
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023

Build A High Performance Remote Storage for Prometheus with Unlimited Time Series

Yang Xiang (向阳)
Yunshan Networks (云杉网络)

TSDB: Time Series DataBase

Prometheus Samples

```
cpu_total{instance=1.2.3.4:80, job=k8s-pod, cluster=prod, cpu=1, state=system} 83 1234567890
```

Metric Name + Lables

Metric Value & Timestamp

Prometheus

High Cardinality Problem

Cardinality =~ N(metrics_name) * N(label_1_values) * N(label_2_values) * .....

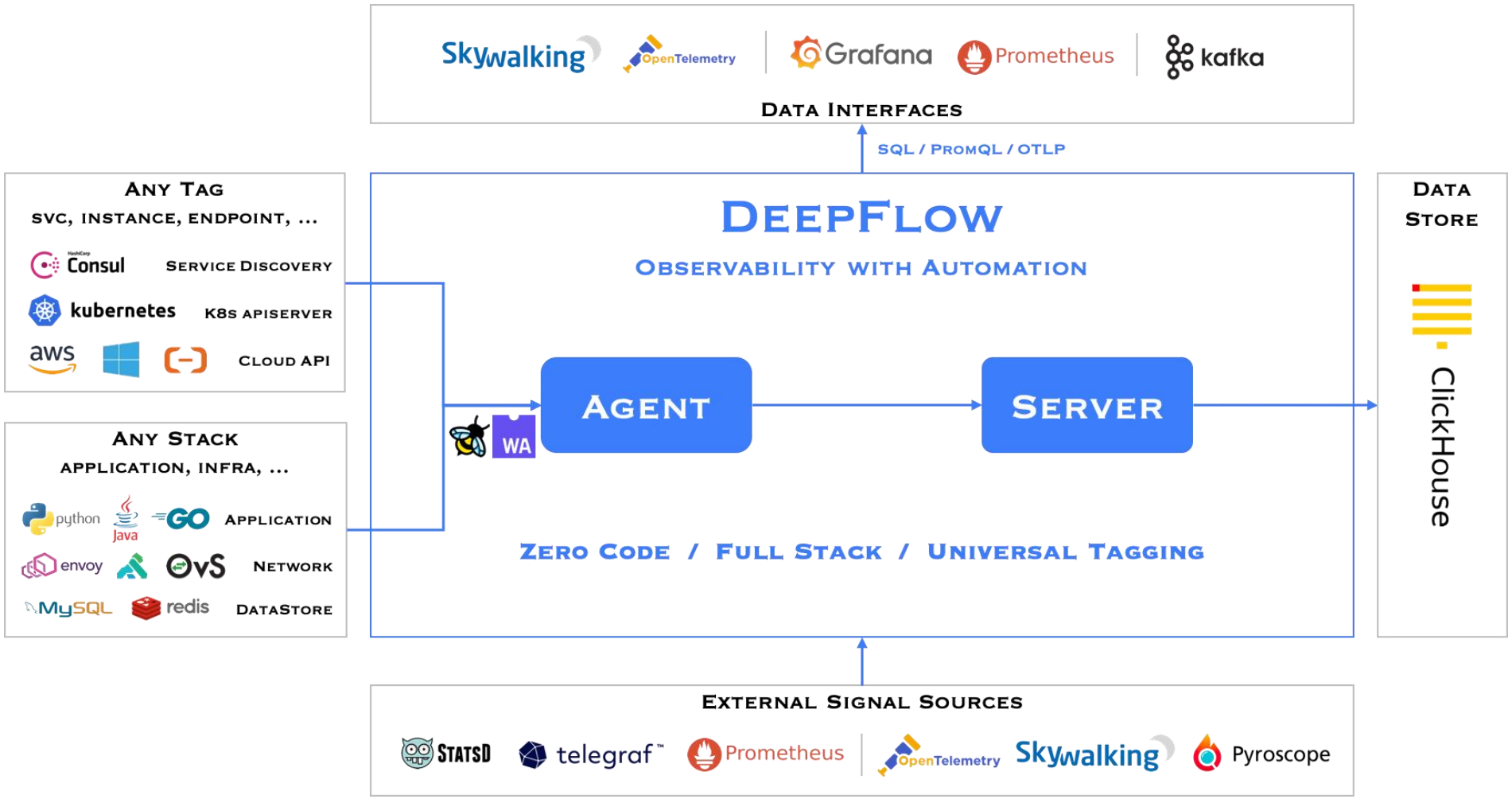
N(Series) = 20M, Label Values:

name	c	max(value)
id	172878	ygc
container_id	169890	docker://ffffdbc070dc1e00000db448390c605c19e7b9e625741f2490624b4c8c174185
name	144573	x01Vehicle
mountpoint	63361	/var/lib/kubelet/pods/fe02be86-35a9-4a6e-bd4d-e900000fbd33/volumes/kubernetes.io~projected/kube-api-access-x52sh
uid	46462	fffd733b-bbd4-421b-9afa-8885b282c846
pod	38254	yearlyreport-58d9849b-5rbcl
pod_name	28501	yearlyreport-58d9849b-5rbcl
address	24578	fe:ff:f3:00:00:00
device	20703	xxxx_nodev:/inf-hdmap
pod_ip	20282	172.00.000.99
replicaset	20200	yearlyreport-c6b57fd94
uri	19837	root
path	19140	stats
route_path	16372	获取认证字符串,使用正式私钥签名
routeId	14606	service1
owner_name	13380	yearlyreport-58d9849b
routeUri	13151	webjars/springfox-swagger-ui/*
image	12438	xxxxxxxxcar.xxxxxxxcloudcr.com/xxxxxxx/wp-weixin-provider@sha256:f30536baf1bcb0d47e000009088baef914d5a8b251b77192e66811d5ad5c660e
instance	9629	prometheus.kube-system.svc:9093
created_by_name	8890	yearlyreport-58d9849b
cluster_ip	7690	None
podip	7330	10.28.99.56
podname	7256	xxx-recommend-api-bffb7b7d5-ltrgx
label_rollouts_pod_template_hash	7165	ffff9bbcc
resource	6707	xxxxolumes.zfs.openebs.io
image_id	6190	docker-pullable://xxxxxxxxcar.xxxxxxxcloudcr.com/xxxxxxx/wp-weixin-provider@sha256:f30536ba000000047eb97949088baef914d5a8b251b77192e66811d5ad5c660e
image_spec	6144	xxxxxxxxcar.xxxxxxxcloudcr.com/xxxxxxx/wp-weixin-provider:1.0.0-230725.1059-32d4dfed
filename	5848	vsp-recommend-api.log
host	5175	yearlyreport.dev.k8s.chj.cloud
endpoint	5012	yearlyreport

30 rows in set (1.40 sec)

Metrics Storage in DeepFlow

Metrics Lables: client_ip, server_ip, client_port, server_port, protocol, endpoint, url, process_name, pod, ...
We use ClickHouse as a long-term storage.



An Intuitive Idea


```
CREATE TABLE prometheus_samples
(
    `time` DateTime('Asia/Shanghai') CODEC(DoubleDelta),
    `metric_name` String,
    `tag_names` Array(String),
    `tag_values` Array(String),
    `value` Float64
)
```

```
SELECT tag_values[indexOf(tag_names, 'host')] AS `host`
FROM prometheus_samples
WHERE (tag_values[indexOf(tag_names, 'host')]) = 'some-host'
```

Too Many Lables, Very Slow Query

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2023

```
Query id: 78d4852c-b61c-4acc-a732-ba7dd8904e3c

Row 1:
──────
metric_name:     prometheus.container_memory_failures_total
any(tag_names):  ['container','failure_type','id','image','name','namespace','pod','scope','beta_kubernetes_io_arch','beta_kubernetes_io_instance_gpu','beta_kubernetes_io_instance_type','beta_kubernetes_io_os','cce_baidubce_com_kubelet_dir','cluster','cluster_id','cluster_role','failure_domain_beta_kubernetes_io_region','failure_domain_beta_kubernetes_io_zone','feature_node_kubernetes_io_cpu_cpuid_ADX','feature_node_kubernetes_io_cpu_cpuid_AESNI','feature_node_kubernetes_io_cpu_cpuid_AVX','feature_node_kubernetes_io_cpu_cpuid_AVX2','feature_node_kubernetes_io_cpu_cpuid_AVX512BITALG','feature_node_kubernetes_io_cpu_cpuid_AVX512BW','feature_node_kubernetes_io_cpu_cpuid_AVX512CD','feature_node_kubernetes_io_cpu_cpuid_AVX512DQ','feature_node_kubernetes_io_cpu_cpuid_AVX512F','feature_node_kubernetes_io_cpu_cpuid_AVX512IFMA','feature_node_kubernetes_io_cpu_cpuid_AVX512VBMI','feature_node_kubernetes_io_cpu_cpuid_AVX512VBMI2','feature_node_kubernetes_io_cpu_cpuid_AVX512VL','feature_node_kubernetes_io_cpu_cpuid_AVX512VNNI','feature_node_kubernetes_io_cpu_cpuid_AVX512VPOPCNTDQ','feature_node_kubernetes_io_cpu_cpuid_CMPXCHG8','feature_node_kubernetes_io_cpu_cpuid_FMA3','feature_node_kubernetes_io_cpu_cpuid_FXSR','feature_node_kubernetes_io_cpu_cpuid_FXSROPT','feature_node_kubernetes_io_cpu_cpuid_GFNI','feature_node_kubernetes_io_cpu_cpuid_HYPERVISOR','feature_node_kubernetes_io_cpu_cpuid_IBPB','feature_node_kubernetes_io_cpu_cpuid_LAHF','feature_node_kubernetes_io_cpu_cpuid_MOVBE','feature_node_kubernetes_io_cpu_cpuid_OSXSAVE','feature_node_kubernetes_io_cpu_cpuid_SCE','feature_node_kubernetes_io_cpu_cpuid_SHA','feature_node_kubernetes_io_cpu_cpuid_VAES','feature_node_kubernetes_io_cpu_cpuid_VMX','feature_node_kubernetes_io_cpu_cpuid_VPCLMULQDQ','feature_node_kubernetes_io_cpu_cpuid_WBNOINVD','feature_node_kubernetes_io_cpu_cpuid_X87','feature_node_kubernetes_io_cpu_cpuid_XSAVE','feature_node_kubernetes_io_cpu_hardware_multithreading','feature_node_kubernetes_io_cpu_model_family','feature_node_kubernetes_io_cpu_model_id','feature_node_kubernetes_io_cpu_model_vendor_id','feature_node_kubernetes_io_kernel_config_NO_HZ','feature_node_kubernetes_io_kernel_config_NO_HZ_FULL','feature_node_kubernetes_io_kernel_version_full','feature_node_kubernetes_io_kernel_version_major','feature_node_kubernetes_io_kernel_version_minor','feature_node_kubernetes_io_kernel_version_revision','feature_node_kubernetes_io_pci_0300_1013_present','feature_node_kubernetes_io_system_os_release_ID','feature_node_kubernetes_io_system_os_release_VERSION_ID','feature_node_kubernetes_io_system_os_release_VERSION_ID_major','instance','instance_group_id','job','kubernetes_io_arch','kubernetes_io_hostname','kubernetes_io_os','kubernetes_io_tor','node_kubernetes_io_instance_type','topology_kubernetes_io_region','topology_kubernetes_io_zone','container_name','pod_name','receive_replica','role']
any(tag_values): [
f
6c
'true'
'kubernetes
]
tag_count:       79
```

Too Many Columns?

Number of columns is not limited explicitly. 1000 columns will work Ok.

Jun 28, 2016

Google Groups
https://groups.google.com › clickhouse

Are there any limit on number of columns in ClickHouse

But actually we don't need so many columns,
this will be a very sparse table.

Align Columns by Metric Name

One metric usually has no more than a few hundred labels

```
CREATE TABLE prometheus_samples
(
    `time` DateTime('Asia/Shanghai') CODEC(DoubleDelta),
    `metric_name` String,
    `label_name_1` String,
    `label_name_2` String,
    // ...
    `label_name_256` String, // Assume that metric will not have
                             // more than 256 labels
    `value` Float64
)

SELECT host, value
FROM prometheus_samples
WHERE host = 'some-host'
```

Can Columns be Reduced Further?

app labels

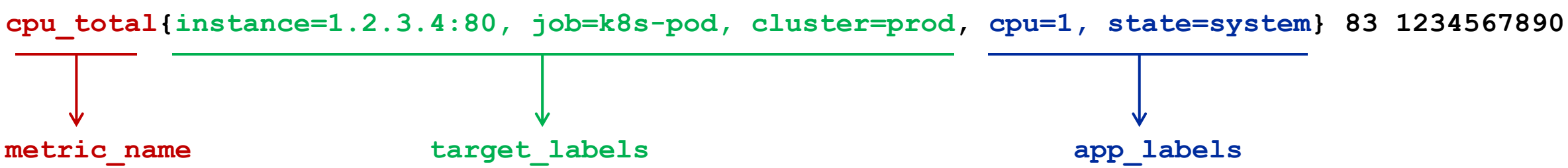
app
owner
service
namespace
endpoint
...

Pods are often rebuilt

infra labels

pod
node
cluster
env
region
...

Distinguish Target Labels an App Lables



Cardinality(all series) = 20M

v.s.

N(targets) = 14K
Cardinality(app labels) = 1M

Distinguish Target Labels an App Lables


```
CREATE TABLE prometheus_samples
(
    `time` DateTime('Asia/Shanghai') CODEC(DoubleDelta),
    `metric_name` String,
    `target_id` Uint32,
    `label_name_1` String,
    // ...
    `label_name_128` String, // Assume that metric will not have
                             // more than 128 app labels
    `value` Float64
)
```

Can Columns be Reduced Further? / 2

meta label

pod

container

```
apiVersion: v1
items:
- apiVersion: v1
  kind: Pod
  metadata:
    annotations:
      cni.projectcalico.org/containerID: b85f57e1d0a98d274293cefd4b9...
      cni.projectcalico.org/podIP: 100.89.13.129/32
      cni.projectcalico.org/podIPs: 100.89.13.129/32
    labels:
      k8s-app: calico-kube-controllers
      pod-template-hash: 57fbd7bd59
  spec:
    containers:
    - env:
      - name: KUBE_CONTROLLERS_CONFIG_NAME
        value: default
      - name: DATASTORE_TYPE
        value: kubernetes
      - name: ENABLED_CONTROLLERS
        value: node
      - name: KUBERNETES_SERVICE_HOST
        value: 10.96.0.1
      - name: KUBERNETES_SERVICE_PORT
        value: "443"
```

Meta Labels & Unlimited Custom Labels


```
CREATE TABLE prometheus_samples
(
    `time` DateTime('Asia/Shanghai') CODEC(DoubleDelta),
    `metric_name` String,
    `target_id` Uint32,
    `pod` String,
    `container` String,
    `label_name_1` String,
    // ...
    `label_name_16` String, // Assume that metric will not have
                            // more than 16 labels injected by instrumentation
    `value` Float64
)
```

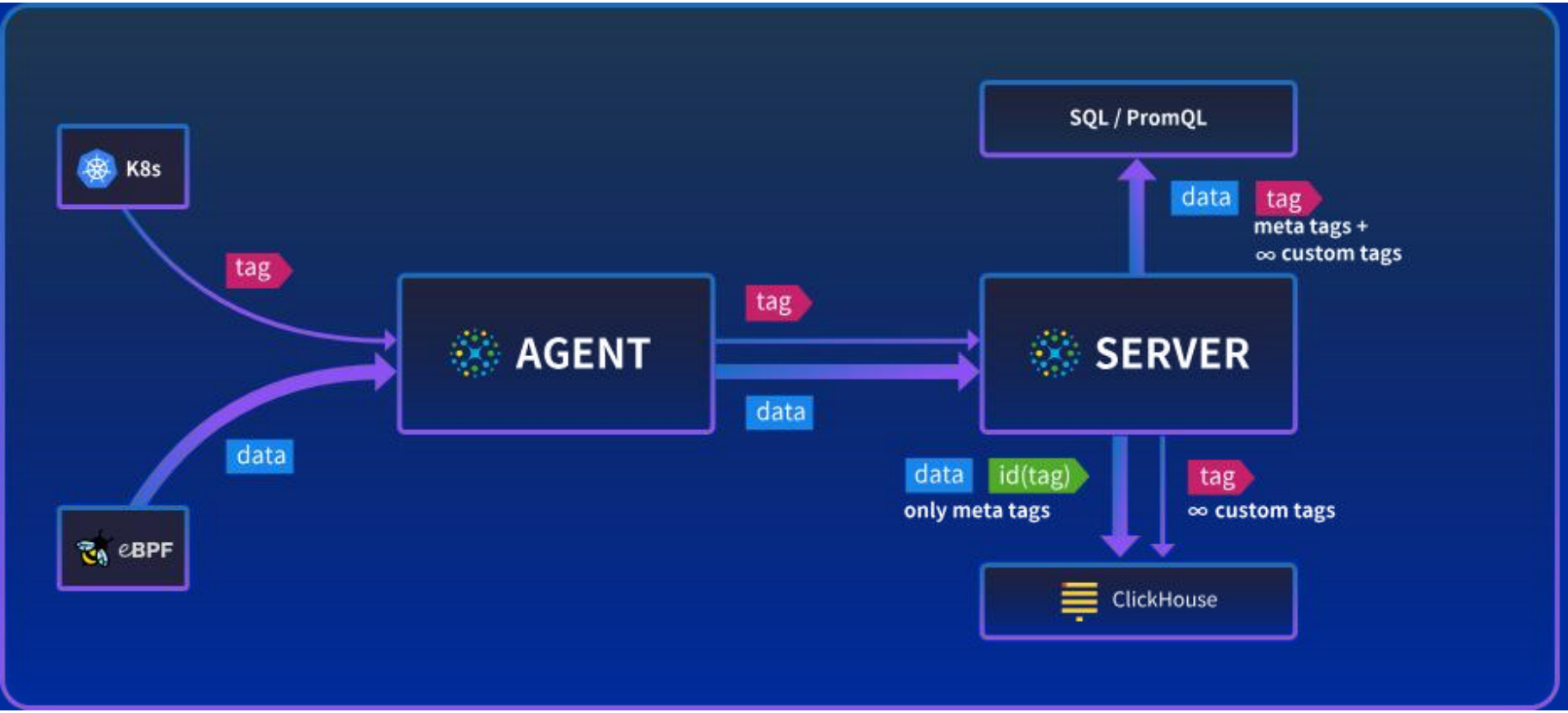
Query Custom Lables

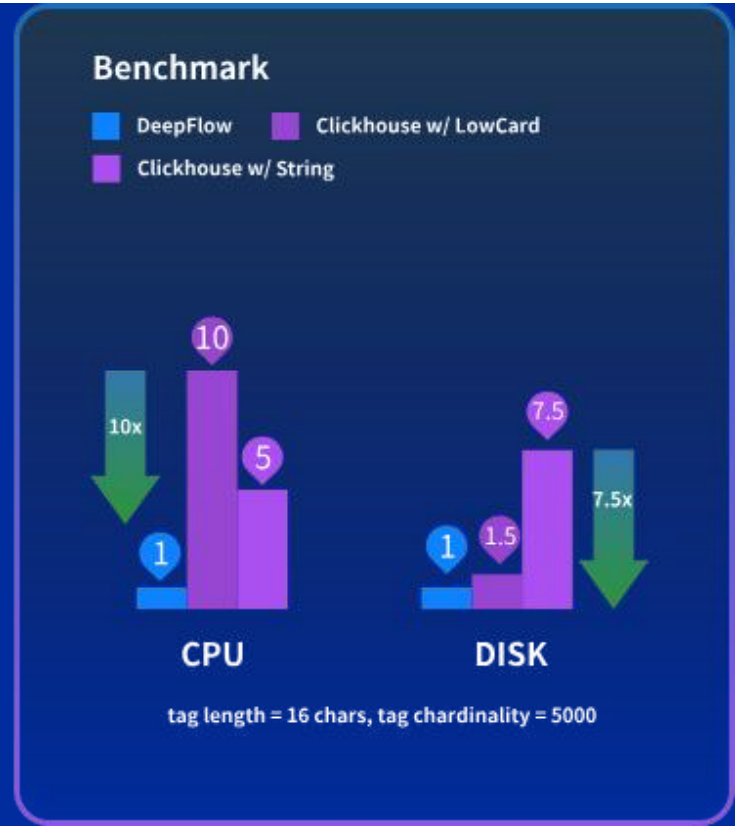

```
CREATE TABLE prometheus_samples
(
    `time` DateTime('Asia/Shanghai') CODEC(DoubleDelta),
    `metric_name` String,
    `target_id` Uint32,
    `pod` String,
    `container` String,
    `label_name_1` String,
    // ...
    `label_name_16` String, // Assume that metric will not have
                            // more than 16 custom labels (by instrumentation)
    `value` Float64
)
```

```
CREATE DICTIONARY custom_labels.k8s_label
(
    `pod_id` Uint64,
    `key` String,
    `value` String,
)
PRIMARY KEY pod, key
LIFETIME(MIN 0 MAX 60)
LAYOUT(COMPLEX_KEY_HASHED())
```

```
SELECT dictGet(
    custom_labels.k8s_label,
    'value',
    (pod_id, 'app')
) AS `app`
FROM `cpu_total`
WHERE `app` = 'some-app'
LIMIT 1
```

SmartEncoding in DeepFlow





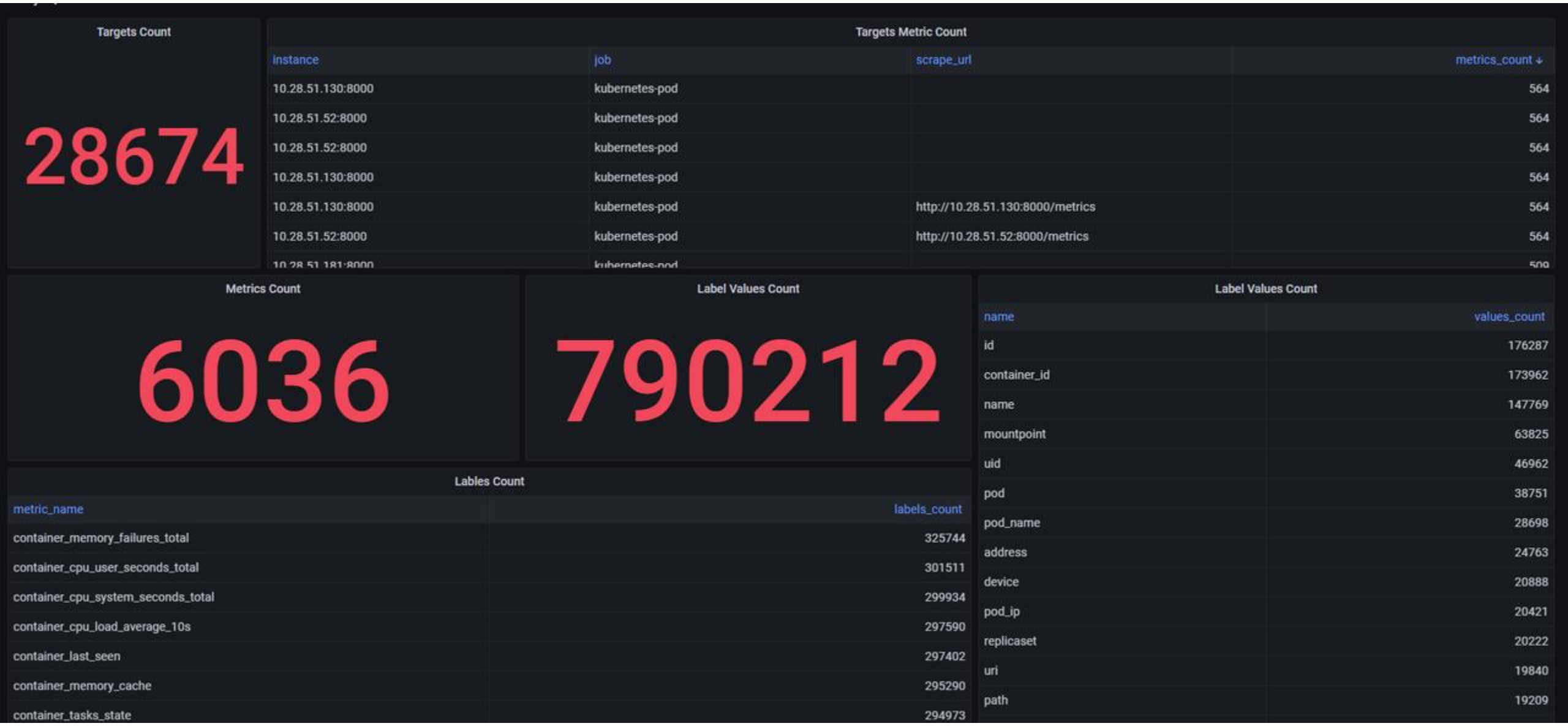
Benchmark (v.s. VictoriaMetrics)

Similar write resource consumption to VictoriaMetrics,
but supports unlimited labels and has
an order of magnitude better query performance when querying high-cardinality series.

	总资源 [1]	压测参数 [2]	读写压力	CPU 均值	CPU 峰值	内存均值	内存峰值
DeepFlow (DF-Server 及 CK)	**24** U **96** GB	targetCount = **4200** scrapeInterval = 15s queryInterval = **60**s	367k samples/s 5.8M series 1.5 QPS	12.0 U	15.3 U	12.1 GB	18.2 GB
VictoriaMetrics (不含 VMInsert)	44 U 208 GB	targetCount = 6000 scrapeInterval = 15s queryInterval = 15s	367k samples/s 5.8M series 1.5 QPS	12.1 U	13.8 U	14.8 GB	18.6 GB
Mimir	43.2 U 283 GB	targetCount = 6000 scrapeInterval = 15s queryInterval = 15s	367k samples/s 5.8M series 1.5 QPS	20.3 U	24.1 U	102 GB	120 GB

Benchmark (one of our prod. env)

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2023



Targets Count

28674

Targets Metric Count

instance	job	scrape_url	metrics_count ↓
10.28.51.130:8000	kubernetes-pod		564
10.28.51.52:8000	kubernetes-pod		564
10.28.51.52:8000	kubernetes-pod		564
10.28.51.130:8000	kubernetes-pod		564
10.28.51.130:8000	kubernetes-pod	http://10.28.51.130:8000/metrics	564
10.28.51.52:8000	kubernetes-pod	http://10.28.51.52:8000/metrics	564
10.28.51.181:8000	kubernetes-pod		509

Metrics Count

6036

Label Values Count

790212

Label Values Count

name	values_count
id	176287
container_id	173962
name	147769
mountpoint	63825
uid	46962
pod	38751
pod_name	28698
address	24763
device	20888
pod_ip	20421
replicaset	20222
uri	19840
path	19209

Lables Count

metric_name	labels_count
container_memory_failures_total	325744
container_cpu_user_seconds_total	301511
container_cpu_system_seconds_total	299934
container_cpu_load_average_10s	297590
container_last_seen	297402
container_memory_cache	295290
container_tasks_state	294973

Benchmark (one of our prod. env)

400K samples/s ingestion

CPU: 740%~1100% (3200% Total), MEM: 29G~37G (128GB Total) // only deepflow-server

Benchmark (one of our prod. env)

+1500K QPS
CPU: +600% // only deepflow-server

Thanks !

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2023

OpenSourced under the Apache 2.0 License
https://github.com/deepflowio/deepflow
https://deepflow.io

Our WeChat Group
Discord? Check our GitHub repo!