The “Have I been pwned?” Microsoft Azure Ecosystem

The UI is responsive across devices of all sizes. It’s a first class experience from small smartphones to large, high DPI displays. This is achieved through CSS media queries combined with extensive use of scalable vector graphics.

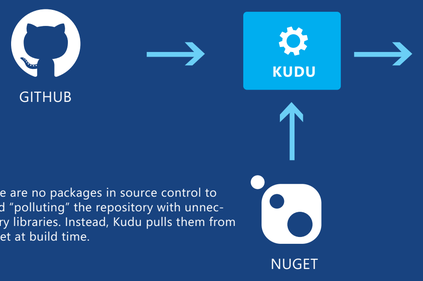
Breach and paste search features are fully implemented by a publicly facing API. It’s free for use without authentication or rate limits and drives a number of mobile apps and other community projects.

Deployment

Deployment of the web site is entirely automated using GitHub and the native constructs available within the Azure website service.

All code goes into GitHub in feature branches. The “master” branch remains production ready and when it’s time to release, is merged into a “deploy” branch.

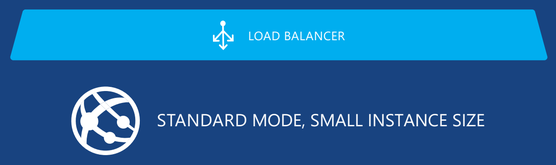
Kudu automatically picks up any changes in the “deploy” branch and publishes them to a staging deployment slot in seconds. Once tested and ready for go live, the staging slot is swapped with the production one and becomes live almost instantaneously.



There are no packages in source control to avoid “polluting” the repository with unnecessary libraries. Instead, Kudu pulls them from NuGet at build time.

Website

The website is the face of HIBP and it runs on Azure’s website as a service offering. The one site hosts both the HTML interface built in MVC 5 and the API back end in Web API 2.



The website runs on the smallest possible instance size to keep cost at a minimum. Autoscale is configured by CPU with a target range between 60% and 80%. It will autoscale all the way out to 10 instances in response to demand. It can be manually scaled up to medium or large instances which further double or quadruple the capacity.

A scheduled WebJob sends reminder notices to subscribers who have not yet verified their email address. It sends 3 days after the original email and tailors the message to further reduce the chance of being caught by spam filters.

WEBJOB

CDNs

Multiple CDNs are used to both get content closer to customers for performance gains and move bandwidth to free providers.

The Azure CDN is used for all breached company logos. The logos are loaded into Blob Storage as scalable vector graphics and automatically distributed across the globe to each CDN endpoint.

Public CDNs are used for jQuery, Bootstrap and Font Awesome. This keeps the bandwidth costs off Azure and serves the assets from the fastest possible location for the user. It also provides them the opportunity to load them from cache if already loaded via another website using the same CDN service.

Monitoring and alerts

Different monitoring services are used to track resource utilisation in real time down to a very fine grain. Alerts automatically advise when noteworthy events occur that may require attention.

Azure Monitoring

Azure’s built-in monitoring is used extensively to gauge resource utilisation and performance. The native alerts for when thresholds are hit are used to provide early warning of events such as increased traffic load. This also serves as advance notice of when the monthly bill may increase.

New Relic

New Relic runs in-process within the website to monitor internal transaction times as well as in the browser to monitor client timings. The data is then aggregated on New Relic’s portal. The service also monitors availability and sends alerts when an outage occurs.

Raygun.io

All unhandled exceptions are automatically logged by Raygun.io and triaged in their portal. Notifications are automatically sent when new errors appear or previously resolved ones reappear.

Google Analytics

All website activity is logged via Google Analytics and analysed to better understand usage patterns. This is predominantly to look at traffic sources and trends over time.

Breach Processing

Breaches are manually verified for accuracy which requires a degree of pre-processing in an isolated environment. Different logical machines are used for analyses and then importing into Table Storage.

This VM runs a console app which takes in a cleansed file of unique email addresses from the breach. It imports them into the Table Storage facility then sends notifications to impacted subscribers.

In cases there the breach is very large (100M+ records), a high performance VM with SQL Server is used for the analysis. It’s expensive, but it runs for very short durations to keep cost low.

Some breaches come in forms that could contain malicious software. These are analysed and records extracted in a sandboxed VM that doesn’t directly touch the broader HIBP service.

Storage

The data tier spreads storage across both a relational SQL Azure database (SQL as a service) and Azure Table Storage (NoSQL data store). Each has its own advantages in terms of speed and cost.

A SQL database is used for relational data which benefits from being queried in an ad hoc fashion. This includes a list of all breaches and pastes, subscriber data for notifications and domain monitoring subscriptions.

SQL Database

The SQL Azure backup is a combination of the 7 day point in time restore feature of the Basic tier and an automatic export done weekly with a 28 day retention period.

Backup to Blob Storage

Table Storage

Table Storage is used for storing all breaches and pastes. It’s massively fast for looking up records by a key and comes at a fraction of the cost of a SQL database.

Backup of Table Storage is a manual periodic backup using Red Gate’s Azure Backup service.

Backup to Blob Storage

Paste Service

The paste service automatically imports dumps that appear on Pastebin and other paste services in response to tweets from the @dumpmon account. This is often the earliest public indictor of a breach.

The @dumpmon Twitter account is continuously monitored by a worker role. New tweets about pastes with email addresses in them have the paste URL extracted and stored in a queue.

The queue is continuously monitored for new paste URLs. When a new message appears, it’s popped off the queue, retrieved from Pastebin, saved to storage then deleted from the queue. Median time to first email being searchable after the paste is posted is 33 seconds.

Additional services

Mandrill
Email

Email is used extensively for subscription verification and reminders. It’s also used for domain search verification and notifications when a breach or paste impacts a subscribed account.

FeedBurner
RSS

RSS is exposed via FeedBurner which consumes the source feed from the HIBP website. This keeps the load and the bandwidth on Google’s free service with only a very small overhead on the Azure website.

WHOIS
API

Before searching for breaches and pastes across an entire domain, verification of domain ownership is required. One verification channel is an email to the domain contact for which a service is required to retrieve registered email addresses.

UserVoice

Ideas and feature requests are shared and voted for via UserVoice as a means of crowd-sourcing future development and priorities.

dnsimple

All DNS and registration services for HIBP domains are managed by DNSimple.