MONITORING 25,000 BLACK BOX ENDPOINTS & PROVING THE SRE TEAM'S VALUE

Aaron Wieczorek | United States Digital Service

THE U.S. DIGITAL SERVICE

@_a12k

THE UNITED STATES DIGITAL SERVICE

- Founded in **2014** after Healthcare.gov fell over
- Teams at places like **VA, DHS, HHS, GSA**
- Implement **citizen facing services** for the federal government

OUR MISSION

- To deliver **better government** services to the people through **technology and design**
- **Engineers, Designers, PMs**
- Find and solve **big problems**

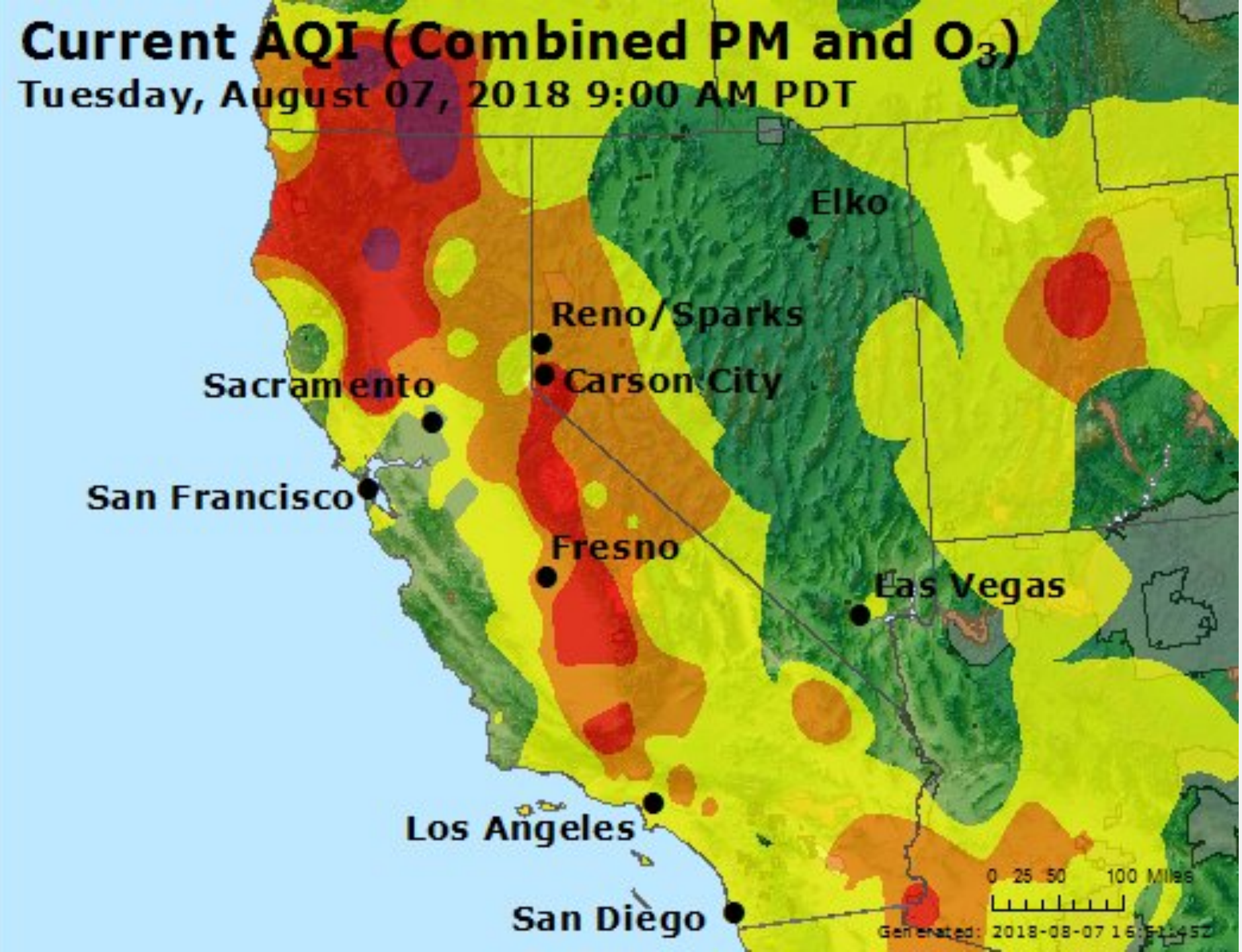
SERVICE DISCOVERY IN THE .GOV SPACE

- **How** do we find the biggest problems?
- We're a **small agency**, people might not know we exist
- USDS is **mobile** – Can be on site tomorrow, for free
- Go **where the work is**
- Best solution so far:

  Step in when there is a crisis, when we know there is a crisis…

EPA'S AIRNOW.GOV

- **2018 California wildfires** – Deadliest and most destructive
- People looking at **particulate matter, ozone** at Airnow.gov
- Influx during the wildfires meant the **service fell over** under the load



AIRNOW.GOV:
2018

@_AI2K

AIRNOW.GOV: *2018*

Resource Not Found

Contact Us Share

You are here: EPA Home » Error Page

Resource Not Found (404)

The requested resource was not found on the EPA's Web Server. Please return to **the previous page** and use the comment link there to report this broken link. If you do not see a link above or your browser does not support the above link, use the comments page to describe your problem to the EPA's Internet Support. When contacting us, please include the following information:

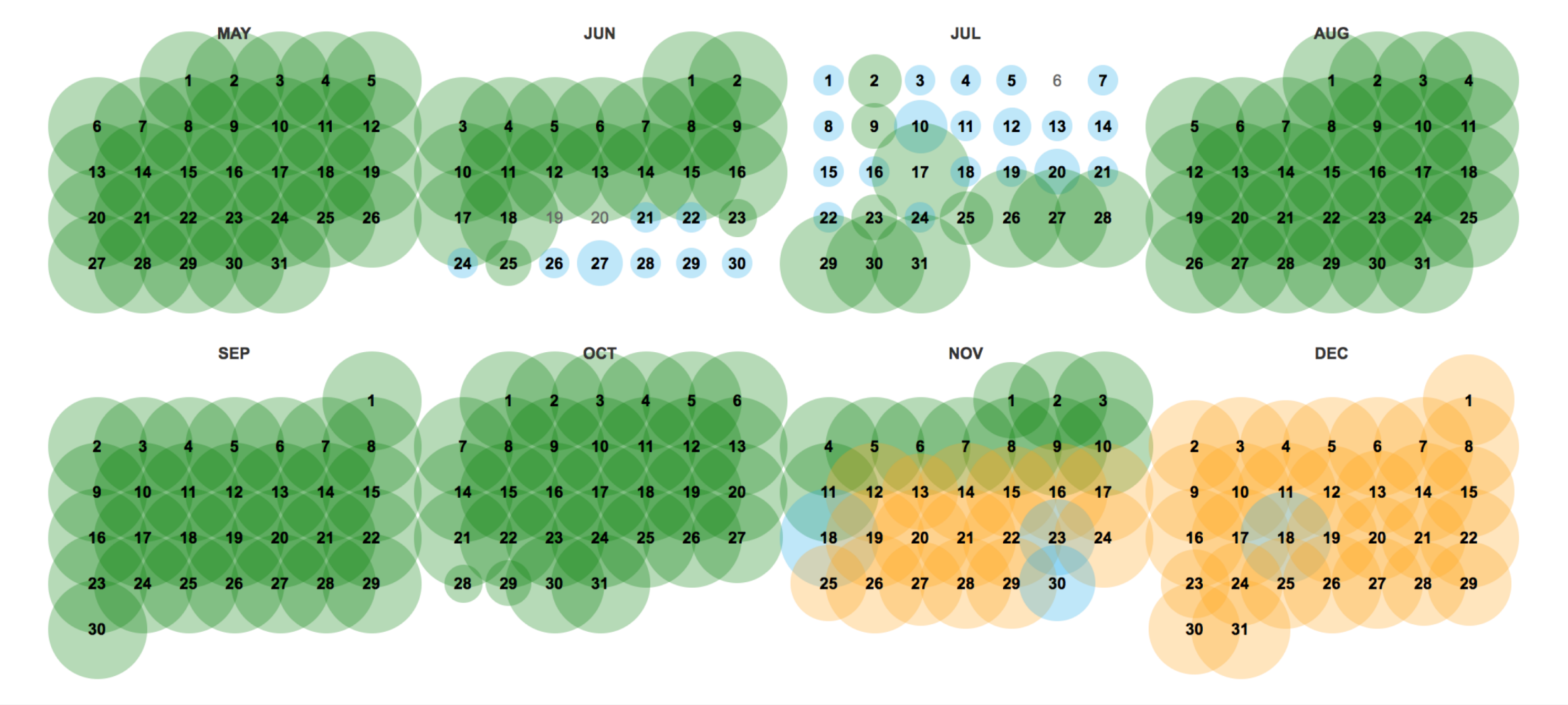
- **the Internet address of the missing file (ex. https://web.archive.org/web/20181127054238/http://airnow.gov/)**

Also include the URL (address) of the page from which you are linking. Please confirm the URL (address) of the page you are trying to reach. This information will aid in expediting your request.

Top of Page

@_AI2K

AIRNOW.GOV - 2018



AIRNOW.GOV: 2018

- One of my colleagues happened to be checking Airnow.gov so they could get a status on the area

USPTO OUTAGE

- **U.S. Patent and Trademark Office** site down August 2018
- August 15-23 – **9 days**
- Estimated cost **$4m/hr** of service disruption
- **~$864,000,000 lost**

USPTO OUTAGE

HOW DO WE GET AHEAD OF THE PROBLEM?

- Idea to **proactively implement monitoring** for every .GOV service so we can step in if a critical service goes down
- **What targets** do we monitor?
- What do we **monitor for**? Uptime? Latency?

LOTS OF SERVICES IN .GOV SPACE

- **How** do you find all these services?

  Office of the Chief Information Officer maintains public records

- **26,049 .GOV services** in the federal government + .mil endpoints

- **2,972 domains** with higher traffic

  Some deprecated, some really small, some larger and more critical

CUSTOM SOLUTION AS MVP

- **Scripts** that send requests
- **Python Requests + CLI**
- Problems:

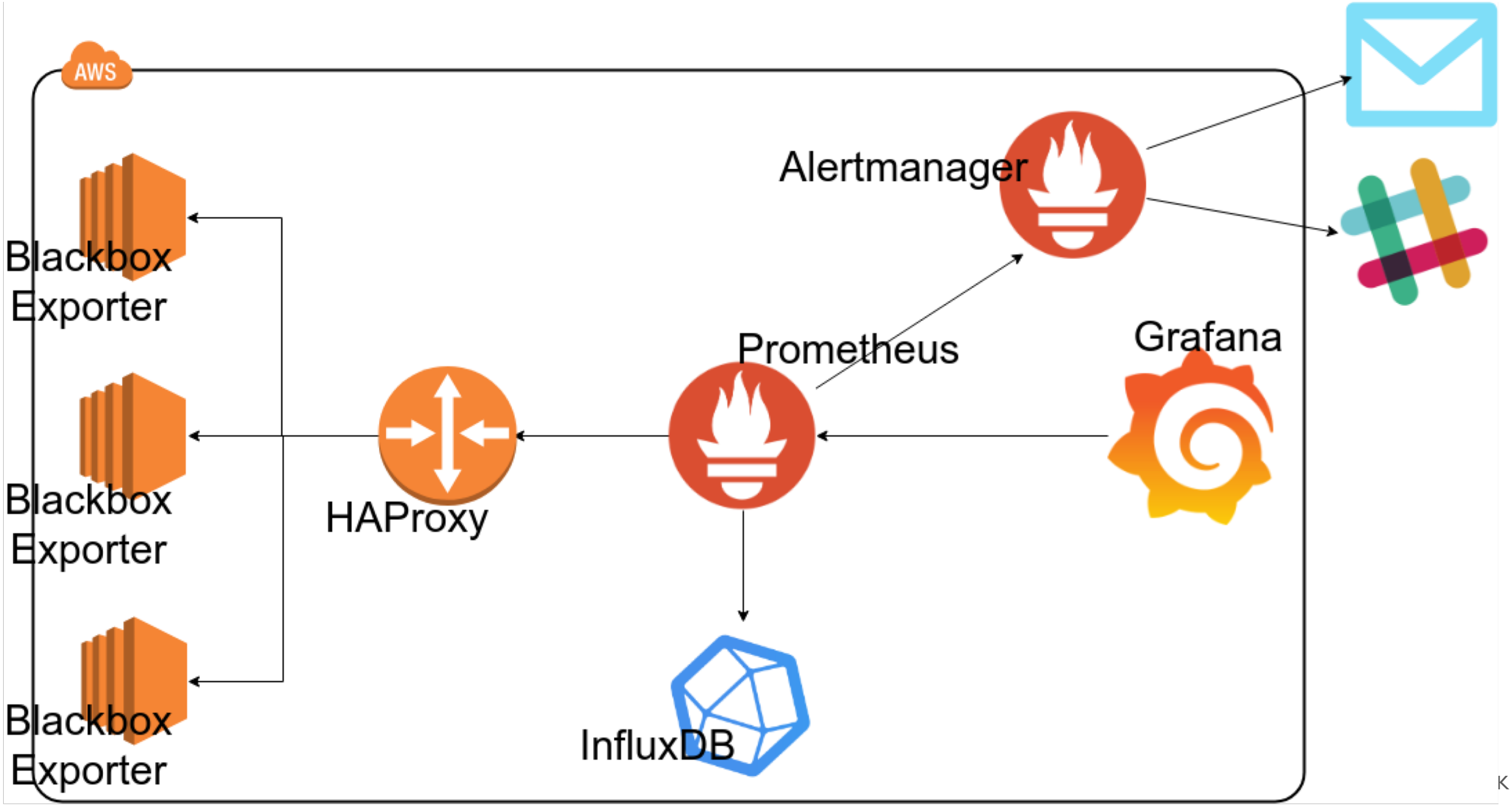
  Building something that was already OSS

  Data wasn't robust enough without lots of additional work

EXPANDING TO A MORE ROBUST ARCHITECTURE

Grafana

influxdb


AWS
Blackbox Exporter
Blackbox Exporter
Blackbox Exporter
HAProxy
Prometheus
Alertmanager
Grafana
InfluxDB

WHAT TO MONITOR

- http 2xx responses
- Performance vs availability

Seek out particular targets to pitch

TUNING PROMETHEUS AND BLACKBOX

- Scrape intervals & timeouts
- Relabel configs vs Proxy connections
- Number of Blackbox instances
- Curating 25,000+ endpoints == hardest part

Some 2xx, some 4xx, some 5xx, 8,000 0s

Prometheus Alerts Graph Status Help

Targets

All Unhealthy

blackbox-exporter (25038/25038 up) show less

Endpoint

TUNING GRAFANA

- Dashboarding with Singlestats
- Careful with Queries
- Segregating information into groups

Amazon EC2 Abuse <ec2-abuse@amazon.com>

Your Amazon EC2 Abuse Report

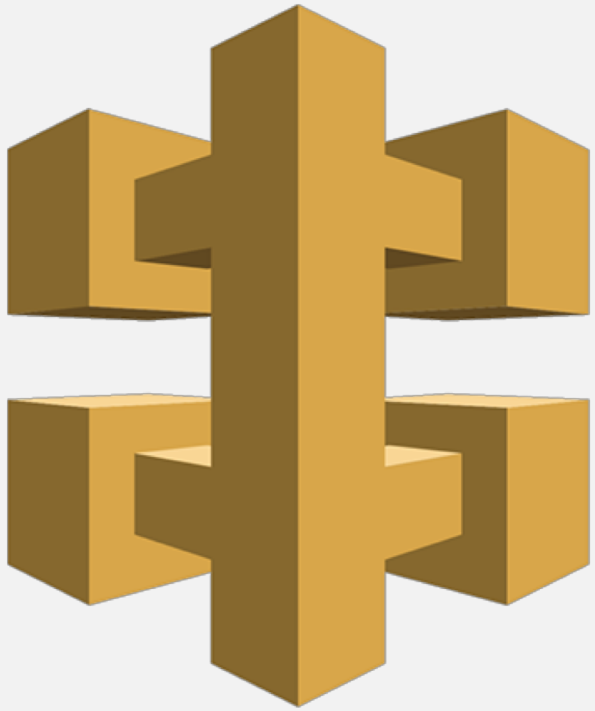
Abuse Type: Web Crawl

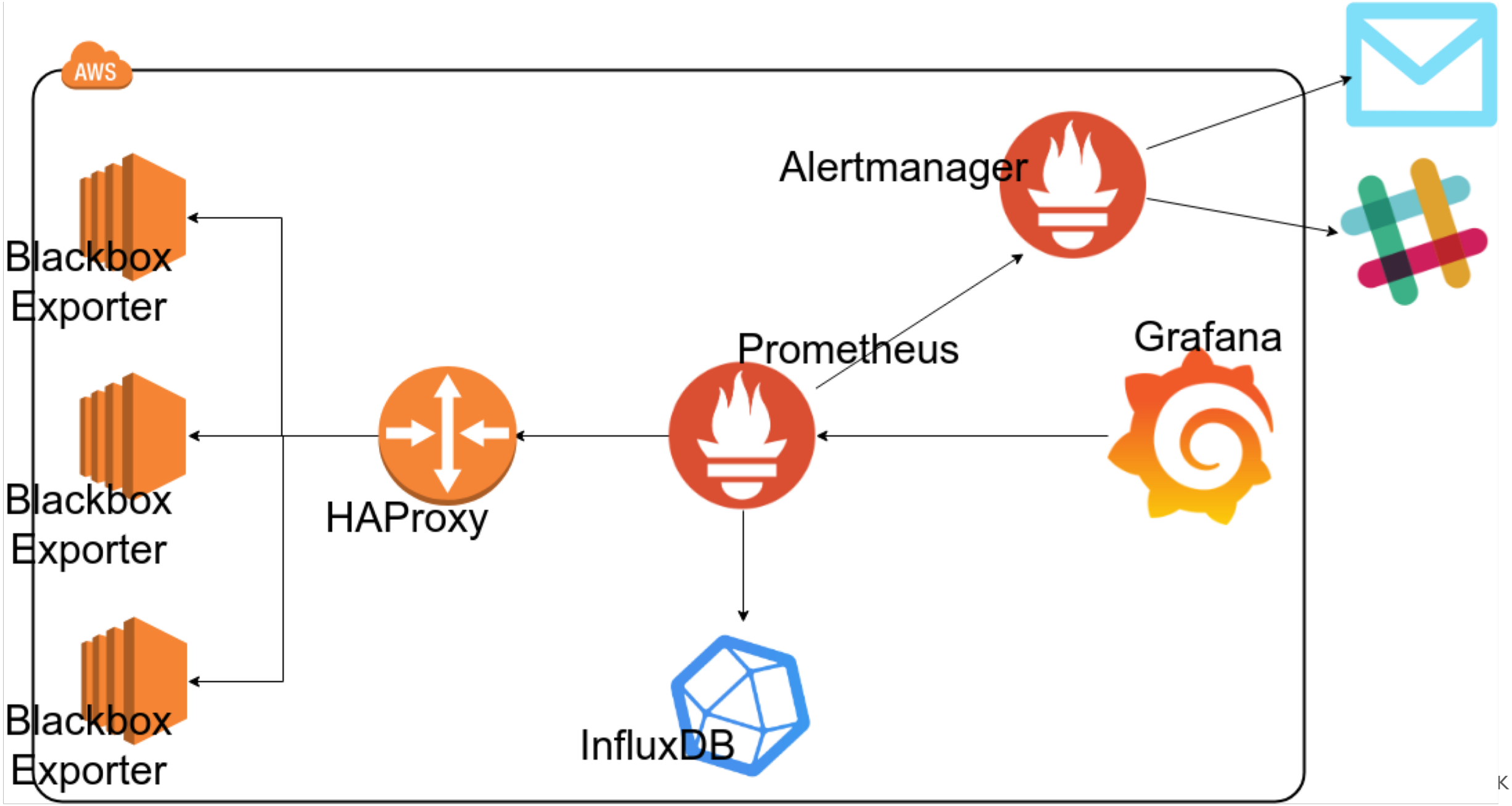
MORE TUNING, NEW ARCHITECTURE

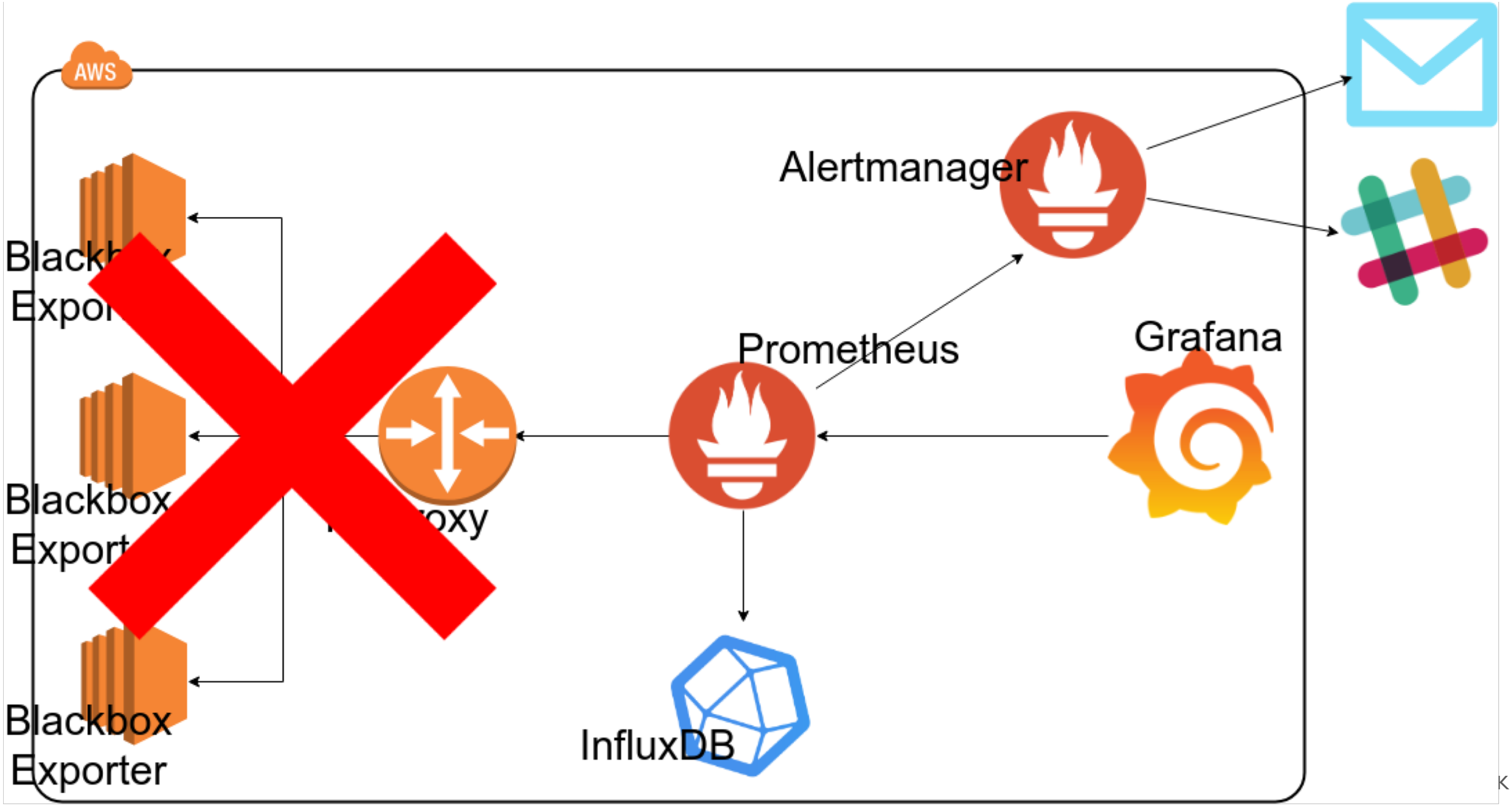
- **Adjust number of instances** of Blackbox Exporters to 5
- **HAProxy** service discovery made this easy
- **Adjust scrape interval** -- 15 minutes/900 seconds
- **Costs are rising** - ~$700/mo for Blackbox + HAProxy

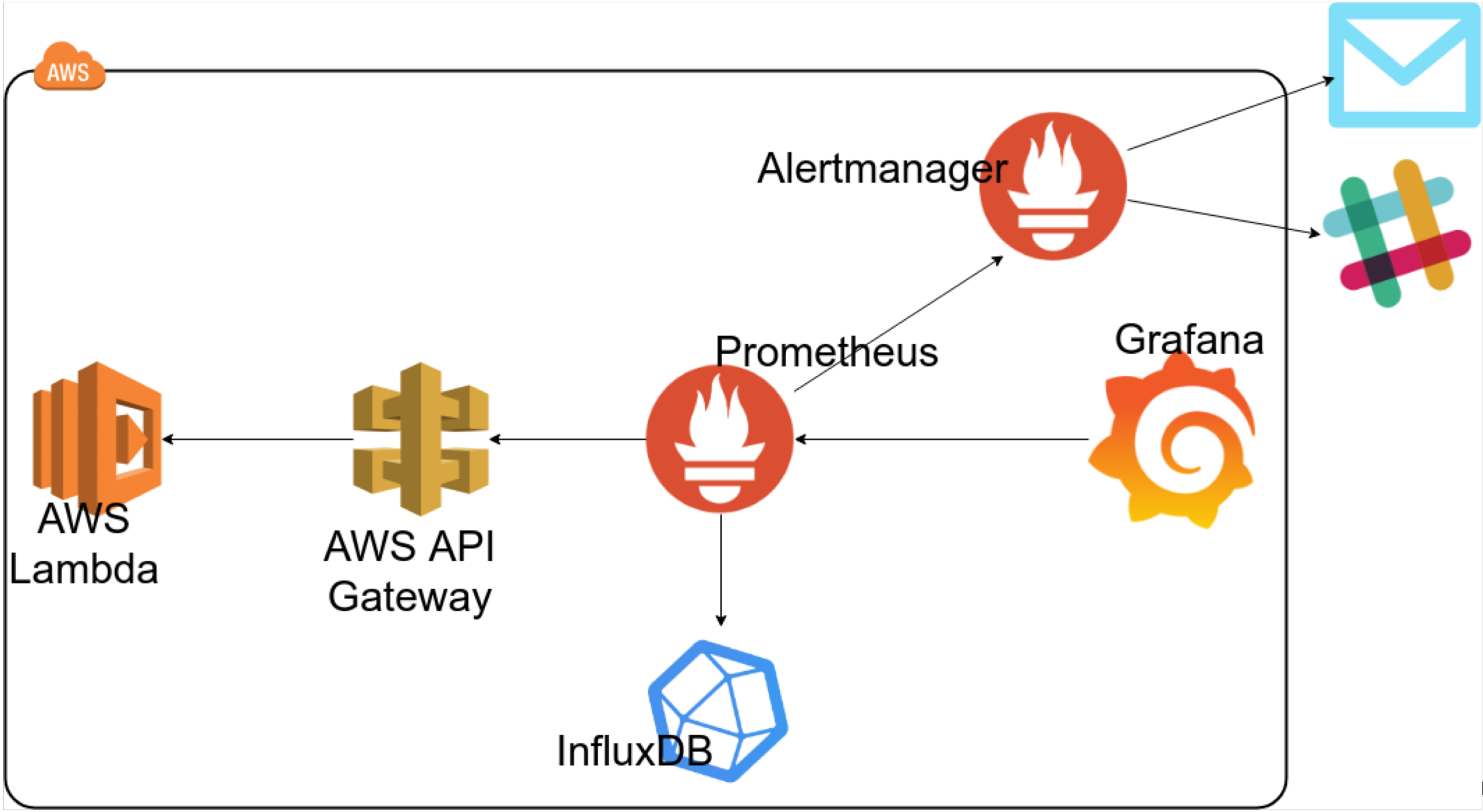
m5.xlarge	4	16	16 GiB	EBS Only	$0.192 per Hour

MORE TUNING, NEW ARCHITECTURE




AWS
Blackbox Exporter
Blackbox Exporter
Blackbox Exporter
HAProxy
Prometheus
Alertmanager
Grafana
InfluxDB


AWS
Alertmanager
Prometheus
Grafana
Blackbox Exporter
Blackbox Exporter
Blackbox Exporter
InfluxDB


AWS
Alertmanager
Prometheus
Grafana
AWS
Lambda
AWS API
Gateway
InfluxDB

WHAT'S NEXT?

- Writing our **own exporters**
- Want to make sure service working **beyond first page**
- **No Health endpoint** we can hit
- Targeted approach:

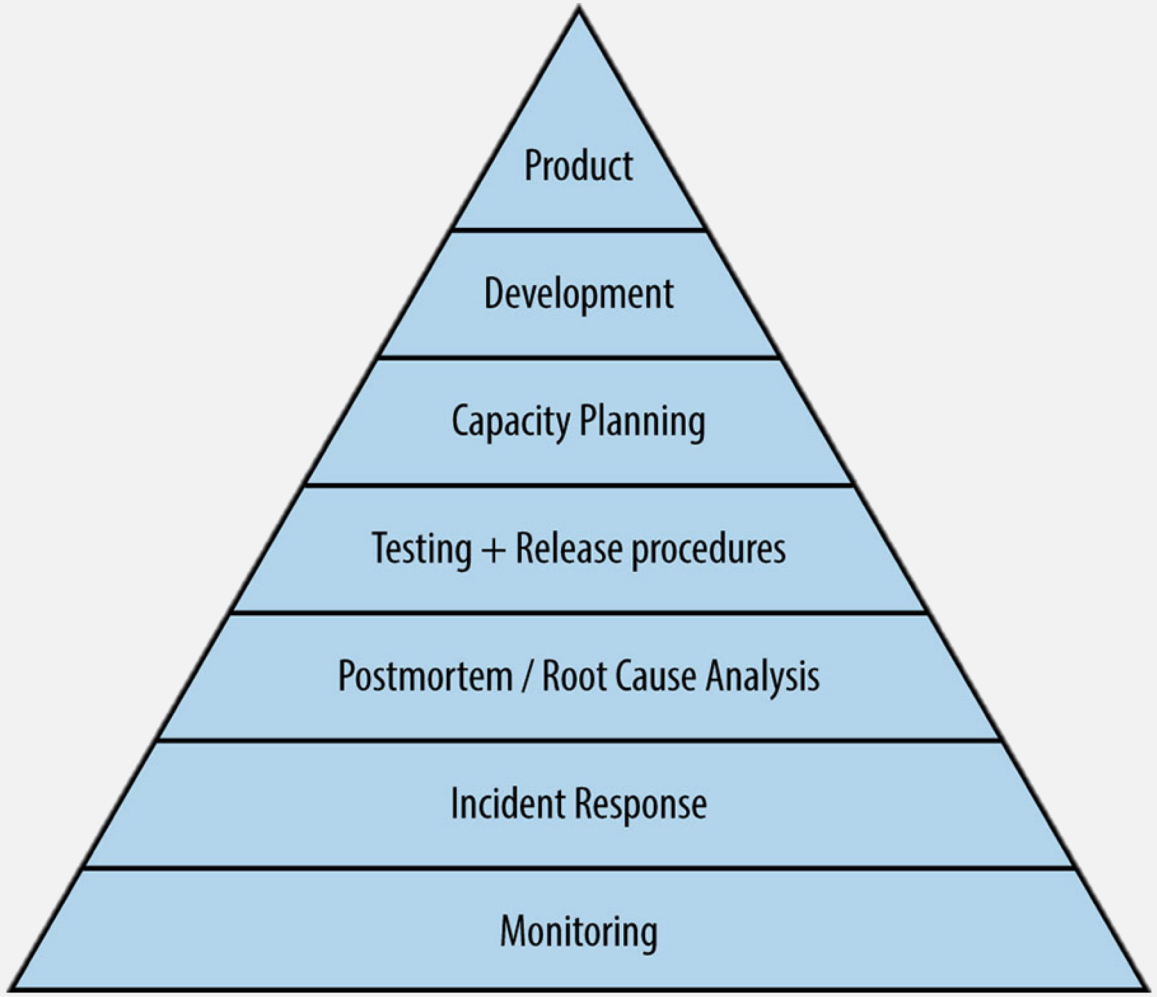
  Build, for example, login flows for sites that may return 2xx but 5xx after login

  Need permission

WHAT'S THE RESULT

- Have data to **show uptime** for services
- Have data to **act quickly** when service goes down
- Have data to show if service is likely to **go down, or high latency, low uptime**
- Take this data **to other orgs** to show them SRE concepts, prove the value of your team, talk monitoring
- **Opens the door** for PMs, Designers to engage as well

WHAT'S THE RESULT



LESSONS LEARNED

- **Proactive monitoring** allows immediate incident response
- **Fix the problem**, shift gears to product teams for long term improvement, build on SRE hierarchy
- **Training teams** that you work with on
- **Prove** the SRE team's value, SRE hierarchy value

LESSONS LEARNED

- Sometimes targets **don't like it** when you are sending lots of requests in rapid succession
- Helped prompt **move to lambda**
- Lambda spreads out IPs, not tied to specific instances
- Saves money, **10% of the cost**

LESSONS LEARNED

- **Dashboards** with this many endpoints are hard
- How do you reason about what's going on

  High level metrics

  Drill down for all services at one agency

  Is everything at one agency down, just one service?

- How do you **not kill your browser** on heavy loads

LESSONS LEARNED

- **Alerting is hard**
- Do we need to be **on call** for someone else's system?
- **Sensible defaults**, bubble up when a service goes down, triage whether its worth stepping in
- **Alert** for large scale outages, or outages on particular systems

LESSONS LEARNED

- **Tuning monitoring settings** for large variety/number of systems is extremely difficult
- Let's **be nice** to the services we're trying to help

  HEAD vs GET requests

  Longer scrape intervals

P.S. WE'RE HIRING!

USDS.GOV/APPLY

THE U.S. DIGITAL SERVICE

Aaron Wieczorek

@_a12k