weaveworks

Cortex: Infinitely Scalable Prometheus

December 2018

What is Cortex ?

Cortex is a time-series store built on Prometheus

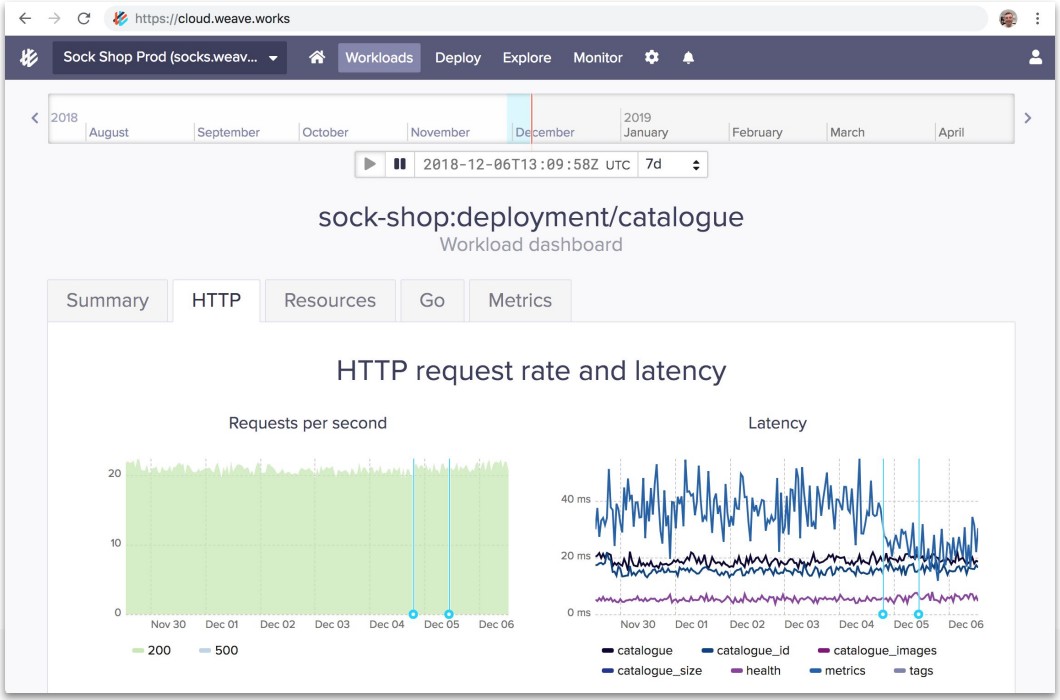
- Horizontally scalable
- Highly Available
- Long-term storage
- Multi-tenant

Cortex is a CNCF incubator project

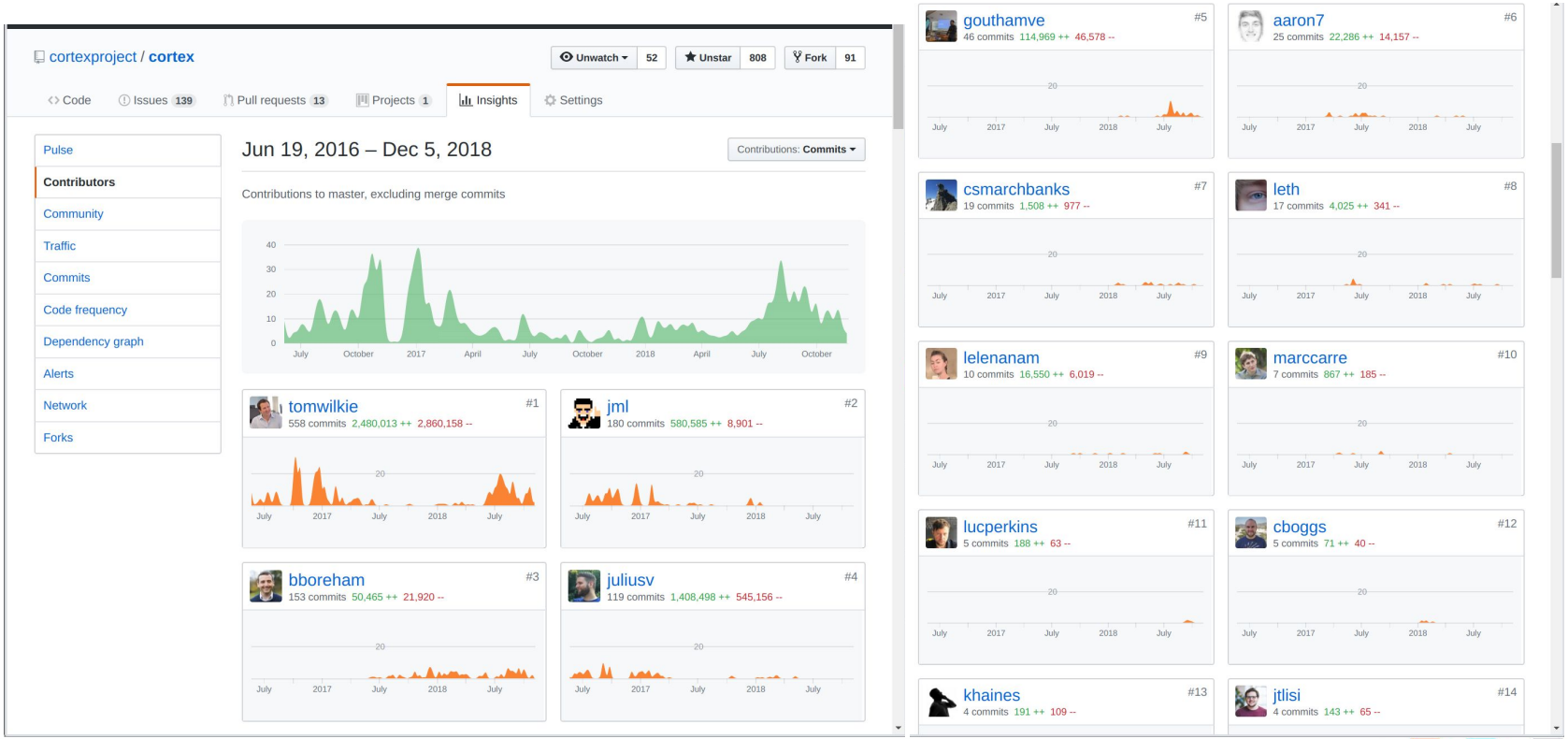
https://github.com/cortexproject/cortex

Why did we build Cortex

Prometheus As A Service on cloud.weave.works



Who wrote Cortex?



Who uses Cortex?

weaveworks

adzuna

Grafana Labs

ELECTRONIC ARTS™

Qordoba

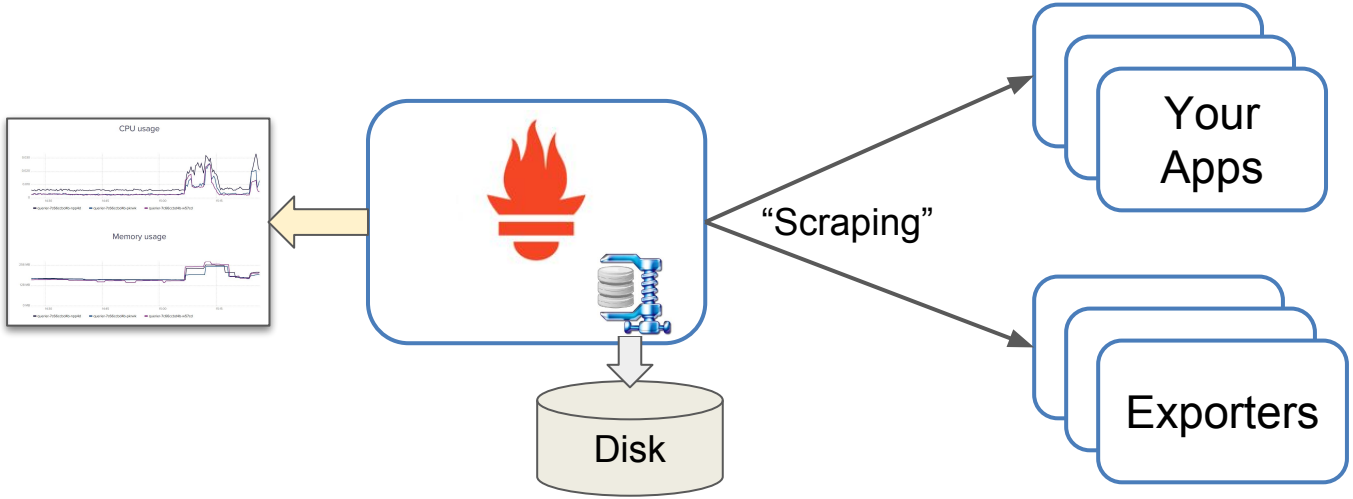
ASPEN MESH

FreshTracks.io

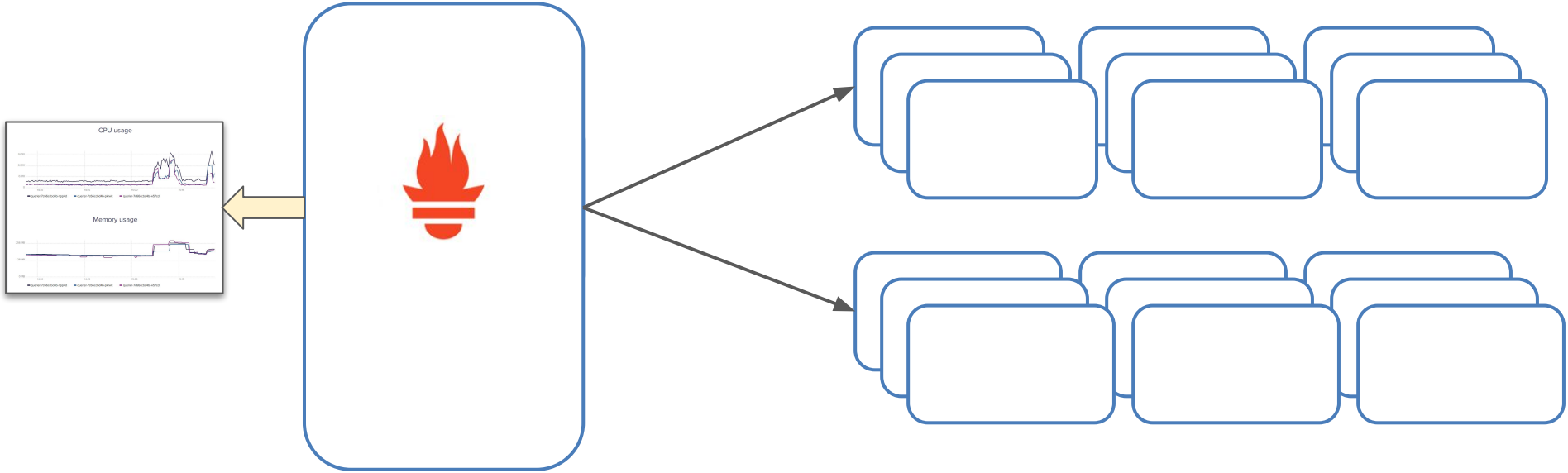
ADORE ME

>2 million samples/s
>100 million timeseries

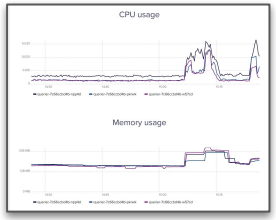
Basic Prometheus operation

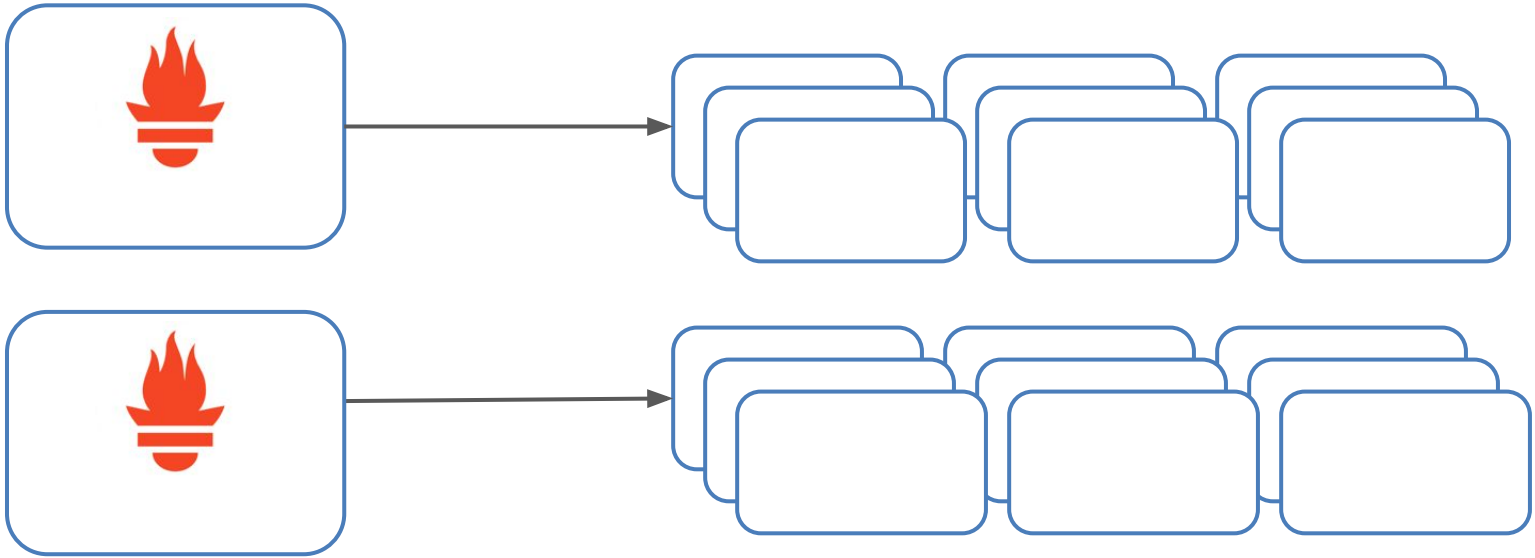


Scaling Prometheus

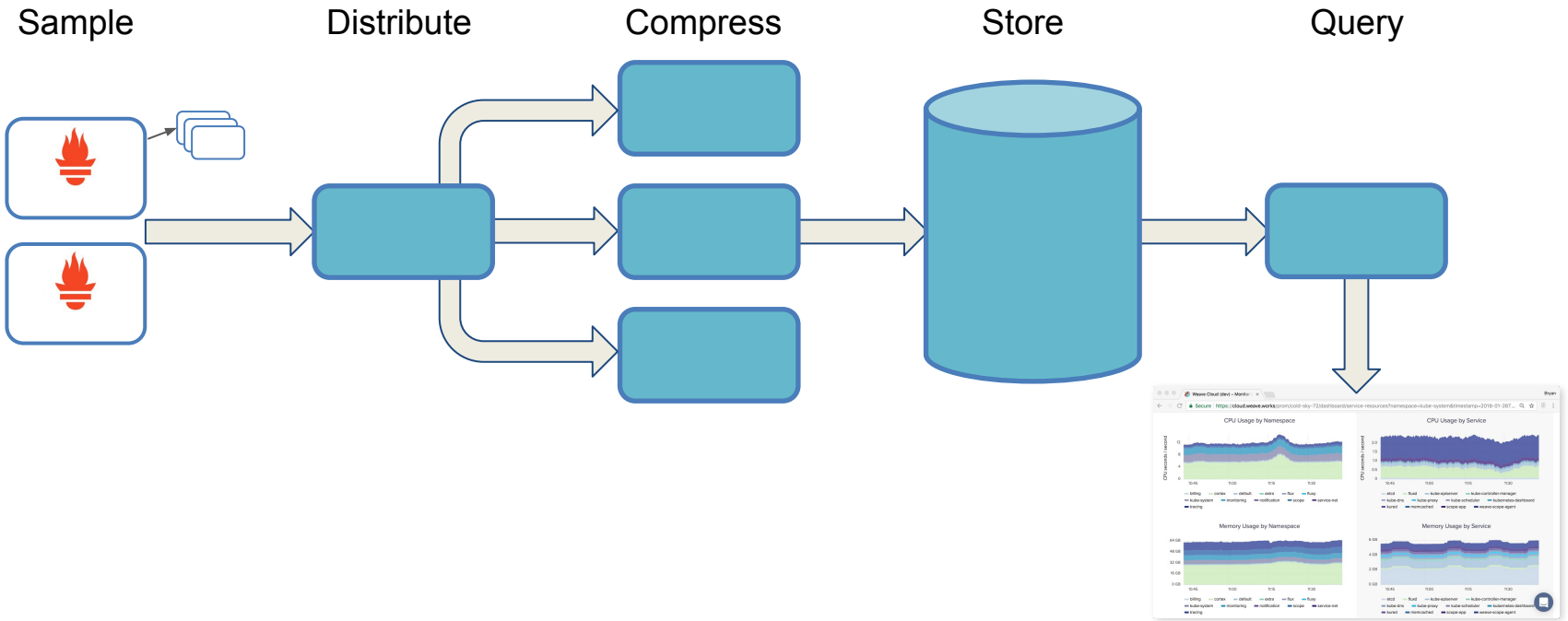


Sharding Prometheus

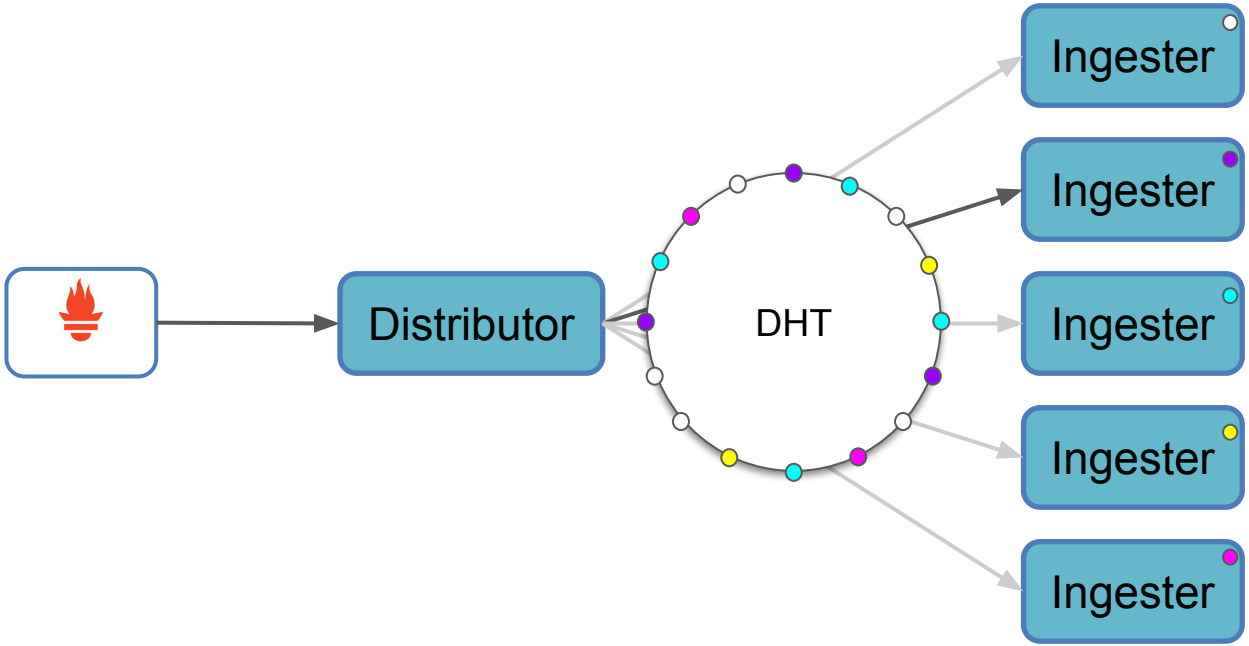




Cortex

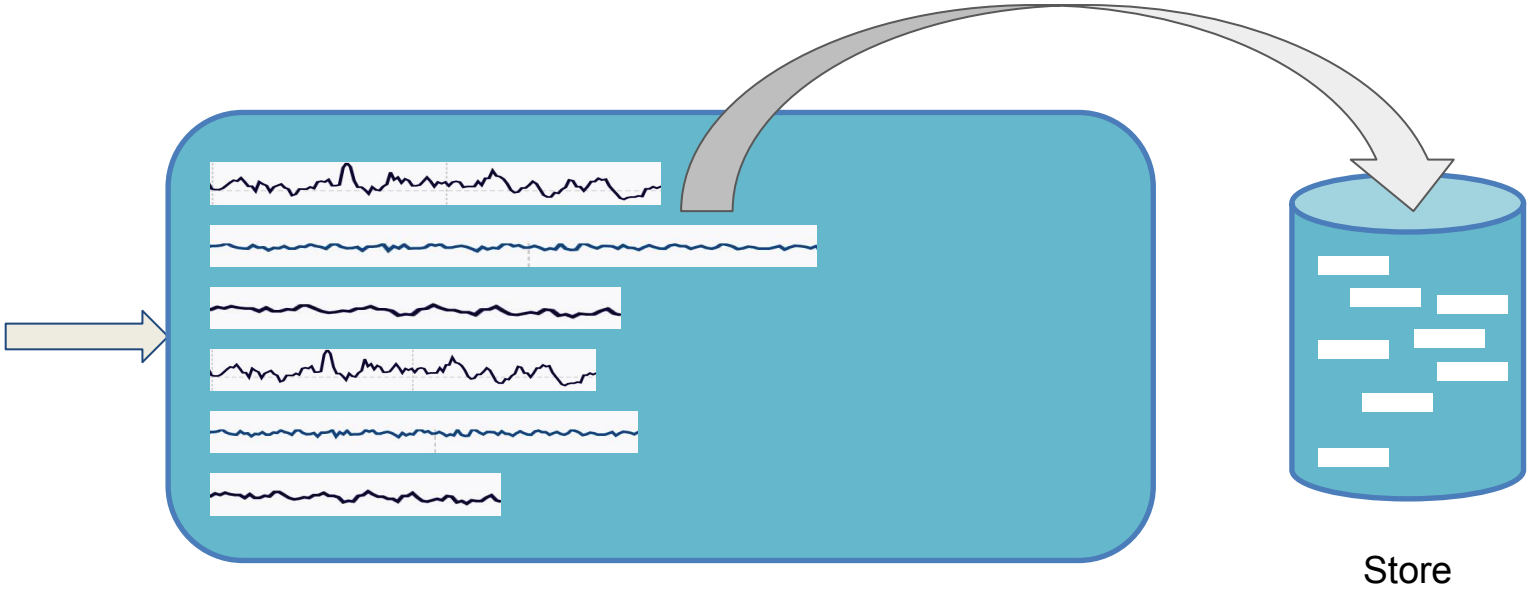

Sample
Distribute
Compress
Store
Query
CPU Usage by Namespace
CPU Usage by Service
Memory Usage by Namespace
Memory Usage by Service

Cortex: Distributing for scalability



DHTs: see http://nms.csail.mit.edu/papers/chord.pdf

Cortex data compression and chunking



Link to paper on Gorilla compression

Long-term storage

Want:

- Scalability
- Speed
- Durability

Long-term Storage

DynamoDB

Google Cloud Bigtable

S3

Google Cloud Storage

Cortex inverted index

Suppose PromQL query is:

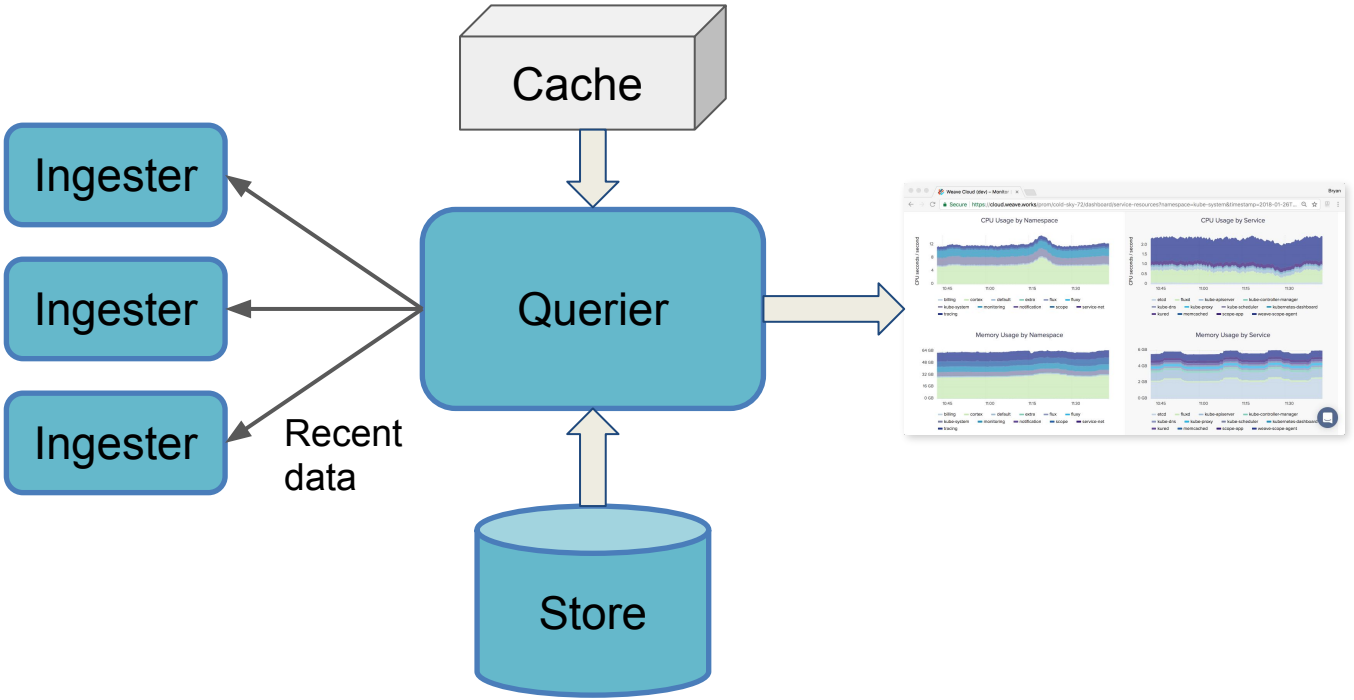
```
http_duration_seconds{job="nginx"}
```

Go to index row `http_duration_seconds:job`

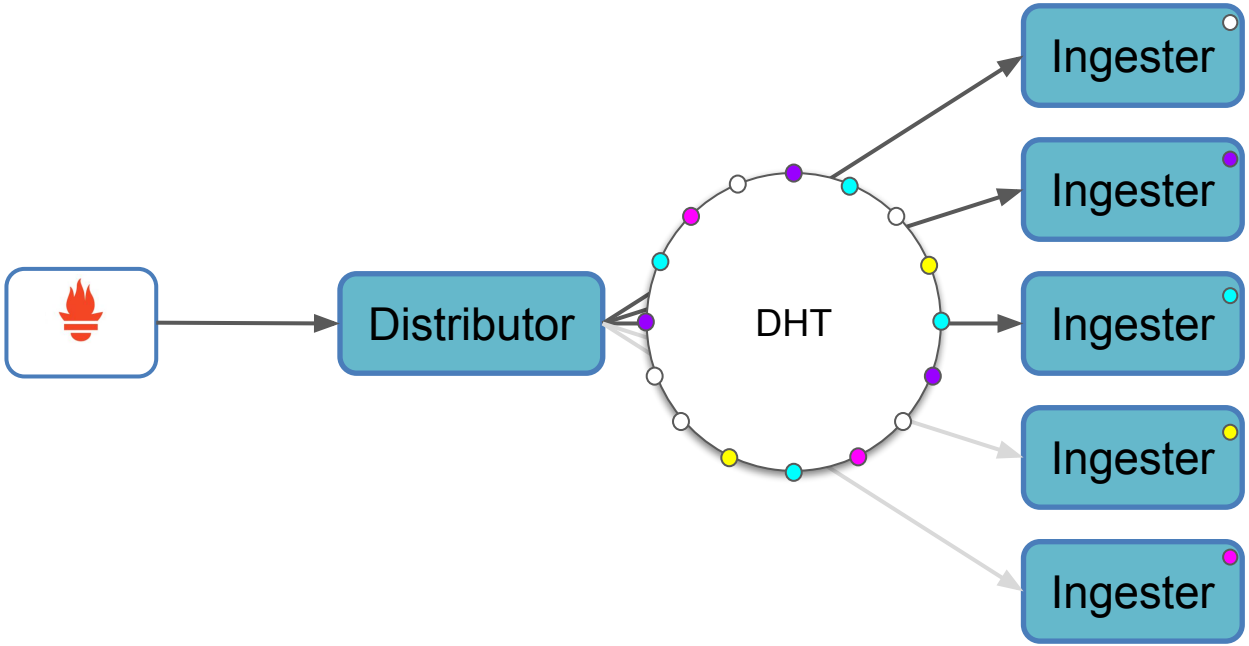
Look up “nginx”

- ➢ set of timeseries
  - ➢ look up each timeseries
    - ➢ set of chunks

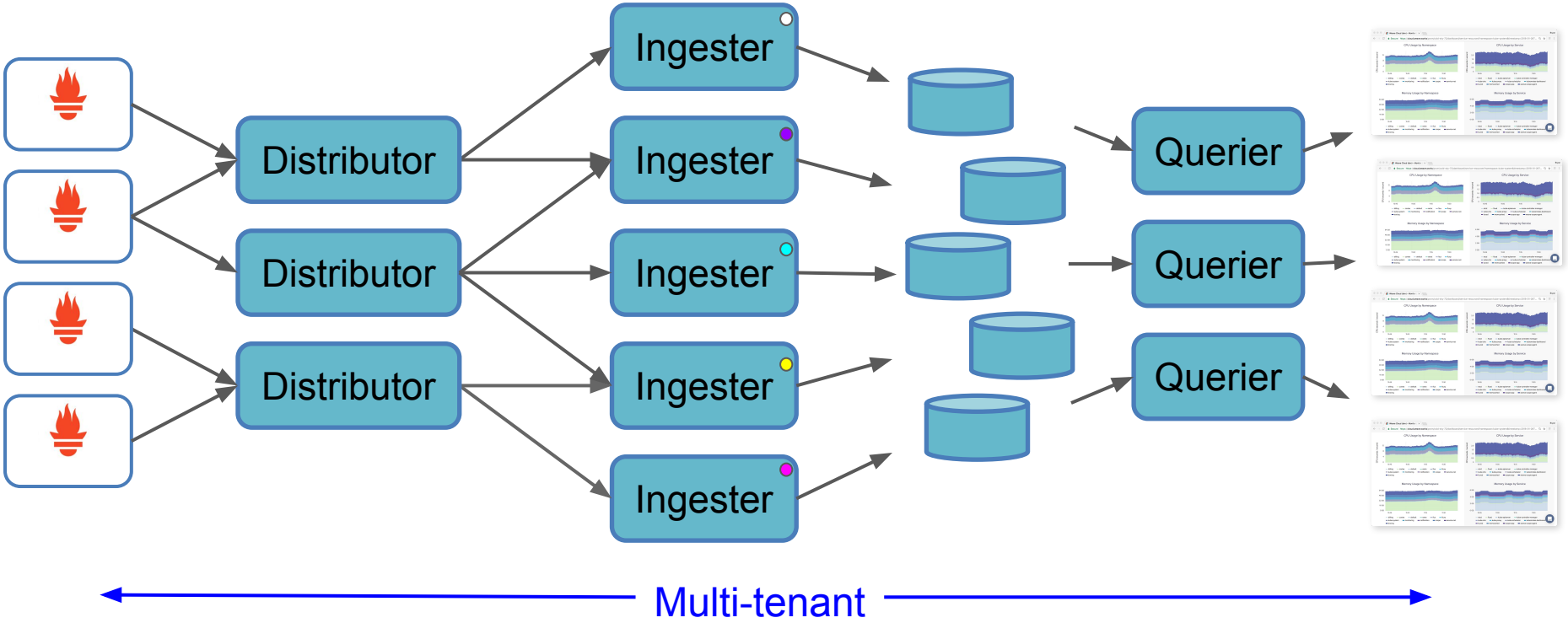
Cortex querier



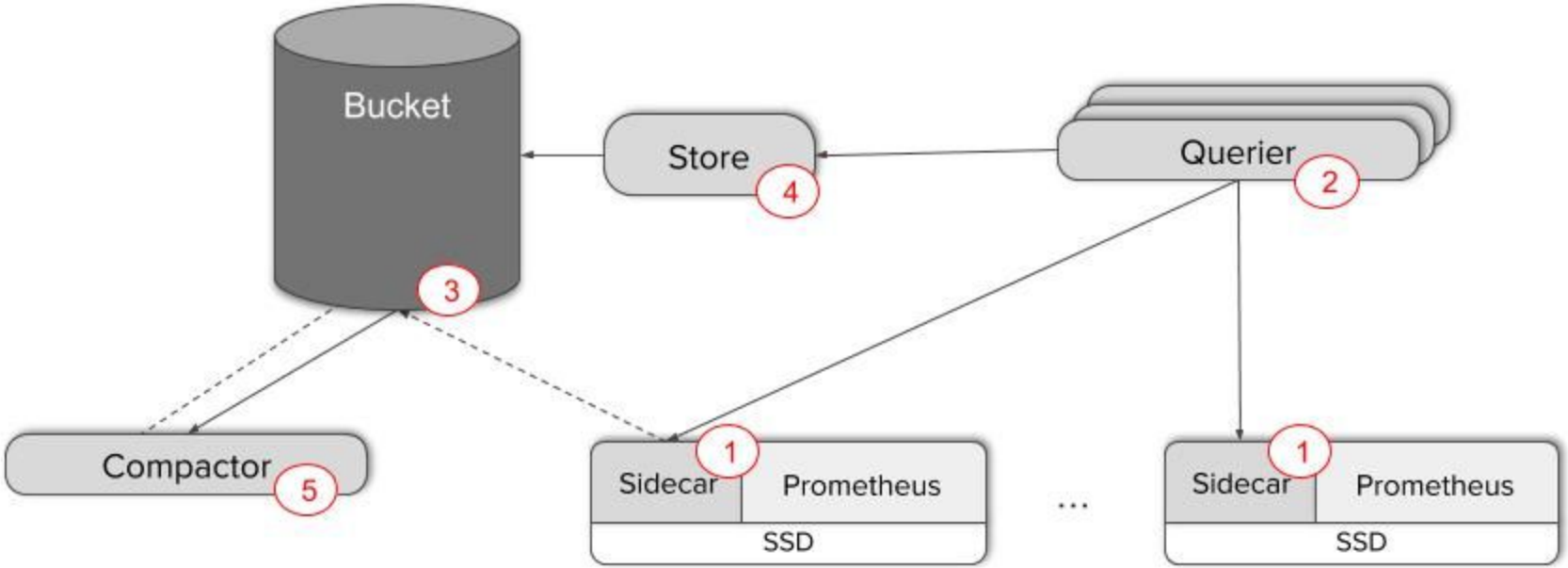
Cortex: Replicating for resiliency



Cortex: Infinitely Scalable Prometheus



Thanos


Bucket
Store
4
Querier
2
3
Compactor
5
1
Sidecar
Prometheus
SSD
...
1
Sidecar
Prometheus
SSD

Cortex similarities to Thanos

Huge re-use of Prometheus code

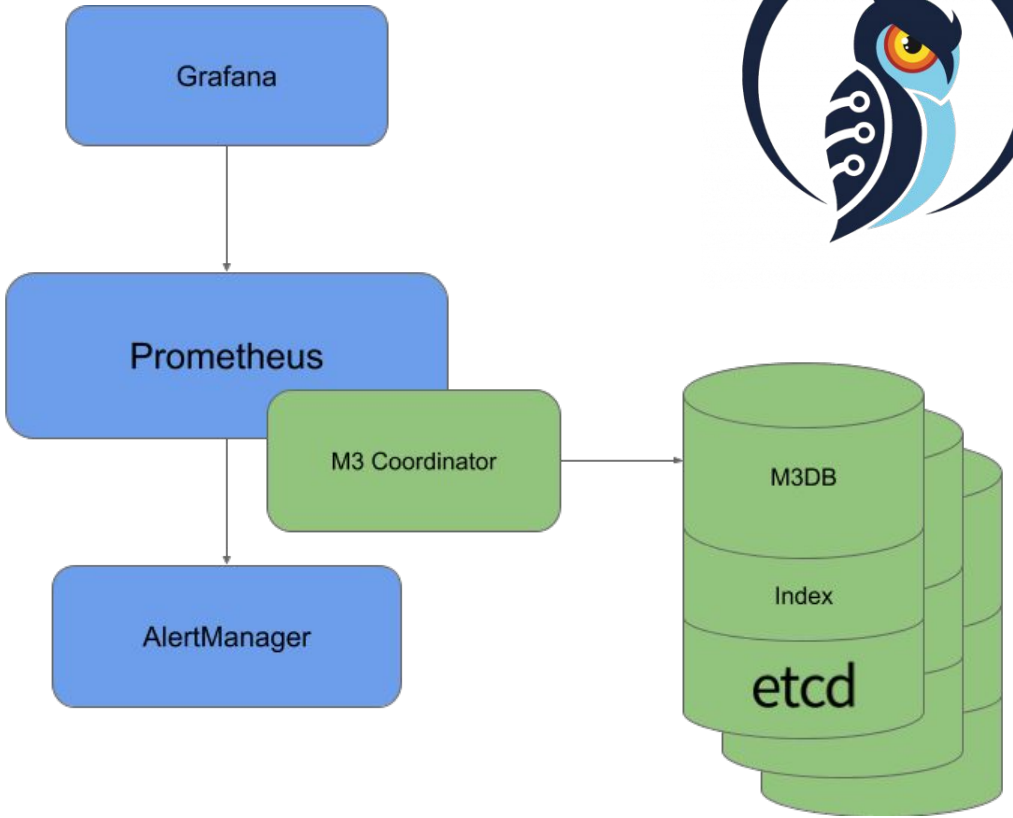
Bring multiple Prometheus' data into global view

Long-term storage in cloud buckets

Multi-component architecture

Cortex differences to Thanos

Multi-tenant	Single-tenant
Remote write API	Sidecar beside Prom
Automatic sharding	Manual sharding
Indexed small chunks	Prom TSDB blocks
	Downsampling


Grafana
Prometheus
M3 Coordinator
AlertManager
M3DB
Index
etcd

M3

Cortex: Experiences in production

We run Cortex as part of cloud.weave.works

Anyone on the Internet can sign up for a free trial

This should be fun...

Cortex: Experiences in production

Getting the best performance out of a NoSQL store

- Parallelising operations to take advantage of scale
- Batching operations to minimise call overheads
- Designing keys to avoid hot-spots
  - Schema has evolved - on v9 today
  - Still have all the code to read older data

Cortex: Experiences in production

Provisioning DynamoDB

- Ingester can queue up writes for many minutes - smooths out peaks
- Balancing capacity over multiple tables is a whole other trick
- Eventually automated the process, based on Cortex metrics for queueing and throttling

Cortex: Experiences in production

Running out of RAM

- Ingesters blowing up when they can’t flush
- Queriers blowing up when they get too many samples in memory

Cortex: Experiences in production

Short-lived timeseries are a significant pinch-point.

- Metadata dwarfs sample data for hours
- Things like Apache Spark create lots of short-lived pods
- cAdvisor (inside kubelet) had bugs creating thousands of spurious series

Cortex: recent enhancements

Caching index lookups

Caching index writes

Parallelising within queries

Bigger Chunks

Looking forward

Write-Ahead Log (WAL)

Simpler configuration

Sharded Ruler

Downsampling

THANK YOU!