2023 SUMMIT | APACHE • SkyWalking
CHINA · SHANGHAI

2023 · 上海
SkyWalking Summit

纵目 tetrate

演讲主题

高洪涛

Tetrate创始工程师 、 Apache SkyWalking核心贡献者

BanyanDB

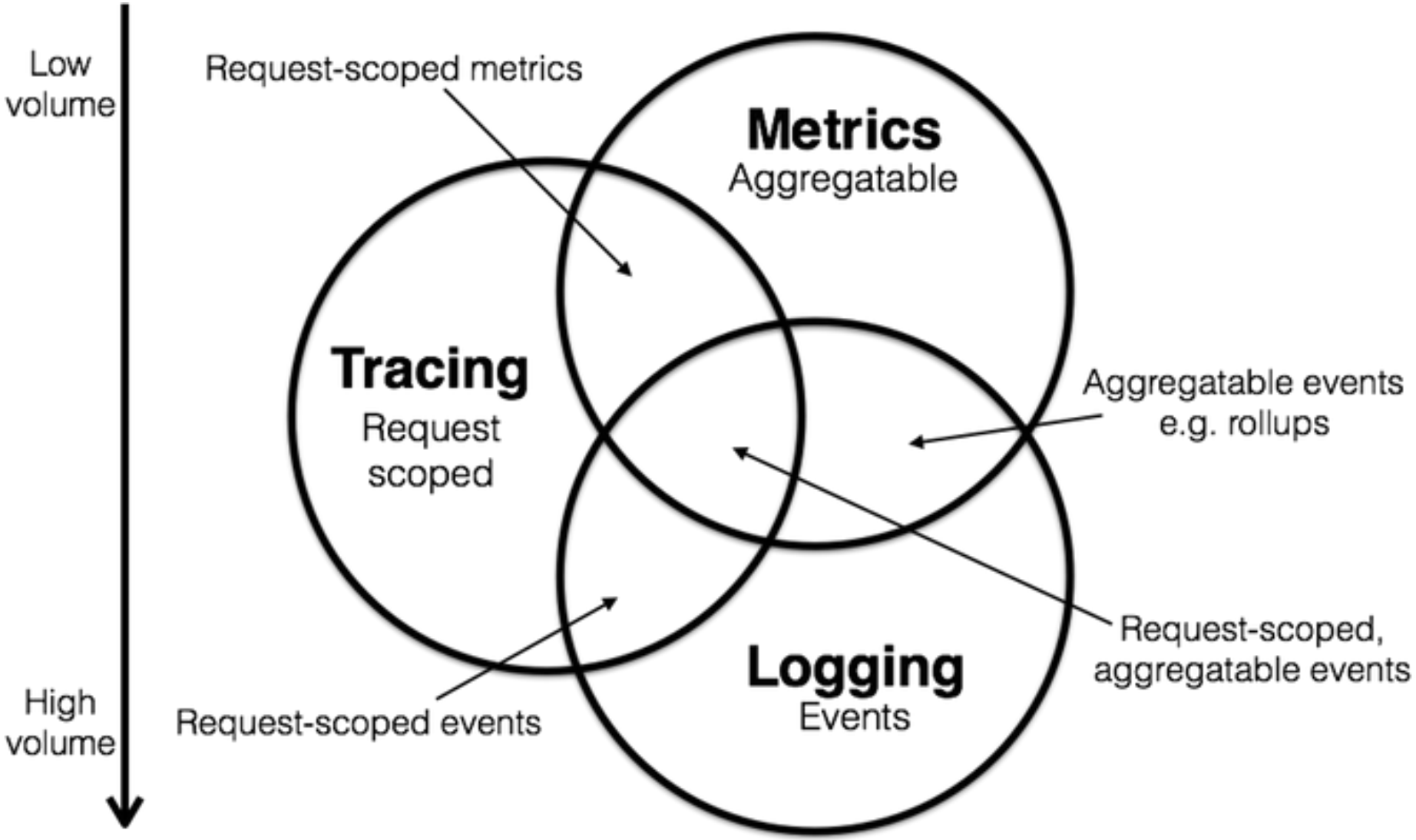
Cloud Native Observability Database

目录
CONTENTS

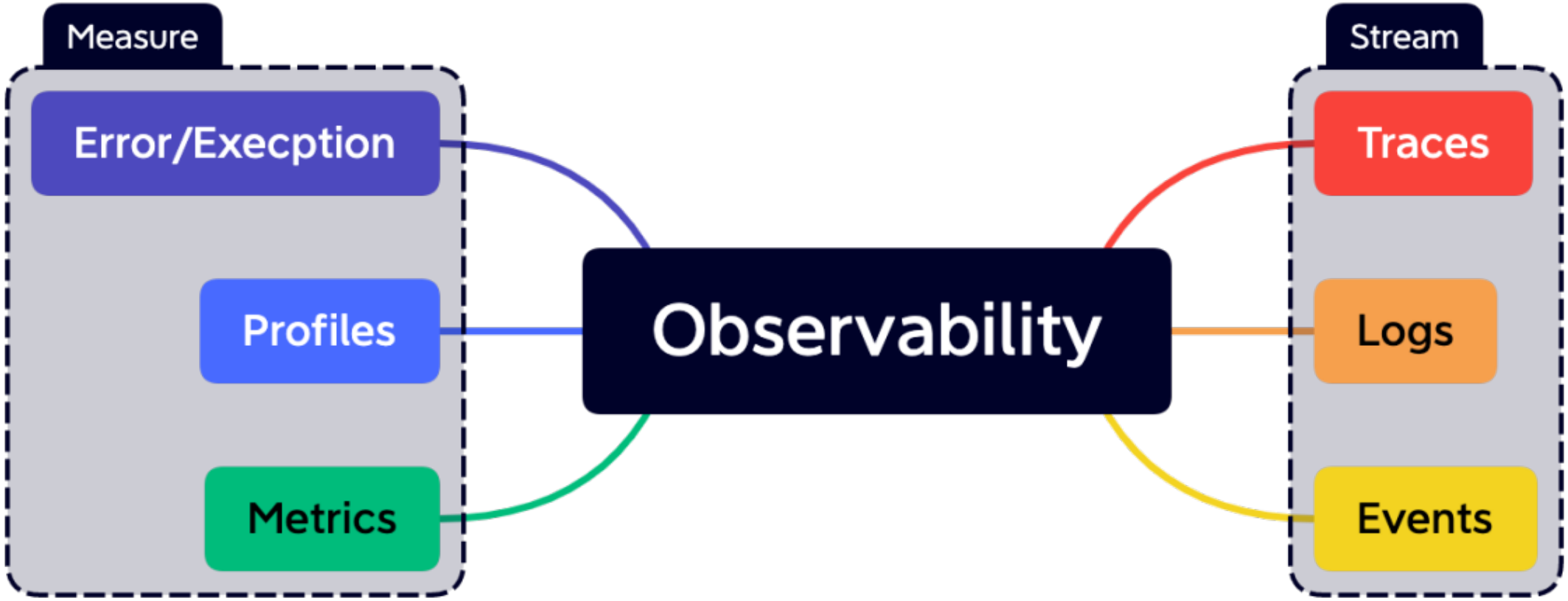
01

Observability & TSDB

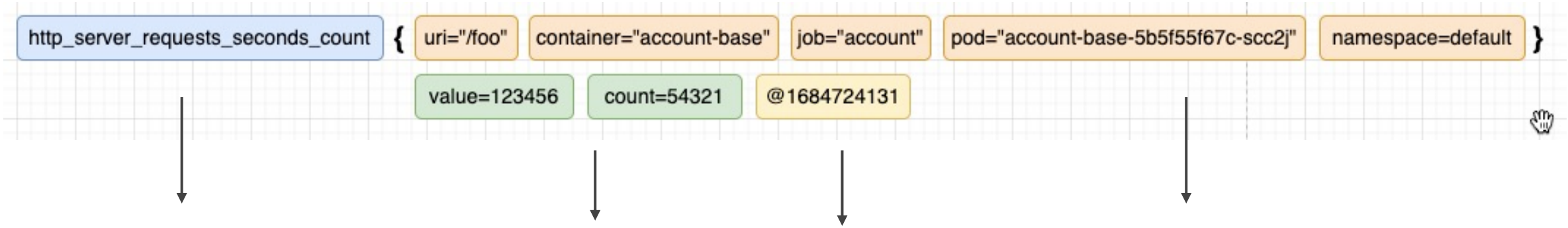
Three Pillars



More Models


Measure
Error/Execption
Profiles
Metrics
Observability
Stream
Traces
Logs
Events

Tags & Fields



metric_name (label: __name__)

Fields: single-valued or multiple values

Timestamp: epoch

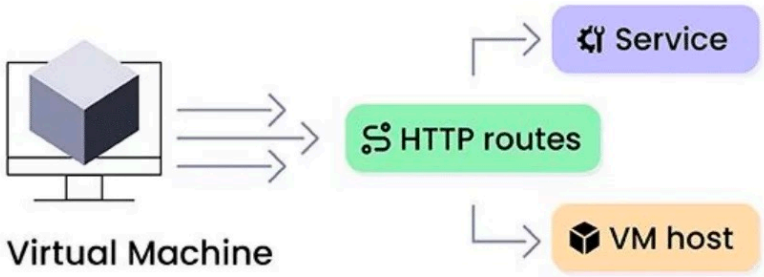
LabelSet (Tags)

```
samples (
  metric_name String,
  labels String,
  timestamp Int64,
  value Float64
) ORDER BY (metric_name, labels, timestamp)
```

High Cardinality

How Quickly Does Cardinality Grow?

VIRTUAL-MACHINE BASED ENVIRONMENT

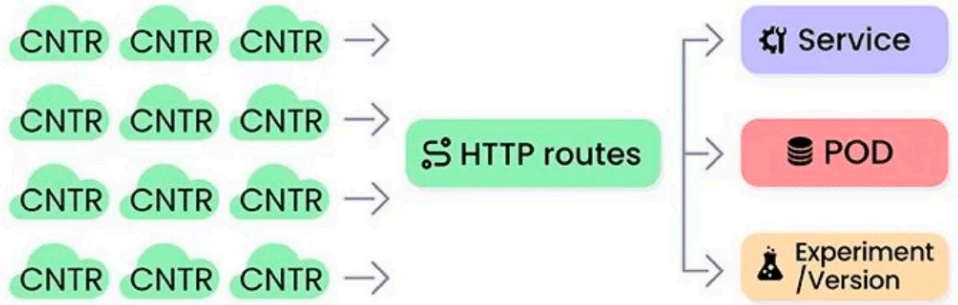


20 HTTP routes
5 HTTP status code types
5 services
300 Vms

CARDINALITY

150 thousand
possible unique
time series

CLOUD-NATIVE ENVIRONMENT



20 HTTP routes
5 HTTP status codes types
5 services
12,000 pods (40x VMs)
25 experiments / versions

CARDINALITY

150 million
possible unique
time series

chronosphere

02

Glance of BanyanDB

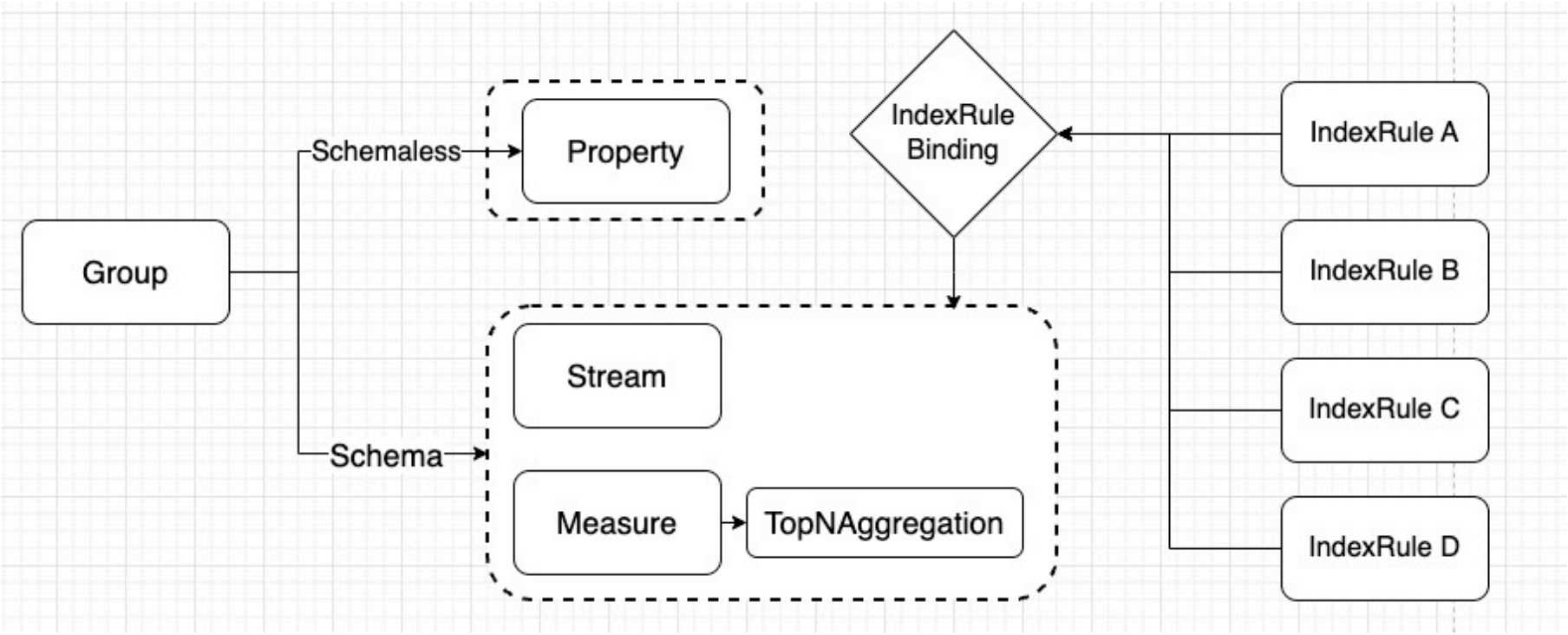
Logical Data Model

Pros:

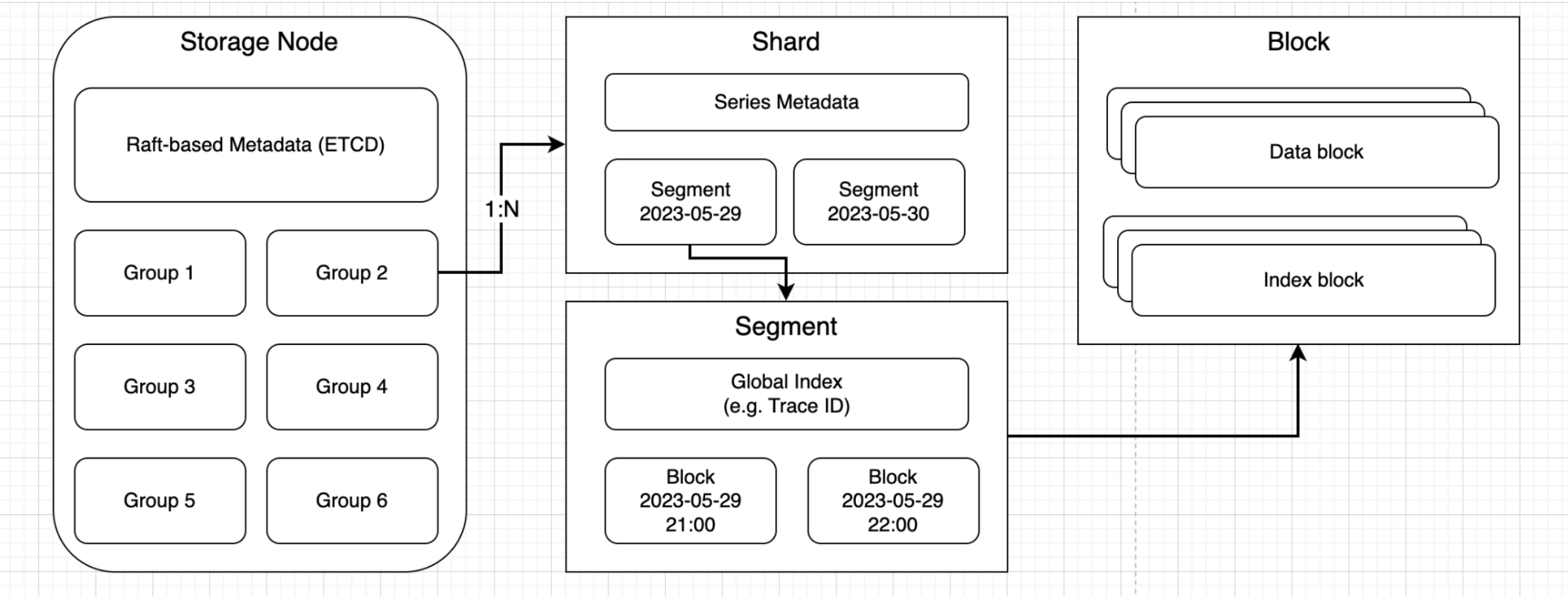
- Reduce High Cardinality
- Validation
- Querying
- Integration

Cons:

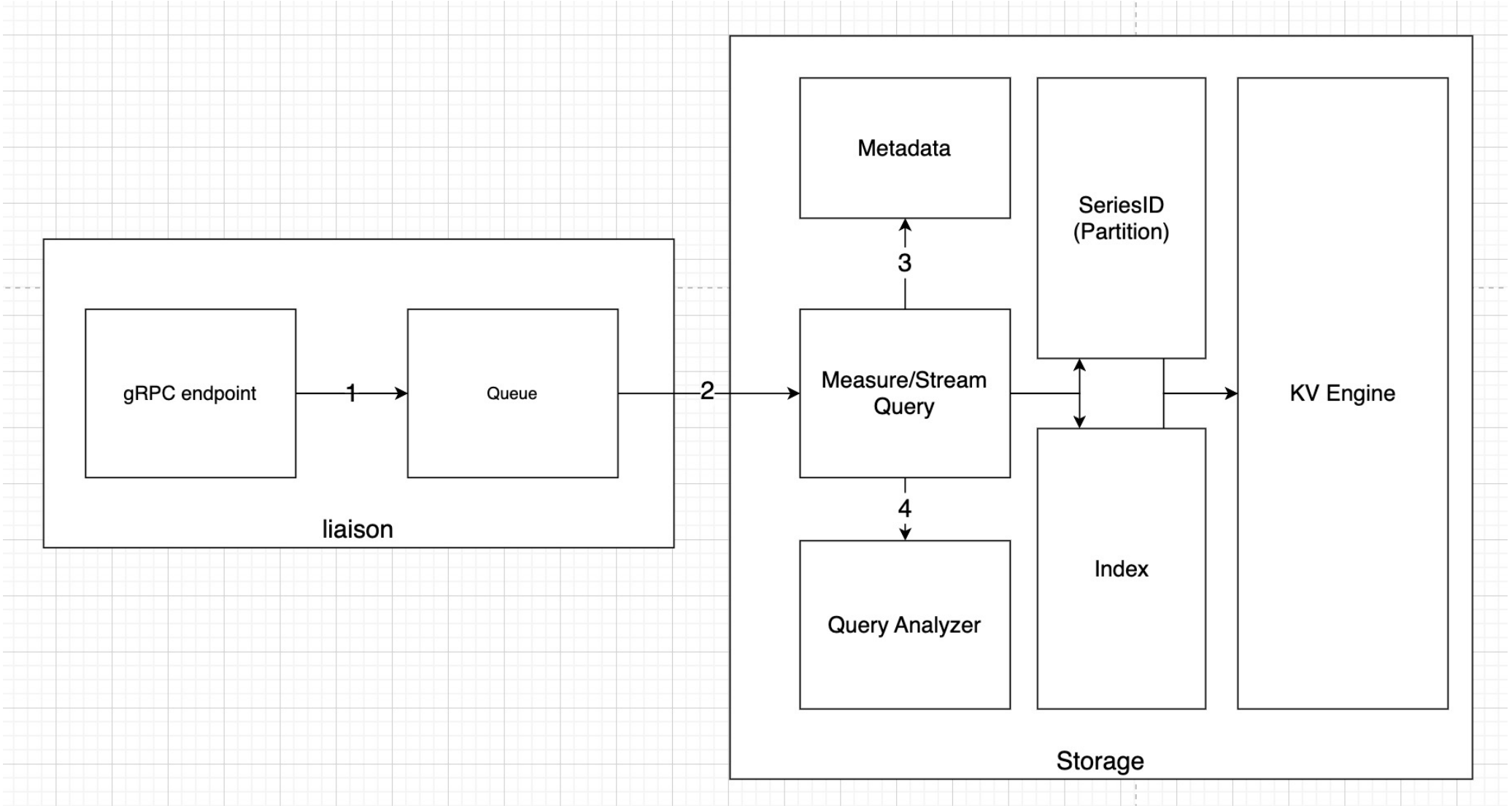
- Complexity
- Overhead
- Maintenance



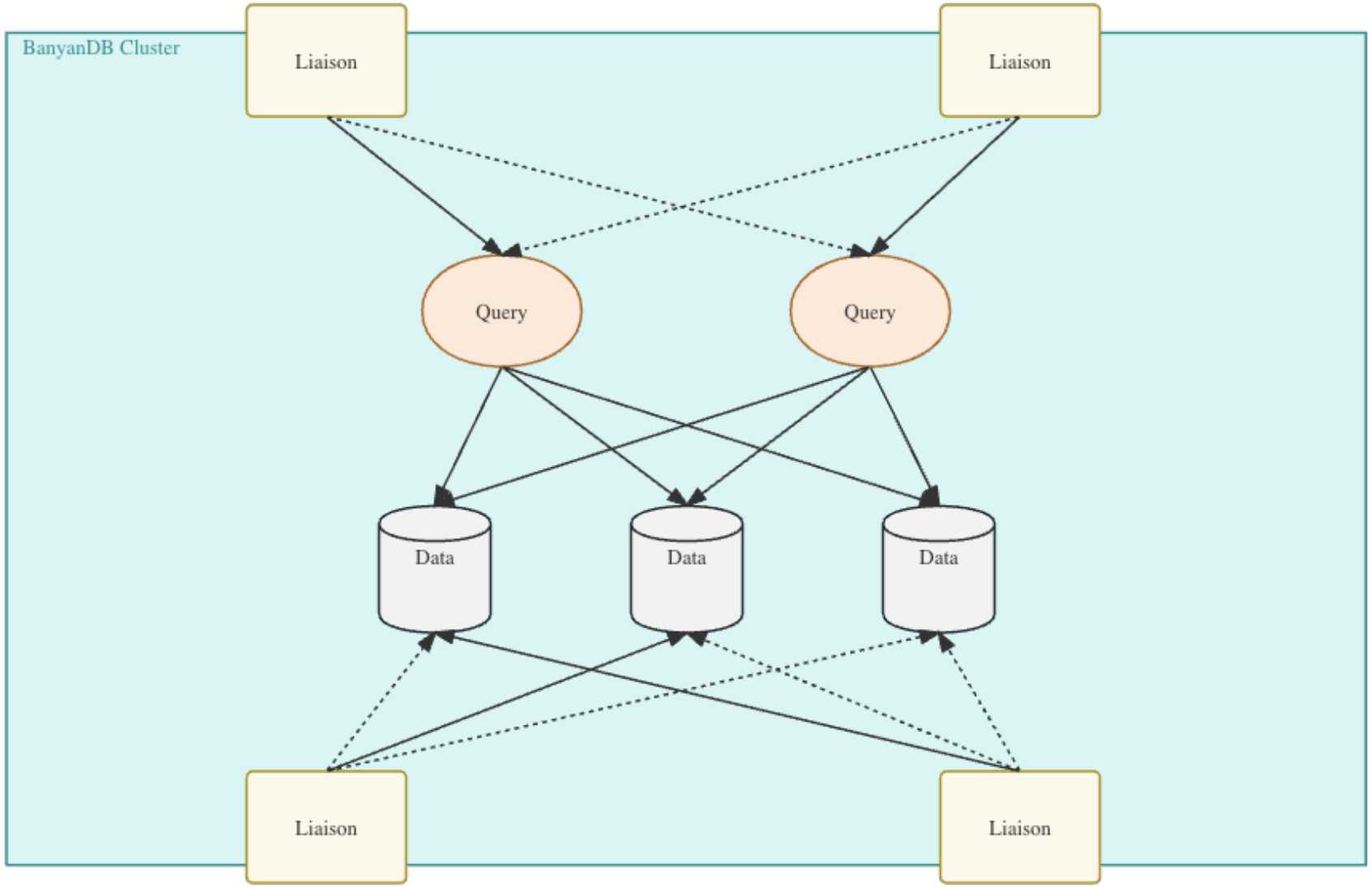
Physical Data Model



Arch


gRPC endpoint
1
Queue
liaison
2
Metadata
3
Measure/Stream Query
4
Query Analyzer
SeriesID (Partition)
Index
KV Engine
Storage

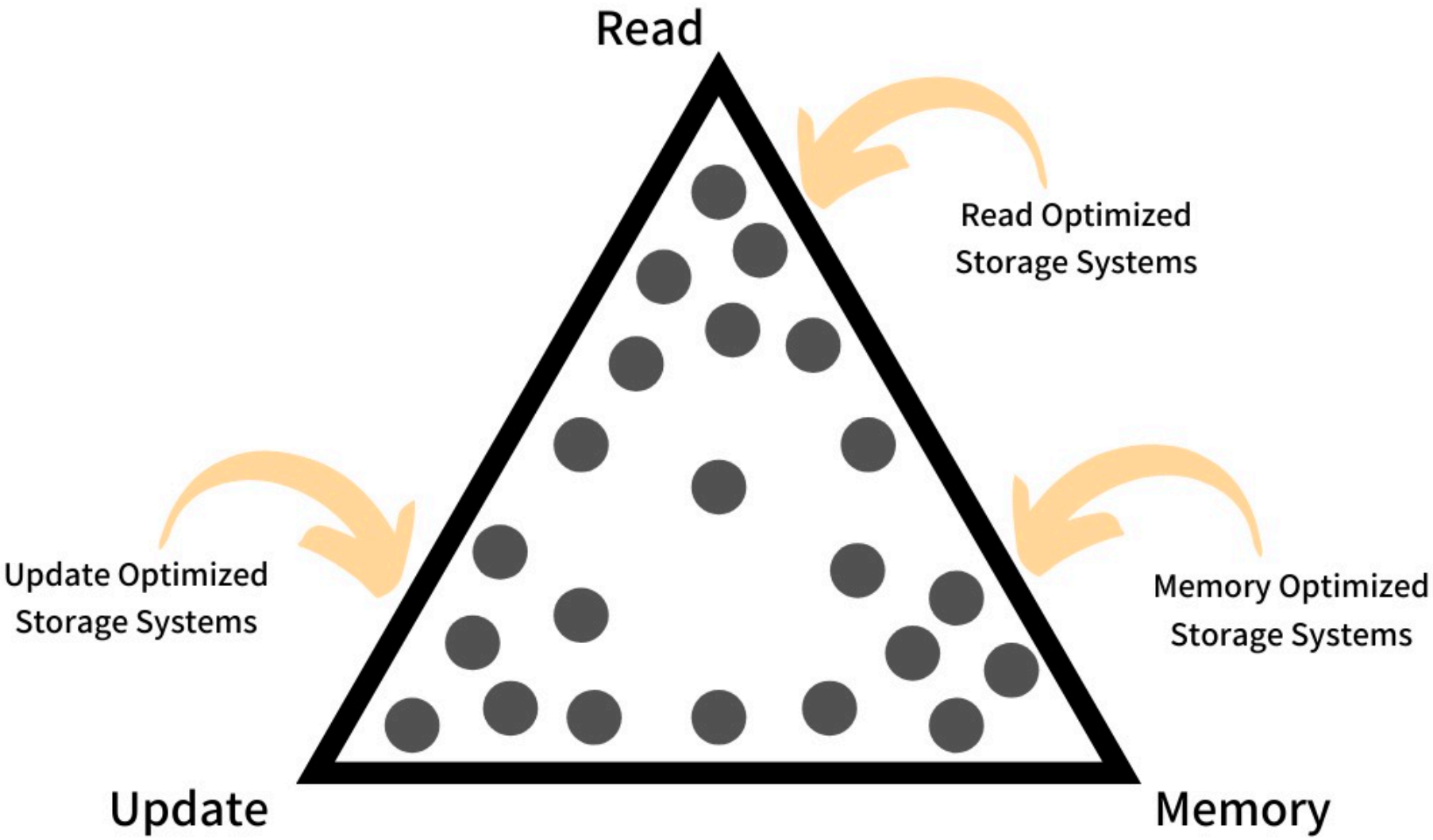
Distributed



03

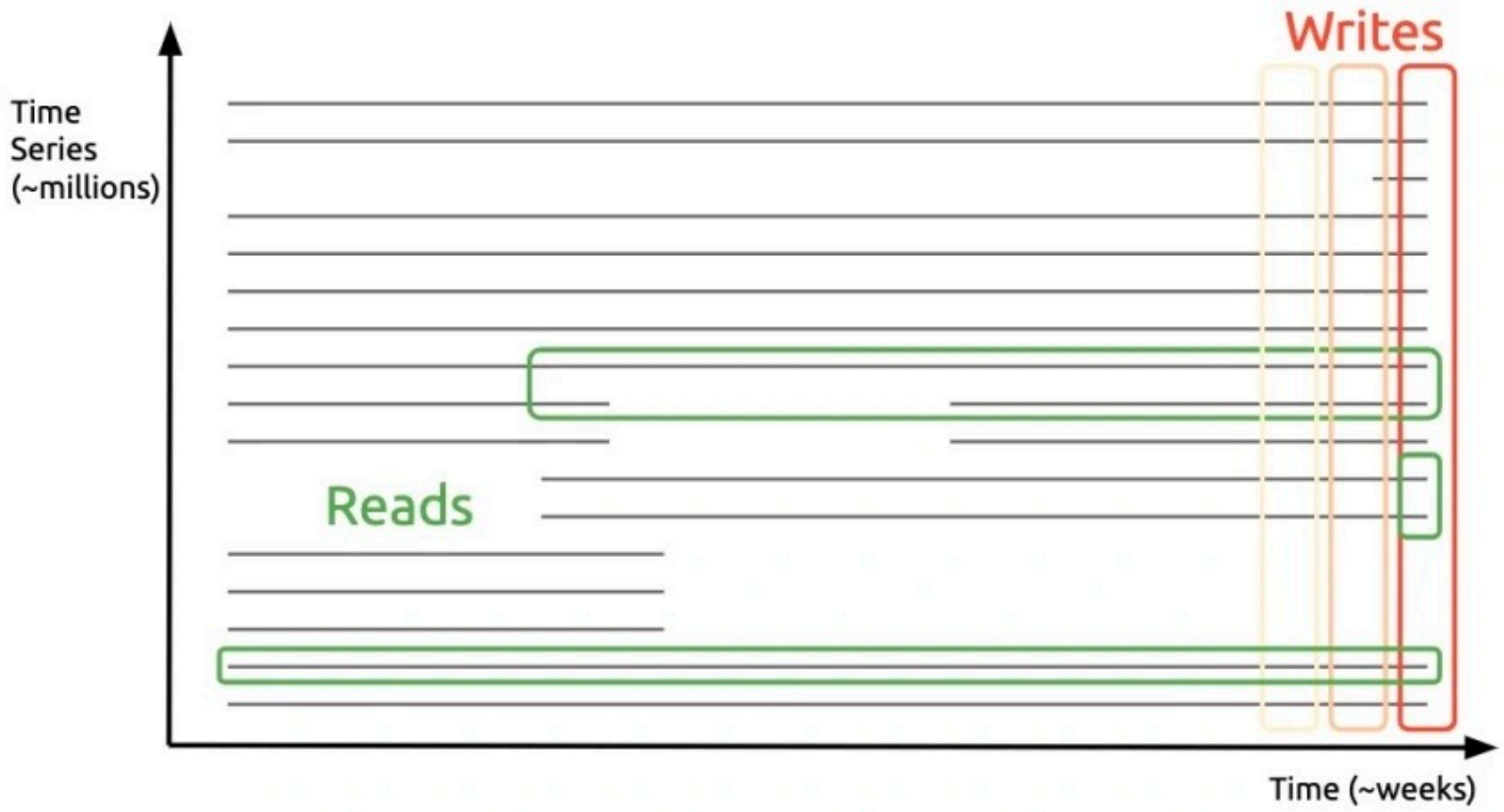
Data Access

Rum Conjecture



BanyanDB Access Method

- Vertical write
- Horizontal read
- Insertions >> lookups
- Old data is less likely to be access



Compression based on Series

{component="printer",location="f2c16",level="error"} "Printing is not supported by this printer"

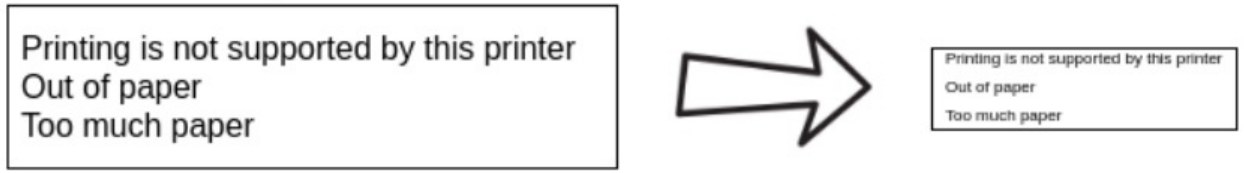
Label key/values hashed to form Stream ID: 3b2cea09797978fc

The log entry is added to a "chunk"

Additional log messages with the same labels are added to the same "chunk":

{component="printer",location="f2c16",level="error"} "Out of paper"
{component="printer",location="f2c16",level="error"} "Too much paper"

Chunks are filled then compressed and stored:



A separate and small index is kept to lookup chunks

Different label keys or values will hash to a different stream and different chunk:

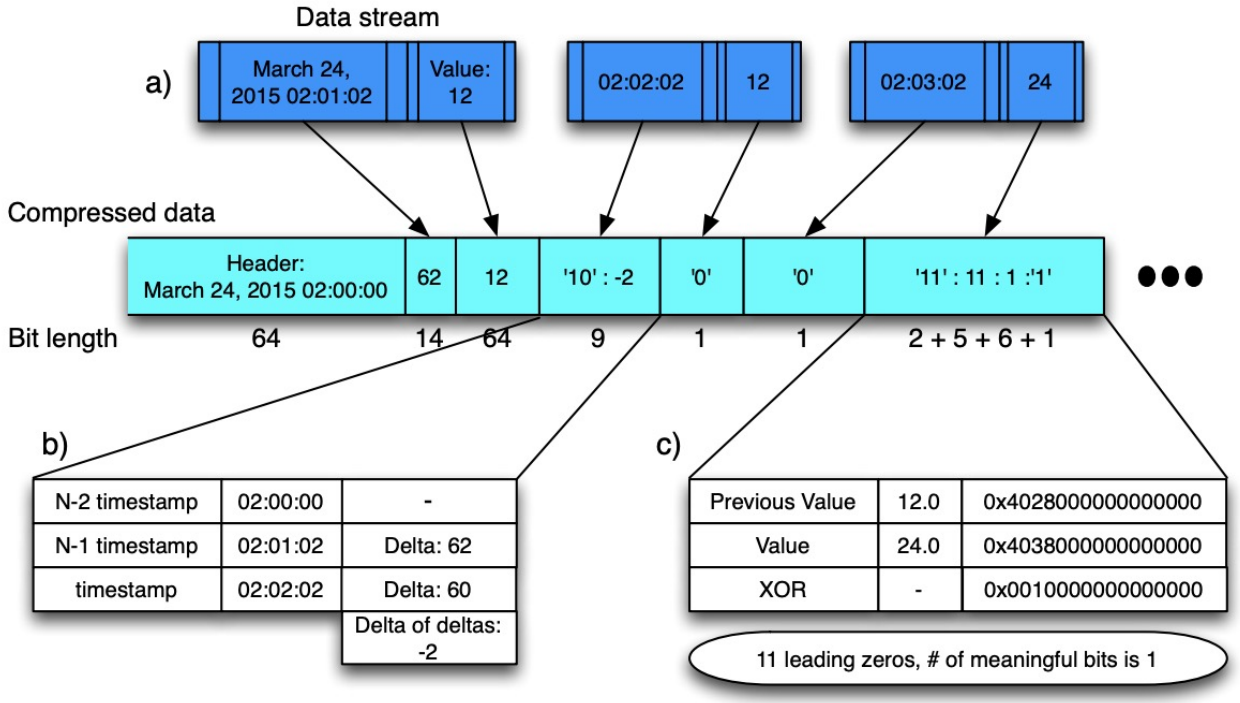
{component="printer",location="f2c16",**level="info"**} "Consider the environment before printing this log message"

fd9a709ddf43a93a

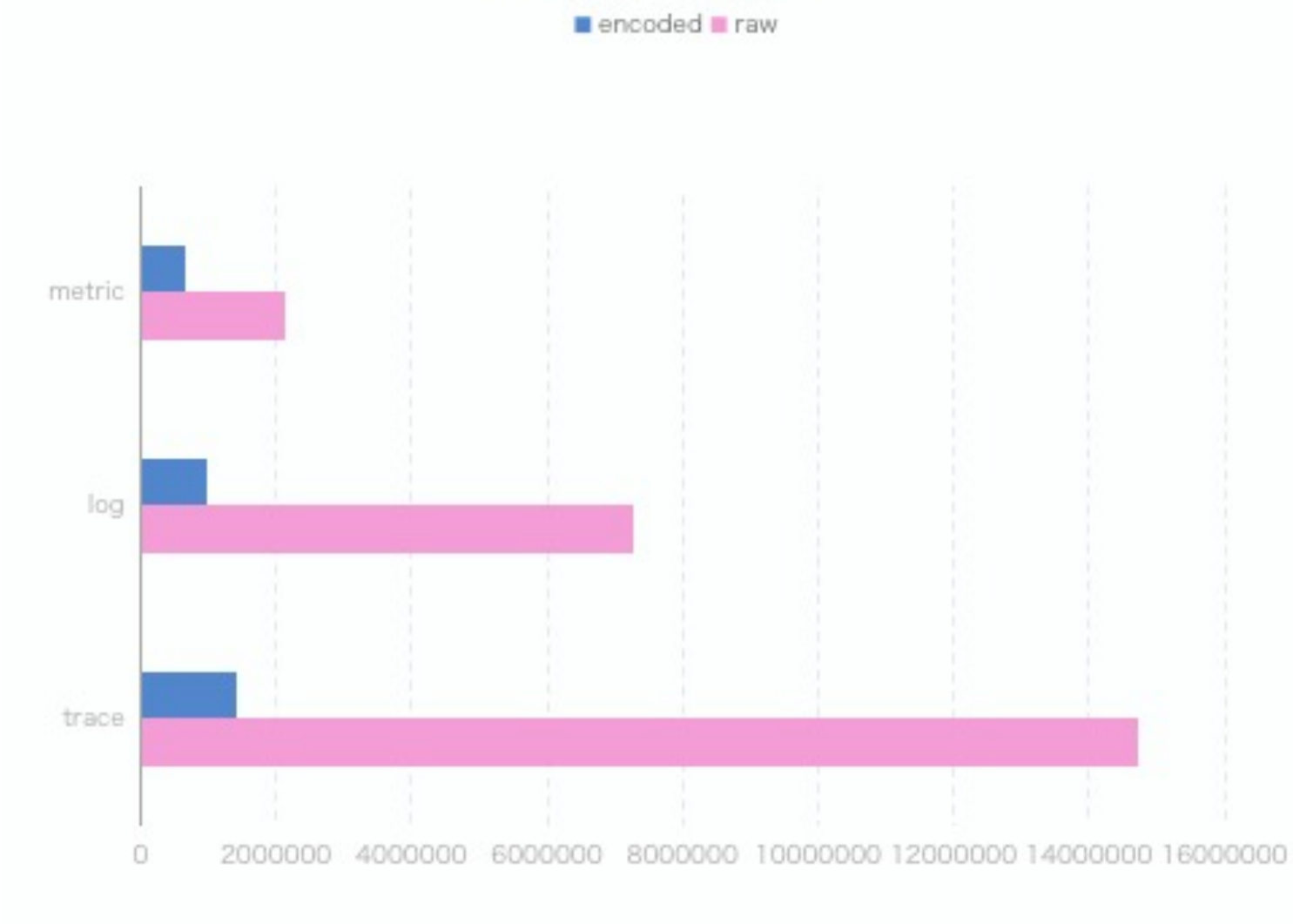
Compression Alogrithm

Enhanced Gorilla

- Facebook Gorilla:
  - Timestamp: fixed interval => derived timestamp
  - Value: XOR
- Compress big chunk



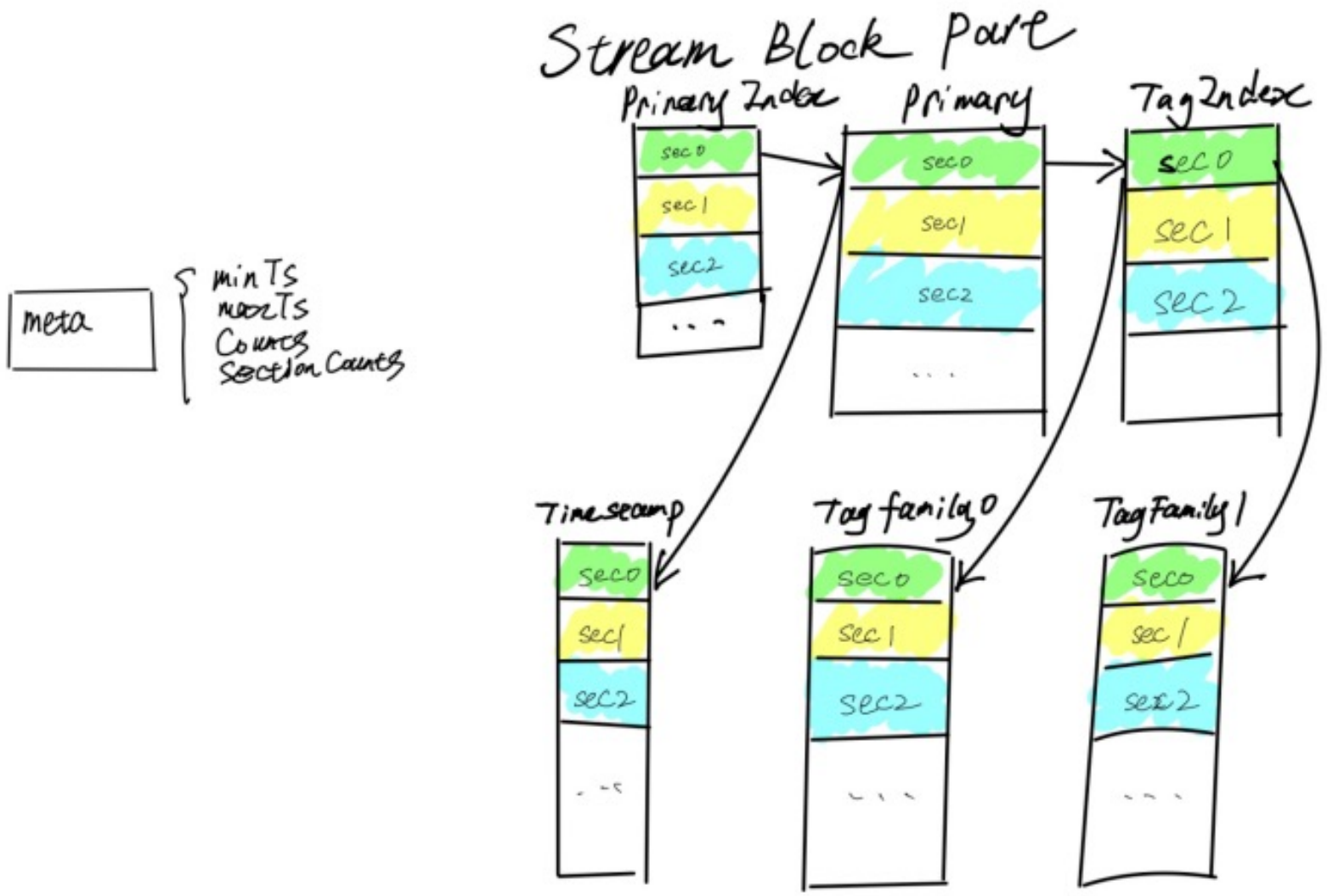
Compression Ratio

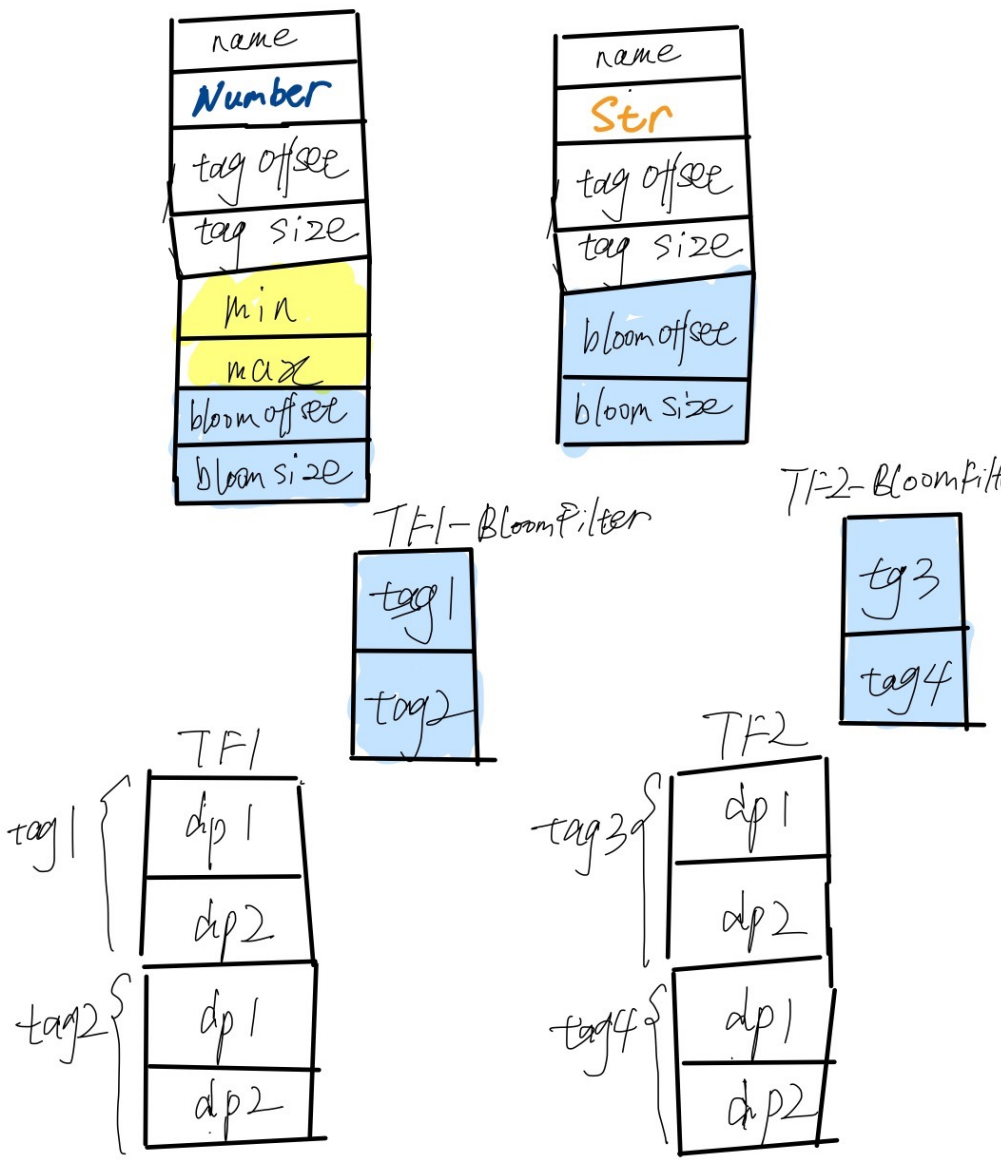


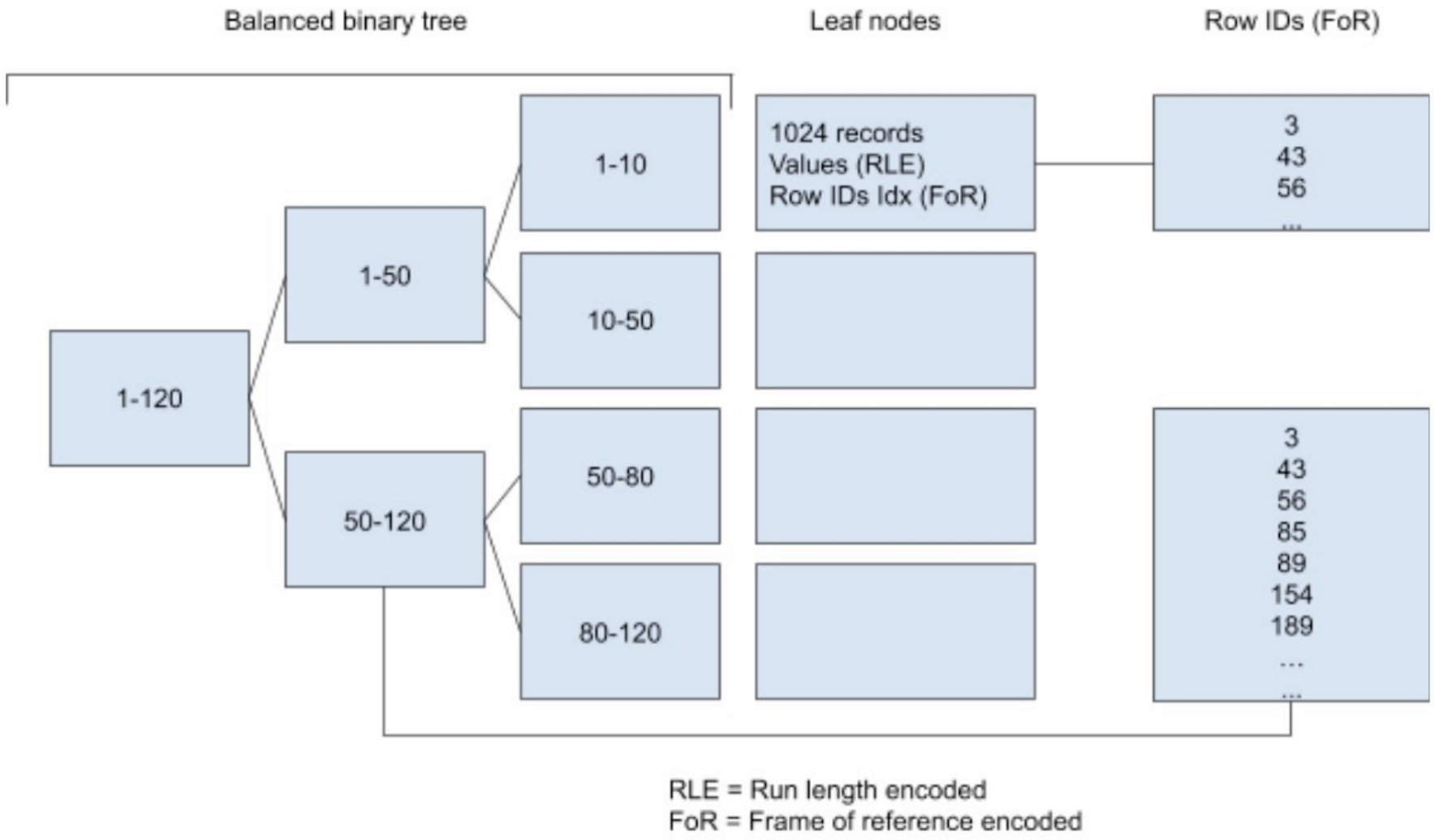
04

What’s Next

Column Storage




name
Number
tag offset
tag size
min
max
bloom offset
bloom size
name
Str
tag offset
tag size
bloom offset
bloom size
TF-1-BloomFilter
tag1
tag2
TF-2-BloomFilter
tg3
tag4
TF1
tag1
dp1
dp2
tag2
dp1
dp2
TF2
tag3
dp1
dp2
tag4
dp1
dp2


Balanced binary tree
Leaf nodes
Row IDs (FoR)
1-120
1-50
50-120
1-10
10-50
50-80
80-120
1024 records
Values (RLE)
Row IDs Idx (FoR)
3
43
56
...
3
43
56
85
89
154
189
...
...
RLE = Run length encoded
FoR = Frame of reference encoded

BanyanDB

https://github.com/apache/skywalking-banyandb/

BanyanDB

BanyanDB, as an observability database, aims to ingest, analyze and store Metrics, Tracing and Logging data. It's designed to handle observability data generated by observability platform and APM system, like Apache SkyWalking etc.

Introduction

BanyanDB, as an observability database, aims to ingest, analyze and store Metrics, Tracing, and Logging data. It's designed to handle observability data generated by Apache SkyWalking. Before BanyanDB emerges, the Databases that SkyWalking adopted are not ideal for the APM data model, especially for saving tracing and logging data. Consequently, There's room to improve the performance and resource usage based on the nature of SkyWalking data patterns.

Q&A

欢迎提问交流

（仅限2位提问）

2023 SUMMIT | APACHE • SkyWalking CHINA · SHANGHAI

2023 · SkyWalking Summit

感谢您的观看

纵目 tetrate