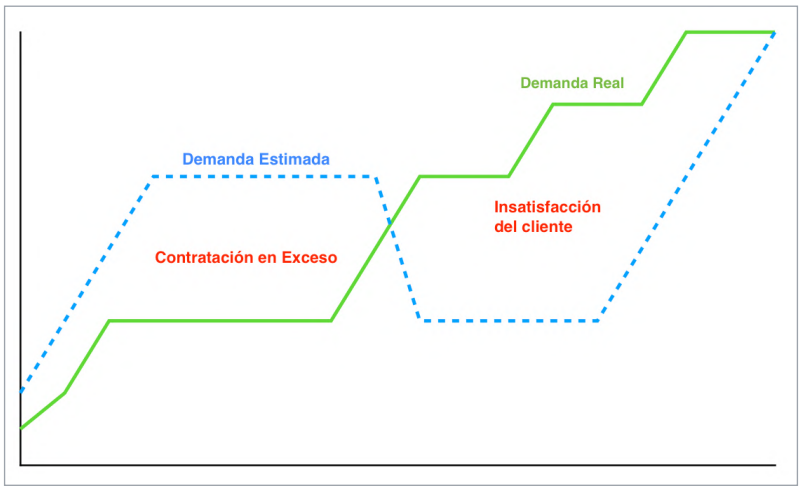
Gestión y optimización de costes en AWS

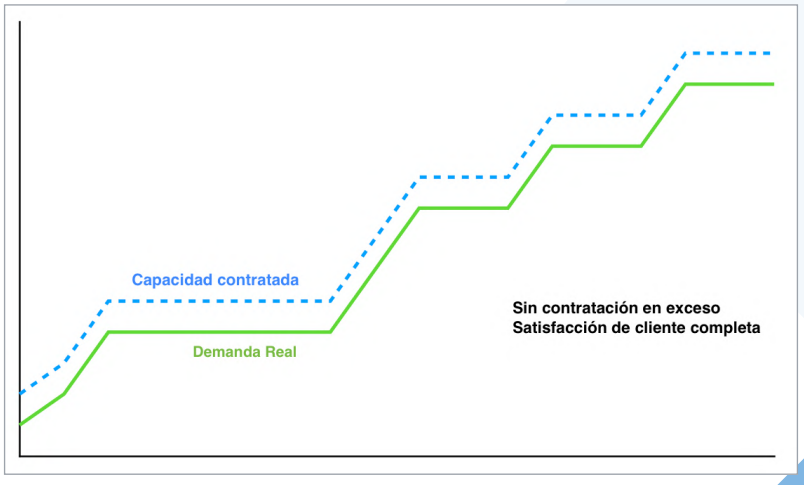
En este artículo haremos un repaso a las diversas opciones de monitorización y optimización de costes operativos que ofrece Amazon Web Services.

1/ Ajustar la capacidad a la demanda

Una de las principales causas de un sobre-coste inesperado es **la contratación de capacidad en exceso respecto a la demanda real de los sistemas**. Esto suele ser a causa del modelo de trabajo clásico llevado a cabo en **instalaciones OnPremise**, en las que la falta de elasticidad y la lentitud de los procesos de ampliación de capacidad derivan en la provisión de capacidad para atender el pico de demanda estimado durante el ciclo de vida del hardware:



Sin embargo, en un entorno elástico como la infraestructura de Amazon Web Services no existen estas restricciones y por tanto **el margen entre la capacidad contratada y la demanda real puede ajustarse de forma prácticamente continua**, reduciendo significativamente el coste de infraestructura durante los periodos de baja demanda de capacidad. Adicionalmente, los recursos recuperados como resultado de este ajuste pueden reinvertirse para solicitar demanda adicional por encima de las previsiones durante periodos punta, garantizando la continuidad del negocio y la satisfacción de los usuarios:



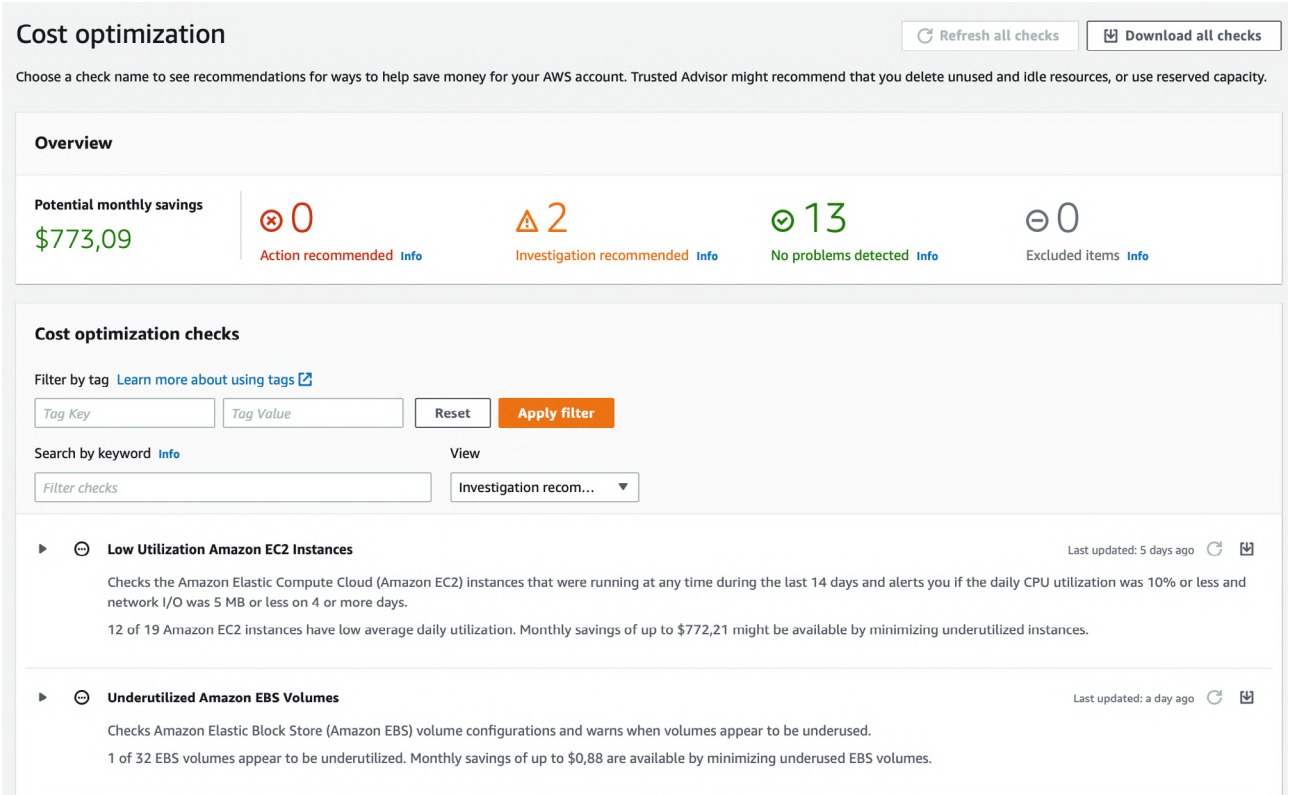
A continuación enumeraremos **algunas buenas prácticas para ajustar correctamente la capacidad contratada a la demanda del negocio en AWS**:

Identificación recursos abandonados

Debido a la naturaleza elástica de la plataforma y la facilidad a la hora de provisionar nuevos recursos de infraestructura, es habitual que muchos de estos recursos permanezcan activos incluso después de finalizar su ciclo de vida productivo, bien sea por olvido de un operador o por necesidades urgentes del negocio que no permiten dedicar el tiempo requerido a las tareas de investigación necesarias para esta tarea.

Para facilitar la identificación de recursos abandonados nuestra recomendación es utilizar la información proporcionada por **AWS Trusted Advisor.** Este servicio incluye múltiples recomendaciones para el seguimiento de buenas prácticas, entre los que se incluyen la reducción de costes operativos. Dentro de la categoría de recomendaciones de control de costes tenemos múltiples recomendaciones en este ámbito:

- Balanceadores de carga sin tráfico
- Instancias RDS o EC2 con baja o nula actividad
- ElasticIPs sin asociar
- Etc.



Un vistazo rápido a los hallazgos de **Trusted Advisor** permitirán a los operadores de la plataforma identificar rápidamente recursos susceptibles de ser eliminados y con ello reduciendo costes que de otra manera no eran sino un gasto inútil de recursos.

Uno de los servicios de AWS que con más asiduidad genera una acumulación significativa de recursos abandonados o huérfanos es **Elastic Block Storage**, y más concretamente los volúmenes sin adjuntar a instancias y los snapshots inútiles que muchas veces realizan los operadores como parte de su trabajo pero que luego nunca se llegan a limpiar una vez terminada una actuación concreta. Para estos casos existen infinidad de herramientas como **aws-amicleaner** o directamente el cliente de línea de **comandos de Amazon**. Con un par de comandos pueden generarse fácilmente informes de recursos huérfanos para facilitar a los operadores las labores de borrado.

Por ejemplo, para identificar volúmenes sin adjuntar una simple búsqueda con el cliente de línea de comandos de AWS:

```
$ aws ec2 describe-volumes --filters Name=status,Values=available | jq -r
‘.Volumes[]|[.VolumeId]’
[
  “vol-50bdbc40”
]
```

aws-amicleaner por su parte proporciona informes de alto valor a la hora de detectar backups abandonados con apenas un par de comandos simples:

```
$ /usr/local/bin/amicleaner --full-report

Retrieving AMIs to clean ...
no-tags (excluded)
+-----------------------+-----------------------+--------------------------+
|         AMI ID        |        AMI Name       |      Creation Date       |
+-----------------------+-----------------------+--------------------------+
| ami-0ccd56c7661805cc1 | qualoom-www-prod.final | 2020-09-03T07:15:17.000Z |
+-----------------------+-----------------------+--------------------------+
AMIs to be removed:
+--------------------+------------+
|     Group name     | candidates |
+--------------------+------------+
| no-tags (excluded) |     1      |

+--------------------+------------+
```

```
$ /usr/local/bin/amicleaner --check-orphans

+------------------------+
|    Orphan Snapshots    |
+------------------------+
| snap-06d16076c7dd4d479 |
| snap-092f68adfa1a1df27 |
| snap-029c5267a89283cc9 |
| snap-002a5d2b8e17e3677 |
| snap-06b794a16aecf647b |
| snap-06c94eb92f5f866b0 |
| snap-0a2eb1f0b5698062e |
| snap-0ece8719e05c77136 |
| snap-0b2daa38be614f4c5 |
| snap-0b073da03fb6f4aa3 |
| snap-0d801e4ad62ca17bf |
| snap-08efc7172f881a4db |
| snap-02946c321c51dd954 |
| snap-0154c833794a78908 |
| snap-019084c74cbeedcf6 |
| snap-097527262b064ece5 |
| snap-0d5b78f603c20b761 |
| snap-0e5dfa72f012d97b5 |
| snap-0ca26cab0be16b3d8 |
+------------------------+
```

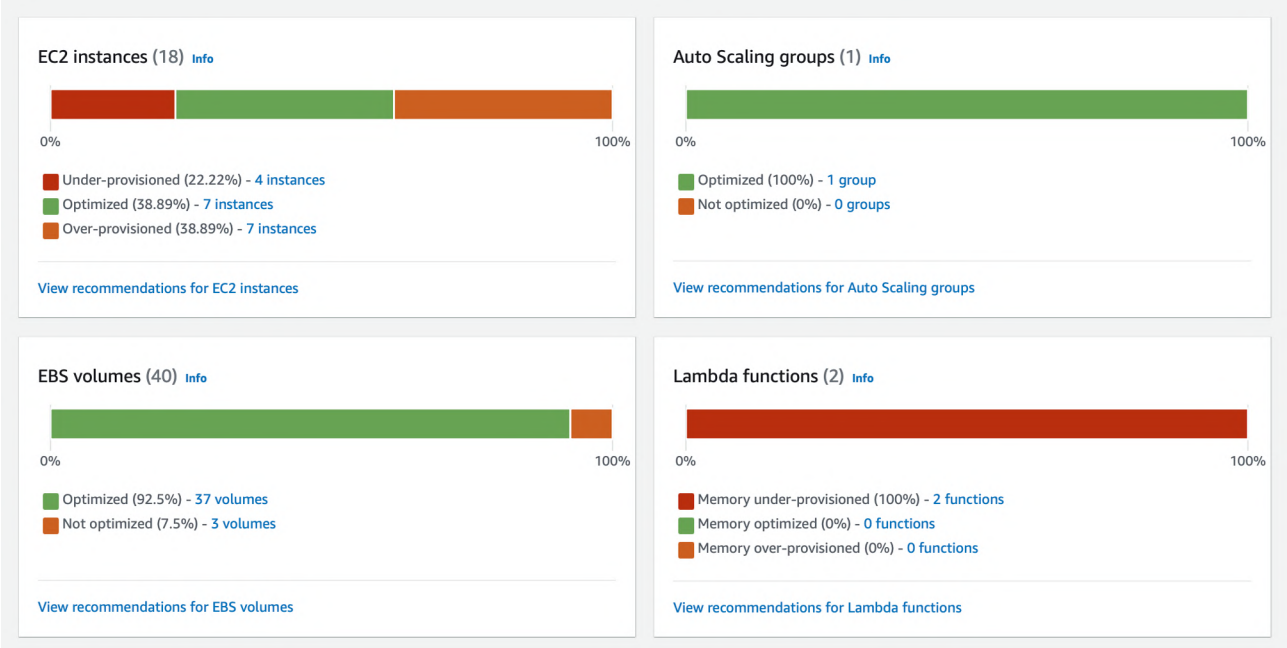
Optimización de recursos con baja utilización

Como ya hemos comentado en el punto anterior, **AWS Trusted Advisor es una excelente herramienta para tomar como punto de partida** a la hora de hacer una revisión de recursos con baja utilización y proceder con las acciones correctivas pertinentes. Sin embargo, lo recomendable siempre es realizar un análisis exhaustivo de la utilización real de los recursos contratados, ya que hay numerosas variables de grano fino que AWS Trusted Advisor no toma en cuenta.

Para ayudar en esta tarea, Amazon proporciona herramientas de propósito específico para la identificación de recursos con baja utilización que complementan a Trusted Advisor, entre las que destaca **AWS Compute Optimizer**. Este servicio **se integra con EC2, EBS, Autoscaling y Lambda** para analizar al detalle toda **la monitorización disponible en Cloudwatch** y con esta información generar recomendaciones de ajustes de capacidad con distintos grados de riesgo en base al histórico de datos analizado.

Estas recomendaciones no solo cubren las instancias sobreprovisionadas sino que, **si se detecta que hay algún recursos sobreutilizado** (lo que puede llevar a un servicio degradado y con ello la insatisfacción de sus clientes), **AWS Compute Analyzer recomendará la ampliación de capacidad necesaria** para garantizar la continuidad y estabilidad del servicio con un uso razonable de recursos:

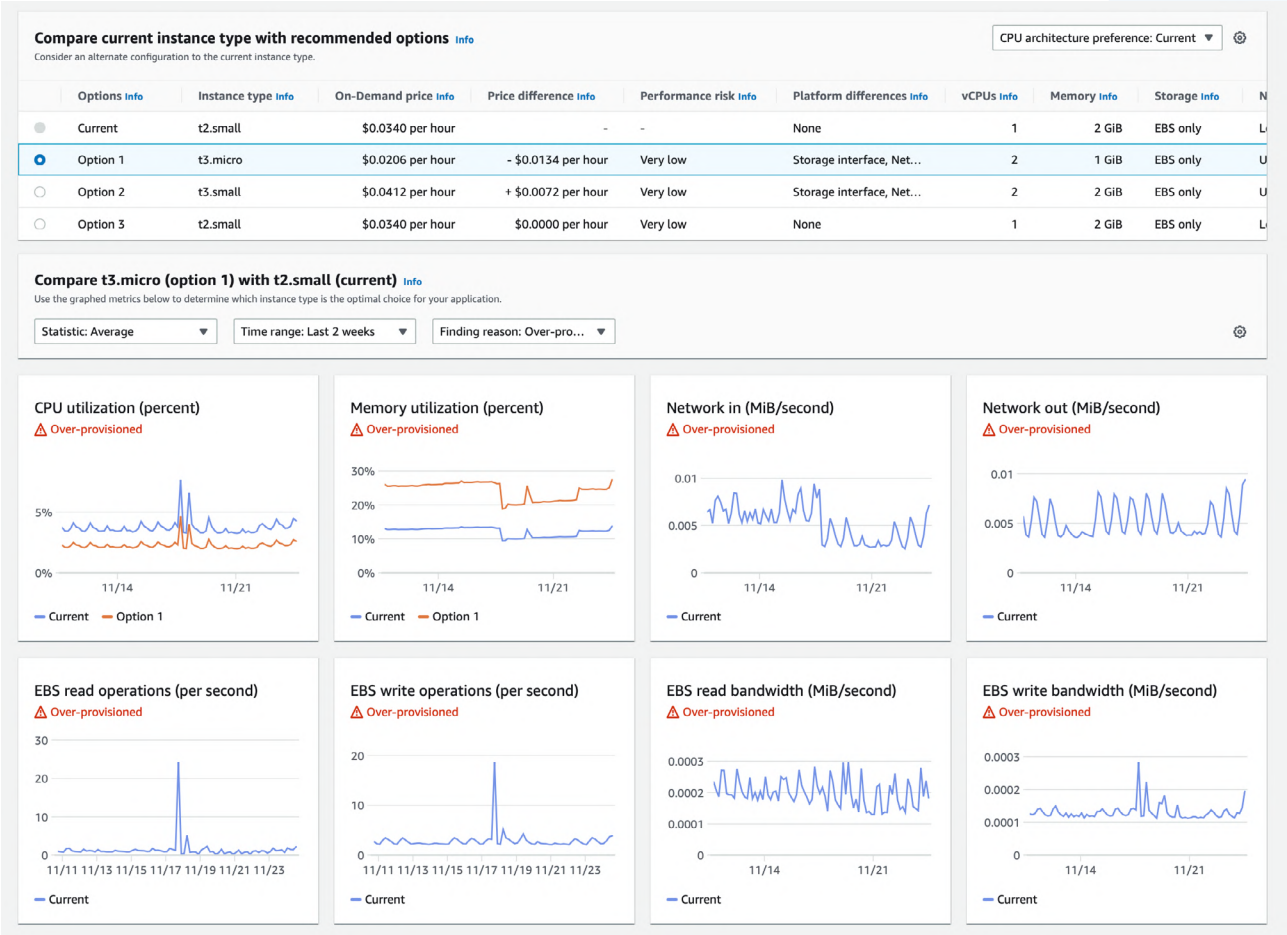
» Dashboard



» Detalle de instancias sobreprovisionadas:

Instance ID	Finding Info	Finding reasons Info	Current instance type
i-959e7d72	Over-provisioned	CPU over-provisioned, Memory over-provisioned, EBS IOPS ov…	t2.small
i-2a9890ce	Over-provisioned	CPU over-provisioned, Memory over-provisioned, EBS IOPS ov…	t2.medium
i-0e9914863b50ca47f	Over-provisioned	CPU over-provisioned, Memory over-provisioned, EBS IOPS ov…	t3a.small
i-0b7e3f9b38878a868	Over-provisioned	Memory over-provisioned, EBS IOPS over-provisioned, EBS th…	t3a.medium
i-08e10f24a817bfa57	Over-provisioned	CPU over-provisioned, EBS IOPS over-provisioned, EBS throug…	c5.large
i-061a4df0078cbe93f	Over-provisioned	Memory over-provisioned, EBS IOPS over-provisioned, EBS th…	t3a.medium
i-05931af6518c27f56	Over-provisioned	Memory over-provisioned, EBS IOPS over-provisioned	r5.xlarge

» Recomendaciones:



Por último, queremos mencionar que todos los servicios de Amazon Web Services emiten numerosas métricas de rendimiento que se recopilan para su análisis pormenorizado en **AWS Cloudwatch**. Nuestra recomendación es **realizar un seguimiento continuo del uso de recursos de cualquier plataforma de TI**, no solo por la reducción de costes sino para garantizar la continuidad del negocio.

Este servicio de monitorización **otorga todas las herramientas necesarias para el análisis, visualización y explotación de métricas**, dotando a los operadores de todos los datos necesarios para valorar el uso de la infraestructura y proponer sus propias optimizaciones, tanto para la reducción de costes operativos como para garantizar la demanda de capacidad.

Optimización de tipologías de máquina

AWS dispone de numerosos servicios de infraestructura que basan su tarificación en la disposición de máquinas virtuales, bien sea **auto-gestionadas como puede ser el caso de EC2 o bien administradas por AWS como las subyacentes a clusters Redshift, RDS, EMR o Elasticache**.

En todos estos casos, la elección de una familia y tipología de instancia adecuada para la carga de trabajo que va a desempeñar es esencial **para garantizar la continuidad del negocio sin incurrir en sobrecostes**. Para ello, Amazon cataloga sus instancias con una terminología que permite su identificación rápida

» Familias:

Familia	Identificador	Característica Principal
Propósito General	T	Instancias de bajo coste con un sistema de créditos para el uso de CPU en ráfagas
	M	4GiB de memoria por vCPU
	A	Procesadores Graviton de primera generación
	Mac	Equipos Apple Mac Mini
Alta CPU	C	2GiB de memoria por vCPU
Alta memoria	R	8GiB de memoria por vCPU
	X	16-22 GiB de memoria por vCPU
	Z	8GiB de memoria por vCPU y procesadores de alto rendimiento
	U	Hasta 55GiB de memoria por vCPU
Alto almacenamiento	H	Hasta 16TB de almacenamiento local HDD por instancia
	I	Hasta 60TiB de almacenamiento local SSD NVMe por instancia
	D	336TB de almacenamiento local HDD por instancia
Informática acelerada	G	GPUs de propósito general
	P	GPUs de propósito específico (Nvidia Tensor Core)
	Inf	Hasta 16 chips AWS Inferentia
	F	Hardware personalizable (FPGA)
	VT	Tarjetas aceleradoras para transcodificación de video

» Arquitectura:

Históricamente, todas las familias de instancia contaban con procesadores Intel. En los últimos años, **AWS ha comenzado a incorporar chips de AMD y ARM**.

Identificador	Arquitectura
i	Intel x86-64 (por defecto en familias sin indicador)
a	AMD64
g	ARM

» Variantes:

Algunas familias como M, C o R están disponibles con variantes en lo relativo a almacenamiento local de la instancia y capacidad de red adicional.

Variante	Arquitectura
d	Almacenamiento local de instancia
n	Alto rendimiento de redes
e	ARM

» Generación:

Identificador numérico de la generación de hardware dentro de una misma combinación de Familia, arquitectura y variante

» Tamaño:

Dentro de una misma familia **se suelen ofrecer distintos tamaños de máquina** con proporciones de CPU, memoria, almacenamiento local, ancho de banda de red y ancho de banda dedicado para EBS. Por lo general, cada salto de tipología supone una duplicación de capacidad respecto del tamaño anterior. Existe un tamaño especial denominado metal, que indica la posibilidad de ejecutar en modo bare-metal con toda la capacidad del hypervisor subyacente dedicada y

Por ejemplo, en el caso de la familia M6gd:

Tamaño	vCPU	Memoria (GiB)	Almacenamiento local (GiB)	Ancho de banda de red (Gbps)	Ancho de banda para EBS (Mbps)
medium	1	4	1 x 59 SSD NVMe	Hasta 10	Hasta 4750
large	2	8	1 x 118 SSD NVMe	Hasta 10	Hasta 4750
xlarge	4	16	1 x 237 SSD NVMe	Hasta 10	Hasta 4750
2xlarge	8	32	1 x 474 SSD NVMe	Hasta 10	Hasta 4750
4xlarge	16	64	1 x 950 SSD NVMe	Hasta 10	4750
8xlarge	32	128	1 x 1900 SSD NVMe	12	9000
12xlarge	48	192	2 x 1425 SSD NVMe	20	13500
16xlarge	64	256	2 x 1900 SSD NVMe	25	19000
metal	64	256	2 x 1900 SSD NVMe	25	19000

Horarios de disponibilidad

En numerosas ocasiones **el negocio solo demanda capacidad de cómputo durante intervalos horarios muy concretos o periódicos**. Entornos pre-productivos y el procesamiento por lotes son solo dos ejemplos en los que no es preciso mantener una infraestructura levantada 24x7.

Para estos casos, y gracias al modelo de facturación por hora y segundo, **la implementación de procesos que paren y arranquen la infraestructura bajo demanda o de forma programada puede ayudar significativamente a reducir costes operacionales**. AWS provee de todo tipo de APIs y SDKs para la implementación de estos procesos con las herramientas de preferencia de cada empresa. Adicionalmente, AWS proporciona **AWS Instance Scheduler**, una solución basada en servicios serverless para la implementación de procesos de apagado/arranque de instancias EC2.

NOTA: Si el horario de disponibilidad no es lo suficientemente reducido, es probable que se obtenga una mayor reducción de costes con un despliegue 24x7 y contratación de instancias reservadas. Como norma general, si el tiempo de disponibilidad supera el 75% diario no suele rentar respecto de la reserva de instancias.

Autoscaling

Numerosos servicios de AWS tienen soporte para el escalado horizontal automático mediante **Autoscaling**, también puedes ver aquí su **guía de usuario**. Esta funcionalidad **permite adaptar el número de servidores disponibles en un momento dado en función de cualquier métrica de rendimiento disponible en Cloudwatch**. Autoscaling se integra con Elastic Load Balancing para le gestión de tráfico y permite el uso de instancias de menor tamaño en favor de instancias monolíticas que habitualmente aportan capacidad en exceso al sistema.

Además de la pura reducción de costes operativos como consecuencia de estar continuamente adaptando la capacidad contratada a la demanda, **Autoscaling proporciona múltiples beneficios adicionales**:

- Posibilidad de **balancear la contratación entre el modelo bajo demanda y Spot** para equilibrar aún más los costes.
- Posibilidad de **elegir múltiples tipos de instancia para prevenir bloqueos** en el escalado ascendente por falta de capacidad dentro de una familia concreta.
- **Reemplazo automático de instancias fallidas** en caso de fallo
- Facilita la implementación de **despliegues blue/green**
- Facilita el **reparto de carga en múltiples zonas de disponibilidad**

En general, **es recomendable hacer uso de Autoscaling en cualquier aplicación**. Sin embargo, es cierto que impone ciertos cambios operacionales y de arquitectura que hacen que dar el paso requiera cierta carga de trabajo.

Arquitecturas Serverless

Muchos casos de uso son susceptibles de refactorizarse **para migrar a servicios sin servidores de Amazon**, los cuales no solo simplifican la operación de la plataforma sino que permiten una **reducción significativa de costes al pasarse de una facturación por horas/segundos de servidores a una facturación por uso**. Mientras no haya actividad en el negocio, no se aplicarán cargos en la facturación, minimizando al máximo el coste operacional respecto de la demanda real. Algunos ejemplos de arquitecturas de TI sin servidores son:

- **Aplicaciones web**
  - Páginas web estáticas hospedadas en S3 y servidas mediante Cloudfront.
  - APIs REST con AWS Lambda y AWS API Gateway
  - APIs GraphQL con AWS AppSync
- **Procesamiento por lotes**
  - Ejecución de contenedores Docker con AWS Fargate.
  - Colas de mensajes de Amazon SQS
  - Orquestación de tareas con AWS Step Functions
- **Ingestión de eventos**
  - Procesamiento orientado a eventos con AWS Lambda
  - Etiquetado de imágenes con Amazon Rekognition
  - Conversión de texto a voz con Amazon Polly
  - OCR de documentos con Amazon Comprehend
  - Envío de notificaciones con Amazon SNS
- **Servicios de almacenamiento**
  - Almacenamiento de objetos a escala masiva sobre Amazon S3
  - Tablas de almacenamiento clave/valor en DynamoDB con provisión de capacidad bajo demanda
  - BDs relacionales compatibles con MySQL y PostgreSQL sobre Amazon Aurora Serverless

2/ Elección del modelo de contratación

Los principales servicios de procesamiento de AWS permiten hacer uso de modelos de contratación especiales que permiten obtener beneficios económicos muy sustanciales a costa de compromisos de ciertos compromisos operacionales por parte del cliente.

Instancias Reservadas

Soportado para su contratación **en EC2, RDS, Elasticache, Redshift y Opensearch**, el modelo de **instancias reservadas** consiste en la **adquisición de compromisos a 1 o 3 años** durante los cuales un número determinado de instancias permanecerá en uso 24x7 durante periodos de 1 o 3 años. A cambio de este compromiso de permanencia, **Amazon ofrece descuentos sobre el precio bajo demanda que alcanzan el 72%**.

Este tipo de contratación **se recomienda solo para las cargas de trabajo más estables** o, en el caso de infraestructuras con Autoscaling, para la porción mínima de capacidad que permanece activa en todo momento.

Cuando se realiza una compra de instancias reservadas, el cliente elige el número de instancias, la familia, tamaño, duración del compromiso y el modelo de pago inicial que se quiere realizar. En función de la elección se aplicarán automáticamente los descuentos sobre la facturación, por ejemplo para la contratación de una **instancias EC2 Linux c6i.16xlarge en la región de Irlanda** tendríamos la siguiente tabla de precios:

Tipo de reserva	Pago inicial	Pago inicial	Importe / h	Importe / mes	Importe / año	Importe a 3 años	Descuento
Precio bajo demanda	No aplica	0,00 US$	2,92 US$	2130,43 US$	25565,18 US$	76695,55 US$	0 %
Reserva a 1 año	Sin pago inicial	0,00 US$	1,94 US$	1418,54 US$	17022,43 US$	51067,30 US$	33%
	Pago parcial	8059,00 US$	0,92 US$	671,60 US$	16118,20 US$	48354,60 US$	37%
	Pago completo	15796,00 US$	0,00 US$	0,00 US$	15796,00 US$	47388,00 US$	38%
Reserva a 3 años	Sin pago inicial	0,00 US$	1,33 US$	971,63 US$	11659,56 US$	34978,68 US$	54%
	Pago parcial	16188,00 US$	0,62 US$	449,68 US$	21584,16 US$	32376,48 US$	58%
	Pago completo	30434,00 US$	0,00 US$	0,00 US$	30434,00 US$	30434,00 US$	60%

Existen algunas **variables más avanzadas como las reservas convertibles**, que permiten el cambio de tamaño dentro de una misma familia durante la vigencia de la contratación, pero los descuentos obtenidos son significativamente menores. **El uso de estas variables se recomienda cuando la reducción de costes a medio/largo plazo es imperativa a pesar de tener gran incertidumbre en las previsiones de demanda.**

NOTA: Es habitual que muchos clientes de AWS piensen que estas contrataciones se asocian con instancias concretas. Esto no es así, AWS aplica el descuento una vez en la fase de facturación. Si en el periodo de vigencia de una reserva se detecta que hay instancias candidatas para el descuento este se aplica automáticamente, aplicándose la tarifa bajo demanda para el exceso.

Planes de ahorro

Los **planes de ahorro de AWS** son una evolución de las instancias reservadas. En este modelo se ofrecen precios más bajos en comparación con los precios bajo demanda, a cambio de comprometerse a un uso especifico (medido en USD/hora) durante un periodo de uno o tres años. Los planes de ahorro están disponibles en EC2, Lambda, Fargate, Cloudfront y SageMaker.

NOTA: Los descuentos aplicables en el caso de planes de ahorro de EC2 son equivalentes a los de las instancias reservadas. Sin embargo, el modelo basado en contratación por dólares/h en lugar de por volumen de máquinas suele resultar más fácil para muchos clientes de AWS.

Contratación Spot

Las **instancias spot de Amazon EC2** permiten hacer uso de la capacidad sobrante disponible en los datacenters de AWS a un precio que se determina en función de un sistema de subasta que, por lo general, permite acceder a capacidad de cómputo **con descuentos de hasta el 90% respecto del precio bajo demanda**. Las contra-prestaciones de este modelo son dos:

1. Si la demanda global de capacidad reservada o bajo demanda de los datacenters de AWS lo requiere, las instancias spot son susceptibles de ser detenidas sin previo aviso.

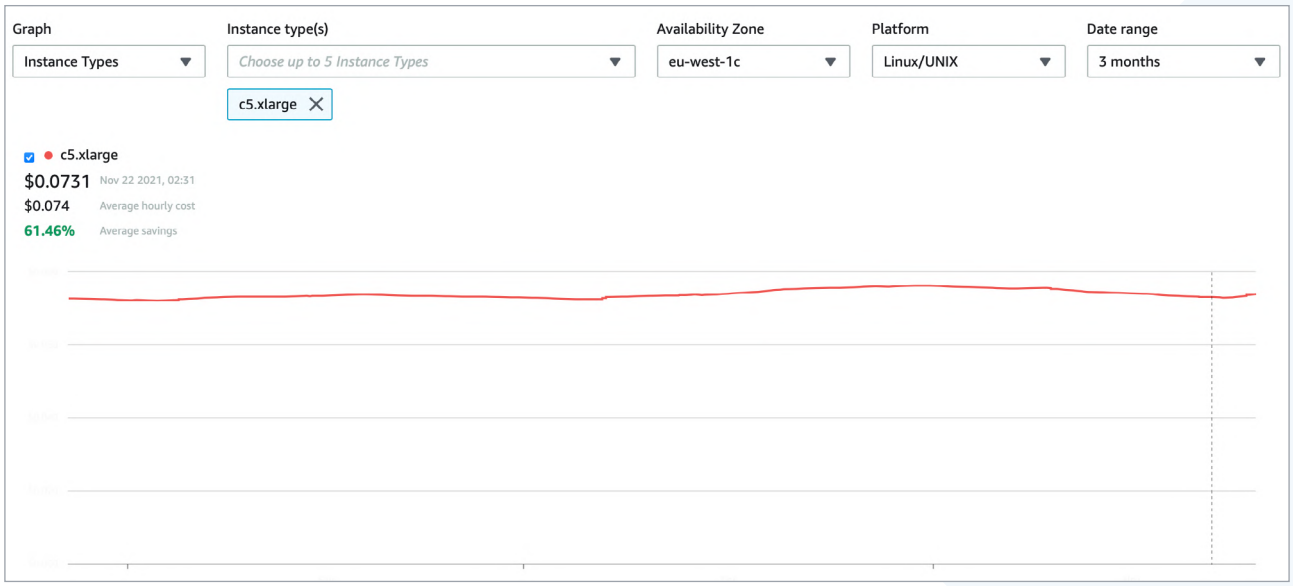
2. Si el precio de la subasta spot sube por encima del límite establecido durante la contratación de instancias, estas son susceptibles de ser detenidas sin previo aviso.

Aunque pueda parecer que este modelo de trabajo tiene muchos riesgos, existen numerosos casos de uso en los que tienen cabida:

- Entornos de desarrollo y pruebas
- BigData
- Aplicaciones web sin estado y tolerantes a errores
- Pipelines CI/CD

La contratación spot **se puede combinar fácilmente con contratación bajo demanda mediante el uso de Autoscaling**, balanceando la contratación de capacidad entre distintas familias y modelos de contratación y utilizando siempre el más ventajoso económicamente en cada momento. Si se da la situación que la contratación Spot no está disponible, Autoscaling realizará la contratación bajo demanda a coste de un pico de facturación mientras dure el evento de demanda global en el datacenter de ejecución.

Desde la consola de contratación de instancias spot de EC2 **se puede monitorizar el precio de la subasta para el tipo de instancia deseado para determinar el precio máximo que se está dispuesto a pagar**. Por ejemplo, para el tipo c5.xlarge Linux en los últimos 3 meses de la zona de disponibilidad eu-west-1a vemos cómo el precio spot se ha mantenido estático en torno a los 0,074$, lo que supone un ahorro promedio del 61%%:



3/ Clases de Almacenamiento

Además de los modelos de contratación especial aplicables a capacidad de cómputo, AWS ofrece descuentos muy significativos en sus servicios de almacenamiento, principalmente a costa de restringir los patrones de acceso al dato. A continuación veremos las particularidades de cada servicio:

EBS

Amazon Elastic Block Storage es el servicio que permite dotar a las instancias del cliente con almacenamiento persistente. Este almacenamiento no solo cuenta con redundancia del dato dentro de una misma zona de disponibilidad para garantizar su durabilidad sino que **permite ciertas características adicionales como encriptación o snapshots**.

Dentro de este servicio, **Amazon proporciona numerosas clases de almacenamiento**, cada una con sus características de rendimiento, capacidad y coste por GB. La elección de una clase de almacenamiento subóptima puede provocar que los costes en este concepto se disparen innecesariamente. Veamos a continuación una comparativa del coste mensual que suponen 10TB y una provisión de 3000 IOPS para los volúmenes que lo permiten, con cada una de las clases disponibles:

Tipo	Precio por GB	Precio IOPS/mes	Total/mes
HDD frío (sc1)	0,02 US$	0,00 US$	168,00 US$
HDD optimizados (st1)	0,05 US$	0,00 US$	500,00 US$
SSD de uso general (GP3)	0,09 US$	0,00 US$	880,00 US$
SSD de uso general (GP2)	0,11 US$	0,00 US$	1100,00 US$
SSD de IOPS provisionadas (io1/io2)	0,14 US$	216,00 US$	1596,00 US$

Como se puede observar, el rango para los mismos 10TB de información varía entre 168 y 1596 dólares al mes. Es por tanto que **se recomienda hacer un análisis adecuado de las necesidades de cada carga de trabajo con el fin de determinar la clase de almacenamiento que permita el desempeño correcto de la actividad empresarial al menor coste**.

EFS

El servicio de sistemas de ficheros NFS altamente disponibles y sin servidores **Amazon Elastic File System** permite hacer uso de distintos tramos de facturación en función de múltiples variables adicionales al tamaño de los datos en sí:

- Redundancia zonal
- Frecuencia de acceso al dato
- Ancho de banda dedicado

Dependiendo de las necesidades y los patrones de acceso **podemos optar a unos descuentos muy significativos respecto de la facturación estándar de almacenamiento**. Supongamos por ejemplo un sistema de ficheros de 1TB y sin ancho de banda provisionado y distintas opciones de patrones de acceso y redundancia

Standard	IA	Redundancia de zona	Importe/mes	Descuento
100 %	0 %	FALSO	330,00 US$	0,00 %
100 %	0 %	VERDADERO	176,00 US$	46,67 %
50 %	50 %	FALSO	177,50 US$	46,21 %
50 %	50 %	VERDADERO	94,65 US$	71,32 %
20 %	80 %	FALSO	86,00 US$	73,94 %
20 %	80 %	VERDADERO	45,84 US$	86,11 %

Dentro de un mismo nivel de redundancia, simplemente por el hecho de activar el mecanismo de transición automática a clases de almacenamiento infrecuente pueden obtenerse reducciones muy sustanciales, sobretodo teniendo en cuenta que lo habitual es que el porcentaje de datos de uso habitual sea una fracción del total en sistemas de ficheros de gran tamaño.

S3

De forma muy similar a EFS, **Amazon S3** ofrece clases de almacenamiento con distintos grados de durabilidad y patrones de acceso. A continuación incluimos un ejemplo del coste de almacenamiento para 50TB con cada una de estas clases de almacenamiento junto con una breve descripción de los casos de uso para los que están destinados:

Clase	Importe/mes	Descuento	Casos de uso
Sandard	1150,00 US$	0,00 %	Almacenamiento de propósito general con acceso frecuente
Standard-IA	625,00 US$	45,65 %	Almacenamiento a largo plazo con acceso poco frecuente (recuperación en ms)
OneZone-IA	500,00 US$	56,52 %	Almacenamiento de datos recreables de acceso poco frecuente
Glacier	200,00 US$	82,61 %	Archivado a largo plazo con recuperación entre 1min y 12h
Glacier Deep Archive	49,50 US$	95,70 %	Archivado a muy largo plazo con recuperación en 12h

Adicionalmente, AWS proporciona una clase de almacenamiento especial denominada Intelligent Tiering, que permite optimizar automáticamente la transición de datos entre Standard y Acceso Infrecuente cuando los patrones de acceso son desconocidos. Esta clase tiene un coste adicional por la monitorización de los patrones de acceso pero permite adoptar clases de almacenamiento de bajo coste en escenarios de alta incertidumbre.

Las contrapartidas de las clases de acceso infrecuente son principalmente dos:

1. **Mayor coste de las peticiones de recuperación de datos.** A menor coste por GB, mayor coste de las operaciones de lectura; es por esto que es importante determinar bien los patrones de acceso antes de optar por una u otra clase de almacenamiento, ya que una mala elección puede conllevar sobrecostes inesperados si el número de operaciones de lectura es muy elevado.

2. **Penalización en la inmediatez de las operaciones de recuperación.** Las clases Glacier y Glacier Deep Archive no permiten leer la información de manera inmediata sino que requieren de hacer solicitudes de recuperación con distintos niveles de servicio. Son el equivalente a los robots de cinta clásicos de los entornos OnPremise.

Opensearch

El servicio de analítica **Amazon Opensearch** permite, además de la contratación de reservas de capacidad y utilización de múltiples clases de almacenamiento EBS, hacer uso de instancias con almacenamiento UltraWarm. Estas instancias permiten expandir el tamaño máximo del almacenamiento de un cluster por una fracción del precio por GB. La contraprestación en este caso es la necesidad de implementar procesos de rotación de índices a dicha clase de almacenamiento.

El backend de almacenamiento detrás de los nodos UltraWarm es S3 en lugar de EBS, lo que obviamente **penaliza el rendimiento de las operaciones de búsqueda**. Sin embargo, al presuponerse que los índices rotados a esta clase de almacenamiento serán de acceso menos frecuente, esta penalización es asumible respecto del coste que tendría mantener todo el dataset en nodos estándar.

Los nodos UltraWarm están disponibles en dos tipologías, cada una con una asignación de capacidad de cómputo y almacenamiento administrado. Estos nodos únicamente **sirven para la desconexión de índices rotados a UltraWarm y servir de interfaz a los nodos principales del cluster** cuando se realizan operaciones de búsqueda contra índices de acceso infrecuente.

Adjuntamos a continuación un ejemplo del coste de almacenamiento de un dataset de 5TB utilizando distintos porcentajes de almacenamiento UltraWarm. No se valoran los costes de los nodos, únicamente el almacenamiento en sí:

Standard	UltraWarm	Importe/mes	Descuento
100 %	0 %	149,00 US$	0,00 %
75 %	10 %	114,15 US$	23,39 %
50 %	20 %	79,30 US$	46,78 %
25 %	50 %	49,25 US$	66,95 %

Optimizando los procesos de rotación de índices y provisionando un número adecuado de nodos UltraWarm, podemos desplegar clusters Opensearch con un tamaño contenido y una capacidad de almacenamiento muy superior de lo que sería necesario utilizando nodos de alta capacidad de memoria con almacenamiento EBS, todo ello a una fracción del coste normal.

4/ Herramientas de detección y monitorización

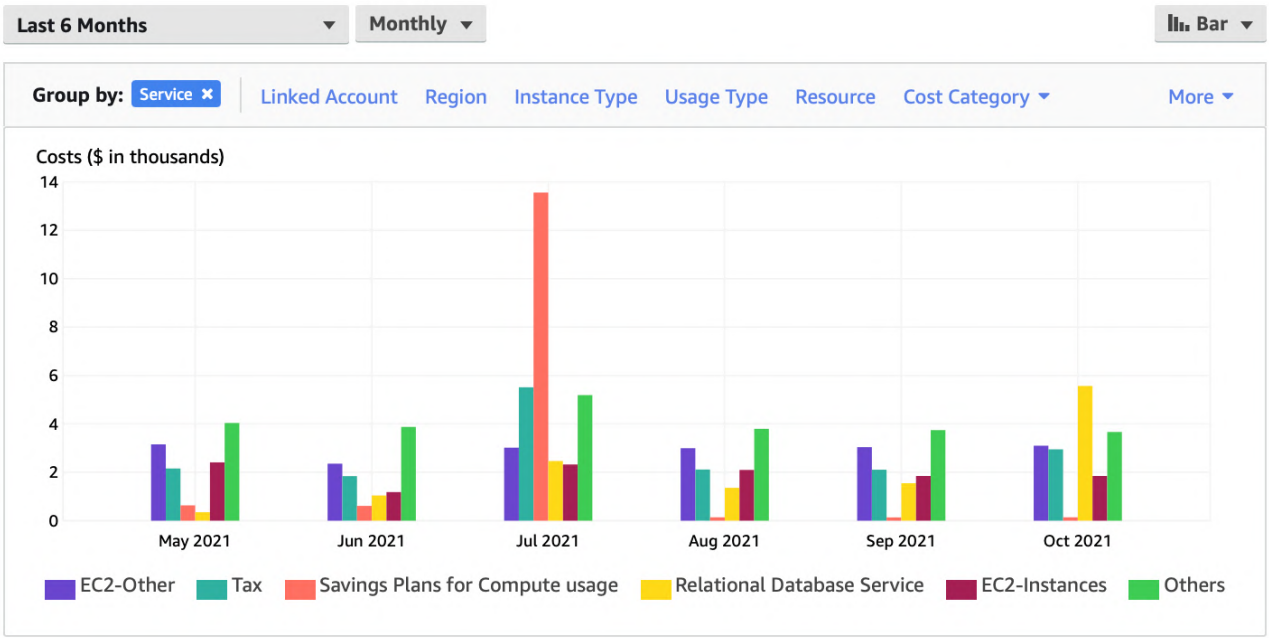
A lo largo de este documento ya hemos mencionado algunas herramientas de alto valor a la hora de detectar y corregir el dimensionamiento de la plataforma de servidores, tales como Trusted Advisor, Compute Optimizer o Cloudwatch. Además de estas herramientas, **AWS proporciona dos servicios especialmente diseñados para el control de costes operativos**: AWS Cost Explorer y AWS Budgets.

AWS Cost Explorer

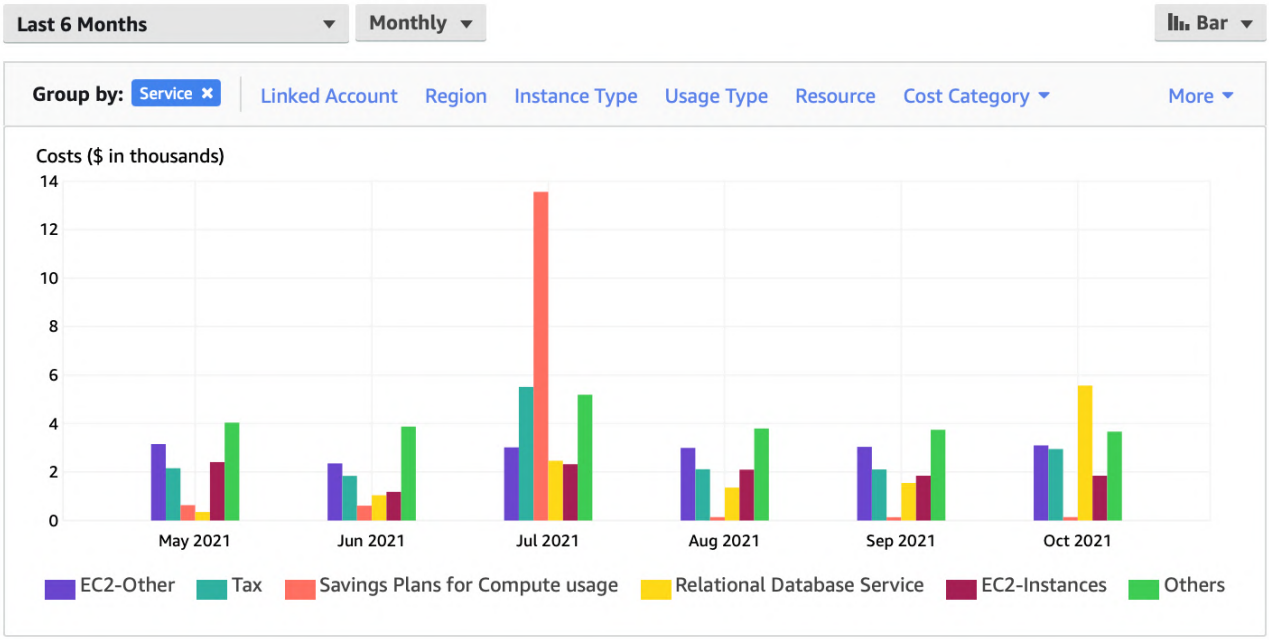
Este servicio **es la principal herramienta para la visualización y análisis de los costes operacionales dentro de Amazon Web Services**. Dentro de este recursos se consolida todo el detalle acerca de las cuentas, regiones, servicios, tipo de utilización, opciones de compra y multitud de opciones avanzadas que generan costes económicos dentro de su organización.

Cost Explorer incluye por defecto un conjunto de informes que ayudarán en la obtención de la información necesaria para determinar los principales factores de coste y tendencias de uso dentro de su cuenta u organización de cuentas en AWS. Se permiten informes con granularidad mensual, diaria o incluso horaria (esta última con coste adicional). Con la configuración adecuada incluso podrán realizarse rupturas por etiquetas o conjuntos de etiquetas, habilitando con ello (siempre y cuando la política de etiquetado se siga correctamente) la elaboración de informes con rupturas por conceptos de negocio, como por ejemplo centros de costes, entornos, unidades de negocio, etc.

Informe de coste mensual por servicio:



Informe de coste diario por tipología de instancia EC2:



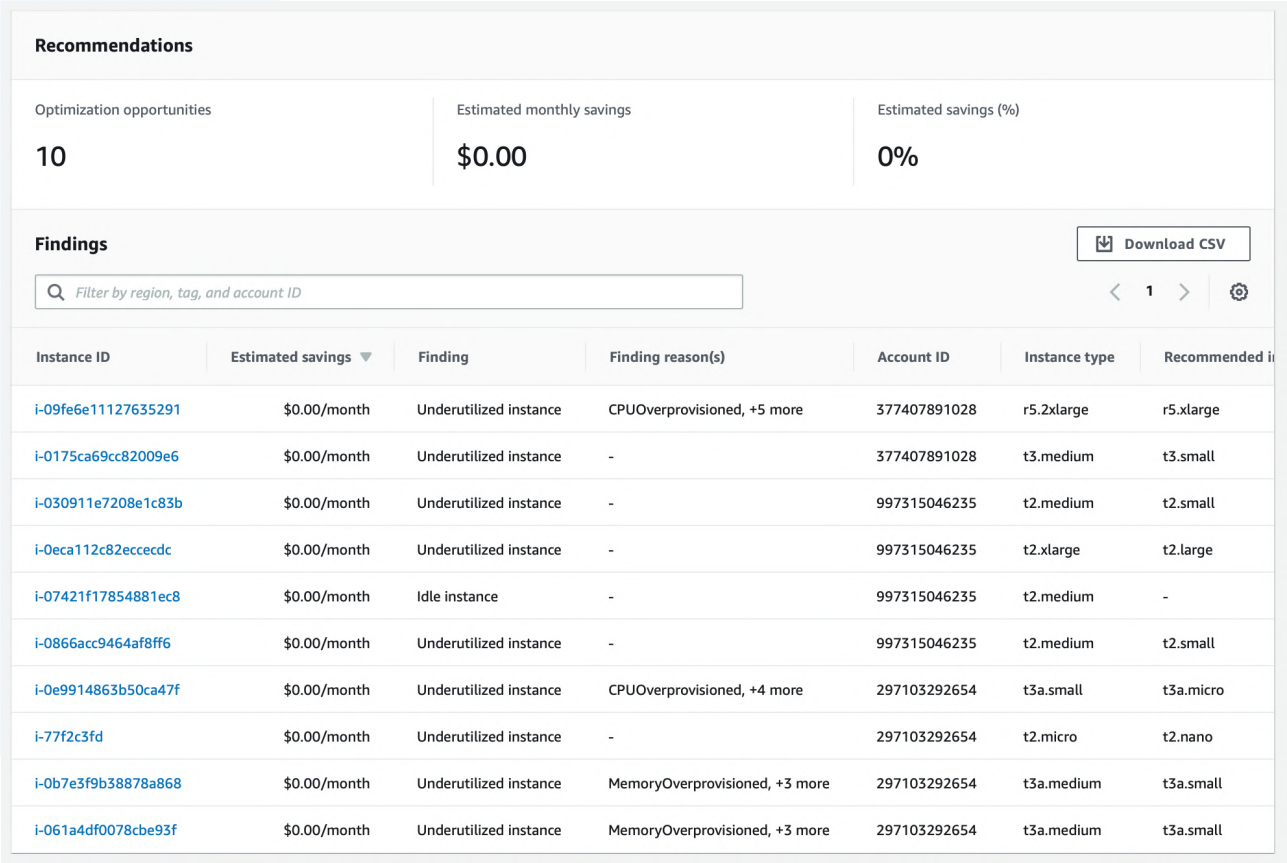
También se incluyen dentro de este servicio capacidades de detección de anomalías mediante Machine Learning para la detección temprana de picos de facturación inesperados que podrían ser resultado de un uso malintencionado, compromiso de credenciales o simples descuidos por parte de ciertos equipos de trabajo:

Anomalies detected (47) Info

Find detected anomalies by property or value | Last 90 days (all) | 1 2 3 4 5

Detection date	Severity	Duration	Service	Account ID	Total cost impact	Assessment
2020-11-11	Low	14 days	Amazon Elastic Compute Cloud - Compute	9[illegible]0	$3,348,575.96	Not submitted
2020-10-21	Low	6 days	Amazon Elastic Compute Cloud - Compute	9[illegible]0	$1,046,571.62	Not submitted
2020-10-08	Low	8 days	Amazon Elastic Compute Cloud - Compute	-	$953,814.27	Not submitted
2020-11-15	Low	3 days	Amazon Elastic Compute Cloud - Compute	-	$898,037.87	Not submitted
2020-10-26	Low	3 days	Amazon Elastic Compute Cloud - Compute	9[illegible]0	$669,155.71	Not submitted
2020-11-29	Low	1 day	Amazon Elastic Compute Cloud - Compute	9[illegible]0	$390,791.80	Not submitted
2020-11-22	Low	1 day	Amazon Elastic Compute Cloud - Compute	9[illegible]0	$346,158.89	Not submitted
2020-09-29	Low	1 day	Amazon Elastic Compute Cloud - Compute	-	$70,607.67	Not submitted
2020-10-27	Low	4 days	Amazon Elastic Block Store	6[illegible]5	$39,900.74	Not submitted
2020-10-23	High	3 days	Amazon Virtual Private Cloud	9[illegible]0	$29,734.04	Not submitted

Cost Explorer se integra con EC2 Compute Optimizer y Trust Advisor para la detección de recursos en desuso o con dimensionamientos sub-óptimos:



Recommendations

Optimization opportunities	Estimated monthly savings	Estimated savings (%)
10	$0.00	0%

Findings

Download CSV

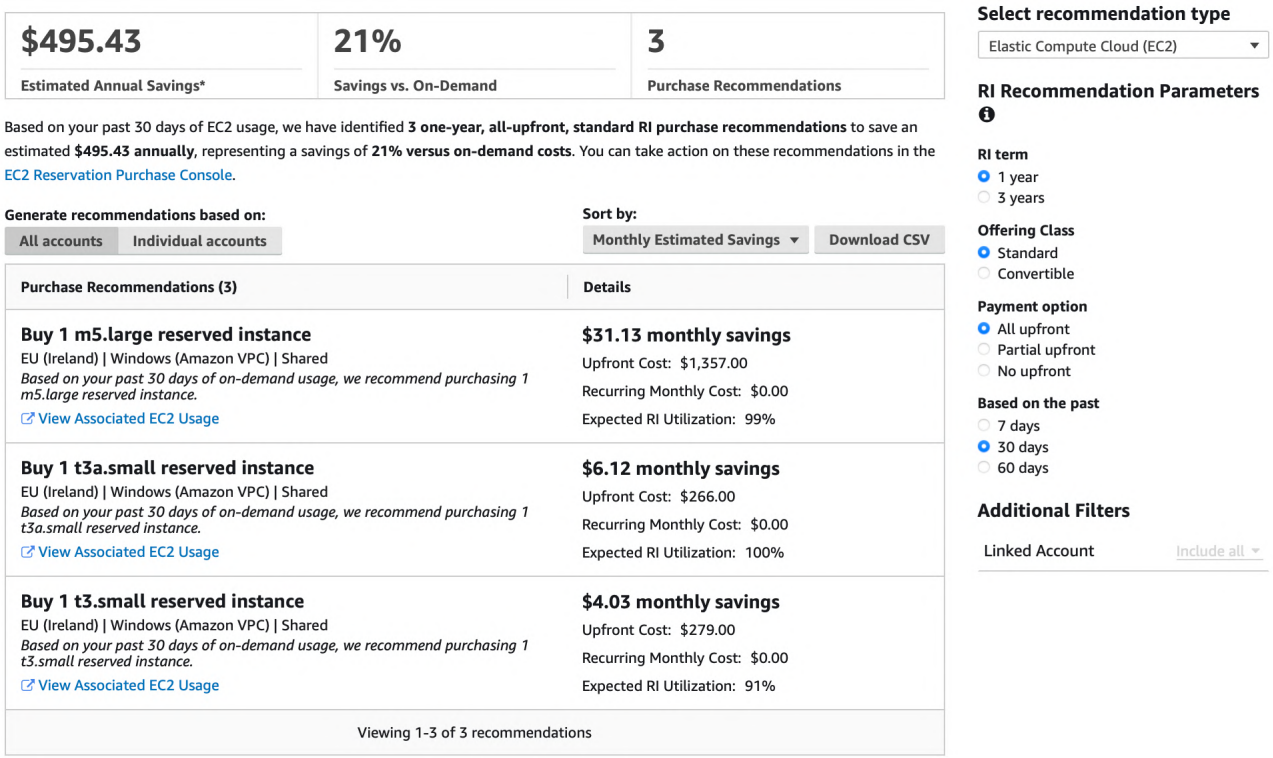
Filter by region, tag, and account ID

1

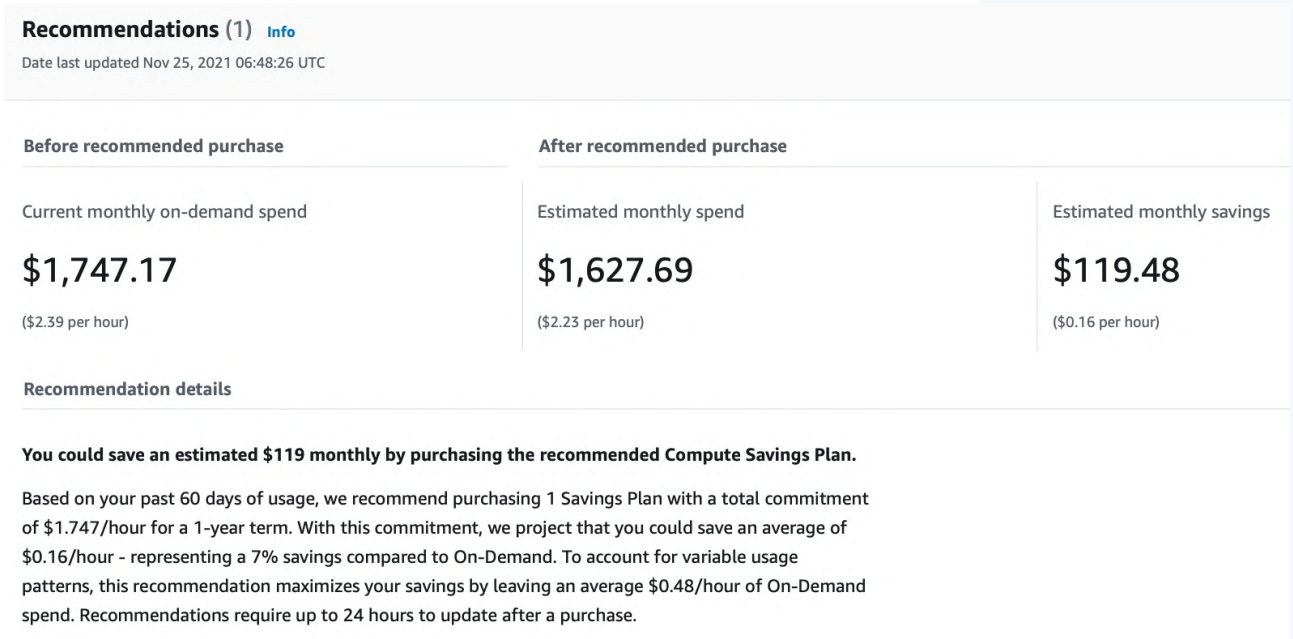
Instance ID	Estimated savings	Finding	Finding reason(s)	Account ID	Instance type	Recommended i
i-09fe6e11127635291	$0.00/month	Underutilized instance	CPUOverprovisioned, +5 more	377407891028	r5.2xlarge	r5.xlarge
i-0175ca69cc82009e6	$0.00/month	Underutilized instance	-	377407891028	t3.medium	t3.small
i-030911e7208e1c83b	$0.00/month	Underutilized instance	-	997315046235	t2.medium	t2.small
i-0eca112c82eccecdc	$0.00/month	Underutilized instance	-	997315046235	t2.xlarge	t2.large
i-07421f17854881ec8	$0.00/month	Idle instance	-	997315046235	t2.medium	-
i-0866acc9464af8ff6	$0.00/month	Underutilized instance	-	997315046235	t2.medium	t2.small
i-0e9914863b50ca47f	$0.00/month	Underutilized instance	CPUOverprovisioned, +4 more	297103292654	t3a.small	t3a.micro
i-77f2c3fd	$0.00/month	Underutilized instance	-	297103292654	t2.micro	t2.nano
i-0b7e3f9b38878a868	$0.00/month	Underutilized instance	MemoryOverprovisioned, +3 more	297103292654	t3a.medium	t3a.small
i-061a4df0078cbe93f	$0.00/month	Underutilized instance	MemoryOverprovisioned, +3 more	297103292654	t3a.medium	t3a.small

Por último, dentro del mismo panel de Coste Explorer **se ofrecen multitud de informes y recomendaciones automáticas para la contratación de reservas de capacidad y planes de ahorro**, así como el grado de cobertura actual. Esta información puede ser clave a la hora de elaborar presupuestos para la compra inicial de reservas y dar visibilidad al departamento de contabilidad sobre los ahorros estimados a medio/largo plazo:

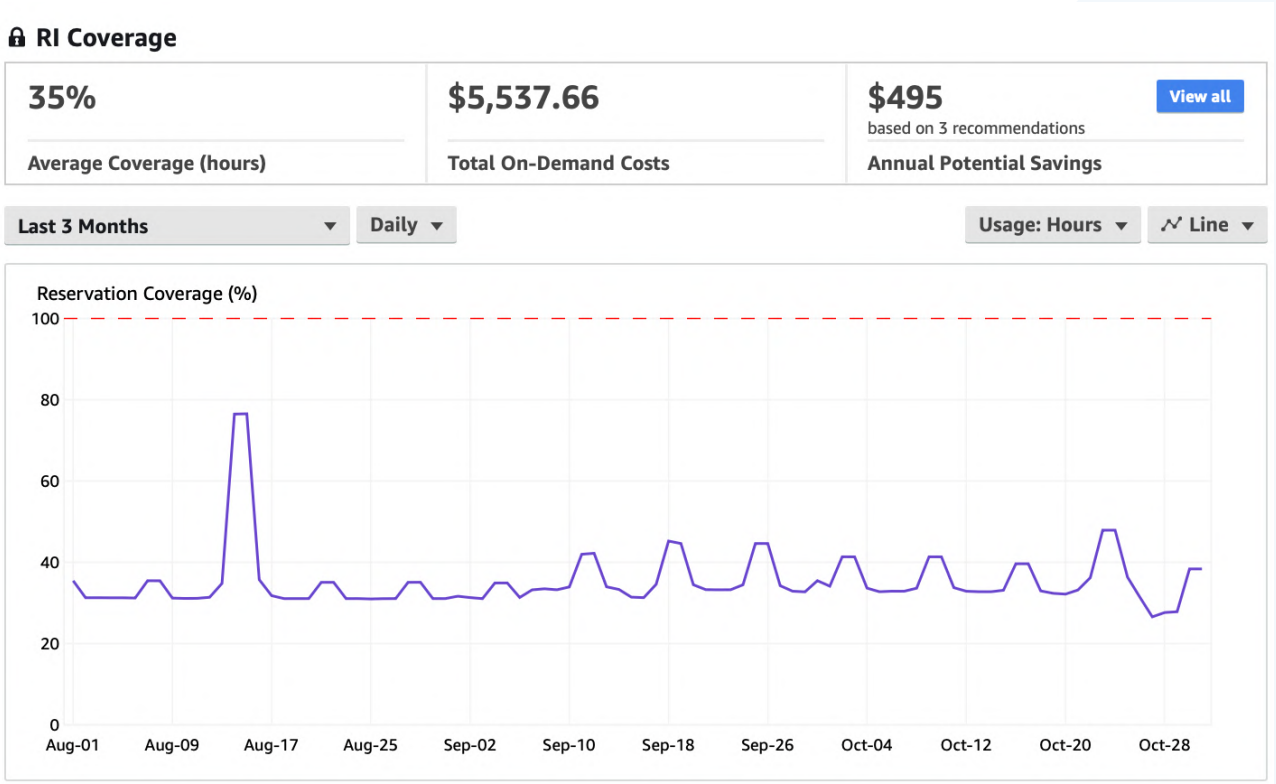
Recomendaciones para instancias reservadas de EC2:



Recomendaciones para planes de ahorro:



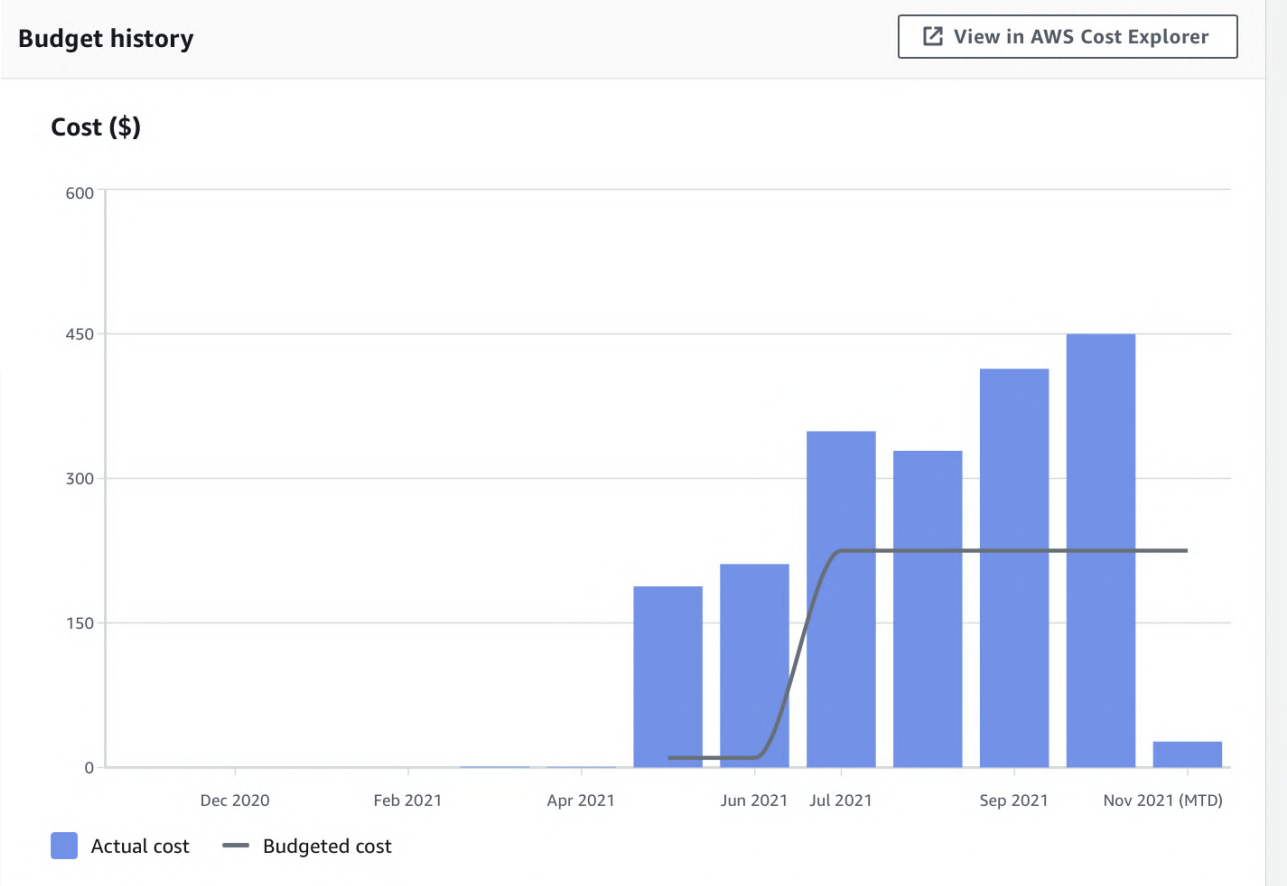
Informe de cobertura de reservas EC2:



AWS Budgets

Este servicio **se complementa con Coste Explorer para la generación de presupuestos para el control de costes operativos**. Estos presupuestos permiten definir un coste máximo para periodos de amortización habituales (día, mes, trimestre...), para generar informes, previsiones de gasto y alarmas cuando la progresión de costes dentro del marco del presupuesto excedo ciertos umbrales.

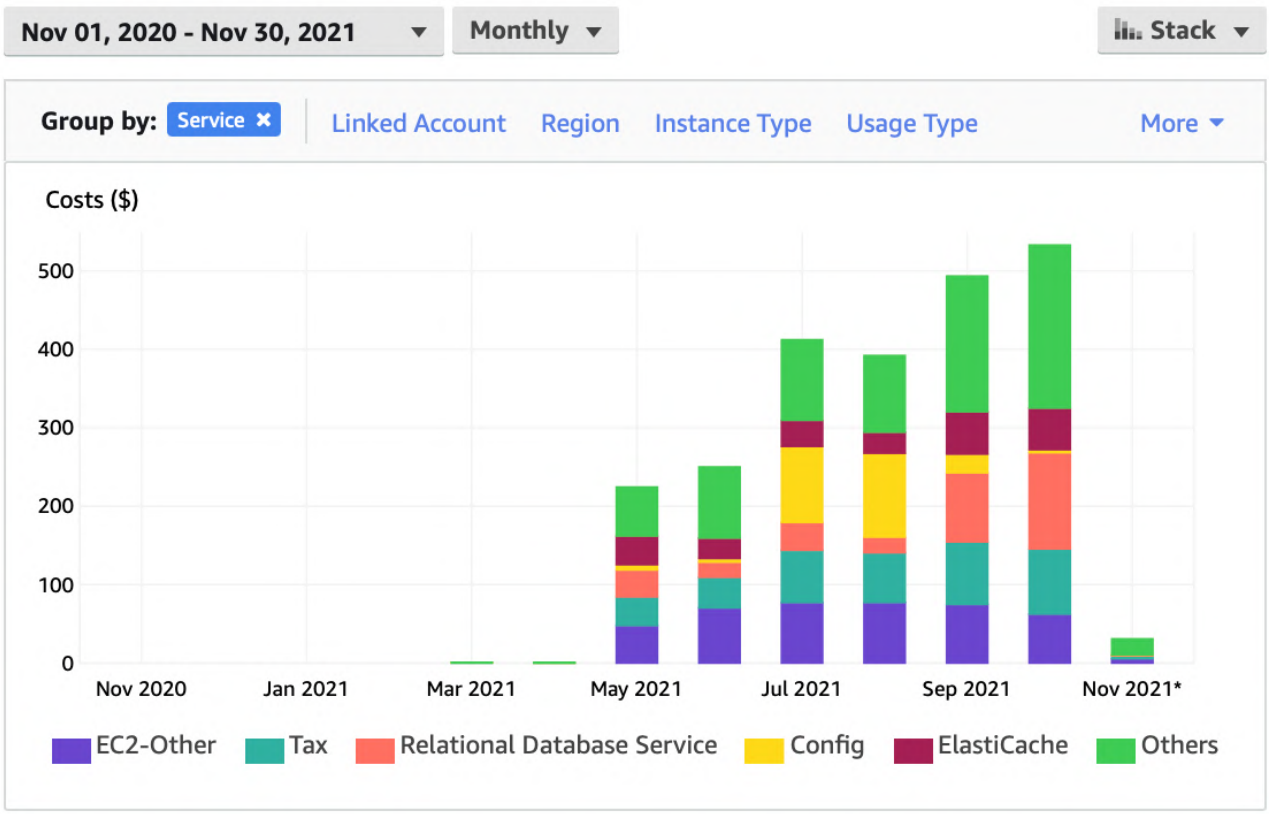
En el siguiente ejemplo se ha definido un presupuesto de 10$/mes (que tras un par de meses subió a 225$/mes) para una cuenta de la organización, con un umbral de alarma del 80% sobre el importe real consumido:



Monthly costs history Download as CSV

Date	Actual	Budgeted	Budget variance (USD)	Budget variance (%)
November 2021 (MTD)	$26.89	$225.00	$198.11	88.05%
October 2021	$449.79	$225.00	-$224.79	-99.91%
September 2021	$413.72	$225.00	-$188.72	-83.88%
August 2021	$328.60	$225.00	-$103.60	-46.04%
July 2021	$348.85	$225.00	-$123.85	-55.04%
June 2021	$211.07	$10.00	-$201.07	-2,010.72%
May 2021	$187.93	$10.00	-$177.93	-1,779.32%

Gracias a este informe, se pudo comprobar cómo la infraestructura de la cuenta sobrepasaba con creces la asignación de presupuesto estimada entre los meses de abril y octubre de 2021. Los encargados del control de facturación recibieron mensualmente notificaciones alertando de la situación, gracias a las cuales durante el mes de noviembre las anomalías que provocaban esta desviación quedaron finalmente resueltas, devolviendo la situación a los niveles inicialmente estimados. La integración directa con AWS Cost Explorer permite analizar al detalle los costes específicos de este presupuesto con un solo click:



5/ Conclusiones

A lo largo de este artículo hemos comentado y analizado multitud de herramientas, opciones y procedimientos para la optimización de costes operativos en infraestructuras corriendo en la plataforma de Amazon Web Services.

Con una dedicación de trabajo adecuada para realizar un análisis de situación, cualquier usuario de AWS comprobará fácilmente que **reducir sustancialmente sus costes operativos es algo perfectamente factible**, mejorando la rentabilidad de la plataforma y habilitando con ello la inversión de recursos en procesos de mejora y servicios para sus clientes.

Es importante tener siempre en mente que la nube de A**WS siempre supondrá una reducción de costes operativos respecto de OnPremise** si se hace un uso adecuado de los recursos disponibles. Manejar un sistema tan flexible y escalable como este sin tener las suficientes precauciones puede conllevar más de una sorpresa a la hora de revisar la factura mensual.

Si este whitepaper te ha parecido interesante y quieres estar informado de nuevos contenidos relacionados con el Cloud e infraestructuras, visita nuestra web para descubrir estos y otros contenidos de valor.

Descubrir más

Si necesitas ayuda para entender correctamente los costes de tu infraestructura/servicios, tanto Cloud como onPremise, no lo dudes, contacta con un equipo especialista que te ayudará a confirmar si tu estrategia es la correcta y existe la posibilidad de optimizarlos.

Contactar