caicloud 才云 | IT大咖说 知识分享平台

Monitoring your Kubernetes cluster with Prometheus

唐鹏程 才云科技

Why we need a monitoring system?

- To know whether things are running smoothly
- To locate faults when something goes wrong
- To send fire alarms to on-calls

What should a monitoring system provide?

- Define metrics data model
- Define interface to collect metrics
- Store metrics
- Provide query interface

Prometheus to the rescue

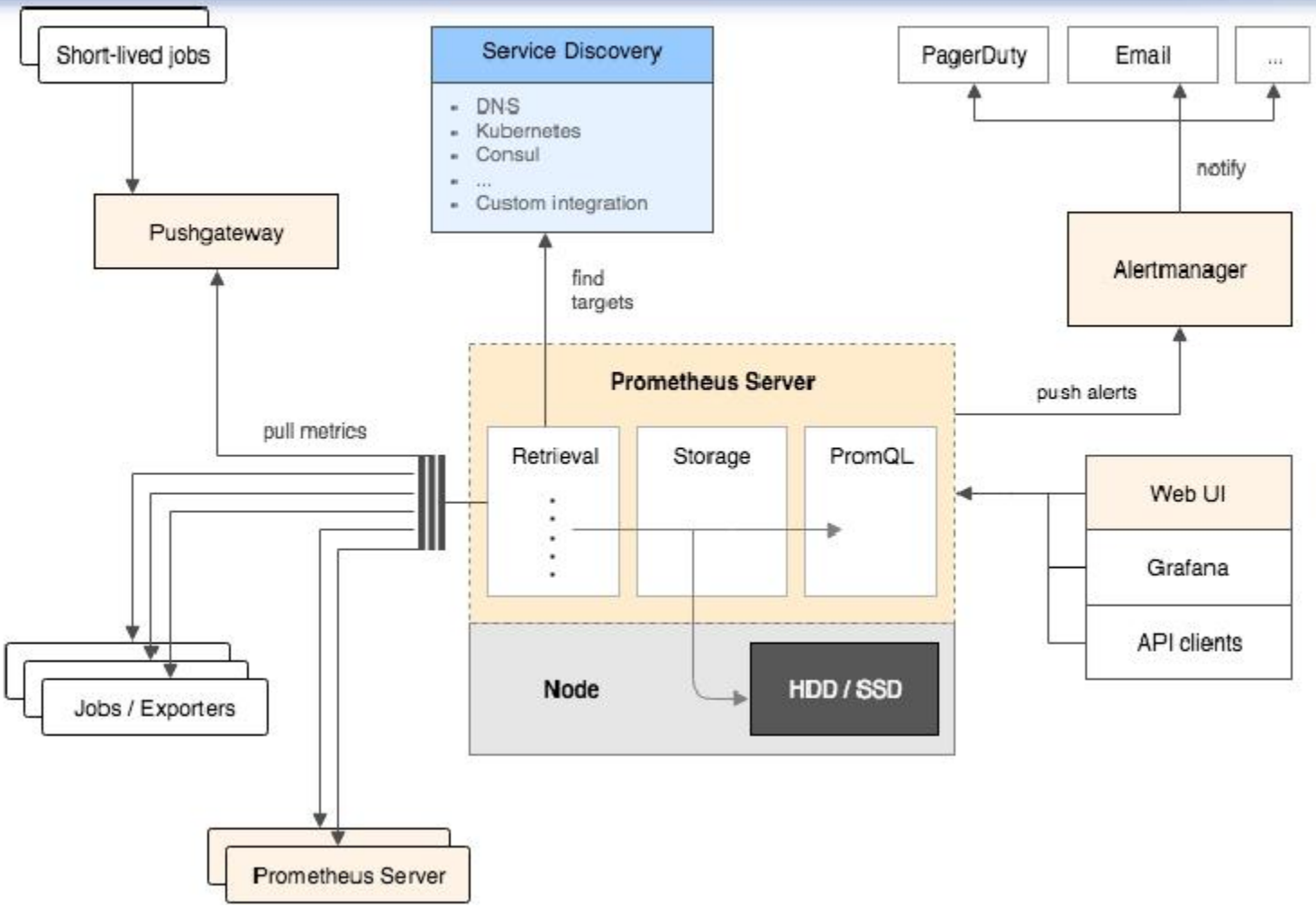
- Inspired by Google’s Borgmon
- A CNCF project initially developed at SoundCloud
- Pull metrics data points via HTTP
- Designed to be self-sustained without external dependencies

Prometheus

Prometheus IS:
A monitoring system for distributed system and infrastructure

Prometheus is NOT:
A long term persistent backend store for BI reporting or data mining

Prometheus Overview



Data model

<metric name>{<label name>=<label value>, ...}

Element	Value
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="nodes",verb="POST"}	3
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="pushs",verb="WATCHLIST"}	398
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="replicasets",verb="WATCHLIST"}	605
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="serviceaccounts",verb="LIST"}	1
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="replicationcontrollers",verb="LIST"}	1
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="pods",verb="POST"}	4
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="configmaps",verb="LIST"}	17
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="services",verb="LIST"}	385
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="pods",verb="WATCHLIST"}	135929
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="ippools",verb="WATCHLIST"}	132
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="limitranges",verb="WATCHLIST"}	608
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="configmaps",verb="GET"}	14604
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="pods",verb="CONNECT"}	5
apiserver_request_latencies_count{instance="kubernetes.default.svc:443",job="kubernetes-cluster",resource="globalconfigs",verb="WATCHLIST"}	132

Metrics Have Types

- Counter (request count, error count, cpu time)
- Gauge (memory usage, network i/o)
- Histogram
- Summary

Prometheus.yml

```
global:
   scrape_interval: 10s # duty interval
rule_files:
   [ - <file path> - ] # path to all the rule files (e.g /etc/prometheus/*.rule)
scrape_configs: # configurations of scrape jobs
 - job_name:  “k8s-exporter”
   static_configs:
      - targets: [ ‘localhost:9100’ , ‘localhost:9101’ ]
 - job_name:  “prometheus”
   static_configs:
      - targets: ['localhost:9090']
```

Prometheus.yml

```
global:
   scrape_interval: 10s # duty interval
rule_files:
   [ - <file path> - ] # path to all the rule files (e.g /etc/prometheus/*.rule)
scrape_configs: # configurations of scrape jobs
 - job_name:  “k8s-exporter”
   static_configs:
      - targets: [ ‘localhost:9100’ , ‘localhost:9101’ ]
 - job_name:  “prometheus”
   static_configs:
      - targets: ['localhost:9090']
```

Rule evaluation

Recording rule:

container_cpu_usage_seconds_total:rate_1m =
rate(container_cpu_usage_seconds_total[1m])

```
sum(container_cpu_usage_seconds_total:rate_1m{pod_name=~"kube-dns.+"}) by (pod_name)
```

Execute | - insert metric at cursor -

Graph | Console

Element	Value
{pod_name="kube-dns-v21-1006896387-t6qc2"}	0.009825499579999359
{pod_name="kube-dns-v21-1006896387-rm72m"}	0.008095682946837693
{pod_name="kube-dns-v21-1006896387-s93hb"}	0.0112235026197915
{pod_name="kube-dns-autoscaler-v21-1598293328-hvq4o"}	0.009281624365280745
{pod_name="kube-dns-v21-1006896387-sg736"}	0.011011953822772582
{pod_name="kube-dns-v21-1006896387-23cj5"}	0.009780900349964945
{pod_name="kube-dns-v21-1006896387-48v0n"}	0.008344431362745517

Prometheus.yml

```
global:
   scrape_interval: 10s # duty interval
rule_files:
   [ - <file path> - ] # path to all the rule files (e.g /etc/prometheus/*.rule)
scrape_configs: # configurations of scrape jobs
 - job_name:  “k8s-exporter”
   static_configs:
      - targets: [ ‘localhost:9100’ , ‘localhost:9101’ ]
 - job_name:  “prometheus”
   static_configs:
      - targets: ['localhost:9090']
```

Prometheus kubernetes service discovery

```
scrape_config:
   - job_name:   “kubernetes_nodes”
     kubernetes_sd_config:
     - role: node
     relabel_config:
     - action: labelmap
       regex: __meta_kubernetes_node_label_(.+)
```

Prometheus kubernetes service discovery

```
scrape_config:
  - job_name:   “kubernetes_nodes”
    kubernetes_sd_config:  # support other service discovery
    - role: node
    relabel_config:
    - action: labelmap
      regex: __meta_kubernetes_node_label_(.+)
```

Prometheus kubernetes service discovery

```
scrape_config:
  - job_name:   “kubernetes_nodes”
    kubernetes_sd_config:
    - role: node        # can be node, pod, endpoint or service
    relabel_config:
    - action: labelmap
      regex: __meta_kubernetes_node_label_(.+)
```

Prometheus kubernetes service discovery

```
scrape_config:
  - job_name:   “kubernetes_nodes”
    kubernetes_sd_config:
    - role: node
    relabel_config:  # take actions with regard to labels
    - action: labelmap
      regex: __meta_kubernetes_node_label_(.+)
```

Exporters

Exporter exports existing metrics from third party systems in prometheus format.

- Databases
- Hardware related
- Storage systems
- HTTP
- …

https://prometheus.io/docs/instrumenting/exporters/

Exporters

Mongo exporter

db.serverStatus() =>

```
> db.serverStatus()
{
        "host" : "storage-mongo-v3.0.5-fix-permission-1440045134-znv1s",
        "version" : "3.0.14",
        "process" : "mongod",
        "pid" : NumberLong(1),
        "uptime" : 355395,
        "uptimeMillis" : NumberLong(355395090),
        "uptimeEstimate" : 348748,
        "localTime" : ISODate("2017-05-04T05:50:01.109Z"),
        "asserts" : {
                "regular" : 0,
                "warning" : 0,
                "msg" : 0,
                "user" : 0,
                "rollovers" : 0
        },
        "backgroundFlushing" : {
                "flushes" : 5923,
                "total_ms" : 207835,
                "average_ms" : 35.08948168158028,
                "last_ms" : 23,
                "last_finished" : ISODate("2017-05-04T05:49:52.488Z")
        },
        "connections" : {
                "current" : 2,
                "available" : 838858,
                "totalCreated" : NumberLong(4)
        },
        "cursors" : {
                "note" : "deprecated, use server status metrics",
                "clientCursors_size" : 0,
                "totalOpen" : 0,
                "pinned" : 0,
                "totalNoTimeout" : 0,
                "timedOut" : 0
        },
```

Mongo exporter

Mongo Exporter

--mongodb.uri=<ip>:<port>
--web.listen-address=:9001
--web.metrics-path=/metrics

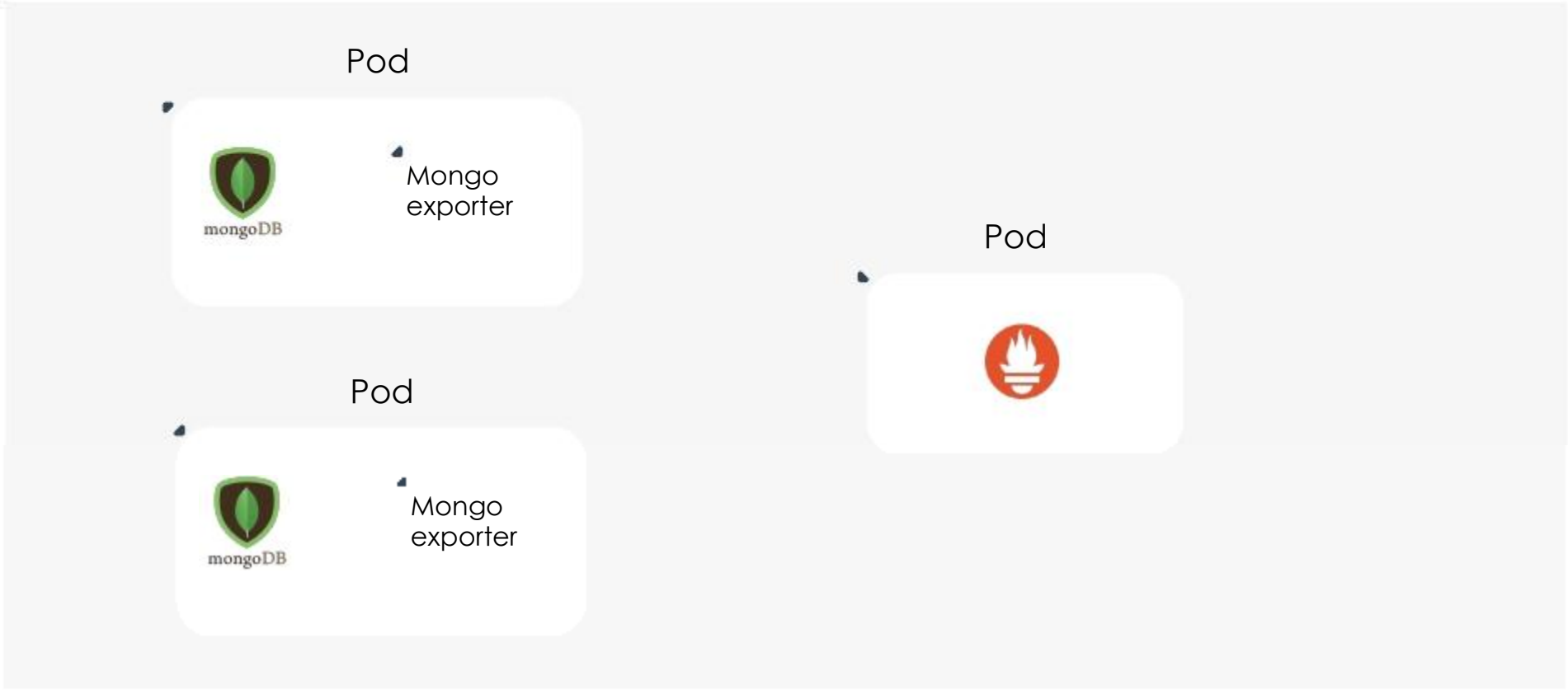
Exporters

Mongo exporter

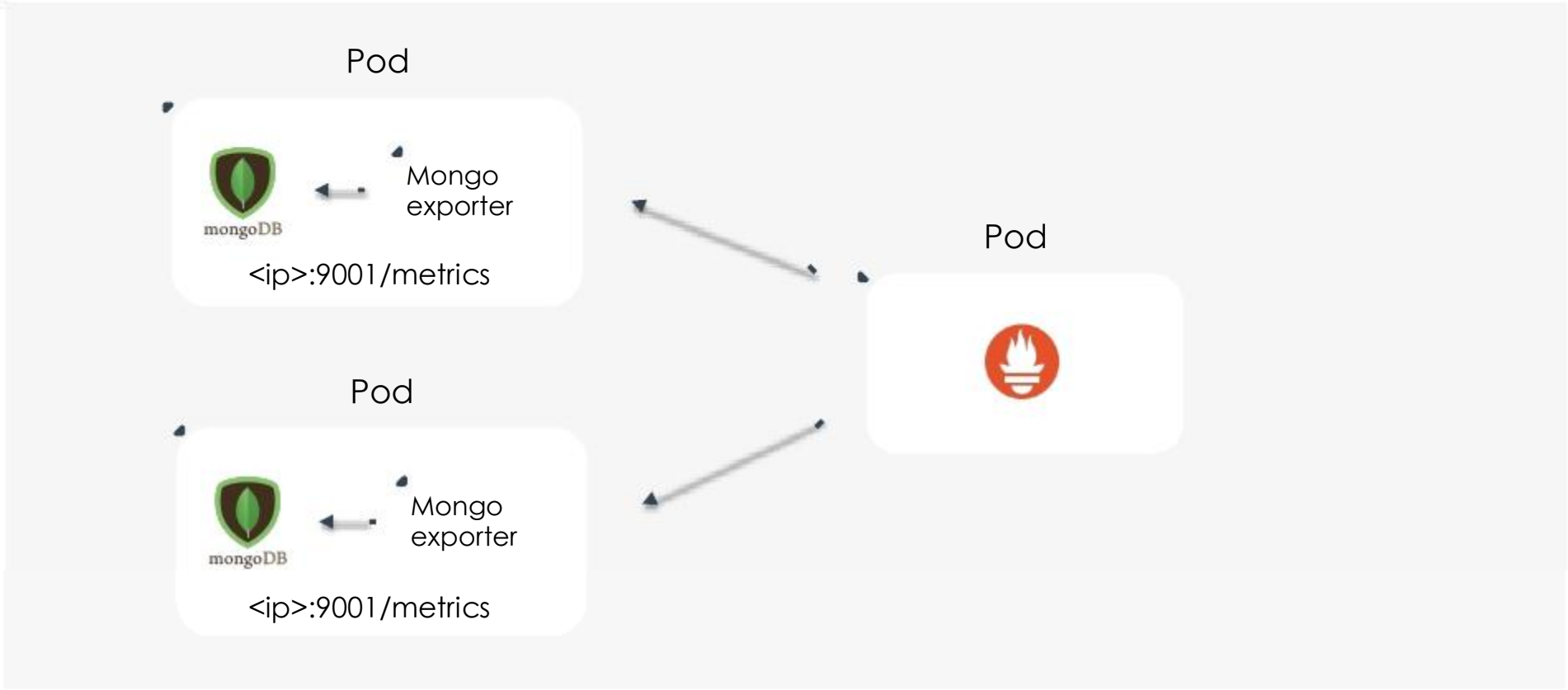
```
> db.serverStatus()
{
        "host" : "storage-mongo-v3.0.5-fix-permission-1440045134-znv1s",
        "version" : "3.0.14",
        "process" : "mongod",
        "pid" : NumberLong(1),
        "uptime" : 355395,
        "uptimeMillis" : NumberLong(355395090),
        "uptimeEstimate" : 348748,
        "localTime" : ISODate("2017-05-04T05:50:01.109Z"),
        "asserts" : {
                "regular" : 0,
                "warning" : 0,
                "msg" : 0,
                "user" : 0,
                "rollovers" : 0
        },
        "backgroundFlushing" : {
                "flushes" : 5923,
                "total_ms" : 207835,
                "average_ms" : 35.08948168158028,
                "last_ms" : 23,
                "last_finished" : ISODate("2017-05-04T05:49:52.488Z")
        },
        "connections" : {
                "current" : 2,
                "available" : 838858,
                "totalCreated" : NumberLong(4)
        },
        "cursors" : {
                "note" : "deprecated, use server status metrics",
                "clientCursors_size" : 0,
                "totalOpen" : 0,
                "pinned" : 0,
                "totalNoTimeout" : 0,
                "timedOut" : 0
        },
```

```
# HELP mongodb_asserts_total The asserts document reports the number of asserts on the database
the log file for the mongod process for more information. In many cases these errors are trivio
# TYPE mongodb_asserts_total counter
mongodb_asserts_total{type="msg"} 0
mongodb_asserts_total{type="regular"} 0
mongodb_asserts_total{type="rollovers"} 0
mongodb_asserts_total{type="user"} 0
mongodb_asserts_total{type="warning"} 0
# HELP mongodb_background_flushing_average_milliseconds The average_ms value describes the rela
to disk. The larger flushes is, the more likely this value is likely to represent a "normal," t
# TYPE mongodb_background_flushing_average_milliseconds gauge
mongodb_background_flushing_average_milliseconds 1.2912523175459296
# HELP mongodb_background_flushing_flushes_total flushes is a counter that collects the number
s of time
# TYPE mongodb_background_flushing_flushes_total counter
mongodb_background_flushing_flushes_total 5933
# HELP mongodb_background_flushing_last_finished_time The last_finished field provides a timest
relative to your server's current time and accounting for differences in time zone, restarting
# TYPE mongodb_background_flushing_last_finished_time gauge
mongodb_background_flushing_last_finished_time 1.4938773340+09
# HELP mongodb_background_flushing_last_milliseconds The value of the last_ms field is the amou
e current performance of the server and is in line with the historical data provided by average
# TYPE mongodb_background_flushing_last_milliseconds gauge
mongodb_background_flushing_last_milliseconds 0
# HELP mongodb_background_flushing_total_milliseconds The total_ms value provides the total num
his is an absolute value, consider the value offlushes and average_ms to provide better context
# TYPE mongodb_background_flushing_total_milliseconds counter
mongodb_background_flushing_total_milliseconds 7661
# HELP mongodb_connections The connections sub document data regarding the current status of in
apacity requirements of the server
# TYPE mongodb_connections gauge
mongodb_connections{state="available"} 838859
mongodb_connections{state="current"} 1
# HELP mongodb_connections_metrics_created_total totalCreated provides a count of all incoming
# TYPE mongodb_connections_metrics_created_total counter
mongodb_connections_metrics_created_total 58805
# HELP mongodb_cursors The cursors data structure contains data regarding cursor state and use
# TYPE mongodb_cursors gauge
mongodb_cursors{state="pinned"} 0
mongodb_cursors{state="timed_out"} 0
mongodb_cursors{state="total_no_timeout"} 0
mongodb_cursors{state="total_open"} 0
```

How do all these fit together?



How do all these fit together?



Exporters

mongodb_op_counters_total

Load time: 413ms
Resolution: 14s

Execute - insert metric at cursor -

Graph Console

Element	Value
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="delete",version="v3.0.5"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="getmore",version="v3.0.5"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="command"}	359788
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="update",version="v3.0.5"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="insert",version="v3.0.5"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="query",version="v3.0.5"}	1
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="command",version="v3.0.5"}	11374
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="delete",version="v3.0.5"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="query",version="v3.0.5"}	1
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="update",version="v3.0.5"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="delete"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="insert",version="v3.0.5"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="getmore"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="update"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="insert"}	0
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.10.70:9001",job="kubernetes-pods",namespace="kube-system",pod_name="test-mongo-v3.0.5-148hq",type="getmore",version="v3.0.5"}	0
mongodb_op_counters_total{instance="10.100.76.57:9001",job="kubernetes-service-endpoints",kubernetes_io_cluster_service="true",namespace="default",service_name="test-mongo",type="query"}	1
mongodb_op_counters_total{caicloud_app="test-mongo",instance="10.100.76.57:9001",job="kubernetes-pods",namespace="default",pod_name="test-mongo-v3.0.5-x35f1",type="command",version="v3.0.5"}	359794

Scaling and High availability

"A single Prometheus server can easily handle millions of time series. That's enough for a thousand servers with a thousand time series each scraped every 10 seconds."

-- Brian Brazil

Scaling and High availability

"A single Prometheus server can easily handle millions of time series. That's enough for a thousand servers with a thousand time series each scraped every 10 seconds."

-- Brian Brazil

What if we go beyond that?

Prometheus -> Prometheis

- Splitting by use
- Horizontal Sharding

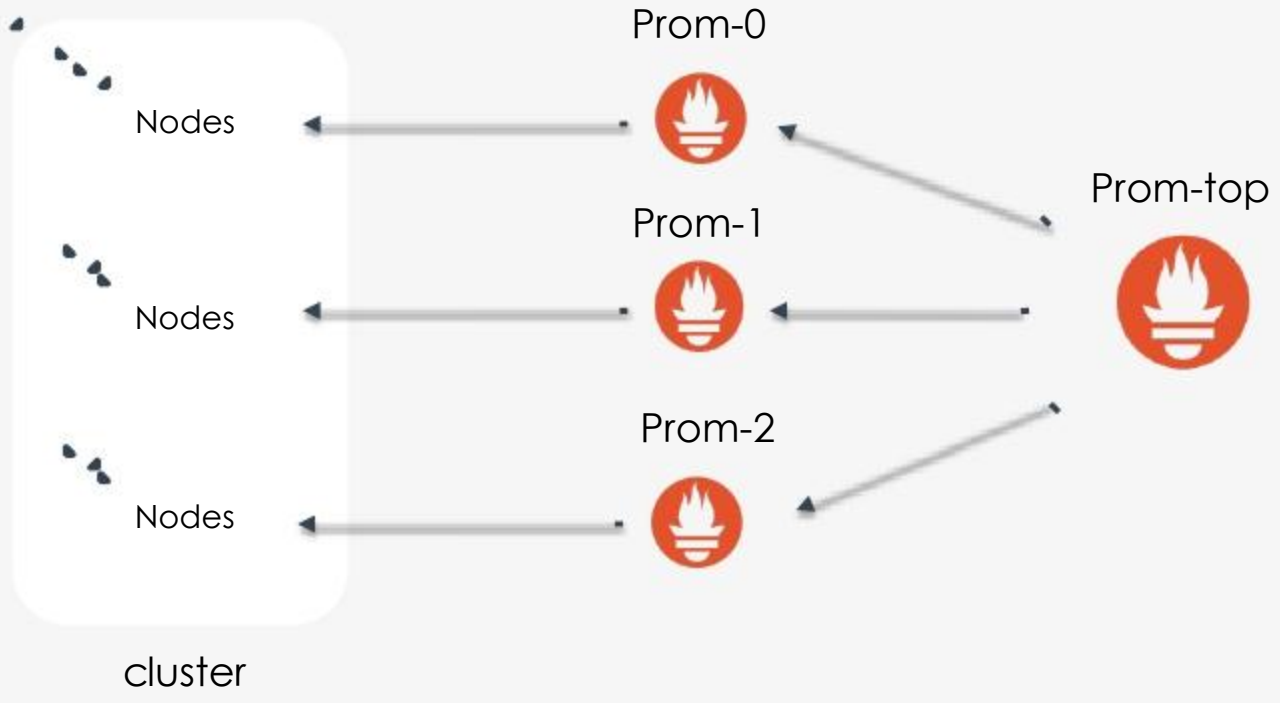
Splitting by use

Cassandra Prometheus

Elasticsearch Prometheus

Hadoop Prometheus

Horizontal Sharding



Slave configuration

```
global:
  external_labels:
     slave: 1 # This is the 2nd slave. This prevents clashes between slaves.
scrape_configs:
  - job_name: some_job
  # Add usual service discovery here, such as static_configs
  relabel_configs:
  - source_labels: [__address__]
    modulus: 3 # 3 slaves
    target_label: __tmp_hash
    action: hashmod
  - source_labels: [__tmp_hash]
    regex: ^1$ # This is the 2nd slave
    action: keep
```

Top level configuration

```
- scrape_config:
  - job_name: slaves
    honor_labels: true
    metrics_path: /federate
    params:
      match[]: -  '{__name__=~ "^slave:.*" }'   # Request all slave-level time series
    static_configs:
      - targets:
        - Prom-0:9090
        - Prom-1:9090
        - Prom-2:9090
```

High Availability

As simple as running duplicate prometheis with the same configuration.

- ~~Consistency~~

- Availability

- Partition Tolerance

caicloud
才云

谢谢大家！