Clickhouse

as long term storage
for metrics, events, logs
from K8s

exness

Platform Team

About us

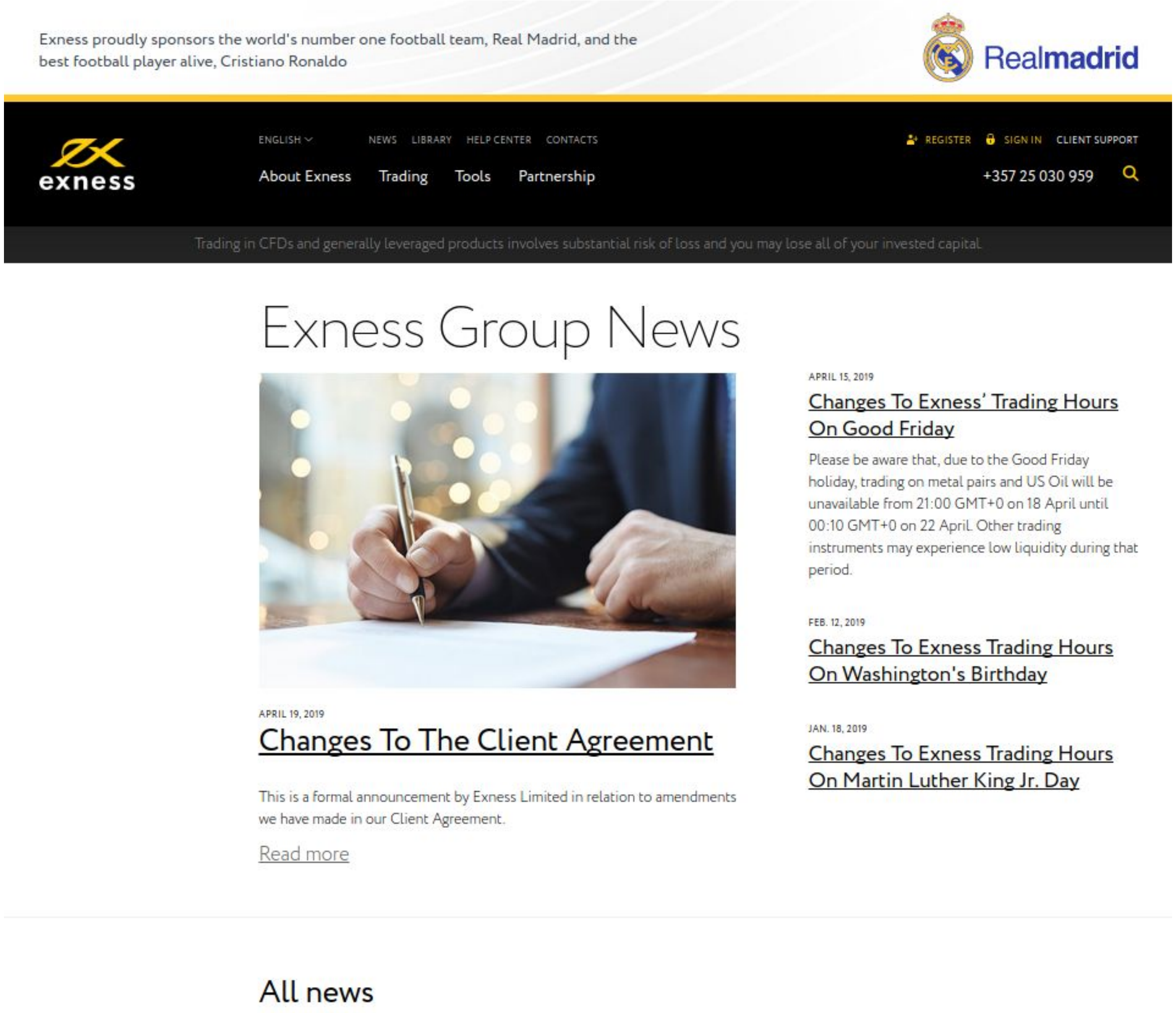
- Platform architecture
- Operations and maintenance
- Troubleshooting
- Deployments

* и Танцы с бубнами - это про нас)

exness

About

www.exness.com



Exness proudly sponsors the world's number one football team, Real Madrid, and the best football player alive, Cristiano Ronaldo

Trading in CFDs and generally leveraged products involves substantial risk of loss and you may lose all of your invested capital.

Exness Group News

APRIL 19, 2019

Changes To The Client Agreement

This is a formal announcement by Exness Limited in relation to amendments we have made in our Client Agreement.

Read more

APRIL 15, 2019

Changes To Exness' Trading Hours On Good Friday

Please be aware that, due to the Good Friday holiday, trading on metal pairs and US Oil will be unavailable from 21:00 GMT+0 on 18 April until 00:10 GMT+0 on 22 April. Other trading instruments may experience low liquidity during that period.

FEB. 12, 2019

Changes To Exness Trading Hours On Washington's Birthday

JAN. 18, 2019

Changes To Exness Trading Hours On Martin Luther King Jr. Day

All news

exness

Agenda

- Introduction
- Long time ago)
- 1nd implementation (Rancher)
- 2nd implementation (K8s)
- Questions

exness

About

what we have now

- **in 2 datacenters**
- **500+ service**
- **2500+ containers**
- **10Krps metrics up to 200+Krps**
- **2Krps Logs up to 100+Krps**

exness

Introduction

Clickhouse in production now

- **3 clusters**
- **10+ servers**
- **200+ cores, 1Tb ram,   20+ Tb SSD**

Introduction

Clickhouse in production now

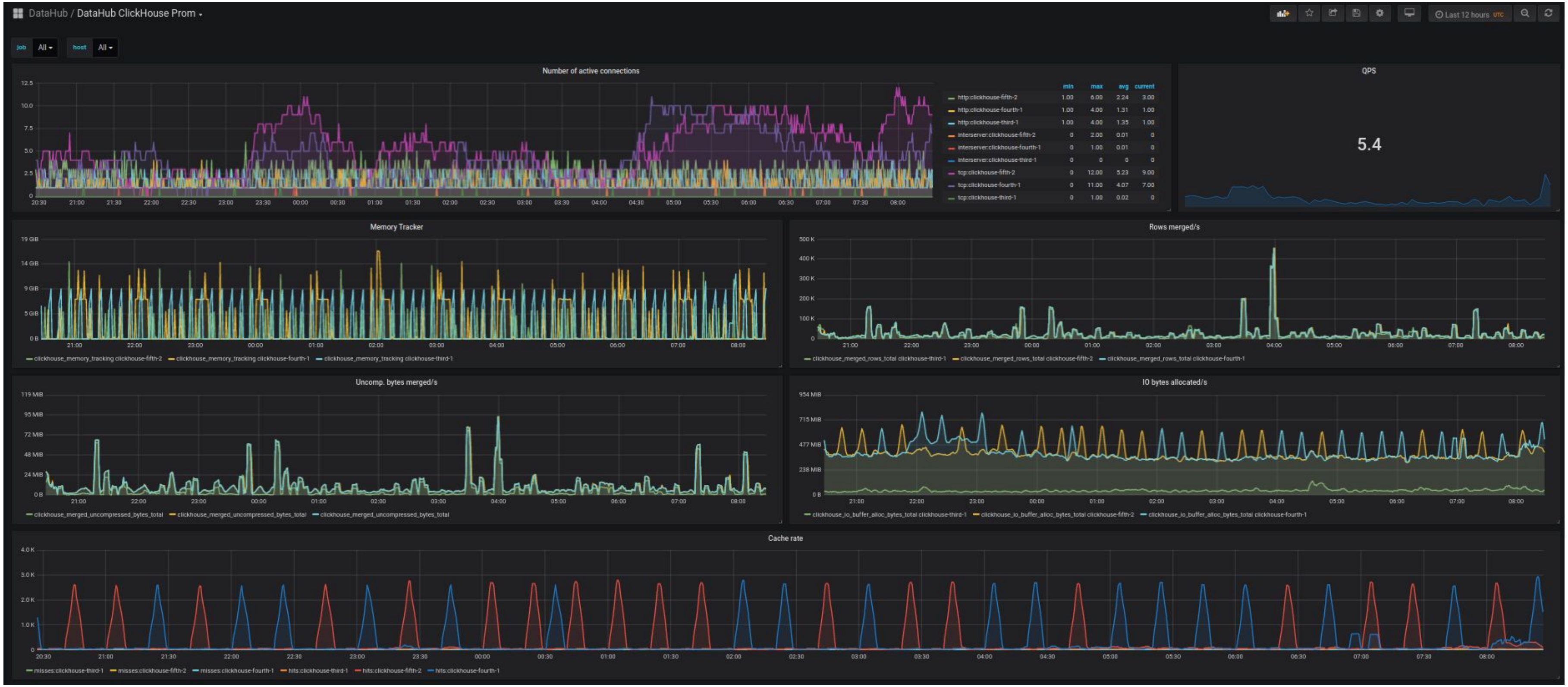
- **Easy to replicate**
- **Easy to shard**
- **Easy to use**
- **Easy to manage**
- **Nice support**
- **K8S Operator :)**

exness

Questions for “full” Clickhouse

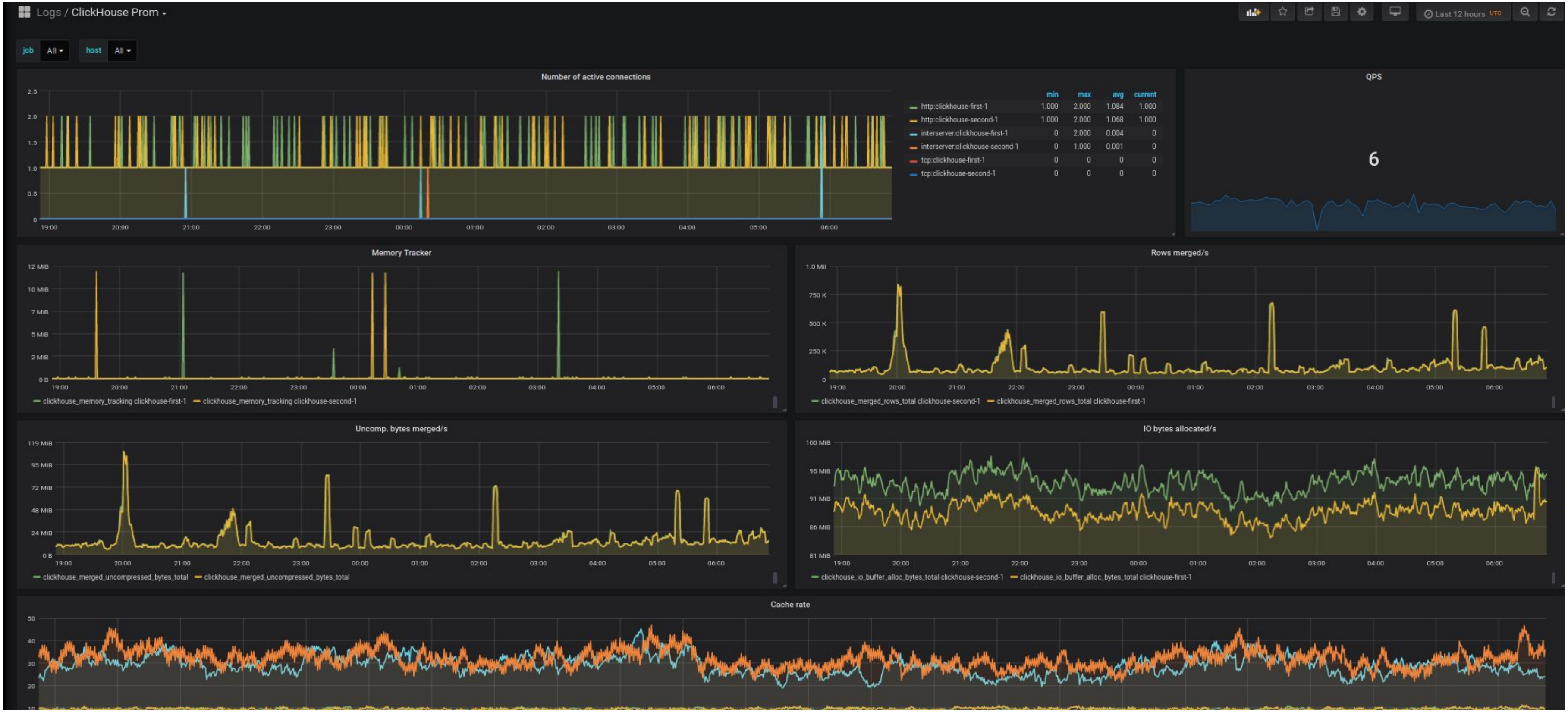
A lot of new versions ???
Access rights so simple
ZooKeeper only ?
Clouds ?
UI access ? (tabix, superset)

1. Zookeeper replacement => Etcd
2. Cloud messages brokers Kinesis, Pub/Sub
3. Auhtorization LDAP, SSO
4. Prometheus metrics exporter
5. GraphQL interface out of the box
6. Clickhouse as Prometheus long-term storage
7. Auto retention policy
8. Detach/Drop/Freeze parts (not only partition as a whole)



	min	max	avg	current
http:clickhouse-fifth-2	1.00	6.00	2.24	3.00
http:clickhouse-fourth-1	1.00	4.00	1.31	1.00
http:clickhouse-third-1	1.00	4.00	1.35	1.00
interserver:clickhouse-fifth-2	0	2.00	0.01	0
interserver:clickhouse-fourth-1	0	1.00	0.01	0
interserver:clickhouse-third-1	0	0	0	0
tcp:clickhouse-fifth-2	0	12.00	5.23	9.00
tcp:clickhouse-fourth-1	0	11.00	4.07	7.00
tcp:clickhouse-third-1	0	1.00	0.02	0

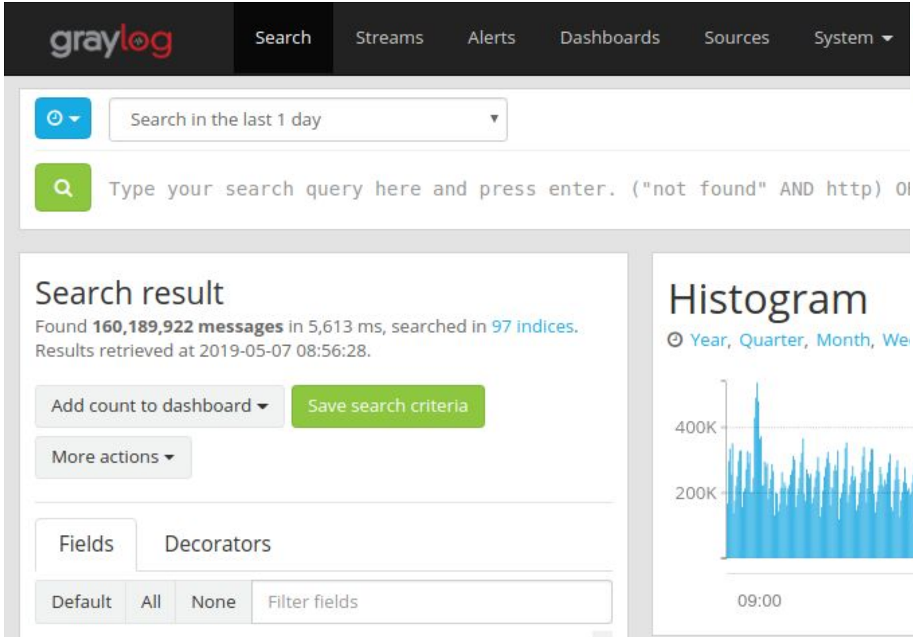
exness


Logs / ClickHouse Prom
Last 12 hours UTC
job All
host All
Number of active connections
2.5
2.0
1.5
1.0
0.5
0
19:00
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00
04:00
05:00
06:00
QPS
6
Memory Tracker
12 MiB
10 MiB
7 MiB
5 MiB
2 MiB
0 B
clickhouse_memory_tracking clickhouse-first-1
clickhouse_memory_tracking clickhouse-second-1
Rows merged/s
1.0 Mil
750 K
500 K
250 K
0
clickhouse_merged_rows_total clickhouse-second-1
clickhouse_merged_rows_total clickhouse-first-1
Uncomp. bytes merged/s
119 MiB
95 MiB
72 MiB
48 MiB
24 MiB
0 B
clickhouse_merged_uncompressed_bytes_total
clickhouse_merged_uncompressed_bytes_total
IO bytes allocated/s
100 MiB
95 MiB
91 MiB
86 MiB
81 MiB
clickhouse_io_buffer_alloc_bytes_total clickhouse-second-1
clickhouse_io_buffer_alloc_bytes_total clickhouse-first-1
Cache rate
50
40
30
20
10

	min	max	avg	current
http:clickhouse-first-1	1.000	2.000	1.084	1.000
http:clickhouse-second-1	1.000	2.000	1.068	1.000
interserver:clickhouse-first-1	0	2.000	0.004	0
interserver:clickhouse-second-1	0	1.000	0.001	0
tcp:clickhouse-first-1	0	0	0	0
tcp:clickhouse-second-1	0	0	0	0

exness

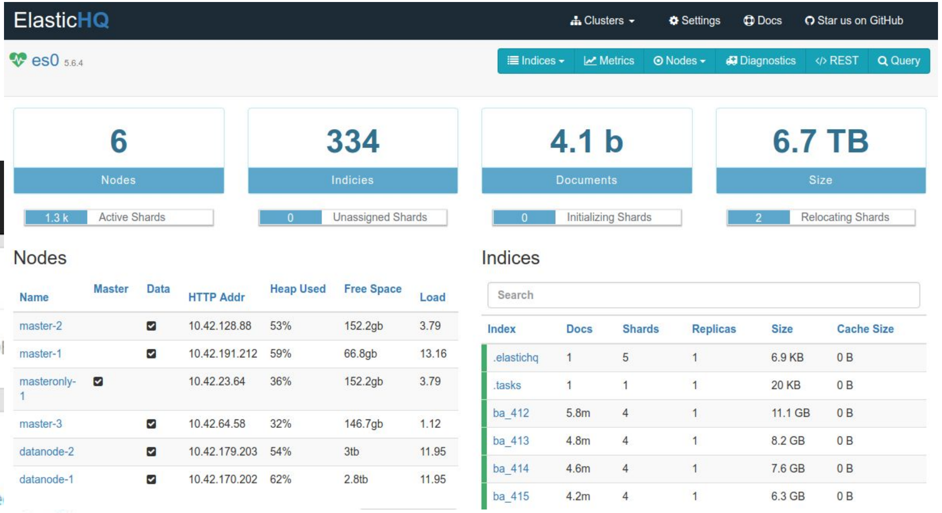
Logs



Search result

Found **160,189,922 messages** in 5,613 ms, searched in 97 indices.
Results retrieved at 2019-05-07 08:56:28.

Histogram



6 Nodes

334 Indicies

4.1 b Documents

6.7 TB Size

1.3 k Active Shards

0 Unassigned Shards

0 Initializing Shards

2 Relocating Shards

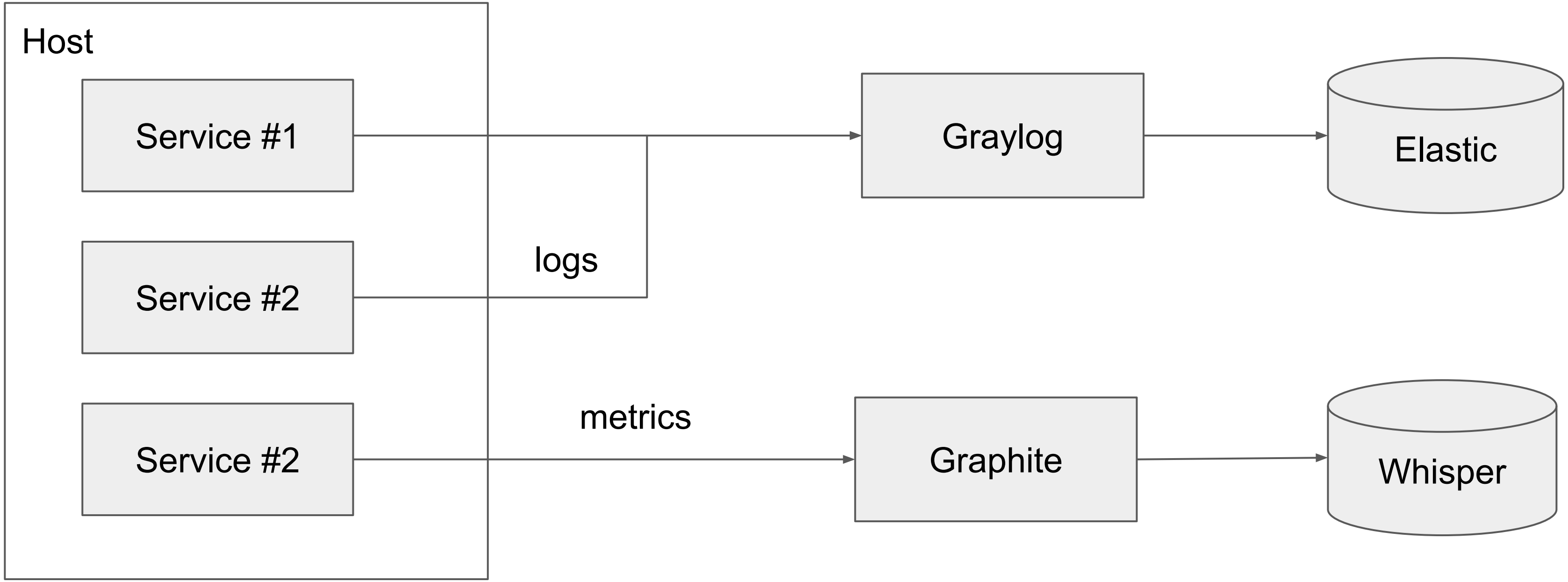
Nodes

Name	Master	Data	HTTP Addr	Heap Used	Free Space	Load
master-2		☑	10.42.128.88	53%	152.2gb	3.79
master-1		☑	10.42.191.212	59%	66.8gb	13.16
masteronly-1	☑		10.42.23.64	36%	152.2gb	3.79
master-3		☑	10.42.64.58	32%	146.7gb	1.12
datanode-2		☑	10.42.179.203	54%	3tb	11.95
datanode-1		☑	10.42.170.202	62%	2.8tb	11.95

Indices

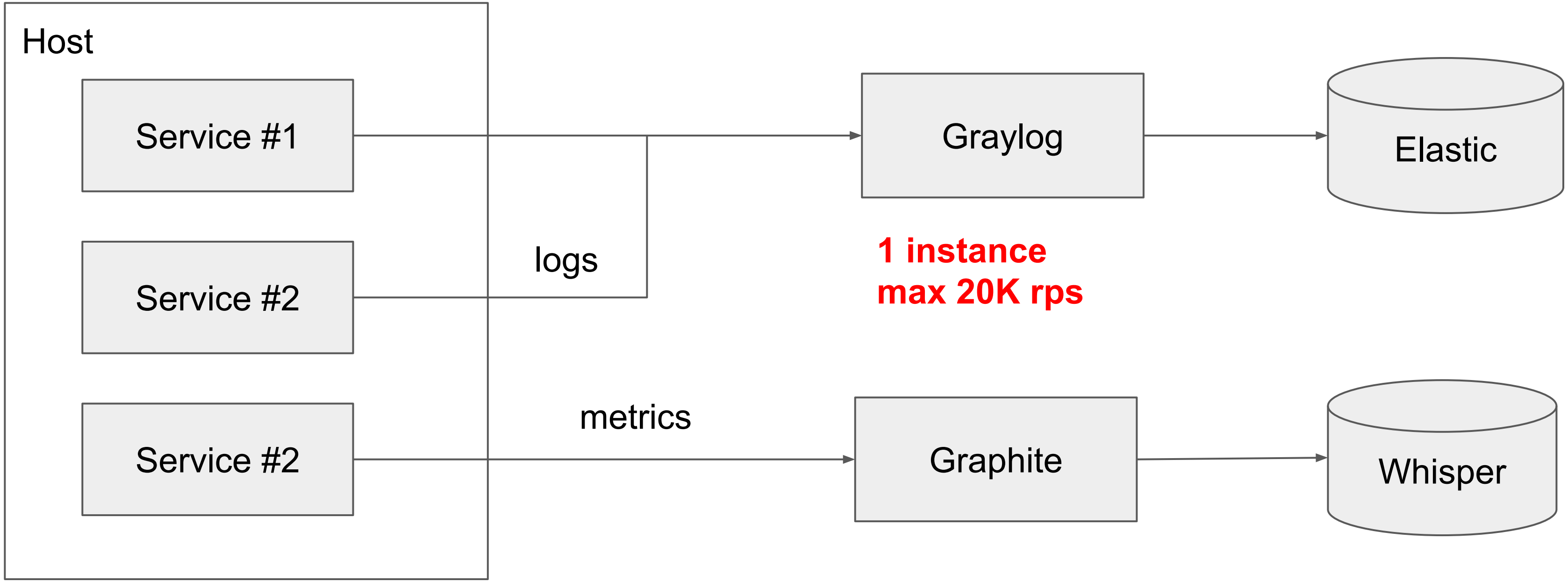
Index	Docs	Shards	Replicas	Size	Cache Size
.elastichq	1	5	1	6.9 KB	0 B
.tasks	1	1	1	20 KB	0 B
ba_412	5.8m	4	1	11.1 GB	0 B
ba_413	4.8m	4	1	8.2 GB	0 B
ba_414	4.6m	4	1	7.6 GB	0 B
ba_415	4.2m	4	1	6.3 GB	0 B

Long time ago)



20+ services

Long time ago)



Host

Service #1

Service #2

Service #2

logs

metrics

Graylog

Elastic

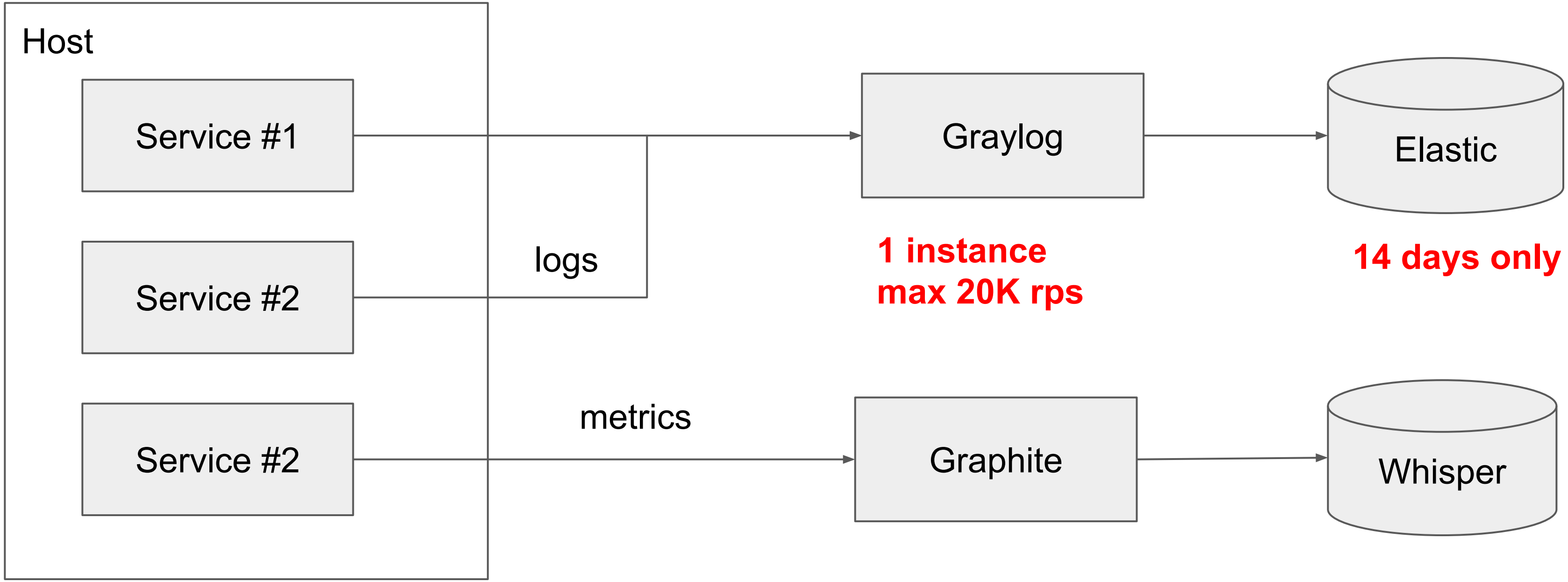
Graphite

Whisper

1 instance
max 20K rps

20+ services

Long time ago)



Host

Service #1

Service #2

Service #2

logs

metrics

Graylog

Graphite

Elastic

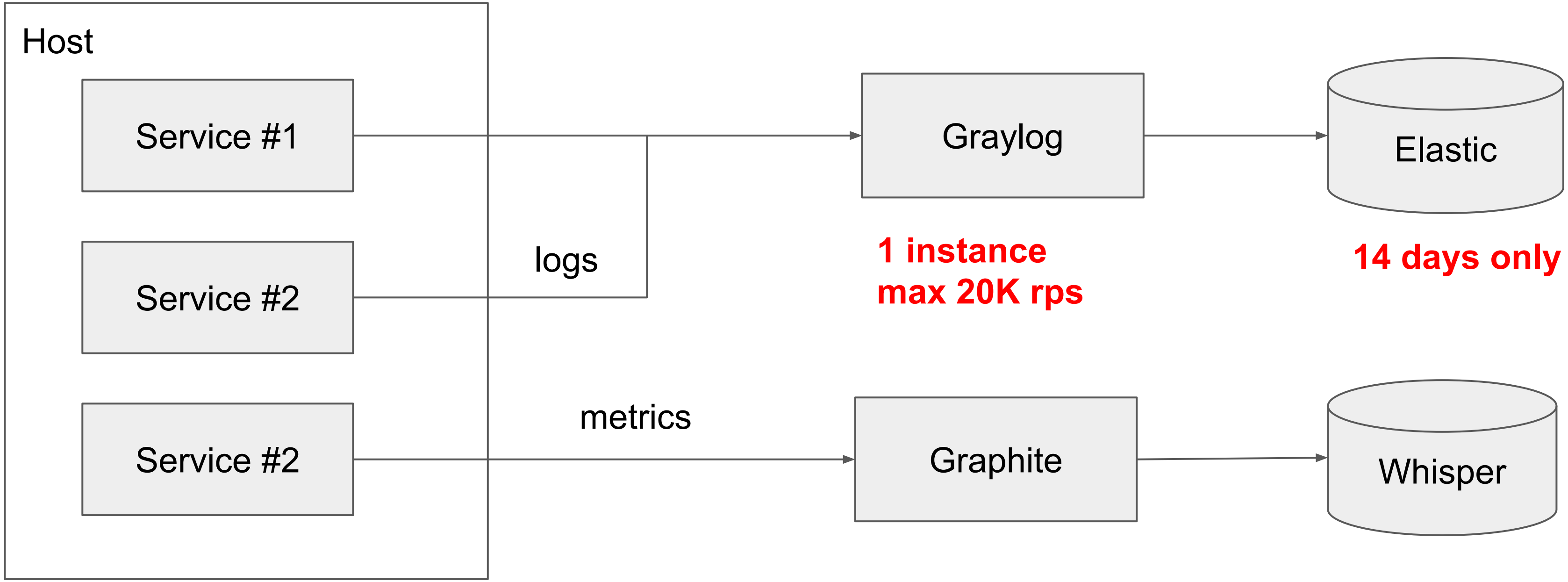
Whisper

1 instance
max 20K rps

14 days only

20+ services

Long time ago)

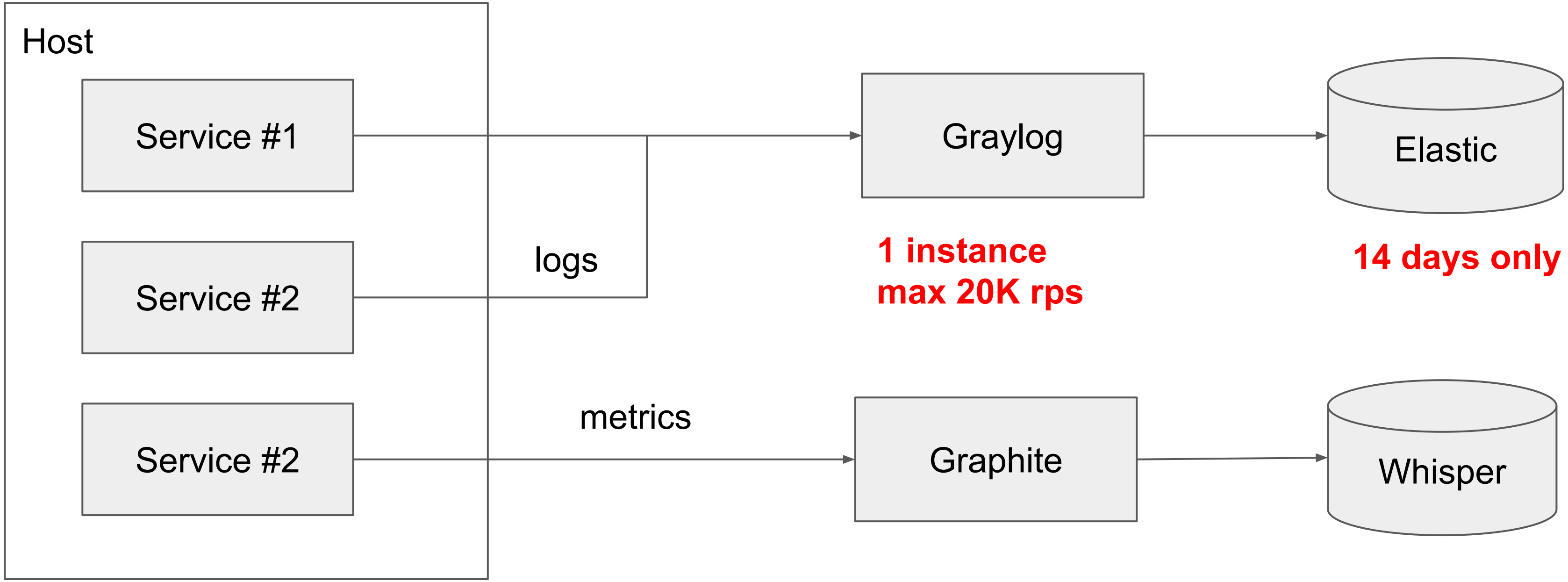


1 instance
max 20K rps

14 days only

200+ services ???

Long time ago)


Host

Service #1

Service #2

Service #2

logs

metrics

Graylog

Graphite

Elastic

Whisper

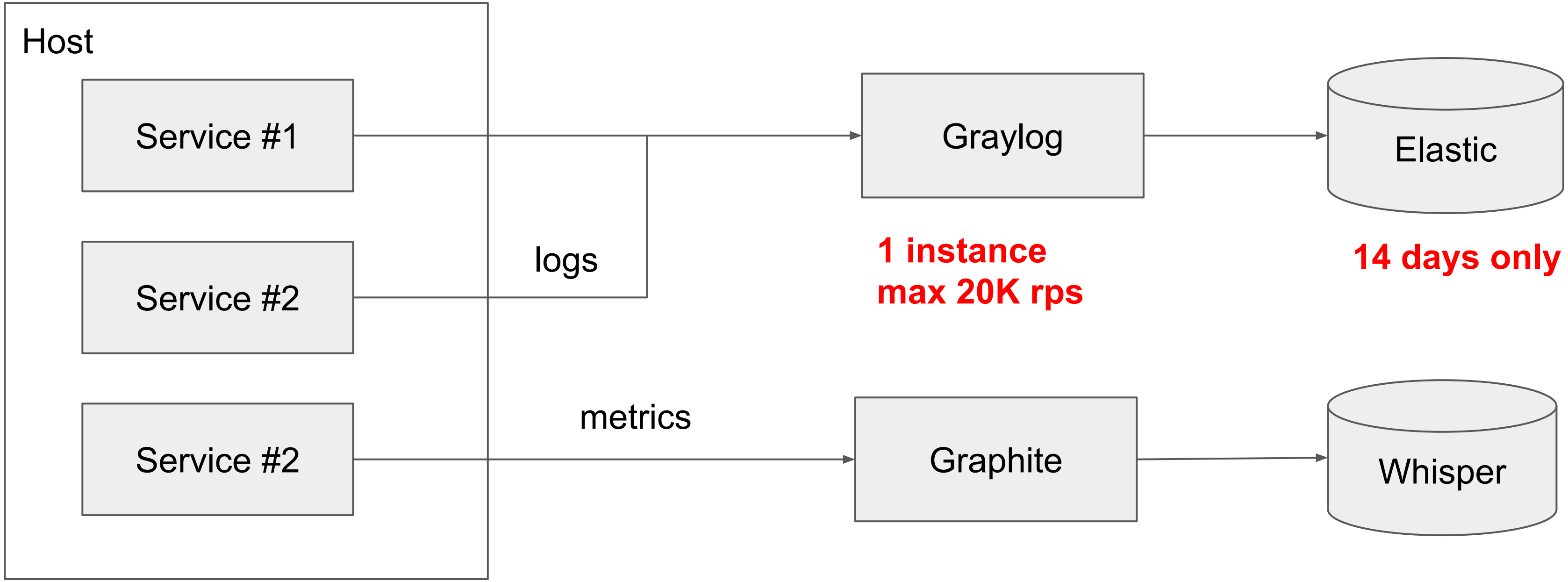
1 instance
max 20K rps

14 days only

200+ services ???

Throttling ?
Maintenance ?
Long term solution for 1 year

Long time ago)



Host

Service #1

Service #2

Service #2

logs

metrics

Graylog

Graphite

Elastic

Whisper

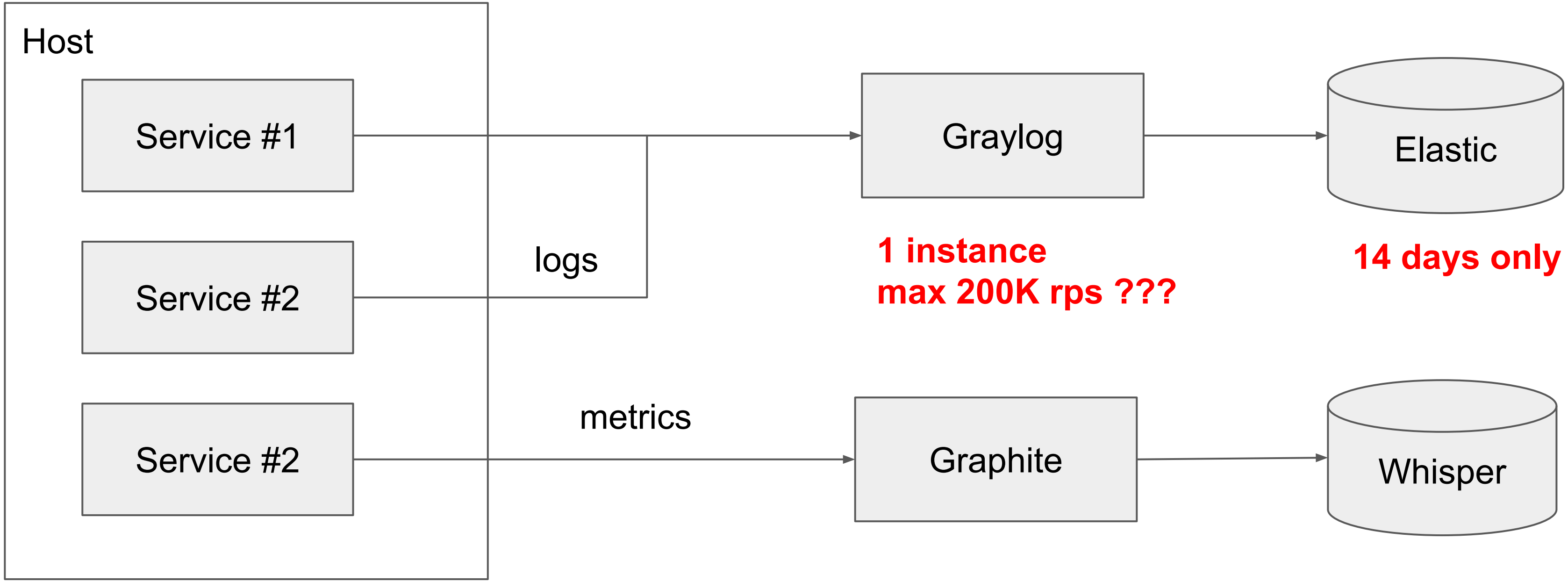
1 instance
max 20K rps

14 days only

2000+ services ???

Throttling ?
Maintenance ?
Long term solution for 1 year

Long time ago)



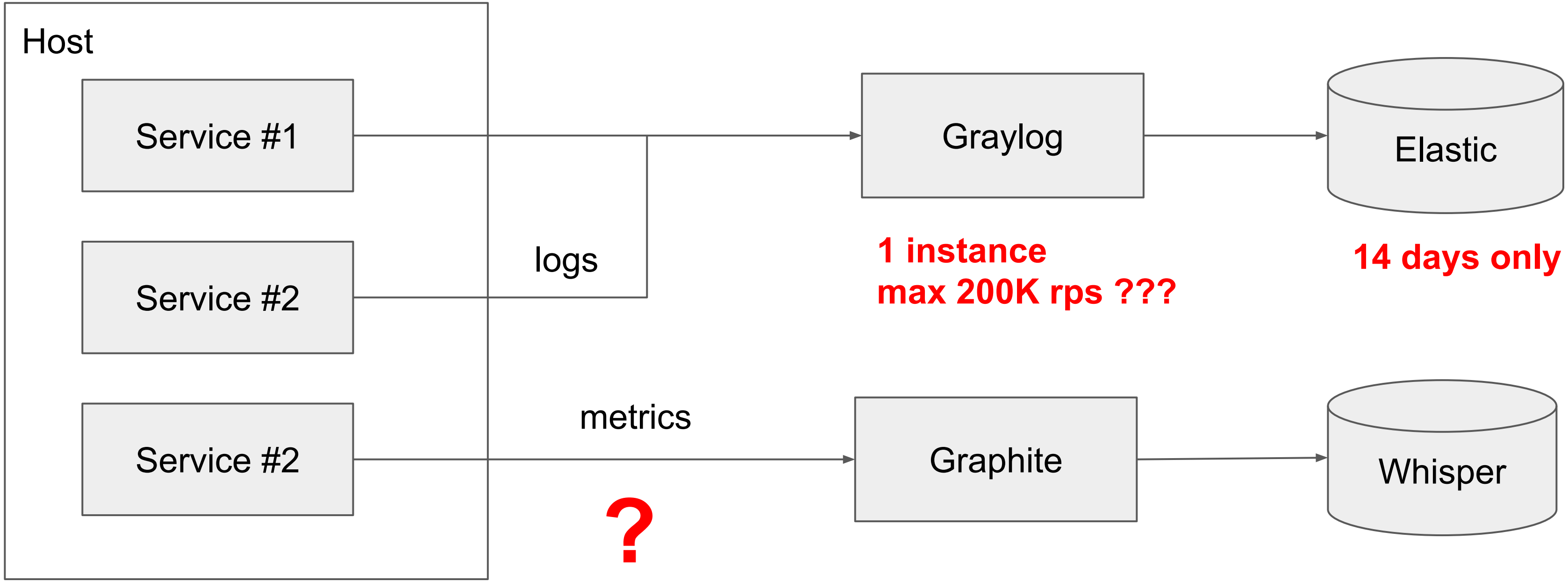
1 instance
max 200K rps ???

14 days only

2000+ services ???

Throttling ?
Maintenance ?
Long term solution for 1 year

Long time ago)



1 instance
max 200K rps ???

14 days only

2000+ services ???

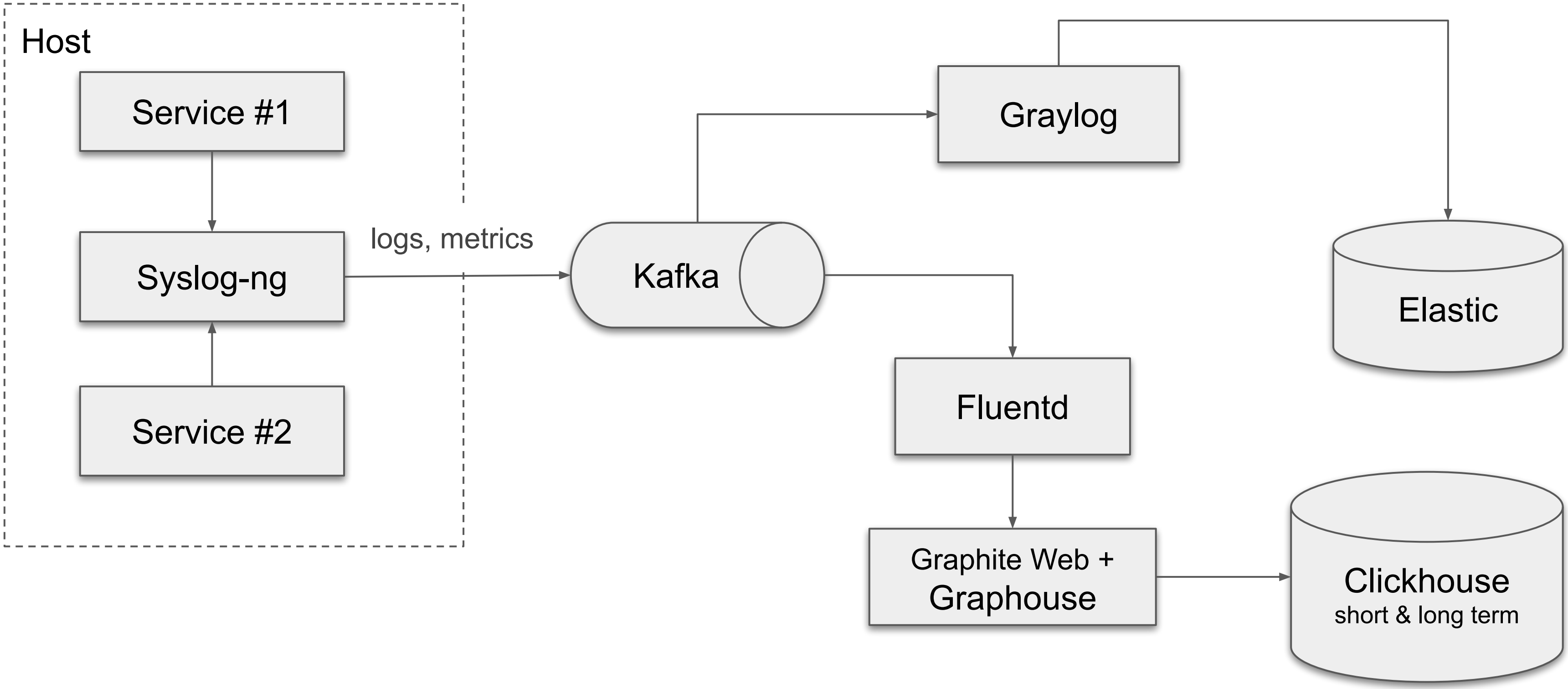
Throttling ?
Maintenance ?
Long term solution for 1 year

Disadvantages

- **UDP: Throttling ?**
- **Rate: 100+ Krps how ?**
- **Elastic: huge load and a lot of resources**
- **Graylog: Java, fixed load on each node**
- **Retention policy: 1 year or more**

1st implementation (Rancher)

1st implementation (Rancher)



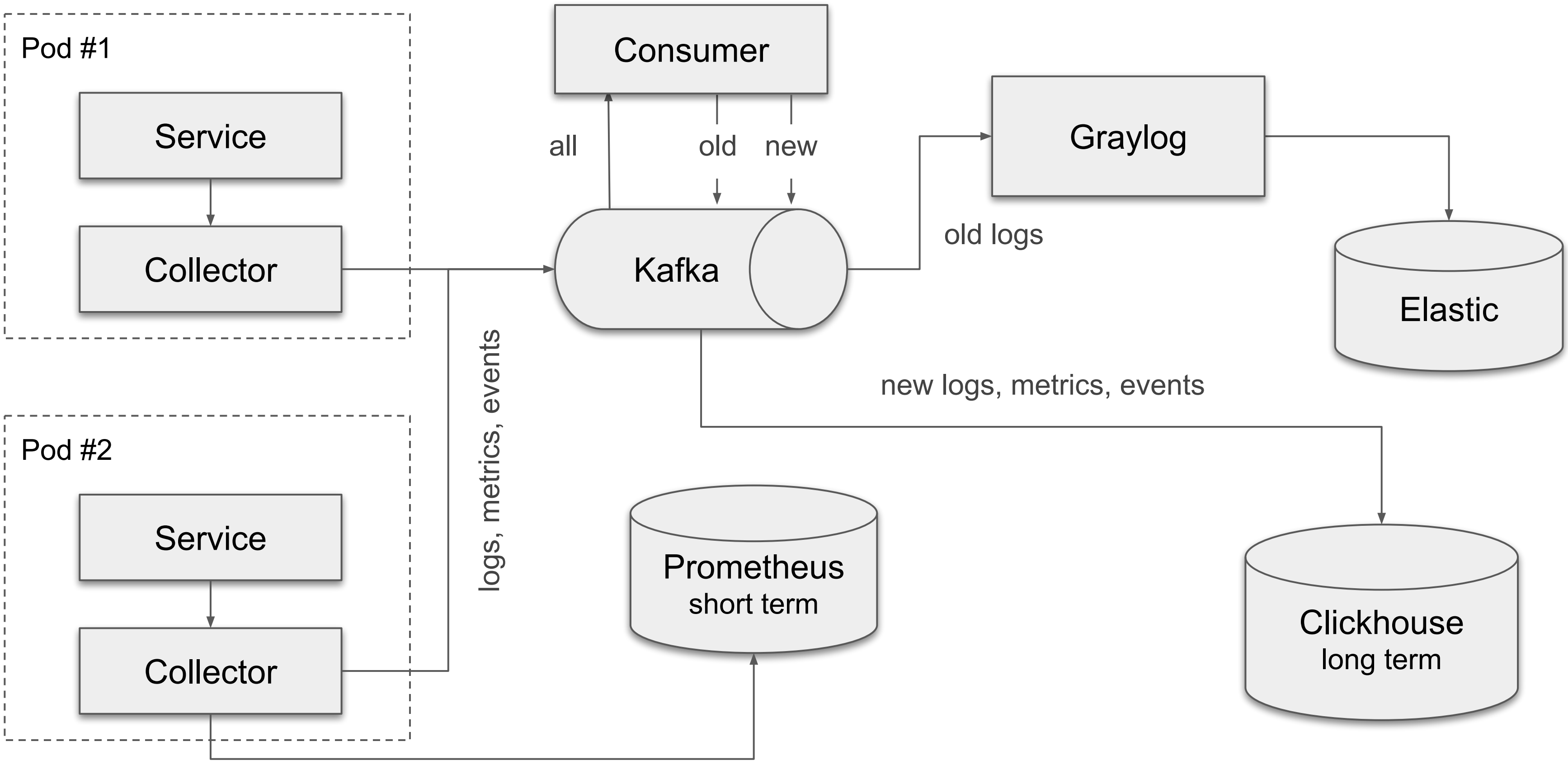
1nd implementation (disadvantages)

- Lack of throttling on service side
- Not supported for tags (Graphite, Grafana, Graphouse)
- Custom message format (hard to understand)
- Syslog-ng (one point of failure on host, module is written on Java)
- Fluentd pitfalls (offsets, not supported groups, configuration)
- Graphite Web is too slow (issue with long term queries)
- Graphouse is fast (but still written on Java, consumes memory)

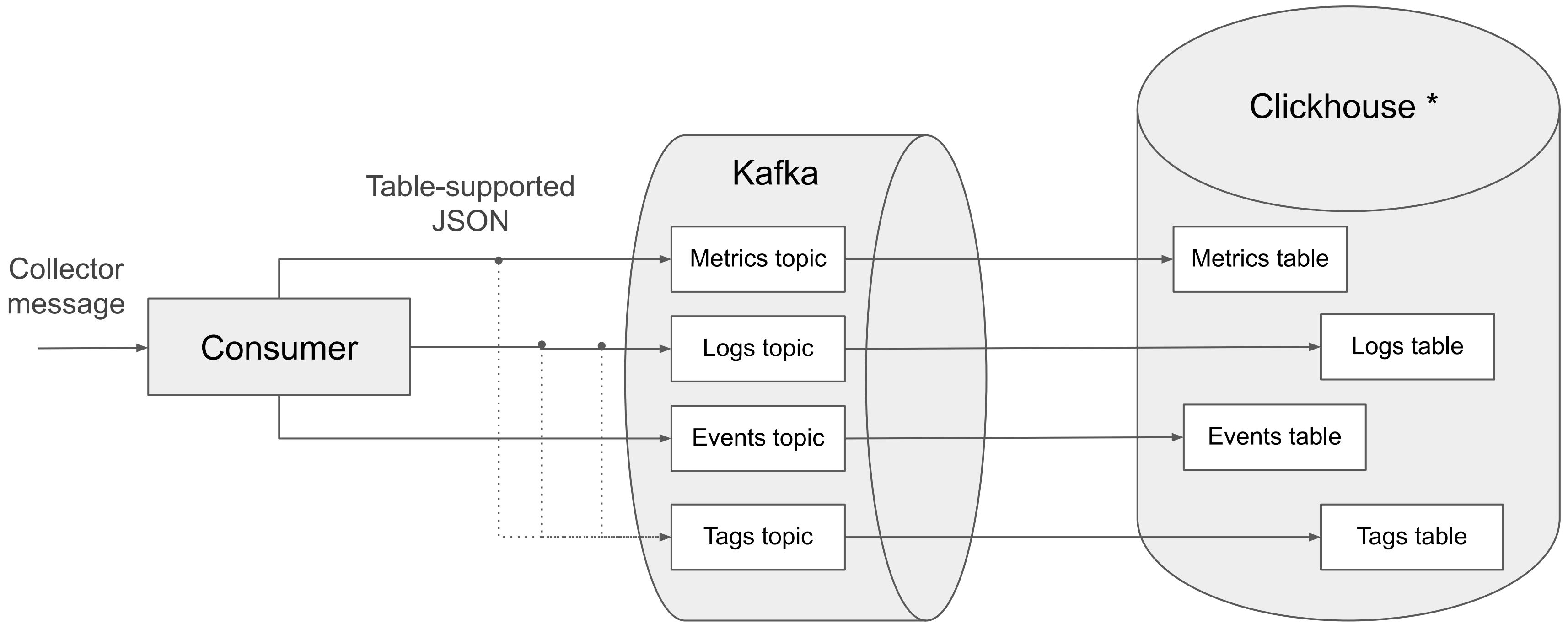
trend to next K8s...

2nd implementation (K8s)

exness



Transformation by Consumer



* K8s namespace-related database with it's own metrics, logs, events (ReplicatedMergeTree) and tags (ReplicatedReplacingMergeTree)

2nd implementation (advantages)

- No one point of failure (collector on board)
- Fast and robust (Collector & Consumer have written on Golang)
- Throttling on service side
- Standard message format & version support (Telegraf)
- Tags support across logs, metrics and events
- Graylog + Elastic for logs, Prometheus for metrics (short term)
- Clickhouse for long term (metrics, logs, events)

Thank you!

Questions?

exness

exness