DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

可观测数据融合处理平台探索

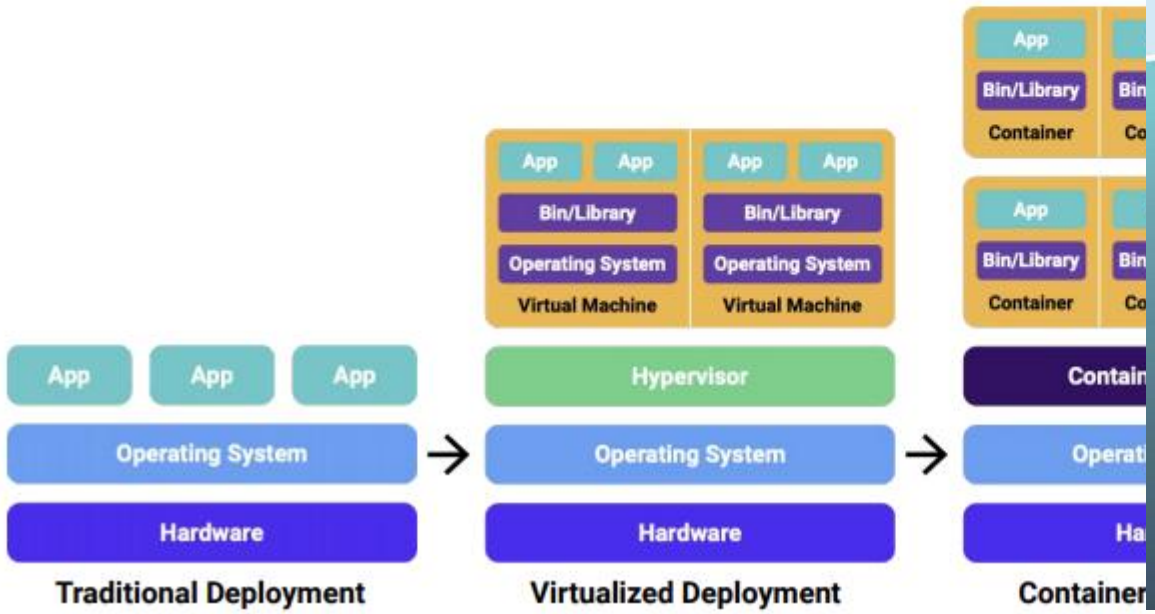
唐恒建

OPPO 高级后端工程师

- 1、可观测要素
- 2、opentelemetry现状
- 3、各类可观测方案对比
- 4、可观测数据融合平台
- 5、总结

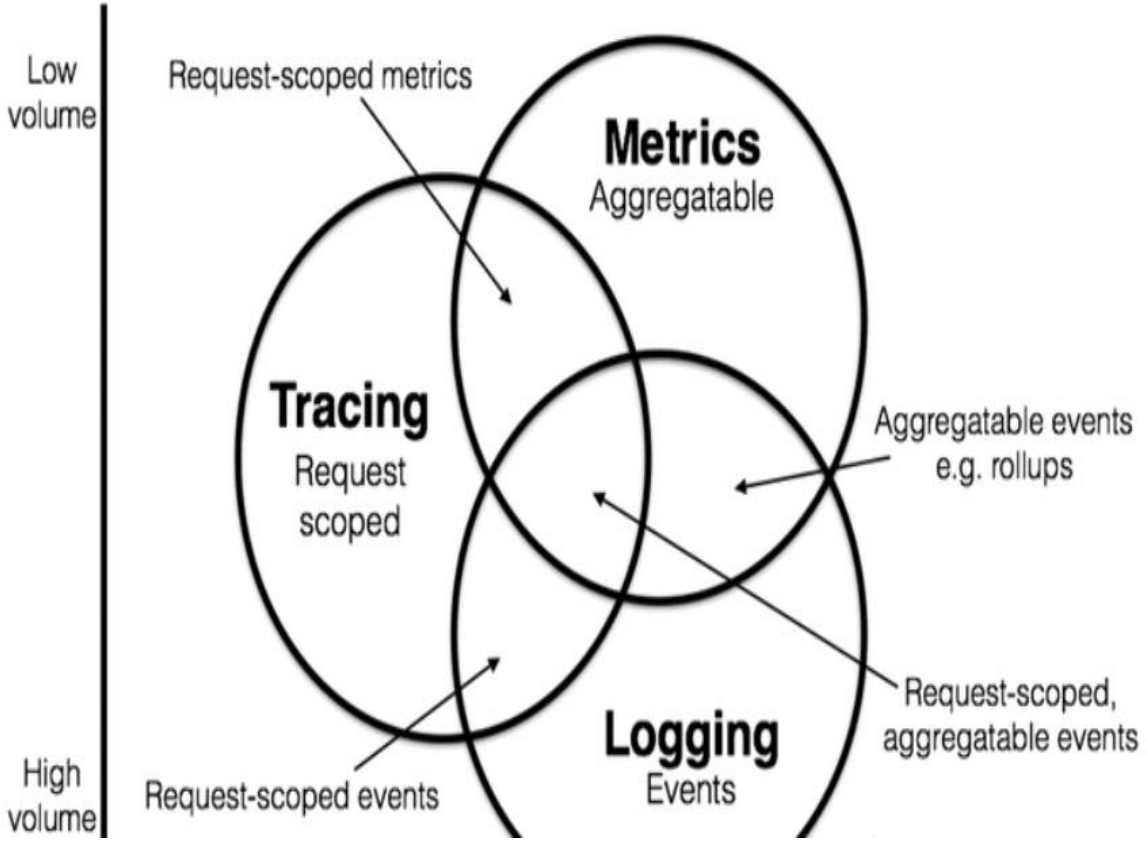
可观测要素

Increasingly complex software deployments

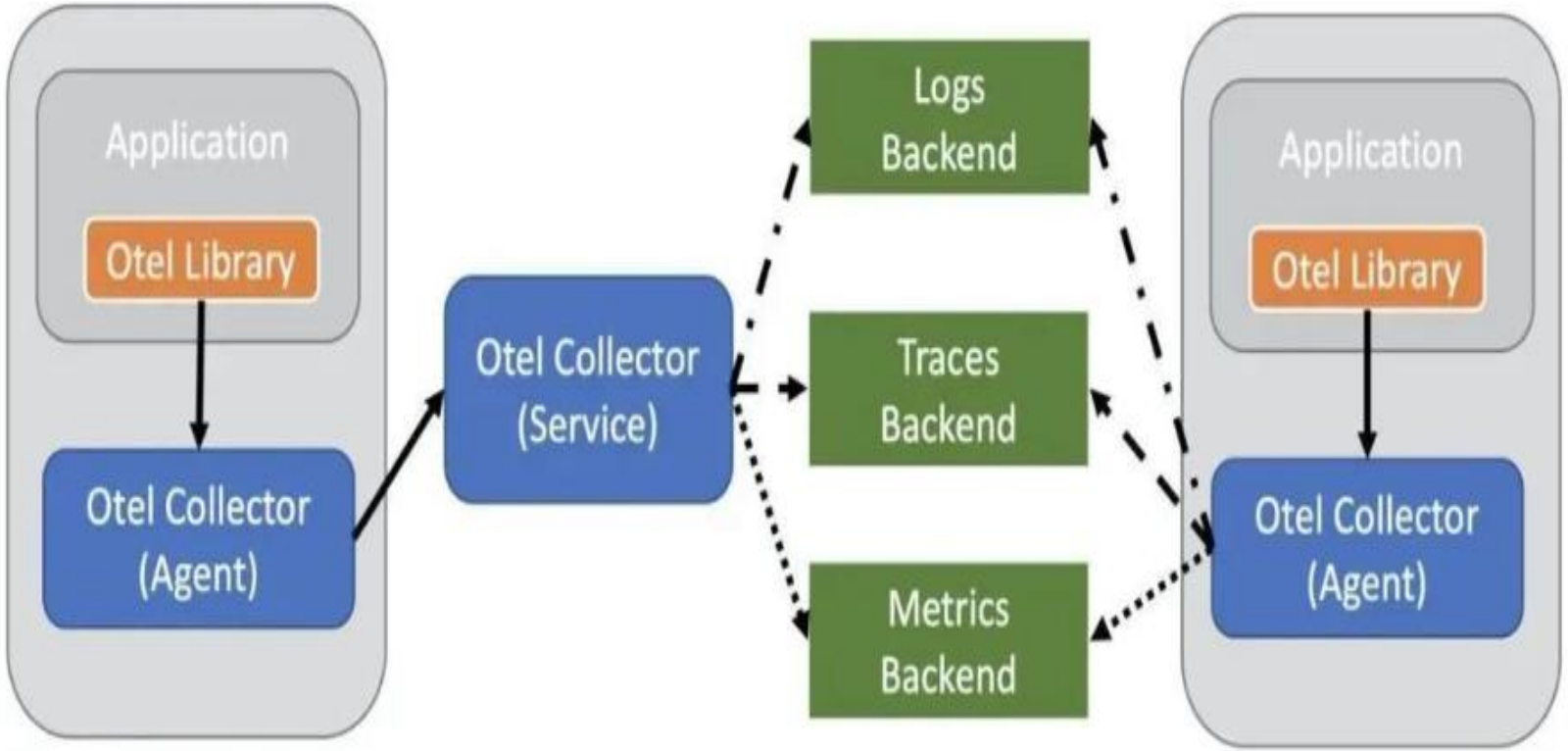


可观测要素


Network monitoring 56.6%
Security monitoring 55.7%
Database monitoring 53.8%
Alerts 51.5%
Infrastructure monitoring 51.0%
Log management 50.3%
Browser monitoring 48.6%
Error tracking 48.5%
Custom dashboards 46.7%
APM 45.4%
Mobile monitoring 43.3%
Serverless monitoring 38.2%
AIOps 37.3%
Distributed tracing 36.3%
Synthetic monitoring 34.2%
ML model performance monitoring 34.1%
Kubernetes monitoring 33.7%
0%
20.0%
40.0%
60.0%


Low volume
High volume
Request-scoped metrics
Metrics
Aggregatable
Tracing
Request scoped
Aggregatable events e.g. rollups
Request-scoped, aggregatable events
Logging
Events
Request-scoped events

opentelemetry现状



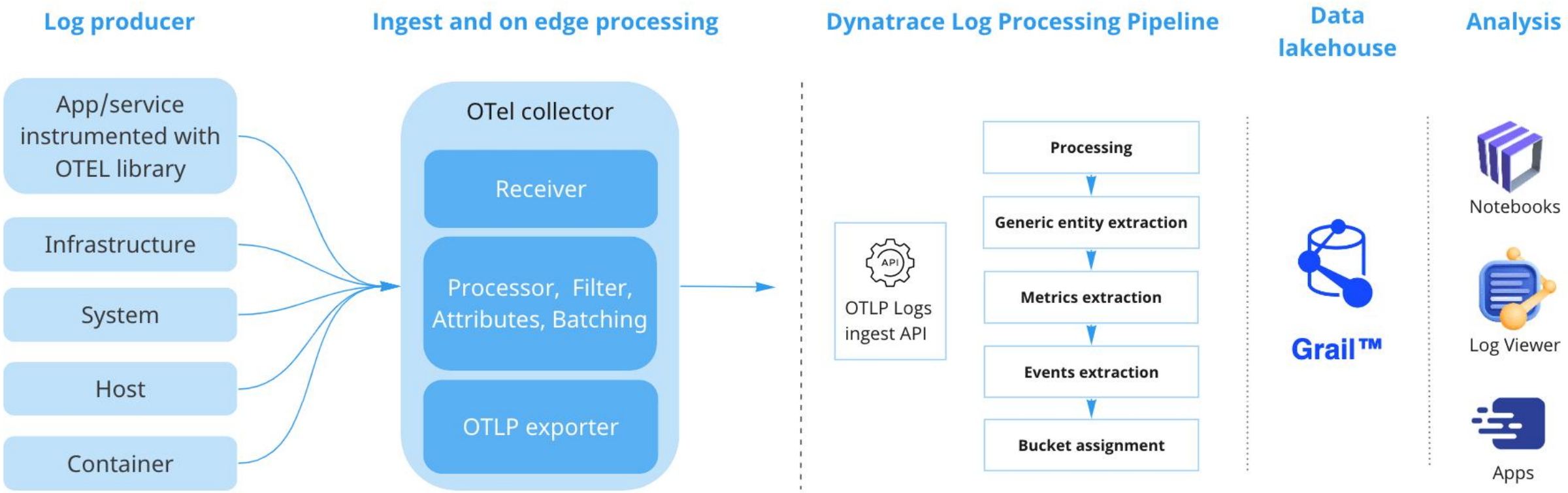
- 维度：metric、trace、log
- 语言：JIT：Java、Rust、Python
  非JIT类：C++、Go
- agent、collect（batch）
- Multi protocol: receiver、exporter

opentelemetry现状

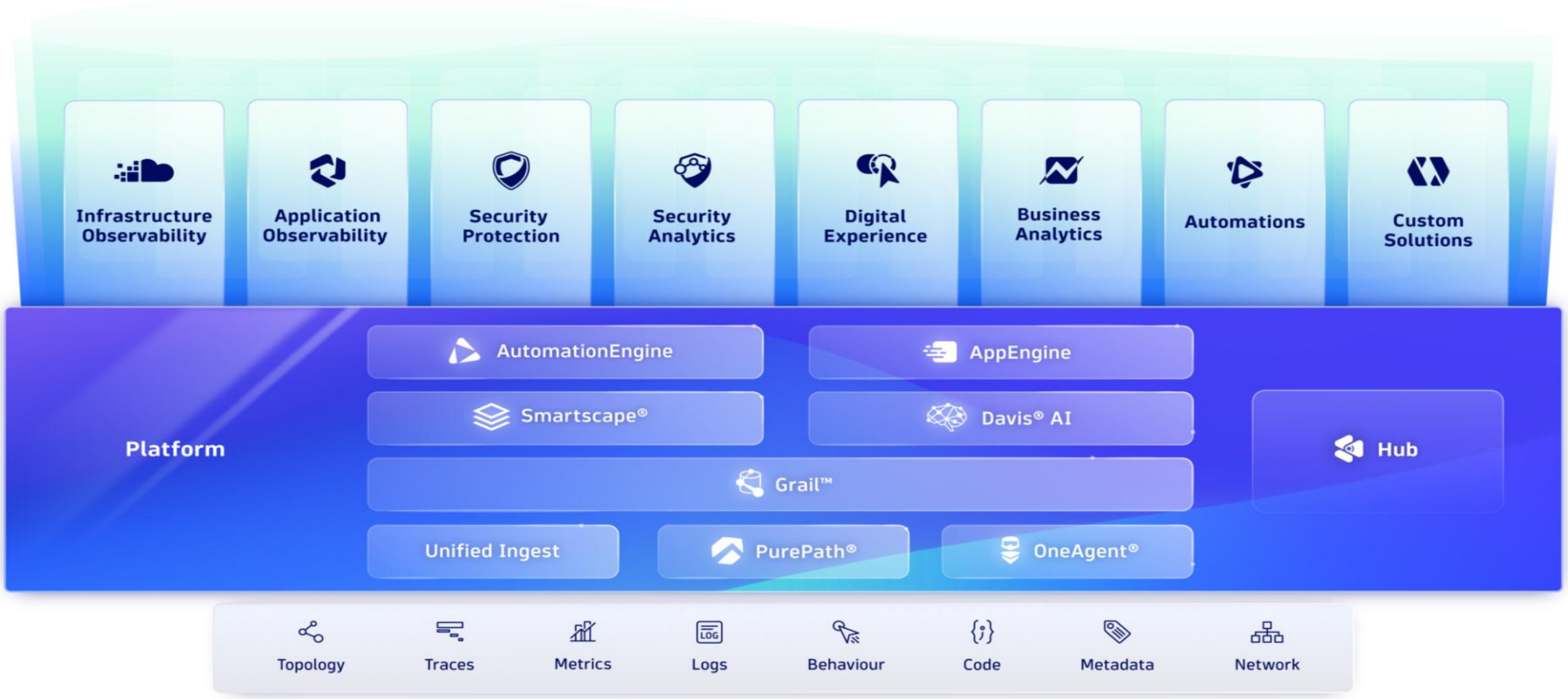

```
receivers:
  otlp:
    protocols:
      grpc:
      http:

  prometheus:
    config:
      scrape_configs:
      - job_name: 'app'
        scrape_interval: 10s
        static_configs:
        - targets: ['app:8080']

exporters:
  otlp:
    endpoint: [redacted]4317
    tls:
      insecure: true

  prometheusremotewrite:
    endpoint: http://[redacted]:8080/api/v1/push
    tls:
      insecure: true
    headers:
      X-Scope-OrgID: demo
```

各类可观测方案对比

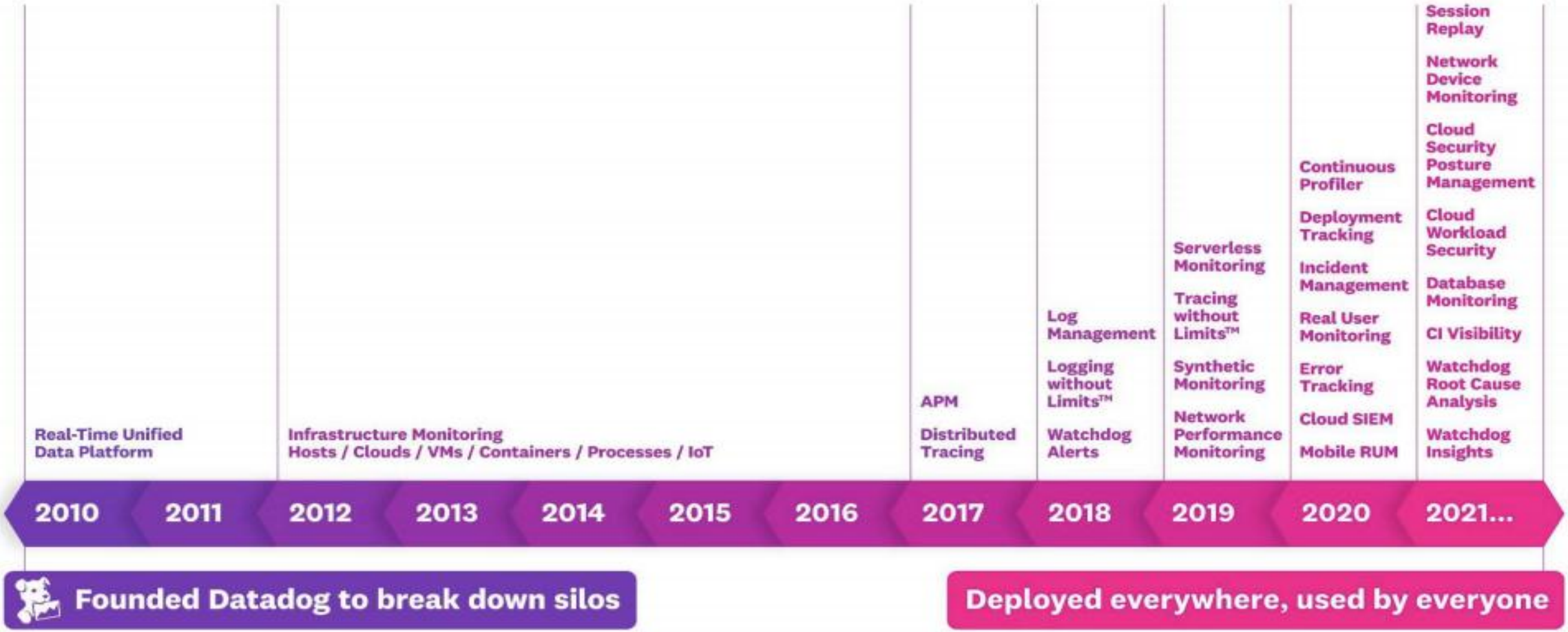
- Dynatrace



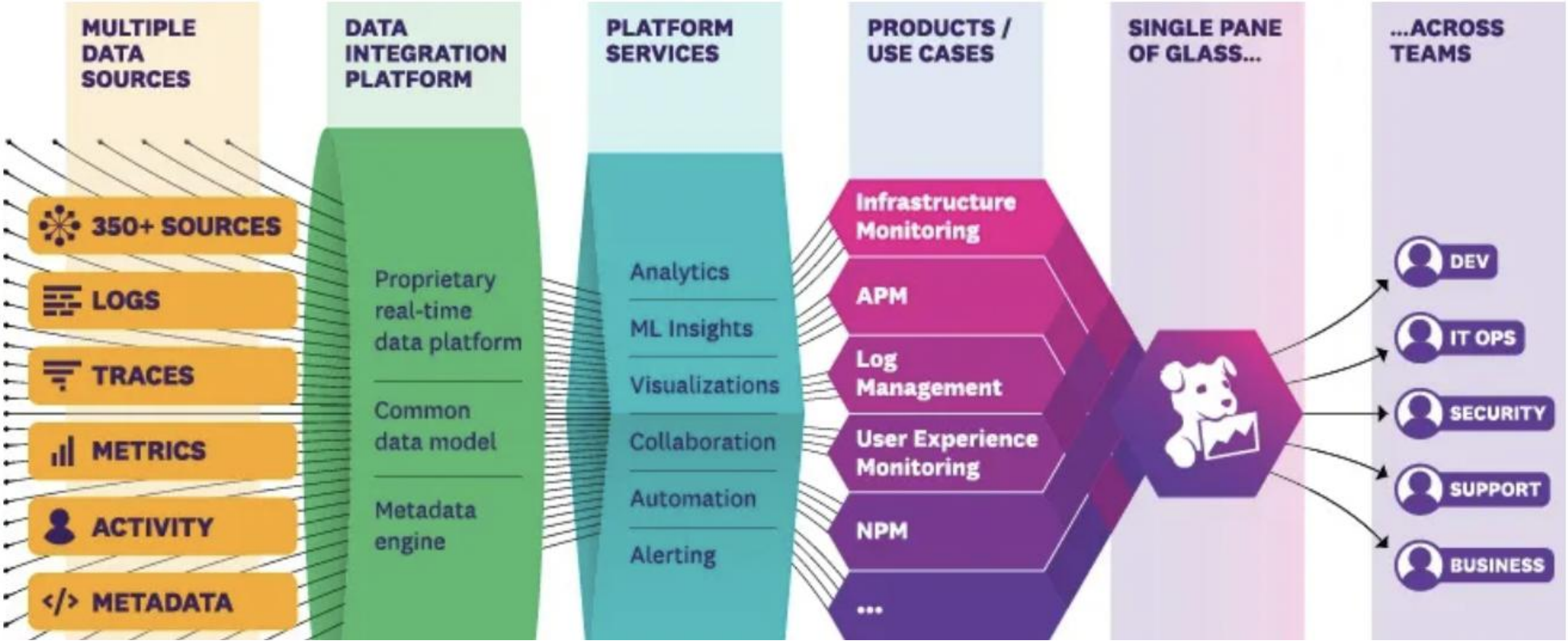
- Dynatrace


Infrastructure Observability
Application Observability
Security Protection
Security Analytics
Digital Experience
Business Analytics
Automations
Custom Solutions
Platform
AutomationEngine
AppEngine
Smartscape®
Davis® AI
Grail™
Hub
Unified Ingest
PurePath®
OneAgent®
Topology
Traces
Metrics
Logs
Behaviour
Code
Metadata
Network

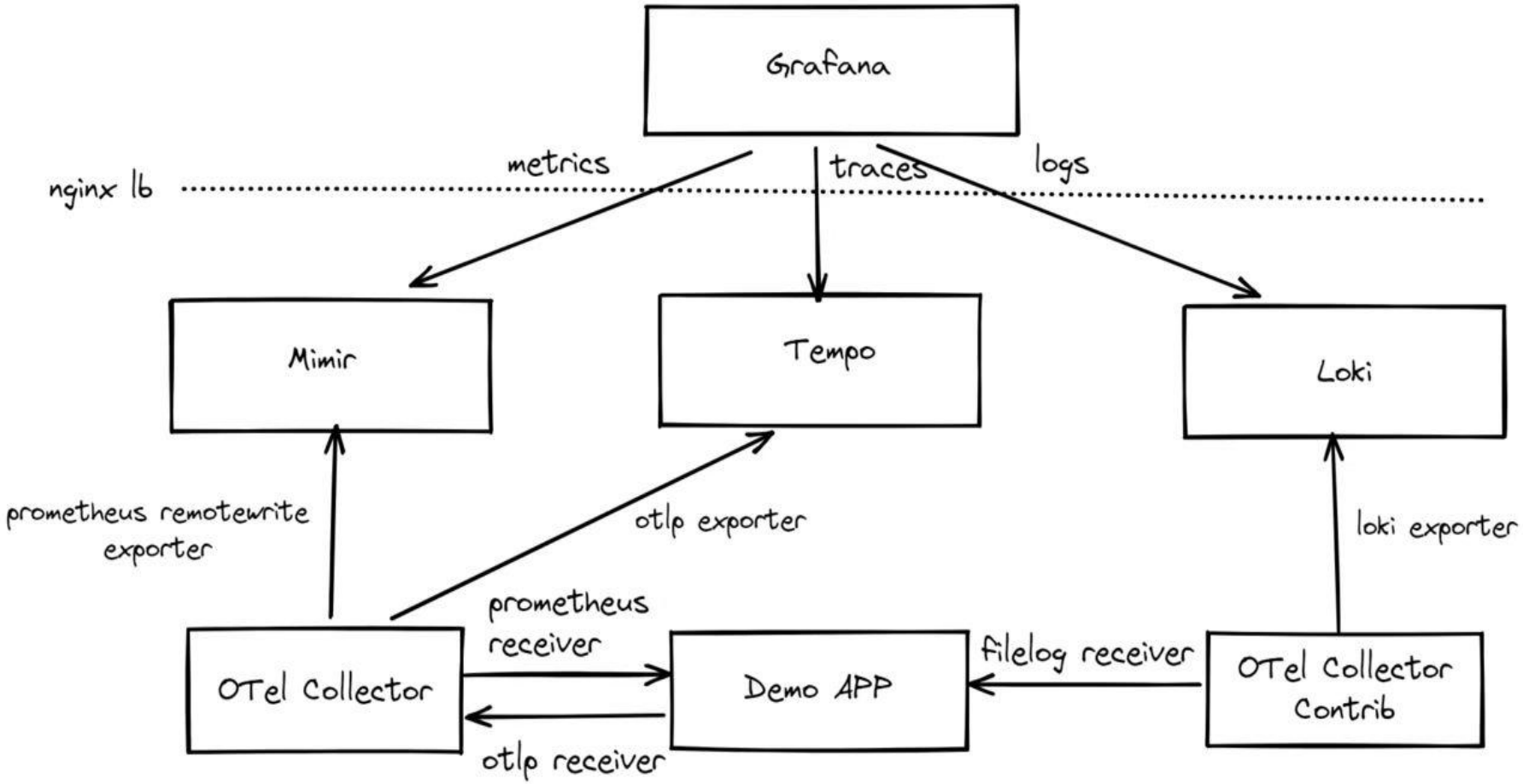
- Datadog


Real-Time Unified Data Platform
Infrastructure Monitoring
Hosts / Clouds / VMs / Containers / Processes / IoT
APM
Distributed Tracing
Log Management
Logging without Limits™
Watchdog Alerts
Serverless Monitoring
Tracing without Limits™
Synthetic Monitoring
Network Performance Monitoring
Continuous Profiler
Deployment Tracking
Incident Management
Real User Monitoring
Error Tracking
Cloud SIEM
Mobile RUM
Session Replay
Network Device Monitoring
Cloud Security Posture Management
Cloud Workload Security
Database Monitoring
CI Visibility
Watchdog Root Cause Analysis
Watchdog Insights
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021...
Founded Datadog to break down silos
Deployed everywhere, used by everyone

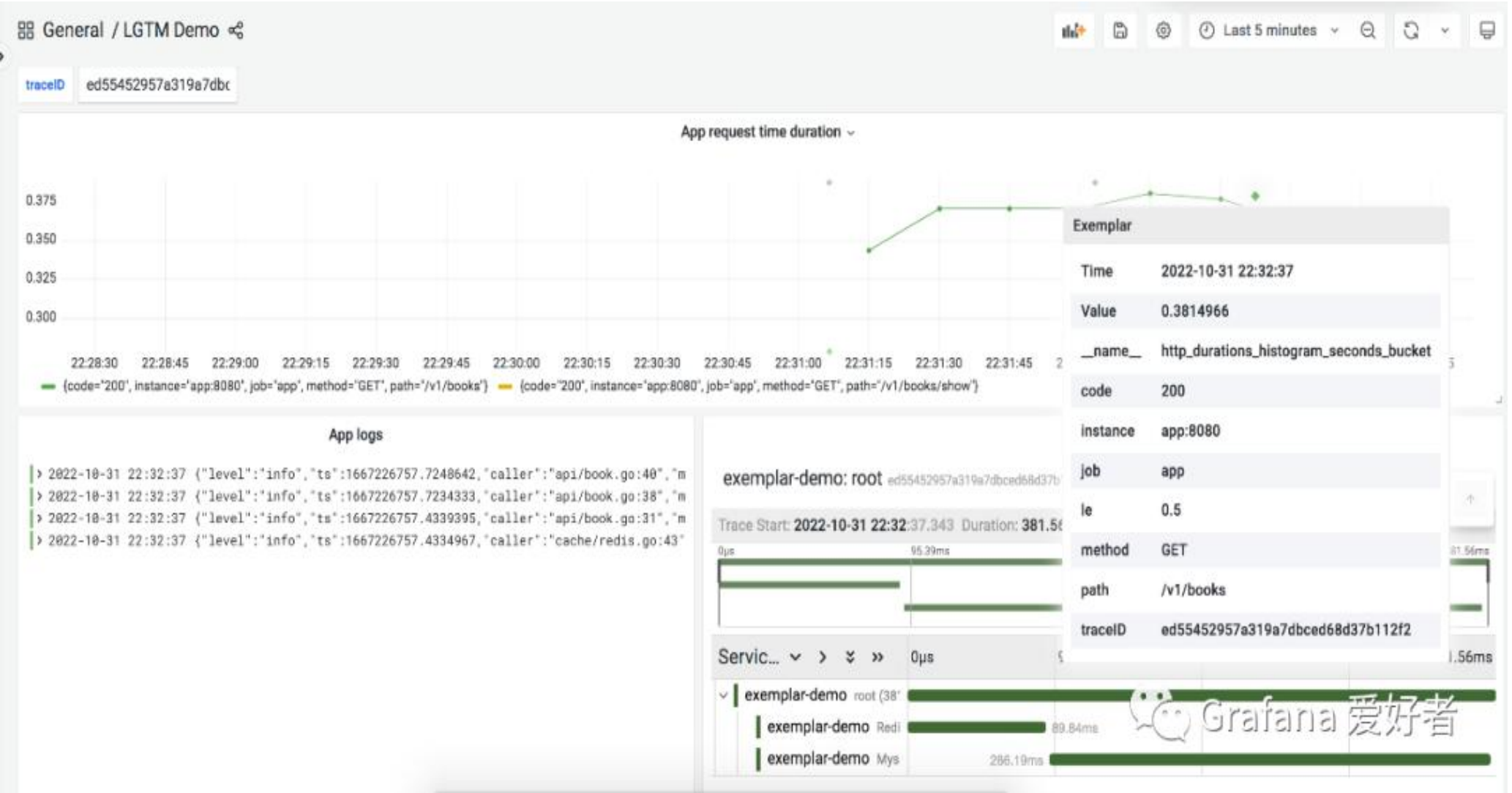
- Datadog


MULTIPLE DATA SOURCES
DATA INTEGRATION PLATFORM
PLATFORM SERVICES
PRODUCTS / USE CASES
SINGLE PANE OF GLASS...
...ACROSS TEAMS
350+ SOURCES
LOGS
TRACES
METRICS
ACTIVITY
METADATA
Proprietary real-time data platform
Common data model
Metadata engine
Analytics
ML Insights
Visualizations
Collaboration
Automation
Alerting
Infrastructure Monitoring
APM
Log Management
User Experience Monitoring
NPM
...
DEV
IT OPS
SECURITY
SUPPORT
BUSINESS

- Grafana

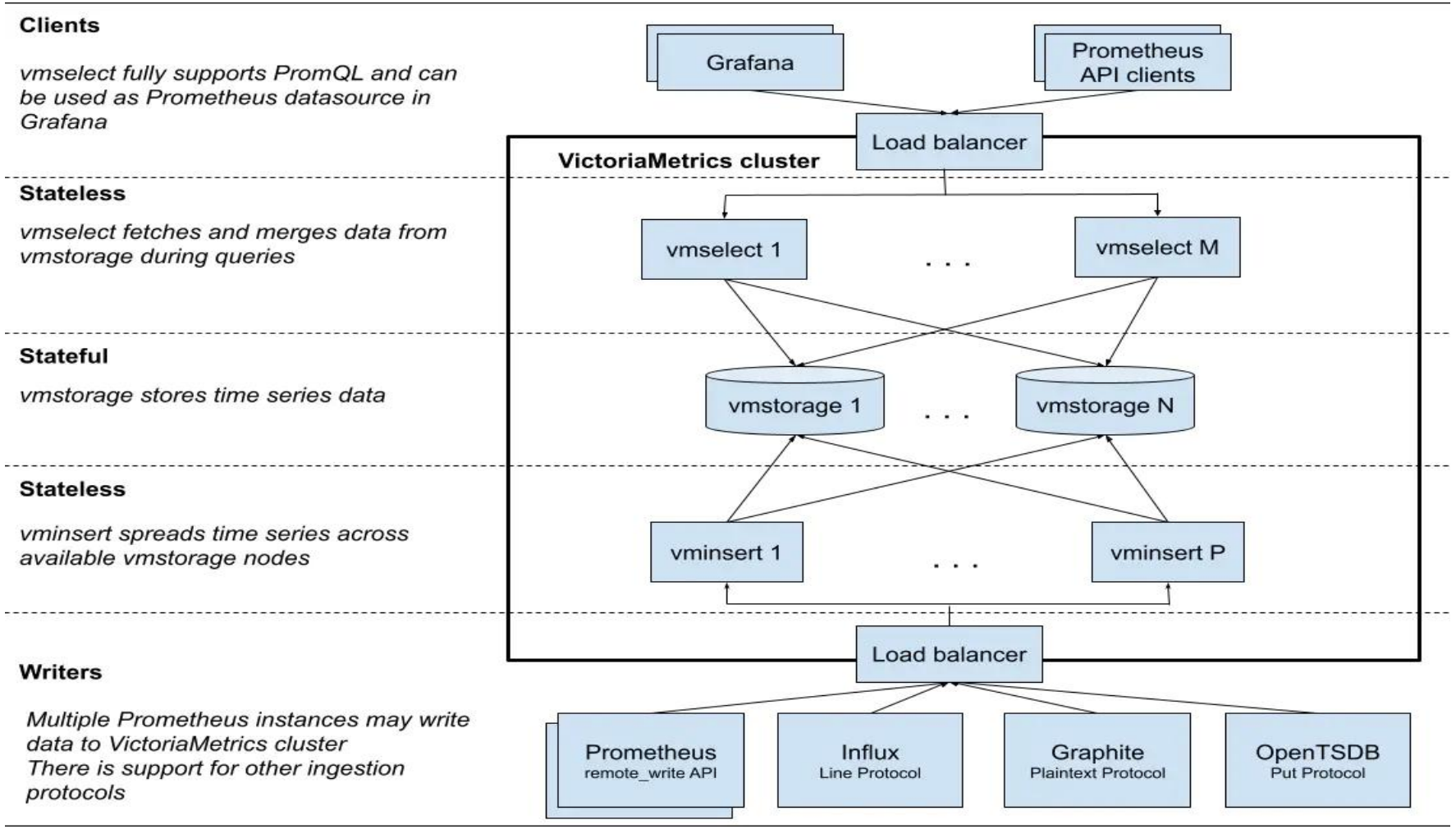

Grafana
metrics
traces
logs
nginx lb
Mimir
Tempo
Loki
prometheus remotewrite exporter
otlp exporter
loki exporter
prometheus receiver
filelog receiver
OTel Collector
Demo APP
OTel Collector Contrib
otlp receiver

- Grafana



- 以exemplar为连接
- 联动metrics、log、trace
- Metrics->logs
- metrics->trace

- Victoriametrics



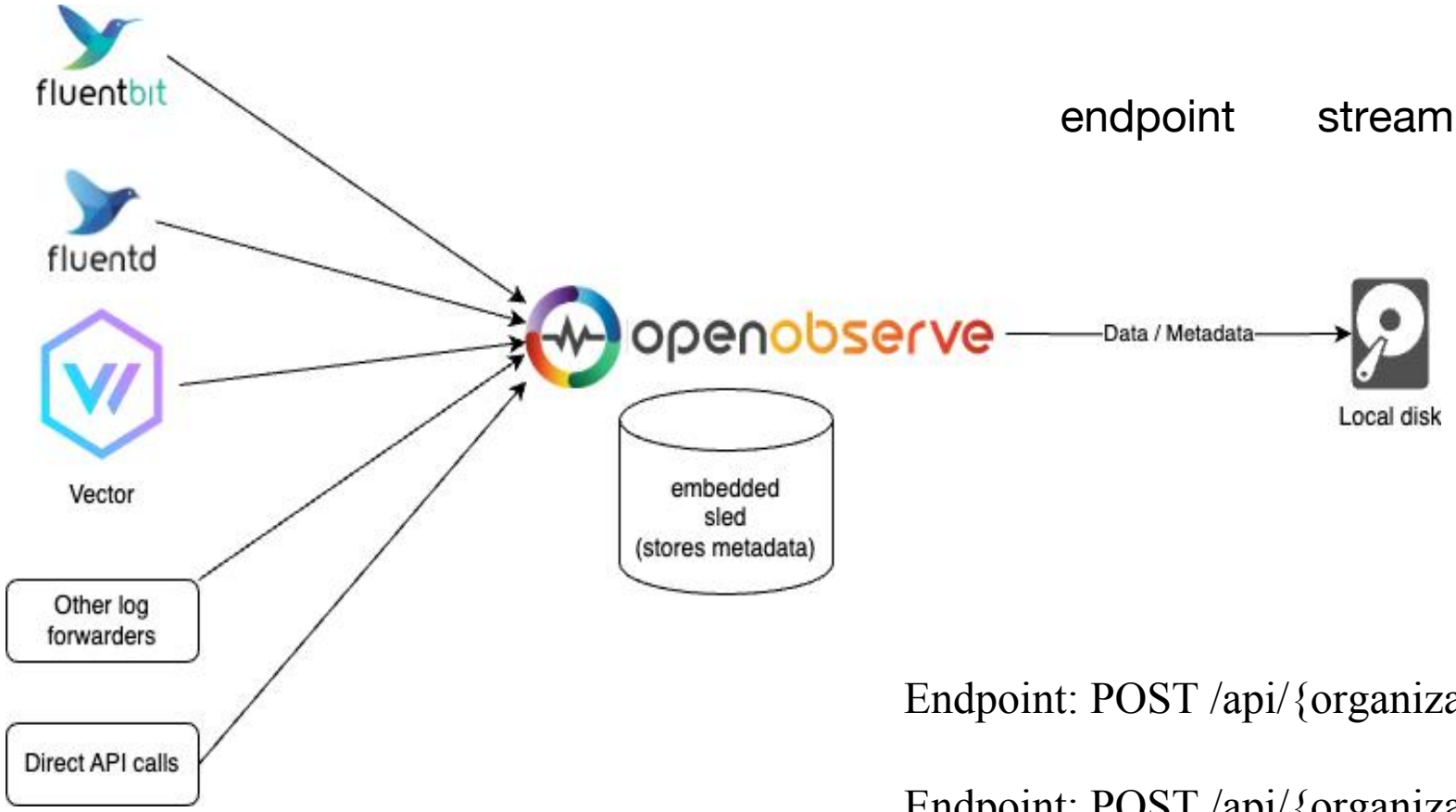
- Less CPU Usage
- Less Disk Usage

- Victoriametrics

VictoriaLogs --- Preview stage

- inspired by ClickHouse architecture
- It uses bloom filters
- encoding and compression for fields with different data types
- logs for the same log stream close to each other.
- maintains sparse index for log timestamps
- high cardinalitymetrics added.

- Openobserve



endpoint stream kvs parquet

Endpoint: POST /api/{organization}/{stream}/_json

Endpoint: POST /api/{organization}/prometheus/api/v1/write

Endpoint: POST /api/{organization}/traces

- Openobserve

```
wal
├── file_list
│   └── 0_2023_08_12_09_7096057797229543424rrz203.json
└── files
    ├── default
    │   └── logs
    │       └── default
    └── org_name
        └── metrics
            ├── avalanche_metric_mmmm_0_0
            │   └── 0_2023_08_12_09_keeping_7096057408967016450 1TZ932.json
            ├── avalanche_metric_mmmm_0_1
            │   └── 0_2023_08_12_09_keeping_709605740895443353 6XxJWc6.json
            ├── avalanche_metric_mmmm_0_10
            │   └── 0_2023_08_12_09_keeping_7096057408987987969D7p6RX.json
```


```
{"key":"files/org_name/metrics/scrape_duration_seconds/2023/08/12/09/7096057797225349120OGLIuf.parquet","meta":
{"min_ts":1691831406600000,"max_ts":1691831997003000,"records":44,"original_size":9412,"compressed_size":3767},
"deleted":false}
{"key":"files/org_name/metrics/scrape_samples_post_metric_relabeling/2023/08/12/09/7096057797221154816eI9FRP.pa
rquet","meta":{"min_ts":1691831406600000,"max_ts":1691831997003000,"records":44,"original_size":9686,"compresse
d_size":3516},"deleted":false}
{"key":"files/org_name/metrics/scrape_series_added/2023/08/12/09/7096057797246320640UNBTLu.parquet","meta":{"mi
n_ts":1691831406600000,"max_ts":1691831997003000,"records":44,"original_size":8890,"compressed_size":3353},"del
eted":false}
{"key":"files/org_name/metrics/scrape_samples_scraped/2023/08/12/09/7096057797258903552MJsMFv.parquet","meta":{
"min_ts":1691831406600000,"max_ts":1691831997003000,"records":44,"original_size":9026,"compressed_size":3380},"
deleted":false}
{"key":"files/org_name/metrics/up/2023/08/12/09/7096057797271486464RMqg56.parquet","meta":{"min_ts":16918314066
00000,"max_ts":1691831997003000,"records":44,"original_size":8140,"compressed_size":3197},"deleted":false}
```


WAL

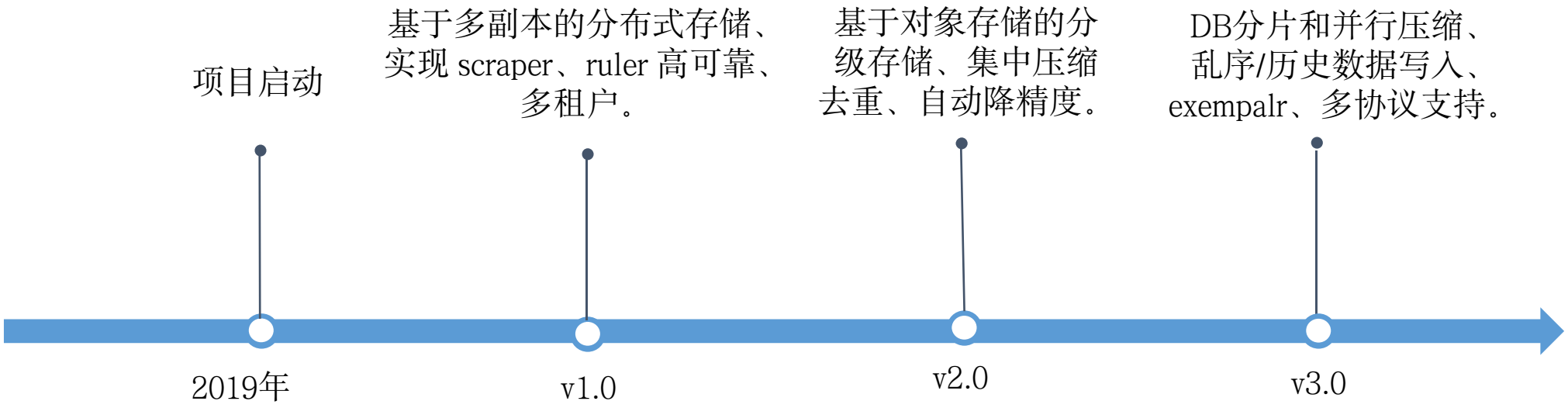
```
files
├── default
│   └── logs
│       └── default
│           └── 2023
│               └── 08
│                   └── 12
│                       └── 08
│                           └── 7096060482146140160LYnjyi.parquet
└── org_name
    └── metrics
        ├── avalanche_metric_mmmm_0_0
        │   └── 2023
        │       └── 08
        │           └── 12
        │               └── 09
        │                   └── 7096059936337166337bly72H.parquet
        ├── metric_mmmm_0_1
        │   └── 12
        │       └── 09
        │           └── 7096059936181977088z6s2s2.parquet
        ├── metric_mmmm_0_10
        │   └── 12
        │       └── 09
        │           └── 7096059936337166336Onot3R.parquet
```

block
s

- Openobserve

可观测数据融合平台

夯实metrics能力

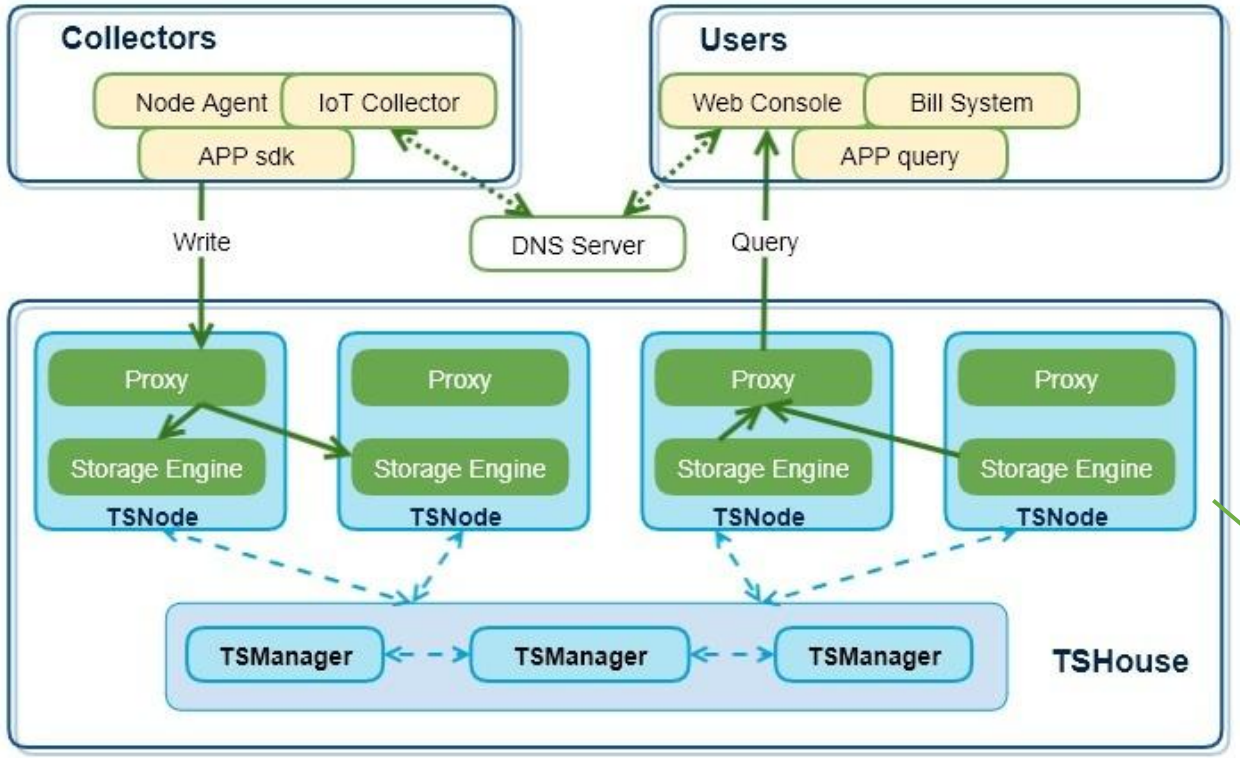


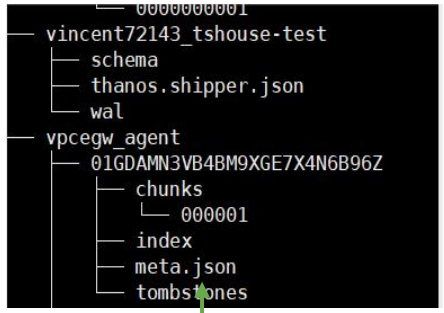
抽象平台层　　分布式、高可用、高基数

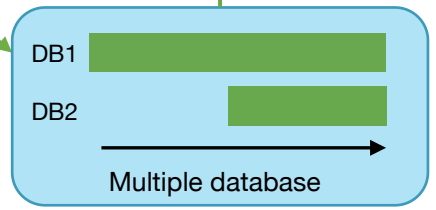
数据融合平台　　扩展log、trace底座能力，作为数据存储层

夯实metrics能力

v1.0 - 基于多副本实现存储高可靠







v1.0 – 小结

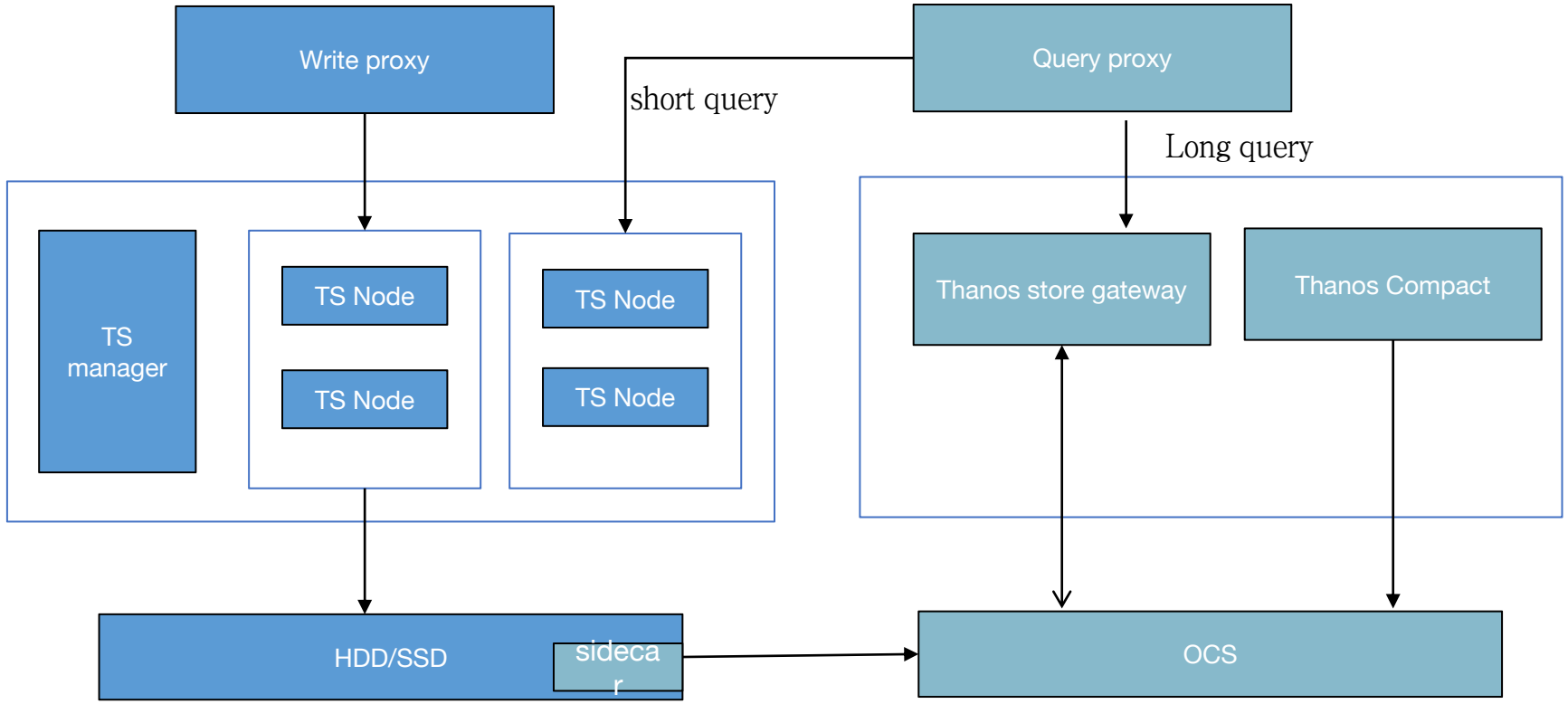
实现：

1. 基于 Prometheus TSDB 构建多租户的分布式存储服务。
2. 租户之间可以配置不同的读写QPS 限制和存储周期。
3. 分布式 scraper 和 ruler 实现数据抓取和 rule 评估的高可靠。

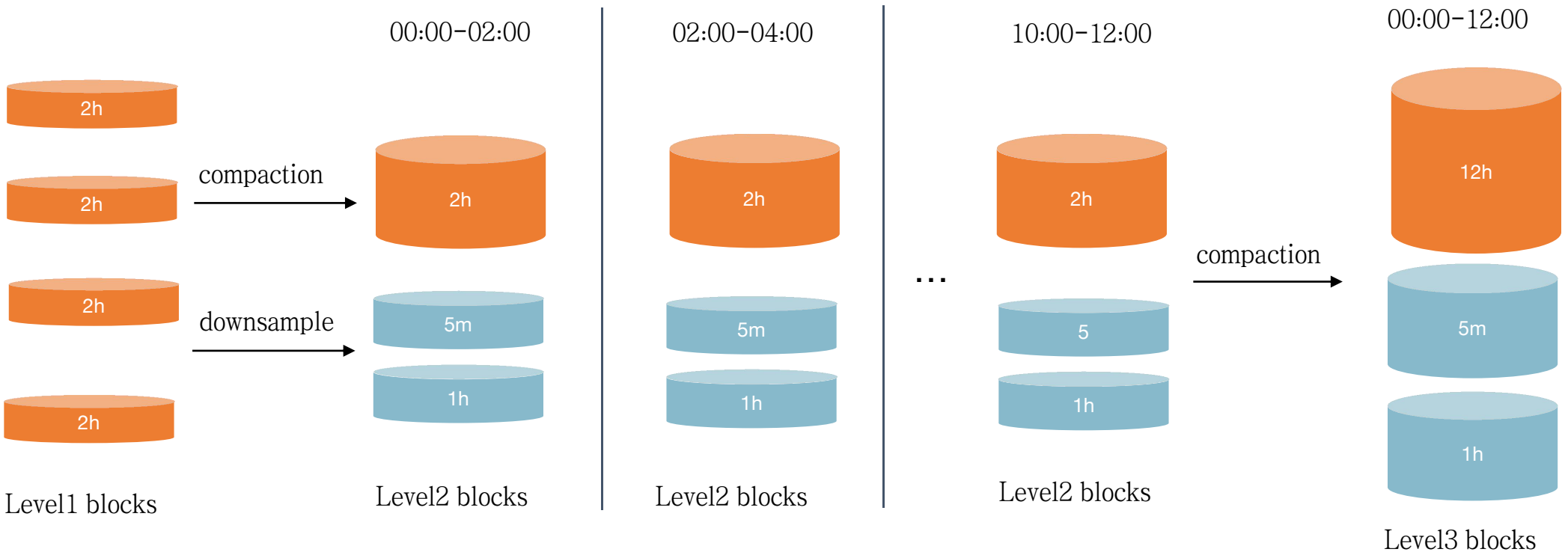
不足：

1. 长时间存储集群运维压力大，扩容和坏盘修复较复杂。
2. 读写没有分离，相互影响，尤其长时间范围查询。
3. 本地存储多副本，对于冷数据（一个月前）没有进行去重，造成存储浪费。

v2.0 - 基于对象存储的分级存储



v2.0 - 集中压缩和自动降精度



v2.0 – 小结

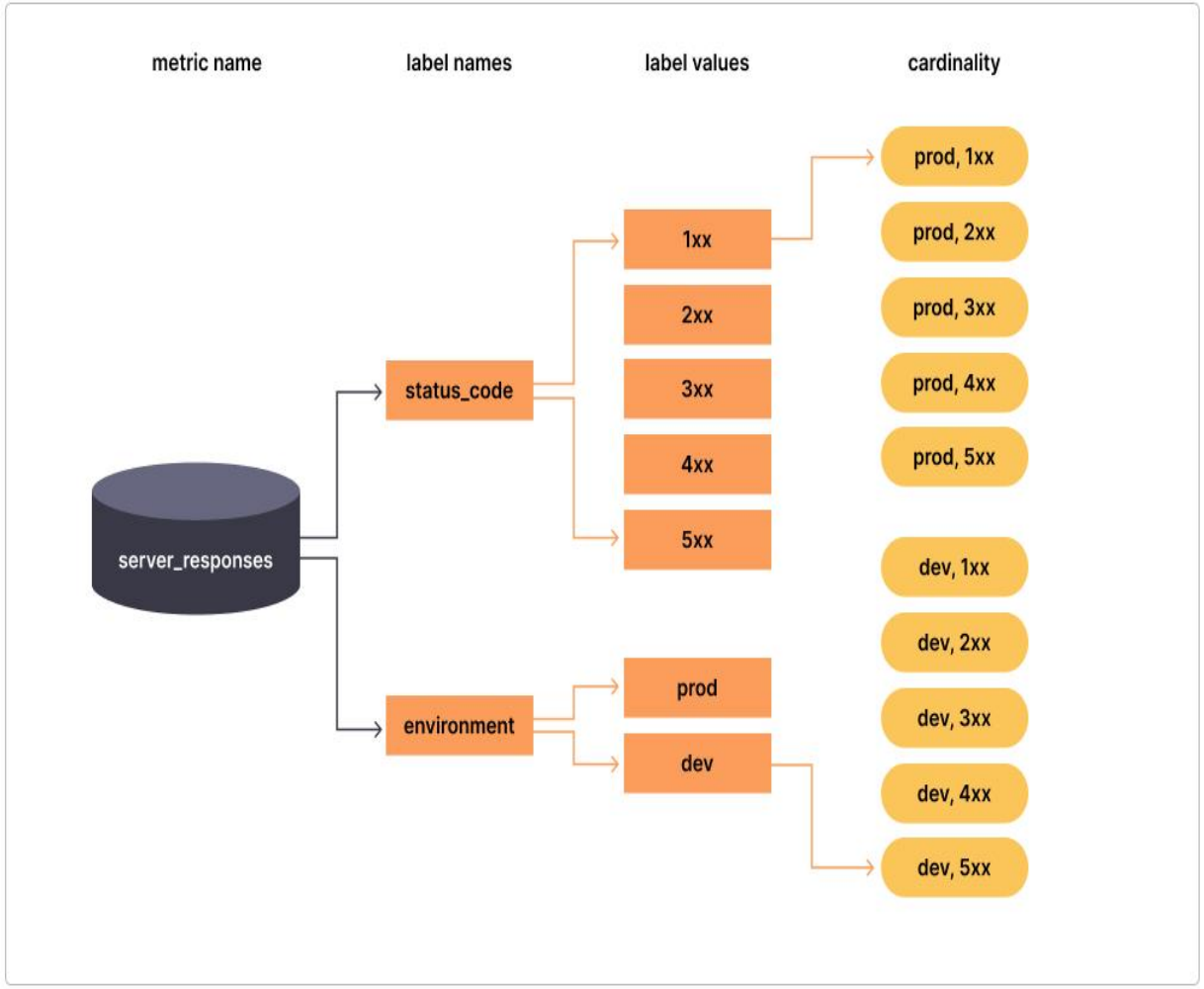
实现：

1. 集中压缩，数据去重，降低成本。
2. 自动降精度，长时间范围查询自动转化，提升查询效率。
3. 依赖对象存储实现长时间数据集中存储。

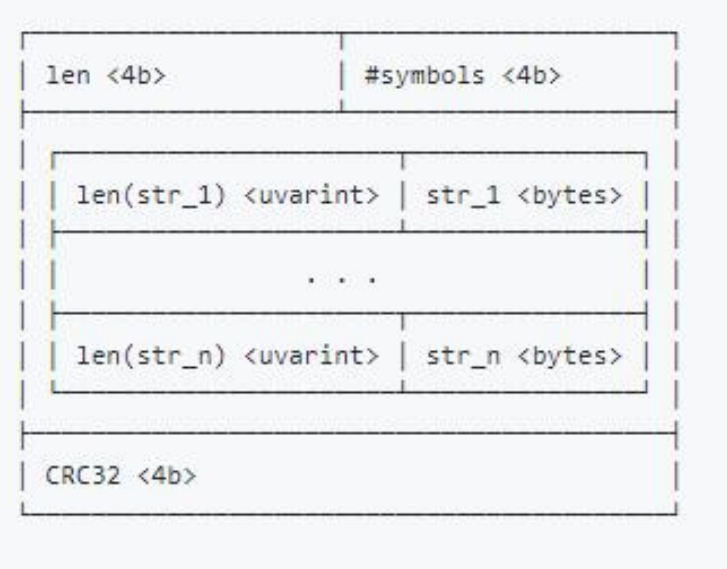
不足：

1. 超大规模数据压缩时间长，占用磁盘和内存大，受限 TSDB 索引中 symbols 长度 64GB 限制，高 level 压缩会导致失败。
2. 历史数据和乱序数据写入支持较差。
3. 仅支持 OpenMetrics/Prometheus 数据格式写入。

v3.0 - 高基数和大规模数据存储挑战

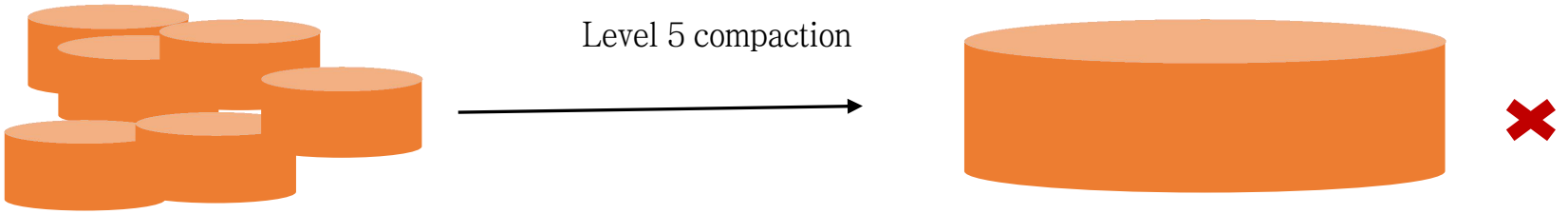


v3.0 - 高基数和大规模数据存储挑战

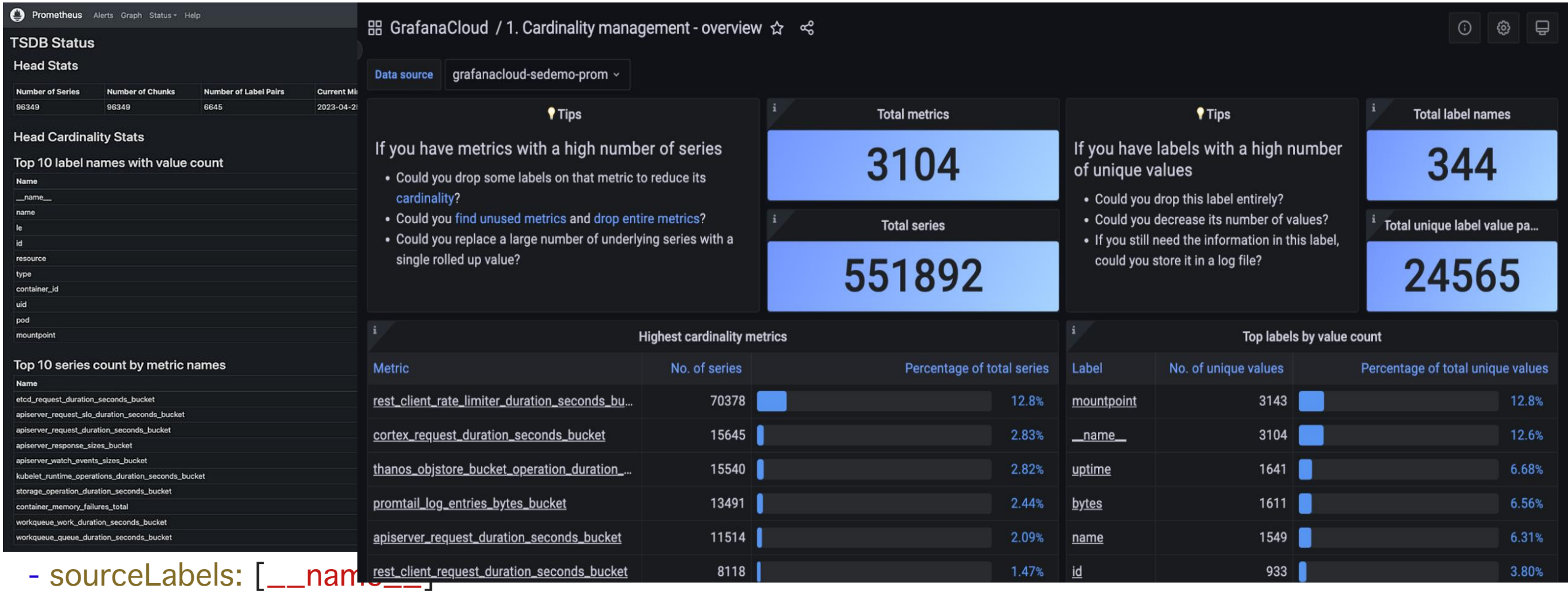



```
523   }
524
525   func (w *Writer) finishSymbols() error {
526 +         symbolTableSize := w.f.pos - w.toc.Symbols - 4
527 +         // The symbol table's <len> part is 4 bytes. So the total symbol table size must be less than or equal to 2^32-1
528 +         if symbolTableSize > 4294967295 {
529 +                 return errors.Errorf("symbol table size exceeds 4 bytes: %d", symbolTableSize)
530 +         }
531 +
532           // Write out the length and symbol count.
533           w.buf1.Reset()
534 +         w.buf1.PutBE32int(int(symbolTableSize))
535           w.buf1.PutBE32int(int(w.numSymbols))
536           if err := w.writeAt(w.buf1.Get(), w.toc.Symbols); err != nil {
537                   return err
```

https://github.com/prometheus/prometheus/pull/9104

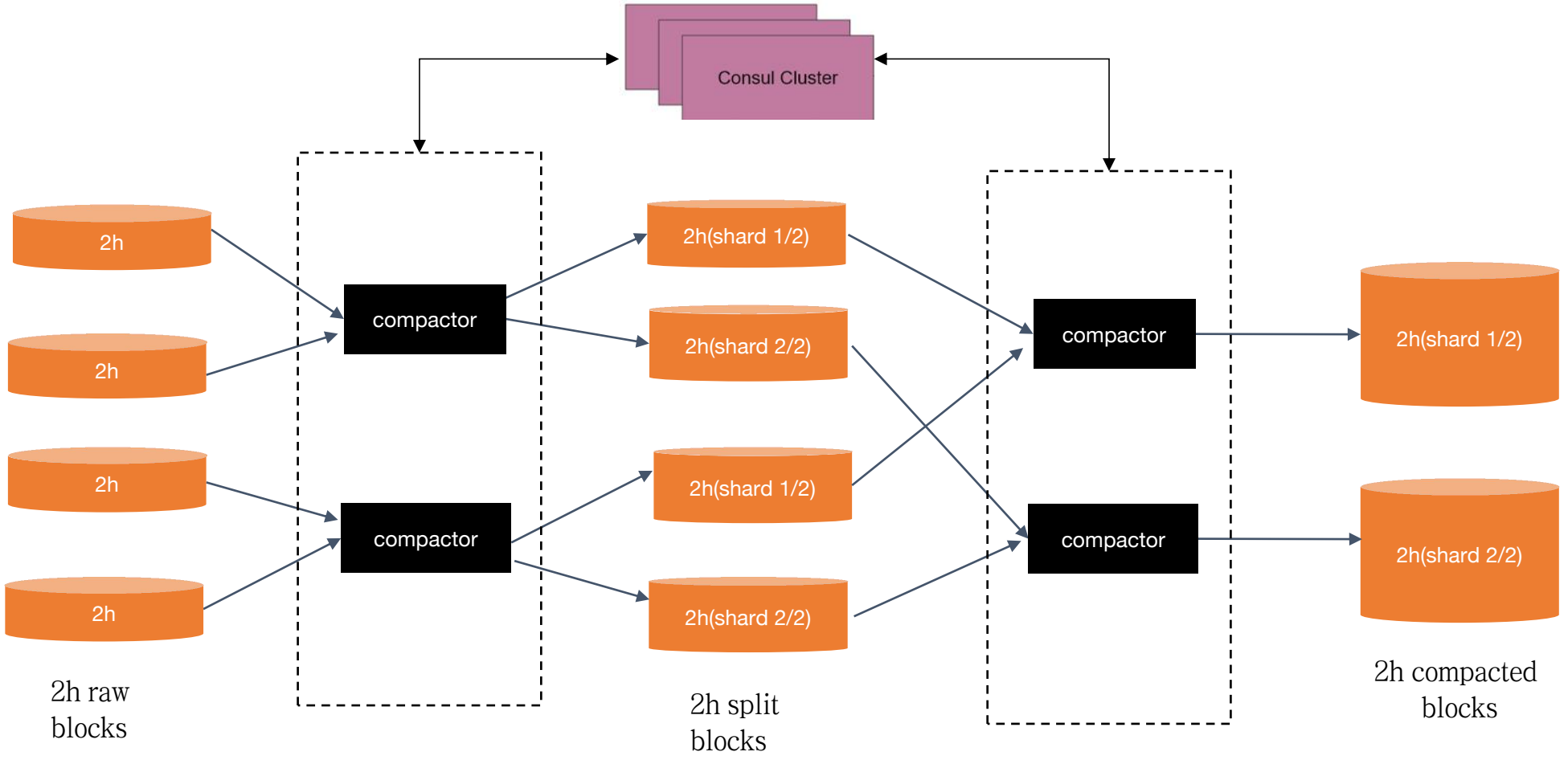


v3.0 - 高基数和大规模数据存储挑战

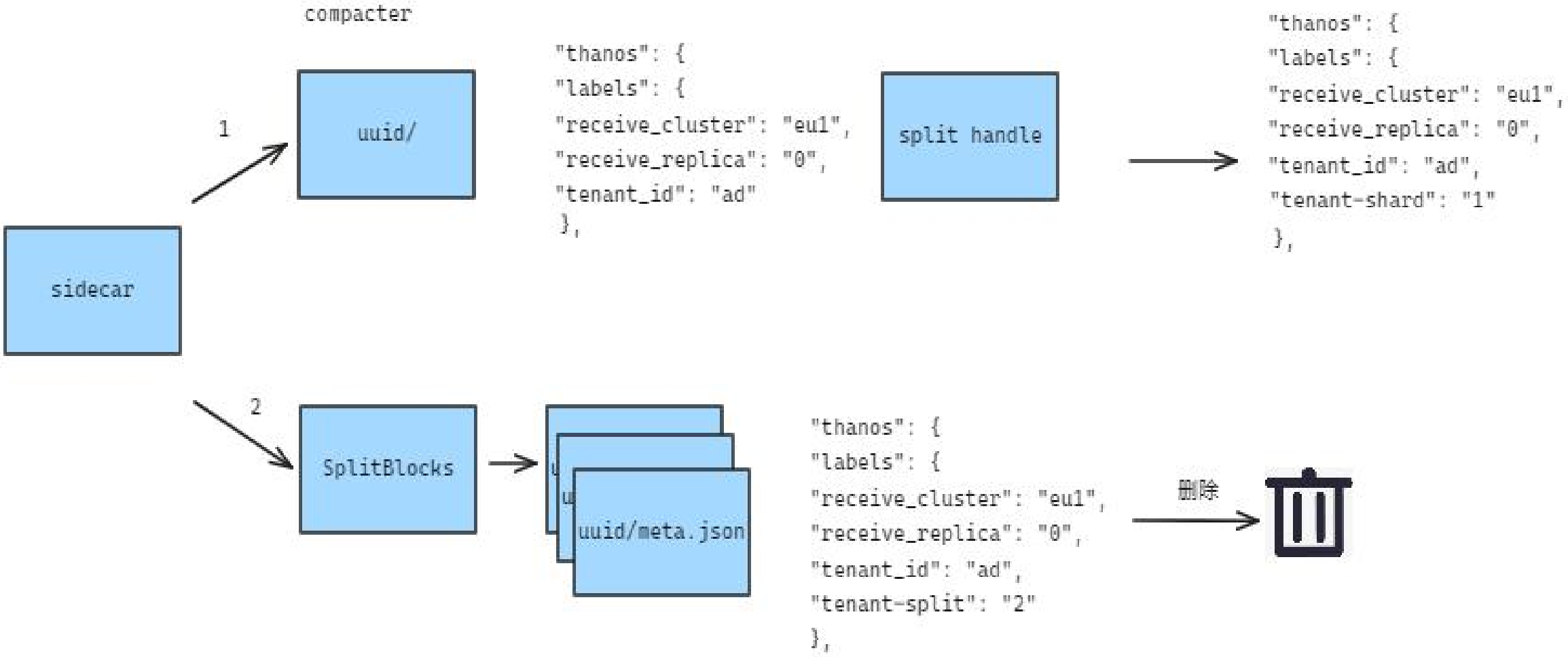



```
- sourceLabels: [__name__]
action: drop
regex: 'node_(nf_conntrack_statl
```

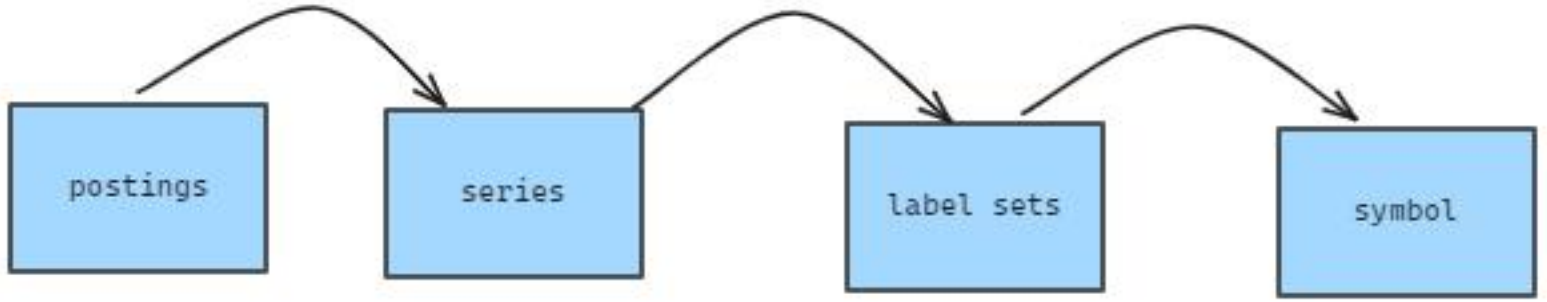
v3.0 - DB 分片与并行压缩

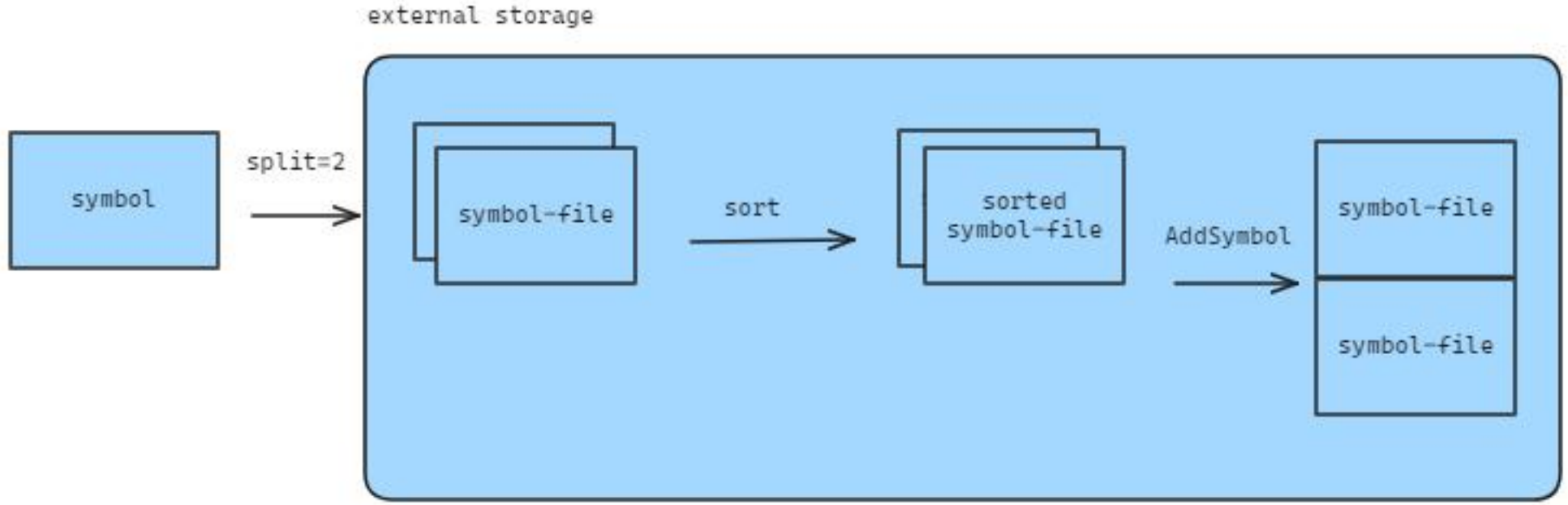


v3.0 - DB 分片与并行压缩

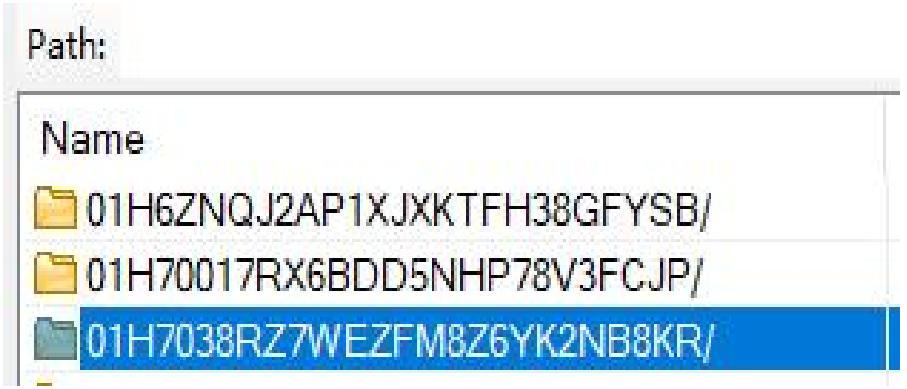


v3.0 - DB 分片与并行压缩

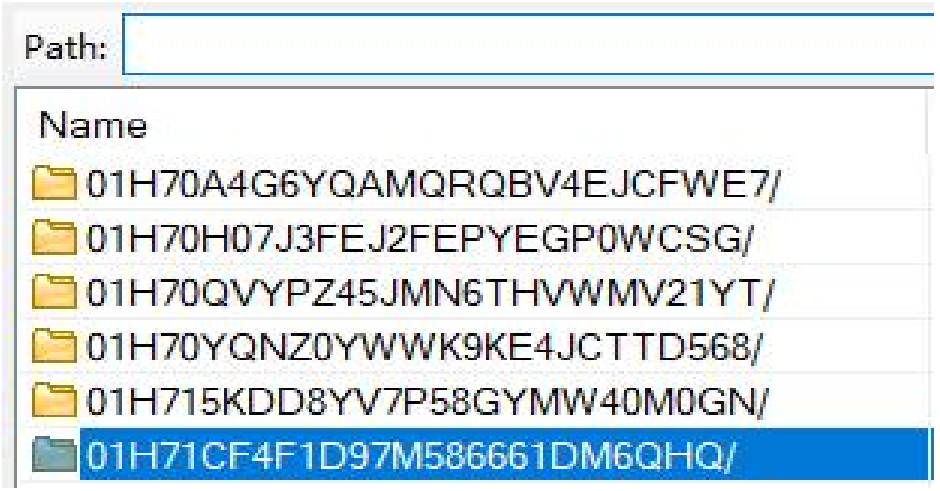




v3.0 - DB 分片与并行压缩

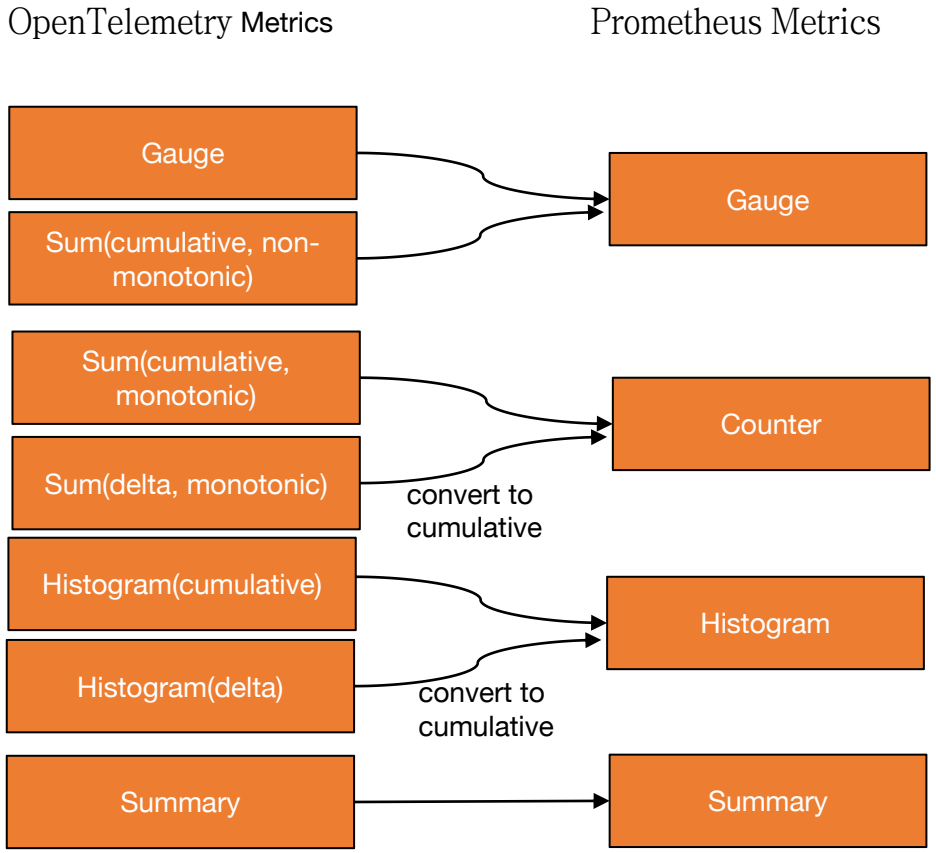



```
"thanos": {
            "labels": {
          "receive_cluster": "eu1",
          "receive_replica": "0",
          "tenant_id": "ad",
          "tenant-split": "2",
        },
```




```
"thanos": {
            "labels": {
          "receive_cluster": "eu1",
          "receive_replica": "0",
          "tenant_id": "ad",
            "tenant-shard": "1",
            "tenant-split": "2",
        },
```

v3.0 - OpenTelemetry 协议支持

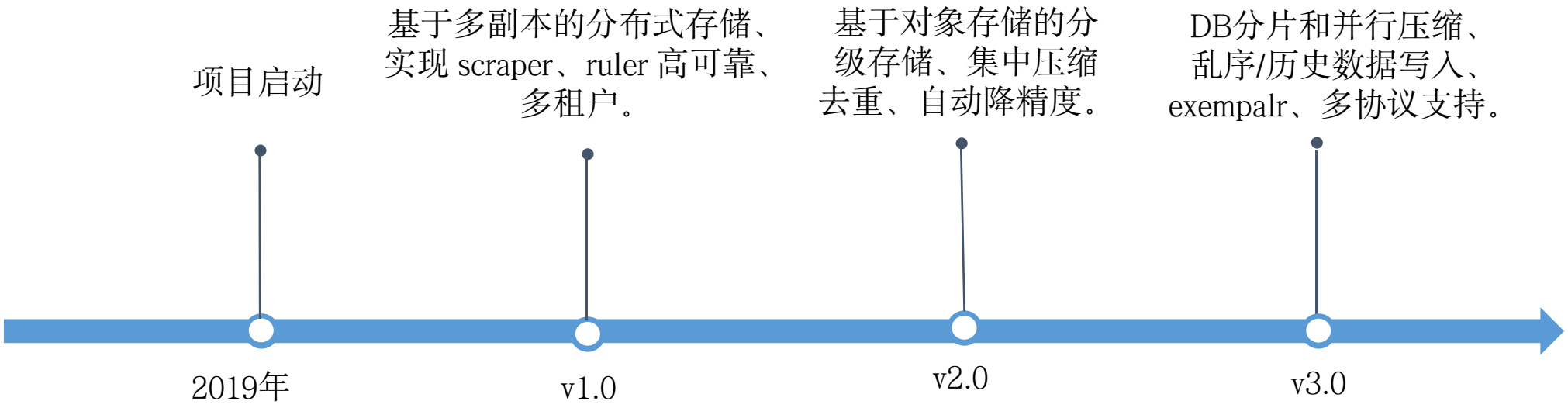



```
Metric #4
Descriptor:
     -> Name: http_durations_histogram_seconds
     -> Description: Http latency distributions.
     -> Unit:
     -> DataType: Histogram
     -> AggregationTemporality: Cumulative
HistogramDataPoints #0
Data point attributes:
     -> code: Str(200)
     -> method: Str(GET)
     -> path: Str(/v1/books/show)
StartTimestamp: 2022-10-19 14:38:52.634 +0000 UTC
Timestamp: 2022-10-19 14:39:22.597 +0000 UTC
Count: 4938
Sum: 94.990601
ExplicitBounds #0: 0.050000
ExplicitBounds #1: 0.100000
ExplicitBounds #2: 0.250000
ExplicitBounds #3: 0.500000
ExplicitBounds #4: 1.000000
ExplicitBounds #5: 2.000000
Buckets #0, Count: 4845
Buckets #1, Count: 26
Buckets #2, Count: 37
Buckets #3, Count: 30
Buckets #4, Count: 0
Buckets #5, Count: 0
Buckets #6, Count: 0
```

```
# HELP http_durations_histogram_seconds Http latency distributions.
# TYPE http_durations_histogram_seconds histogram
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="0.05"} 4845
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="0.1"} 4871
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="0.25"} 4908
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="0.5"} 4938
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="1"} 4938
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="2"} 4938
http_durations_histogram_seconds_bucket{code="200",method="GET",path="/v1/books/show",le="+Inf"} 4938
http_durations_histogram_seconds_sum{code="200",method="GET",path="/v1/books/show"} 94.9906009000003
http_durations_histogram_seconds_count{code="200",method="GET",path="/v1/books/show"} 4938
```

可观测数据融合平台

夯实metrics能力

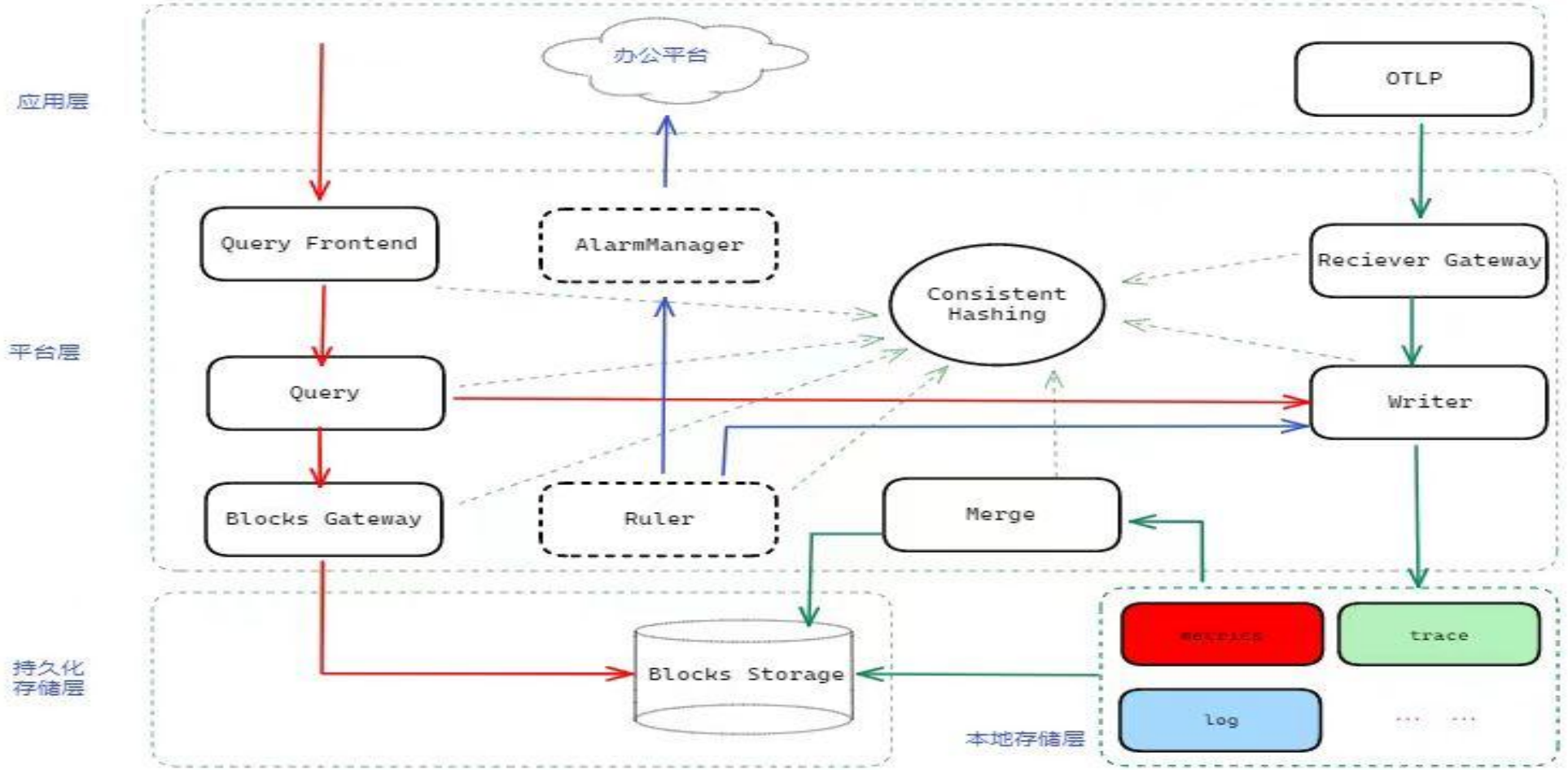


抽象平台层　　分布式、高可用、高基数

数据融合平台　　扩展log、trace等底座能力，作为数据存储层

可观测数据融合平台

抽象平台层



可观测数据融合平台

扩展log、trace能力

TSDB format

- Index
- Chunks
- Head Chunks
- Tombstones
- Wal
- Memory Snapshot

XOR chunk data

num_samples <uint16>	ts_0 <varint>	v_0 <float64>	ts_1_delta <uvarint>	v_1_xor <varbit_xor>	ts_2_dod <varbit_ts>

log chunk data

num_samples <uint16>	ts_0 <int64>	data_len <int64>	log_data <varint>	ts_1 <int64>

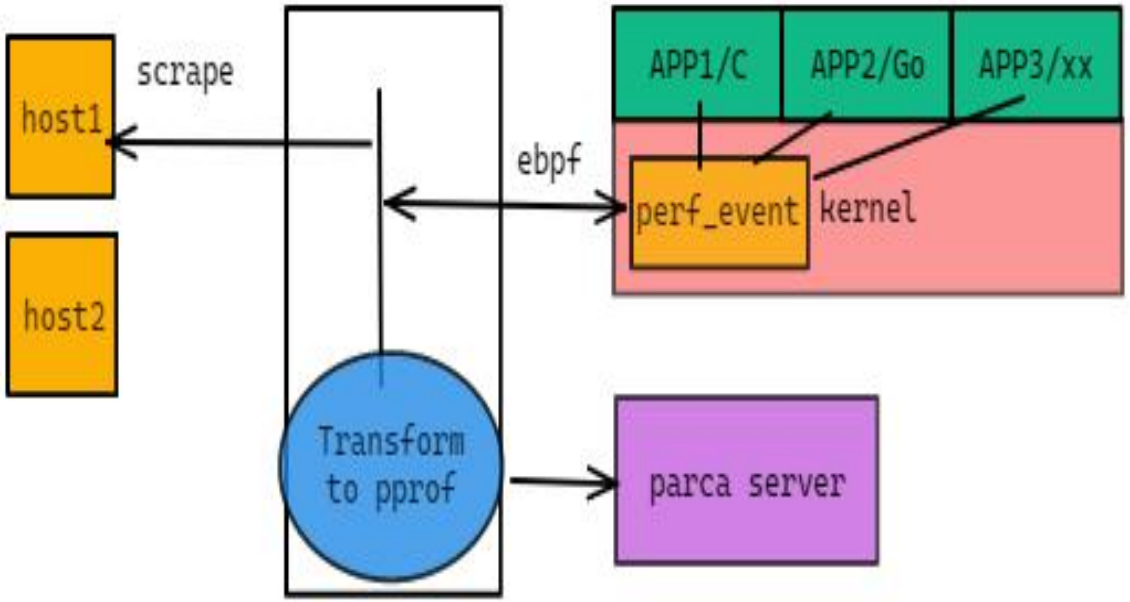
扩展log 能力

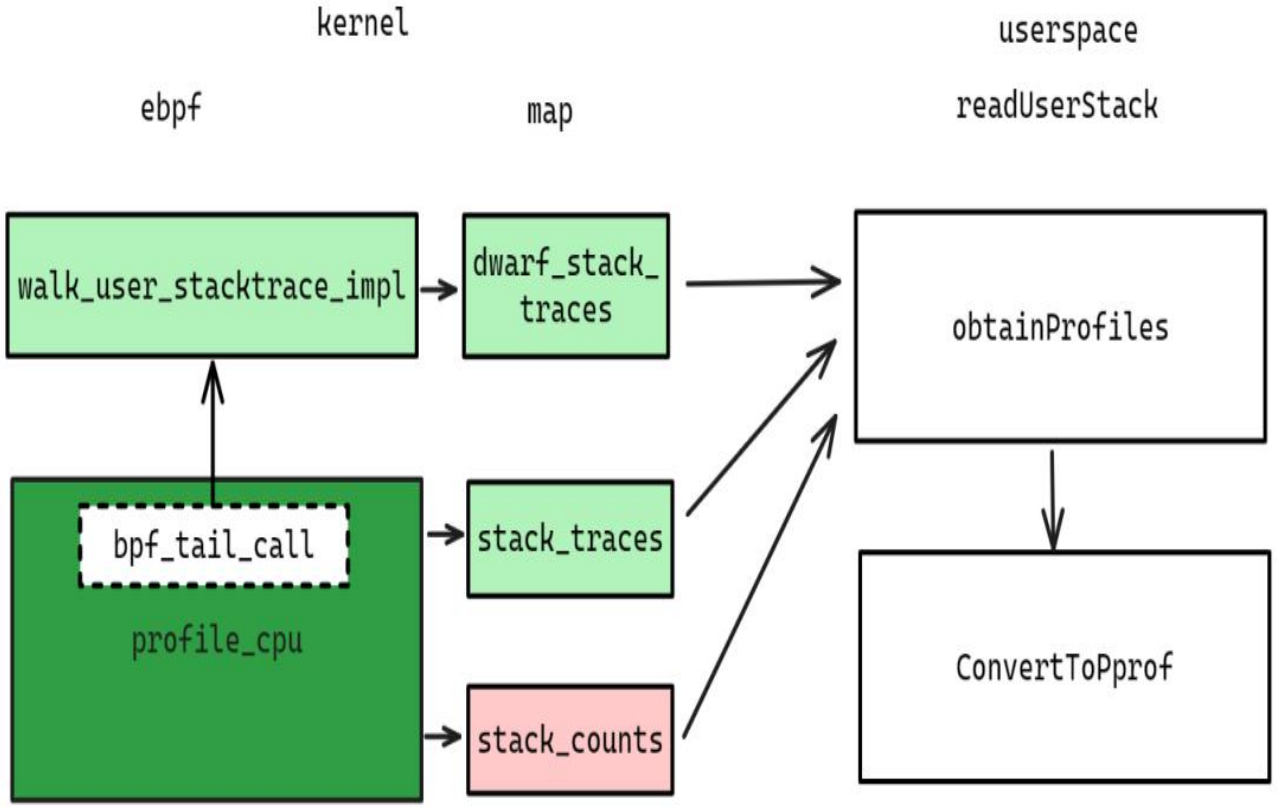

```go
func (a *xorAppender) Append(t int64, v float64) {
    var tDelta uint64
    num := binary.BigEndian.Uint16(a.b.bytes())
    switch num {
    case 0:
        buf := make([]byte, binary.MaxVarintLen64)
        for _, b := range buf[:binary.PutVarint(buf, t)] {
            a.b.writeByte(b)
        }
        a.b.writeBits(math.Float64bits(v),  nbits: 64)
    case 1:
        tDelta = uint64(t - a.t)

        buf := make([]byte, binary.MaxVarintLen64)
        for _, b := range buf[:binary.PutUvarint(buf, tDelta)] {
            a.b.writeByte(b)
        }

        a.writeVDelta(v)

    default:
        tDelta = uint64(t - a.t)
```

```go
func (a *logAppender) AppendLog(t int64, v string) {

    /**
    chunk format:
    t int64;datalen int64;logdata []byte;// default lz4 compression
    t int64;datalen int64;logdata []byte;
    ...
     **/

    // 1. write time section
    buf := make([]byte, 8)
    binary.BigEndian.PutUint64(buf, uint64(t))
    for _, b := range buf {
        a.b.writeByte(b)
    }

    // 2. write datalen section
    lenbuf := make([]byte, 8)
    fmt.Println( a...: "en(v)",len(v))
    binary.BigEndian.PutUint64(lenbuf, uint64(len(v)))
    for _, b := range lenbuf {
        a.b.writeByte(b)
    }
    fmt.Println( a...: "2--a.b.stream:",a.b.stream)
```

eBPF

Parca

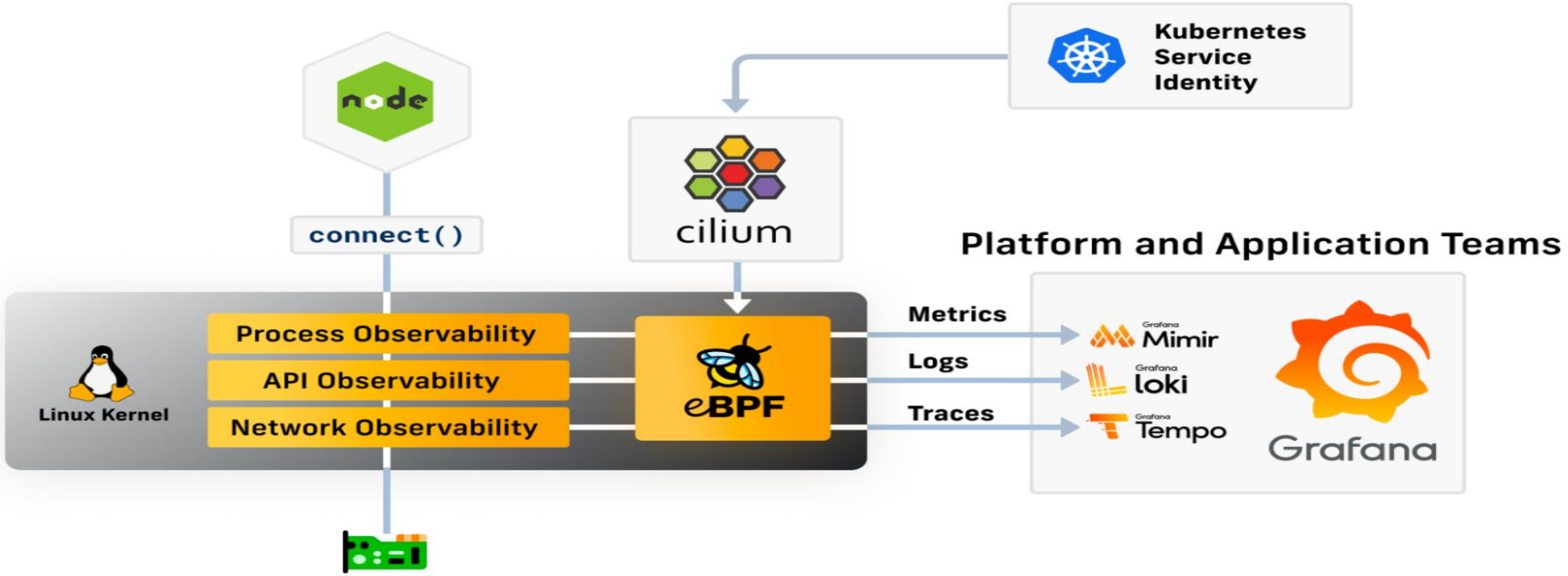




[bpf profile_cpu] "has_unwind_information" works rather than "has_fp" even though a porgrame(C) compiled with frame pointers
https://github.com/parca-dev/parca-agent/issues/1657

OpenTelemetry Auto Instrumentation using eBPF
https://github.com/open-telemetry/opentelemetry-go-instrumentation/pull/149

eBPF


node
connect()
Kubernetes
Service
Identity
cilium
Platform and Application Teams
Linux Kernel
Process Observability
API Observability
Network Observability
eBPF
Metrics
Logs
Traces
Grafana Mimir
Grafana loki
Grafana Tempo
Grafana

总结

- 1、各类方案百花齐放，多调研和灰度验证
- 2、平台聚焦服务业务为主
- 3、找到适合业务的技术方案
- 4、慎重叠加解决方案
- 5、尽量抽象平台能力，扩展底层能力

谢谢

欢迎关注公众号~

Grafana fans

欢迎关注公众号~

THANKS