NSA / PRISM

No, the NSA Won't Restore Your Crashed Hard Drive

So, yes, you still need to back up your data yourself.

By Doug Aamoth | Aug. 30, 2013

Backblaze is an online backup company. In light of jokes making the rounds alluding to the idea that backup services are now unnecessary considering the National Security Agency already has a copy of everyone's data, one of Backblaze's employees decided to put that theory to the test.

According to CNET, Backblaze's Yev Pusin sent a Freedom of Information Act request to the NSA in June asking "for a copy of all the agency's data relating to himself."

Perhaps unsurprisingly, the NSA denied Pusin's request, citing matters of national security; the NSA wouldn't even acknowledge "the existence or non-existence" of such data, according to its letter back to Pusin.

So, yes, you still need to back up your data yourself.

Guess what happened when Backblaze tried using the NSA for data backup [CNET]

https://www.cnet.com/news/guess-what-happened-when-backblaze-tried-using-the-nsa-for-data-backup/

Guess what happened when Backblaze tried using the NSA for data backup

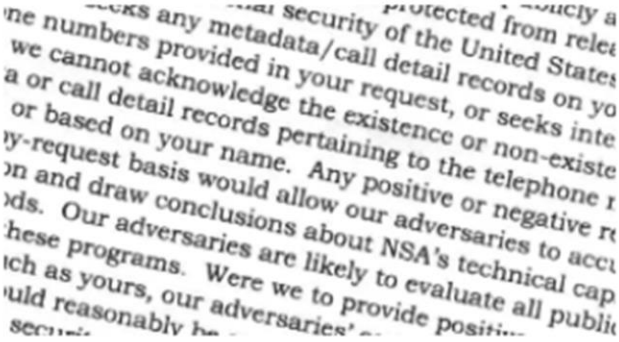
Nobody seriously believes that the NSA would helpfully give you backup if your hard drive failed. But an employee of the online storage company Backblaze asked anyway.

by **Stephen Shankland**

August 29, 2013 1:59 PM PDT

The revelations about the **National Security Agency's extensive snooping** led some **wags** to joke that people don't need to pay online backup services because the government has you covered already.

Nobody was serious about the idea, of course, but one online backup company, Backblaze, tried to see if it really was possible anyway. The answer, of course: no.



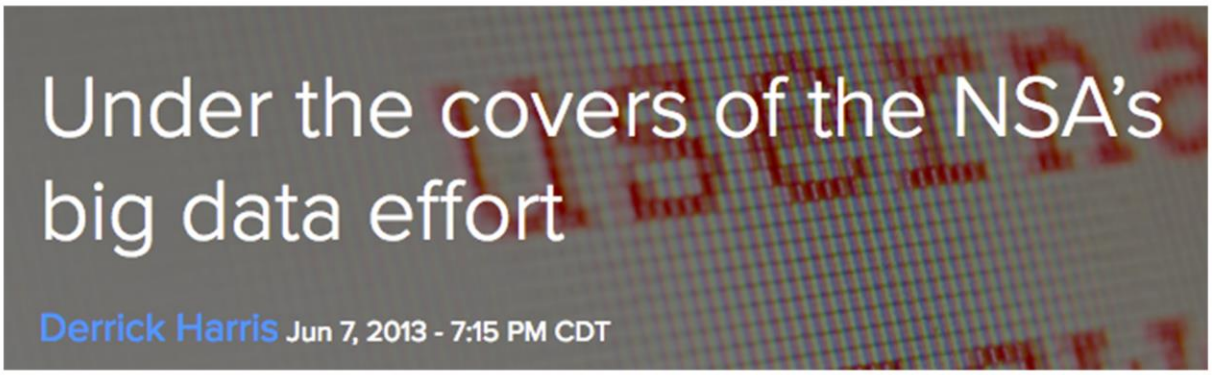
security of the United States
ne numbers provided in your request, or seeks inte
we cannot acknowledge the existence or non-existe
a or call detail records pertaining to the telephone
or based on your name. Any positive or negative
y-request basis would allow our adversaries to acc
on and draw conclusions about NSA's technical cap
ds. Our adversaries are likely to evaluate all publi
hese programs. Were we to provide
ch as yours, our adversaries
uld reasonably

An excerpt of the NSA's rejection of Yev Pusin's request for a copy of any data the agency has about him.

screenshot by Stephen Shankland/CNET

In June, Backblaze employee Yev Pusin sent the NSA a request under the Freedom of Information Act for a copy of all the agency's data relating to himself.

"Your request is denied because the fact of the existence or non-existence of responsive records is a currently and properly classified matter in accordance with Executive Order 13526, as set forth in Subparagraph (c) of Section 1.4," the NSA responded in a letter to Pusin. The letter also said FOIA information requests can be denied when they involve "matters that are specifically authorized...to be kept secret in the interest of national defense or foreign relations."



Under the covers of the NSA's big data effort

Derrick Harris Jun 7, 2013 - 7:15 PM CDT

The NSA's data collection practices have much of America — and certainly the tech community — on edge, but sources familiar with the agency's technology are saying the situation isn't as bad as it seems. Yes, the agency has a lot of data and can do some powerful analysis, but, the argument goes, there are strict limits in place around how the agency can use it and who has access. Whether that's good enough is still an open debate, but here's what we know about the technology that's underpinning all that data.

What is Accumulo?

The technological linchpin to everything the NSA is doing from a data-analysis perspective is Accumulo — an open-source database the agency built in order to store and analyze huge amounts of data. Adam Fuchs knows Accumulo well because he helped build it during a nine-year stint with the NSA; he's now co-founder and CTO of a company called Sqrrl that sells a commercial version of the database system. I spoke with him earlier this week, days before news broke of the NSA collecting data from Verizon and the country's largest web companies.

The NSA began building Accumulo in late 2007, Fuchs said, because they were trying to do automated analysis for tracking and discovering new terrorism suspects. "We had a set of applications that we wanted to develop and we were looking for the right infrastructure to build them on," he said.

Adam Fuchs

The problem was those technologies weren't available. He liked what projects like HBase were doing by using Hadoop to mimic Google's famous BigTable data store, but it still wasn't up to the NSA requirements around scalability, reliability or security. So, they began work on a project called CloudBase, which eventually was renamed Accumulo.

Now, Fuchs said, "It's operating at thousands-of-nodes scale" within the NSA's data centers. There are multiple instances each storing tens of petabytes (1 petabyte equals 1,000 terabyes or 1 million gigabytes) of data and it's the back-end of the agency's most widely used analytical capabilities. Accumulo's ability to handle data in a variety of formats (a characteristic called "schemaless" in database jargon) means the NSA can store data from numerous sources all within the database and add new analytic capabilities in days or even hours.

"It's quite critical," he added.

What the NSA can and can't do with all this data

As I explained on Thursday, Accumulo is especially adept at analyzing trillions of data points in order to build massive graphs that can detect the connections between them and the strength of the connections. Fuchs didn't talk about the size of the NSA's graph, but he did say the database is designed to handle months or years worth of information and let analysts move from query to query very fast. When you're talking about analyzing call records, it's easy to see where this type of analysis would be valuable in determining how far a suspected terrorist's network might spread and who might be involved.

Stewart Baker, former NSA general counsel under George W. Bush, wrote on his blog Thursday that this type of data could also be used for for general pattern recognition — the kinds of stuff that targeted advertisers love to do. Only, instead of the system serving someone an ad because of what they've been searching for and the operating system they're using, Baker presented the hypothetical of "[an] American who makes a call to Yemen at 11 a.m., Sanaa time, hangs up after a few seconds, and then gets a call from a different Yemeni number three hours later."

The big legal question here is around probable cause and whether the government should further investigate this caller based on call patterns similar to those of known terrorists, but the big data question is around false positives. Baker's hypothetical might appear pretty cut and dry but, data scientist Joseph Turian explains, call records in general probably don't offer too strong of a signal and could lead to situations where innocent behavior patterns looks a lot like nefarious ones. "But once you start connecting the dots with other pieces of information you have from other sources," he said via email, "you can start making more predictions."

predictions."

This is where a program like PRISM, the NSA's reported effort to collect data straight from the likes of Google, Facebook and Apple could come into play. If you're able to tie a name or web account to a phone number, you can figure out all sorts of information. If you can prove that certain people are radical Islamists, for example, you can start to infer more things about the others in that social graph.

And if Sqrrl's capabilities are any indicator of what Accumulo is supporting within the NSA, the agency can perform a lot of simpler functions on its data as well. In addition to graph processing, said Ely Kahn, Sqrrl's co-founder and VP of business development, their product includes pre-packaged analytic capabilities around SQL queries and full-text search, and also supports streaming data. This means Sqrrl's version can support any number of interesting use cases — from processing data as it hits the system to keeping a massive index that can be searched in the same way someone searches the web.

How much data is the NSA collecting? Follow the money

We're not quite sure how much data the two programs that came to light this week are actually collecting, but the evidence suggests it's not that much — at least from a volume perspective. Take the PRISM program that's gathering data from web properties including Google, Facebook, Microsoft, Apple, Yahoo and AOL. It seems the NSA would have to be selective in what it grabs.

Assuming it includes every cost associated with running the program, the $20 million per year allocated to PRISM, according to the slides published by the *Washington Post*, wouldn't be nearly enough to store all the raw data — much less new datasets created from analyses — from such large web properties. Yahoo alone, I'm told, was spending over $100 million a year to operate its approxi-

mately 42,000-node Hadoop environment, consisting of hundreds of petabytes, a few years ago. Facebook users are generating more than 500 terabytes of new data every day.

Using about the least-expensive option around for mass storage — cloud storage provider Backblaze's open source storage pod designs — just storing 500 terabytes of Facebook data a day would cost more than $10 million in hardware alone over the course of a year. Using higher-performance hard drives or other premium gear — things Backblaze eschews because it's concerned primarily about cost and scalability rather than performance — would cost even more.

Even at the Backblaze price point, though, which is pocket change for the NSA, the agency would easily run over $20 million trying to store too many emails, chats, Skype calls, photos, videos and other types data from the other companies it's working with.

Actually, it's possible the intelligence community is taking advantage of the Backblaze designs. In September 2011, Backblaze CEO Gleb Budman says, he met with CIA representatives who discussed that agency's five-year plan "to centralize data services into a large private cloud" and how Backblaze's technology might fit into it. Its plans for analyzing this data, as illustrated in the slide below (and discussed by CIA CTO Ira "Gus" Hunt at Structure: Data in March), seem to mirror what the NSA has in mind.