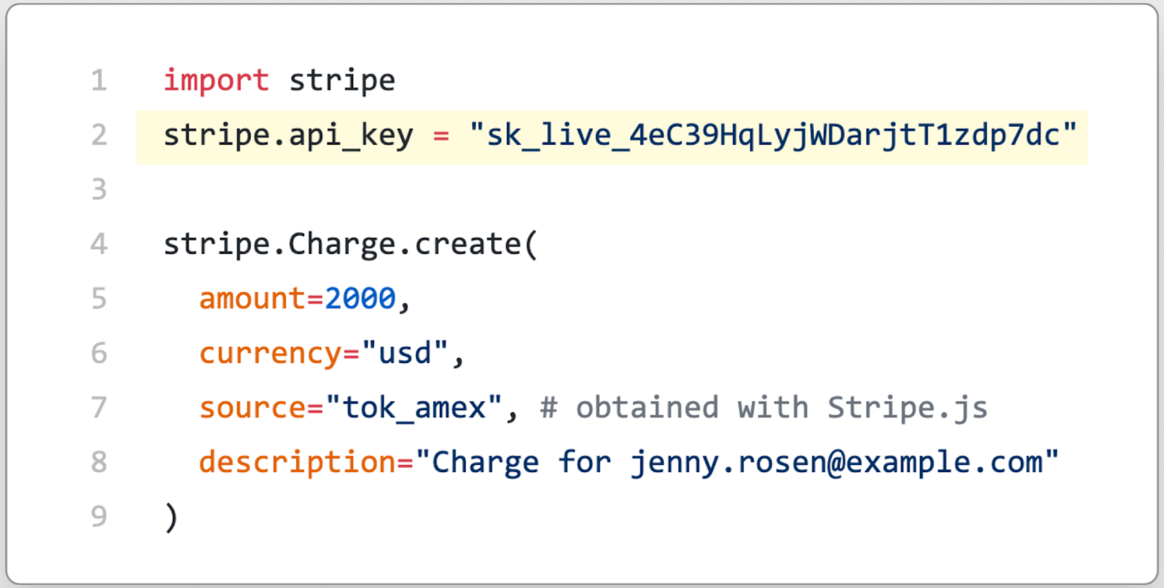
Nightfall

Guide to Data Loss Prevention in GitHub

Data leakage in the context of code repos

As a collaborative distributed version-control platform, Git-based repositories can create environments where secrets & credentials are exposed without notice. Developers can push commits at any time of day, and if no review process is in place they could push code that contains credentials and other sensitive token types.

In tandem with the growth of cloud-based Git repositories such as GitHub, the rise of service-oriented architecture (SOA) and microservices have increased the occurrence of secrets being leaked publicly. With developers making dozens of API calls to a multitude of services within their code, the possibility of secrets exposure can happen through practices like hard-coding PII and secrets or granting repositories overly broad permissions, allowing them to be viewed by unauthorized parties.

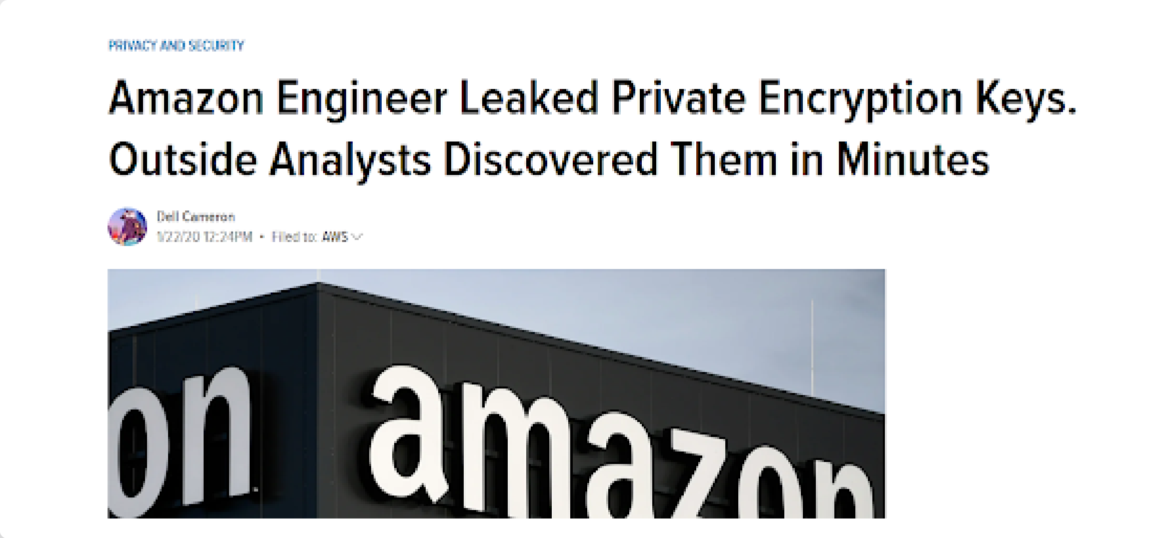


```
import stripe
stripe.api_key = "sk_live_4eC39HqLyjWDarjtT1zdp7dc"

stripe.Charge.create(
  amount=2000,
  currency="usd",
  source="tok_amex", # obtained with Stripe.js
  description="Charge for jenny.rosen@example.com"
)
```

Real world examples of secrets exposure

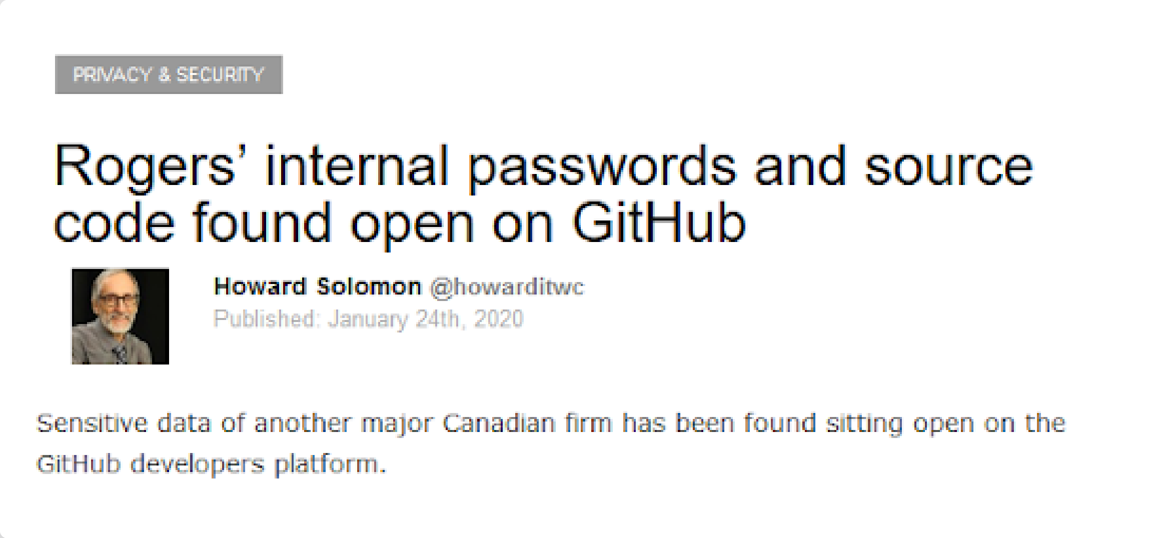
- **In 2016** hackers accessed Uber's private code repositories and used hard-coded credentials to exfiltrate 57 million driver records from an AWS S3 bucket. Uber was fined $148M.
- **In 2019** researchers discovered that thousands of unique secrets leak daily from GitHub repositories.
- **In January 2020** it was discovered that Canadian telecom Rogers Communications had left source code on the open internet for nearly 5 years.
- **Also in January 2020** an AWS DevOps Cloud Engineer committed nearly a gigabyte's worth of data to a personal GitHub repository bearing their own name.
- **In August of 2020** it was revealed that 9 U.S. based healthcare organizations leaked protected health information (PHI) for at least 150,000 patients, illustrating that credentials aren't the only thing at risk of being exposed in repositories.



PRIVACY AND SECURITY

Amazon Engineer Leaked Private Encryption Keys. Outside Analysts Discovered Them in Minutes

Dell Cameron
1/22/20 12:24PM • Filed to: AWS



PRIVACY & SECURITY

Rogers' internal passwords and source code found open on GitHub

Howard Solomon @howarditwc
Published: January 24th, 2020

Sensitive data of another major Canadian firm has been found sitting open on the GitHub developers platform.

What types of credentials & secrets get exposed in repos?

The most commonly exposed token types include:

- API keys & access tokens for 3rd party services, e.g. AWS, Stripe, Twilio, SendGrid, etc.
- Cryptographic keys (SSH, PGP, etc.)
- Certificates (SSL, TLS, etc.)
- Passwords
- Database credentials
- UUIDs, cookies, etc.

How can secrets be detected in repos?

- GitHub provides an automatic token scanning service for limited number of token types for popular services (e.g. AWS, Azure, Alibaba).
- TruffleHog – written in Python, uses regex and entropy-based flagging.
- Gitrob – written in Go, uses keywords and tries for a broader detection range than API keys.
- Git-secrets – limited to searching for AWS keys.

How do most secret detection tools work?

The main two methods used in traditional secrets detection tools are entropy and regular expressions.

Entropy refers to the amount of complexity in a string of characters. Setting thresholds for entropy can help determine the likelihood that a string is a credential as opposed to any other piece of information.

Regular expressions (regex) is used to search for expected characters that are anticipated to be part of a string but regex is bad at capturing variation across different types of services, e.g. AWS or GCP.

Key takeaway

Existing tools tend to result in high false positive rates without extensive fine tuning and observation. Most are not intended for the use case of managing secrets leakage across an entire code base.

How does Nightfall differ from existing tools?

When building Nightfall DLP for GitHub our team sought to develop a novel method leveraging deep learning to overcome the limitations of these methods.

Nightfall DLP is the industry's first cloud-native data loss prevention solution designed to **discover**, **classify**, and **protect** sensitive data in cloud environments. In the context of code repositories Nightfall is trained on features extracted from a broad set of API key patterns and their surrounding context in code. While other tools have attempted to combine context and regex, this context depends on naming conventions. But with natural language processing (NLP) and deep learning, naming conventions don't matter—only meaning does.

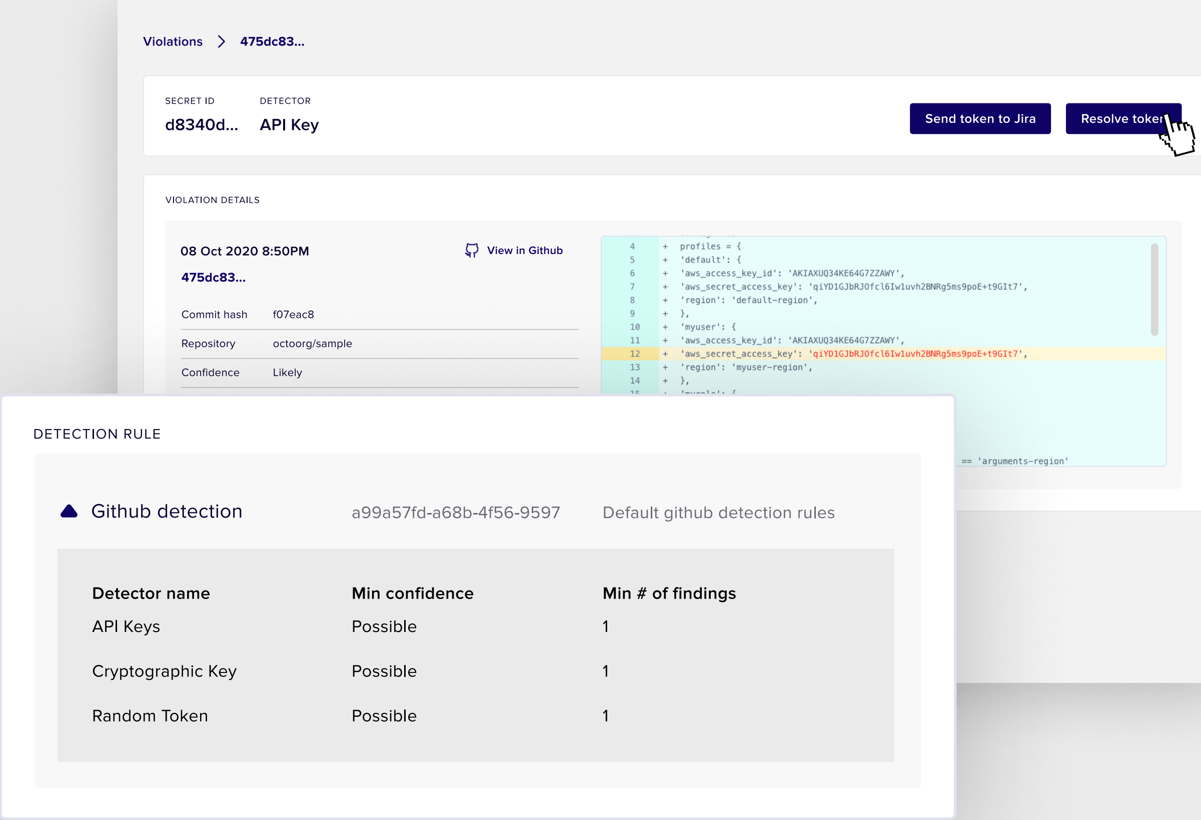
Key takeaway

Nightfall DLP is the industry's first deep learning based platform to detect credentials & secrets in code repositories. Designed to address the low accuracy of tools relying on traditional methods like regexes or entropy thresholds. Nightfall can be used to discover and protect against both PII and credential leakage across your code base.

What are the key features of Nightfall DLP for GitHub?

Nightfall helps organizations:

- Scan entire GitHub organization on every push to detect credentials, PII, and other secrets in public or private repositories via high accuracy machine learning.
- Choose which repos to scan as well as exclude specific tokens, files, and directories from scans via an allow list.
- Leverage pre-tuned detectors to discover secrets from any service or build custom detectors.
- View risk from an intuitive dashboard and inform developers of violations via Jira tickets. Once a secret is rotated, easily resolve all violations with that secret through a simple and intuitive dashboard.
- Send violation alerts to Slack and export results into a SIEM or reporting tool with custom webhooks.
- Drill into each violation to see details on the secret, the code snippet in GitHub, and any other violations with the same secret.



Violations > 475dc83...

SECRET ID d8340d...

DETECTOR API Key

VIOLATION DETAILS

08 Oct 2020 8:50PM

475dc83...

Commit hash	f07eac8
Repository	octoorg/sample
Confidence	Likely

```
+  profiles = {
+  'default': {
+  'aws_access_key_id': 'AKIAXUQ34KE64G7ZZAWY',
+  'aws_secret_access_key': 'qiYD1GJbRJOfcl6Iw1uvh2BNRg5ms9poE+t9GIt7',
+  'region': 'default-region',
+  },
+  'myuser': {
+  'aws_access_key_id': 'AKIAXUQ34KE64G7ZZAWY',
+  'aws_secret_access_key': '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',
+  'region': 'myuser-region',
+  },
== 'arguments-region'
```

DETECTION RULE

Github detection a99a57fd-a68b-4f56-9597 Default github detection rules

Detector name	Min confidence	Min # of findings
API Keys	Possible	1
Cryptographic Key	Possible	1
Random Token	Possible	1

What types of PII does Nightfall detect?

Nightfall can detect the following token types within images via OCR and over 100+ file types, including Google proprietary files types:

Standard PII: Age, Credit Card Number, Email, Ethnic Group, Name, Location, Phone Number

Health: ICD, FDA, DEA, NPI, DOB

Finance: IBAN, SWIFT, CUSIP, Routing Numbers

Crypto: Bitcoin, Ethereum, Litecoin Addresses & Private Keys

Network: IP Address, Hardware ID, MAC Address

Custom: API Keys, your application UUIDs, and much more.

IDs: Driver's License Number, Taxpayer ID, Passport Number, Social Security Number, Vehicle ID

What is Nightfall?

Nightfall is a platform to discover, classify and protect sensitive data across cloud SaaS & cloud infrastructure. Nightfall supports compliance efforts with a number of industry standards like PCI DSS, GDPR, HIPAA, CCPA, and much more. Additionally Google Drive is just one of the many platforms Nightfall secures.

Nightfall works by continuously monitoring data flowing in and out of data silos and classifying that data with machine learning. Data marked as sensitive can be automatically quarantined, deleted, and redacted with workflows.

Key benefits

- Get started quickly - no setup or tuning required.
- Leverage 150+ pre-tuned, standard detectors of PII out of the box.
- Rich analytics to examine all your PII risk, both in real-time and historically.

Discover

Continuously monitor data that is flowing into and out of data silos.

Classify

Machine learning classifies sensitive data & PII automatically

Protect

Automated workflows for quarantine, deletion, redaction, alerts, and more.

Case study: Galileo Health

Industry: Healthcare
Employees: 55
Head of Security and Compliance: Michael Supon
Profile: Leveraging a digital platform and a unique, multi-disciplinary care management system, Galileo provides an effective, scalable solution for delivering medical care to complex and underserved populations. The New York-based startup has seen incredible growth since its inception in 2018, with over 55,000 downloads of the Galileo mobile app.

The Challenge

In addition to navigating the rapidly-changing technology landscape, healthcare startups must contend with a myriad of stringent regulations like HIPAA. Michael Supon, Galileo's Head of Security and Compliance, was no stranger to the challenge of maintaining HIPAA compliance across an entire organization. With years of experience in healthcare technology, Supon knew his team needed an automated solution to protect against potential data breaches.

How we help

Securing sensitive data in both Slack and GitHub was critical for Supon and his team. Within Slack, Supon needed a adaptable tool to help with data policy enforcement, something he found in Nightfall. Before acquiring Nightfall, Supon's team had to spend countless hours monitoring Galileo's repositories, but they also risked leaks of sensitive information in between pull requests. Now, in addition to improved productivity, the Galileo team also enjoys increased protection from the financial liability of a data breach.

"Nightfall's ease of setup and accuracy of identified data are both on point. Nightfall has eased our collective mind."

galileo

Case study: Calgary Public Library

Industry: Public Service (Library)
Employees: 400
Head of Security and Compliance: Michael Supon
Profile: Calgary Public Library is the second largest library system in Canada, serving over 700,000 members each year. During COVID-19, Calgary Public Library has become even more important, providing free access to books, media resources, and programs to its patrons via their website and allowing people to stay connected through the library's virtual portal.

Prioritizing protection for code repos

The IT Interfaces Group had to migrate their code repository from an on-premises Subversion server to GitHub without exposing the entire organization's sensitive data. "We needed to make sure we are not storing any secrets in the code. The only way to do this reliably was to proactively screen the code for potential secrets," says Anton.

How we help

With Nightfall automatically scanning for data that could leak, Calgary Public Library doesn't have to worry about exposing data that could compromise their systems. Our deep learning based detectors deliver higher accuracy & fewer false positives than traditional approaches. Overall, the Calgary Public Library IT team has seen a jump in productivity and confidence in their security with Nightfall. "Our programmers can sleep better at night. Now we can spend our time developing enterprise applications instead of custom solutions to lint our code for secrets," says Anton.

"Nightfall was the quickest and easiest way for us to guarantee we are not committing any passwords, API keys, or other sensitive information to our GitHub repositories."

CALGARY
PUBLIC
LIBRARY

Want to learn more about Nightfall?

To get started with Nightfall, request a demo or email us at **sales@nightfall.ai** with any questions.

Request a demo

About Nightfall

Nightfall is the industry's first cloud-native DLP platform that discovers, classifies, and protects data via machine learning. Nightfall is designed to work with popular SaaS applications like **Slack**, **Google Drive**, & **GitHub** as well as IaaS platforms like **AWS**.

Nightfall