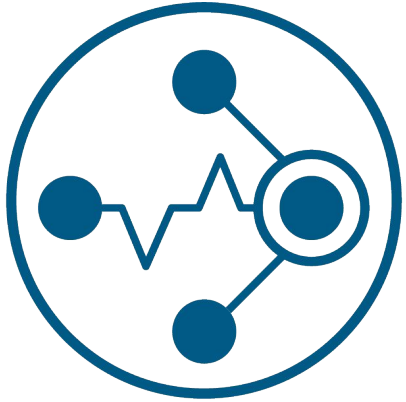
weaveworks

Cortex: Infinitely Scalable Prometheus

Bryan Boreham (@bboreham)



What is Cortex ?

Cortex is a time-series store built on Prometheus

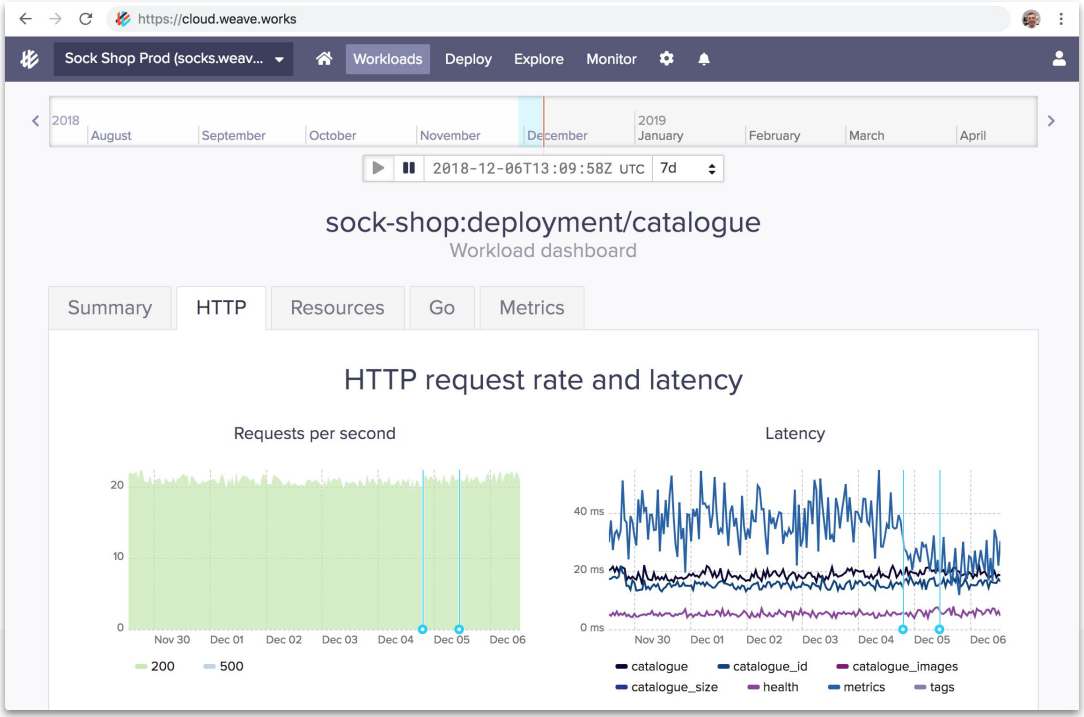
- Horizontally scalable
- Highly Available
- Long-term storage
- Multi-tenant

Cortex is a CNCF Sandbox project

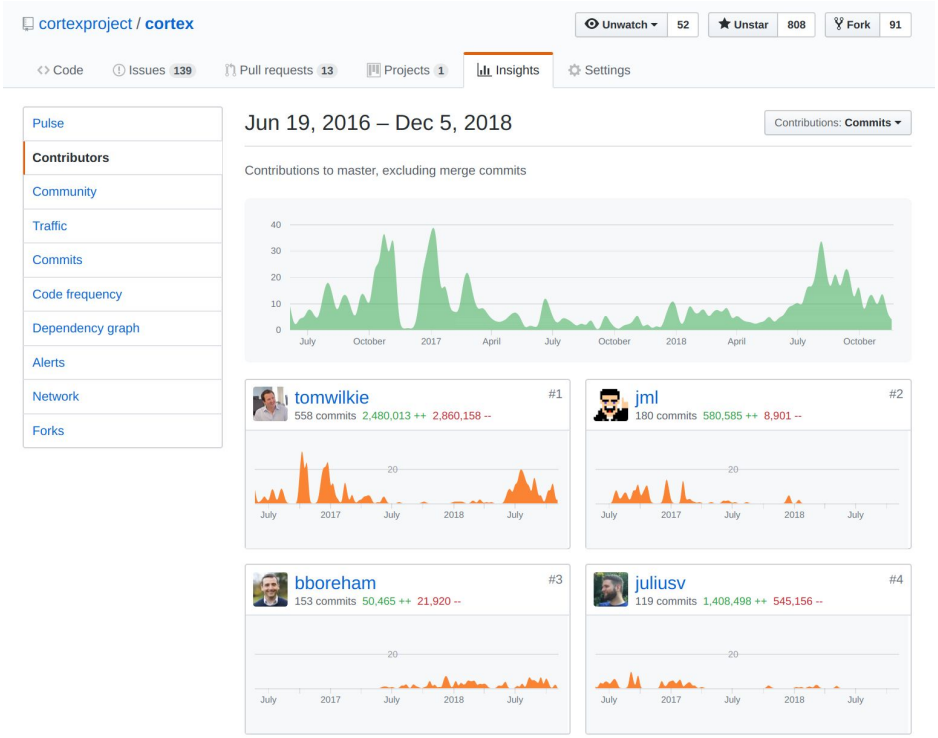
https://github.com/cortexproject/cortex

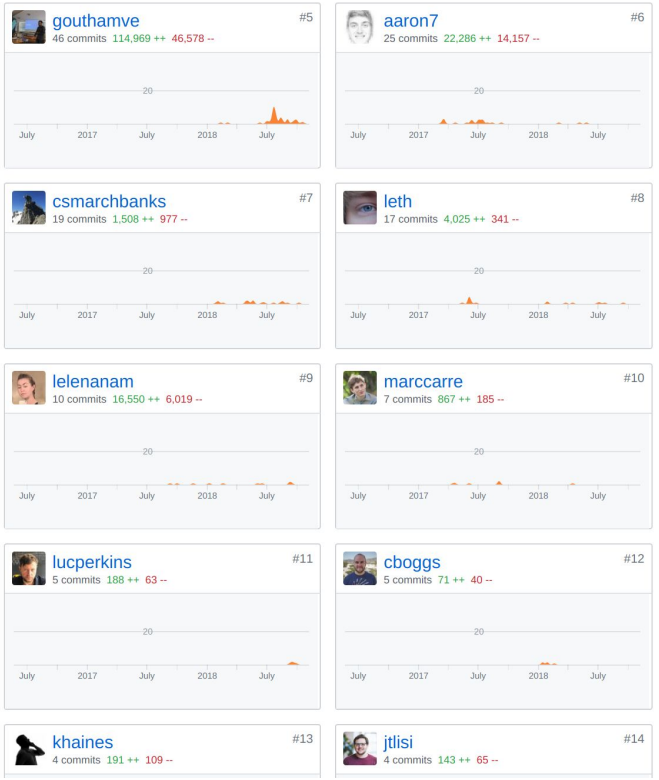
Why did we build Cortex

Prometheus As A Service on cloud.weave.works



Who wrote Cortex?





Who uses Cortex?

weaveworks

ELECTRONIC ARTS™

Qordoba

Grafana Labs

FreshTracks.io

adzuna

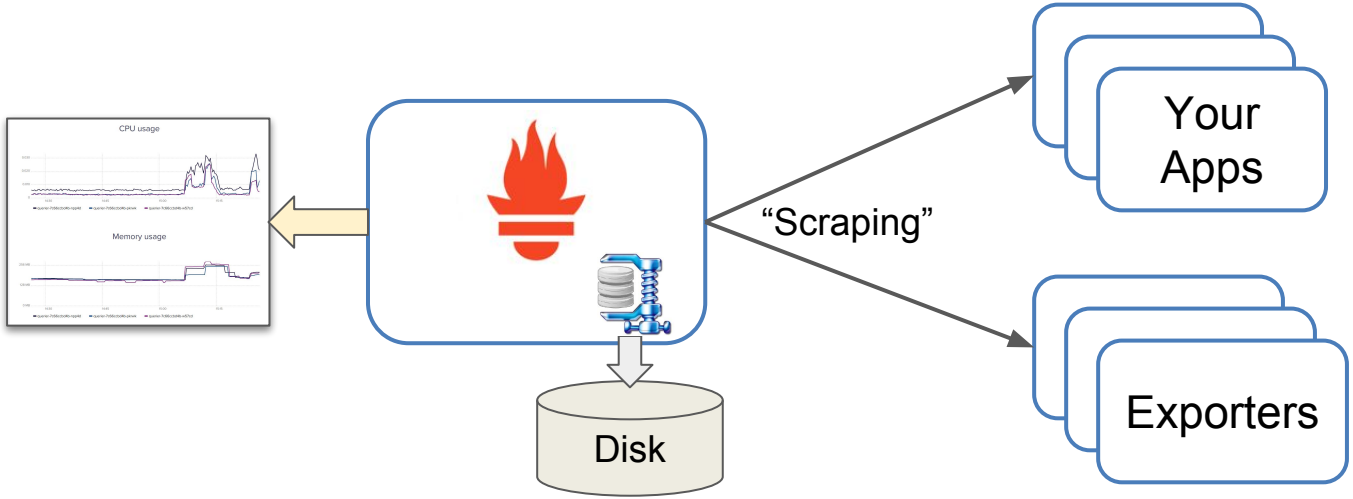
PLATFORM9

ASPEN MESH

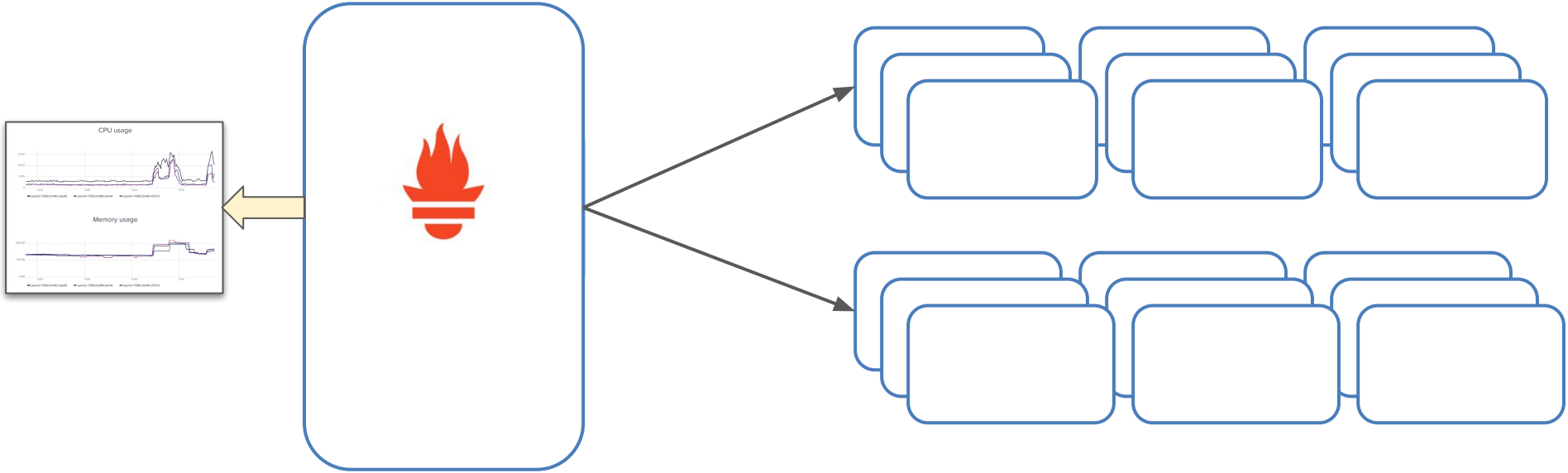
ADORE ME

MAYADATA

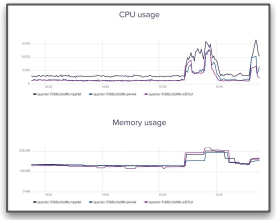
Prometheus: basic operation

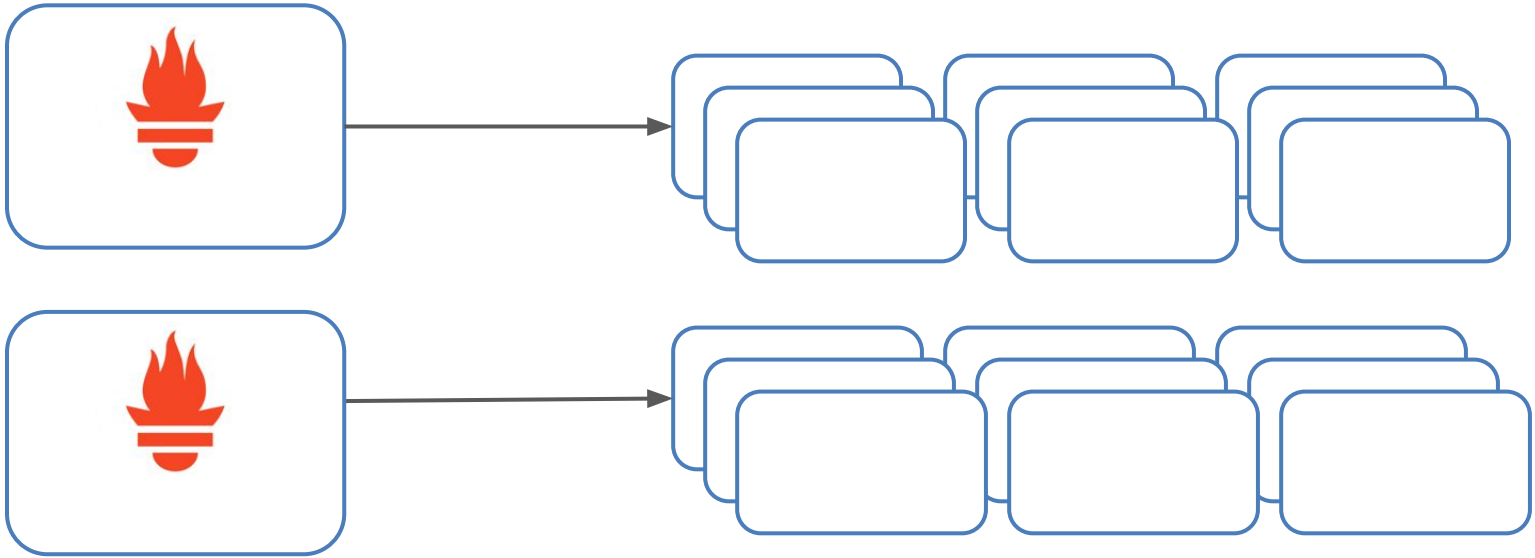


Scaling Prometheus

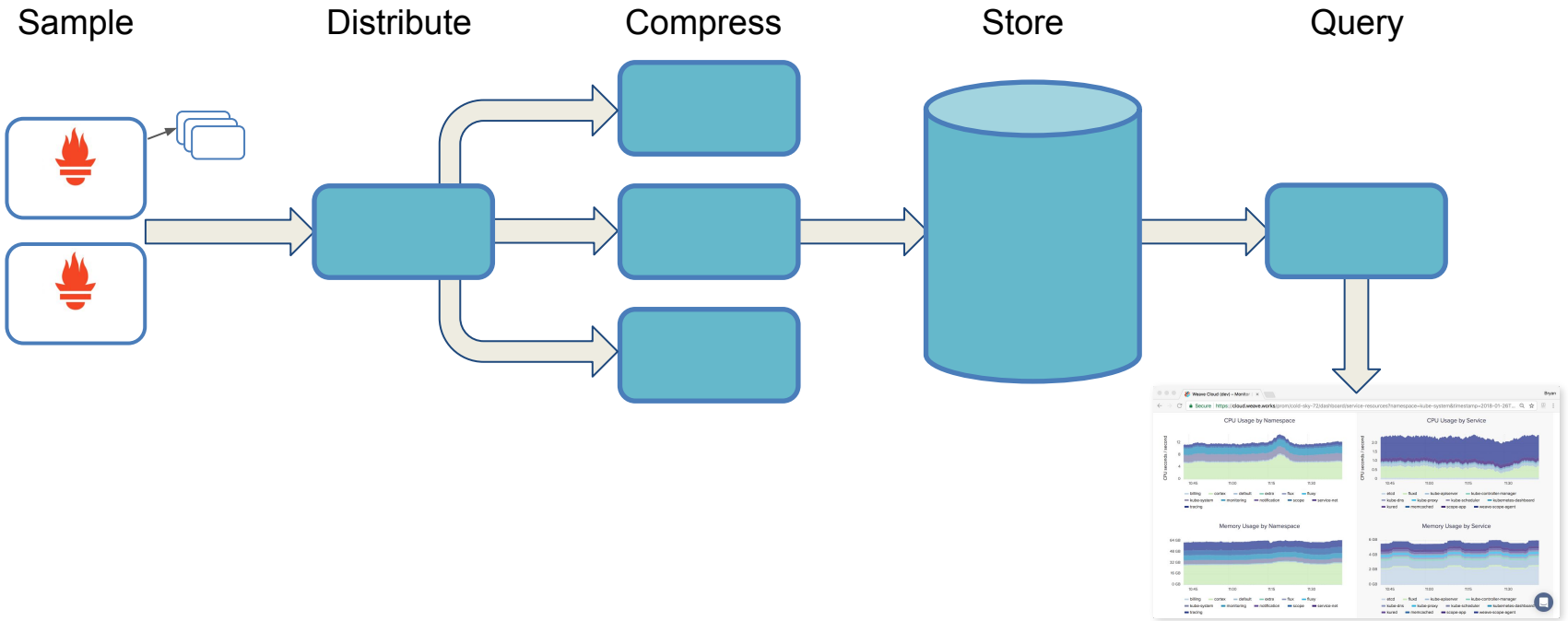


Sharding Prometheus

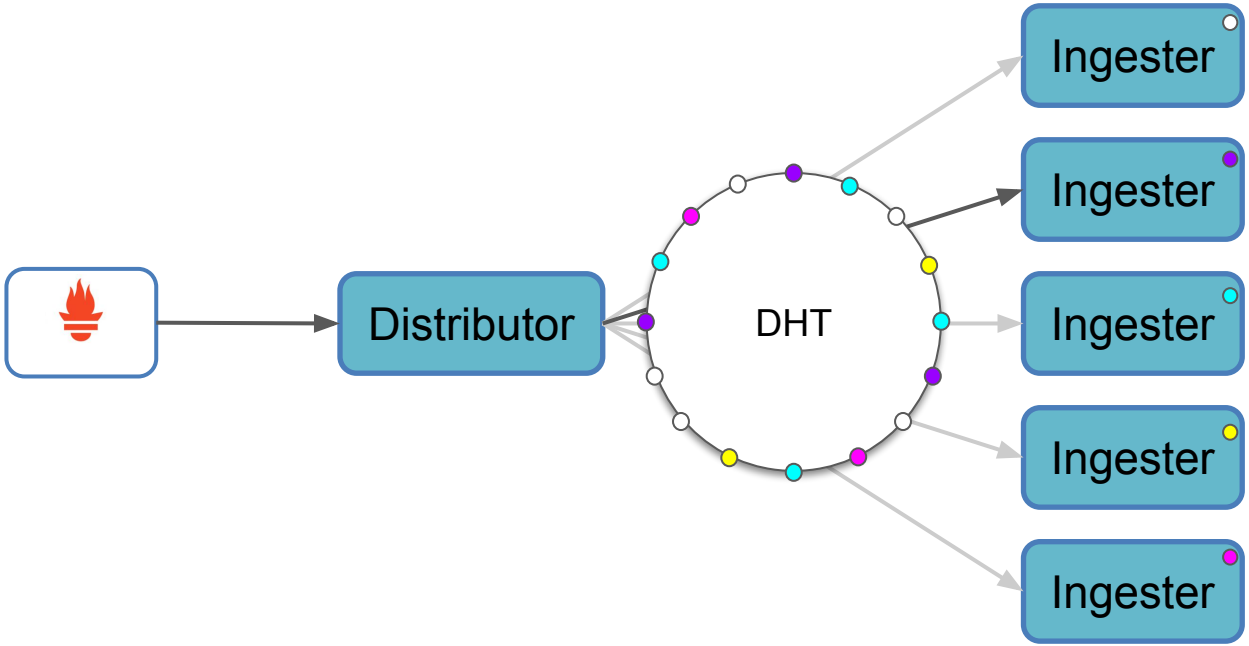




Cortex

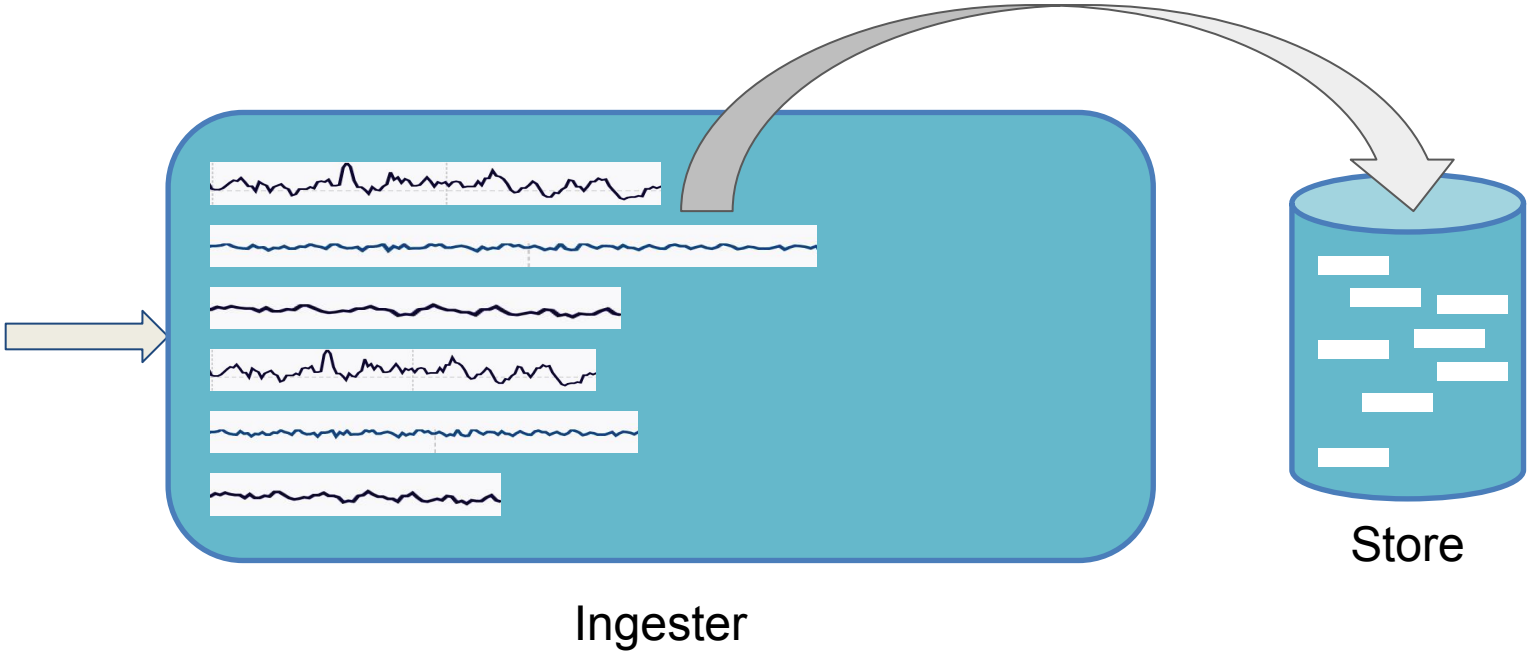

Sample
Distribute
Compress
Store
Query
CPU Usage by Namespace
CPU Usage by Service
Memory Usage by Namespace
Memory Usage by Service

Cortex: Distributing for scalability



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

Cortex data compression and chunking



Gorilla compression: http://www.vldb.org/pvldb/vol8/p1816-teller.pdf

Long-term storage

Want:

- Scalability
- Speed
- Durability

Long-term Storage

DynamoDB

Google Cloud Bigtable

S3

Google Cloud Storage

Cortex inverted index

```
http_duration_seconds{job="shipping",instance="a",path="foo",result="200"}
```

...	
http_duration_seconds:job	orders, shipping, customers, ...
http_duration_seconds:instance	a, b, c, d, ...
http_duration_seconds:path	/foo, /bar, /...
http_duration_seconds:result	200, 401, 402, 404, 501, 503, ...
...	

Cortex index lookup

Suppose PromQL query is:

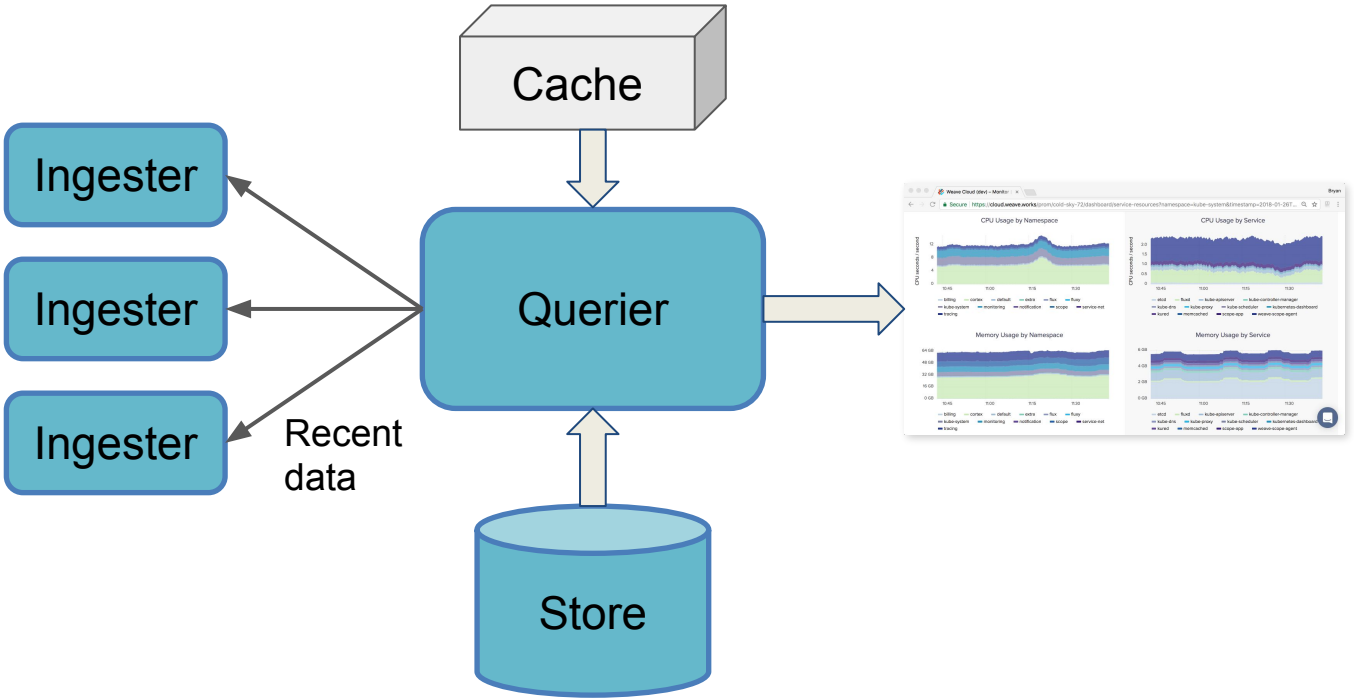
```
http_duration_seconds{job="shipping"}
```

Go to index row `http_duration_seconds:job`

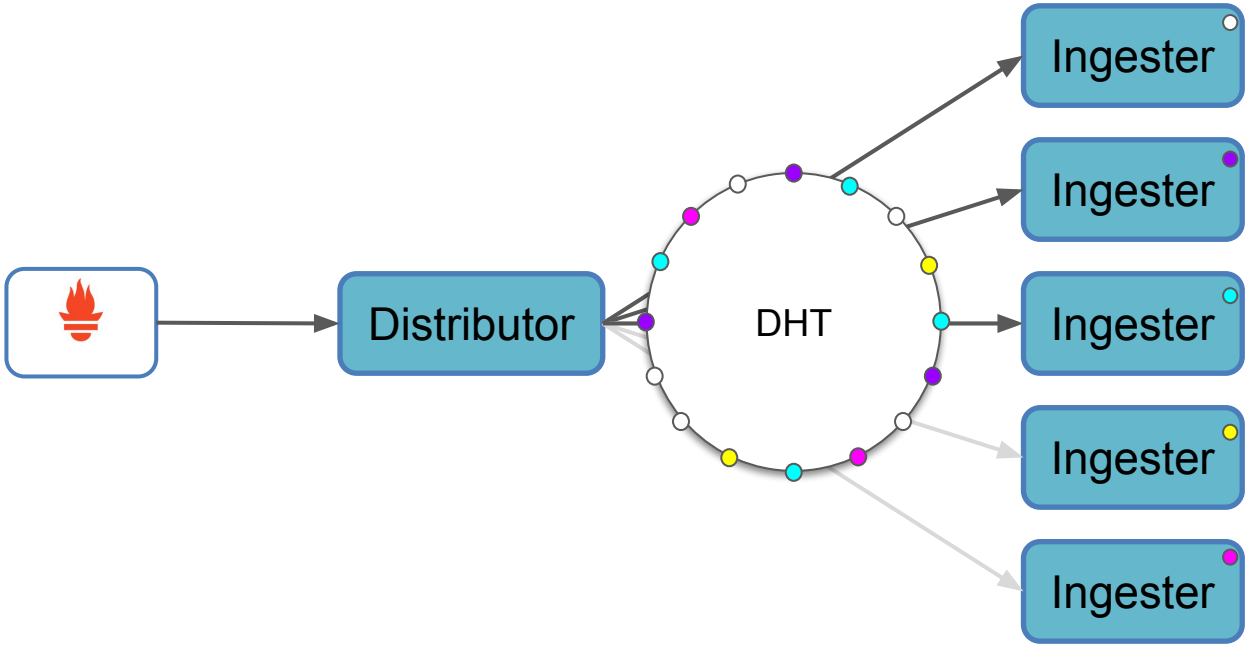
Look up “shipping”

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

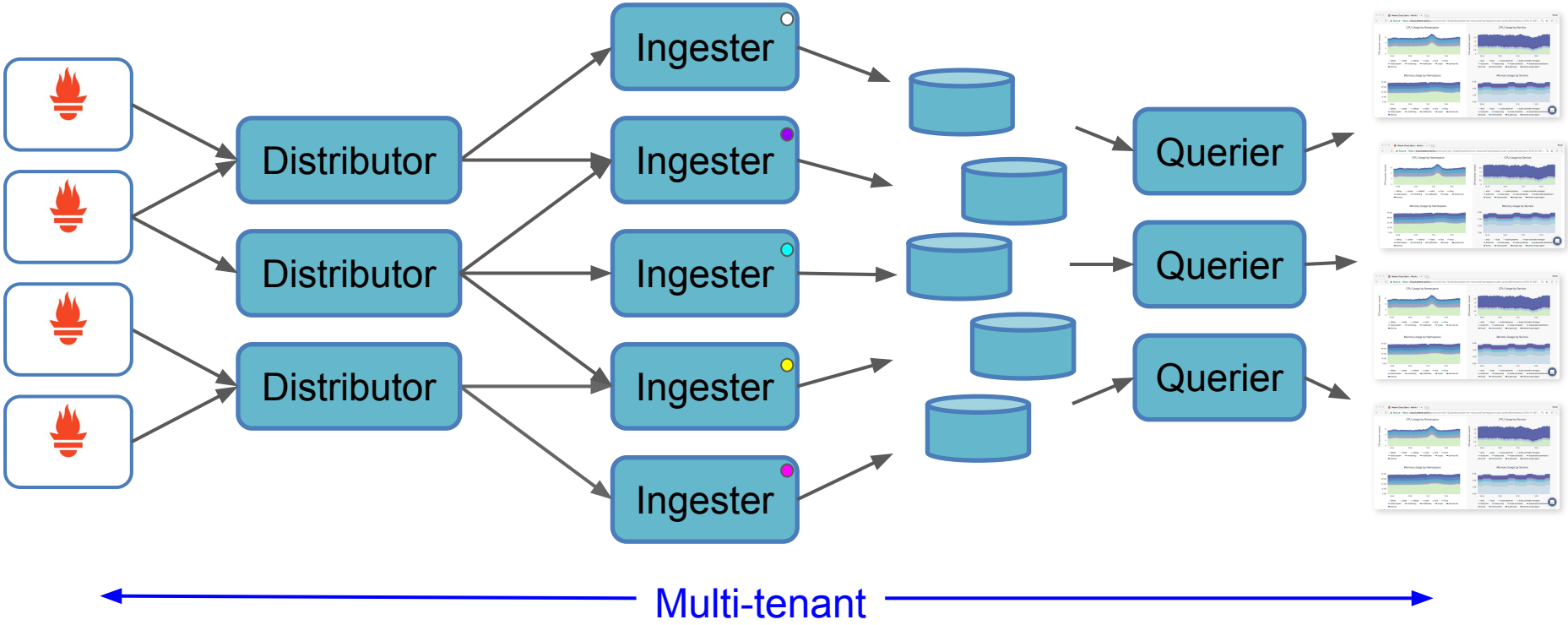
Cortex querier



Cortex: Replicating for resiliency



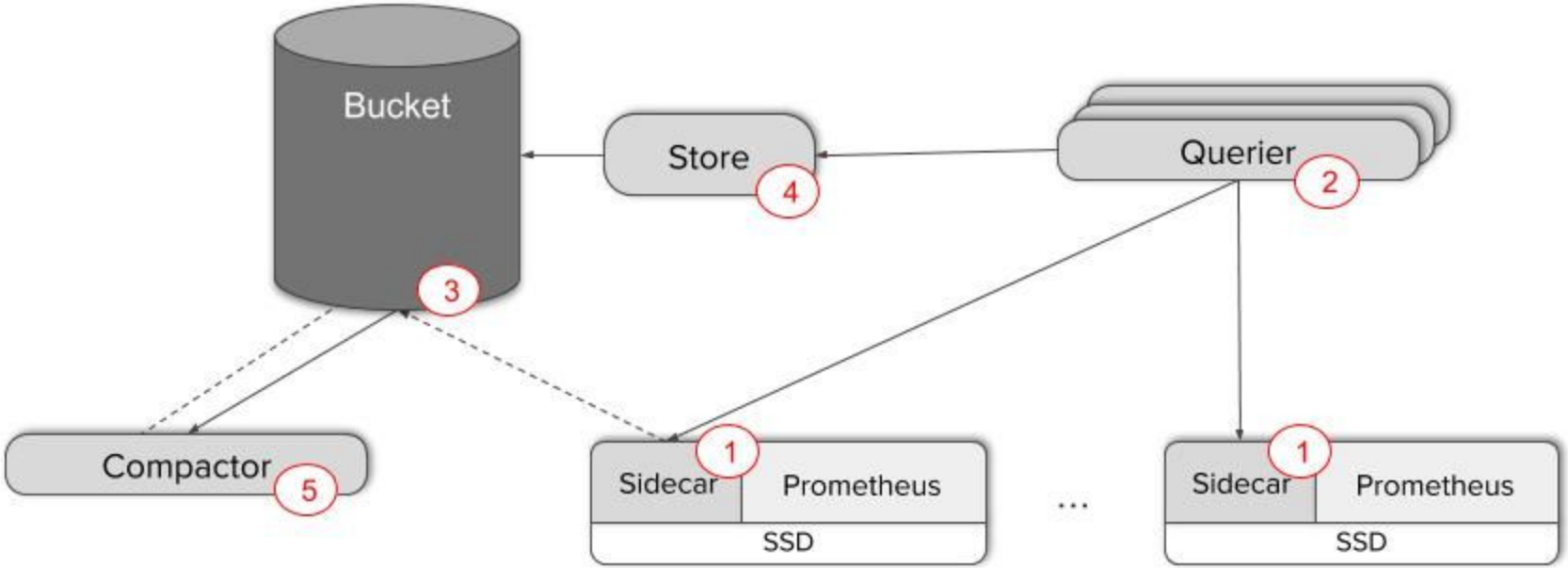
Cortex: Infinitely Scalable Prometheus



Also...

Thanos

“Highly available Prometheus setup with long term storage”



Cortex similarities to Thanos

Huge re-use of Prometheus code

Bring multiple Prometheus' data into global view

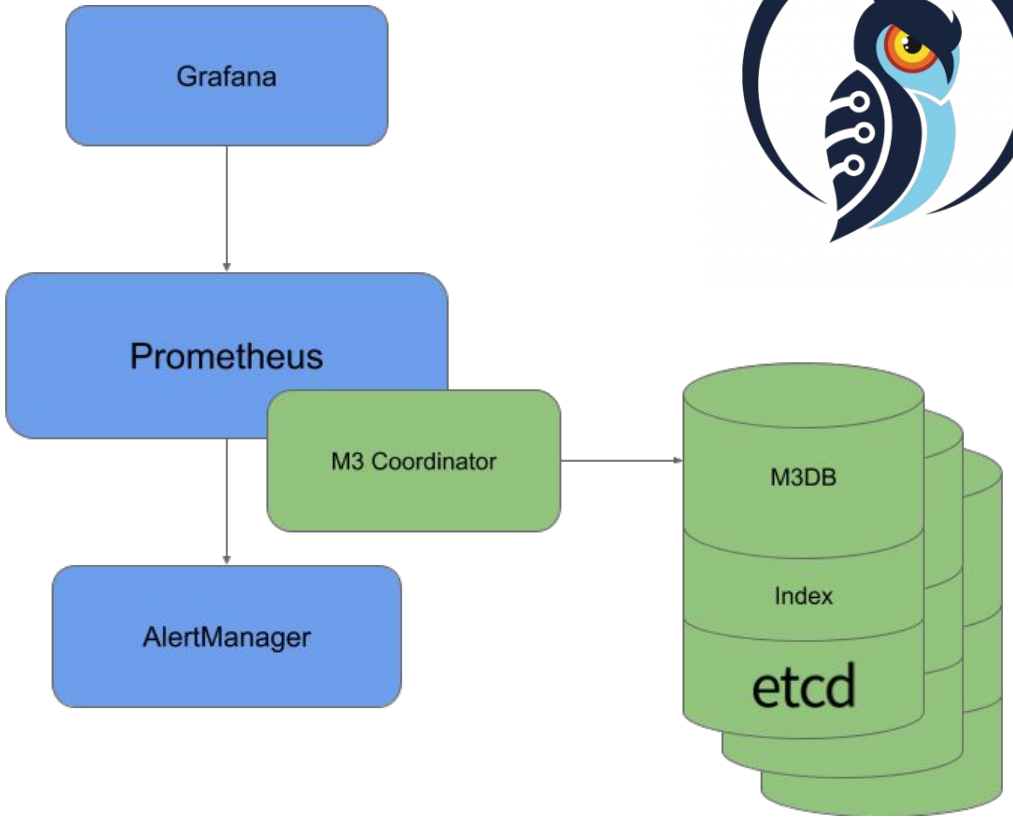
Split between recent data and historic data

Long-term storage in cloud buckets

Multi-component architecture

Cortex differences to Thanos

Multi-tenant	Single-tenant
Automatic sharding	Manual sharding
Indexed small chunks	Prom TSDB blocks
Query sharding	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 large NoSQL store is hard:

- Parallelising to take advantage of scale
- Batching to minimise call overheads
- Tuning index schema 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

Out of memory errors...

- Ingesters blowing up when they can't flush
- Queriers blowing up when they get too many samples in memory
- High-cardinality queries

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

Cortex: Looking forward

Write-Ahead Log (WAL)

Simpler runtime configuration

Sharded Ruler

Downsampling?

More users, more contributors!

THANK YOU!

https://github.com/cortexproject/cortex