Mobile Malware Analyze System using ElasticSearch

Mobile Threat Response Team

White Li

About Me

- 李啸（White Li）.
- Developer & Ops @TrendMicro.
- Response for doing DevOps Job of Trend's Mobile Threat Detection Expert Distributed System.
- WeChat: storm_spark
- white_li@trendmicro.com.cn
- https://github.com/swordsmanli
- Based in Nanking.

TREND MICRO

Topics

- **Overview – Malware Detection System**

Introduction -- What We Do.

Before -- With out Elasticsearch.

After -- Based on Elasticsearch.

Scenarios Introduction.

Data Levels.

- **SQL On Elasticsearch**

Features -- Multi-Tables join / Cache / Priority Scan.

- **User Experiences**

Deploy Topo. -- Optimization.

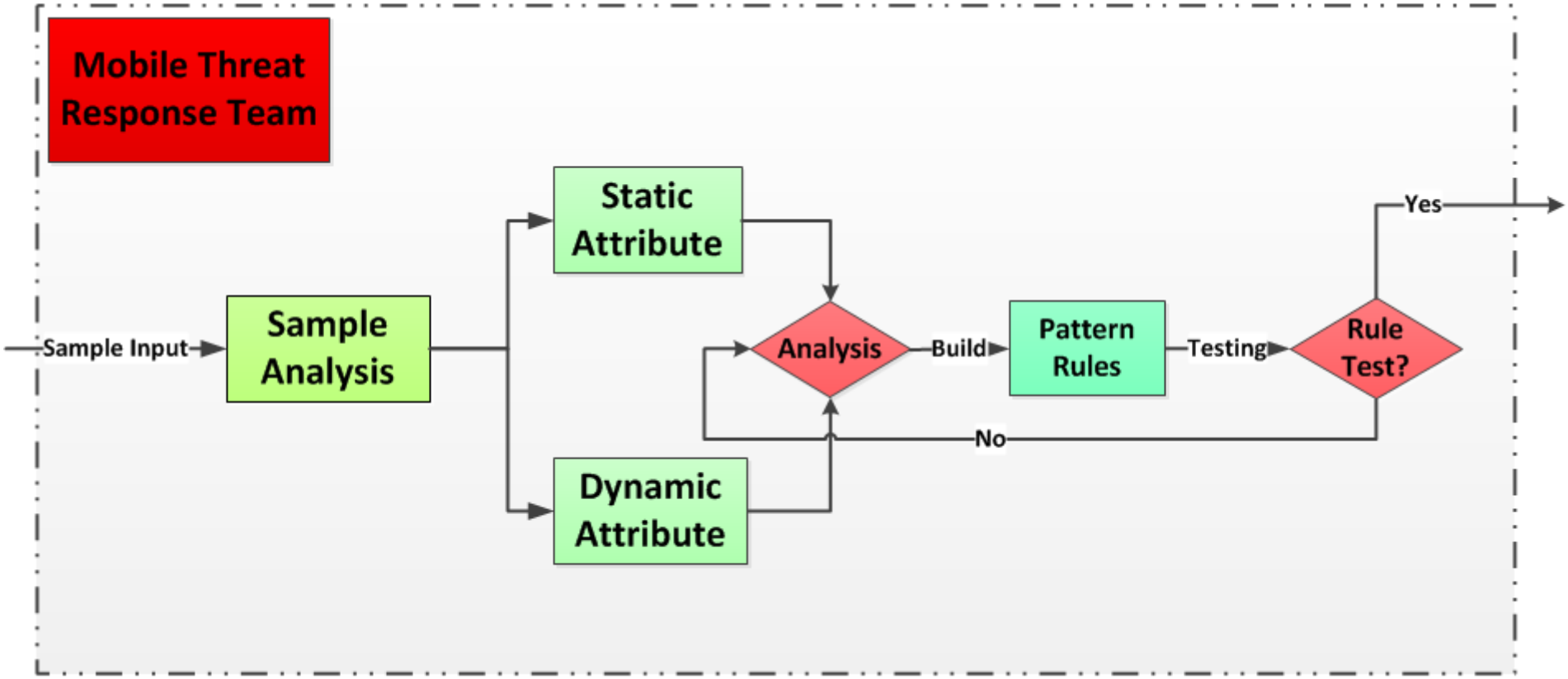
Ops Related. -- JVM Heap ? Admin Tools ?

Database Partition. -- Pros and Cons.

- **ES on Spark**

NRT indexing and search?

What We Do

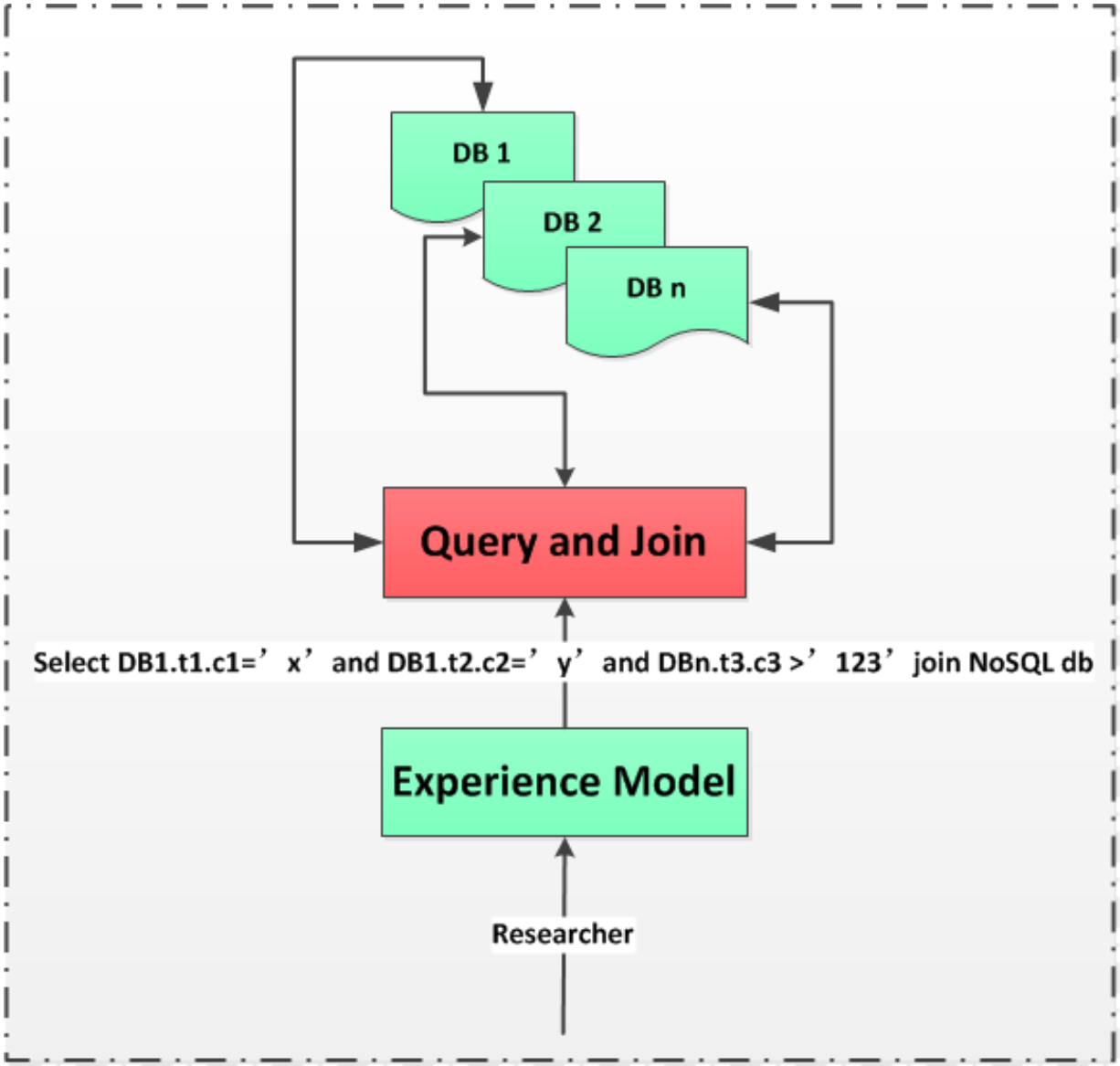


We talk about?

- **Attribute Analysis**
- **Rule Test**

TREND MICRO

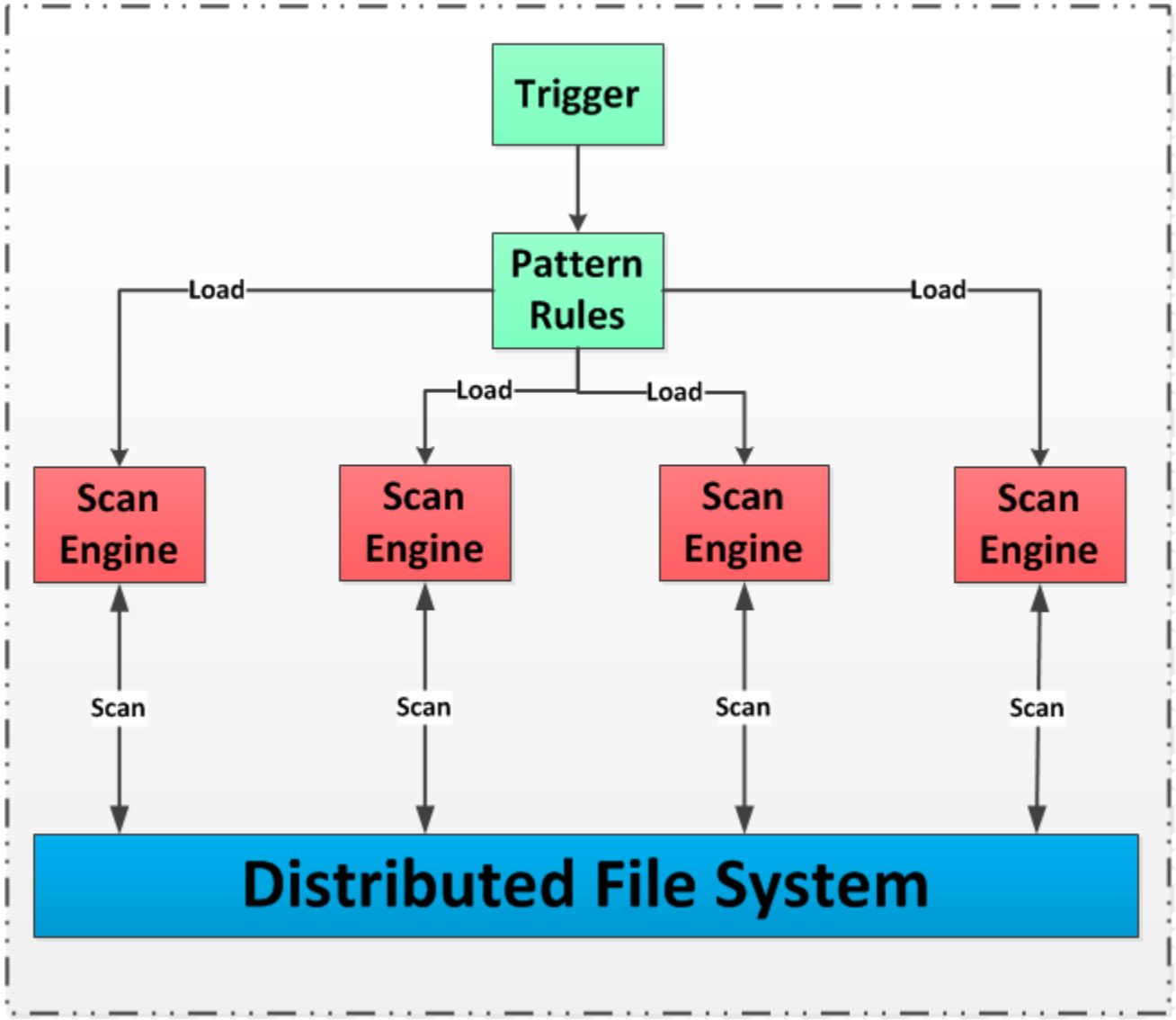
Before—Analysis?



Drawbacks:

- Complexity
- Performance
- Opaque
- Maintenance

Before—Rule Testing?

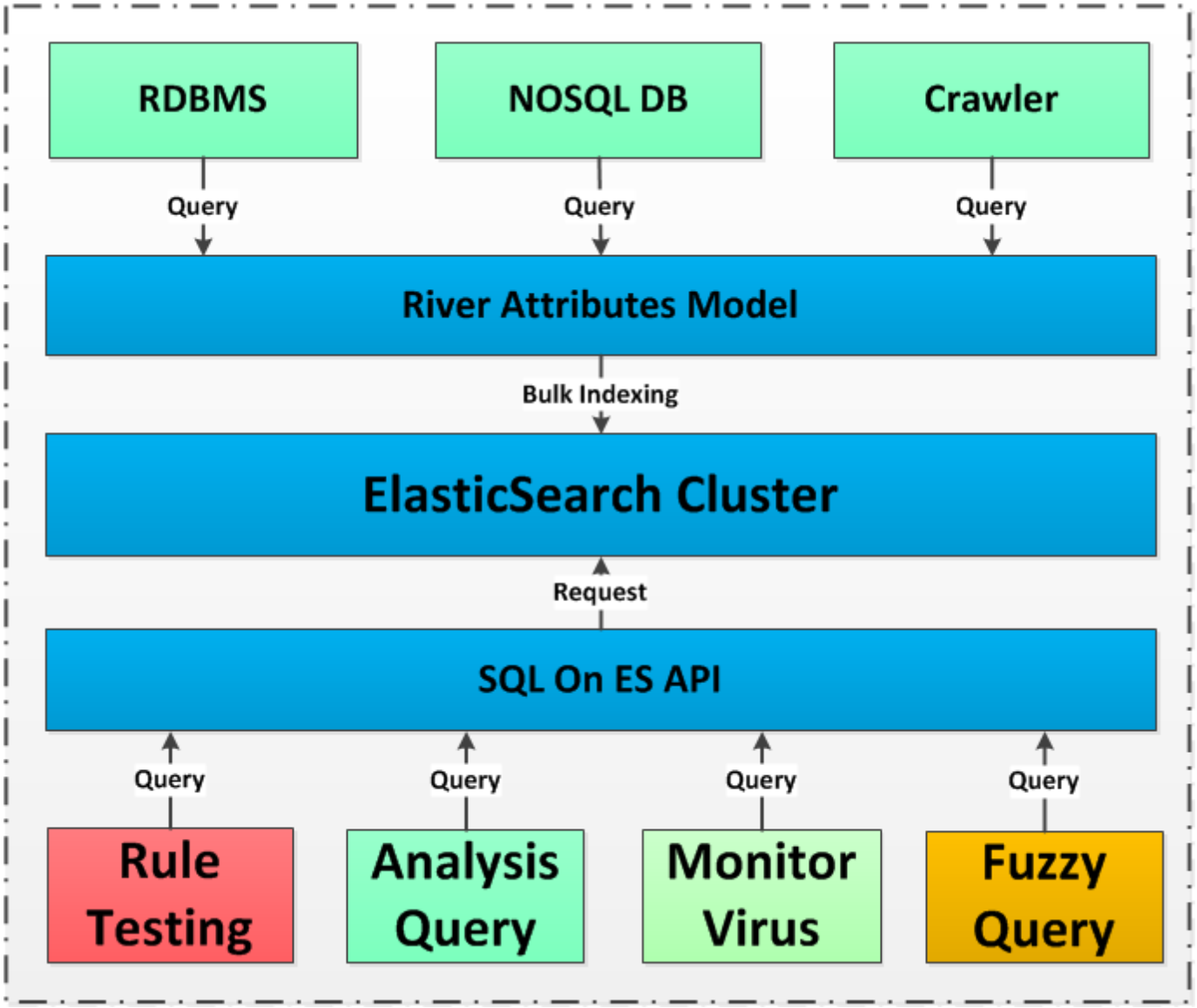


Drawbacks:

- Performance
- Isolation
- Dependency

TREND MICRO

After—Elastic Search Engine

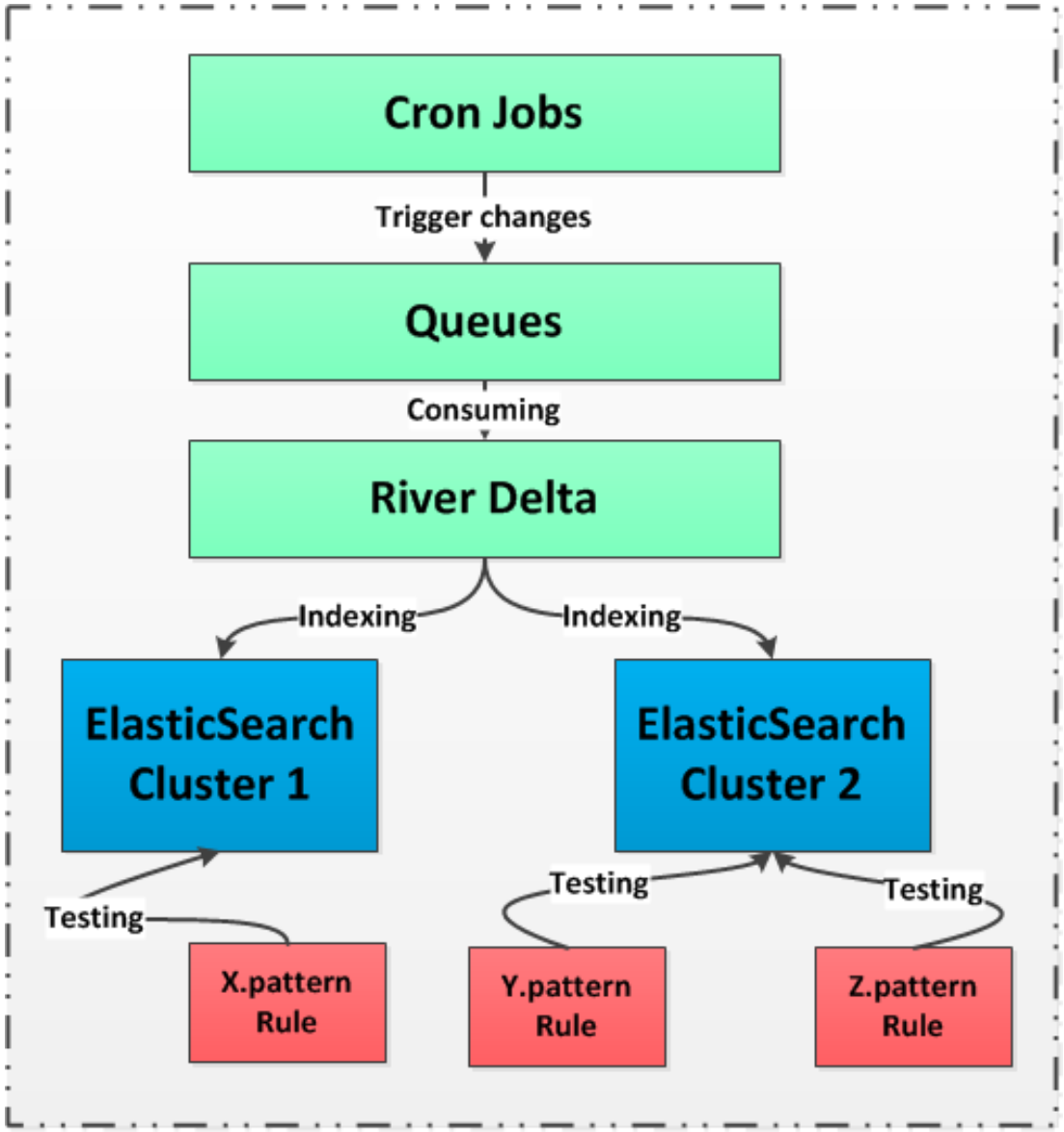


Advantages:

- Performance
- Hardware
- Simplicity
- Scalability
- Fault-Tolerant

TREND MICRO

Scenario—Rule Testing



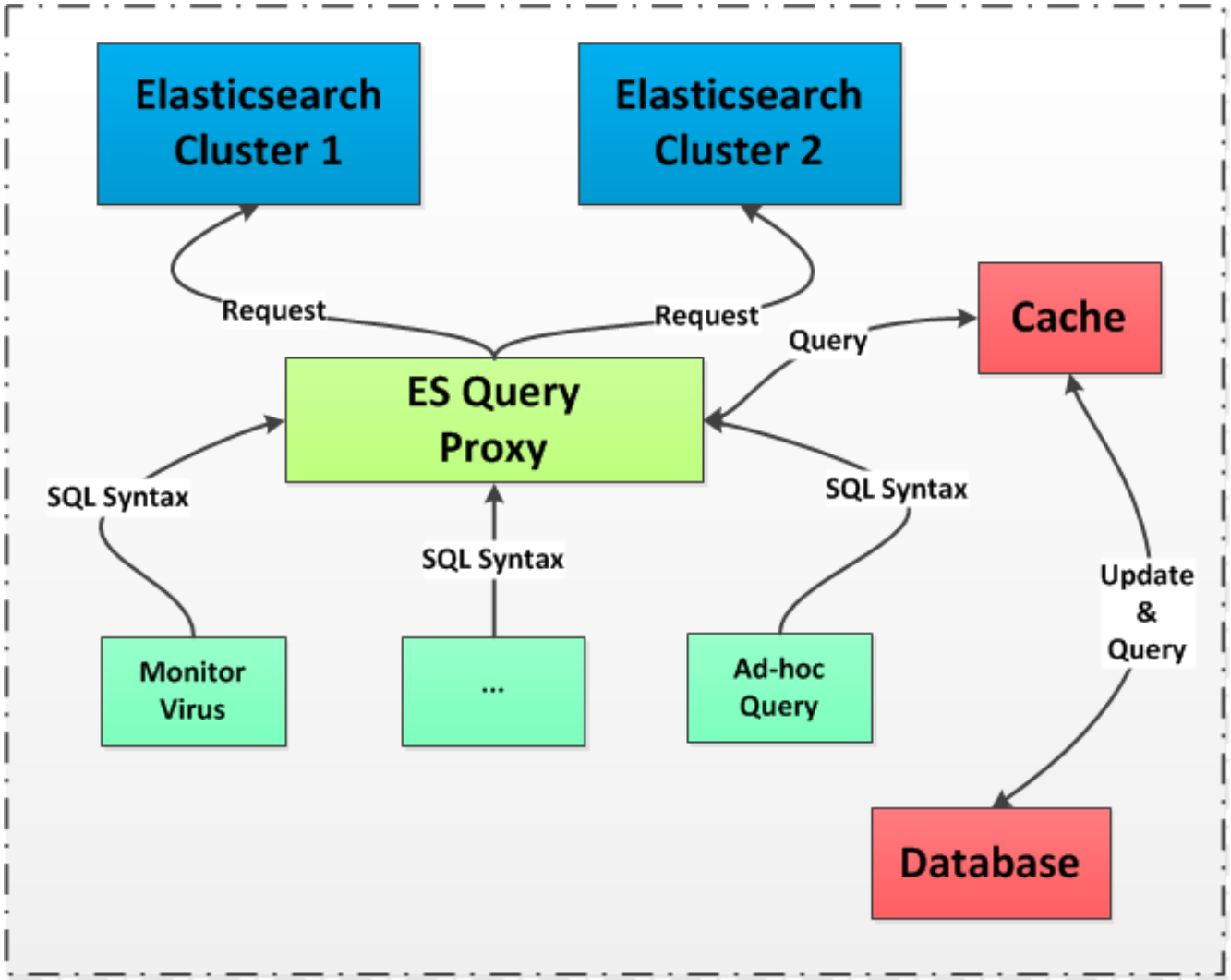
Challenges:

- Fast
- Consistency
- Multi-task

Solutions:

- ES inverted index
- MQ
- ES concurrency

Scenario—Common Query



Challenges:

- Fast
- Simple
- Join

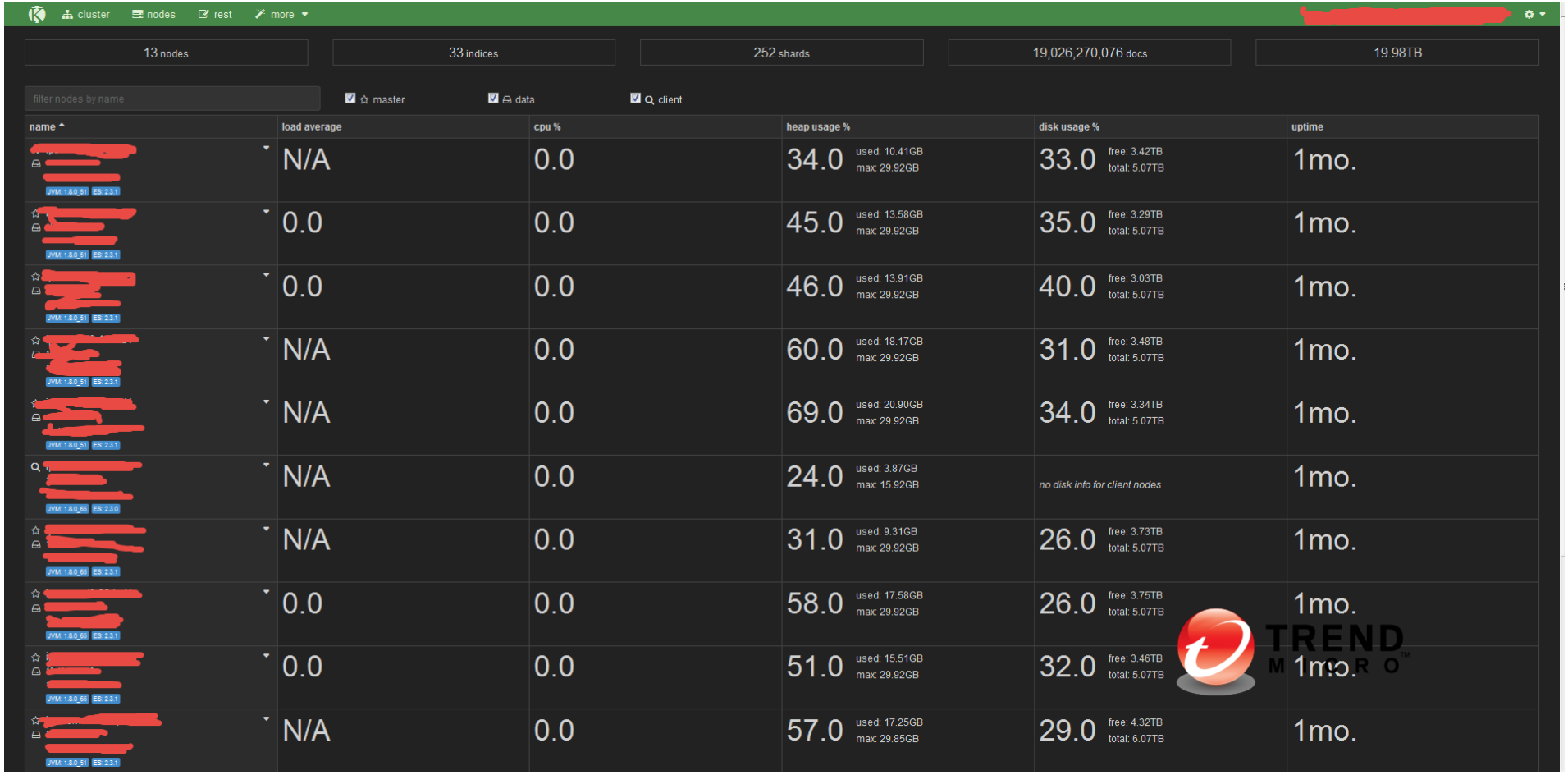
Solutions:

- Cache/Priority
- SQL-Like
- Cross

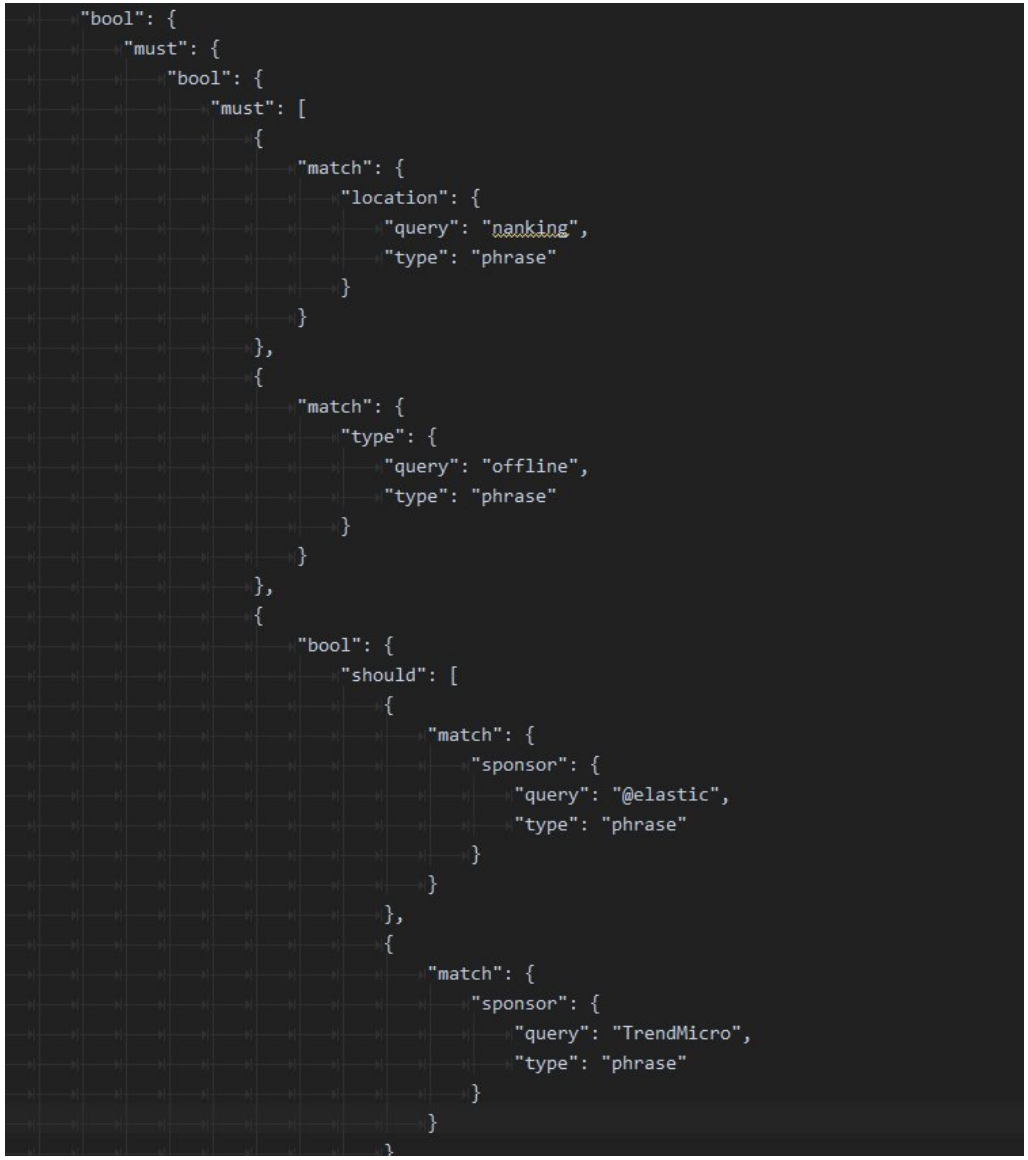
Data Levels

- **20+ nodes**
- **30TB+ data**
- **40 billion+ docs**

Data Levels


cluster
nodes
rest
more
13 nodes
33 indices
252 shards
19,026,270,076 docs
19.98TB
filter nodes by name
master
data
client
name
load average
cpu %
heap usage %
disk usage %
uptime
N/A 0.0 34.0 used: 10.41GB max: 29.92GB 33.0 free: 3.42TB total: 5.07TB 1mo.
0.0 0.0 45.0 used: 13.58GB max: 29.92GB 35.0 free: 3.29TB total: 5.07TB 1mo.
0.0 0.0 46.0 used: 13.91GB max: 29.92GB 40.0 free: 3.03TB total: 5.07TB 1mo.
N/A 0.0 60.0 used: 18.17GB max: 29.92GB 31.0 free: 3.48TB total: 5.07TB 1mo.
N/A 0.0 69.0 used: 20.90GB max: 29.92GB 34.0 free: 3.34TB total: 5.07TB 1mo.
N/A 0.0 24.0 used: 3.87GB max: 15.92GB no disk info for client nodes 1mo.
N/A 0.0 31.0 used: 9.31GB max: 29.92GB 26.0 free: 3.73TB total: 5.07TB 1mo.
0.0 0.0 58.0 used: 17.58GB max: 29.92GB 26.0 free: 3.75TB total: 5.07TB 1mo.
0.0 0.0 51.0 used: 15.51GB max: 29.92GB 32.0 free: 3.46TB total: 5.07TB 1mo.
N/A 0.0 57.0 used: 17.25GB max: 29.85GB 29.0 free: 4.32TB total: 6.07TB 1mo.

DSL is Boring



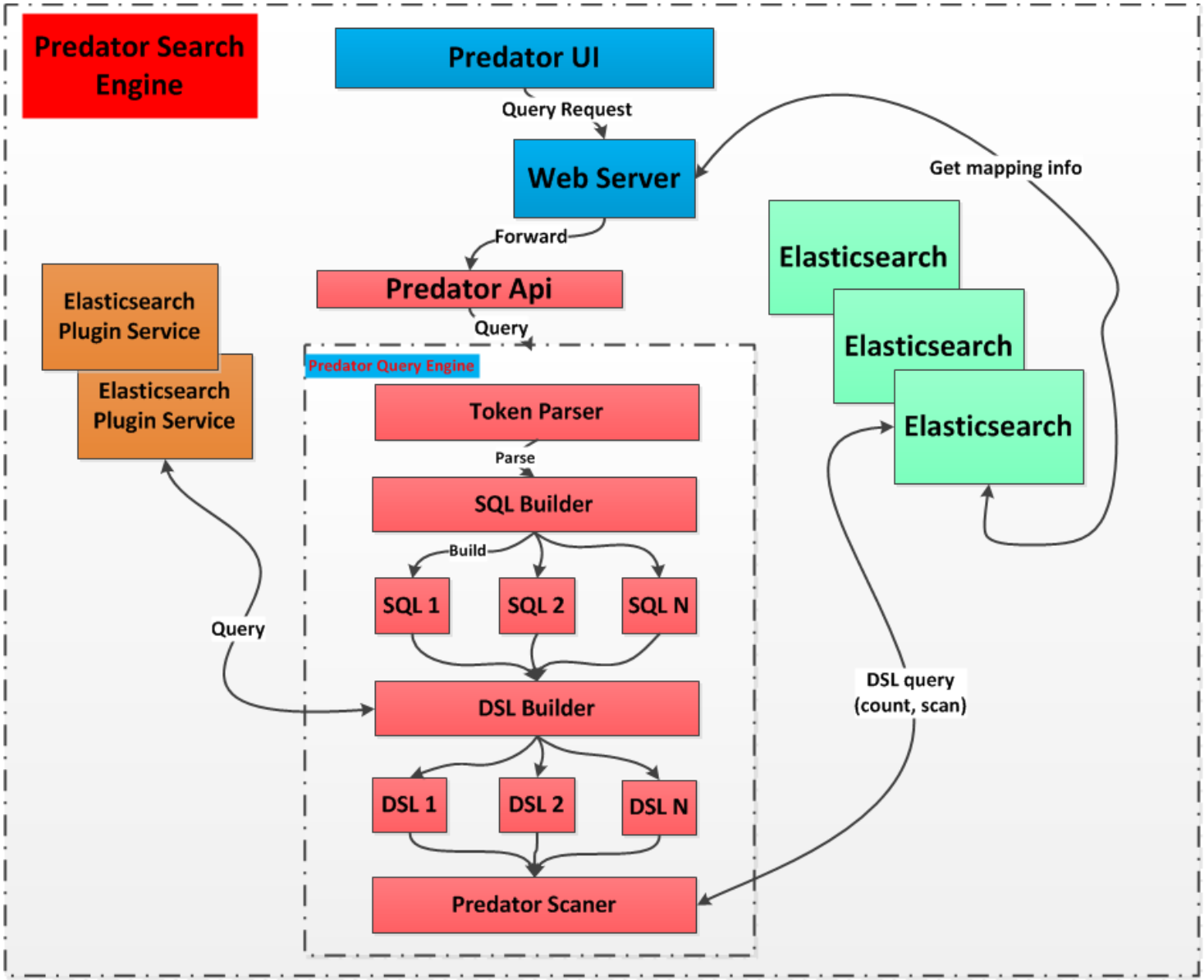
```
"bool": {
    "must": {
        "bool": {
            "must": [
                {
                    "match": {
                        "location": {
                            "query": "nanking",
                            "type": "phrase"
                        }
                    }
                },
                {
                    "match": {
                        "type": {
                            "query": "offline",
                            "type": "phrase"
                        }
                    }
                },
                {
                    "bool": {
                        "should": [
                            {
                                "match": {
                                    "sponsor": {
                                        "query": "@elastic",
                                        "type": "phrase"
                                    }
                                }
                            },
                            {
                                "match": {
                                    "sponsor": {
                                        "query": "TrendMicro",
                                        "type": "phrase"
                                    }
                                }
                            }
```

SQL statement:

SELECT elasticsearch
FROM meetup
WHERE
location="nanking"
AND type='offline'
AND sponsor
IN
('@elastic', 'TrendMicro')

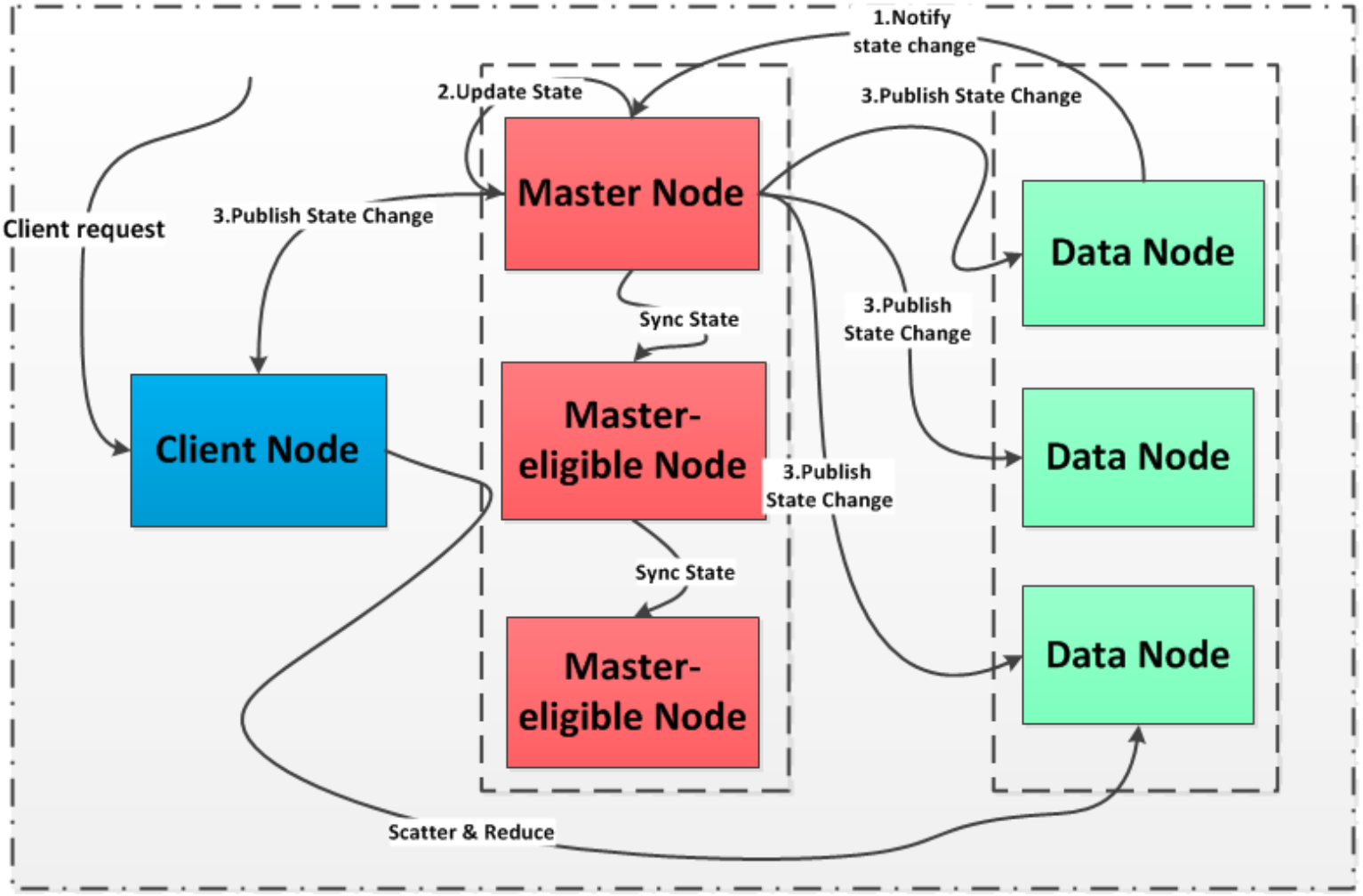
SQL-Like statement is more familiar by Developers, so we need SQLs to express DSLs.

SQL on ElasticSearch



TREND MICRO

Optimization--Topology



Roles:

Master

Data

Client

Control JVM Heap?

1.Segments Memory:

segment(Term Dictionary --> Posting list) + fast search(**term index**).

- ***How***:

delete or close unused indices. curator periodically optimize indices(only increase I/O).

2.Filter Cache:

filter query result cached in memory.

- ***How***:

"**indices.cache.filter.size**: 10%"(experience).

3.Field Data:

sort or aggregation will analyze inverted index and fill in the memory.
Pre-read
ES 2.0 will use **doc_value** for not analyzed field, but analyzed field still has.

- ***How***:

"**indices.fielddata.cache.size**: 10%"(experience).

4. Buffers:

bulk Queue/index buffer.

- ***How***:

Default is OK. Heap (**size** = queue nums * bulk size).

Admin Tools?

- **Head**

http://mobz.github.io/elasticsearch-head/

- **Kopf**

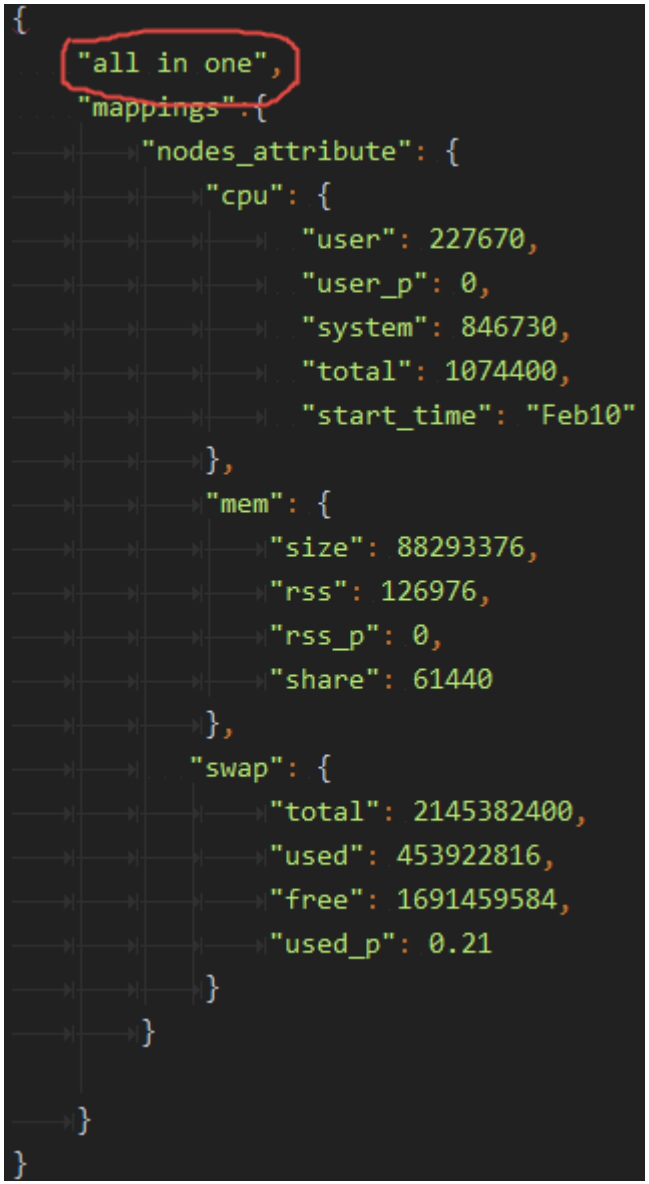
https://github.com/lmenezes/elasticsearch-kopf

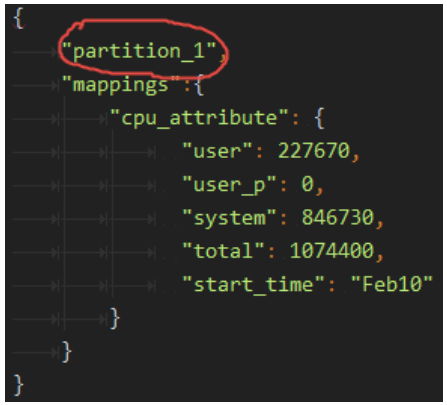
- **Curator**

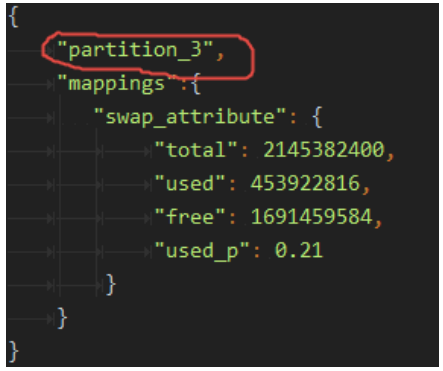
https://github.com/elastic/curator

alias	Index Aliasing
allocation	Index Allocation
bloom	Disable bloom filter cache
close	Close indices
delete	Delete indices or snapshots
open	Open indices
optimize	Optimize Indices
replicas	Replica Count Per-shard
seal	Seal indices (Synced flush: ES 1.6.0+ only)
show	Show indices or snapshots
snapshot	Take snapshots of indices (Backup)

Database Partition?







Problems:

Data sources

Frequent update

Performance

Risk

TREND MICRO

Pros and Cons

Let's talk about ***Pros and Cons*** When we need to join tables.

Pros :

- Code Logical clear.
- Speed up indexing rate.
- Avoid frequently update.
- High available, Scalable which lower down interference.

Cons:

- All in One Doc stay in same shard.
- Drop Posting lists join using **bitset** (filter in-memory).
- Drop Posting lists join using **skip-list** (random access disk).
- Need high performance cross cluster/type supporting tool.

NRT Ad-hoc?

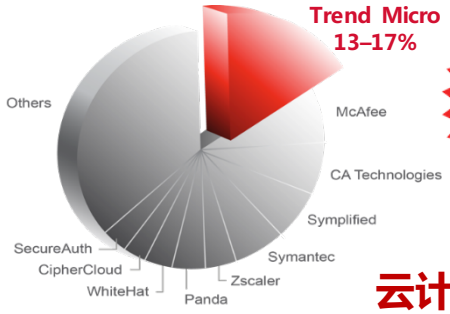
Q & A

We are Hiring

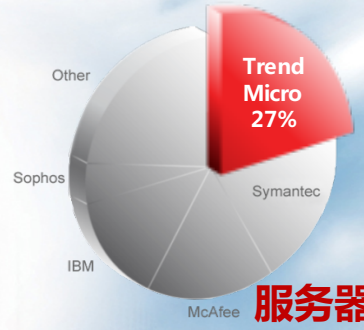
TREND MICRO

趋势科技 是全球最大的独立安全软件提供商

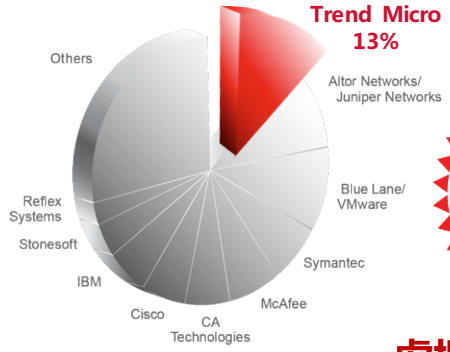
为全球 50 强企业中的 48 家提供安全防护



云计算安全第一名



服务器安全市场第一名

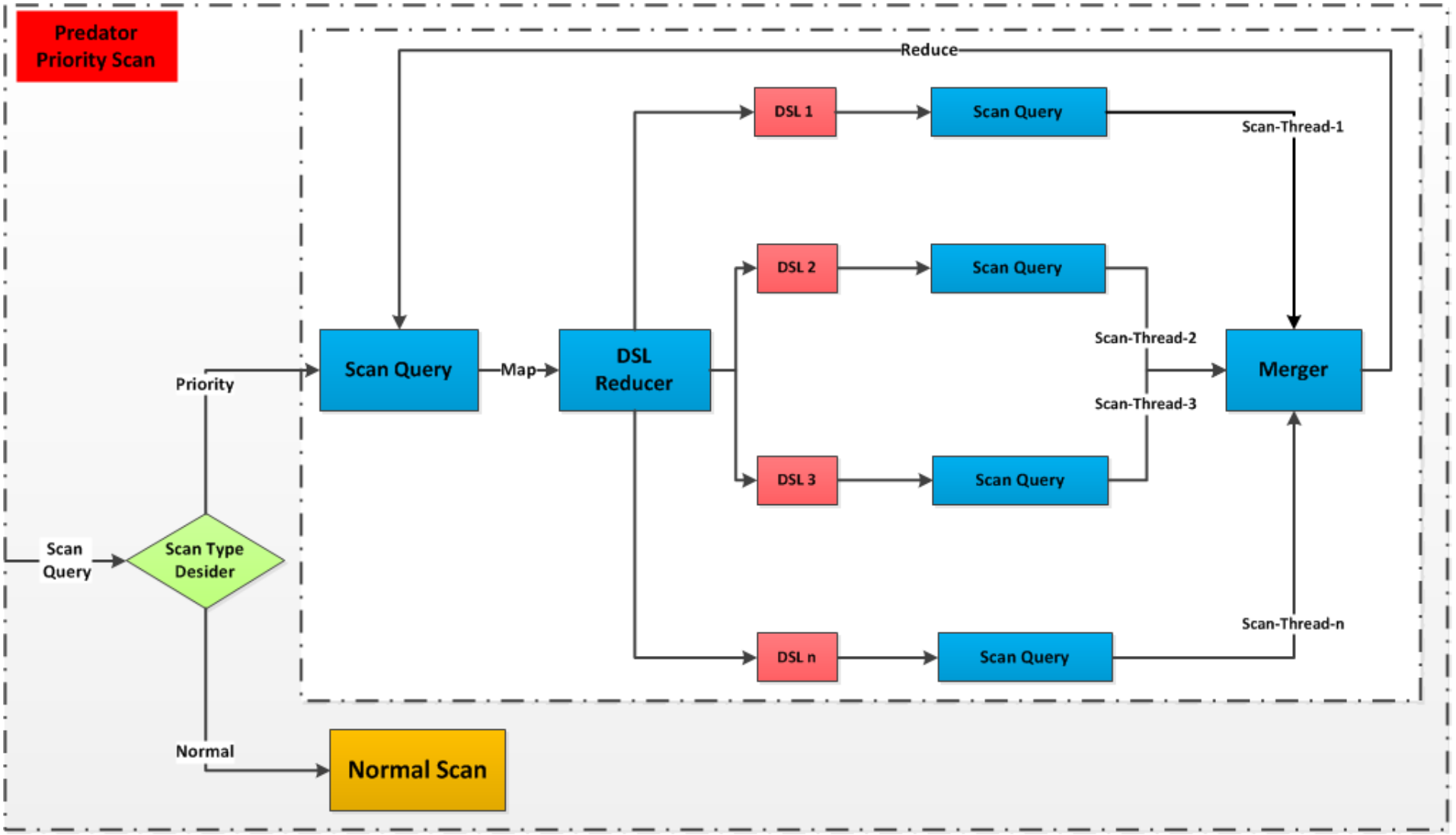


虚拟化安全第一名

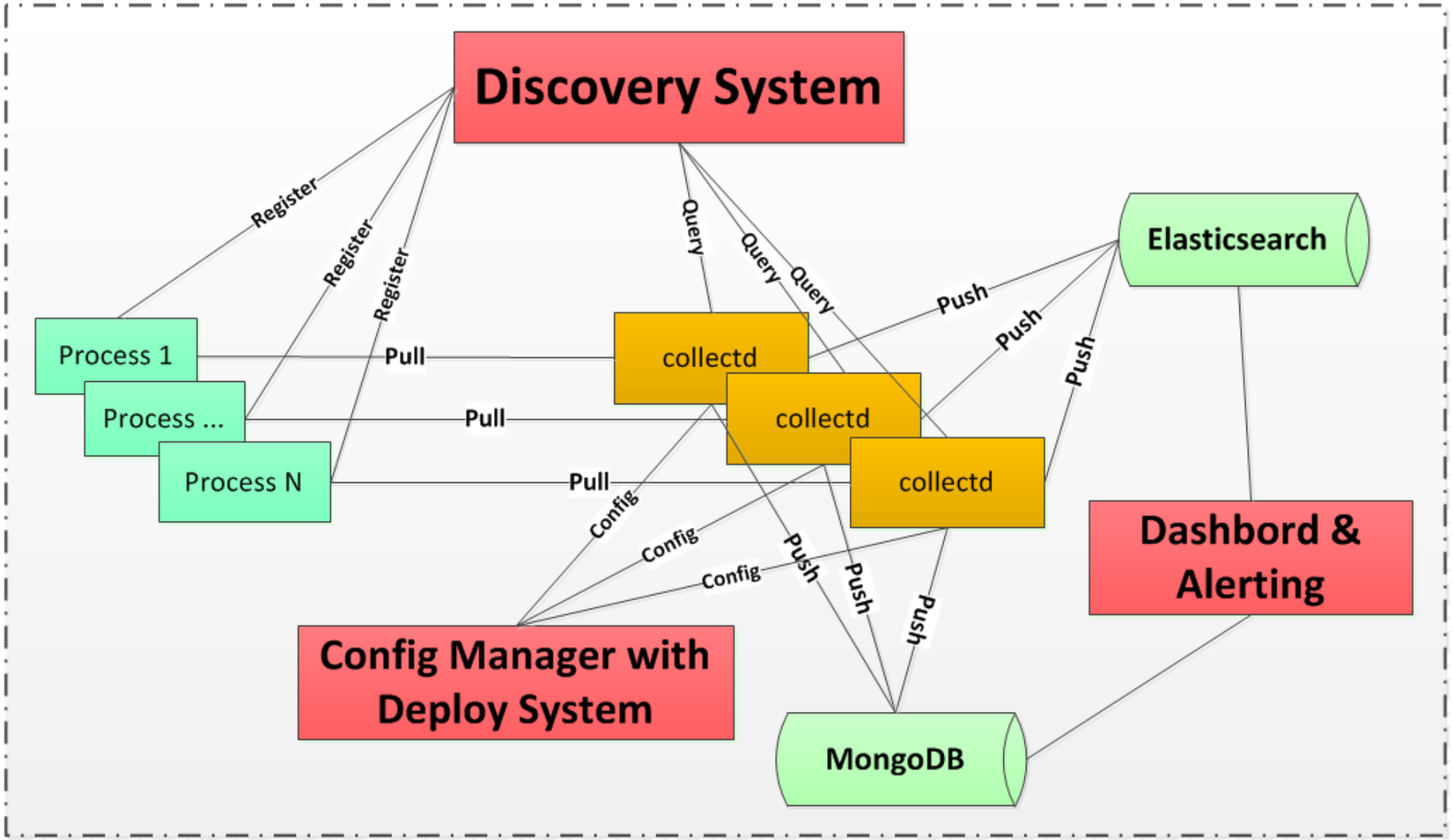
全球小企业安全市场第一名

Backup Slide

Backup Slide -- Priority Scan



Backup Slide -- Metrics



TREND MICRO