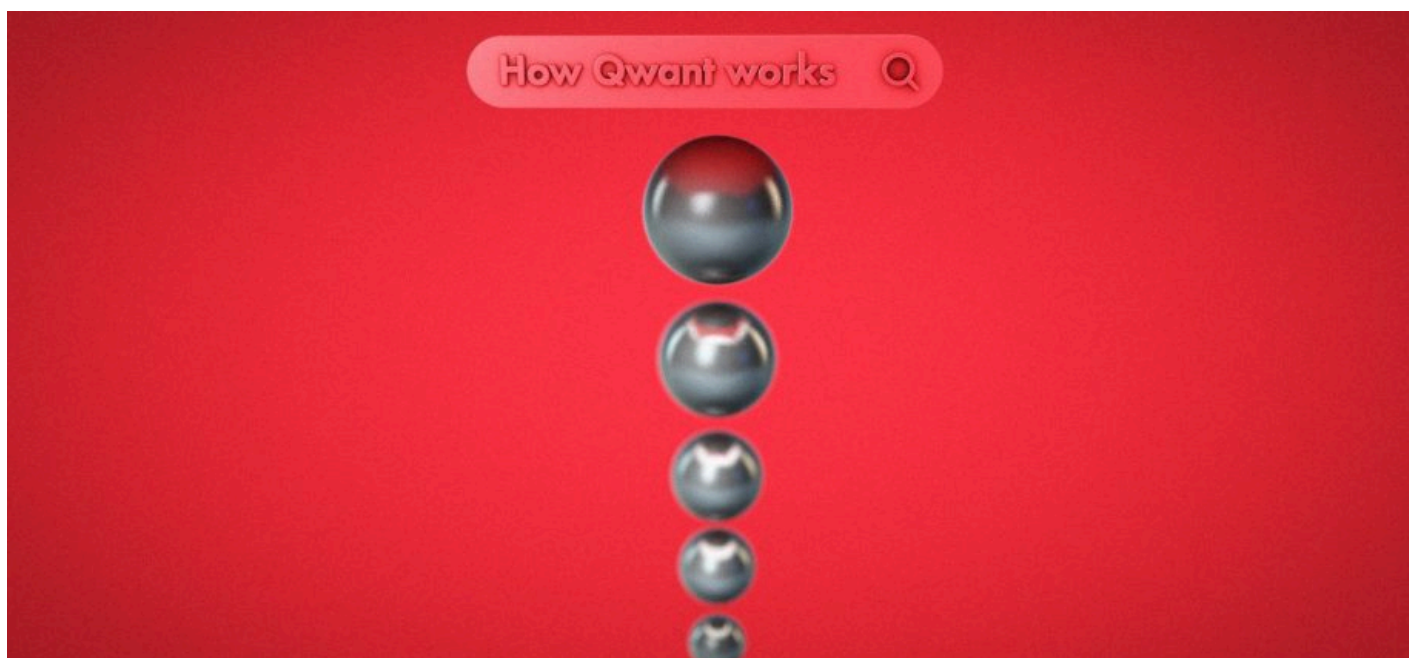


Qwant News, Qwant News

Comment les outils Microsoft renforcent Qwant



A l'occasion du salon VivaTech 2019, Qwant et Microsoft ont annoncé ensemble un partenariat pour accompagner la forte croissance de Qwant tout en continuant à respecter la vie privée de ses utilisateurs.

Team Qwant 14 juin 2019 ⌚ 7min

A l'occasion du salon VivaTech 2019, Qwant et Microsoft ont annoncé ensemble un partenariat pour accompagner la forte croissance de Qwant tout en continuant à respecter la vie privée de ses utilisateurs. Ainsi, Qwant pourra louer à Microsoft de

la puissance de calcul et de stockage qui sera utilisée sur des données pour indexer le Web.

Qu'une chose soit claire : l'intérêt pour Microsoft est de fournir de la puissance de calcul à Qwant pour répondre à sa croissance exponentielle. C'est tout. 🗨️

[@JP_Courtois](#) [#VivaTech](#) pic.twitter.com/W7cHzwyleY

— Qwant (@Qwant_FR) [May 17, 2019](#)

Cette annonce est le fruit de l'intérêt que vous portez à nos services : la croissance des requêtes traitées par nos serveurs est en perpétuelle augmentation, et à ce rythme nos infrastructures seront bientôt sous l'eau et lentes à répondre ! Nous avons donc décidé de revoir toute notre infrastructure pour concentrer nos serveurs sur la protection de vos données, et migrer une partie de ce qui n'est pas sensible sur des machines en cloud. Puisque nous cherchons continuellement à améliorer notre qualité, cette stratégie nous permettra aussi d'augmenter notre couverture du Web avec des moyens supplémentaires pour notre indexation.

Aujourd'hui, pour concevoir le moteur de recherche que vous connaissez, nous gérons plus de 400 serveurs dont nous sommes propriétaires et que nous avons installés dans un data-center en région parisienne. Ils servent à plusieurs usages : le front (ce que vous voyez), le crawl (la découverte et mise à jour des contenus), le stockage de textes et d'images, l'indexation, les cartographies, les actualités, nos outils internes et bien d'autres. Il va sans dire que près des deux tiers de ce parc de machines est réservé au bon fonctionnement de Qwant Search (que ce soit la production ou la pré-production). Cependant, nos ressources techniques ne sont pas infiniment extensibles, ce qui est un frein à notre développement alors que l'achat et le renouvellement de matériel est très onéreux. Ainsi, nous allons profiter des ressources à la demande de Microsoft Azure que nous louerons pour absorber et accroître notre développement. En parallèle, nous profiterons de cette souplesse pour continuer à investir dans notre infrastructure souveraine.

Pour comprendre ce qui va se passer avec ce partenariat, je vous propose de suivre le voyage des pages du web dans nos serveurs.

25 machines visitent 12 000 pages web par seconde

La première étape pour créer un moteur de recherche est de découvrir ce qui fait internet, de quels contenus il se compose. En effet, mis à part les noms de domaine et les adresses IP, rien n'est standardisé. Vous pouvez créer le site avec la structure et le contenu que vous souhaitez sans limitation. Pour que nous en ayons connaissance, nous devons faire fonctionner des robots d'exploration, ou crawlers dans notre jargon. Aujourd'hui c'est environ 25 machines qui sont dédiées à cette tâche, elles sont capables de visiter 12.000 pages web par secondes, soit 1 milliard par jour. En plus de la découverte de nouveaux sites, les robots doivent mettre à jour les contenus déjà visités, un cas primordial pour les sites d'actualités.

À la sortie du crawler les données extraites du web sont brutes, elles doivent être analysées avant de rentrer dans nos bases de données. C'est ici que nous utilisons des programmes informatiques d'analyse que nous nommons micro-services. Ils nous servent à déterminer par exemple : la langue de la page, son domaine sémantique, mais aussi des scores approchés de graphe. Nous allons utiliser le service Kubernetes d'Azure pour déployer facilement et opérer des micro-services d'analyse de production et de recherche. Ces derniers permettront d'améliorer la pertinence du moteur à l'avenir.

20 milliards de documents indexés

Lorsque les données crawlées sont analysées et mises en forme il faut alors les indexer. C'est-à-dire les structurer comme dans les index que l'on trouve à la fin des livres, pour être capable de les retrouver. L'index dans un livre donne la correspondance entre un mot et le numéro de page où il apparaît. Pour internet l'application est quasiment identique à la différence du nombre de documents que nous devons indexer. Aujourd'hui notre parc de serveurs nous permet d'atteindre 20 milliards de documents indexés, soit plusieurs centaines de téraoctets de données.

Cependant nos crawlers découvrent plus de documents qu'il n'est possible d'en indexer à cause de la limite physique du nombre de serveurs disponibles. Le

crawler télécharge près de 10 gigaoctets par seconde. Ainsi actuellement nous priorisons l'indexation en fonction par exemple de la langue et du score de graphe du document. Les pages web qui ne sont pas indexées sont stockées dans un système d'archivage, cela représente près de 2 pétaoctets de données. En utilisant, les solutions de stockage d'Azure nous pourrions libérer chez nous de la place pour augmenter la taille de nos index et par la suite augmenter notre qualité et notre couverture d'internet. De plus, nous pourrions utiliser les ressources Azure pour tester de nouvelles méthodes d'indexation ou de nouvelles langues avant de les passer en production. C'est une souplesse nécessaire pour délivrer des services fonctionnels facilement.

Muscler notre capacité de calcul du graphe

Quand l'index est prêt nous pouvons facilement retrouver tous les documents qui sont liés à un ou des mots clés. Cependant, prenons l'exemple de « chaton » qui va faire remonter des dizaines de millions de pages web de nos index. Pour réordonner tout cela, nous devons ordonner ces pages, en utilisant les scores des graphes et des méthodes de machine learning.

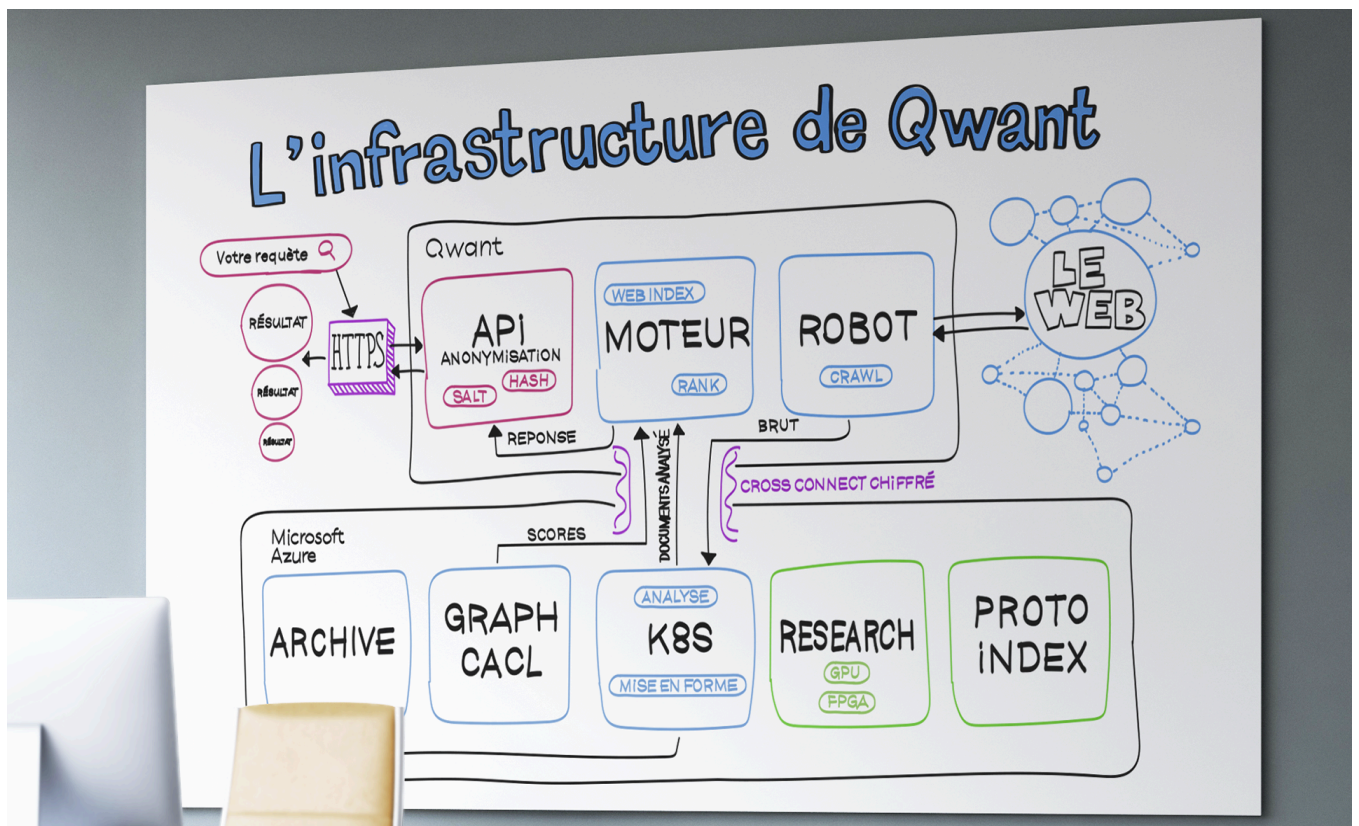
Le calcul des scores de graphe est très onéreux, d'une part car le graphe du web représente des téraoctets de données. De plus car il nécessite plusieurs serveurs travaillant en parallèle pour l'extraire. Nous pourrions utiliser les ressources Azure à la demande pour calculer le graphe ou entraîner nos modèles de Learning To Rank.

Une équipe de chercheurs spécialisés pour améliorer l'index

Depuis près d'un an et demi, nous avons renforcé notre équipe de chercheurs spécialisés (en imagerie, calculs haute performance, traitement des langues, cryptographie...) pour améliorer le moteur de recherche Qwant. Nous utilisons d'ores et déjà des cartes graphiques très puissantes dans les serveurs de type DGX de Nvidia. Ces serveurs nous servent à entraîner des modèles de machine learning, pour la traduction automatique, la recherche d'images, l'optimisation des

algorithmes de graphe. Nous travaillons aussi sur des technologies d'analyse de documents ou la recherche d'information en utilisant des FPGA.

Aujourd'hui, un projet avancé est la recherche d'images qui s'intègre à notre crawler et qui doit passer à l'échelle de la production. Nous devons ainsi stocker des centaines de millions d'images, soit plusieurs centaines de téraoctets d'images. Ici, le stockage objet qu'offre Azure sera très utile pour augmenter la taille de notre index d'images et faciliter l'apprentissage des modèles.



L'utilisateur n'échange pas avec Microsoft

Enfin parlons de vous, qui nous utilisez. Lorsque vous vous connecterez à Qwant, vous serez toujours exclusivement en relation avec des machines dont nous avons la propriété, que nous installons et administrons nous-mêmes. Jamais vous ne serez connecté sur des machines du cloud d'Azure, et jamais vos données personnelles n'y seront envoyées. Nous utilisons du cloud Azure pour le seul « back-office » de Qwant. Par ailleurs et pour mémoire, dès l'établissement de votre connexion nous réalisons une anonymisation de vos données en particulier votre

IP qui est salée et hachée. C'est-à-dire que nous rajoutons du bruit et un découpage de cette IP pour la rendre anonyme lorsque nous demandons à afficher des publicités ou devons stocker des données dans nos logs. Seules ces données anonymes circulent dans notre réseau interne.

Dans la même thématique

 Qwant  [INSTALL](#)

 Firefox  [INSTALL](#)

 DuckDuckGo Privacy Browser  [INSTALL](#)

 Microsoft Edge  [INSTALL](#)

 Ecosia Browser - Fast & ...  [INSTALL](#)

 Opera Browser with free ...  [INSTALL](#)

 Brave Browser  [INSTALL](#)

 Mozilla Firefox Browser  [INSTALL](#)

[Qwant News](#)

Choisir Qwant sur Android, c'est désormais possible

Team Qwant

13 septembre 2021 ⌚ 3 min

PORNOGRAPHIE

La plateforme d'information à destination des parents pour lutter contre l'exposition des mineurs aux contenus pornographiques en ligne.

[DÉCOUVRIR EN VIDÉO](#)

[Qwant News](#)

Qwant réaffirme son soutien à la plateforme jeprotegemonenfant.gouv.fr

Team Qwant

22 septembre 2021 ⌚ 2 min



Qwant News

Qwant renouvelle son équipe dirigeante avec l'arrivée de trois managers expérimentés

Team Qwant

28 février 2022 ⌚ 5 min



Le moteur de recherche dont vous êtes l'utilisateur, pas le produit !

Découvrir le service

Services

Search

Junior

App Android

App iOS

À propos de Qwant

Pourquoi Qwant ?

Better Web

Nous rejoindre

Centre d'aide

Contact

Suivez-nous !



[Politique de confidentialité](#)

[Conditions d'utilisation](#)

[Mentions légales](#)

© 2026 Qwant. Tous droits réservés.