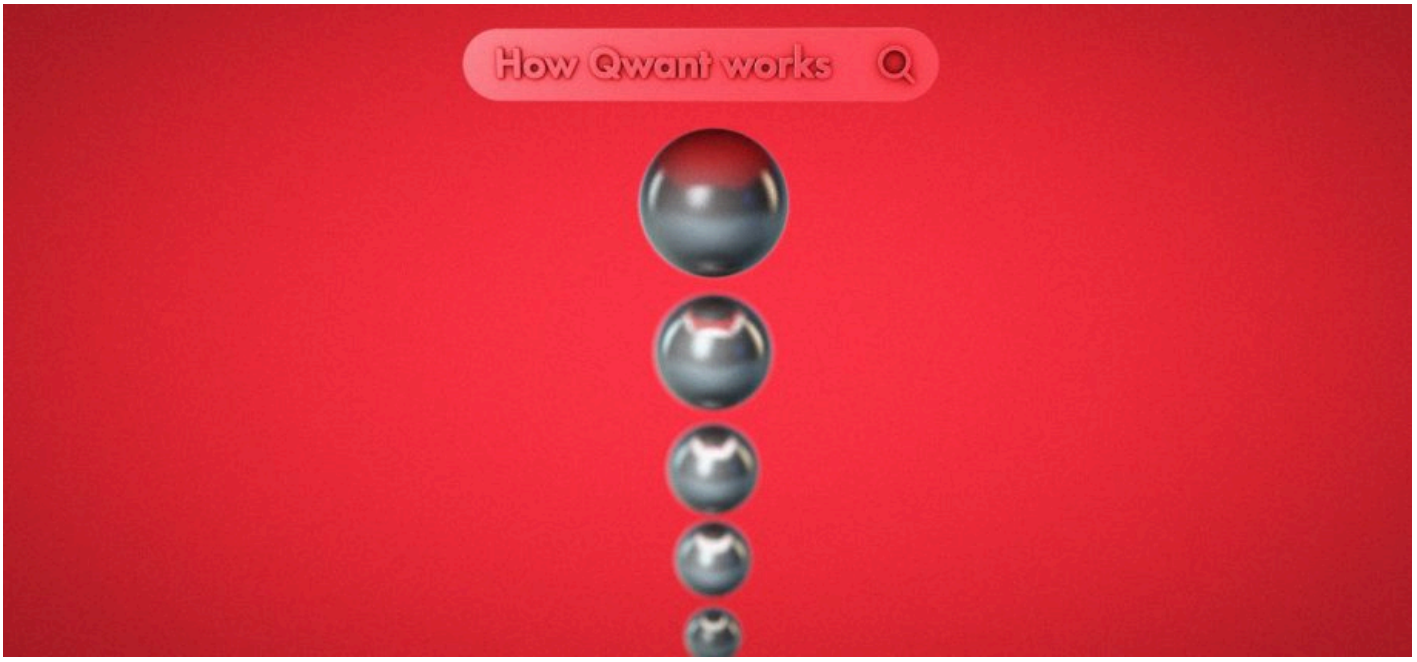
Qwant
Better Web

Better Web

How Microsoft tools strengthen Qwant



At VivaTech 2019, Qwant and Microsoft jointly announced a partnership to support Qwant's strong growth while continuing to respect the privacy of its users, as they have done since the creation of the search engine.

Team Qwant 25 June 2019 7min

At the VivaTech 2019 trade show, Qwant and Microsoft have announced a partnership to support Qwant's amazing growth while preserving user privacy. With this partnership, Qwant will use Microsoft Azure Cloud resources to compute its Web indexes in order to serve more accurate results.

Qwant pitched us their ambition: create an European alternative to the big tech platform that embodies European values. @Microsoft wants to help them to reach that great goal respecting citizens privacy! #VivaTech pic.twitter.com/T0y2oKEuEr

— Qwant (@QwantCom) May 17, 2019

Qwant's growth is amazing and this means we need to adapt our server infrastructure so that we can keep up with the growth while improving the quality of our search results. This is why we have decided to update our server infrastructure so that all our servers are used to interact with users (so we know for sure that no user data is being collected) while we're moving to Azure's Cloud server services that are not dealing with user data. As we are always looking to improve the quality of our results, this strategy will also allow us to increase our coverage of the Web by indexing more pages.

To create the search engine you know and love, we own and run more than 400 servers that are located in a data center near Paris. They are used for several purposes: the front (the user interface what you see), the crawl (discovering and updating content), the storage of texts and images, indexing, maps, news, our internal tools and many others. It is important to notice that nearly two thirds of this equipment is used to run Qwant Search (whether it is production or pre-production). However, our technical resources are not infinitely expandable. The purchase and renewal of materials is very expensive and it is a constraint to our development. Thus, we will rent resources from Microsoft Azure to support and increase our development while continuing to invest in our own infrastructure.

To understand in more detail what will happen thanks to this partnership, here is an explanation on how the web pages work in our servers.

25 servers visit 12,000 web pages per second

The first step in creating a search engine is to discover which content is on the Internet. Indeed, apart from domain names and IP addresses, nothing is standardized. You can create the site with the structure and content you want

without limits. In order for us to be informed of the existence of all websites, we must use crawlers. Today we have about 25 machines that are dedicated only to this task, they are able to visit 12,000 web pages per second, which is a little more than a billion pages per day. In addition to discovering new sites, crawlers must regularly revisit already known sites to see if their content has changed since last visit, sometimes that happens very often for news sites.

Once the crawler has done its work, the data extracted from the web is raw and must be analyzed before it can be entered into our databases and shown to our users. This is where we use computer analysis programs that are available in the form of micro-services. They are used to determine, for example, the language of the page or the trust scores of web pages via calculations on the web graph (the relationship between pages). Thanks to the partnership with Microsoft, we will take advantage of Azure's Kubernetes service to easily deploy and operate these micro-services for production analysis and research. The more we can deploy such micro-services, the more our engine will take into account quality signals to provide the best possible results and thus improve the relevance of the engine in the future.

20 billion indexed documents

When the crawl data is analyzed and formatted, it must be indexed. In other words, we need to create an index, similar to what you may find at the end of books: a list of words and on which page they appear in the book (except that we're talking about billions of pages). Today, our server base allows us to reach 20 billions of indexed documents, or several hundred terabytes of data.

However, our crawlers discover more documents than it is possible to index because of the physical limit of the number of servers available. The crawler downloads nearly 10 gigabytes per second. Currently we prioritize indexing according to, for example, the language and graph score (popularity) of the document. Web pages that are not indexed are stored in an archiving system, this represents nearly 2 petabytes of data. By using Azure's storage solutions we will be able to free up space at home to increase the size of our indexes and then increase the quality of our results and increase our web coverage. In addition, we will be

able to use Azure Cloud computing resources to test new indexing methods or new languages before moving them into production. This gives us much-needed flexibility to deliver functional services more easily and quickly.

Strengthen our computation capacity of the graph

When the index is ready we can easily find all the documents that are linked to a search. However, let us take the example of “cat” which will bring up tens of millions of web pages from our indexes. Which pages should be highlighted for this keyword? To show the most relevant pages first, we must rank them, using a multitude of factors (including the famous web graph scores) and machine learning methods to combine all the available information into a single decision-making process for each web page.

The calculation of graph scores is very expensive, on the one hand because the web graph represents terabytes of data, and on the other hand because it requires several servers working at the same time. Thanks to the partnership with Microsoft, we will be able to use Azure on demand resources to calculate the graph and the associated scores. We will also be able to use GPU and then FPGA-based infrastructures to train our Learning To Rank models.

A team of expert researchers to improve the index

For nearly a year and a half, many researchers (language processing, imaging, cryptography, etc.) have joined Qwant. We are already using very powerful graphics cards in Nvidia’s DGX servers. These servers are used to train machine learning models, for machine translation, image search, optimization of graph algorithms, etc. We are also working on document analysis and information retrieval technologies using FPGAs.

Today, one of our most ambitious projects is the image search* that integrates into our crawler and must be scaled up to production. We must therefore store hundreds of millions of images, or several hundred terabytes of images. The

storage that Azure will offer us will be very useful to increase the size of our image index and facilitate the learning of models.

The user is not talking to Microsoft

At last, let's talk about you, users. When you use Qwant, you will always be connected to machines that we own and operate directly. You will never be connected to Azure's cloud machines, and your personal data is never shared with third parties. We use Azure for Qwant's back office purposes, namely computing the index of the Web. We take this opportunity to remind you that as soon as you connect to our search engine, our servers anonymize your data, especially your IP address which is salted and hashed. That is, we add noise and a breakdown of this IP address to make it anonymous when we ask to display ads or have to store data in our logs. Only this anonymous data is being used in our internal network.

* https://research.qwant.com – Scientific paper: Image search using multilingual texts: a cross-modal learning approach between image and text, M. Portaz et al., https://arxiv.org/abs/1903.11299

Relative posts

Better Web

Europe must protect our personal data!

Team Qwant

29 January 2021 12 min

Better Web

#BetterWeb 11: Gaëla Lahnine – 12 million French people with disabilities concerned by digital accessibility

Team Qwant

15 November 2021 10 min

Better Web

#BetterWeb 03: Guillaume Chaslot – AlgoTransparency to break filter bubbles

Team Qwant

2 April 2021 1 min

The search engine where you are the user, not the product!

Discover the service

Services

Search

Junior

App Android

App iOS

About Qwant

Why Qwant?

Better Web

Join Us

Help Center

Contact Us

Press

Follow us !

Privacy Terms of Service Legal notices Ranking

© 2026 Qwant. All right reserved.