

ASAPQuery: 100x Drop-in Acceleration for Metrics Observability

Milind Srivastava

PhD student @ Carnegie Mellon University



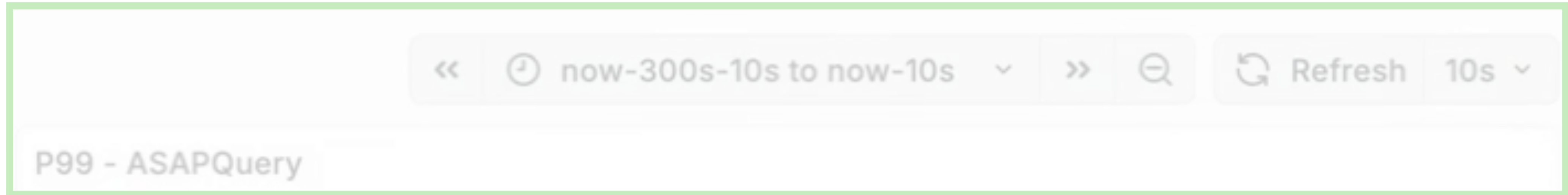
github.com/ProjectASAP/ASAPQuery

May 22, 2026

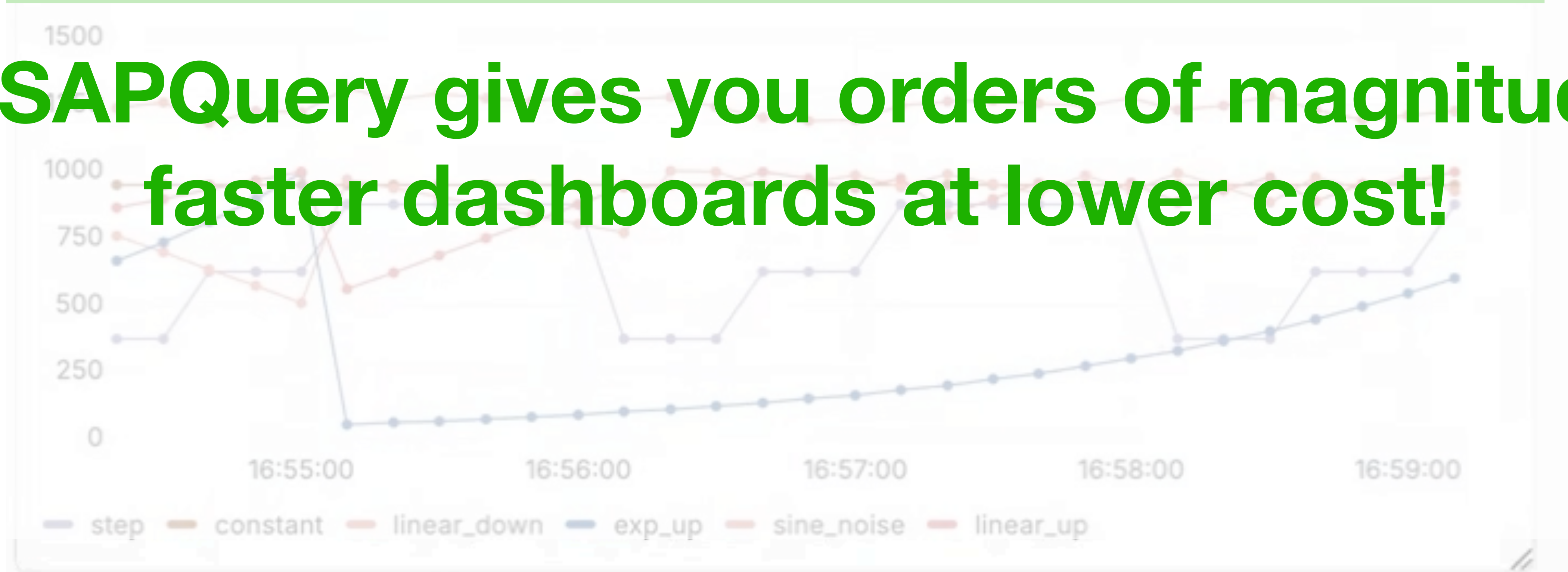
Slow Grafana dashboards – looks familiar?



What if we could have this?



ASAPQuery gives you orders of magnitude faster dashboards at lower cost!



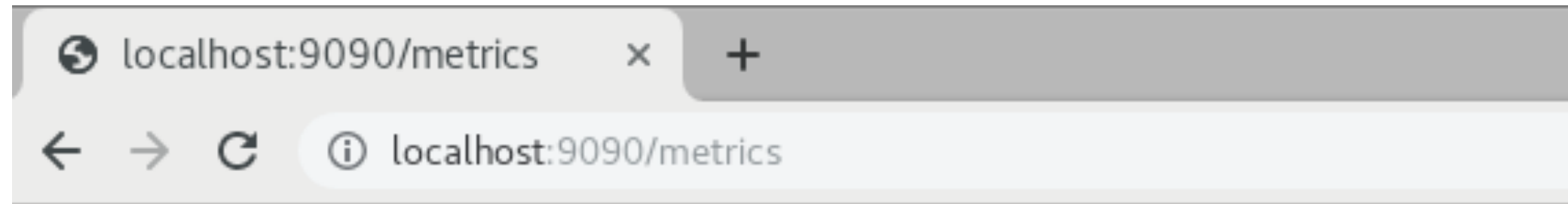
Who am I



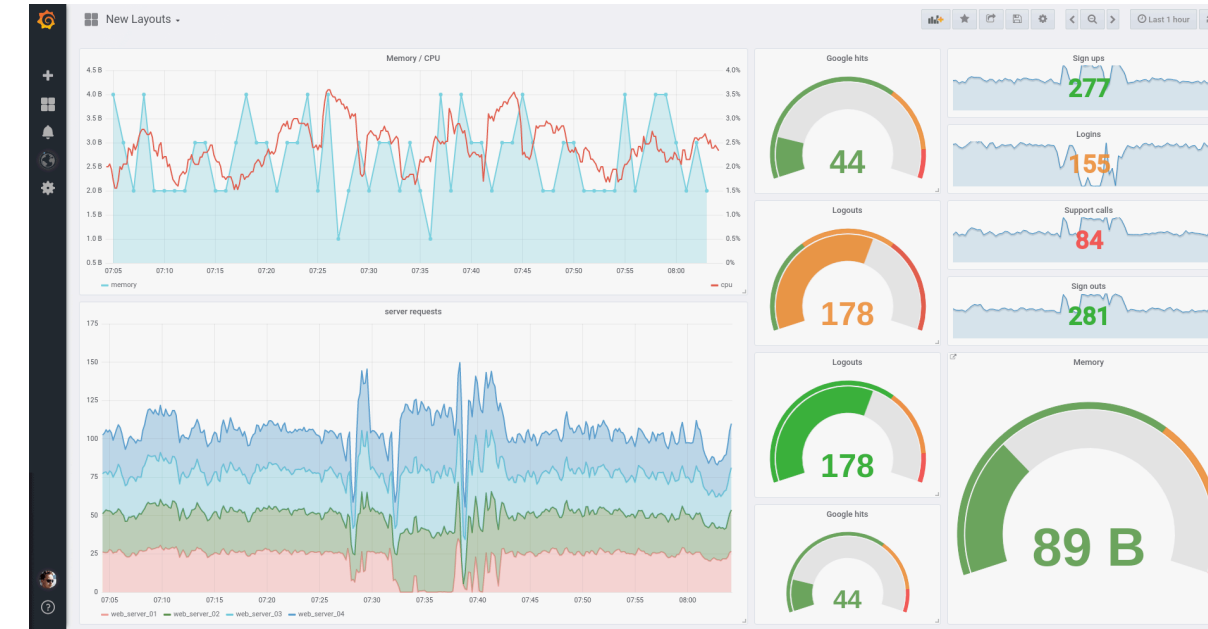
Milind Srivastava

- PhD student @ Carnegie Mellon University
- Lead for ProjectASAP: projectasap.github.io
- We build systems to make observability cheaper and faster
- Our research democratizes the benefit of cool algorithms for practitioners like you

ASAPQuery: A drop-in accelerator for Prometheus-Grafana



```
# HELP go_gc_duration_seconds A summary of the GC invocation durations.  
# TYPE go_gc_duration_seconds summary  
go_gc_duration_seconds{quantile="0"} 1.097e-05  
go_gc_duration_seconds{quantile="0.25"} 1.8263e-05  
go_gc_duration_seconds{quantile="0.5"} 3.7672e-05
```



exporter-1

exporter-2

exporter-3

Pull metrics

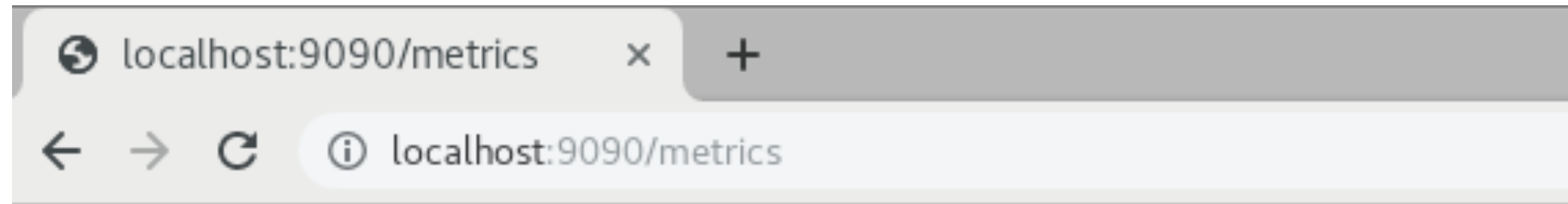


Prometheus

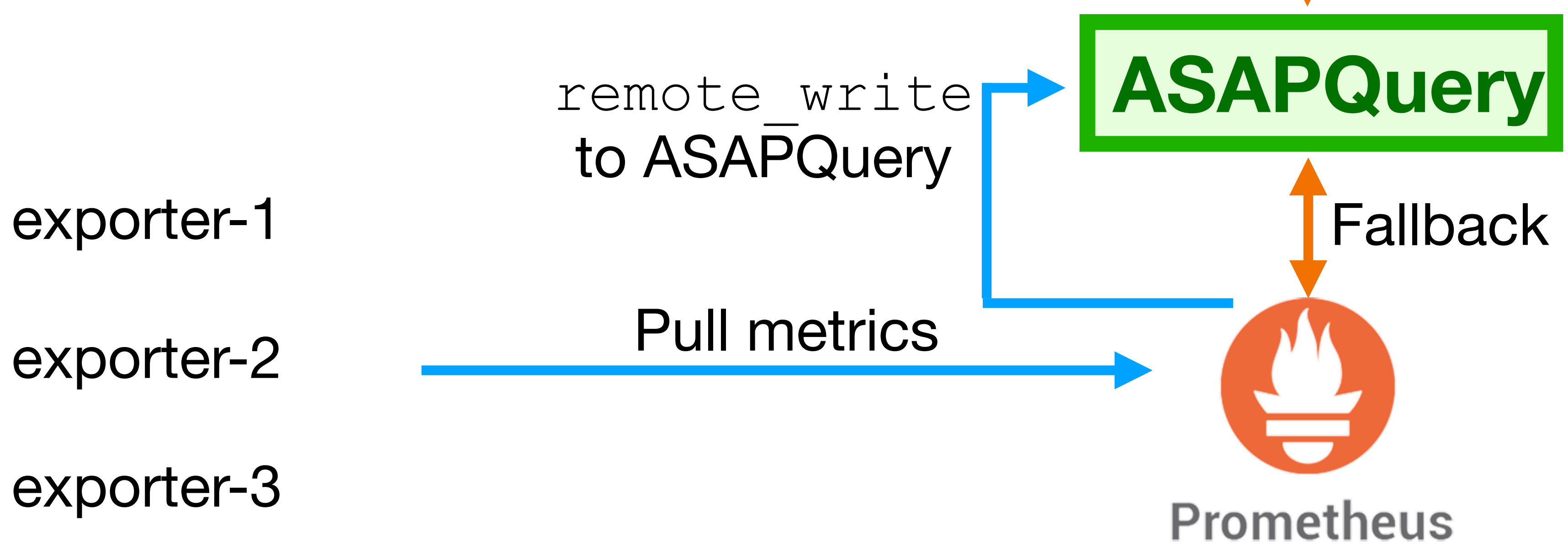
Query



ASAPQuery: A drop-in accelerator for Prometheus-Grafana

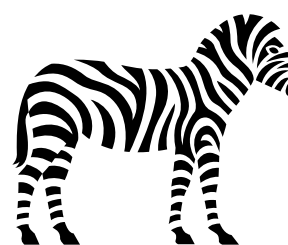
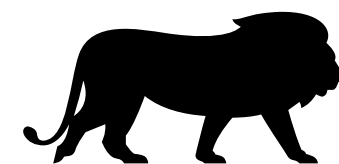
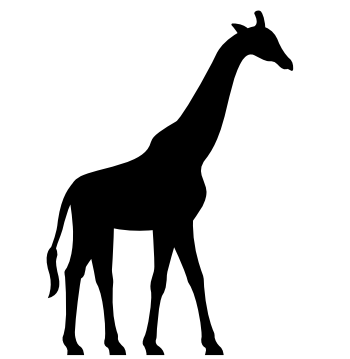
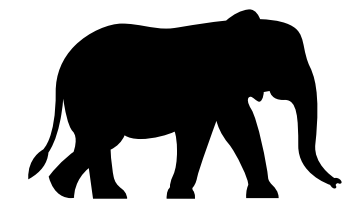
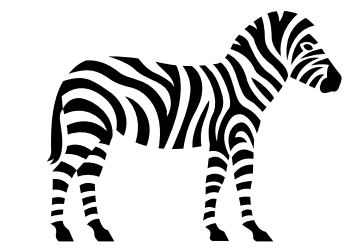
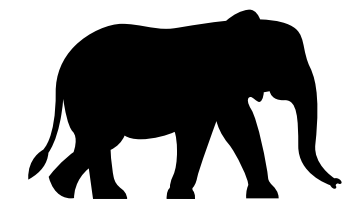


```
# HELP go_gc_duration_seconds A summary of the GC invocation durations.  
# TYPE go_gc_duration_seconds summary  
go_gc_duration_seconds{quantile="0"} 1.097e-05  
go_gc_duration_seconds{quantile="0.25"} 1.8263e-05  
go_gc_duration_seconds{quantile="0.5"} 3.7672e-05
```

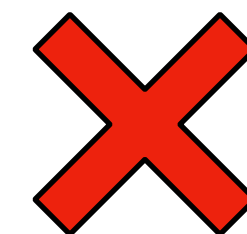
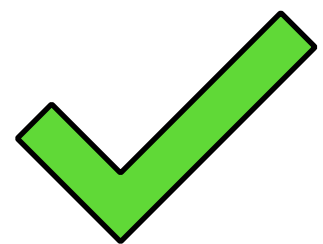
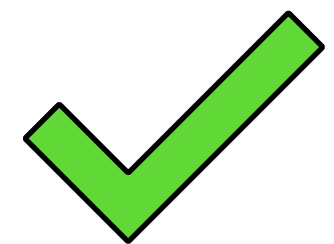
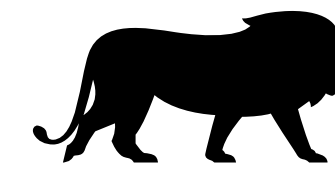
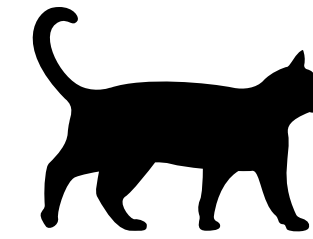
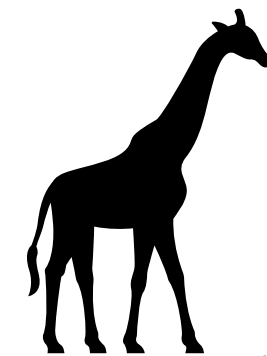
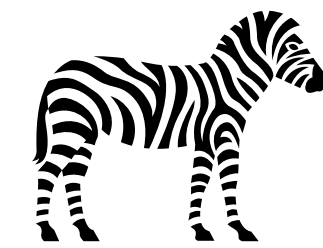
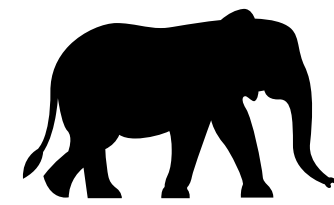


The secret sauce: Sketches

Query: Did you see a giraffe?

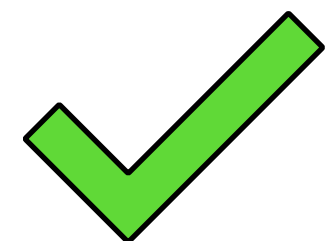
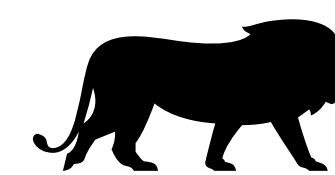
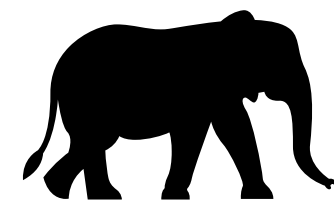


exact



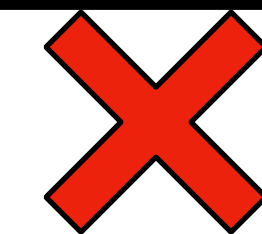
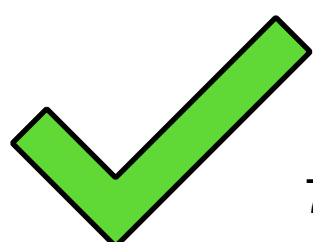
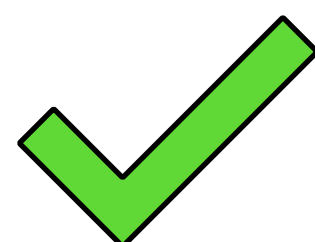
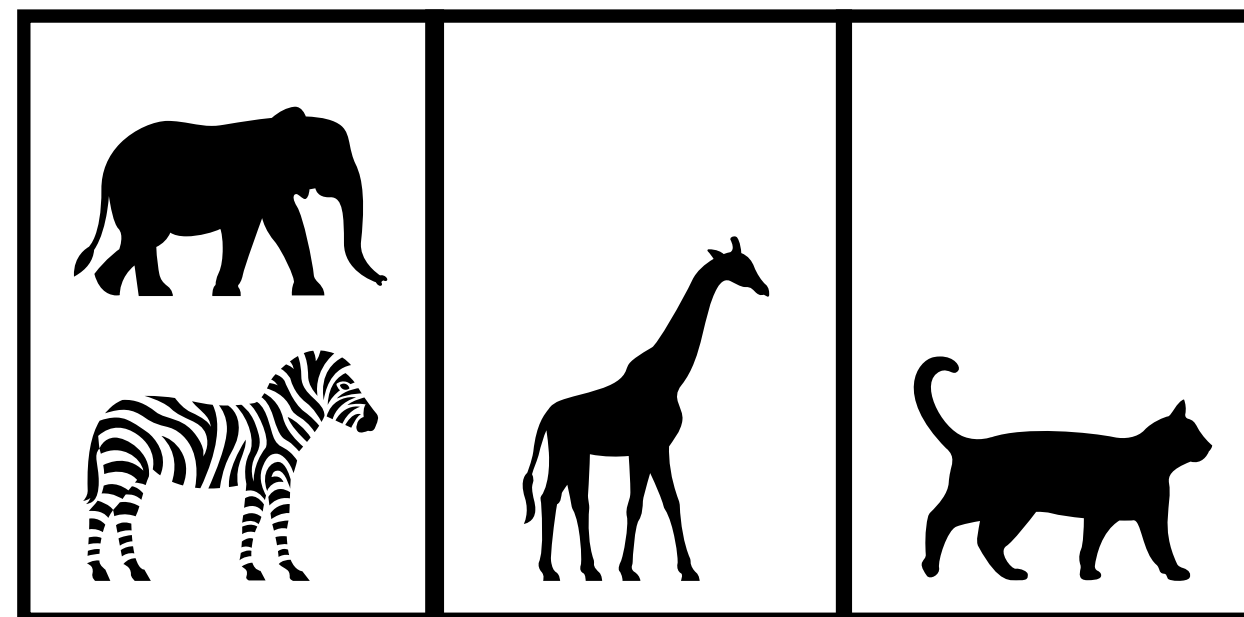
\$\$\$

sample

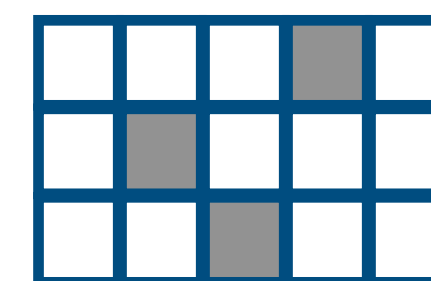


inaccurate

hash()

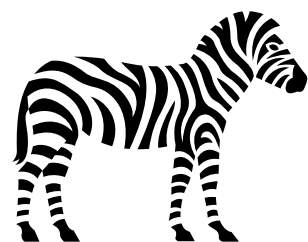
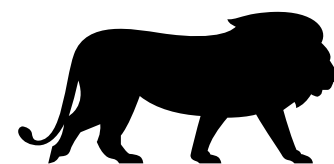
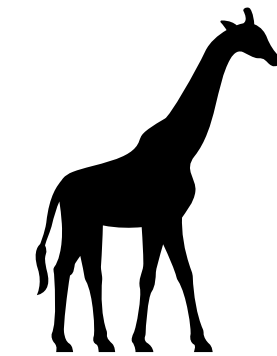
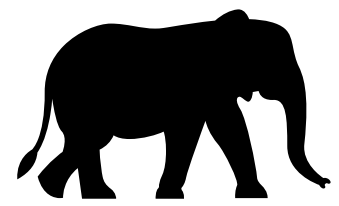
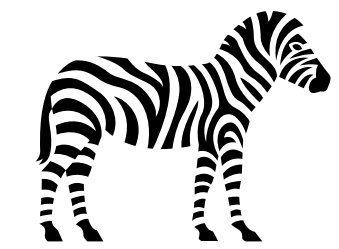
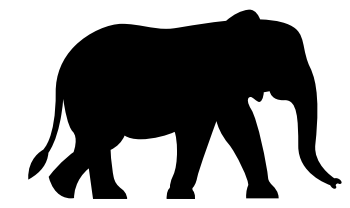


“sketch”

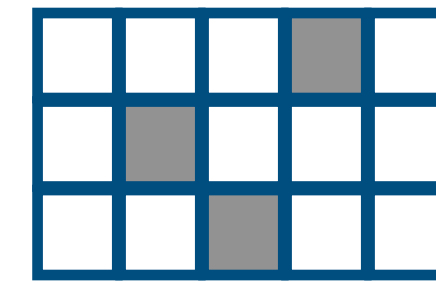
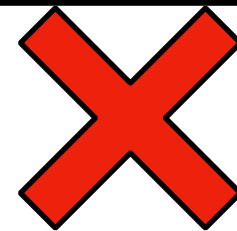
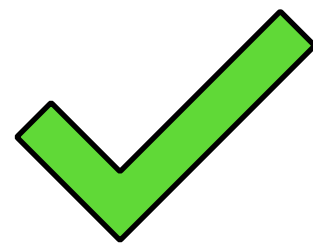
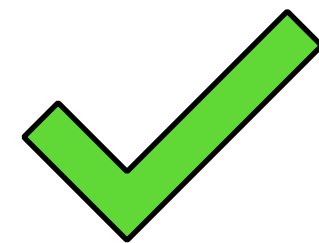
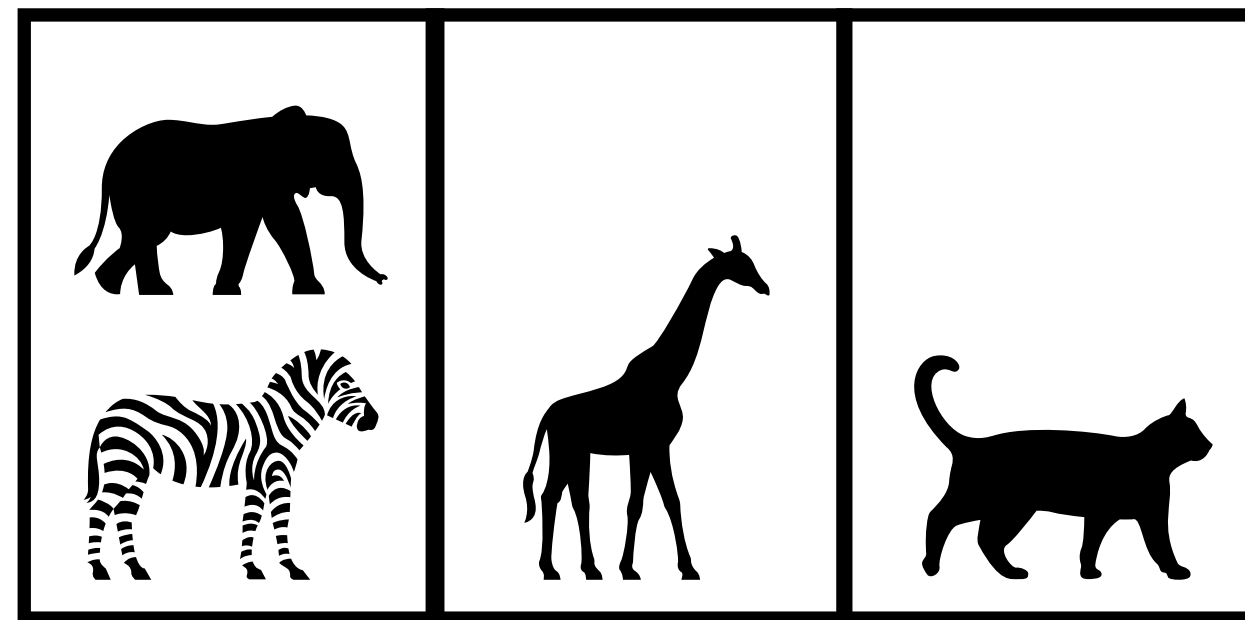


\$, high accuracy

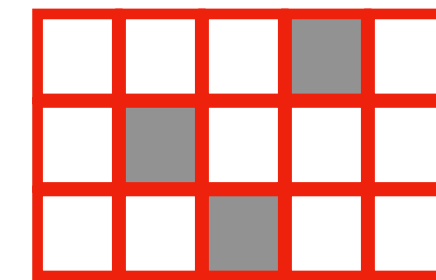
Sketches can answer many queries



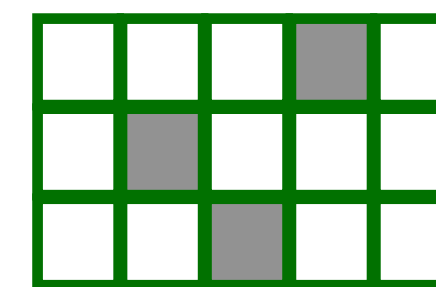
Bloom filter sketch



CountMin Sketch



CountMin Sketch + heap



Queries on animals

Did I see a giraffe?

How many times did I see a lion?

What are the top 5 animals I saw?

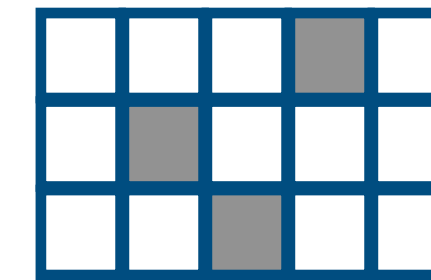
From animals to metrics

Metric: CPU usage

h1, login, 5%



Bloom filter Sketch



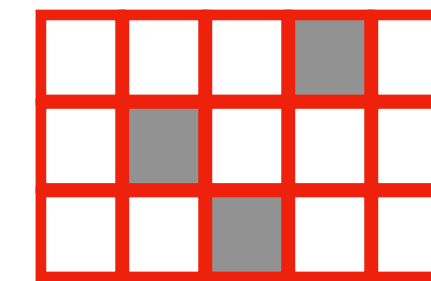
Is host X up?

h1, web, 20%

h2, proxy, 30%



CountMin Sketch

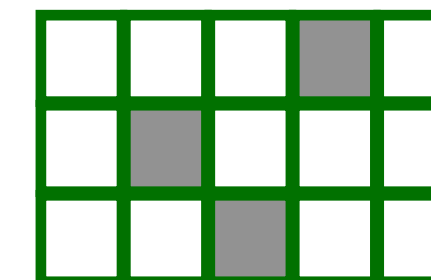


Avg CPU usage
across services

h3, load-
balancer, 50%



CountMin Sketch + heap



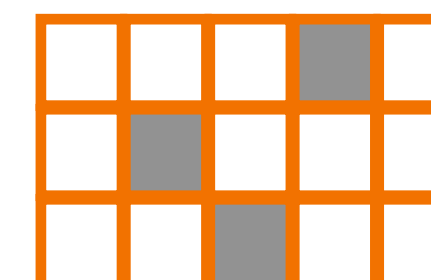
Top 5 hosts with
highest CPU

h1, web, 30%

...



KLL Sketch



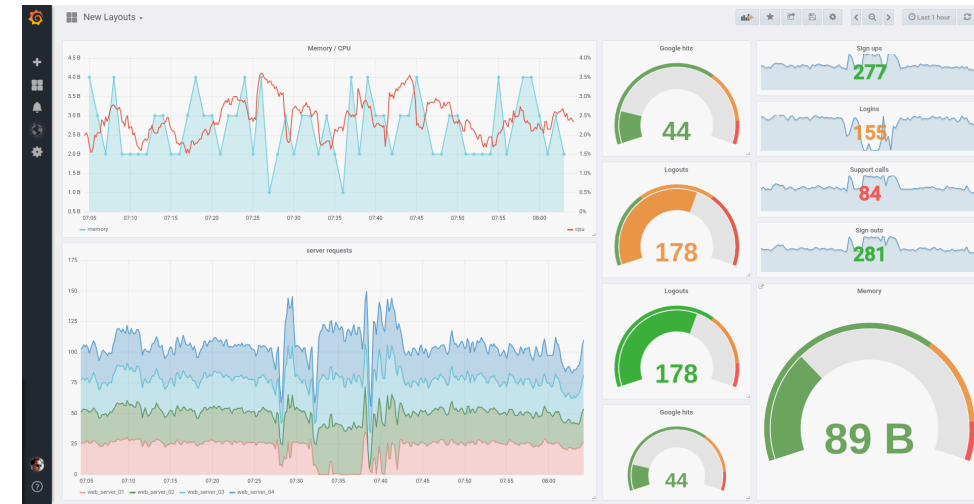
p90 CPU usage
across hosts

...

...

ASAPQuery makes sketches easy to use for observability

Metrics observability



Grafana



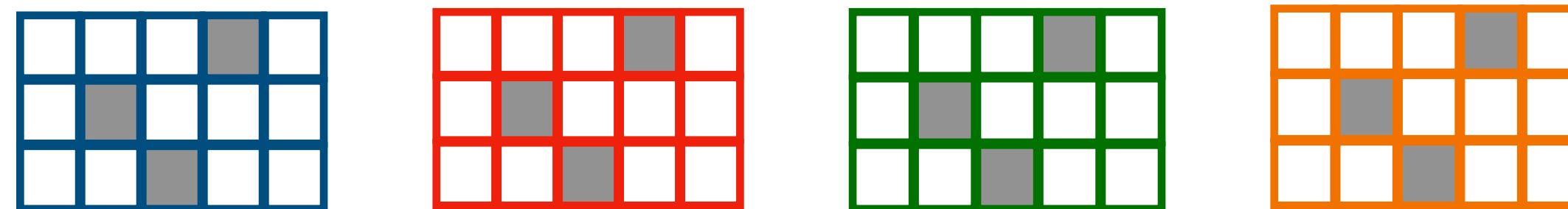
Prometheus

ASAPQuery

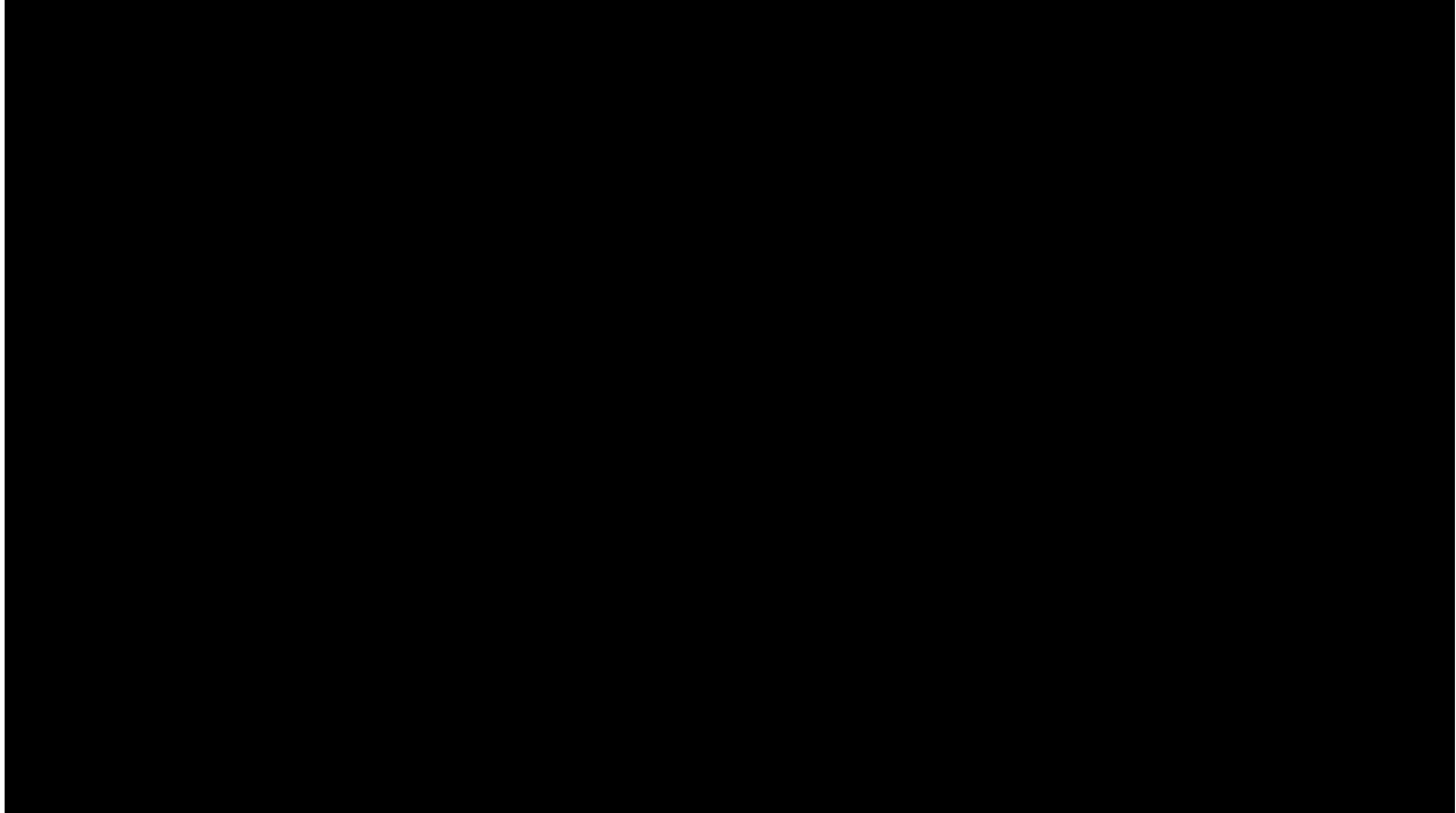
Faster dashboards

Lower CPU & memory
==
Lower \$\$\$

Sketches



Live demo of ASAPQuery



Try out ASAPQuery and share your thoughts!

If this is useful, I would appreciate a star on the repo :)



[github.com/ProjectASAP/
ASAPQuery](https://github.com/ProjectASAP/ASAPQuery)

- ASAPQuery modes (instructions in Github repo)
 - quickstart: 5 min to see benefits
 - drop-in: works with your Prometheus-Grafana stack
- Upcoming blog posts:
 - Library of sketches used in ASAPQuery
 - Architecture of ASAPQuery
- Email me for help or feedback: milindsr@andrew.cmu.edu



projectasap.github.io