NEWS

Human error root cause of November Microsoft Azure outage

The Nov. 18-19 outage was caused by an update that Microsoft personnel assumed to be sound, but wasn't

By Joab Jackson

U.S. Correspondent, IDG News Service

Shutterstock

Human error was the culprit for a November outage of the Microsoft Azure cloud storage service. The company is hoping that recent updates that automate formerly manual processes will help prevent similar outages in the future.

"Microsoft Azure had clear operating guidelines but there was a gap in the deployment tooling that relied on human decisions and protocol," wrote Jason Zander, Microsoft vice president for Azure, in a blog post Wednesday detailing the outage. "With the tooling updates the policy is now enforced by the deployment platform itself."

This is not the first time Azure has been bedeviled by human failure.

In February 2013, a lapsed security certificate led to a major Azure outage .

Both cases show how even small errors can have a huge impact in a service as large as Azure and seem to have reinforced for Microsoft the importance of automating manual processes as thoroughly as possible.

This latest Azure outage happened late in the evening of Nov. 18, Pacific time, due to intermittent failure from some of the company's storage services.

Other Azure services that relied on the storage service also went offline, most notably the Azure Virtual Machines.

The outage stemmed from a change in the configuration of the storage service, one that was made to improve the performance of the service.

Typically, Microsoft, like most other cloud providers, will test a proposed change to its cloud services on a handful of servers. This way, if there is a problem with the configuration change, engineers can spot it early before a large number of customers are affected. If the change works as expected, the company will then roll out the change to larger numbers of servers in successive waves, until the entire system is updated.

In the case of this particular change, however, an engineer assumed that the update had already been tested in a number of waves (or "flights" in Microsoft parlance), and so went ahead and applied the change across the rest of the system.

The configuration, however, contained an elusive bug that would cause the storage service software to go into an infinite loop, preventing further communications with other components of the system.

Microsoft engineers quickly pinpointed the problem and issued fixes. By 10:50 a.m. UTC, the storage service was completely back online, though restoring all of the virtual machines, a small number of which were isolated from the network due to the outage, would take another two days.

In the weeks that followed, Microsoft investigated in detail what went wrong, as well as looked into ways to make sure the outage wouldn't happen again. As a result, the company has updated its deployment system so that it now enforces the testing and flighting policies before new code or a change goes live across the entire system.

"With the tooling updates the policy is now enforced by the deployment platform itself," Zander wrote.

In the outage of February 2013, a failure in manual protocols was also to blame. Parts of the system went offline due to lapsed security certificates. The process to apply the updated certificates to Azure machines was scheduled with a larger routine update, a decision made by engineers who were unaware that the new certificates would not be delivered until after the old ones had expired .

After investigating the November incident, Microsoft wanted to share its "root cause analysis" with customers, Zander wrote, in hopes that users would find the act of transparency to be proof of Microsoft's commitment to providing quality cloud hosting services.

Overall, the act of posting of the root cause analysis seemed to please at least some Azure users and the IT community as well, despite the additional negative publicity it could bring Microsoft.

"I've seen several companies where analysis like this would be for management only. I guess it's just human nature to want to sweep mistakes and accidents under the rug, but it does also speak volumes about the culture in such companies. Kudos to Microsoft and every other big player that communicates these things," wrote a user on the Hacker News aggregation site.

Joab Jackson covers enterprise software and general technology breaking news for The IDG News Service. *Follow Joab on Twitter at @Joab_Jackson. Joab's e-mail address is Joab_Jackson@idg.com*

Related: Cloud Computing | Server | Disaster Recovery