisionar recursos deberán tener un alto nivel de confianza de que necesitarán utilizar dicho recurso por el periodo contratado: el proveedor de servicio cobrará por dicho recurso, sin importar si el mismo es utilizado o no (Singer, 2010).

$$Costo\ OD = p * utilizacion$$
$$Costo\ RI = p * desc * duracion\ del\ compromiso$$

$$Costo\ RI < Costo\ OD$$
$$p * desc * duracion\ del\ compromiso < p * utilizacion$$
$$\boldsymbol{desc < \frac{utilizacion}{duracion\ del\ compromiso}}$$

Siendo *p* el precio del recurso bajo demanda, y *desc* el descuento por reservar dicho recurso, siempre convendrá optar por una reserva cuando la esperanza de utilización sea mayor al descuento multiplicado por la duración del compromiso de la reserva. En un ejemplo simple, para un descuento *del 30%*, es conveniente reservar el recurso, siempre y cuando se utilice más del *70%* del tiempo dentro del compromiso asumido.

A su vez, existen distintos tipos de Reservas, que pueden variar de acuerdo a los distintos proveedores de nube, siendo las siguientes las más usuales (Awati, 2024):

- **Reservas Estándar:** Son aquellas con el mayor descuento y la menor flexibilidad. Usualmente el cliente opta por un tipo de recurso, que no podrá ser cambiado durante la duración de la reserva.

- **Reservas Convertibles:** Poseen un descuento menor en comparación con las Reservas Estándar, pero cuentan con la capacidad de ser modificadas. Las modificaciones más usuales están asociadas al tipo de recurso requerido, por ejemplo, cambiar un recurso “chico” por uno más potente abonando únicamente la diferencia.

De acuerdo con el proveedor, los recursos reservados cuentan con la posibilidad de ser abonados de distintas maneras, siendo las siguientes las opciones más comunes:

- **Adelanto Total:** En esta modalidad, el cliente paga la totalidad del recurso al momento de contratación, evitando pagos futuros por el mismo, durante el período comprometido. Es la modalidad de pago que otorga mayor descuento.

- **Sin Adelanto:** En esta modalidad, el cliente paga el recurso contratado en cuotas fijas e iguales, devengadas mensualmente. Es la modalidad de pago que otorga el menor descuento.

- **Con Adelanto Parcial:** Esta modalidad es una mezcla de las dos anteriores. Aquí, el cliente paga por adelanto ~50% del valor total del recurso, y el restante ~50%, lo abona en cuotas iguales y fijas, devengadas mensualmente durante el período comprometido.

Aquellas entidades que opten por contratar Recursos Reservados deberán evaluar la mejor estrategia para abonar los mismos. Como regla general, cuando el descuento ofrecido por el proveedor sea mayor a la *tasa de descuento* del contratante, se deberá optar por pagos *Upfront* (adelanto total o parcial). Por el contrario, cuando el descuento ofrecido por el proveedor sea menor a la tasa de descuento, se optará por pagos que no impliquen adelantos de capital.

Adicionalmente a los beneficios comerciales, los Recursos Reservados poseen una ventaja operativa adicional: el proveedor garantizará su disponibilidad, incluso si el recurso tuviera más demanda que la capacidad instalada. Esto significa, que el cliente de un Recurso Reservado tendrá el mismo siempre disponible durante el tiempo de contratación. De esta forma, se elimina el riesgo de necesitar un recurso, y que el proveedor no tenga la capacidad/disponibilidad de vender ese recurso. Esto es particularmente útil para alojar aquellos procesos que sean críticos para el negocio (Ravhon, 2024).

En conclusión, los Recursos Reservados ofrecen un mecanismo estratégico para asegurar eficiencia comercial en la contratación de recursos, y garantizar disponibilidad de los mismos a lo largo del tiempo. Sin embargo, para asegurar que su aplicación sea verdaderamente eficiente, aquellos clientes que opten por Recursos Reservados, deberán tener un entendimiento preciso de la esperanza de utilización de dichos recursos, y la relación a largo plazo que deseen entablar con el proveedor de nube utilizado.

La importancia de utilizar Recursos Reservados, como RI/SP[6] se refuerza a través de las observaciones de expertos (Simonassi), relevadas durante las entrevistas a especialistas del sector, donde este mecanismo es identificado como estratégico a la hora de optimizar los gastos en nube pública. A su vez, esta herramienta resultó identificada por una porción significativa de los especialistas que respondieron la encuesta, identificando la misma como extremadamente efectiva a la hora de eficientizar la operación de cloud y reducir el costo total de la misma, sin comprometer la operación del negocio.

9.3. Utilización de Recursos Efímeros

En su esencia, los proveedores de Nube deben contar con grandes Centros de Cómputos (tradicionales) para luego poder alocar los distintos recursos, bajo demanda, a sus distintos clientes. Toda vez que un usuario de servicios de nube provisiona una instancia de cómputo, un volumen de disco, una base de datos, o espacio de almacenamiento, esos recursos son provistos por servidores y equipos de *storage*, propiedad del proveedor de nube, alojados en un centro de cómputos como los descriptos en el cuerpo teórico del presente trabajo. Esto significa que dichos proveedores, como Amazon (AWS), Google (GCP), o Microsoft (Azure), entre otros, sufren de las mismas dificultades al administrar sus centros de cómputos.

Una de las dificultades descriptas con anterioridad, corresponden a los costos de oportunidad que se generan por los recursos ociosos, o sea por el exceso de capacidad de dichos centros. Una de las formas más eficaces que han encontrado los proveedores de nube para mitigar este problema, es lo que se conoce como mercado de “recursos efímeros”: toda capacidad que en cualquier momento dado se encuentre ociosa, es generalmente puesta disposición de su base de usuario a un precio significativamente menor al precio de lista. Los recursos efímeros más comunes corresponden a las instancias de cómputo: en AWS conocidas como “*Spot Instances*”, en GCP como “*Preemptible* VMs”, y en Azure como “*Low-Priority* VMs”. Dichas instancias suelen tener descuentos de hasta ~80% en comparación con el precio de ese mismo recurso contratado en modalidad “*On-Demand*” (Isobe, 2021).

[6] *Reserved Instances / Savings Plans*

Esto significa que un usuario de nube que hace uso de estos recursos efímeros puede obtener exactamente el mismo recurso, con la misma capacidad y potencia, a una fracción del precio en comparación con la modalidad de contratación estándar. La gran diferencia entre un recurso bajo demanda y un recurso efímero, es que el proveedor puede recuperar dicho recurso en cualquier momento, con un corto aviso, no garantizado, que varía según el proveedor (de 0 a 2 minutos, en general) (Xie, 2017). Es decir, dicho recurso efímero estará disponible para su uso, al precio acordado, hasta el momento en que el proveedor requiera esa capacidad, para entregarla a otro usuario que esté dispuesto a pagar el precio estándar por ese mismo recurso.

La importancia de utilizar Recursos Efímeros, como instancias *spot*, *pre-emptibles machines*, o *low-priotiy instances*, se refuerza a través de las observaciones de expertos (Pampliega), relevadas durante las entrevistas a especialistas del sector, donde este mecanismo es identificado como estratégico a la hora de optimizar los gastos en nube pública. A su vez, esta herramienta resultó identificada por una porción significativa de los especialistas que respondieron la encuesta, identificando la misma como extremadamente efectiva a la hora de eficientizar la operación de cloud y reducir el costo total de la misma, sin comprometer la operación del negocio.

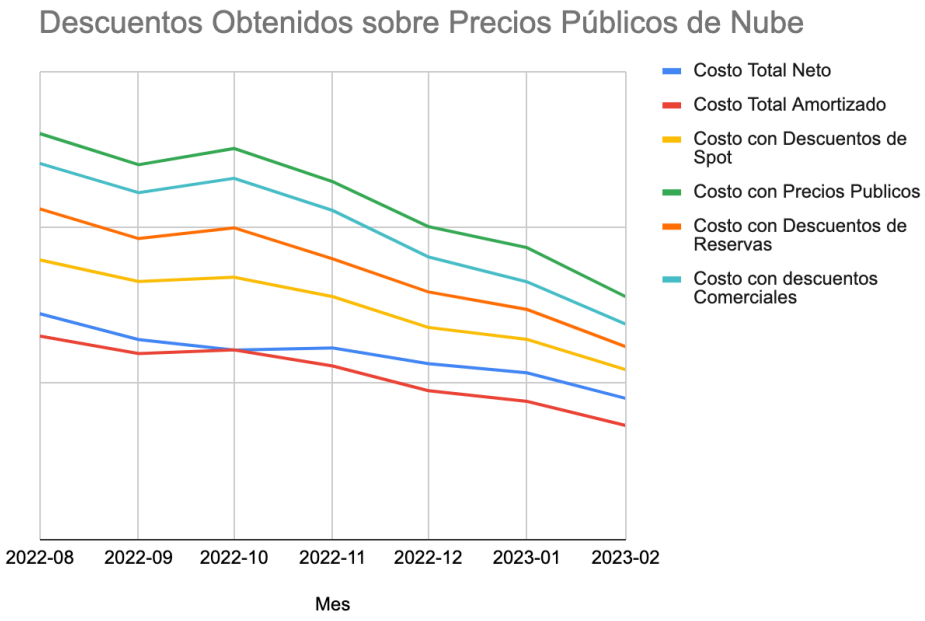


Ilustración 23 - Descuentos sobre Precios Públicos (Fuente: Elaboración propia)

Si bien los valores absolutos de costo mensual están intencionalmente ausentes, en la ilustración anterior puede observarse como se logra un descuento aproximado del 50% sobre los precios públicos del proveedor de nube[7], tan solo aplicando los mecanismos descriptos hasta el momento.

[7] Equivalente a provisionar la totalidad de los recursos en modalidad bajo demanda (*on-demand*), sin invertir esfuerzo en optimizar dichos recursos.

CAPÍTULO X: MECANISMOS DE EFICIENCIA TÉCNICOS

10.1. Aprovechando la Flexibilidad Inherente de los Servicios de Nube

Una de las mayores ventajas de utilizar servicios de nube pública es la capacidad de provisionar recursos, para luego des-provisionar los mismos luego de su utilización, y pagar el costo de dichos recursos únicamente por la fracción de tiempo utilizado (modalidad *on-demand*).

Para lograr un uso eficiente de los recursos *cloud* será clave escalar y des-escalar la infraestructura requerida para operar el negocio de acuerdo con la demanda en cada momento determinado. Esto significa aumentar la cantidad de recursos disponibles a medida que la demanda aumenta, y achicar la cantidad de recursos, en la medida que la demanda disminuye. Estos ciclos son muy comunes en aquellos negocios de alta estacionalidad diaria, semanal o mensual. Como ejemplo, puede pensarse en el *Marketplace* de MercadoLibre: debido a que opera únicamente en Latinoamérica, la demanda durante el día será mucho mayor que la demanda en períodos nocturnos, donde gran parte de los usuarios se encontrará durmiendo. Es por esto que la necesidad de recursos de infraestructura será significativamente mayor durante el día que durante la noche. Por ende, será de gran importancia configurar las aplicaciones en **grupos de escalado** que sigan aquellas métricas que miden la utilización de recursos y permitan incrementar o disminuir la cantidad de infraestructura disponible de forma automática y proporcional a la cantidad de demanda para cada período de tiempo (Lorido-Botran, 2014).

Cabe destacar que habrá que prestar particular atención a los parámetros de velocidad de escalado y des-escaldo. Dado que el provisionamiento de recursos no es instantáneo (generalmente requiriere algunos pocos minutos), habrá que tener en cuenta la variabilidad de cada aplicación para asegurar que siempre se cuenta con recursos ociosos disponibles para garantizar la operación, incluso en aquellos periodos de tiempo donde ocurran variaciones bruscas en la demanda.

La flexibilidad del cloud resulta una característica inherente al servicio de nube pública que fue observada por los tres especialistas entrevistados (ver sub-capitulo 6.2). En

dicho análisis se destacó la importancia de adoptar y utilizar esta característica como uno de los mayores diferenciales positivos en comparación con los modelos tradicionales de Datacenter con infraestructura propia, permitiendo ajustar los recursos en tiempo real, según las necesidades del negocio.

10.2. Eligiendo el Nivel de Almacenamiento Adecuado

Todos los proveedores de nube pública ofrecen servicios para almacenar información. Sin embargo, es muy común encontrar múltiples servicios con este mismo fin. Dado que existen diversos mecanismos para almacenar información, como pueden ser los discos de estado sólido, los discos mecánicos, o incluso almacenamiento a través de cintas magnéticas, la elección de uno sobre otro dependerá de las necesidades del cliente.

Todos los mecanismos son realmente efectivos para almacenar *data*. Sin embargo, la velocidad de lectura y/o escritura (*throughput*) dependerá del medio que se utilice para almacenar dicha información. Cabe de esperar por ende que medios de acceso más rápidos sean más costosos en comparación a sus pares más lentos.

Los equipos que deseen contar con una utilización eficientes de medios de almacenamiento, deberán entender cuidadosamente tres variables, de modo de poder elegir la configuración de *storage* correcta para sus aplicaciones (Erradi, 2020):

- **Tiempo de Recupero:** Cuánto tiempo de espera puede soportar una aplicación para tener disponibilidad sobre la información almacenada. Normalmente, esta variable va desde acceso instantáneo, hasta varias horas (o incluso días) de espera.

- **Velocidad de Lectura/Escritura:** Es la variable que corresponde con el ancho de banda que tendrá el hardware para leer o escribir cierta cantidad de información por período de tiempo.

- **Cantidad de lecturas y/o escrituras**: Se corresponde a la cantidad de veces que esa información será accedida (tanto para lectura como para modificaciones) en un cierto período de tiempo.

Una vez definidos los requerimientos aplicativos, se podrá proceder a elegir el servicio que ofrezca las capacidades mínimas, al mejor precio posible. Cabe aclarar que este análisis deberá repetirse con cierta periodicidad, dado que dichos parámetros son susceptibles a sufrir cambios a lo largo del tiempo.

10.3. Transferencia de Datos

Al igual que ocurre con el almacenamiento, existen múltiples mecanismos para transmitir información entre aplicaciones. Como regla general, cuanto más próximos se encuentren los clientes de una comunicación (emisor y receptor), más barato será el costo de transferencia.

Para comunicaciones entre servidores propios, deberá priorizarse, en la medida de lo posible, que dichos servidores se encuentren en la misma red (VPC). Cuando esto no sea posible, deberá priorizarse la comunicación con servidores que se encuentren en la misma región. Por último, la comunicación que deberá tratar de evitarse es la comunicación entre servidores propios a través de internet. Estas elecciones tendrán un impacto directo en performance, dado que la latencia en la comunicación será significativamente menor cuanto más cerca se encuentren el receptor y emisor de los mensajes de red. Adicionalmente, tendrá un impacto en costos, dado que el tráfico dentro de una misma VPC suele ser gratuito, pero no así el tráfico saliente de una VPC (Degani A. , 2023).

Para aquellos casos donde deban distribuirse archivos estáticos a clientes a través de internet, será altamente recomendable el uso de una CDN (*Content Distribution Network*) que minimizará el tiempo de entrega y el costo de dichas operaciones.

10.4. Alocación de Costos

Uno de los pilares para asegurar el uso eficiente de recursos de nube es la correcta alocación de costos. Si bien es común encontrar equipos centralizados y

especializados en "*Cloud Economics*" dentro de las empresas que hacen consumo extensivo de recursos *cloud*, la responsabilidad sobre el gasto y la eficiencia debe ser un parámetro claro y preciso, con un alto grado de capilaridad y granularidad.

En aquellos casos donde la responsabilidad sobre la eficiencia se encuentra en un equipo único y centralizado, ocurre que la escalabilidad de dicho patrón resulta insuficiente para lograr altos niveles de eficacia: ser altamente eficiente en la nube requiere de conocimiento extensivo del ámbito técnico, pero también demanda conocimiento concreto de negocio y lógica aplicativa sobre el flujo o aplicación que se busca eficientizar. En la práctica, resulta posible contar con un equipo centralizado que posea el conocimiento técnico suficiente para asegurar la eficiencia. Sin embargo, es virtualmente imposible lograr que este equipo mantenga un conocimiento extenso y actualizado de la lógica aplicativa y de negocio sobre las distintas aplicaciones de la empresa. Es especialmente cierto en grandes empresas que cuentan con cientos, miles, o decenas de miles de aplicaciones o microservicios distintos.

Por otro lado, cuando la responsabilidad se asigna de forma suave (o imprecisa) sobre todos los desarrolladores, se cae en un escenario donde la obligación es de todos, pero a su vez es de nadie. Como consecuencia, se logra un compromiso pobre, o en el mejor de los casos, inconsistente, sobre la responsabilidad de asegurar altos niveles de eficiencia en el consumo y provisión de recursos. En estos casos, es común ver como los empleados de empresas que utilizan infraestructura en la nube, perciben los costos asociados como un "impuesto corporativo", un costo que usualmente desconocen, y que en ocasiones consideran insignificantes para el negocio (Wang, 2020).

El escenario ideal en estas ocasiones es lograr que los empleados se sientan responsables de lograr que los proyectos en los que trabajan tengan un RoI (retorno sobre la inversión) positivo. Para esto, deben conocer el impacto esperado de sus proyectos en el negocio, información que normalmente estiman los gerentes de producto, y también deben conocer los costos asociados a estos mismos proyectos. El primer paso para garantizar que esto ocurra, es comenzar por lograr una asignación de costos precisa a los distintos actores, equipos y unidades de negocio,

para luego generar una visualización clara sobre los mismos, que impulse el sentido de responsabilidad y *accountability*.

Existen diversas estrategias para encarar el desafío de lograr que los desarrolladores accedan a una visualización clara de los costos generados por sus acciones (o sea, por las aplicaciones que desarrollan y alojan en la nube). Entre ellas, las más comunes consisten en enviar reportes consolidados, con una cadencia fija, detallando el uso, el consumo y costo de los recursos de nube utilizados durante el periodo comprendido. En estos casos, se suele enviar un reporte a cada equipo de la organización, incluyendo las variables económicas asociadas a los consumos de nube. En otros casos, existen empresas que complementan esta estrategia con actividades de gamificación, donde se incentiva a los distintos equipos de la organización a competir entre ellos por conseguir las mejores métricas asociadas al uso eficiente de la infraestructura. Resulta común utilizar una o más métricas de las descriptas en capítulos anteriores. Otras empresas emplean mecanismos más tradicionales, como alocar dichos costos en los sistemas financieros, reportando los gastos de nube como parte del P&L de cada equipo, unidad de negocio, y/o proyecto. En estos casos, la empresa se apalanca sobre mecanismos ya existentes, donde dichos gastos se agregan en “Centros de Costos”, que son monitoreados por las áreas de Planificación y Finanzas.

Dependiendo de la precisión que se busque en la alocación de los costos de nube, se pueden emplear distintos algoritmos. Desde los más sencillos, como distribuir los costos de forma proporcional a la cantidad de empleados (o *revenue*) de cada área, hasta los más sofisticados, que consisten en desarrollar un motor de facturación interno que distribuya los gastos basados en indicadores de consumo. Existen organizaciones que incluso modifican estos costos previo a su alocación para incentivar (subsidiando) o desalentar (añadiendo primas) la utilización de ciertos recursos o servicios.

Independientemente de la estrategia elegida, la misma hará uso de herramientas comunes para lograr el objetivo de distribución de costos, siendo las más usuales i) utilización de cuentas o proyecto; ii) etiquetado de recursos; y iii) la distribución de costos compartidos.

En primer lugar, la **utilización de cuentas o proyectos** será de vital importancia para lograr una estructura de costos que sea compatible con la distribución de costos general de la empresa. Para lograr esto se deberá asegurar una alineación y un entendimiento común entre el área de finanzas, y el equipo de *FinOps* responsable por la alocación de costos de recursos de nube pública. Será altamente recomendable y efectivo elaborar presupuestos anuales definidos contra dichas cuentas y/o proyectos, de manera de poder detectar desvíos respecto de los objetivos financieros del negocio.

Continuando con el **etiquetado de recursos**, consiste en un mecanismo que permite adjuntar *metadata* a la mayoría de los recursos de los proveedores de nube. Como su nombre lo indica, las etiquetas o *tags*, consisten en un par de <*clave, valor*> que pueden ser definidos por el usuario, y adjuntados a los recursos en cuestión. De esta manera, todos los recursos, y por ende sus costos asociados, pueden ser fácilmente filtrados por una o más claves definidas por la organización.

Resulta de vital importancia definir y alinear una estructura de etiquetado que sea funcional a la organización, definiendo, por ejemplo, un tag de "*owner*" o "proyecto" que indicará el área o centro de costos al cual será alocado el gasto que genere dicho recurso. La complejidad de dicha estructura radicará, en este caso, en los distintos valores que pueda adoptar esta clave. La lista permitida de valores deberá corresponder con la estructura organizacional de la propia empresa.

Dado que las estructuras organizacionales de las empresas no son estáticas a lo largo del tiempo, resulta importante contar con un proceso que garantice que las etiquetas sean actualizadas, o correctamente mapeadas, a medida que se produzcan cambios en la organización. Por ejemplo, es una práctica común distribuir aquellos costos de naturaleza operacional por equipo (ingeniería, QA, finanzas, RRHH, etc.). Es también común que nuevos equipos sean creados, o equipos existentes desaparezcan, se renombren o transformen. En estos casos, será necesario garantizar que la estructura de etiquetado sea mantenga consistente con los cambios organizacionales. De otra forma, la alocación de costos perderá efectividad debido al hecho que existirán recursos, generando gastos, que no tendrán dueño, y por ende, carecerán de un

responsable claro de garantizar que dichos costos tengan sentido para el negocio, y que los recursos asociados a ellos estén optimizados de forma correcta.

Cabe aclarar que la utilización de estas etiquetas no es de uso exclusivo asociado a la alocación de costos efectiva, sino que su versatilidad permite su utilización para otros ámbitos de naturaleza diversa. Usos comunes que generalmente se encuentran en las etiquetas son, entre otros: definición del ambiente al que pertenece un recurso (*test*, *staging*, *QA*, producción), el dueño del recurso (que equipo es responsable de mantenerlo y operarlo), indicadores de confidencialidad (si el recurso almacena o procesa información sensible), entre muchos otros. Las posibilidades de uso de las etiquetas son virtualmente infinitas, y dependerán de la necesidad de cada usuario y/o empresa.

Por último, la **distribución de costos compartidos** consistirá en contar con los mecanismos requeridos para poder distribuir los costos de aquellos recursos atómicos que sean utilizados por más de un proyecto o cliente interno. Un ejemplo de esto podría ser un servidor utilizado como nodo en un clúster de Kubernetes. Esta tecnología permite alojar múltiples aplicaciones de distinta naturaleza en un mismo servidor virtual. Debido a que el proveedor de nube ofrecerá la posibilidad de identificar costos a nivel de servidor, será responsabilidad del cliente desarrollar los mecanismos necesarios para poder distribuir ese costo entre las distintas aplicaciones que utilizan dicho servidor. Cabe aclarar que el uso de recursos de una aplicación puede variar enormemente respecto de otra, en el mismo servidor.

Las estrategias más comunes para lograr una distribución de costos compartidos efectiva y precisa, consistirán en identificar aquellas variables que mejor representen la utilización de recursos de infraestructura. Para el ejemplo anterior de Kubernetes, es común encontrar algoritmos de distribución de costos basados en las reservas de CPU y memoria por aplicación. Por el contrario, para casos de aplicativos *multi-tenant* que sirven múltiples clientes, suelen utilizarse cantidad de transacciones o *requests* realizados, o tráfico de red utilizado (Yousafzai, 2017).

La correcta alocación de costos de los consumos cloud fue identificada como un mecanismo fundamental por dos de los tres expertos entrevistados (Simonassi y

Pampliega). En dicho análisis se destacó la importancia, no solo de alocar correctamente los costos incurridos, sino de complementar dicha herramienta con la promoción de una cultura que intensifique la visibilidad y la transparencia, y que promueva la austeridad, prudencia y eficiencia de los recursos tecnológicos, como un pilar en los equipos de ingeniería.

10.5. Mercados Secundarios

Algunos proveedores cuentan con mercados secundarios de reservas. Si la organización utilizara uno de estos proveedores de nube (por ejemplo, AWS), existirá la posibilidad de utilizar dichos mecanismos para revender recursos reservados. Este será un mecanismo efectivo para recuperar dinero sobre aquellas reservas que fueran compradas en el pasado, pero no pudieran ser convertidas, y actualmente se encuentren subutilizadas o en completo desuso.

El mecanismo de reventa suele ser realmente simple, pero cuenta con múltiples restricciones, como por ejemplo, tiene máximos permitidos por año (medidos en dólares americanos), y limitaciones a la hora de definir el precio máximo de reventa permitido (Ambati, 2020).

Es interesante destacar el hecho de que dicho mecanismo no fue identificado ni observado en las encuestas a especialistas, ni en las entrevistas a expertos. Esto se debe principalmente a que los mercados secundarios cuentan con poco volumen transaccional a nivel global, por lo que es común encontrarse con dificultades o demoras para ejecutar una venta. Es por esto que, si bien este mecanismo puede ser un recurso válido para mitigar un error de planificación, debe considerarse que su uso es normalmente limitado y restringido a volúmenes pequeños.

CONCLUSIONES

La aparición y subsecuente adopción de la nube pública ha revolucionado la forma en que las empresas gestionan su infraestructura tecnológica, ofreciendo escalabilidad, flexibilidad y eficiencias nunca antes vistas. Sin embargo, el cambio de paradigma llegó con nuevos desafíos, como la necesidad de contar con nuevos conocimientos, la seguridad de los datos, las fuertes dependencias con distintos proveedores, y por sobre todo, la necesidad de una gestión de costos efectiva. A medida que la tecnología continúa avanzando, es fundamental que las empresas evolucionen para aprovechar al máximo las ventajas de la nube pública, sin dejar de adoptar un enfoque estratégico para mitigar los riesgos asociados a la misma (que pueden tener consecuencias devastadoras para un negocio).

La industria de los juegos móviles, como se ejemplifica en el caso de Wildlife Studios, ha experimentado un crecimiento exponencial que solo fue posible gracias a la nube pública. La capacidad de escalar rápidamente, adaptarse a las demandas del mercado y llegar a una audiencia global ha sido fundamental para el éxito de la misma. Sin embargo, también enfrenta desafíos únicos, como la necesidad de optimizar el rendimiento de los juegos en diferentes dispositivos, sin dejar de brindar una experiencia de usuario fluida, y garantizar un esquema de costos que sea compatible con el negocio.

En última instancia, la nube pública se ha convertido en un habilitador clave para la innovación y el crecimiento en diversos sectores. A medida que las empresas continúan adoptando esta tecnología, es esencial que se centren en la gestión eficiente de los costos, la seguridad de los datos y la optimización del rendimiento para maximizar sus beneficios. La colaboración entre los proveedores de nube, las empresas y sus expertos en tecnología, será fundamental para impulsar la innovación y superar los desafíos del futuro.

La correcta optimización de costos en ambientes de Nube Pública se ha vuelto un requisito fundamental (y cuasi-obligatorio) para aquellas empresas que adoptaron esta tecnología y vieron sus volúmenes operativos escalar hasta alcanzar los millones

de dólares por año en costos, siendo los equipos de *FinOps* (o *Cloud Economics*) necesarios para poder operar el negocio bajo márgenes positivos, y uno de los más fáciles de justificar a nivel de retorno de inversión.

Teniendo en cuenta los argumentos teóricos, los casos testigos descriptos en el presente trabajo, y la opinión de diversos especialistas y expertos en la materia reflejados en las respuestas a los ejercicios de encuestas y entrevistas, se identifica, con total seguridad, que esta práctica es fundamental a la hora de adoptar servicios de nube pública, y será una disciplina cada vez más necesaria (y demandada) en el futuro.

BIBLIOGRAFÍA

J. R. Storment, M. F. (2020). *Cloud FinOps: Collaborative, Real-Time Cloud Financial Management.* O'Reilly Media.

Gaca, A. (29 de March de 2023). *Company cloud migration examples: companies that migrated to the Cloud*. Obtenido de Future Processing: https://www.future-processing.com/blog/company-cloud-migration-companies-that-migrated-to-the-cloud/

Metz, C. (14 de March de 2016). *The Epic Story of Dropbox's Exodus From the Amazon Cloud Empire*. Obtenido de Wired: https://www.wired.com/2016/03/epic-story-dropboxs-exodus-amazon-cloud-empire/

Richter, F. (2024). *Cloud Infrastructure Market.* Statista.

IDC. (12 de March de 2020). *IDC*. Obtenido de Worldwide Server Market Revenue: https://www.idc.com/getdoc.jsp?containerId=prUS46132420

HGInsights. (2024). *The Google Cloud Platform Ecosystem in 2024.* HGInsights.

Jackson, K. G. (2018). *Architecting cloud computing solutions : build cloud strategies that align technology and economics while effectively managing risk.* Packt Publishing.

Orban, S. (2017). *Ahead in the cloud : best practices for navigating the future of Enterprise IT.* Createspace Independent Publishing Platform.

Iyar, S. (2007). *Why buy the cow?* Lulu.com.

Wang, B. (01 de September de 2020). *Cost Allocation Blog Series #1: Cost Allocation Basics That You Need to Know*. Obtenido de AWS Cloud Financial Management: https://aws.amazon.com/blogs/aws-cloud-financial-management/cost-allocation-basics-that-you-need-to-know/

Henshaw, K. (13 de October de 2015). *Submer*. Obtenido de Datacenter Cooling Methods. The Importance of Choosing the Right Cooling Method: https://submer.com/blog/datacenter-cooling-methods/

statcounter. (2024). *StatCounter GlobalStats*. Obtenido de Mobile Operating System Market Share Worldwide: https://gs.statcounter.com/os-market-share/mobile/worldwide

AWS. (19 de December de 2024). *Amazon Elastic Compute Cloud User Guide*. Obtenido de Reserve compute capacity with EC2 On-Demand Capacity Reservations: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html

AWS. (2013). *What is cloud computing?* Obtenido de Amazon Web Services: https://aws.amazon.com/what-is-cloud-computing/

Azure. (2022). *What is Cloud Computing?* Obtenido de Microsoft Azure: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-computing/

Google. (2016). *Google Cloud Computing*. Obtenido de Google Cloud overview: https://cloud.google.com/docs/overview/

IronSource from Unity. (2021). *IronSource.* Obtenido de Ad Mediation: https://www.is.com/glossary/ad-mediation/

Telecommunications Industry Association. (2024). *TIA Online.* Obtenido de TIA-942-C DATA CENTER INFRASTRUCTURE STANDARD: https://tiaonline.org/resource/tia-942-c-data-center-infrastructure-standard/

Uptime Institute. (2017). *Uptime Institute*. Obtenido de Tier Certification Overview: https://uptimeinstitute.com/tier-certification

Wikipedia. (2024). *Snake (1998 video game)*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Snake_(1998_video_game)

Wikipedia. (2024). *Tetris*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Tetris

Wildlife. (2020). *Who we Area*. Obtenido de Wildlife Studios Homepage: https://wildlifestudios.com/who-we-are/

Agar, J. (2017). *Turing and the Universal Machine: The Making of the Modern Computer.* Icon Books Ltd.

Formation. (2024). *A guide to lighting data centers.* Formation Data.

IEC. (01 de 2023). International Standard 61537.

Nilsson, M. a. (2014). Advantages and challenges with using hypoxic air venting as fire protection. *Fire and materials, 38(5)*, (págs. 559-575).

Haghshenas, K. S. (2023). Enough hot air: the role of immersion cooling. *Energy Informatics, 6(1)*, (pág. 14).

Pilz, K. &. (2023). Compute at Scale--A Broad Investigation into the Data Center Industry. *arXiv preprint arXiv:2311.02651.*

Andrae, A. &. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. (págs. 117-157). Challenges.

Cho, J., & Kim, B. S. (2011). Evaluation of air management system's thermal performance for superior cooling efficiency in high-density data centers. *Energy and Buildings*, vol. 43 iss. 9.

Mauricio Arregoces, M. P. (2003). *Data Center Fundamentals.* Cisco Press.

D. Gmach, J. R. (2007). Capacity Management and Demand Prediction for Next Generation Data Centers. *IEEE International Conference on Web Services* (págs. 43-50). Salt Lake City: IEEE.

Allspaw, J. (2008). *The Art of Capacity Planning.* O'Reilly.

Suwan, B. (13 de 05 de 2024). *Challenges Facing the Refurbished IT Hardware Market*. Obtenido de Inside Systems: https://insidesystems.com/blog/challenges-facing-the-refurbished-it/

Espin. (07 de 08 de 2024). *Digital Industries Nightmare: The Impacts of IT outages*. Obtenido de E-Spin Corp: https://www.e-spincorp.com/digital-industries-nightmare-the-impacts-of-it-outages/

Menascé, D. A. (2009). Understanding Cloud Computing: Experimentation and Capacity Planning. *Int. CMG conference.*

Ling, G. (04 de 02 de 2021). *Google Play Store and Apple App Store fees, (+12 other stores)*. Obtenido de App Radar: https://appradar.com/blog/google-play-apple-app-store-fees

Martens, B. W. (2012). Costing of cloud computing services: A total cost of ownership approach. *45th Hawaii International Conference on System Sciences* (págs. 1563-1572). Hawaii: IEEE.

Perry, R. H. (2009). *Force.com cloud platform drives huge time to market and cost savings.* TechRepublic.com.

X. Guerron, S. A.-D.-L.-D.-G. (2020). Taxonomy of Quality Metrics for Cloud Services. *IEEE Access, vol. 8*, 131461-131498.

Storment, J. (09 de 2015). *Apptio*. Obtenido de Red Line vs. Green Line: Reporting on Reserved Instance Coverage and Waste: https://www.apptio.com/blog/red-line-vs-green-line-reporting-on-reserved-instance-coverage-and-waste/

Dar Juan, L. (2024). *RI Utilization vs RI Coverage: Difference Between these Amazon EC2 Reserved Instance Metrics*. Obtenido de TutorialsDojo: https://tutorialsdojo.com/ri-utilization-vs-ri-coverage-difference-between-these-amazon-ec2-reserved-instance-metrics/

Novotný, A. (12 de 7 de 2022). *What is a Cache Hit Ratio and How do you Calculate it?* Obtenido de StormIT: https://www.stormit.cloud/blog/cache-hit-ratio-what-is-it/

Lee, I. (2019). Pricing schemes and profit-maximizing pricing for cloud services. *Journal of Revenue and Pricing Management, 18*, 112-122.

Krishnakumar, V. (22 de 08 de 2024). *Why Should You Consider AWS EDP for Your Organization?* Obtenido de CloudOptimo: https://www.cloudoptimo.com/blog/why-should-you-consider-aws-edp-for-your-organization/

Ambati, P. I. (2020). No Reservations: A First Look at Amazon's Reserved Instance Marketplace. *12th USENIX Workshop on Hot Topics in Cloud Computing.*

Singer, G. L. (2010). Towards a model for cloud computing cost estimation with reserved instances. *Proc. of 2nd Int. ICST Conf. on Cloud Computing.*

Awati, R. (05 de 2024). *AWS Reserved Instances*. Obtenido de TechTarget: https://www.techtarget.com/searchaws/definition/AWS-Reserved-Instances-Amazon-Reserved-Instances

Ravhon, R. (22 de 07 de 2024). *AWS Reserved Instances: Pros/Cons, Types & Use Cases*. Obtenido de FinOut: https://www.finout.io/blog/aws-reserved-instances-pros-cons-types-use-cases

Isobe, G. A. (2021). An evaluation of the discount rates for spot instances on Amazon EC2. *Bulletin of Networking, Computing, Systems, and Software, 10(1)*, 27-29.

Xie, J. (2017). *What to do During a Spot Instance Interruption?* Obtenido de MemVerge: https://memverge.com/blog/what-to-do-during-a-spot-instance-interruption/

Lorido-Botran, T. M.-A. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. *Journal of grid computing, 12*, 559-592.

Erradi, A. &. (2020). Online cost optimization algorithms for tiered cloud storage services. *ournal of Systems and Software, 160*, 110457.

Degani, A. (30 de 04 de 2023). *AWS Data Transfer Pricing: Hidden Network Transfer Costs and What to Do About Them*. Obtenido de NetApp: https://bluexp.netapp.com/blog/aws-cvo-blg-aws-data-transfer-costs-solving-hidden-network-transfer-costs

Yousafzai, A. G. (2017). Cloud resource allocation schemes: review, taxonomy, and opportunities. *Knowledge and information systems, 50*, (págs. 347-381).

Mohammed, B. (21 de 08 de 2024). *SAN vs NAS Storage: A Detailed Comparison*. Obtenido de VPSServer: https://www.vpsserver.com/san-nas-network-storage/

Tony Chan. (2024). *AWS EDP 2024 Negotiation Guide*. Obtenido de Cloudforecast: https://www.cloudforecast.io/blog/aws-edp-guide/

ANEXOS

ANEXO A: ENCUESTAS

A.I Preguntas de encuesta sobre el uso de nube pública en empresas de tecnología

Uso de Nube Pública en Empresas de Tecnología

Encuesta anónima para el estudio de adopción y optimización de servicios de Nube Pública

Acceder a Google para guardar el progreso. Más información

* Indica que la pregunta es obligatoria

1.- ¿Está su organización actualmente utilizando servicios de nube publica? *

- ◯ Sí, 100% nube
- ◯ Sí, en un modelo híbrido entre Nube Pública e Infraestructura On-Prem
- ◯ No, 100% Infraestructura On-Prem

2.- ¿Hace cuántos años que su organización utiliza servicios de nube? *

- ◯ Menos de 1 año
- ◯ Entre 1 y 3 años
- ◯ Entre 3 y 5 años
- ◯ Entre 5 y 10 años
- ◯ Más de 10 años

3.- ¿Qué nivel de madurez cree Ud. que tiene su organización respecto de la utilización de servicios de nube pública? *

- ◯ Inicial
- ◯ Fundacional
- ◯ Definido
- ◯ Medido
- ◯ Optimizado

4.- ¿Qué proveedores de nube utiliza actualmente? *

- ☐ AWS
- ☐ GCP
- ☐ Azure
- ☐ Otro

5.- ¿Tiene su organización contratos de largo plazo firmados con alguno de sus proveedores de nube? *

- ◯ No
- ◯ Sí, menor o igual a 2 años
- ◯ Sí, menor o igual a 3 años
- ◯ Sí, menor o igual a 5 años
- ◯ Sí, mayor a 5 años

6.- ¿Cuánto gasta su organización anualmente en servicios de nube publica? *

- ◯ USD 10k por año, o menos
- ◯ USD 100k por año, o menos
- ◯ USD 500k por año, o menos
- ◯ USD 1M por año, o menos
- ◯ Más de USD 1M por año

7.- ¿Está preocupado por los costos de nube publica de su organización? *

- ◯ No, no estoy preocupado en absoluto
- ◯ Levemente preocupado
- ◯ Preocupado
- ◯ Muy preocupado
- ◯ Extremadamente preocupado

8.- ¿Tiene su organización algún tipo de iniciativa en curso para optimizar los costos de nube pública? *

- ◯ Sí
- ◯ No

9.- ¿Tiene su organización personal dedicado a optimizar los costos de nube pública? *

- ◯ No
- ◯ Sí, 1 persona
- ◯ Sí, 5 personas o menos
- ◯ Sí, 10 personas o menos
- ◯ Sí, 10 personas o más

10- ¿Qué tan efectiva considera que ha sido su organización en optimizar los costos de utilización de nube? *

- ◯ Muy poco efectiva
- ◯ Algo efectiva
- ◯ Muy efectiva, pero aún queda trabajo por hacer
- ◯ Extremadamente efectiva, y ya no quedan oportunidades significativas por capturar

11.- ¿Cuánto más alto considera que serían sus costos de nube pública, si no hubiera realizado ninguna tarea de optimización (incluyendo negociación de contratos y cualquier proyecto cuyo objetivo sea aumentar la eficiencia operativa)? *

- ◯ Iguales a los actuales
- ◯ 10% mayores, o menos
- ◯ 25% mayores, o menos
- ◯ 50% mayores, o menos
- ◯ Mayores a un 50%

12.- ¿Consideraría dejar de utilizar servicios de nube pública? *

- ◯ Sí
- ◯ No
- ◯ Quizá en el futuro

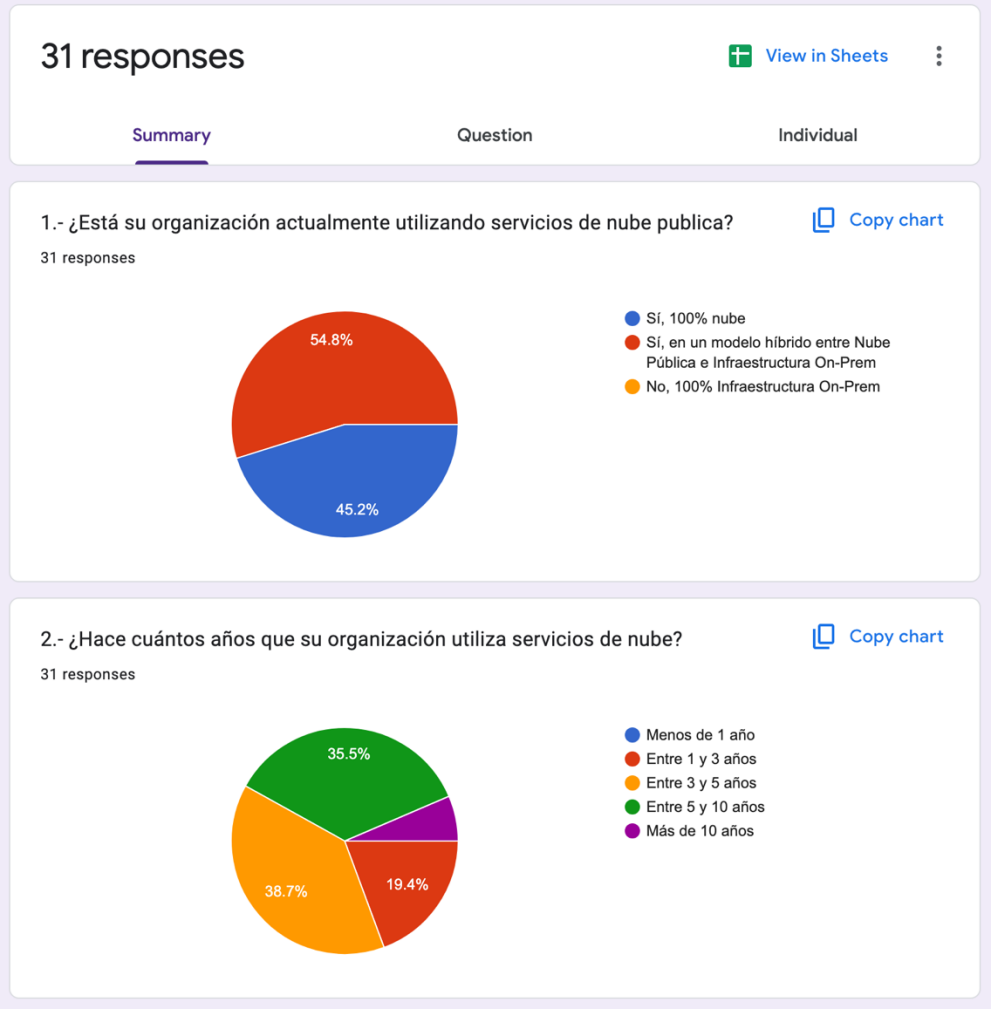
13.- ¿Cuáles son los mecanismos de eficiencia que mayores resultados obtuvieron?

Tu respuesta

Enviar Borrar formulario

Nunca envíes contraseñas a través de Formularios de Google.

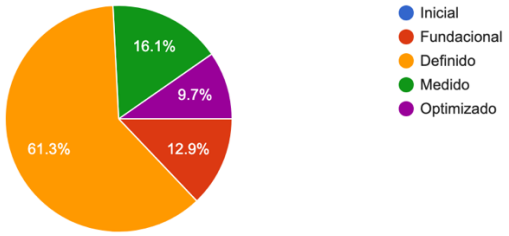
A.II Respuestas de encuesta sobre el uso de nube pública en empresas de tecnología



3.- ¿Qué nivel de madurez cree Ud. que tiene su organización respecto de la utilización de servicios de nube pública?

Copy chart

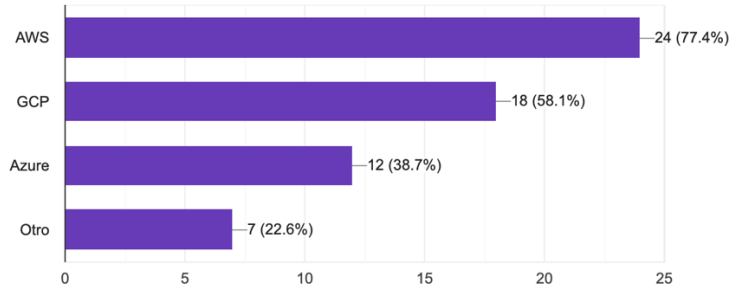
31 responses



4.- ¿Qué proveedores de nube utiliza actualmente?

Copy chart

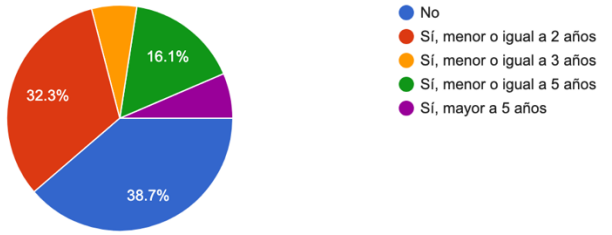
31 responses



5.- ¿Tiene su organización contratos de largo plazo firmados con alguno de sus proveedores de nube?

Copy chart

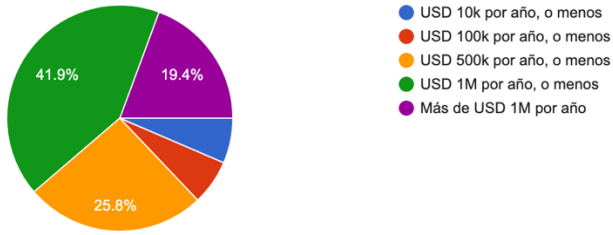
31 responses



6.- ¿Cuánto gasta su organización anualmente en servicios de nube publica?

Copy chart

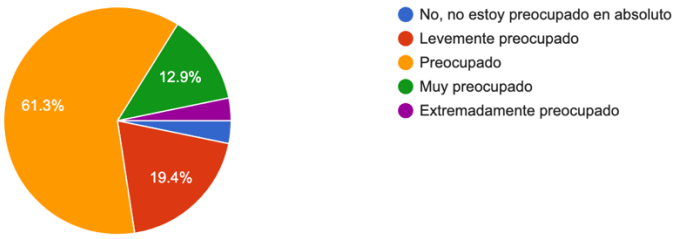
31 responses



7.- ¿Está preocupado por los costos de nube publica de su organización?

Copy chart

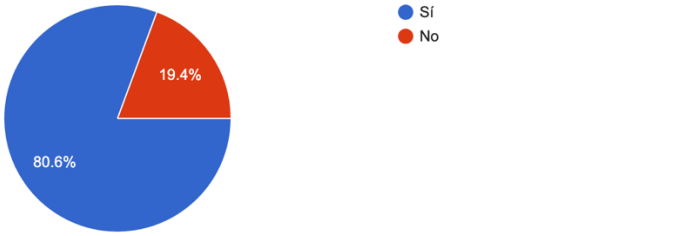
31 responses



8.- ¿Tiene su organización algún tipo de iniciativa en curso para optimizar los costos de nube pública?

Copy chart

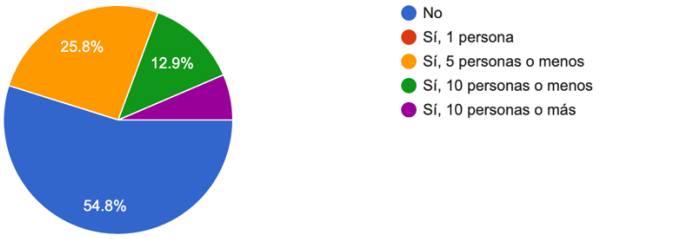
31 responses



9.- ¿Tiene su organización personal dedicado a optimizar los costos de nube pública?

Copy chart

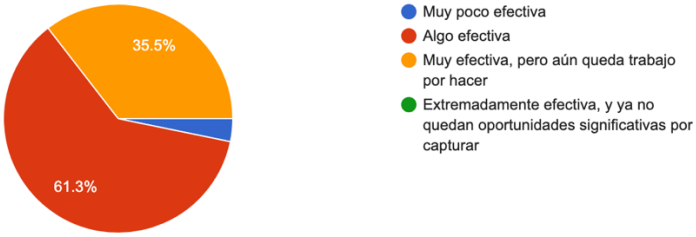
31 responses



10- ¿Qué tan efectiva considera que ha sido su organización en optimizar los costos de utilización de nube?

Copy chart

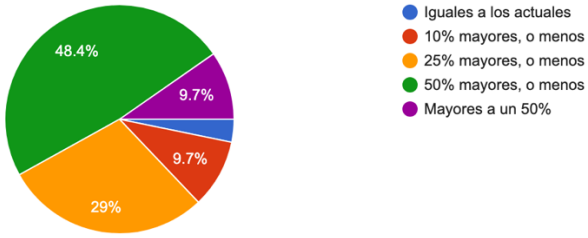
31 responses



11.- ¿Cuánto más alto considera que serían sus costos de nube pública, si no hubiera realizado ninguna tarea de optimización (incluyendo negociación de contratos y cualquier proyecto cuyo objetivo sea aumentar la eficiencia operativa)?

Copy chart

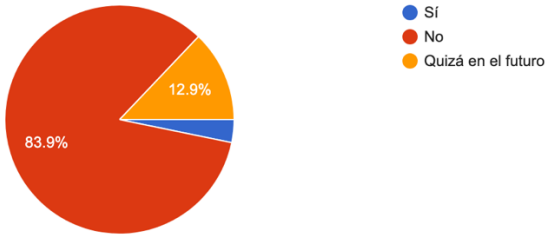
31 responses



12.- ¿Consideraría dejar de utilizar servicios de nube pública?

Copy chart

31 responses



13.- ¿Cuáles son los mecanismos de eficiencia que mayores resultados obtuvieron?

8 responses

Reservas y Spot Instances

reservas

Por lejos, la utilizacion de instancias spot.
Luego diria Savings Plans
Siguiente, rightsizing de recursos
Por ultimo hacer housekeeping, y eliminar o apagar cosas sin usar.

Empezamos a ver una disminucion acalerada de los cstos en cuanto seteamos mejorar niveles de eficiencias como un objetivo de negocio.
Adicionalmente, nos apalancamos mucho en el soporte del equipo tecnico de AWS

savings plans, descuentos comerciales, perseguir a los developers

budgets y control estricto

sponsorship del equipo ejecutivo + seguimiento constante

ANEXO B: ENTREVISTAS

B.I Preguntas de entrevistas a expertos en infraestructura y nube pública

1) ¿Cuáles son las principales ventajas de nube pública vs infraestructura tradicional?
2) ¿En qué circunstancias recomendarías un modelo vs el otro?
3) ¿Cuáles son las barreras de entrada más significativas para comenzar a utilizar servicios de nube pública?
4) ¿Cuáles son las principales dificultades que se encuentran en la adopción de nube pública en el mediano y largo plazo?
5) ¿Cuál es tu predicción respecto del futuro de la infraestructura tradicional? ¿Pierde/Gana terreno vs nube pública? ¿Desaparece?
6) ¿En tu experiencia personal, que tan importante es la correcta gestión de los costos (incluyendo mecanismos de optimización), dentro de la estrategia de adopción de nube pública?
7) ¿Cuáles son los principales mecanismos para garantizar una operación eficiente en la nube?
8) ¿Cuáles son las principales métricas utilizadas para medir la eficiencia en la nube?
9) ¿Considera que la práctica de *FinOps / Cloud Economics* está suficientemente desarrollada en empresas de tecnología en Latinoamérica?
10) ¿Qué recomendarías a aquellas empresas que estén evaluando construir la práctica de FinOps dentro de su organización?
11) ¿Cuáles dirías que son los principales problemas o riesgos que aún no tienen una solución efectiva dentro de la utilización de servicios de nube?

B.II Respuestas de entrevistas a expertos en infraestructura y nube pública

B.II.I Respuestas de *Darío Simonassi* (Experto en Plataformas de Ingeniería, CEO, Null Platform)[8]

1) La lista es larga. Destaco algunas obvias, como la flexibilidad del cloud, y la simplicidad para levantar recursos con extrema sencillez y velocidad. Resulta realmente barato para proyectos chicos, con un bajísimo *overhead*, incluso gratis en algunos proveedores que ofrecen *tiers* de prueba sin costo, ilimitados en el tiempo.

   Adicionalmente, existen otras ventajas que son menos reconocidas, como la innovación continua que introducen los grandes proveedores de nube, poniendo a disposición, de forma permanente, nuevos servicios y mejoras constantes en sus plataformas, que pueden ser fácilmente aprovechadas por cualquier usuario.

   Por último, pero no menos valioso, los servicios de nube pública de un mismo proveedor se destacan por la facilidad de integración: conectar una base de datos con un servidor, o asignar permisos a un servicio existente es extremadamente sencillo, con una experiencia de usuario muy pulida. Es realmente muy difícil, y por demás costoso, replicar este tipo de experiencias con el modelo de infraestructura propia.
2) El modelo de nube pública es recomendable cuando la agilidad y la velocidad son claves, y cuando los costos fijos deben mantenerse bajos.

   El modelo de infraestructura tradicional es más compatible con grandes empresas que quieren minimizar sus costos operativos, haciendo *insourcing* de la gestión de infraestructura, o con aquellas empresas que tengan requerimientos específicos de negocio.
3) La migración de sistemas *legacy* suele ser un desafío para las organizaciones, donde las aplicaciones deben ser re-escritas para funcionar de forma correcta en un entorno cloud. Adicionalmente, para aquellas operaciones que requieran de tiempos de latencia extraordinariamente bajos, las migraciones de aplicaciones deben estar acompañadas por migraciones de datos, lo que

[8] https://www.linkedin.com/in/simonassiluisdario/

muchas veces complejiza este tipo de transformaciones, convirtiéndolos en verdaderos retos.

Por último, los aspectos regulatorios generalmente son un obstáculo para aquellas empresas que operan en países o industrias fuertemente regulados, lo que los obliga a tener mucho control sobre la infraestructura sobre la que montan sus servicios.

4) Sin dudas, mantener el gobierno, el control y los estándares sobre la infraestructura y los servicios desplegados. La adopción del cloud abre muchísimas opciones para crear y configurar servicios diseñados para cumplir objetivos similares. Si la organización no crea los mecanismos y procesos necesarios para mantener un estándar en su ecosistema, pronto se encontrará con un ambiente sumamente heterogéneo, que será muy difícil de mantener y expandir de forma sostenible.
5) El modelo de infraestructura clásica no desaparecerá, pero sin lugar a dudas seguirá perdiendo terreno versus el modelo de cloud. El modelo *on-prem* será una opción costo-efectiva para aplicaciones específicas, pero generalmente una muy inferior para la mayoría de los casos.
6) Es fundamental para crecer de forma controlada, especialmente en aquellas empresas que tengan limitaciones presupuestarias. Crear una cultura organizacional con métricas claras, donde todos los equipos tengan disciplina y se sientan responsables por los objetivos de negocio, incluyendo los financieros, es fundamental para garantizar que los costos operativos se mantengan en niveles sanos, y respeten los ratios operativos recomendados de la industria.
7) Considero que para mantener una operación eficiente en la nube se debe contar con solidos mecanismos centralizados para garantizar la eficiencia en aquellas oportunidades que puedan realizarse sin el involucramiento directo de los desarrolladores. Por ejemplo, tener una estrategia de reservas óptima. En segundo lugar, contar con los mecanismos para identificar aquellas ineficiencias atribuibles al accionar de los desarrolladores, y a partir de estos, generar un *loop* que incluya esfuerzo de los equipos que son dueños de las aplicaciones ineficientes, de forma de corregir o capturar aquellas oportunidades detectadas. Por ejemplo, *rightsizing* de recursos, o la implementación de *lifecycle policies* para la información almacenada. Por

último, es importante contar con las herramientas que permitan acercar la información de consumos a todos los niveles de la organización, de forma de habilitar a los distintos equipos a entender si el valor de las aplicaciones que desarrollan justifica su costo actual en infraestructura.

8) En el pasado, me ha dado muy buen resultado crear y monitorear métricas que relacionen los costos de infraestructura con el tamaño del negocio. Por ejemplo, entender el costo total por usuario concurrente, o el costo total por operación. De esta forma, resulta muy sencillo empujar a los equipos técnicos a que los costos unitarios de infraestructura sean cada vez más bajos, a medida que se logra escala y se itera sobre la eficiencia.
9) Creo que la mayoría de las empresas aún no entienden completamente el valor de invertir recursos en un equipo de FinOps de manera temprana. Usualmente, se delegan estas tareas en los equipos de infraestructura de forma débil, y muchas veces se crean objetivos que pueden ser ineficaces, o incluso contradictorios. A la larga, si se descuida la práctica de FinOps, es el equipo de finanzas el que ejerce presión en la organización para priorizar la eficiencia operativa de los recursos de infraestructura.
10) Diría que la mayoría de las empresas que consideran esta práctica es porque ya la necesitan. Por ende, que sean ágiles y decididos a la hora de avanzar. Además, haría hincapié en empezar por ganar visibilidad respecto de dónde y en qué servicios se gasta el dinero, e identificar qué función cumplen los mismos para el negocio. Idealmente, cada dólar gastado debe tener un dueño responsable de justificar ese gasto, puertas adentro.
11) Monitorear costos en tiempo real sigue siendo un desafío con la mayoría de los proveedores. Esto significa que es muy difícil detectar picos de consumo, que muchas veces no responden a una necesidad real de negocio, en el momento que ocurren. Esto se debe a que muchos proveedores de cloud aún tienen un *delay* muy importante, muchas veces de muchas horas, entre el consumo efectivo de un recurso, y la facturación del mismo.

    Adicionalmente, sigue siendo costoso construir capas de abstracción que disminuyan el acoplamiento entre el cliente y el proveedor cloud, y por ende genere independencia del mismo. Esto genera dificultades a la hora de migrar hacia otras soluciones (por ejemplo, infraestructura tradicional), u otros proveedores, dado que la mayoría de los cambios requerirá modificaciones en

la lógica aplicativa; lo cual es extremadamente costoso en términos de esfuerzo.

B.II.II Respuestas de *Juan Martin Pampliega* (Experto en Data, CEO, Mutt Data)[9]

1) - Reducción drástica del *time-to-market*: los servicios de cloud permiten una agilidad superior, al combinar infraestructura virtualmente ilimitada, con complejos servicios de tecnología gestionados, disponibles en cuestión de segundos.

   - Globalización inmediata: para aquellas necesidades donde se requiera desplegar infraestructura en distintos países, regiones o continentes, ya sea por cuestiones de latencia o regulaciones, la nube pública permite provisionar infraestructura y servicios en prácticamente cualquier región del mundo, en instantes.

   - Simplicidad: los proveedores de nube abstraen de forma excelente la complejidad de operar un Datacenter, con todo lo que ello implica, e incluso servicios como bases de datos, orquestación de clústeres de cómputo, servicios de mensajería, entre otros. Esta simplicidad reduce enormemente los requerimientos de inversión y conocimiento de los clientes que utilizan estos servicios.

2) Para aquellos casos de uso donde no haya una excelente razón para no hacerlo, la recomendación por defecto será utilizar la nube pública. Hay casos específicos donde la infraestructura tradicional suele ser conveniente, o incluso mandatoria. Si tuviera que nombrar ejemplos, diría que el sector bancario es un gran exponente de esto, aunque esta regulación está siendo cada vez más flexible en muchos países.

   También existen algunos casos de nicho, donde utilizar infraestructura propia puede tener mucho sentido desde un punto de vista económico: hace unos 2 años aproximadamente, ayudamos a un cliente *cloud-native* a armar una granja de servidores Mac en sus oficinas, dado que los costos de hacerlo en la nube eran 4x mayor.

[9] https://www.linkedin.com/in/juan-martin-pampliega/

3) - Seguridad y confianza: si bien en la mayoría de los casos no está justificado desde un punto de vista técnico, muchas organizaciones tradicionales aún tienen preocupaciones acerca de la seguridad de sus datos en la nube.

   - Falta de un plan estratégico: sin una estrategia clara, la adopción de la nube puede ser ineficiente y costosa. Es muy común ver empresas que migran sus aplicaciones con una estrategia de *shift & lift*. Si bien es una estrategia válida para iniciar el *journey* hacia la nube, es fundamental entender que, si no se repiensa la arquitectura de las aplicaciones para correr en un entorno cloud, resultará muy difícil obtener beneficios que superen los costos de una hipotética migración.

4) - Optimización continua: se requiere de esfuerzo y dedicación constante para garantizar una operación óptima en la nube.

   - Fragmentación tecnológica: Es muy común ver empresas que adoptan la nube de manera parcial (una parte en la nube, otra en sus Datacenters). Si bien es la forma de empezar, habrá que ser consciente que en la medida que se tenga un ambiente hibrido, incrementará la complejidad de la operación, al tener que conectar las partes, poseer conocimientos de tecnologías y procesos para cada ambiente, y se aumentará la cantidad de puntos de falla en la operación.

   - Evolución tecnológica constante: la tecnología avanza y todos los proveedores de nube aplican actualizaciones que en muchos casos traen *breaking changes*. Para ser justos, en la mayoría de los casos, la notificación suele venir con muchos meses de anticipación, pero a diferencia de la infraestructura tradicional, donde el dueño tiene control absoluto, habrá que mantenerse actualizado de forma permanente para evitar incidentes.

5) Para jugadores pequeños y medianos, la infraestructura tradicional dejará de ser una opción. De hecho, es realmente difícil hoy en día encontrar empresas que elijan no estar en la nube. Para jugadores grandes, mi predicción es que los modelos híbridos ganarán cada vez más terreno, principalmente por una cuestión de optimización de costos, complementando a la nube para casos específicos.

6) Para empresas pequeñas o en etapas iniciales, poco importante. Para empresas que tengan un costo significativo en recursos de nube, muy importante (cuanto más grande el gasto, más importante la gestión de costos).

Para empresas realmente grandes, como MercadoLibre, Nubank, TiendaNube, etc., absolutamente crítico. En estos casos, no gestionar adecuadamente los costos y la eficiencia, tiene impactos medidos en los millones de dólares por año.

7) - Contar con un equipo responsable por la eficiencia en la nube, idealmente dedicado.

   - Generar una cultura de austeridad, prudencia y eficiencia respecto de los recursos tecnológicos

   - Negociar contratos comerciales de forma agresiva con el/los proveedores de cloud elegidos.

   - Contar con mecanismos de etiquetado que permitan la correcta asignación de costos

8) - Niveles de utilización (CPU, Memoria, Discos, etc.), para entender si los recursos provisionados tienen el tamaño y capacidad efectivamente requeridos.

   - Maximizar la utilización de recursos spot en aquellas aplicaciones *stateless* (idealmente, la mayoría) que no sean criticas para el negocio.

   - Utilización y cobertura de *Committed Resources*, para entender la eficiencia en la modalidad de contratación de los recursos de cómputo, buscando un balance óptimo entre compromisos y descuentos.

9) No. Hay empresas líderes, como MercadoLibre, que lo hacen muy bien. Sin embargo, la mayoría de las empresas no cuentan con el conocimiento ni la cultura adecuada para garantizar una utilización eficiente de la nube pública. Sin embargo, en los últimos años la conciencia y el nivel general en esta práctica mejoró considerablemente. Estimo que esta tendencia se mantendrá en el futuro.

10) - Evaluar y analizar herramientas que simplifican enormemente la identificación y captura de las oportunidades “genéricas”. Como por ejemplo spot.io, o cloudhealth, ente muchas otras. Generalmente, estas plataformas cobran un porcentaje sobre lo ahorrado, por lo que siempre tendrán un RoI positivo.

    - Integrar o acercar los equipos de tecnología con finanzas: existen muchas herramientas y procesos dentro de los departamentos de finanzas que son

realmente útiles para la práctica de FinOps, como por ejemplo definición y seguimiento de presupuestos.

- Contratar gente con experiencia en el área: hace algunos años era prácticamente imposible, ya que ninguna empresa regional tenía conocimientos ni equipos en FinOps. Hoy el panorama el destino y es posible encontrar especialista que acelerarán considerablemente la adopción de buenas prácticas en esta área.

11) - Transparencia en costos: aún sigue siendo realmente difícil estimar con precisión los costos de una solución de nube. Generalmente, existen costos ocultos que son realmente muy difíciles de anticipar si no se cuenta con experiencia extensiva en el o los servicios en cuestión. Los proveedores de cloud aún tienen una deuda para con este problema.

- Despliegues en la nube en regiones restringidas: operar en países restringidos como China, aún sigue siendo un desafío debido a las regulaciones locales particulares.

B.II.III Respuestas de *Rodrigo Bustos* (Experto en Infraestructura, Cloud & Platform Senior Manager, MercadoLibre)[10]

1) En primer lugar, diría que la flexibilidad que ofrece la nube pública para ajustar los recursos en tiempo real, según las necesidades del negocio.

   En segundo lugar, las bajísimas barreras de entrada que ofrece la nube pública: cualquier persona puede *hostear* sus aplicaciones en la nube, incluso con inversiones cercanas a $0.

2) Yo diría que la nube es ideal para proyectos personales, *startups* y *scale-ups*, donde mantener la complejidad baja y la velocidad alta es esencial. La nube también es ideal para aquellos *workloads* que tengan una demanda de recursos muy variable. En estos casos, resulta extremadamente conveniente poder escalar los recursos en los momentos de alta carga, y des-escalarlos cuando la demanda baja, provisionando únicamente los recursos necesarios en todo momento.

   La infraestructura tradicional tiene como ventaja el control. Para aquellas empresas que tengan necesidad de un estricto control sobre su operación

[10] https://www.linkedin.com/in/roviati

tecnológica, gestionar sus propios Datacenters puede ser la única solución. Ejemplos de esto son empresas que tengan regulaciones o requerimientos muy estrictos sobre seguridad o *compliance*, entre otros.

3) Si bien las barreras de entrada suelen ser bajas, es necesario contar con el conocimiento para poder operar servicios de nube pública. El know-how requerido es distinto al de un modelo de infraestructura tradicional, y no contar con el mismo puede generar problemas complejos en el futuro.

   Por otro lado, es común en las empresas encontrar resistencia al cambio cuando se introducen nuevos paradigmas. En estos casos, es importante trabajar en la parte cultural para asegurar una transformación efectiva.

4) En el corto plazo, la gestión de los costos en nube pública casi siempre es un desafío que produce mucho dolor en las organizaciones. Sin una estrategia y un monitoreo adecuado, los costos pueden escalar rápidamente hasta cifras muchas veces impensadas.

   En el largo plazo, el *lock-in* con los proveedores de nube suele ser un factor de riesgo importante. Es muy sencillo utilizar los servicios que los proveedores de cloud tienen en sus catálogos, y si uno no planifica con cuidado, luego de varios años resulta realmente muy caro siquiera evaluar una migración hacia otro proveedor.

5) Estimo que la infraestructura tradicional seguirá perdiendo terreno vs nube pública. Un signo de ello es cómo la banca tradicional ya está lentamente migrando hacia cloud. Sin embargo, estoy convencido que la infraestructura tradicional seguirá siendo relevante en algunos nichos, como por ejemplo, ambientes fuertemente auditados o que requieran de altísima seguridad, y para aquellos players que sean realmente grandes y decidan volver parcial o totalmente a administrar su propia infraestructura, debido a que poseen la escala para que sea económicamente conveniente.

6) Absolutamente crucial para la sostenibilidad de largo plazo. Una mala gestión de los mismos puede convertir a la nube en un gasto descontrolado en lugar de una solución eficaz. Aquellos actores que opten por utilizar extensivamente la nube deberán tener una estrategia clara para mantener a raya los costos, con un proceso iterativo y constante que garantice la eficiencia en la utilización de los recursos.

7) Encuentro que los principales mecanismos son tener un monitoreo permanente, y en tiempo real, de los costos; sumado a los mecanismos que prevengan de forma proactiva configuraciones ineficientes, y procesos automáticos que detecten, y en lo posible corrijan, los desvíos que pudieran ocurrir.
8) Utilización de recursos, detectando sobre-provisionamientos y recursos “zombies”. Contar con métricas de costos por proyecto o área, de manera de poder evaluar el retorno de las distintas iniciativas o unidades de negocio.
9) En la gran mayoría de las empresas, no. Aclarando que en empresas pequeñas normalmente no es necesaria. Para el resto, mi experiencia es que la gran mayoría de las compañías deja dinero sobre la mesa, al no dedicar suficiente esfuerzo a las prácticas de *FinOps*.
10) Identificar, a alto nivel, el tamaño de la oportunidad a capturar. Luego, empezar con un equipo pequeño: generalmente, los primeros pasos en optimización suelen ser sencillos, pero con un retorno muy alto. Una vez que se tenga experiencia en la práctica, iterar, y evaluar el tamaño ideal del equipo, de acuerdo con el nivel de oportunidad de ahorro, que será proporcional a la complejidad y tamaño de la organización de tecnología.
11) Creo que los principales problemas con la utilización de nube pública que preocupan a la mayoría de las organizaciones son 1) el *lock-in* con los distintos proveedores; y 2) la gestión eficaz de ambientes multi-cloud.