Livret d'Epargne Cloud

gekko
les ingénieurs du cloud

INTRODUCTION

Vous avez démarré vos premiers workloads sur un ou plusieurs Cloud providers. Vous avez pu savourer le plaisir inégalé d'obtenir des ressources en quelques minutes plutôt que plusieurs semaines, construire des architectures sécurisées et performantes en quelques clics, tester des services innovants pour supporter vos projets digitaux en toute simplicité et sans passer par de multiples process internes faisant intervenir des dizaines d'interlocuteurs. Tout cela, pour quelques euros par heure. Mais ça n'a pas duré.

La facture qui pesait quelques centaines d'euros par mois est passée à plusieurs milliers. Et quand elle a dépassé 5 chiffres, les questions ont surgi : Comment avait-on ouvert ce compte, déjà ? Avec la carte bancaire de qui ? Et c'est sur quel budget, d'ailleurs ? On en a vraiment besoin ? Est-ce que ça ne fait pas double usage avec ce qu'on a dans nos datacenters ? Qui a la main sur la dépense, et qui surveille ce que l'on fait ? Et est-ce qu'on ne peut pas optimiser tout ça ? Outre que ne pas y répondre signifie souvent la fin de l'expérience - et c'est dommage -, il faut bien admettre que ce sont de bonnes questions : nous parlons ni plus ni moins de la mise en place d'une gouvernance opérationnelle et financière d'un composant appelé à jouer un rôle majeur dans l'IT de toutes les entreprises.
Or c'est compliqué, pour au moins deux bonnes raisons … et deux mauvaises !

››› LES CAUSES DE LA COMPLEXITE

Les bonnes tout d'abord

Au-delà des usages qu'il permet, le Cloud a profondément bouleversé les processus d'achat et d'exploitation de l'IT en tant que ressource.
Les entreprises fonctionnent par budget, et rationalisent leurs dépenses via leur fonction achats ; on achète au meilleur prix en maximisant le levier du volume et donc de la centralisation.

Par conséquent, les achats d'infrastructures reposent sur une estimation de capacité à long terme : on s'équipe pour pouvoir traiter les besoins d'aujourd'hui et de demain, et pour supporter la crête de charge aussi bien que le train-train quotidien.

C'est pourquoi les infrastructures des datacenters sont souvent excédentaires et faiblement chargées, allumées jour et nuit, hébergeant des projets actifs ou non. Tout ceci ne posant pas de problème tant que l'on reste dans le volume d'infrastructures qui a été planifié et provisionné.

Or le Cloud est par construction doublement orthogonal à ce modèle :

- D'un point de vue achat, plusieurs personnes ont la main sur la dépense, et celle-ci devient plus difficile à anticiper, réguler et optimiser. De plus, on provisionne à la demande et non à long terme ; les coûts sont variables et non plus fixes.

- D'un point de vue exploitation, la tentation sera grande pour les équipes d'appliquer les mêmes principes de dimensionnement, de tolérance à l'inutilisation et de fonctionnement jour et nuit que sur les datacenters ; ce qui dilue les bénéfices du Cloud.

Outre les bien connus glissements de modèle « Capex vs. Opex » et « Make vs. Buy », le Cloud amène donc une gestion de l'IT orientée «flux» plutôt que «stocks» : là où l'on cherchait l'équilibre entre trop (incidence financière) et pas assez (incidence business) d'un stock d'IT, on s'intéresse à présent à la régulation de la consommation d'un flux d'IT en fonction du besoin.
Tout ceci nécessite de l'outillage et de la pédagogie, c'est dans cette lignée que s'inscrit le Livret d'Epargne Cloud qui peut être vu comme un véritable outil :
un bon reporting opérationnel et une bonne ventilation des coûts seront le commencement de la sagesse...

LA COMPLEXITÉ DE L'INFORMATION

Et c'est là qu'intervient la première mauvaise raison. Le Cloud est un outil fabuleux en termes de puissance, de possibilités, d'innovation.... Cependant, nous pensons que ce ne sera pas faire offense aux Cloud providers que d'estimer qu'ils ne facilitent pas vraiment la tâche des gestionnaires. La quantité d'information est colossale sur beaucoup de métriques différentes ; arriver à avoir une vision globale nécessite de traiter cette information afin de la rendre compréhensible.

Quiconque s'est déjà escrimé sur des exports de coûts pour retraiter les chiffres en vue d'une analyse de consommation ou de réallocation des coûts n'est pas étranger avec cette complexité : le manque d'informations clés, la profusion d'informations parasites, les multiples conventions de nommage inconstantes, la lourdeur des fichiers, etc.
Les différentes factures et rapports sont par ailleurs souvent insuffisants, et il est difficile d'avoir une vue d'ensemble regroupant services et régions.
Ce qui n'est déjà pas évident pour un compte devient quasi-impossible dès que l'on gère une flotte de plusieurs comptes au profit de plusieurs projets ou métiers de l'entreprise. D'ailleurs, tous les Cloud providers ne proposent pas forcément de vue consolidée (notamment des mesures d'usage et de performance), ce qui ajoute une difficulté supplémentaire.

In fine, des outils natifs existent et permettent souvent de faire beaucoup de choses. Cependant, leur prise en main est assez complexe et nécessite par moment de faire appel à des compétences techniques. Cela représente un investissement en temps et souvent en main d'œuvre.

En fonction des besoins et des objectifs recherchés, un outillage additionnel (du marché ou maison) peut donc s'avérer utile. Chaque outil est fortement spécialisé et il faut bien définir son besoin avant de le choisir.

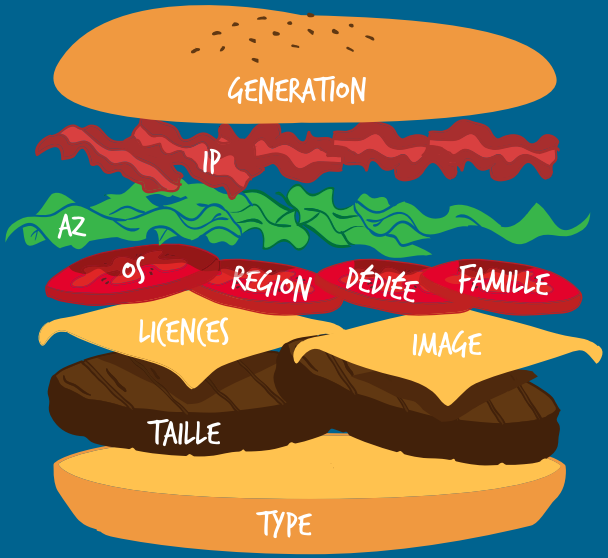
SURABONDANCE DE CHOIX

La seconde mauvaise raison part d'une bonne intention : c'est la surabondance du choix.
Chez Gekko nous l'appelons l'effet «Subway», du nom de la célèbre chaîne de restauration, par analogie avec la perplexité dans laquelle nous plonge la perspective de devoir faire notre choix dans une combinatoire de 5 types de pains, 4 variétés de viande, un grand nombre de crudités et légumes, plusieurs sortes de sauces et condiments, le tout en taille de 10, 15 ou 30 cm…

C'est forcément plus riche que «jambon / crudités» ou «saucisson sec / cornichons», mais cela est potentiellement générateur de frustration, et impose donc plus de réflexion…

Quel rapport avec le Cloud ?

Prenons l'exemple d'AWS en août 2019, dans le service de calcul EC2, nous avons le choix entre une dizaine de familles d'instances, réparties en plusieurs catégories (générales, optimisées pour le calcul, pour la mémoire, pour le stockage ou le calcul intensif), existant chacune sur 1 à 5 générations, et disponibles en 3 à 7 tailles selon les modèles ; si on ne retient que les générations actives (N et N-1), ceci ne représente environ «que» 70 types d'instances. Celles-ci sont disponibles en versions Windows, versions Linux, Red Hat Enterprise Linux, SUSE, avec ou sans licence pour du SQL, dans 22 régions (dont une spéciale pour l'administration US) et 69 zones de disponibilité. Enfin au-delà de la technique, il faudra choisir entre une tarification à la demande ou par réservation ; et dans ce dernier cas, il faudra décider si l'on veut que les instances soient convertibles ou non, avec une option régionale ou pas, et opter pour une durée de réservation (1 an ou 3 ans), ainsi que pour un


GENERATION
IP
AZ
OS
REGION
DÉDIÉE
FAMILLE
LICENCES
IMAGE
TAILLE
TYPE

mode de paiement (upfront, partial upfront, no upfront). Si l'on regarde la totalité de l'offre dans le monde entier, le nombre de combinaisons possibles pour la configuration d'une instance EC2 dépasse 2 millions. On comprend facilement que pour le non-initié cela puisse être une source de confusion. Et EC2 n'est qu'un des services, parmi la centaine offerte par AWS (165 en août 2019). Comme pour notre fameux sandwich, il est donc difficile de choisir ; et une fois choisi, il est difficile de savoir que l'on n'aurait pas pu faire mieux. Mais heureusement, à la différence des sandwichs, il existe là encore des outils pour nous aider à objectiver et optimiser nos choix, et c'est ce que nous verrons plus loin.

››› LE MULTICLOUD

A ces multiples sources de complexité vient s'en ajouter une nouvelle : le multicloud. A l'heure de faire ce choix cornélien, rares sont les entreprises à décider d'un mouvement généralisé vers un seul et même fournisseur Cloud. Cela est compréhensible, de la même manière que les industriels mettent un point d'honneur à diversifier leurs fournisseurs, avoir plusieurs fournisseurs Cloud revient à ne pas mettre tous ses œufs dans le même panier.

Cependant, de la même manière que pour l'industrie, cela a des avantages comme des inconvénients. On trouve dans la première catégorie :

L'optimisation du sourcing au cas par cas

En effet, si les prix des principaux services sont globalement alignés entre les différents Cloud providers, il existe de légères différences qui peuvent être très avantageuses pour certaines applications aux besoins spécifiques. Ainsi, si les VMs d'utilisation générale ont sensiblement voire exactement le même prix pour les puissances équivalentes, des services spécialisés comme les requêtes API sont à 3,5$ par million pour Azure et AWS, là où GCP dispose de seuils commençant à 3$ et descendant jusqu'à 1,5$ par millions. Cela représente un réel avantage de coûts dans ce cas d'utilisation.

L'accès à des services particuliers	Si AWS est actuellement le leader en matière d'éventail de services proposés, les autres s'alignent rapidement et surtout, certains font le choix de se spécialiser. Ainsi GCP offre des services de machine learning très poussés et continue à créer des services très innovants répondant à des besoins très spécifiques.
Les zones de disponibilité	Certaines entreprises sont assez sensibles à l'emplacement où sont stockées leurs données notamment en matière de sauvegarde d'archive. Certains préfèrent garder la donnée dans leur pays.
Les conflits collatéraux	Beaucoup de décisions sont clivantes dans les milieux technologiques, ce qui participe à la volatilité du marché. Ainsi, être sur plusieurs Clouds, c'est s'assurer d'avoir un Cloud toujours compatible avec tel ou tel logiciel et ne pas être dépendant de l'alignement de la stratégie commerciale des éditeurs de licence avec celle des Cloud providers (forte augmentation des prix de licence Windows, désaccords entre Amazon et Oracle, et bien d'autres).
	Dans la seconde catégorie, ce sont surtout des désavantages financiers :
L'existence de programmes de réduction des prix en fonction des volumes	Répartir ses besoins Cloud sur plusieurs providers, c'est diviser la charge pour chacun de ceux-ci. Ainsi, les programme de réduction à la volumétrie sont moins intéressants et donc les économies sont réduites.
Une multitude de factures	Comprendre les factures issues d'un provider est une chose mais multiplier cet exercice peut rapidement devenir une tâche mensuelle conséquente.

La redondance de certains coûts et services

Chaque Cloud dispose de coûts fixes mensuels. Avoir plusieurs Cloud correspond à multiplier ces coûts par autant de fois que l'on a de Cloud. Par exemple, les coûts mensuels de support pour les entreprises s'élèvent à une dizaine de milliers d'euros. Ainsi, pour une facture mensuelle d'une centaine de milliers d'euros, le multicloud est un investissement conséquent.

Augmentation de la complexité

Avoir plusieurs Cloud signifie assurer le même niveau de sécurité partout, créer des connexions entre les Clouds, etc. Cela complexifie énormément le principe même du Cloud. Cela complexifie aussi la vie des développeurs qui auront besoin de faire une étude sur le Cloud à utiliser avant de se lancer dans la création d'une application. C'est un pas en arrière par rapport à l'agilité du Cloud.

Un besoin d'expertise plus large

Si les grands Cloud providers ont une offre similaire, ils utilisent des technologies différentes dans un certain nombre de sujets. Ainsi, être multicloud signifie posséder les compétences nécessaires pour chacun des Cloud utilisés, cela peut vite devenir compliqué à la vitesse où évoluent les technologies.

Le multicloud n'est donc pas une évidence. On peut même penser qu'en-dessous d'une taille critique, son intérêt est limité. Mais au-delà d'une certaine taille, il peut dans certains cas s'imposer, et nécessite alors un suivi FinOps poussé, ne serait-ce que pour être capable de comparer les équations économiques d'un Cloud par rapport à un autre.

LE CLOUD FINOPS : C'EST QUOI ?

Un métier est donc né, absolument clé dans le processus d'adoption du Cloud dans les entreprises car garant de l'adhérence entre le design technique, le comportement de consommation et la finance. Nous l'appelons pour notre part Cloud FinOps, et nous y distinguons 4 activités complémentaires :

Pilotage budgétaire	Business case, suivi du réel vs le budget (généralement avec un étalonnage car les activités basculées diffèrent de ce qui était prévu), analyse des changements, alertes si écarts importants, estimations, etc.
Allocations des coûts	Showback & chargeback, suivi détaillé de la consommation Cloud (par compte, par service, par appli, etc.), ventilation des coûts (avec ou sans refacturation interne), etc.
Cost optimisation	Actions de baisse des coûts par élimination du gaspillage, saturation des ressources, sélection des options tarifaires ou des services les plus appropriés, etc.
Conduite du changement	Formation et sensibilisation afin d'aligner et impliquer les différents acteurs de l'entreprise dans un plan d'amélioration continue, communication aux métiers et fonctions support, etc.

Chez Gekko, nous avons eu la chance de participer à quelques-uns des premiers projets de Cloud FinOps, et d'échanger avec les principaux acteurs internationaux de ce domaine qui suivent notre retour d'expérience, avec un focus sur l'optimisation financière. Il ne tient plus qu'à vous de venir l'enrichir avec nous.

››› L'OPTIMISATION

Une célèbre styliste britannique disait à propos des vêtements : «Buy less. Choose well. Make it last.».
Nous trouvons pour notre part que cette maxime résume bien les fondamentaux de n'importe quelle méthodologie de réduction de coûts.

Appliqué au Cloud, nous classons les initiatives en 4 catégories.

ACTIONS «QUICK WINS»

Acheter moins

Cloud economics

Utiliser mieux

Rightstizing, tiering, élimination gaspillage

Payer moins cher

Réservations, remise sur volume

ACTIONS STRUCTURELLES

Consommer mieux

Modifications d'architecture : lambda, Aurora, RDS, spot, services managés

ACHETER MOINS

Sur ce sujet, on pense d'abord spontanément au gaspillage, et c'est effectivement la première action évidente à entreprendre. Toutes les sociétés ne sont pas logées à la même enseigne sur ce point, mais les datacenters hébergent assez naturellement des ressources oubliées ou inutiles (chez Gekko, nous les avons affectueusement nommées les «zombis») : volumes non attachés, VMs (voire serveurs) inutilisées, équipements réseau inactifs, etc. Ces zombis se développent d'autant plus facilement que les environnements sont instables ou dynamiques (environnements de test ou d'intégration, expérimentations d'applications, métiers innovants, etc.), au gré des tests, des projets suspendus jusqu'à nouvel ordre, des ressources commissionnées «au cas où» ou «en attendant que», etc.

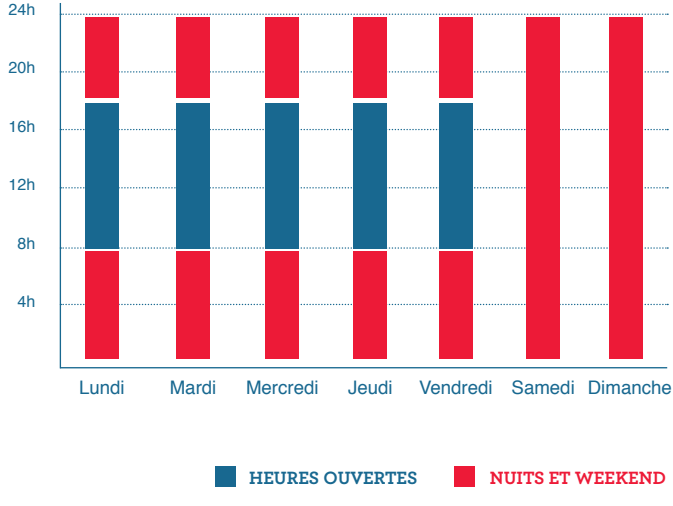
Cela n'est pas très gênant dans un datacenter dans la mesure où les zombis mobilisent des infrastructures existantes et utilisées par ailleurs, et génèrent donc un coût marginal faible. Dans le Cloud en revanche, ces ressources inutilisées représentent un coût direct inutile et qui peut être évité ; leur identification et leur élimination quotidienne (manuelle ou automatique) sont donc indispensables.

Au-delà de la chasse au gaspillage, le deuxième levier de réduction de la quantité consommée repose sur une utilisation maximale d'une des caractéristiques fondamentales du Cloud - **l'élasticité** :

- Il y a 168 heures dans une semaine.
- Le week-end représente 48 heures, soit près de 30% de l'ensemble ; une application pouvant être arrêtée le week-end devrait donc générer une économie de 30% sur les ressources associées.

- Sur la base d'une journée travaillée de 10 heures, les nuits représentent 70 heures, soit près de 60% de la semaine hors week-end. Il y a donc un nouveau gisement d'économies pour les applications ne supportant que les heures ouvrées : de manière globale, on a sur une semaine un gain de plus de 65% atteignable sur les applications qui peuvent être éteintes en heures non ouvrées.

- Et il est possible d'aller plus loin en déclinant les jours fériés (une dizaine par an en France, selon les années et la coïncidence avec un week-end ou non), les vacances ou tout particularisme saisonnier d'une entreprise.



UN EXEMPLE

Chez Azure, prenons une machine virtuelle D4 V3, c'est-à-dire l'équivalent d'un serveur généraliste de 4 vCPUs et 16 Gb de RAM. Elle coûte environ 24 centimes par heure *(août 2019, Europe de l'ouest, Linux).*

Quel sera son coût annuel si :

- elle n'est **jamais éteinte**, comme sur un datacenter : $2 096,64
- elle est **éteinte le week-end :** $1 497,60
  Soit 29% d'économies
- elle est également **éteinte la nuit :** $624,00
  Soit 70% d'économies
- elle est **éteinte 10 jours fériés et 3 semaines de vacances :** $564,00
  Soit 73% d'économies

Multiplié par quelques dizaines de machines virtuelles (par exemple sur des environnements de test ou de développement) l'impact n'est pas négligeable.

UTILISER MIEUX

Si la première des habitudes à changer concerne l'extinction des ressources lors des périodes d'inutilisation, la seconde porte sur le dimensionnement des serveurs.Sur ce point, le raisonnement se fondait sur la quantité de mémoire, le nombre et la puissance des processeurs, ainsi que sur les variations de charge dans une journée ou sur une période donnée. Il s'y ajoutait souvent une composante de montée en charge ou d'évolution dans le temps : on dimensionnait ainsi sur la crête à venir - en intégrant le fait que « les serveurs devront tenir un certain temps » - souvent 3 ans. Résultat, l'utilisation moyenne des serveurs dans les datacenters est souvent faible, cumulant ces multiples réserves de puissance.

Le problème est que de nombreux dimensionnements dans le Cloud intègrent ces pratiques, alors qu'il y a de multiples raisons de ne pas le faire :

- Facilité à changer d'instance en cas d'évolution du besoin dans le temps - pas de nécessité de prévoir à trop long terme ;
- Pour les applications supportant le scale-out (passage à l'échelle), possibilité d'ajouter des ressources - éventuellement via auto-scale ;
- Pour les applications ne le supportant pas, possibilité de «burster», c'est à dire d'augmenter temporairement la capacité des ressources utilisées ; ce sont les instances T chez AWS, l'utilisation intensive des instances F1 chez GCP et la série B des VMs d'Azure

UN EXEMPLE

Allons chez GCP cette fois, prenons une n1-standard-4, équivalent de la D4 V3 de chez Azure. Son coût annuel est de $2 138,57 (août 2019, Europe de l'ouest 3, Linux) soit un coût mensuel d'environ $178,21.

- Si elle a été dimensionnée trop largement dans la perspective d'une évolution du business, elle fonctionnera sans doute sur la machine directement inférieure (deux fois plus petite) : la n1-standard-2 coutera la moitié ; par exemple, commencer sur une n1-standard-2 pendant 9 mois et ne passer en n1-standard-4 uniquement lorsque c'est nécessaire coutera $1 336,61 sur un an, soit 37% d'économies.

- Si elle a été dimensionnée trop largement dans la perspective d'un appel de charge à hauteur de 1 à 2 heures par jour, mais que l'application autorise le scale-out, elle pourra également tourner sur une n1-standard-2 la majorité du temps et 2 n1-standard-2 lorsque nécessaire (coût annuel $1 158,39 ; soit 46% d'économies). Si l'application ne le permet pas, elle pourra peut-être tourner sur une machine personnalisée de type f1-micro aux performances assez proches, et «burster» lorsque nécessaire afin de délivrer de la puissance supplémentaire occasionnellement.

Nous voudrions conclure cette section par un type de rightsizing propre au Cloud, qui est la chasse aux instances d'anciennes générations. Par exemple, AWS fait évoluer rapidement sa gamme d'instances, les nouvelles offrant généralement des performances meilleures pour un coût légèrement inférieur.

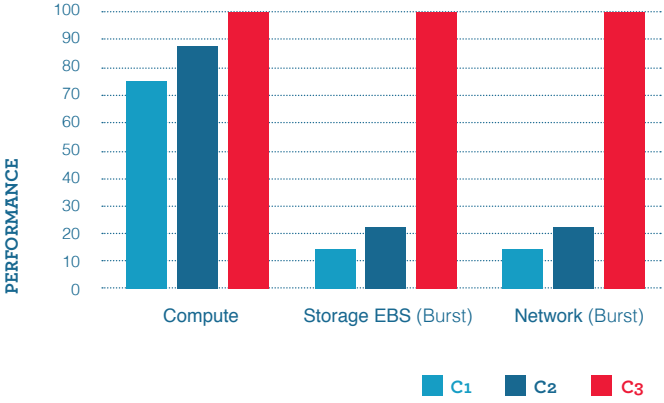
Prenons les instances de la famille « calcul optimisé » (type c), très répandues dans les configurations client : les instances c5 ont été lancées en fin d'année 2017, les parcs sont désormais constitués de c5, mais on trouve encore des c4 (les c3 ont complétement disparu).

- **Prix d'une c3.xlarge** (7.5 Gb RAM, 4 vCPUs) avec Linux en Irlande : $0,239/heure (dernier prix en date).
- **Prix d'une c4.xlarge** (7.5 Gb RAM, 4 vCPUs) avec Linux en Irlande : $0,226/heure (août 2019), soit 5% moins cher
- **Prix d'une c5.xlarge** (8 Gb RAM, 4 vCPUs) avec Linux en Irlande : $0,192/heure (août 2019), soit 15% moins cher

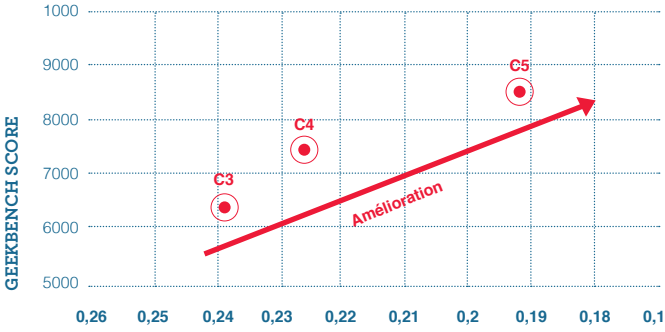
Or les c5 sont plus récentes et plus performantes que les c4, lesquelles le sont plus que les c3. Par exemple, en appliquant un benchmark avec l'outil Geekbench 3 (21 tests integer, floating point et memory), nous obtenons les résultats suivants :

- La **c3** obtient un score de 6421
- La **c4** obtient un score de 7504 (17% plus puissante)
- La **c5** obtient un score de 8485 (32% plus puissante que la c3 et 13% plus puissante que la c4)

COMPARATIF DES PERFORMANCES INTERGÉNÉRATIONNELLES (famille c)



COMPÉTITIVITÉ ÉCONOMIQUE ($/heure)



Il y a donc possibilité de gagner de la puissance en baissant la facture (voire de baisser beaucoup la facture, par exemple en profitant pour passer sur une c5.large si le taux d'utilisation le permet). Notons néanmoins que la question se pose moins dans les régions récentes : par exemple, la région Europe de l'Ouest 3 Paris, ouverte fin 2017, dispose uniquement de c5 pour le calcul optimisé (mais beaucoup de clients français utilisent la région Irlande par habitude).

PAYER MOINS CHER

La dernière étape de l'optimisation est la réservation. Cela concerne uniquement les services de calcul (serveurs) qu'ils soient sous forme de machines, de bases de données ou encore de services plus managés comme des conteneurs.

Les réservations sont une mécanique du modèle tarifaire qui fonctionne de manière différente pour chacun des Cloud providers :

- **L'offre AWS est la plus complète :** les réductions sont les plus importantes mais la multitude de paramètres augmente la complexité
- **Azure dispose d'une offre du même type,** plus simple, avec des réductions plus faibles
- En ce qui concerne GCP, **les réservations sont détachées des réductions.** Les réservations sont faites pour réserver la capacité, cela garantit donc d'avoir la ressource disponible contre un engagement financier sur une durée. S'ajoute à cela des engagements de consommation donnant lieu à des réductions.

Illustrons cela : je réserve une instance n1-standard-1, cela veut dire que tant que je ne supprime pas ma réservation, GCP m'alloue la puissance de calcul et je la paie, utilisée ou non. En parallèle, je peux m'engager avec GCP sur 1 an par exemple, leur promettant de leur payer chaque heure de cette année, en retour, j'obtiens une réduction pour cette instance.

	AWS	AZURE	GCP
Frais initiaux	Sans / Partiels / Totaux	Totaux	Sans
Caractéristiques	Durée, Région, Type d'instance, OS	Durée, Région, Type d'instance	Région, # de vCPUs, # de GBs de RAM
Capacité de changement	Convertible : forte capacité ; Standard : faible capacité	Forte capacité	Changements automatiques
Annulation	Aucune	12% de pénalité	Aucune
Cumul des réductions	Volume des réservations, EDP	Aucun	Aucun

Intéressons-nous au cas le plus complexe d'AWS.

Le Cloud d'Amazon offre la possibilité de bénéficier d'une réduction tarifaire significative : de 30 à 60% selon la durée et les modalités de paiement en contrepartie d'un engagement pour une période donnée (1 ou 3 ans), sur un type de ressources donné.

Voici par exemple ce que cela donne pour une instance m5.xlarge avec Linux en France (août 2019) :

- Son coût horaire est de $0,224/heure, soit environ $1 962/an ;
- Si on réserve sur un an (mode « standard ») en payant tout en avance, on paiera $1 286 pour un an, soit 34% d'économies ;
- Cette même réservation sur un an en payant partiellement d'avance donne un premier paiement de $656 et un paiement horaire de $0,076/heure, soit un total annuel de $1 313, soit une économie de 33% ;
- Enfin, la réservation sur 1 an sans paiement upfront donne un paiement horaire de $0,157, soit un total annuel de $1 375 et une économie de 30%.

Il y a bien un engagement pris sur un an (dans l'exemple) à payer la totalité des heures, qu'elles soient consommées ou pas. Comme il y a engagement à payer toutes les heures, il est évident qu'il faut être assez certain de la permanence du besoin. De plus, la réservation offre un tarif préférentiel sur un tarif figé dans le temps, or les prix sont souvent revus à la baisse, ce qui réduit les économies faites avec les réservations.

On peut donc se dire à première vue que c'est intéressant mais totalement contradictoire avec les objectifs de flexibilité précédemment vus.En fait non, car AWS a introduit une grande dose de subtilité (mais aussi de complexité) qui permet de combiner flexibilité et réservation. Les réservations sont consommables par toutes les instances du type choisi et sont partagées dans une organisation. Ainsi, plusieurs instances sur des comptes différents peuvent se relayer afin de consommer la même réservation.

Le concept de la réservation dans le Cloud Computing porte d'abord sur un engagement financier. Ainsi, les providers permettent de plus ou moins les modifier en gardant toujours à l'esprit qu'une modification doit entrainer un engagement financier supérieur :

Pour Azure, Il existe un type de réservations (All upfront avec une durée de 1 ou 3 ans), cependant, cela réinitialise le terme. Azure permet même d'annuler une réservation moyennant un coût de 12%. Côté Google, Les engagements ont une granularité toute autre. Ils se basent sur les processeurs virtuels et la mémoire RAM. Ainsi, la promesse est sur une utilisation de ressource qui est vraiment très flexible.

Enfin, AWS dispose de deux offres de réservations :

- Les **réservations standards** qui offrent les économies les plus fortes mais qui sont peu flexibles
- Les **réservations convertibles** qui sont fortement modifiables mais offrent des économies moindres.

Le tableau ci-dessous illustre la situation à ce jour (août 2019) des réservations AWS

Type	Standard	Convertible
Durée	1 an - 3 ans	1 an - 3 ans
Paiement	Sans frais initiaux / Frais initiaux partiels (50%) / Tous les frais initiaux	
Économie	~40-60%	~30-50%
Changement OS	Non	Oui
Changement taille	Non (sauf Linux)	Oui
Changement famille	Non	Oui
Éligible aux baisses de prix	Non	Oui

Comment se repérer et faire son choix dans tout cela ?

Chaque entreprise dispose de processus différents dans les choix budgétaires et leur mise en place. Ainsi, il semble nécessaire de définir un ensemble de règles personnalisées afin de bénéficier au maximum des réservations. D'une manière générale, la question peut se poser pour des environnements très flexibles et fréquemment éteints (dev & test, fonctionnement limité aux heures ouvrées) mais les instances réservées sont globalement rentables dès que la charge est stable dans le temps.

PETITE DIGRESSION NON FINANCIÈRE SUR LES INSTANCES RÉSERVÉES

Avant d'être une option tarifaire, les instances réservées sont.... Réservées.
Il faut se souvenir que le Cloud ne garantit pas à un client que les ressources qu'il souhaite acheter soient effectivement disponibles «en stock» : il peut arriver que tel ou tel modèle d'instance ne soit pas disponible à un instant donné sur telle zone de disponibilité, dans telle région (c'est rare, mais nous l'avons vu). On peut d'ailleurs imaginer des situations (sinistre de zone) générant un afflux de demandes simultanées qui ne pourraient être servies.
La réservation d'instance permet de pallier à cela et de garantir l'accès à la ressource, puisqu'elle est effectivement «réservée» dans le capacity planning des providers ; c'est donc une caractéristique intéressante en soi, par exemple pour une solution de Disaster Recovery.

Ceci n'est bien sûr qu'un survol du sujet des réservations ; pour être exhaustif, il faudrait aussi parler du mode de fonctionnement des coupons, de l'importance de la gouvernance des réservations, des optimisations possibles grâce aux convertibles, de la mutualisation… Sujets que nous aborderons volontiers avec vous mais qui nécessiteraient trop de pages pour avoir leur place dans ce livret. En ce qui concerne les réservations, souvenez-vous que « l'inaction coûte souvent plus cher qu'une petite erreur ». Il ne faut donc pas tergiverser !

ACTIONS STRUCTURELLES

L'application des mesures quick wins fait généralement gagner 20-25% de la facture lors des actions initiales, et permet surtout de maintenir l'efficacité dans le temps. Il y a toutefois de nombreux autres leviers d'action, et nous proposons ci-dessous une liste non limitative :

Dépasser les machines virtuelles

Les coûts engendrés par les instances de calcul, les principaux services de « compute », représentent généralement une grande partie de la facture Cloud mais cela ne signifie pas qu'il n'y a rien d'autre à optimiser : le tiering du stockage offre de nombreuses possibilités, la tarification par réservation ne s'applique pas qu'aux serveurs virtuels (aussi aux bases de données) et les coûts de réseau justifieraient un document dédié à eux seuls (minimisation du transfert inter-régions, utilisation d'adresses IP privées, etc.).

Buy VS Make

Utiliser le Cloud suppose déjà une ouverture à consommer un service plutôt que le construire. Les Cloud providers permettent d'aller très loin dans cette approche, avec de très nombreux services applicatifs qui produisent des économies de licences et de montée en charge.
Sur de nombreux services managés, il n'existe quasiment pas de différences entre les Cloud providers. Par exemple, au niveau des infrastructures, les services de bases de données sont sensiblement identiques. Ils permettent de créer des moteurs de base de données sécurisés, incluant la redondance et les sauvegardes. Outre la simplification apportée, cela génère des économies directes (notamment au niveau des licences, grâce au modèle de paiement à l'usage) et indirectes (charge de service évitée pour les backups, par exemple) ; nous n'évoquons pas ici les

migrations de base de données (Oracle vers Aurora, Cloud SQL ou Azure SQL Databases, par exemple), mais c'est évidemment une source d'économies considérables.
Au niveau du développement, il est possible de maximiser l'usage des ressources via l'utilisation de conteneurs, ou même de s'affranchir de l'infrastructure par une approche «function as a service» avec les différents services serverless (AWS Lambda, Google Cloud Functions ou encore Azure Functions). Les niveaux d'intégration sont différents :

	Conteneurs	Serveurless
Infrastructure	Création et gestion	Abstraction
Langages de programmation	Totalement libre	Astreint à une certaine liste
Reversibilité (possibilité de retour en arrière)	Possible	Faible
Avantage économique	Via densification	Facturation à l'événement

Les résultats en termes de facilité et de rapidité de déploiement ou d'économies sont en revanche dans les deux cas importants. Cela semble donc très avantageux. Bien évidemment cela doit être étudié au cas par cas et dépend des caractéristiques de l'application.

Oser les régimes particuliers

Le Cloud offre une multitude de possibilités, notamment en proposant les instances disponibles (par rapport à la demande générale) à très bas prix. Les mécanismes sont différents pour chaque Cloud provider. Pour AWS, le prix

des « Spot Instances » dépend de l'offre et la demande - on constate généralement une fourchette de 60% à 90% d'économies par rapport au prix à la demande. Pour Google, les « instances préemptives » coûtent 20% du prix mais ne peuvent durer plus de 24h. Azure et ses instances basses priorités suivent la stratégie de Google. Il y a évidemment une contrepartie : ces instances peuvent être interrompues avec un préavis très court (quelques minutes) C'est donc un usage réservé à des charges interruptibles ou (et surtout) à des charges répartissables par design sur une flotte d'instances (des automatismes permettent de définir de tels groupes).

Consolider les comptes

La plupart des services managés offrent un discount selon le volume.Ceci vaut également pour les frais de support (facturation en pourcentage de la facture globale avec des tranches). Il y a donc un intérêt évident à lier tous les comptes dans un compte de facturation consolidé - la difficulté consistant souvent à identifier ceux qui ont peut-être été ouverts hors des circuits traditionnels de l'IT.

››› LA GOUVERNANCE

Comme nous l'avons relevé plus haut, une des difficultés du Cloud FinOps vient de la multiplicité des acteurs, des comptes, des types de services, etc. Bref, ce qui semble évident au niveau individuel est rendu complexe par le nombre de combinaisons possibles et le volume de cas à traiter. Dès qu'une organisation grandit, il n'y a pas d'alternative : il faut structurer l'approche du Cloud. C'est ce que nous appelons la Gouvernance du Cloud.


EUH ... ON L'A FAIT ÇA OU PAS ??
EUH ... QUI DEVAIT S'EN OCCUPER DÉJÀ ?

LE SHOWBACK : PREMIERE ETAPE VERS UNE DEPENSE CLOUD CONTROLEE

Les dépenses dont personne ne se sent responsable grandissent vite. Il est à ce titre salutaire de fournir à chaque classe d'utilisateurs une information claire sur les ressources dont il a la charge et sur lesquelles il peut agir :

- Suivi quotidien des ressources sous ou pas utilisées, pour les équipes techniques,
- Suivi technique et financier régulier sur les différents services utilisés, et leur tendance, pour le responsable de production,
- Bilan financier au minimum hebdomadaire pour les responsables de centres de coût (applications, domaines ou métiers), et le contrôleur de gestion,
- Heatmap (matrice présentant des taux d'utilisation pour un groupe de ressources par jour/heure) pour des responsables applicatifs afin d'aider le dimensionnement.

Sans suivi spécifique, cette information est trop souvent mensuelle, à l'occasion de la facturation, ce qui fait perdre l'occasion d'enclencher des actions au fil de leur nécessité (c'est d'ailleurs pour cela que l'optimisation de coûts Cloud n'est pas une action «coup de poing» mais bien une discipline quotidienne à mettre en place et faire vivre...).

LA SCORECARD CLOUD FINOPS, UN EXCELLENT OUTIL DE COMMUNICATION ET D'ALIGNEMENT AVEC LES MÉTIERS ET LES FONCTIONS SUPPORT.
Plus la facture Cloud grandit, plus il devient fondamental de démontrer aux différents acteurs de l'entreprise que cet argent n'est pas dépensé en vain. Il est même judicieux de démontrer qu'au contraire, la quantité de workloads ou le nombre de services ont augmenté comparativement plus vite que la facture. Pour cela, la solution idéale est de définir une ou plusieurs métriques « business », sur lesquelles indexer la facture ou certains de ses composants : coût du clic ou d'un affichage de message pour des services internet, coût de traitement d'une transaction ou d'un processus, coût de l'utilisateur, etc.
D'autres KPIs peuvent être suivis avec intérêt : taux de couverture RI, taux de perte d'heures RI, ratio d'élasticité (coût horaire moyen en jour contre la nuit, semaine contre week-end), etc.La définition d'objectifs chiffrés et d'actions associées, ainsi que leur suivi régulier, sont les meilleurs moyens d'assurer une amélioration continue de la gestion financière du Cloud.

En somme, le showback est la première étape de la prise de conscience de la dépense Cloud. Dans certaines entreprises, ce reporting peut se transformer en chargeback, c'est-à-dire de réallouer les coûts aux différentes entités qui commandent la dépense. Faire comprendre le FinOps en montrant la dépense est une chose, le faire en allant directement chercher dans le porte-monnaie en est une autre.

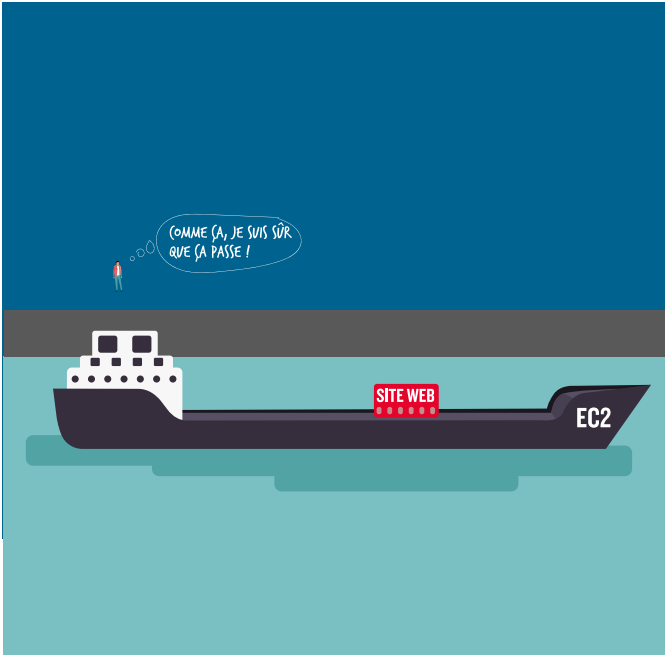
DÉFINIR UNE MATRICE DE RESPONSABILITÉ

Conséquence du point précédent, nous relevons que les clients qui sont les plus performants sur le sujet sont ceux qui en ont clairement confié la responsabilité à quelqu'un. Cela ne signifie pas que c'est un travail à plein temps (mais cela peut l'être chez les grands utilisateurs de Cloud, et les gains le justifient pleinement), mais compter sur la bonne volonté ou le comportement vertueux des uns et des autres ne suffit généralement pas.

Les différentes décisions à prendre dans le cadre du FinOps nécessitent différentes connaissances / compétences :

Le rightsizing

Ce dernier implique de modifier les ressources et peut potentiellement avoir une influence sur les charges de travail. Il faut donc avoir une bonne idée de ce qui se passe avant de modifier quoi que ce soit. Il faut aussi être sûr que la personne dont c'est la tâche le fasse en connaissance de cause.



Les réservations

Elles nécessitent une vue à plus long terme de l'évolution des applications, une connaissance des moyens financiers mobilisables, et dans la mesure du possible une standardisation de l'approche au niveau de l'entreprise. Ainsi, il n'est pas du rôle du développeur de se pencher sur ce type de question. Cela n'empêche pas de faire remonter des KPI pertinents à un chef de projet, l'amenant à réfléchir sur des optimisations possibles et donc à des possibles réservations sur les ressources que son équipe utilise.

De nombreux modèles de gouvernance sont envisageables, et doivent être mûrement réfléchis afin de permettre autant que possible à l'organisation d'atteindre ses objectifs : motiver l'ensemble des équipes à s'impliquer dans le FinOps, maximiser l'agilité de l'organisation, aller vers plus de standardisation, garder la main en central sur l'allocation des coûts Cloud…

DÉFINIR LES BUDGETS ET LES SUIVRE

Un financier considérera l'étape de l'établissement d'un budget comme nécessaire. Elle est cependant encore assez peu répandue dans le domaine du Cloud. A la complexité du catalogue de services s'ajoute la difficulté d'anticiper le dimensionnement de certains paramètres (notamment concernant les flux de données), l'implication d'acteurs pas toujours formés aux estimations budgétaires…

Toutes ces plus ou moins bonnes raisons de se passer d'un budget ont 2 conséquences désagréables :

- La première est une absence totale de visibilité financière : on avance à tâtons !
- La seconde est que, sans estimation originelle, il est difficile d'identifier une déviation, et donc de l'adresser ; il existe donc un risque de sous-optimisation !

Il faut donc créer un processus budgétaire pour le Cloud afin de permettre un suivi. Et la difficulté s'est déplacée par rapport au modèle on-premise. Là où celui-ci présentait des coûts globaux assez aisément planifiables : les infrastructures étant généralement renouvelées tous les 5 ans, la difficulté résidait dans la manière d'allouer ces coûts aux différents utilisateurs.

Dans le Cloud, les coûts sont majoritairement directs ; néanmoins le nombre d'écueils à éviter est très important. La facturation à la minute voire à la seconde impose d'estimer la consommation de l'application précisémment sans pour autant disposer d'autres informations que l'architecture. Cela implique de ne rien oublier ni de compter deux fois la même chose !
Ce n'est donc pas un travail évident et il faut se rapprocher au maximum du réel.

DÉFINIR DES RÈGLES ET LES PROCESSUS D'AUTOMATISATION QUI EN DÉCOULENT

Les règles

La création de nouvelles ressources est d'une rapidité déconcertante sur le Cloud, c'est un avantage considérable pour les protagonistes techniques. Cependant, c'est un cauchemar pour ceux chargés de réguler la dépense. .

Le nombre d'instances croit d'autant plus que les gens adoptent la méthodologie Cloud. Le nombre de services et le volume de chacun ne fait qu'augmenter et si suivre quelques instances est assez facile à la main, suivre plusieurs milliers d'éléments devient un défi. De la même manière que l'on définit des **standards** dans l'industrie, pour bénéficier tant des retours d'expériences opérationnelles que des éventuels leviers financiers, il faut en définir pour le Cloud.

Un certain type de charge de travail doit systématiquement appeler une certaine configuration. Et pour s'assurer que ces standards sont bien appliqués, rien de tel que de créer une nomenclature des profils de performance pour les ressources et ainsi, pour chaque profil, décider d'appliquer ou non une optimisation.

Cela passera la plupart du temps par des **tags**. Les Cloud providers permettent effectivement de définir des tags pour chaque ressource, couvrant aussi bien des informations techniques que fonctionnelles.
Il est donc possible de relier chaque ressource à différents agrégats : application, environnements (exemples : développement, intégration, production, etc.), service utilisateur (et donc payeur), etc.

Les difficultés se situent dans la taxonomie des tags (cohabitation de «dev», «DEV», «dévelop» etc.), dans leur non-systématisation (dans une perspective de facturation, il faut que chaque ressource ait un tag permettant l'allocation de coûts - faute de quoi on s'oblige à de pénibles retraitements de coûts sous-absorbés), et dans la consolidation d'informations provenant de plusieurs comptes selon des systèmes de tags communs.
A noter que des automatisations peuvent être mises en place afin de « forcer » le tagging par les utilisateurs puis à terme de le mettre en place de manière systématique. Une couche de tags virtuels peut éventuellement être mise en place afin de faciliter l'association des ressources.

Compte tenu de la rigidité qu'apporte la création de ces règles, il est impératif de les avoir réfléchies au regard des objectifs visés, et partagée avec les différents protagonistes du Cloud. La liberté des développeurs est essentielle à l'agilité de l'entreprise et il n'est pas question de leur imposer un processus long et fastidieux pour créer une instance de test.

Cependant, rien n'empêche de vérifier que tout est utilisé correctement quelques temps après la création.

L'AUTOMATISATION

D'une manière générale, le Cloud perd son avantage de passage à l'échelle sans automatisation... Et pour aller plus loin, faire du management de Cloud sans automatisation semble être une complication peu nécessaire.

Les FinOps n'échappent pas à la règle et il y a de nombreux sujets à creuser : détection automatique de ressources insuffisamment utilisées ou orphelines, détection de ressources non taguées, envoi d'alertes automatisées, automatisation de start/stop, surveillance et automatisation de recréation d'instances spot, automatisation de changement de taille d'instances, etc.

Si certaines de ces fonctions sont facilitées par des outils du marché, la plupart peuvent être mises en place avec les outils natifs des Cloud providers moyennant un effort de mise en place plus ou moins élevé en fonction de leur complexité.
Il peut être jugé préférable pour certaines actions, au moins dans un premier temps, de rester en mode manuel pour garder un meilleur contrôle, mais le passage à l'échelle impliquera nécessairement une dose d'automatisation, faute de quoi les représentants FinOps seront intégralement absorbés par l'application des décisions.

DÉFINIR L'OUTILLAGE NÉCESSAIRE COMPTE TENU DES OBJECTIFS

Tout est une question de temps dans le Cloud. Consommer moins longtemps, c'est réduire ses coûts. Ainsi, il existe plusieurs outils sur le marché dont la principale qualité est de permettre une prise en main (et donc un impact financier) plus rapide : CloudHealth, CloudCheckr, Cloudability et Teevity (seul français de la liste, construit sur la base de la plateforme opensource ICE issue du projet FinOps de Netflix).

Ils ont tous leurs forces ou faiblesses (orientation reporting ou optimisation, couverture multicloud, fonctionnalités, ergonomie, tarification, souplesse et personnalisation…) et correspondent à des typologies différentes de taille de projet client : on ne cherchera pas le même outil pour rationaliser l'usage d'un environnement mono-compte de 15 ou 20 k€ / mois que pour établir une facturation multicloud / multi-comptes de plusieurs centaines de k€ mensuelles, avec introduction d'unités d'œuvre métier.
Chez Gekko, nous en utilisons plusieurs, en fonction des cas que nous rencontrons ; nous avons mis en place l'un d'entre eux afin d'administrer la fonction FinOps (reporting, facturation, optimisation) de certains de nos clients.

Nous voyons également quelques cas (généralement de grands comptes), où des outillages «maison» ont été mis en place - effectivement, des outils natifs existent et moyennant un peu d'huile de coude, il est possible d'en tirer les bonnes conclusions. Le risque avec de tels outillages réside principalement dans leur maintenabilité et leur évolutivité (et le coût associé).

VERS UNE APPROPRIATION DE LA CULTURE CLOUD FINOPS PAR TOUS LES ACTEURS

De manière complémentaire, le travail de Cloud FinOps ne s'arrête pas aux optimisation financières à court terme. Il s'agit aussi d'accompagner les entreprises dans leur passage à une mentalité Cloud FinOps. C'est-à-dire d'apprendre à consommer les services managés de manière optimale, de connaître les cas d'usage, de perdre moins de temps à choisir parmi les multiples offres, etc. Tout cela représente une base de connaissance considérable que le FinOps a pour mission de partager.

Dans cette vue plus large, le représentant FinOps a aussi une mission de conduite du changement : former les équipes aux bonnes pratiques Cloud afin de tendre vers le comportement vertueux de chacun, de l'architecte au financier, du développeur à l'acheteur.

EN CONCLUSION

Nous avons rencontré beaucoup de clients qui souhaitaient une analyse de leur facturation assortie de recommandations, mais qui voyaient mal l'intérêt
de s'outiller dans la durée. Il est vrai que la campagne initiale est généralement spectaculaire, avec ses 20-25% de gains. L'expérience montre qu'ils peuvent être vite perdus. D'abord, ils peuvent tout simplement ne jamais être mis en œuvre… La nature humaine étant ce qu'elle est, un suivi dans la durée favorise l'exécution effective des actions, qui auraient pu être vite oubliées sinon. Ensuite, à moins d'avoir une production totalement stable (mais alors, faut-il vraiment être dans le Cloud ?!), de nouveaux services vont arriver, de nouveaux composants vont être intégrés et il faudra les optimiser ; tout ceci sans même parler de tirer profit des évolutions tarifaires des différents Cloud providers. Notre sujet relève donc bien de la course de fond et non du sprint.


AH OUAIS ON SE SENT BIEN LÀ
GOUVERNANCE
MIGRATION
DEVOPS
FINOPS
SERVERLESS
AUTOMATISATION
MULTICLOUD
VM
BUILD
LOAD BALANCER
DISASTER RECOVERY

Nous espérons que ce petit document vous aura intéressé, ou au moins convaincu de l'importance du sujet. Si c'est le cas et que nous pouvons vous aider, contactez-nous, nous serons ravis de le faire.
Si vous souhaitez contribuer au développement de nos réflexions et expériences sur le sujet, contactez-nous également ! L'utilisation du Cloud à grande échelle dans les entreprises est encore un phénomène récent, la discipline de Cloud FinOps est encore jeune et nous sommes pleinement conscients qu'il reste beaucoup à faire et à co-construire.

Et si ce n'est pas le cas, peut-être serez-vous plus intéressé par notre anti-guide final ?!

5 MOYENS SÛRS DE PERDRE DE L'ARGENT AVEC LE CLOUD :

1

Dimensionnez comme d'habitude

«Nos applications ont toujours eu besoin de 8 vCPUs et 32 Gb de RAM sur le datacenter, donc si on veut être sûrs que ça marche toujours, il faut la même chose dans le Cloud. D'ailleurs on sera bien contents s'il y a un gros pic comme on avait eu ce fameux jour l'an dernier».

2

Faites comme d'habitude

«Notre IT fonctionne en 24x7, il n'y a rien qu'on peut arrêter. Bon peut-être le développement et les tests la nuit et le week-end. Et puis cette appli aussi.
Mais c'est compliqué. Et on n'est pas sûrs que ça redémarrera. Et puis ça ne va pas chercher bien loin».

3

Comptez sur la discipline des utilisateurs

«OK ils ont démarré plein d'instances, mais c'était pour essayer - d'ailleurs le Cloud c'est fait pour ça, non ? Et nos équipes savent se tenir, ils les arrêteront quand ils n'en auront plus besoin».

4

Attendez que la situation soit stable pour réserver des instances

«Là on a bien ces instances qui sont 'on' depuis 6 mois, c'est dommage on aurait déjà gagné de l'argent avec des réservations, mais on n'est pas sûrs qu'on ne va pas en changer donc on attend ; et il faut encore qu'on y voit clair avec les futurs besoins avant de s'engager.»

5

Laissez les budgets où ils sont

« Ce compte est associé au projet big data, c'est noyé dans l'ensemble et porté par le métier ; celui-ci est dans le budget de développement, tout le monde contribue ; et celui-là est dans les frais généraux, personne n'y verra rien. »

À PROPOS DE GEKKO

Société de conseil et d'intégration, Gekko est spécialisée dans la stratégie et la réalisation des migrations vers le Cloud.

Nous accompagnons nos clients dans la conception, le déploiement et la maintenance d'une infrastructure Cloud flexible, connectée et sécurisée, 100% DevOps. En combinant notre expertise technique pointue et notre profonde connaissance de la production informatique, nous vous permettons de tirer le meilleur parti du Cloud.
Créée en 2015, Gekko est Advanced Partner AWS habilitée Well-Architected, elle est la première société française à obtenir la Competency Storage AWS et compte aujourd'hui plus de 100 consultants spécialisés AWS, DevOps et micro-services.

gekko
les ingénieurs du cloud

GEKKO
12 rue d'Alsace,
92300 Levallois-Perret
T + 33 (0) 158 744 600

WWW.GEKKO.FR