La latence est la nouvelle panne

POURQUOI LA VITESSE EST LE NOUVEL ENJEU SUR LA TABLE

Sommaire général

Étant donné que les entreprises exploitent de plus en plus d'applications hébergées sur différentes infrastructures, dans différentes zones géographiques et auprès de différents fournisseurs, la vitesse à laquelle les utilisateurs finaux peuvent accéder à ces applications est mise sous pression.

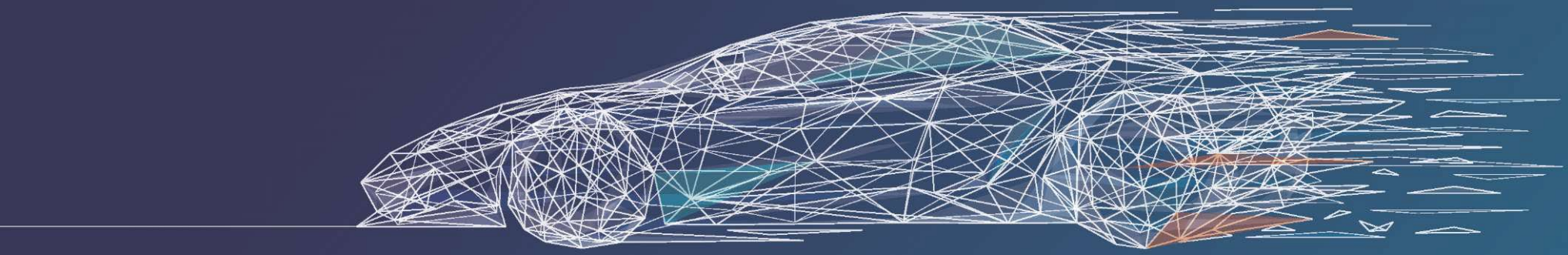
Si l'on ajoute à cela le fait que les applications sont aujourd'hui constituées de bien des composants différents, on obtient la recette d'une expérience utilisateur dégradée.

Cette double caractéristique, à savoir modularité des applications et complexité des infrastructures, peut avoir comme conséquence directe une mauvaise qualité des applications.

C'est pour cela que la vitesse de la couche de données, la couche horizontale commune à toute l'application, est essentielle.

Pouvoir utiliser une couche de données géographiquement répliquée, tout en évitant les problèmes d'incohérence des données, est un défi que tous les responsables informatiques doivent résoudre.

En exploitant une couche de données qui unifie vos données à travers le cloud et le monde, les entreprises peuvent surmonter certaines limites inhérentes qui ont mis au défi les équipes technologiques depuis des décennies et offrir de meilleures expériences à leurs utilisateurs finaux.



Introduction

Les équipes digitales ont passé la dernière décennie à s'assurer que leurs actifs numériques étaient disponibles à toutmoment, et elles y sont largement parvenues ! La haute disponibilité est désormais la norme.

Les entreprises ont, en partie, atteint ce haut niveau de dématérialisation et de disponibilité en tirant parti des avantages qu'apporte le cloud : facilité d'évolutivité, modularité des services, modèles architecturaux plus raffinés. Ces caractéristiques permettent toutes d'obtenir des résultats positifs, mais elles le font avec, comme revers de la médaille, une complexité accrue. Cette complexité a d'abord eu un impact important en termes de disponibilité et a donné lieu à ce que nous appelons l'Ère de la disponibilité. Cependant, au fur et à mesure que les entreprises comprennent mieux comment offrir une haute disponibilité, elles constatent qu'il y a encore d'autres problèmes à résoudre.

Toutefois, à présent, l'Ère de la disponibilité commence à s'estomper, car les entreprises cherchent de plus en plus à réduire la latence, prochaine clé pour déverrouiller les résultats recherchés.

Elles comprennent de plus en plus que les produits et services lents peuvent tout aussi bien ne pas du tout être disponibles, la latence étant la nouvelle panne.

Malheureusement, la résolution des problèmes de latence est souvent plus difficile que de créer une haute disponibilité. Si la disponibilité peut être améliorée grâce à une bonne ingénierie, de plus grands niveaux de redondance, et une meilleure surveillance et visibilité, le problème de la latence est limité par les lois mêmes de la physique.

Pour réduire la latence autant que faire se peut, les entreprises doivent comprendre ce qu'elle est et les facteurs qui y contribuent, et disposer de directives claires et définitives pour réduire, autant que possible, celle qui caractérise leurs applications et sites web.

Si la latence est la nouvelle panne, voici l'intelligence dont vous avez besoin pour proposer le plus faible niveau physiquement possible.

Les entreprises comprennent de plus en plus que les produits et services lents peuvent tout aussi bien ne pas du tout être disponibles, la latence étant la nouvelle panne.

Ce ne sont pas les gros qui mangent les petits, c'est le rapide qui absorbe le lent

Aller vite est la nouvelle normalité. Si l'analyse et la prudence étaient autrefois de mise, la réalité opérationnelle d'aujourd'hui est que, pour rester en tête de la concurrence, les entreprises doivent innover plus vite que jamais. Le paysage opérationnel de chaque entreprise évolue rapidement, et ce sont celles qui sauront le mieux réagir à ce dynamisme qui réussiront.

POURQUOI L'AGILITÉ EST SI IMPORTANTE : LE TEMPS EST ESSENTIEL

Le monde dans lequel nous vivons aujourd'hui est nettement différent de celui d'il y a seulement quelques années, etle rythme du changement ne cesse de s'accélérer. Dans ce contexte, donner aux entreprises la capacité d'agir rapidement est plus important que jamais. Il est essentiel de comprendre les changements qui se produisent dans la société pour mieux saisir à quel point il est important d'agir vite.

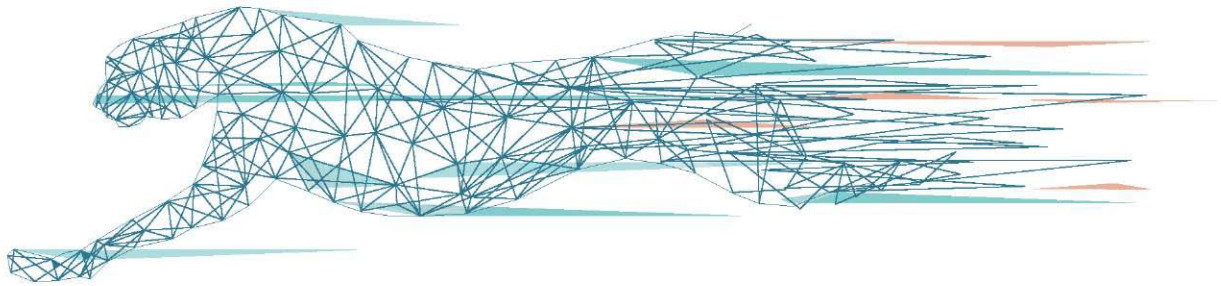
En 2011, le célèbre entrepreneur, investisseur et membre du comité directeur Marc Andreessen (l'inventeur du navigateur web Netscape) a écrit un désormais célèbre article d'opinion pour ***Le Wall Street Journal,*** expliquant Pourquoi le logiciel mange le monde.

Dans son essai, Andreessen présente sa théorie sur l'ampleur, la portée et la vitesse de ce changement, suggérant que : « Nous sommes au beau milieu d'un vaste et spectaculaire changement technologique et économique dans lequel les sociétés de logiciels sont prêtes à prendre le contrôle de vastes pans de l'économie ».

Mais si le fait que ce changement est très important, c'est la vitesse de ce changement qui est la plus pertinente ici. La capacité de procéder à des modifications, de tirer parti d'une variété d'outils et de fournir la meilleure expérience possible à l'utilisateur final est un élément fondamental de la capacité à agir vite. Si tout cela ressemble beaucoup au mode de fonctionnement des entreprises de la Silicon Valley, c'est logique. Le fait est qu'une grande partie des entreprises qui réussissent àperturber les secteurs traditionnels ressemblent et se sentent de plus en plus comme des entreprisesde la technologie .

Il y a presque dix ans, Andreessen écrivait cet essai, et nombre de ses prédictions se sont réalisées. Bien qu'il soit devenu stéréotypé d'utiliser Tesla, Uber, Lyft, Netflix et Airbnb comme exemples de la perturbation numérique, on peut dire sans risque que les cadres des entreprises de taxi et d'hôtellerie ont été frappés par un raz-de-marée dont les proportions sont sans précédent. Au-delà du cliché, ce qu'il faut retenir, c'est l'ampleur des efforts déployés par ces entreprises pour offrir l'expérience client la plus rapide possible sur leurs applications : la vitesse compte vraiment.

Il va sans dire que, pour aller vite dans une entreprise, il faut fournir des expériences numériques qui présentent elles-mêmes ces attributs de vitesse. La latence est le nouveau facteur qui bloque la transformation organisationnelle.



Changer le monde en passant au numérique

On trouve de très nombreux exemples d'entreprises traditionnelles qui fondent leurs espoirs de réussite future sur un passage au numérique. Il est intéressant de se pencher sur quelques exemples pour se faire une idée de cette ampleur.

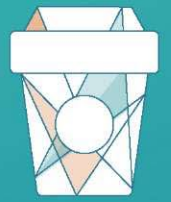
DIGITAL JOE : STARBUCKS PASSE AVANT TOUT AU NUMÉRIQUE

Le PDG de Starbucks, Kevin Johnson, était autrefois cadre chez Microsoft. L'expérience qu'il a acquise au sein de ce géant de la technologie, également implanté dans la région de Seattle, l'a aidé à appliquer la pensée numérique à son nouveau rôle au sein d'un type d'entreprise très différent.

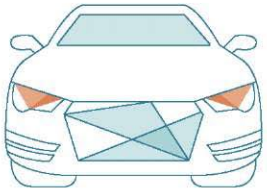
Johnson parle du parcours numérique de Starbucks en toute franchise : « Là où d'autres tentent de créer une application mobile, Starbucks a construit une plateforme consommateur de bout en bout ancrée autour de la fidélité ».

La principale innovation numérique de l'entreprise gravite autour de son application Mobile Order and Pay. Se concentrer sur l'application est fondamentalement une stratégie axée sur le client, car elle répond aux besoins fondamentaux du consommateur : la commodité, les coupe-files et la vitesse d'exécution, etc... Associée à son vaste programme de fidélisation, l'application offre à Starbucks l'occasion idéale de pratiquer la vente incitative et le marketing auprès des consommateurs. Élément tout aussi important, l'application renvoie à l'entreprise d'énormes volumes de données utilisateurs, ce qui lui permet de mieux comprendre les habitudes et les désirs de ses clients.

Starbucks a beaucoup investi dans la création de points de contact numériques pour ses clients et grâce à sa présence massive dans le monde, la disponibilité de l'application, en termes de temps de disponibilité et de latence bruts, était essentielle.



Se focaliser sur l'application est fondamentalement une stratégie axée sur le client, car elle répond aux besoins fondamentaux du consommateur.

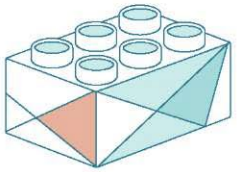


AUDI : CONSTRUCTEUR AUTOMOBILE OU CORPORATION ?

Déjà très concurrentielle, l'industrie automobile est confrontée à une force de perturbation massive à court et moyen terme. Les nouveaux modèles de vente, l'essor des véhicules électriques et la conduite autonome changent la donne pour les constructeurs automobiles. Face à ces défis, Audi a changé le modede vente de ses véhicules.

Lancé en 2012, Audi City offre une expérience de marque approfondie qui permet aux visiteurs d'explorer virtuellement toute la gamme Audi, même dans les concessions de centres-villes qui n'ont pas suffisamment de place pour accueillir des salles d'exposition.

Audi est une marque de luxe, et la décision de la société de perturber son propre canal de vente n'a pas été prise à la légère. Audi a beaucoup investi dans la création d'une expérience de vente au détail virtuelle qui soit aussi authentique que l'expérience physique. Ce processus a consisté en partie à utiliser divers points de contact, des facteurs de formulaires de demandes et des approches de présentation. Réaliser tout cela en respectant les attentes des utilisateurs en matière de rapidité d'exécution était un défi technologique qui nécessitait une nouvelle réflexion.



LEGO : DES BRIQUES EN PLASTIQUE AUX BLOCS NUMÉRIQUES

Le groupe LEGO est le célèbrefabricantdanois des jouets pour enfants du même nom. Après une longue période d'essor de 1970 à 1991, LEGO a connu un déclin régulier de son activité de 1992 à 2004. Si bien qu'en 2004, l'entreprise s'est retrouvée au bord de la faillite.

Arrivée à un point de basculement, LEGO a été contraint d'entamer une restructuration majeure. Sa transformation numérique s'est concentrée sur le développement de nouvelles sources de revenus provenant des films, des jeux mobiles et des applications mobiles.

Quand LEGO s'est lancée dans ce processus, l'une des principales restrictions que la société a dû surmonter a été l'impact sur les performances de dizaines de milliers d'enfants utilisant en même temps toutes leurs applications et jeux LEGO. La direction a décidé que la rapidité de l'innovation et de la mise à disposition de ses produits numériques était une exigence non négociable.

La direction de LEGO a décidé que la rapidité de l'innovation et de la mise à disposition de ses produits numériques était une exigence non négociable.

Les deux époques de la prestation numérique

Lesentreprises ont connu deux époques dans le numérique. D'abord, elles ont dû faire face à l'Ère de la disponibilité. Aujourd'hui, la disponibilité devenant un problème largement résolu, elles entrent dans l'Ère de la vitesse.

L'ÈRE DE LA DISPONIBILITÉ : LE TEMPS DE DISPONIBILITÉ EST LA CLÉ

Avec l'avènement d'Internet et la création d'entreprises comme Amazon, eBay et Netflix, les corporations ont commencé à explorer le potentiel de ces nouvelles technologies et de ces nouveaux modèles économiques. Aux premiers jours de la transformation numérique, les équipes informatiques recherchaient principalement un seul indicateur : le temps de disponibilité. Les entreprises qui se lancent dans le monde numérique poursuivaient un seul objectif : garantir que leurs sites Web et leurs applications soient disponibles partout, à tout moment. Cette époque, que nous appelons l'Ère de la disponibilité, était caractérisée par des outils et des approches qui assuraient la fiabilité des sites.

L'Ère de la disponibilité a favorisé une énorme quantité d'innovations, tout cela dans le but d'augmenter le nombre de 9 dans l'indicateur de pourcentage du temps de disponibilité. Le regroupement des fonctions de développement et d'exploitation dans le rôle combiné du DevOps consistait à accélérer le développement des applications et à accroître la fiabilité. De puissants outils de surveillance des applications et des infrastructures et des plates-formes ont été créés afin d'atteindre ce Saint Graal : des pourcentages de disponibilité toujours plus élevés dans un environnement qui évolue toujours plus vite.

En effet, si l'objectif des « cinq neuf » est facile à comprendre, il est important de savoir ce qu'implique réellement un temps de disponibilité de 99,999 % : pas plus de seulement 26 secondes de temps d'arrêt par mois. Comme de plus en plus d'entreprises s'approchent ou atteignent des statistiques de temps de disponibilité de ce type grâce à une ingénierie de haute qualité et à une compréhension approfondie de ce qu'il faut faire pour planifier les pannes, les directeurs informatiques ont pu se concentrer sur d'autres domaines à améliorer. Ainsi, des domaines autrefois ignorés, deviennent aujourd'hui critiques.

La disponibilité étant un problème largement résolu, les entreprises entrent dans l'Ère de la vitesse.

LES STATISTIQUES QUI INDIQUENT LA FIN DE L'ÈRE DE LA DISPONIBILITÉ

Les entreprises ont passé les dix ou vingt dernières années à se faire dire, que, à mesure qu'elles multipliaient les points de contact numériques avec les clients, la disponibilité fondamentale de ces points de contact était essentielle. Une génération entière de praticiens de l'informatique a été obsédée par les indicateurs de disponibilité et les moyens de les améliorer.

Il existe cependant des facteurs fondamentaux qui changent la donne pour ces praticiens. Outre la complexité accrue que leurs propres efforts pour assurer la disponibilité ont posé, il y a aussi des facteurs externes qui entraînent des exigences critiques de latence qui soit la plus faible possible.

À mesure que les consommateurs migrent en masse vers les points de contact mobiles, la manière même dont ils consomment les données, et leurs exigences d'immédiateté, changent. Les consommateurs utilisent leurs appareils mobiles pour mieux s'informer sur les produits et services qui sont importants pour eux. 80 % des consommateurs consultent les informations sur les produits, les avis et les prix sur leur smartphone lorsqu'ils font leurs achats dans un magasin physique.

Et cette tendance à consommer de l'information n'est qu'un début. Les consommateurs effectuent également des transactions de nouvelles manières. Un bon tiers de tous les achats effectués pendant la période des fêtes de fin d'année en 2018 ont été effectués sur smartphone.

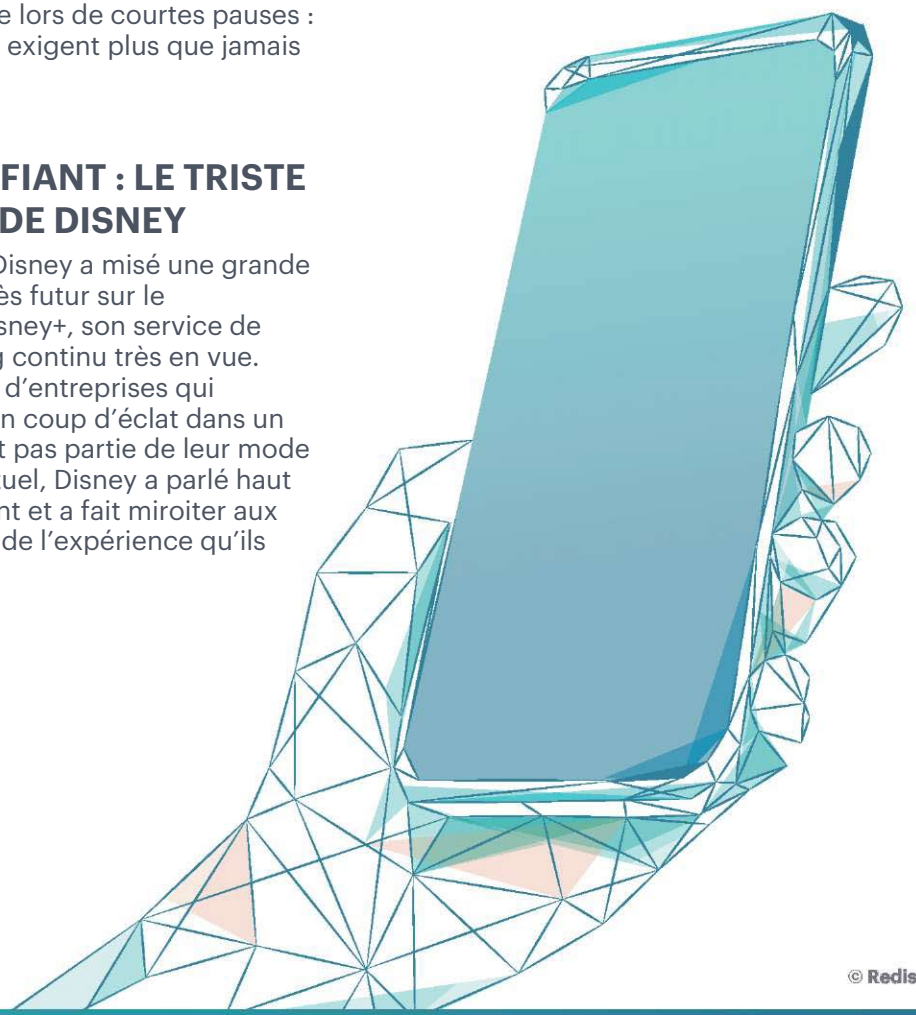
Malheureusement, les entreprises ont tendance à surestimer leur propre capacité à offrir de bonnes expériences. Une étude de Qualtrics a révélé que si 60 % des entreprises pensent offrir une bonne expérience sur mobile, seuls 22 % des consommateurs sont du même avis.

Tout cela montre qu'il faut une navigation mobile rapide dans des contextes différents, de la navigation fixe, en marchant, dans un magasin, ou encore lors de courtes pauses : tous ces contextes exigent plus que jamais la rapidité.

UN RÉCIT ÉDIFIANT : LE TRISTE LANCEMENT DE DISNEY

L'année dernière, Disney a misé une grande partie de son succès futur sur le déploiement de Disney+, son service de vidéo en streaming continu très en vue. Comme beaucoup d'entreprises qui cherchent à faire un coup d'éclat dans un domaine qui ne fait pas partie de leur mode de prestation habituel, Disney a parlé haut et fort du lancement et a fait miroiter aux clients les qualités de l'expérience qu'ils allaient vivre.

Malheureusement, dès le lancement de Disney+, les utilisateurs ont commencé à se plaindre de la mauvaise qualité du service : les mises en mémoire tampon prolongées, les pertes de connexion et la latence générale ont été autant de problèmes qui ont entravé ce qui aurait pu être un jour de lancement jubilatoire. Les critiques étaient claires : un service qui propose une vitesse médiocre est tout aussi mauvais qu'un service totalement indisponible.



L'ÈRE DE LA VITESSE : SOMNOLEZ ET VOUS PERDEZ

Ces dernières années, la plupart des entreprises ont acquis une bonne compréhension du temps de disponibilité. Pendant ce temps, leurs fournisseurs de services ont fait beaucoup pour intégrer de multiples redondances dans leurs plateformes, veillant à ce que le chemin vers une disponibilité presque parfaite soit facile à parcourir. Les outils de surveillance, les pratiques d'ingénierie sur la fiabilité des sites, et l'adoption de la résilience en cas de défaillances inévitables ont contribué à fournir ce que les utilisateurs finaux attendent désormais : des sites Web et des applications qui sont disponibles à chaque fois que l'on en a besoin.

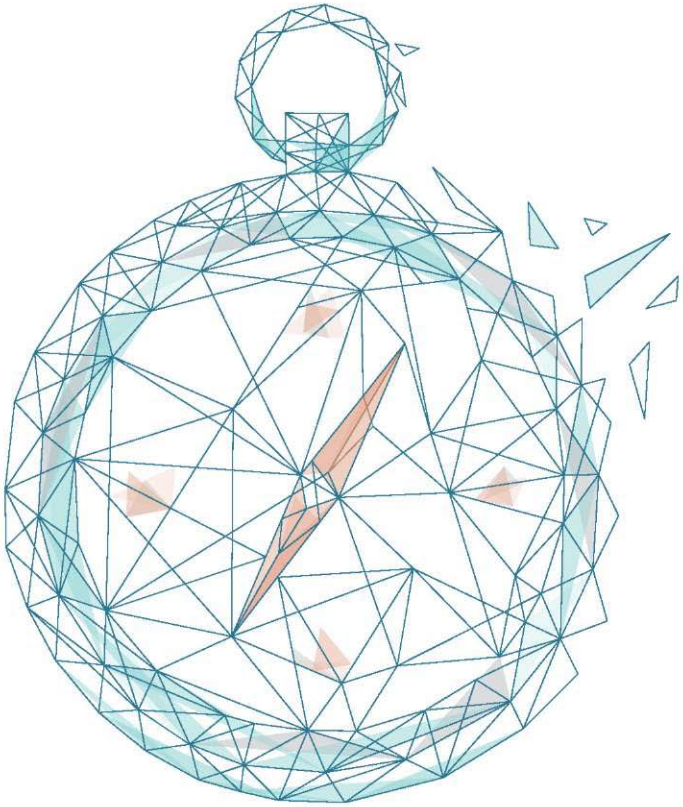
Mais toute cette ingénierie supplémentaire et l'utilisation d'architectures toujours plus complexes dans le but de fournir les applications les plus résilientes a posé de nouveaux défis, qui sont tout aussi critiques que le temps de disponibilité.

Il est clair que nous entrons dans une deuxième ère, une ère dans laquelle la fiabilité est devenue l'enjeu du moment tandis que la vitesse est désormais le facteur de différenciation concurrentiel. Les décisions des clients, autrefois prises moyennant du temps et de l'analyse, sont de plus en plus prises en un clin d'œil. Et si votre site met plus de temps que celui d'un battement de cil à se charger, ou si votre service de streaming souffre de blocages et de pauses tampons, vous perdez forcément.

Si vous pensez que l'insatisfaction des clients n'a pas d'impact sur leurs habitudes de consommation, détrompez-vous. Comme le détaille un article de Forbes de 2019 (How Fast Is Fast Enough ?) Mobile Load Times Drive Customer Experience and Impact Sales): « Une page qui se charge lentement sur un appareil mobile ne met pas seulement la patience des consommateurs à rude épreuve. Elle peut être « l'échec de l'expérience client qui vous coûte une vente. C'est la principale conclusion du rapport 2019 sur la vitesse des pages ... L'étude, qui explore les attitudes de 1 150 consommateurs et entreprises, constate que la vitesse des pages est un facteur décisif dans le comportement d'achat ».

Et l'impact d'une mauvaise vitesse de page n'est pas sans conséquence : « Près de 70 % des consommateurs déclarent que la vitesse des pages a un impact sur leur volonté d'acheter. De plus, un temps de chargement lent diminue également les chances que ces consommateurs reviennent ensuite. Une analyse des données révèle que 22 % des acheteurs ont déclaré qu'ils fermeraient l'onglet, 15 % ont expliqué qu'ils visiteraient le site d'un concurrent et 12 % parleraient à un ami de leur expérience négative ».

Si la nouvelle ère est définie par la nécessité de veiller à ce que la latence soit la plus faible possible, quelles sont les choses auxquelles les entreprises doivent penser pour atteindre cet objectif ?



La rapidité d'exécution dans un monde complexe

Dans son billet fondateur de 2013 sur L'entreprise composable™, Jonathan Murray, ancien directeur technique de Warner Music Group, a décrit l'avenir de la technologie dans le contexte des demandes de rapidité et d'agilité formulées par les entreprises. S'appuyant sur l'expérience qu'il a eue toute sa vie de la mise en œuvre de stratégies numériques dans les grandes entreprises, Murray a décrit l'Entreprise composable comme ceci : « Les fonctions, processus, organisations, relations avec les fournisseurs et la technologie de l'entreprise doivent être considérés comme des blocs de construction qui peuvent être reconfigurés si nécessaire pour faire face à l'évolution du paysage concurrentiel ».

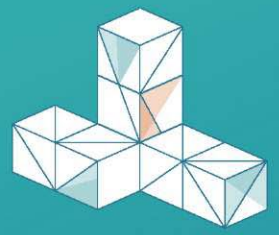
Ce nouveau Modèle d'exploitation par composants (COM) exige une approche de type « Lego » pour concevoir et mettre en œuvre des processus et les entreprises qui les soutiennent. La mise en œuvre d'une approche fondée sur le COM aura des répercussions profondes sur la structure des entreprises, la nature du travail.

Les conceptions d'entreprise basées sur le modèle COM créeront des contraintes importantes pour les infrastructures et entreprises traditionnelles spécialisées dans l'informatique. Nos services informatiques actuels ont été construits pour servir un modèle d'exploitation statique et souvent cloisonné d'un point de vue fonctionnel. L'informatique doit devenir beaucoup plus adaptable de manière dynamique pour suivre le rythme d'activité d'aujourd'hui.

« Une nouvelle approche fondée sur le Modèle d'architecture par composants (CAM) de l'infrastructure informatique, des applications et des services sera nécessaire pour faire en sorte que l'informatique puisse répondre aux besoins des entreprises. Le délai entre l'identification d'un besoin commercial et la mise à disposition de la solution informatique requise doit être une question d'heures et de jours plutôt que de mois et d'années ».

Rédigé il y a plusieurs années, ce billet prémonitoire de Murray décrit la nouvelle normalité au sein des entreprises. Nous avons assisté, ces dernières années, à un séisme dans l'utilisation de l'infrastructure et la création des applications. Avec l'essor des conteneurs, des microservices, des architectures, des outils d'applications modulaires discrètes, etc., faire fonctionner une application et s'assurer qu'elle fonctionne bien implique de jongler entre des dizaines de services, régions, zones géographiques, fournisseurs de services et autres éléments.

Si cette composabilité favorise la productivité des développeurs et l'agilité de l'entreprise, elle a un coût. Il semblerait que proposer une faible latence dans ces conditions soit une chimère.



Nous avons assisté, au cours des dernières années, à un séisme dans l'utilisation de l'infrastructure et la création des applications.

Rapidité des données dans un environnement informatique distribué

Comme nous l'avons vu dans le travail précurseur de Murray sur la composabilité dans les applications et les infrastructures modernes, nous ne disposons plus d'une simple pile monolithique sur laquelle les applications sont construites. Au contraire, dans le but d'offrir aux développeurs et à leurs entreprises la plus grande flexibilité et la plus grande vélocité possible, nous tirons parti d'un grand nombre de services modulaires pour développeurs, de différents modèles d'infrastructures, de diverses approches de l'hébergement et d'une répartition géographique massive des applications. Pendant tout ce temps, nous essayons de fournir ces applications aussi rapidement que possible aux utilisateurs du monde entier.

Dans cette ère de la complexité massive, il serait facile de penser qu'il n'existe pas de tissu commun sur lequel les entreprises peuvent s'appuyer : leur monde semble perpétuellement fluide et en constante évolution.

Il existe toutefois un fil conducteur qui se retrouve dans toutes les activités de l'entreprise : les données. En considérant la couche de données comme un fil conducteur cohérent et unifié sur lequel s'appuient toutes les autres parties de la pile, nous permettons aux organisations de donner du sens au chaos.

Et en choisissant une couche de données qui est conçue pour les environnements distribués, qui affiche les temps de traitement les plus rapides et qui offre la meilleure résilience de sa catégorie, nous sommes à même de fournir exactement ce dont une entreprise a besoin.

L'un des principaux moyens pour les entreprises de s'assurer que leurs applications sont à la fois résilientes et rapides consiste à proposer un modèle de couche de données cohérent. Et pour avoir des données cohérentes, il faut commencer par une base de données capable d'atteindre des objectifs apparemment impossibles : des architectures distribuées, de la cohérence, de la flexibilité et de la rapidité.

Les approches modernes pour réduire la latence

Comme nous l'avons vu, les applications sont de plus en plus créées à l'aide de microservices : en tirant parti d'une multitude de composants différents, avec différentes approches de l'infrastructure, un hébergement dans une variété de lieux différents, consommées par des utilisateurs venant de partout et distribuées sur de nombreuses plateformes différentes.

Avec des données situées dans autant d'endroits et transmises sur autant de réseaux différents, il n'est pas surprenant que les possibilités de conflits soient nombreuses. Pour traiter ces conflits, des types de données répliquées sans conflit (CRDT) ont été développés pour permettre aux données d'être répliquées sur plusieurs sites.

Avec les CRDT, les répliques individuelles peuvent être mises à jour indépendamment et simultanément sans aucune coordination entre elles. Sans CRDT, des mises à jour simultanées de plusieurs répliques de mêmes données, sans coordination entre les ordinateurs hébergeant ces répliques, peuvent entraîner des incohérences entre les répliques.

Par contre, avec les CRDT, toutes les incohérences qui résultent de cette approche distribuée peuvent être résolues. Les CRDT ont été au départ utilisés dans des situations où la distribution de masse est la norme : les systèmes de discussion en ligne, les jeux d'argent sur Internet et les services de streaming audio et vidéo, mais ils sont de plus en plus utilisés dans des applications plus génériques.

Il y a une importante technologie sous-jacente à l'intégration d'un CRDT, mais le moyen le plus simple de le voir est qu'un CRDT fournit une couche de données grâce à laquelle les répliques peuvent agir de manière autonome tout en assurant la cohérence.

Avec des données situées dans autant d'endroits et transmises sur autant de réseaux différents, il n'est pas surprenant que les possibilités de conflits soient nombreuses.

Aller de l'avant avec le cache

Dans les modèles de base de données traditionnels, le site d'une base est séparé du cache. Considérez la base de données comme la bibliothèque municipale principale et le cache comme la bibliothèque locale, où les livres les plus populaires sont conservés pour répondre aux demandes les plus courantes des emprunteurs. Si les livres les plus populaires sont cohérents, cela peut fonctionner sans problème, mais au fur et à mesure que les habitudes de lecture évoluent, que de nouveaux livres paraissent et qu'ils deviennent moins populaires, les choses se compliquent.

Et cette notion de vérification rapide et constante de différents éléments d'information est simplement la métaphore des applications modernes : toute la composabilité dont Murray a parlé fait qu'il faut accéder aux données à partir de la base de données depuis de nombreux services et différents endroits et ce, à différents moments.

Dans un monde où les services sont de plus en plus discrets et, donc, où les endroits où une erreur est susceptible de se produire sont de plus en plus nombreux, le modèle traditionnel n'est pas la panacée (voir le diagramme ci-dessous). Et pour les applications où les modèles de données consistent davantage à transférer de nombreux morceaux d'informations de la taille d'une bouchée, le modèle de cache n'est peut-être pas le moyen le plus rapide d'amener les données là où elles doivent aller.

C'est ici qu'intervient la notion de couche de données unique : en exploitant une couche unique pour remplacer la combinaison base de données/cache, la complexité de la couche de données est réduite. En retour, ce qui est turbocompressé, c'est l'application distribuée et modulaire qui est la norme aujourd'hui.

L'avantage supplémentaire est que la réduction du nombre de parties au niveau de la couche de données réduit également la latence. Si les différentes parties d'une couche de données complexe peuvent être rapides, le fait de disposer d'un seul magasin de données réduit le nombre de sauts de réseau qui réduit invariablement les processus.

Par conséquent, à la place d'un cache, de nombreuses bases de données modernes exploitent des techniques en mémoire où la mémoire sert de stockage, au lieu des disques externes. C'est essentiel car, comme pour tout contenu stocké en mémoire, la vitesse n'est pas limitée par de multiples couches de stockage. Avec un modèle basé sur le cache, ce qui est stocké dans le cache devient le goulot d'étranglement qui limite la vitesse globale.

Dans un modèle traditionnel, l'accès aux données signifie que l'application doit :

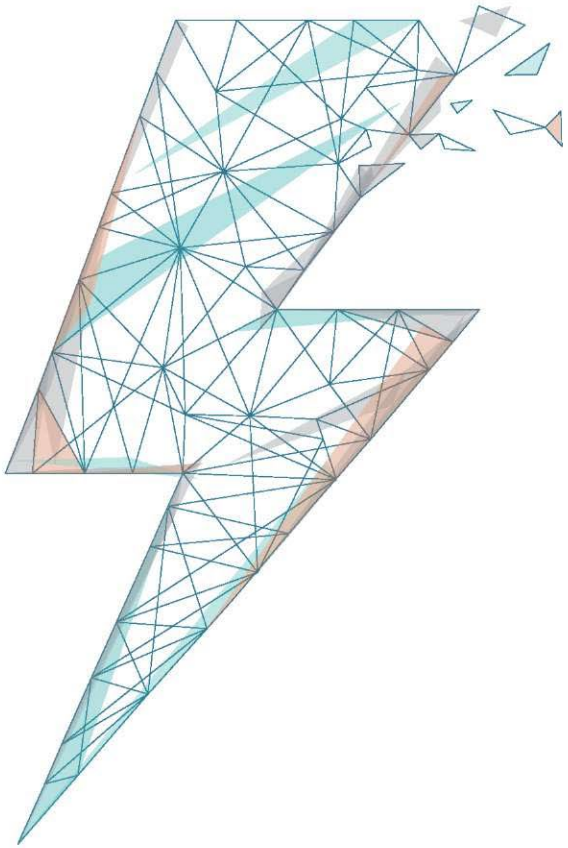
Vitesse : un octet à la fois

Les bases de données traditionnelles, comme nous l'avons vu plus haut, s'appuient sur une mémoire externe pour leur cache. Jusqu'à très récemment, et ces 50 dernières années, le stockage se faisait sur des disques physiques tournants et donc, la plupart des approches traditionnelles des bases de données étaient optimisées par rapport à cela.

Mais, comme les disques durs sont des dispositifs physiques, ils ont des contraintes fixées par le monde physique. Pour contourner ces contraintes physiques, un certain nombre de contraintes d'exploitation ont été créées. Bien qu'il s'agisse d'un détour technique, la minutie de la technologie des disques physiques impacte grandement la vitesse de la base de données.

Le nœud du problème, cependant, est que la ligne qui sépare le stockage moderne de la mémoire s'estompe. L'essor des disques durs à état solide (SSD) et d'autres nouvelles approches du stockage impliquent que ces solutions techniques de contournement, conçues pour un monde limité par la vitesse physique des dispositifs mécaniques, ne sont plus nécessaires. Cela signifie également que le stockage peut être hiérarchisé, de sorte que toutes les données peuvent être conservées dans un stockage rapide et que le besoin d'un cache séparé disparaît.

Le résultat net, pour ceux qui essayent de concevoir en fonction de la vitesse, est une couche de données plus rapide servant de base à la conception de nos applications.



L'essor de nouvelles approches de stockage impliquent que ces solutions techniques de contournement conçues pour un monde limité par la vitesse physique des dispositifs mécaniques, ne sont plus nécessaires.

Mettre à l'échelle sans nuire à la vitesse

C'est très bien de créer une application qui fonctionne rapidement avec une utilisation limitée, mais que se passe-t-il lorsque votre débit augmente considérablement ? C'est le problème que tout développeur d'applications, à la recherche de l'adoption et de la viralité, espère rencontrer.

Mais la mise à l'échelle se fait de deux façons : ***de manière verticale,*** en termes de quantité de données transférées à travers la couche de données, mais aussi horizontale par rapport à la simple quantité d'informations existante.

Les entreprises doivent construire une couche de données qui permet cette mise à l'échelle de manière progressive et transparente. Cela implique de réfléchir à un certain nombre de facteurs différents : la possibilité d'exécuter la couche de données à plusieurs endroits, la possibilité d'utiliser différents types de mémoire et de stockage, la possibilité de hiérarchiser les données en fonction de leur régularité d'utilisation, et, enfin, la possibilité d'évoluer de manière globale.

Passons à ce dernier point. Toute cette capacité à stocker et à traiter en mémoire est une bonne chose, mais si votre application doit se diffuser dans le monde, pouvez-vous encore profiter de ce même faible niveau de latence ?

C'EST UN MONDE MULTICŒUR

Les traitements modernes s'effectuent de plus en plus souvent à l'aide d'un contact multicœur. Le multicœur fait simplement référence à l'informatique où deux ou plusieurs unités de traitement individuelles se trouvent dans une seule et même unité centrale. Les instructions envoyées à l'unité centrale peuvent être traitées sur des cœurs séparés en même temps, ce qui augmente la vitesse globale.

L'exploitation des architectures multi-cœurs peut être un défi. Les entreprises qui souhaitent exploiter une couche de données capable d'évoluer de la manière la plus performante possible ( ) doivent y réfléchir. Votre couche de données est-elle capable d'évoluer sur un seul cluster pour fournir la meilleure échelle avec la latence la plus faible ??

MISE À L'ÉCHELLE VERTICALE

MISE À L'ÉCHELLE HORIZONTA

Les entreprises doivent mettre en place une couche de données qui permette cette mise à l'échelle de manière progressive et transparente.

UNE PROMENADE DANS LES RUES DU CAP

Étant donné que ce document sera inévitablement utilisé par ceux qui aspirent à construire des applications distribuées mondialement qui affichent des performances similaires à celles des applications locales, il est utile d'examiner certaines contraintes autour des couches de données distribuées.

Il y a environ 20 ans, l'informaticien Eric Brewer a développé le théorème CAP, qui concerne les applications distribuées et, précisément, les données que ces applications créent et consomment.

Le théorème CAP, dans les termes les plus simples, affirme que tout système de données partagées en réseau ne peut avoir que deux propriétés sur les trois souhaitables ; cohérence (C) (consistency en anglais) qui désigne l'existence d'une seule copie unique et à jour des données, haute disponibilité (A) (availability en anglais) des données et la tolérance au partitionnement (P) (partition tolerance en anglais).

Et, étant donné que, dans ces premiers jours où une recherche de la vitesse était tout, le théorème CAP signifiait que les approches les plus susceptibles de produire les vitesses et la disponibilité des applications les plus rapides (partitionnement du réseau et haute disponibilité) entraîneraient forcément une incohérence des données.

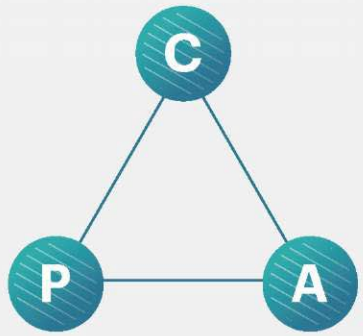
Toutefois, au cours des décennies qui ont suivi l'introduction du théorème CAP, de nouvelles approches de la gestion des systèmes distribués ont été développées qui permettent cet exploit théoriquement impossible : cohérence des données, disponibilité et tolérance au partitionnement. L'essor des nouvelles approches en matière de données signifie que nous pouvons avoir une faible latence sans renoncer à la cohérence.

Bien que ce ne soit pas le lieu pour un traité technique définitif, il est important pour les personnes responsables de l'application de leur entreprise de comprendre les rudiments du fonctionnement des applications modernes.

Comme indiqué, dans un monde où les applications sont, par nécessité, distribuées, il y aura plusieurs nœuds inclus dans de nombreuses applications individuelles. Dans cette situation multi-nodale, il existe deux grandes options : **Les données Active-Passive** ou **les données Active-Active.**

LE THÉORÈME CAP

COHÉRENCE
Cela équivaut à avoir une seul et même copie actualisée des données



PARTITIONS
La tolérance au partitionnement réseau

DISPONIBILITÉ
La haute disponibilité de ces données

L'essor des nouvelles approches en matière de données signifie que nous pouvons avoir une faible latence sans renoncer à la cohérence.

COUCHE DE DONNÉES UNIFIÉE

Les plans de données, la partie du logiciel qui traite les demandes de données, peuvent être Actifs-Actifs ou Actifs-Passifs.

L'approche Active-Active (aussi parfois appelé double active) est une approche par laquelle chaque nœud a accès à une base de données répliquée donnant à chaque nœud l'accès et l'utilisation d'une seule application. Cette technologie permet de maintenir la cohérence des données pour vos applications dans différents environnements (serveurs, hybride, multi-cloud) et même pour des applications distribuées dans le monde entier. Dans un système actif-actif, toutes les demandes voient leur charge répartie sur toute la capacité de traitement disponible. Lorsqu'une panne se produit sur un nœud, un autre nœud du réseau prend sa place.

Un cluster actif-actif est généralement composé d'au moins deux nœuds, tous deux exécutant activement et simultanément le même service. Étant donné qu'il y a plus de nœuds pouvant servir, il y aura également une amélioration marquée du débit et des temps de réponse par rapport à une approche Active-Passive.

ACTIF-PASSIF

Un cluster actif-passif se compose également d'au moins deux nœuds. Toutefois, comme l'indique le nom « Actif- Passif », tous les nœuds ne sont pas actifs. Dans un cluster à deux nœuds, par exemple, si le premier nœud est déjà actif, le second nœud doit être passif ou en veille. Le nœud passif (alias basculement) sert de sauvegarde et est prêt à prendre le relais dès que le serveur (alias primaire) se déconnecte ou est incapable de fonctionner.

Lorsque les clients se connectent à un cluster à deux nœuds en configuration Active-Passive, ils se connectent à un seul serveur. En d'autres termes, tous les clients se connectent au même serveur. Comme dans la configuration Active- Active, il est important que les deux serveurs aient exactement les mêmes paramètres. C'est ce qu'on appelle la redondance, et c'est ce qui garantit que les données peuvent se répliquer sans problème entre les nœuds.

Si des modifications sont apportées aux paramètres du serveur primaire, ces modifications doivent être transmises au serveur de basculement. Ainsi, lorsque le basculement prend le relais, les clients ne sont pas en mesure de faire la différence.

Si la latence est la nouvelle panne, il est clair que plus un nœud est proche de l'utilisateur de l'application, plus les chiffres de latence seront bas. Nous devons donc trouver un moyen de distribuer les applications de manière globale (puisque la distribution des nœuds à proximité des utilisateurs d'applications réduit la latence), tout en garantissant la cohérence. Heureusement, nous pouvons compter sur un peu d'aide à cet égard.

CONCEVOIR AVEC LA VITESSE À L'ESPRIT

La réplication sans conflit est une notion qui permet à plusieurs copies (répliques) de données d'exister à plusieurs endroits de manière cohérente. C'est une méthode très importante pour garantir une faible latence pour les applications distribuées, mais il y a d'autres aspects à prendre en compte. Comme indiqué plus haut, les bases de données modernes conçues pour offrir la plus faible latence aux applications modernes stockent les données en mémoire. En supprimant le besoin d'un cache externe, nous pouvons réduire la quantité de trafic de données nécessaire.

Alors que les bases de données traditionnelles ont été conçues pour des cas d'utilisation où un temps de traitement de 10 ou 100 millièmes de seconde était acceptable, dans le monde d'aujourd'hui, qui nécessite des temps de réponse instantanés, les performances inférieures au millième de seconde sont une nécessité.

L'ÉCHEC EST ACCEPTABLE QUAND ON FAIT LES CHOSES RAPIDEMENT

Le basculement, comme son nom l'indique, est un système automatisé par lequel, en cas de défaillance d'un nœud pour une quelconque raison, un autre nœud répliqué prend le relais. S'il est facile de concevoir le basculement, la vitesse de ce basculement détermine l'impact de la panne sur l'utilisateur final.

Pour garantir une latence minimale dans un monde où les défaillances nodales peuvent être inévitables, il est important que la couche de données multinodale puisse assurer le basculement aussi rapidement que possible.

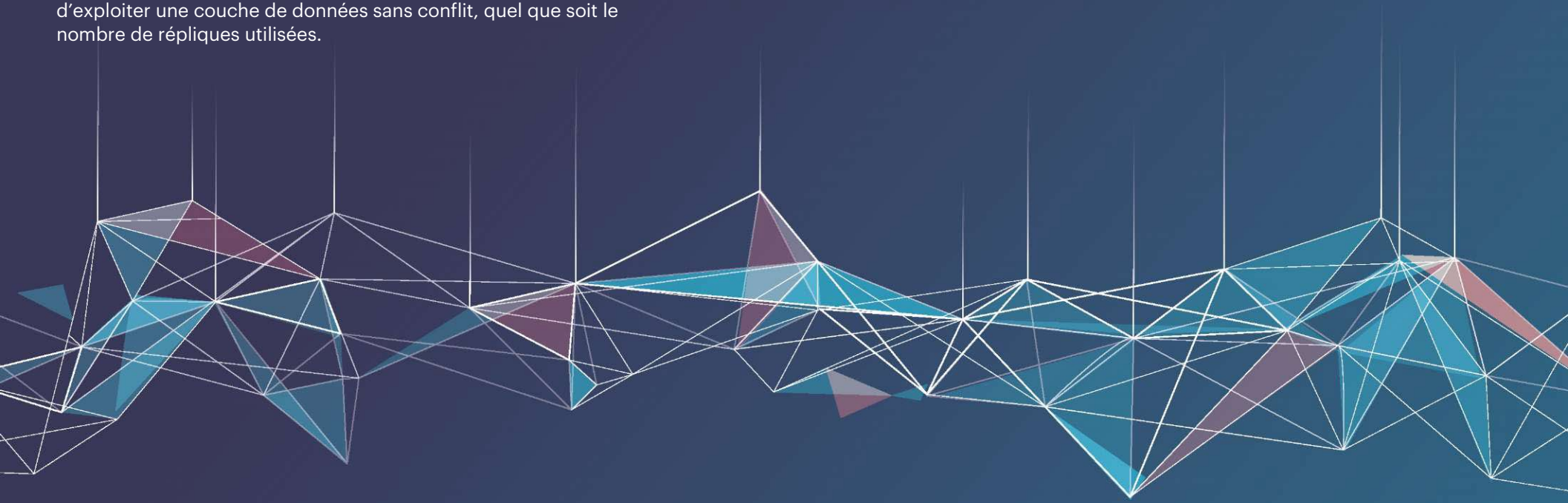
Résumé

Dans le monde moderne, les entreprises, poussées à proposer des expériences numériques, doivent s'assurer que leurs parties prenantes peuvent utiliser les applications quand et où elles le souhaitent. Mais les utilisateurs d'aujourd'hui exigent non seulement un accès continu, mais aussi des performances quasi instantanées. Dans un monde qui passe de l'ère de la disponibilité à celle de la vitesse, la latence peut être aussi grave que l'indisponibilité des applications.

Heureusement, nous disposons aujourd'hui d'options qui n'existaient tout simplement pas il y a dix ans. De nombreux obstacles qui empêchaient de fournir des applications rapides, dont le théorème CAP, ont été surmontés. Désormais, les entreprises ont la possibilité d'exploiter une couche de données sans conflit, quel que soit le nombre de répliques utilisées.

En exploitant des bases de données qui fonctionnent entièrement en mémoire, et en les exécutant de manière active, nous fournissons des bases de données plus rapides que celles qui étaient disponibles auparavant, et nous offrons la faible latence que les utilisateurs d'applications d'aujourd'hui exigent.

Cette question devrait être considérée comme urgente pour chaque entreprise : vos concurrents et vos perturbateurs fournissent des applications rapides et vos clients les exigent, vous n'avez pas le luxe de vous accorder du temps.



À propos de l'auteur - Ben Kepes

Ben Kepes est un analyste, commentateur et consultant en technologie. Au cours de ces quinze dernières années, il s'est forgé une réputation d'expert mondialement reconnu dans les domaines du cloud computing, des technologies d'entreprise et de la transformation numérique.

Les commentaires de Ben ont été largement publiés dans des médias tels que Forbes, Wired et The Guardian, et il a été invité à prendre la parole lors de nombreuses conférences sur la technologie, les affaires et l'intérêt général.

@benkepes

À propos de Redis

Les entreprises modernes dépendent de la puissance des données en temps réel. Avec Redis, les entreprises offrent des expériences instantanées d'une manière hautement fiable et évolutive.

Redis est le berceau de Redis, la base de données en mémoire la plus populaire au monde et le fournisseur commercial de Redis Enterprise, qui offre des performances supérieures, une fiabilité incomparable et une flexibilité inégalée pour la personnalisation, l'apprentissage automatique, l'IdO, la recherche, le commerce électronique, les réseaux sociaux et des solutions de mesure dans le monde entier.

Régulièrement classé parmi les leaders dans les principaux rapports d'analystes sur les bases de données NoSQL, les bases de données en mémoire, les bases de données opérationnelles et les bases de données en tant que service (DBaaS), Redis bénéficie de la confiance de plus de 7 400 entreprises clientes, dont cinq sociétés classées au Fortune 10, trois des quatre émetteurs de cartes de crédit, trois des cinq premières sociétés de communication, trois des cinq premières sociétés de santé, six des huit premières sociétés technologiques et quatre des sept premiers marchands.

Proposé en tant que service dans les clouds publics et privés, en tant que logiciel téléchargeable, dans des conteneurs et destiné à des déploiements hybrides cloud/sur site, Redis Enterprise sert dans les cas d'usage populaires de Redis comme les transactions à grande vitesse, la gestion des tâches et des files d'attente, les magasins de session utilisateur, l'ingestion de données en temps réel, les notifications, la mise en cache du contenu et les données de séries chronologiques.

Siège social
700 E El Camino Real Suite 250
Mountain View, CA 94040

tél. : +1 (415) 930-9666
redis.com

Suivez-nous