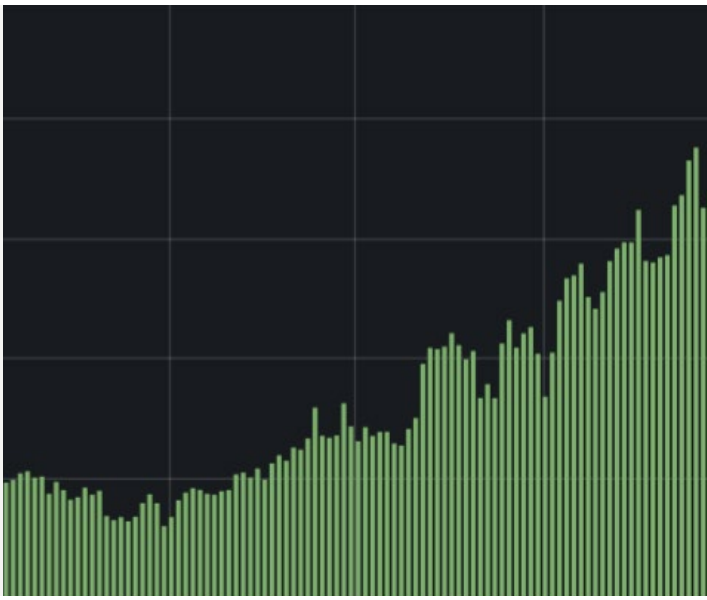
Backblaze

urlscan.io

Threat Analysis Firm Taps Backblaze in the Fight Against Cybercrime



Use Cases
Primary Storage / Content Distribution

Keywords
InfoSec / SaaS / NGINX

1B
Files

700K+
Websites Scanned Per Day

<1s
Retrieval in the 95th Percentile

Situation

Every day, web threat analysis service urlscan.io runs 700,000+ website scans seeking out malicious content, capturing 3.5+ million files and artifacts in the process. With a commitment to keep the results of website scans in perpetuity, and a 60TB-and-growing storage footprint, the company needed capacity and performance that would scale with growth. Unfortunately, their original, multi-platform solution couldn't keep up.

Solution

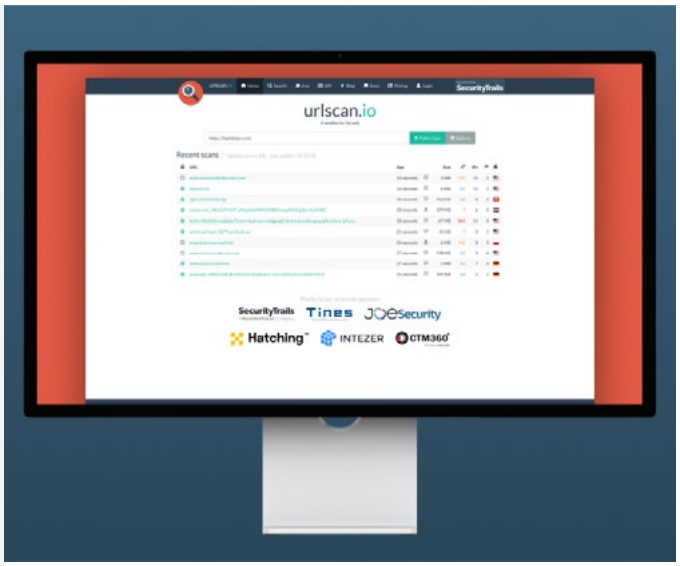
urlscan.io chose Backblaze B2 Cloud Storage for its speed and consistent response times, but the ease of integration sealed the deal: The Backblaze S3 Compatible API made migrating screenshots and document object model (DOM) snapshot data easy, and caching on NGINX with hot storage on Backblaze B2 enables tens of thousands of daily users to simultaneously access files at speed, without needing a CDN.

Result

Now, urlscan.io easily manages retrieving and permanently storing billions of files on a lean stack that scales effortlessly with their daily data growth. Budgeting is simple and performance is right where urlscan.io wants it, with customers reporting a great user experience. Backblaze gives the team "absolute peace of mind," and the freedom to focus on their mission: protecting businesses and individuals from malicious websites.

urlscan.io
A sandbox for the web

Founded in 2016, urlscan.io is a highly regarded threat analysis service used by Fortune 100 companies, governments, and others in the InfoSec industry. It allows anyone to safely and easily analyze unknown URLs, so they can identify whether or not they lead to malicious websites. With the threat of malware and phishing continuing to grow, urlscan.io helps individuals and businesses to better protect their identity, reputation, and systems.



Threat Analysis Firm Uses Backblaze to Catalog "Most Wanted" Malicious URLs

Today, malware is everyone's concern. According to Webroot, four million new high-risk URLs came into existence in 2021. To make matters worse, nearly 66% of these new websites involved some type of phishing. As malicious actors find new ways to lure unsuspecting users through frighteningly deceitful deceptions, malware will continue to be a threat for anyone who uses the internet. Luckily, there are crime-fighting tools available to help anyone, from individuals to major businesses, protect their identities, reputations, and their systems.

In the InfoSec industry, urlscan.io—a sandbox service that gives users a safe way to analyze a suspicious URL to ensure that it leads to a reputable website—is rapidly gaining a reputation as an essential protection against bad actors. Johannes Gilger, urlscan.io's founder, created urlscan.io during his time working in threat intelligence because he was unhappy with the state of tooling at the time. No single tool could run a standard set of website analysis he needed on a regular basis.

What started as a side hustle grew with organic demand, so Gilger launched urlscan.io in 2020 as a formal business and added paid plans, support, and other commercial features. Today, urlscan.io is used globally by Fortune 100 corporate security teams, government agencies, consumer protection organizations, web developers, and individuals. Some companies use urlscan.io to automate email processing, scanning incoming messages for potential risk and maintaining records of malicious URLs. However, urlscan.io is such a flexible tool that it can be used to automate other website analytic workflows or solve a broad range of problems.

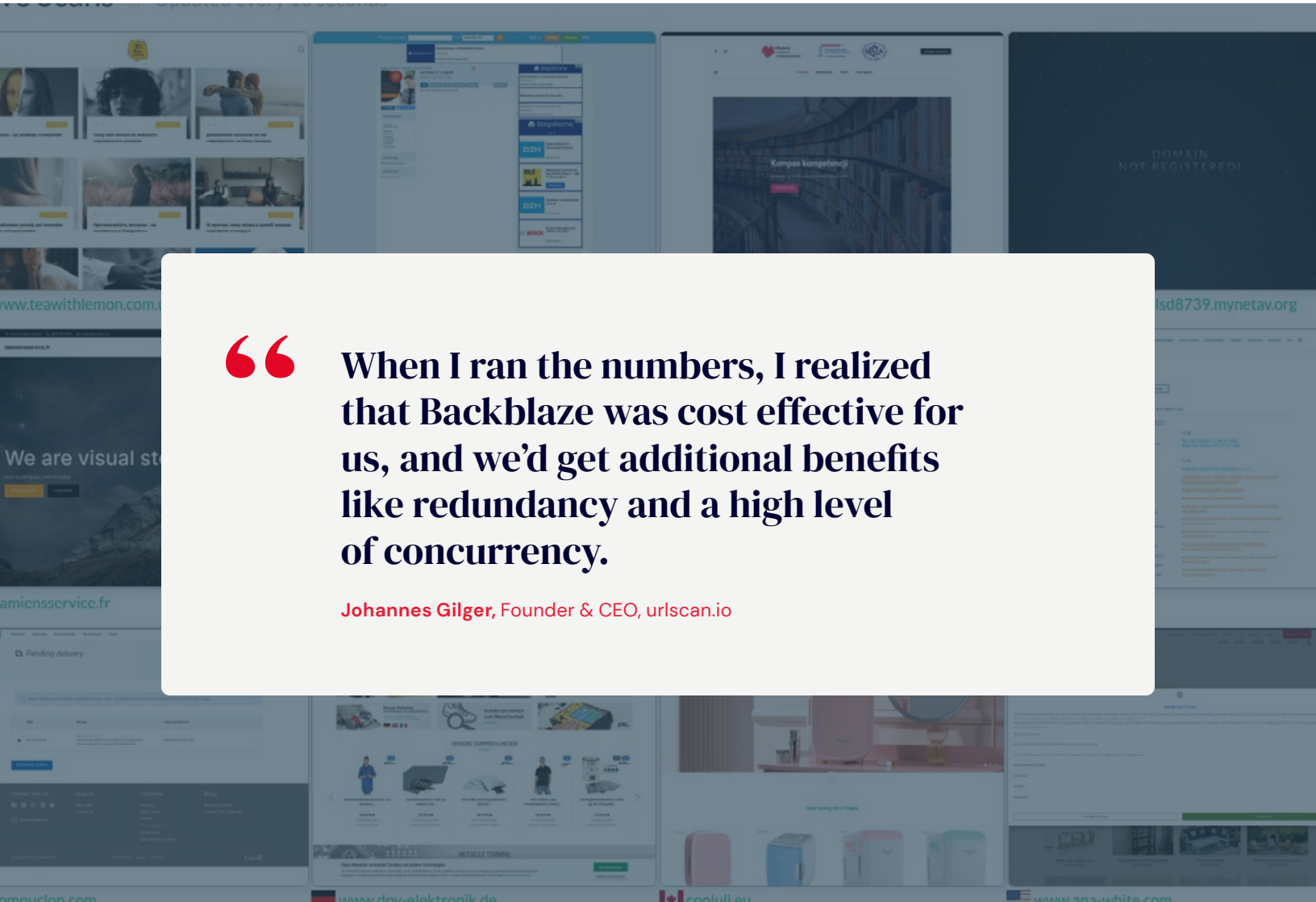
Fun fact: Backblaze uses urlscan.io to flag malicious subdomains and phishing pages being served from B2 Cloud Storage to stop malware before it gets the chance to spread.

Tracking Down the World's Malicious Websites

When a user submits a URL for analysis, urlscan.io runs an automated process that navigates the destination site and captures the activity data, such as the domains and IPs it contacts or the resources it requests. The platform then analyzes this data, compares it to legitimate domains that are commonly faked, and identifies any malicious patterns. Finally, it returns a verdict on the URL's potential risk level and identifies the brand it is attempting to impersonate.

Once scanning is complete, urlscan.io archives the associated files, including a full resolution screenshot of the page, a document object model (DOM) snapshot of the document object tree, the HTML content of the site, certain HTTP responses, such as JavaScript global variables and cookies created by the page, and analytics data.

Every day, urlscan.io runs over 700,000 scans, and with approximately five files produced per scan, that adds up to about 3.5 million new files per day. Screenshots and DOM snapshots are not compressed, and the platform also stores file downloads. Part of urlscan.io's service is to store these files in perpetuity, which means an ever-growing need for storage. To make the venture successful, urlscan.io had to figure out a way to store all of this data efficiently and effectively.



> **When I ran the numbers, I realized that Backblaze was cost effective for us, and we'd get additional benefits like redundancy and a high level of concurrency.**

Johannes Gilger, Founder & CEO, urlscan.io

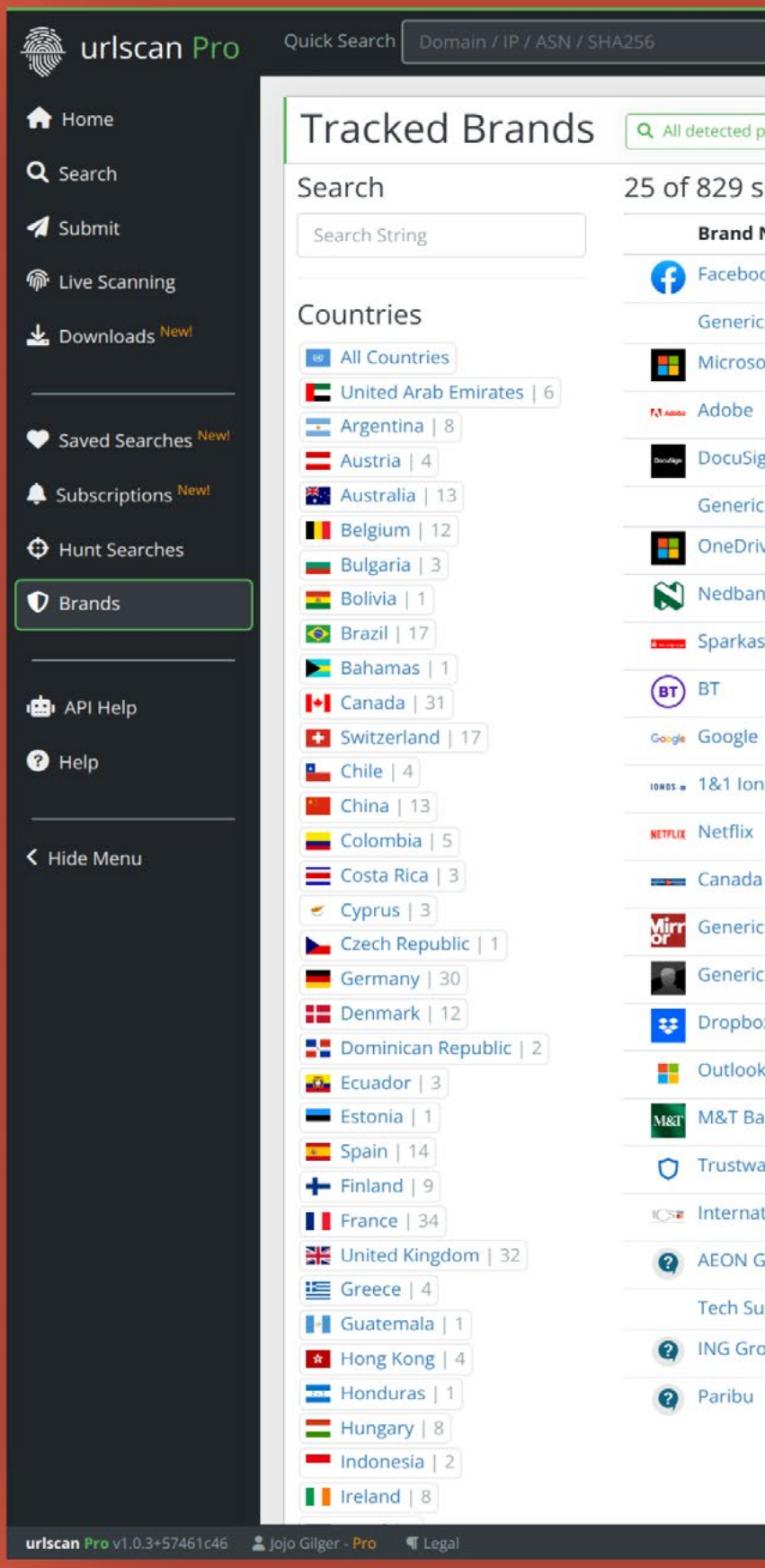
Rounding Up the Usual Suspects

Previously, urlscan.io was storing scan files on both a S3-compatible, MinIO cloud server and on Storage Box, an object storage service from its hosting provider, Hetzner. Screenshots were stored in MinIO and DOM snapshots in Storage Box, with structured data, such as HTTP transactions and hostnames, stored as a JSON blob in MongoDB.

Although the solution was affordable for the fledgling startup, its limited scalability couldn't keep up with the company's growth. "MinIO was inflexible and slow on spinning disks because it's not optimized for small files," said Gilger. "Storage Box had frequent issues and it limited us to 10TB of storage and 10 concurrent connections. Also, we needed a solution that allowed us to add capacity on the fly, which was not possible with either of these existing solutions." In addition, this setup was urlscan.io's single source of truth, with no backup or redundancy for the growing archive.

When it came time to upgrade, the urlscan.io team evaluated a few solutions, including AWS. "Storage pricing was not an immediate concern," said Gilger, "but it was the unknown transfer costs and other hidden fees with AWS that could have been problematic for the business."

Gilger had been aware of Backblaze for a long time, and decided to investigate it as a possible solution. "When I ran the numbers," he said, "I realized that Backblaze was cost effective for us, and we'd get additional benefits like redundancy and a high level of concurrency."



Recruiting Backblaze B2 as a Partner in Crime (Fighting)

The urlscan.io team ran a proof of concept with Backblaze B2 Cloud Storage, primarily to check for speed and latency. They wrote a script and used the Backblaze S3 Compatible API to upload a million files with different levels of concurrency. "We wanted to see how fast we could write data to Backblaze," Gilger said, "and how fast we could get it out. Was it saturating our line speed for uploading files? What was the overall latency distribution?" The team hit their retrieval latency goal in the 95 percentile—below one second—and they were able to saturate their upload queue without having a high number of concurrent requests. Most importantly, the overall consistency of response time was solid.

"We were really happy with the performance of Backblaze B2," said Gilger. "Low latency is a huge requirement for us because we want our users to have a fast, smooth experience with every scan. Our site is very visual where users retrieve dozens of screenshots on almost every page. We were able to download 100 full resolution screenshots simultaneously at full speed on Backblaze."

Once they adopted Backblaze B2, the urlscan.io team used the Backblaze S3 Compatible API to migrate data from MinIO and Storage Box to Backblaze B2. Gilger said, "The Backblaze S3 Compatible API was crucial because we had some code using S3 libraries, so writing to Backblaze B2 meant just changing the bucket names, locations, and files."

> **Low latency is a huge requirement for us because we want our users to have a fast, smooth experience with every scan. We were able to download 100 full resolution screenshots simultaneously at full speed on Backblaze.**
>
> **Johannes Gilger,** Founder & CEO, urlscan.io

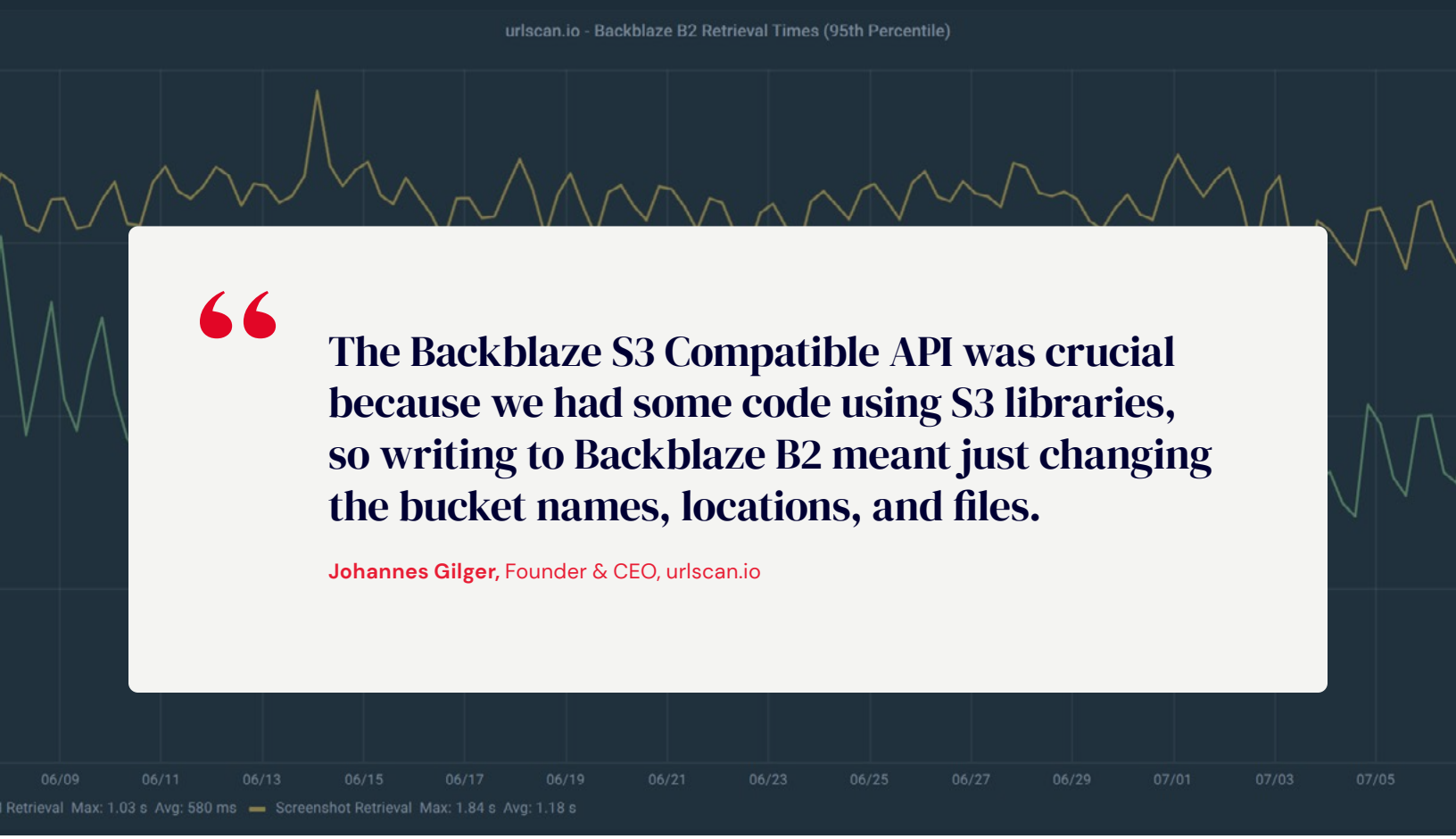
Keeping Trusted Associates to a Minimum

urlscan.io's tech stack is simple and streamlined by design. Built in Node.js on the backend with JavaScript on the frontend, the platform runs on a traditional hosting service with a strictly limited number of third-party integrations. "In general, we are apprehensive about using external services," said Gilger. "We try to keep our footprint of third-party services to a minimum, which helps us stay in control of everything."

urlscan.io doesn't even use a CDN to speed up file transfers across geographies. Instead, they deployed NGINX to provide some light caching in front of Backblaze. To get files out of their Backblaze B2 buckets, the team uses an S3 proxy in a Docker container that generates pre-signed URLs in order to access files transparently from NGINX.

With tens of thousands of daily visitors, most of urlscan.io's use cases benefit from hot storage. Its live page, for example, might have 100 users looking at the same screenshot at any given moment, and it makes sense to cache these for a few minutes. But the platform also caches older scans, so that users can retrieve them faster. For anything else, Backblaze's transfer fees for non-cached items fell well within Gilger's expectations compared to other services like AWS.

Currently, urlscan.io is storing 60TB of data from 1 billion files on Backblaze B2. Eventually, the urlscan.io team plans to move its structured data in MongoDB over to Backblaze B2 as well. The reason, said Gilger, is, "It doesn't make sense for us to store those big JSON blobs in MongoDB because we never update the data. It's more or less write-only, so it's better for them to go into an object storage database."



> **The Backblaze S3 Compatible API was crucial because we had some code using S3 libraries, so writing to Backblaze B2 meant just changing the bucket names, locations, and files.**
>
> **Johannes Gilger,** Founder & CEO, urlscan.io

Keeping Trusted Associates to a Minimum

urlscan.io's tech stack is simple and streamlined by design. Built in Node.js on the backend with JavaScript on the frontend, the platform runs on a traditional hosting service with a strictly limited number of third-party integrations. "In general, we are apprehensive about using external services," said Gilger. "We try to keep our footprint of third-party services to a minimum, which helps us stay in control of everything."

urlscan.io doesn't even use a CDN to speed up file transfers across geographies. Instead, they deployed NGINX to provide some light caching in front of Backblaze. To get files out of their Backblaze B2 buckets, the team uses an S3 proxy in a Docker container that generates pre-signed URLs in order to access files transparently from NGINX.

With tens of thousands of daily visitors, most of urlscan.io's use cases benefit from hot storage. Its live page, for example, might have 100 users looking at the same screenshot at any given moment, and it makes sense to cache these for a few minutes. But the platform also caches older scans, so that users can retrieve them faster. For anything else, Backblaze's transfer fees for non-cached items fell well within Gilger's expectations compared to other services like AWS.

Currently, urlscan.io is storing 60TB of data from 1 billion files on Backblaze B2. Eventually, the urlscan.io team plans to move its structured data in MongoDB over to Backblaze B2 as well. The reason, said Gilger, is, "It doesn't make sense for us to store those big JSON blobs in MongoDB because we never update the data. It's more or less write-only, so it's better for them to go into an object storage database."

Low Latency, Solid Pricing Help Backblaze B2 Earn Its Stripes

Since switching to Backblaze B2, urlscan.io has seen a number of benefits. For instance, Gilger finds that his costs have remained largely unchanged and are well under control. "Backblaze's flat pricing structure allows us to store everything permanently and its performance enables users to access any file with very low latency." This came into focus recently when the urlscan.io team added a new feature and had to re-index their entire archive. Gilger recalls, "It was very reassuring to see that even if we have to retrieve every file, we can still manage the costs very easily."

Latency has continued to be low despite growth. "We're always monitoring latency because it's an SLA for our customers," said Gilger. "With Backblaze, we've actually seen latency improve significantly. People repeatedly tell us that our application is very snappy and very responsive."

In fact, customers are asking for more. The urlscan.io team is expanding their product roadmap with new offerings that serve specific customer needs, such as brand protection or file analysis. Being able to scale without hard bottlenecks imposed by technology is crucial for urlscan.io. By leveraging Backblaze B2, urlscan.io is able to scale up its scanning efforts without hitting limits imposed by its storage layer.

Ultimately though, peace of mind is one of the biggest benefits. "We can sleep at night without worrying about database issues or running out of space some day and all the engineering time and effort of migrating to another solution. Backblaze gives us absolute peace of mind."

About Backblaze

The Backblaze B2 Storage Cloud is purpose-built for ease. It offers always-hot, S3 compatible object storage that supports your workflows via third-party software integrations, APIs, CLI, and web UI. And it's priced for easy affordability at rates a fraction of other cloud providers. Businesses in more than 175 countries use the platform to host content, build and run applications, manage media, back up and archive data, and protect and recover from ransomware.

backblaze.com

Backblaze