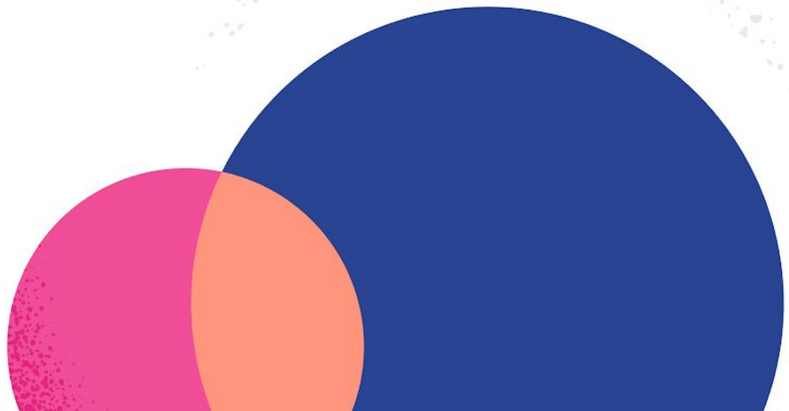
Ingesting Data at Scale into Elasticsearch with Apache Pulsar

Timothy Spann, Developer Advocate

11-Feb-2022



{Tim}

Timothy Spann | Developer Advocate

FLiP(N) Stack = Flink, Pulsar and NiFI Stack

Streaming Systems & Data Architecture Expert

Experience:

15+ years of experience with streaming technologies including Pulsar, Flink, Spark, NiFi, Kafka, Big Data, Cloud, MXNet, IoT and more.

Today, he helps to grow the Pulsar community sharing rich technical knowledge and experience at both global conferences and through individual conversations.

CLOUDERA HORTONWORKS Pivotal Hewlett Packard Enterprise pwc

Stream Native

elastic

Stream Native

- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar’s unified messaging and streaming platform.
- StreamNative Cloud with Flink SQL

Agenda

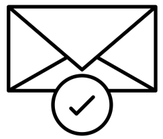
1. Pulsar as a Stream Buffer
2. Data Ingestion
    - Logs, Sensors & Events
3. Let’s Get to Sinking
4. End to End Architecture

Pulsar as a Stream Buffer for Elasticsearch

elastic

Why Apache Pulsar?

Unified Messaging Platform



Guaranteed Message Delivery

Resiliency

Infinite Scalability

Perfect for Buffering

Unified Messaging Model

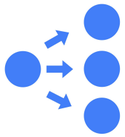
Simplify your data infrastructure and enable new use cases with queuing and streaming capabilities in one platform.

Multi-tenancy

Enable multiple user groups to share the same cluster, either via access control, or in entirely different namespaces.

Scalability

Decoupled data computing and storage enable horizontal scaling to handle data scale and management complexity.



Geo-replication

Support for multi-datacenter replication with both asynchronous and synchronous replication for built-in disaster recovery.

Tiered storage

Enable historical data to be offloaded to cloud-native storage and store event streams for indefinite periods of time.

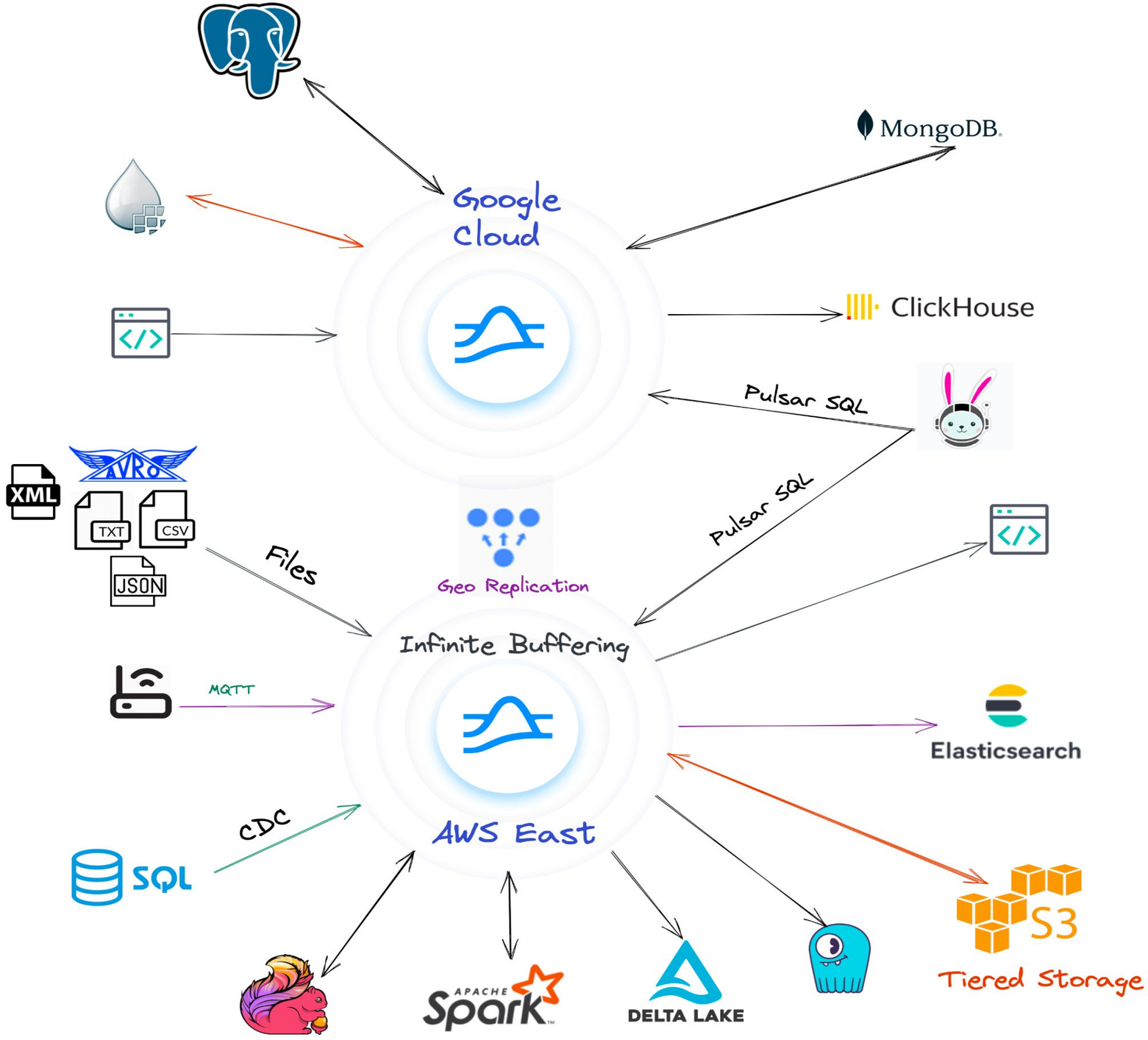
Buffering?

- Time Intervals (Minute, 5 Minutes, 15 Minutes, ...)
- Buffer Batch Size (1MB, 5MB, 100MB, 1GB, ...)
- Batches of Records (1000, 10000, 10000, ...)
- Aggregate or Summarize Data
- Geo-Replication Aggregation Pattern
- Deduplicate data

Stream Buffer It All

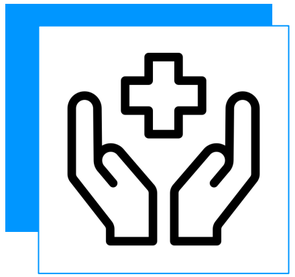
- High throughput
- Massive scalability
- Buffer between many different data producers
- Reduce producer load on Elasticsearch
- Distribute to many downstream systems

- Buffer
- Batch
- Route
- Filter
- Aggregate
- Enrich
- Replicate
- Dedupe
- Decouple
- Distribute



Data Ingestion
(Logs, Sensors & Events)

elastic

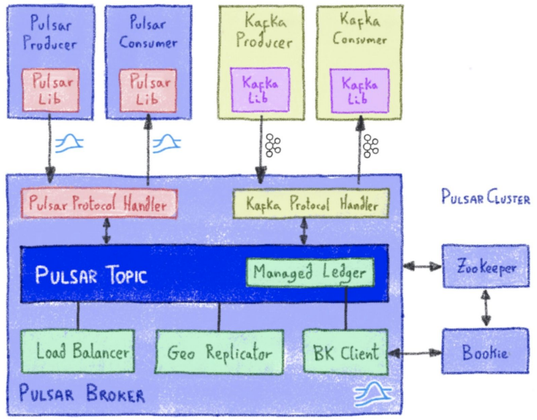


Logs, Sensors & Events

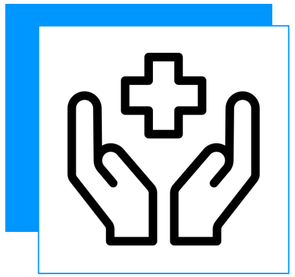
- Netty
- Files
- Apache NiFi Sources
- Sensors
- Canal & Debezium CDC Events
- Kafka, ActiveMQ, RabbitMQ, AMQP, Kinesis, SQS, GCP Pub/Sub

Connectivity

hub.streamnative.io



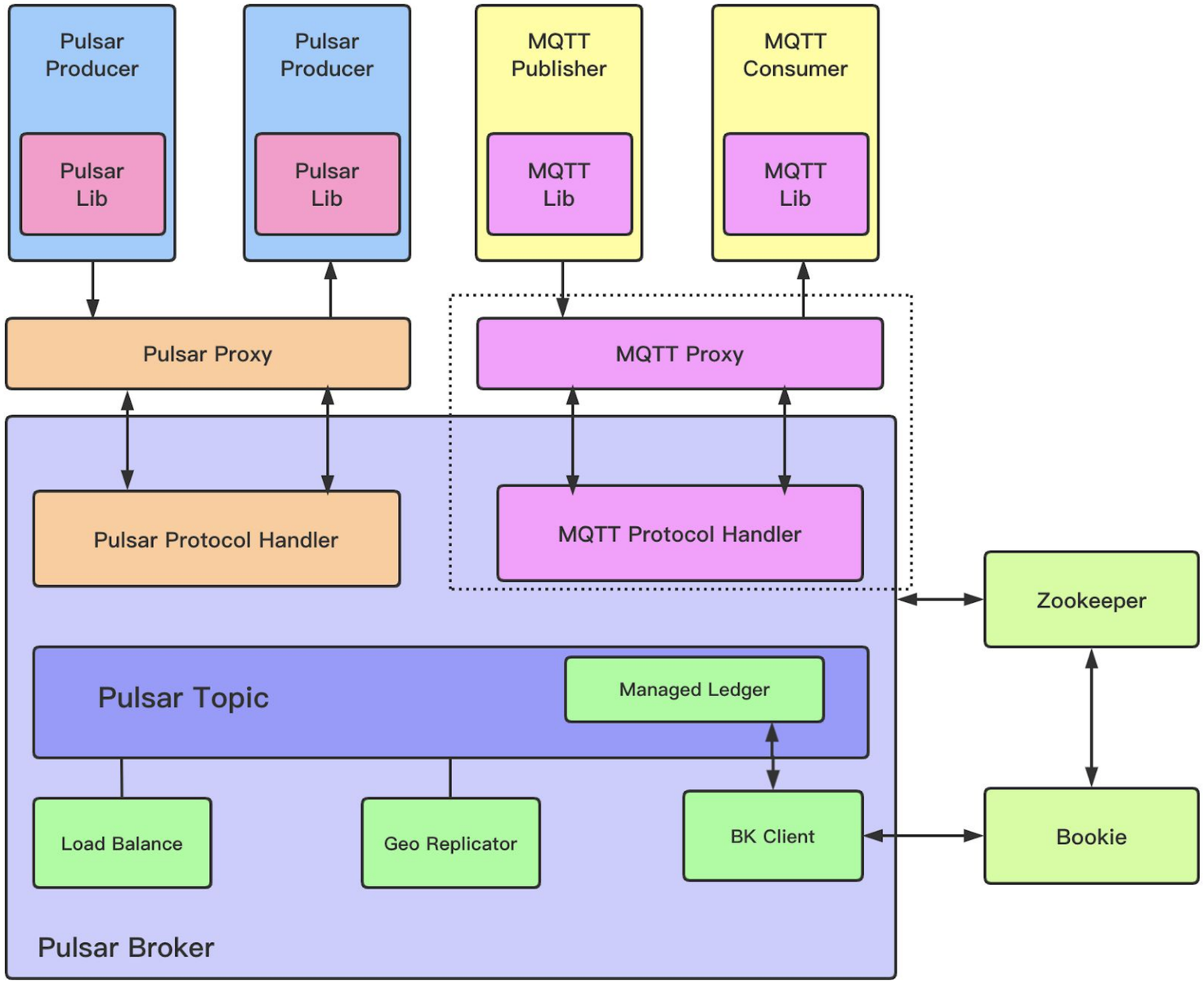
- **Functions** - Lightweight Stream Processing (Java, Python, Go)
- **Connectors** - Sources & Sinks (InfluxDB, Kafka, S3, Kinesis, Lambda, …)
- **Protocol Handlers** - AoP (AMQP), KoP (Kafka), MoP (MQTT)
- **Processing Engines** - Flink, Spark, Presto/Trino via Pulsar SQL
- **Data Offloaders** - Tiered Storage - (S3)



MQTT

On Pulsar

(MoP)



Let’s Get To Sinking

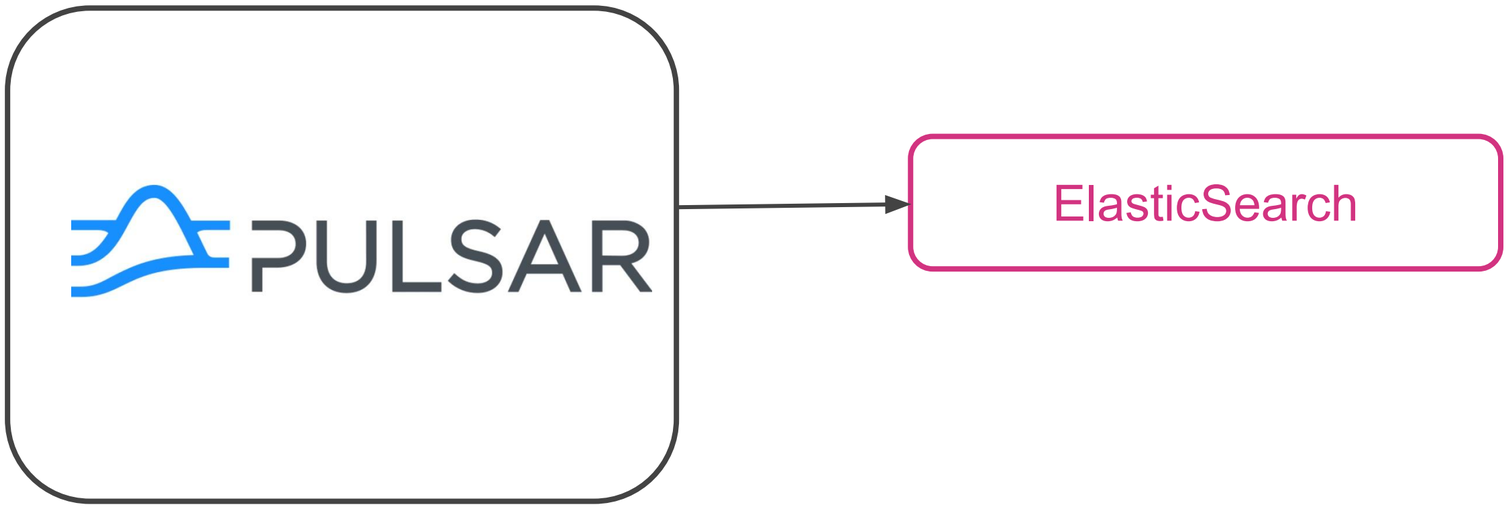
elastic

Moving Data In and Out of Pulsar

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar. https://pulsar.apache.org/docs/en/io-elasticsearch-sink/

- Built on top of Pulsar Functions
- Built-in connectors - hub.streamnative.io

ElasticSearch **Sink** Connector



- **Now with Bulk Index Support**
- **Now with Schema Support**

https://pulsar.apache.org/docs/en/io-quickstart/

End to End Architecture

elastic

Streaming Elastic FLiP Apps - Roll the Demo!!!

PULSAR

elastic

Edge

Gateway

PULSAR

Software script

Protocols

Pulsar

KoP

MoP

Websocket

Apps

CDC

Unified Batch and Stream **STORAGE**

PULSAR
(Queuing + Streaming)

Streaming

Unified Batch and Stream **COMPUTING**

Flink
(Batch + Stream)

Offload

Batch

S3

10 01

Tiered Storage

hadoop HDFS

StreamNative Cloud

APACHE Spark

Pulsar Sink

StreamNative Hub

elastic

visualize

APACHE nifi

APACHE Spark

https://github.com/tspannhw/FLiP-Elastic

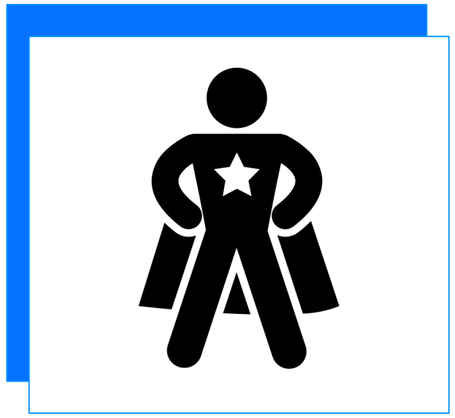
Pulsar 411

elastic

Use Cases

- Unified Messaging Platform
- AdTech
- Fraud Detection
- Universal Data Buffer
- IoT Analytics

elastic

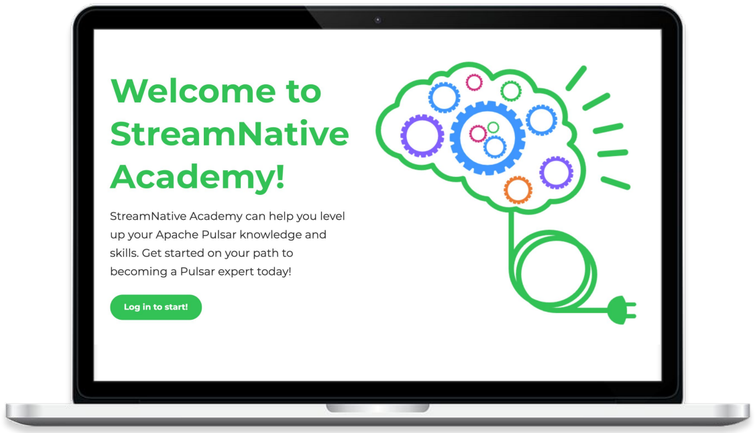


StreamNative Ambassador Program 2022

Learn More

Tell us about your Pulsar experience and what improvements you would like to see!

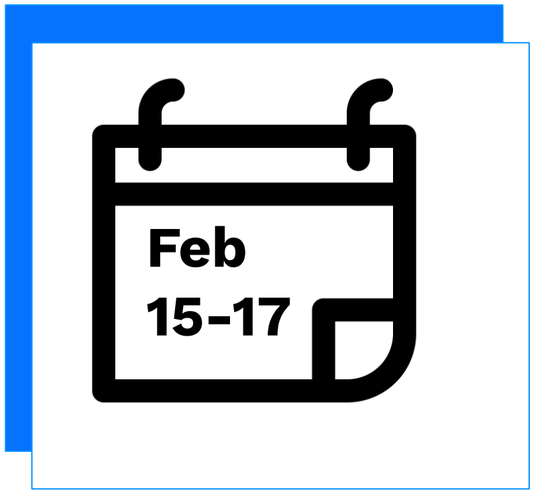
Start Survey



Now Available

On-Demand Pulsar Training

Academy.StreamNative.io



Live 3-day

Developers Training

Save Your Spot!

Times:

- Europe: 3:00 PM CET - 7:00 PM CET
- EasternTime: 9:00 AM - 1: 00 PM EST
- Pacific Time: 6:00 AM - 10 AM PST

FLiP Stack Weekly

This week in Apache Flink, Apache Pulsar, Apache NiFi, Apache Spark, Elasticsearch and open source friends.

https://bit.ly/32dAJft

StreamNative Cloud

Powered by Apache Pulsar, StreamNative provides a cloud-native, real-time messaging and streaming platform to support multi-cloud and hybrid cloud strategies.

Cloud Native

Built for Containers

Flink SQL

Let’s Keep in Touch!

Tim Spann
Developer Advocate

@PaaSDev

https://www.linkedin.com/in/timothyspann

https://github.com/tspannhw

More Elastic Configurations

elastic

Name	Type	Required	Default	Description
elasticSearchUrl	String	true	" " (empty string)	The URL of elastic search cluster to which the connector connects.
indexName	String	true	" " (empty string)	The index name to which the connector writes messages.
schemaEnable	Boolean	false	false	Turn on the Schema Aware mode.
createIndexIfNeeded	Boolean	false	false	Manage index if missing.
maxRetries	Integer	false	1	The maximum number of retries for elasticsearch requests. Use -1 to disable it.
retryBackoffInMs	Integer	false	100	The base time to wait when retrying an Elasticsearch request (in milliseconds).
maxRetryTimeInSec	Integer	false	86400	The maximum retry time interval in seconds for retrying an elasticsearch request.
bulkEnabled	Boolean	false	false	Enable the elasticsearch bulk processor to flush write requests based on the number or size of requests, or after a given period.
bulkActions	Integer	false	1000	The maximum number of actions per elasticsearch bulk request. Use -1 to disable it.
bulkSizeInMb	Integer	false	5	The maximum size in megabytes of elasticsearch bulk requests. Use -1 to disable it.
bulkConcurrentRequests	Integer	false	0	The maximum number of in flight elasticsearch bulk requests. The default 0 allows the execution of a single request. A value of 1 means 1 concurrent request is allowed to be executed while accumulating new bulk requests.
bulkFlushIntervalInMs	Integer	false	-1	The maximum period of time to wait for flushing pending writes when bulk writes are enabled. Default is -1 meaning not set.
compressionEnabled	Boolean	false	false	Enable elasticsearch request compression.
connectTimeoutInMs	Integer	false	5000	The elasticsearch client connection timeout in milliseconds.
connectionRequestTimeoutInMs	Integer	false	1000	The time in milliseconds for getting a connection from the elasticsearch connection pool.

Name	Type	Required	Default	Description
`connectionIdleTimeoutInMs`	Integer	false	5	Idle connection timeout to prevent a read timeout.
`keyIgnore`	Boolean	false	true	Whether to ignore the record key to build the Elasticsearch document `_id`. If primaryFields is defined, the connector extract the primary fields from the payload to build the document `_id` If no primaryFields are provided, elasticsearch auto generates a random document `_id`.
`primaryFields`	String	false	"id"	The comma separated ordered list of field names used to build the Elasticsearch document `_id` from the record value. If this list is a singleton, the field is converted as a string. If this list has 2 or more fields, the generated `_id` is a string representation of a JSON array of the field values.
`nullValueAction`	enum (IGNORE,DELETE,FAIL)	false	IGNORE	How to handle records with null values, possible options are IGNORE, DELETE or FAIL. Default is IGNORE the message.
`malformedDocAction`	enum (IGNORE,WARN,FAIL)	false	FAIL	How to handle elasticsearch rejected documents due to some malformation. Possible options are IGNORE, DELETE or FAIL. Default is FAIL the Elasticsearch document.
`stripNulls`	Boolean	false	true	If stripNulls is false, elasticsearch _source includes 'null' for empty fields (for example {"foo": null}), otherwise null fields are stripped.
`socketTimeoutInMs`	Integer	false	60000	The socket timeout in milliseconds waiting to read the elasticsearch response.
`typeName`	String	false	"_doc"	The type name to which the connector writes messages to. The value should be set explicitly to a valid type name other than "_doc" for Elasticsearch version before 6.2, and left to default otherwise.
`indexNumberOfShards`	int	false	1	The number of shards of the index.
`indexNumberOfReplicas`	int	false	1	The number of replicas of the index.
`username`	String	false	" " (empty string)	The username used by the connector to connect to the elastic search cluster. If `username` is set, then `password` should also be provided.
`password`	String	false	" " (empty string)	The password used by the connector to connect to the elastic search cluster. If `username` is set, then `password` should also be provided.
`ssl`	ElasticSearchSslConfig	false		Configuration for TLS encrypted communication

elastic

Search Elastic

D Dashboard > Example Dashboard

Full screen | Share | Clone | Edit

Count

104

Count of records

CPU

8.437

Average of cpu

Memory

33.606

Average of memory

Top % of ML

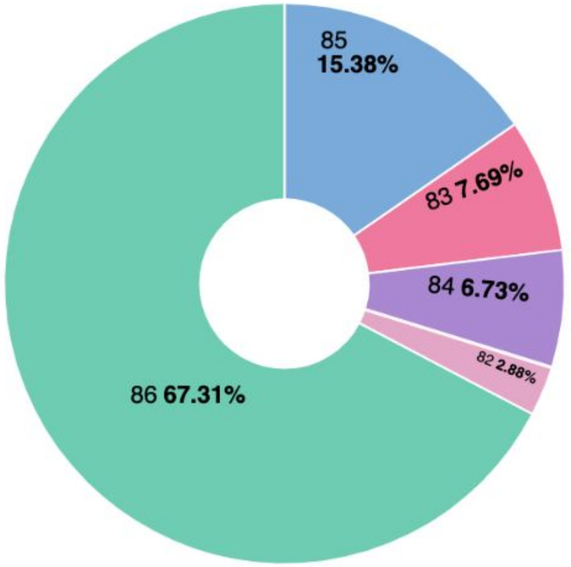
18.842

Median of top1pct

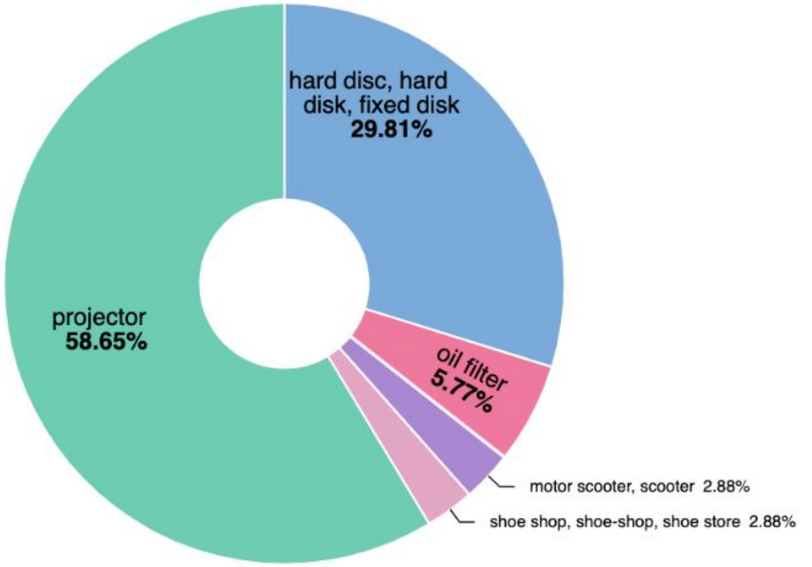
Top Identified Items

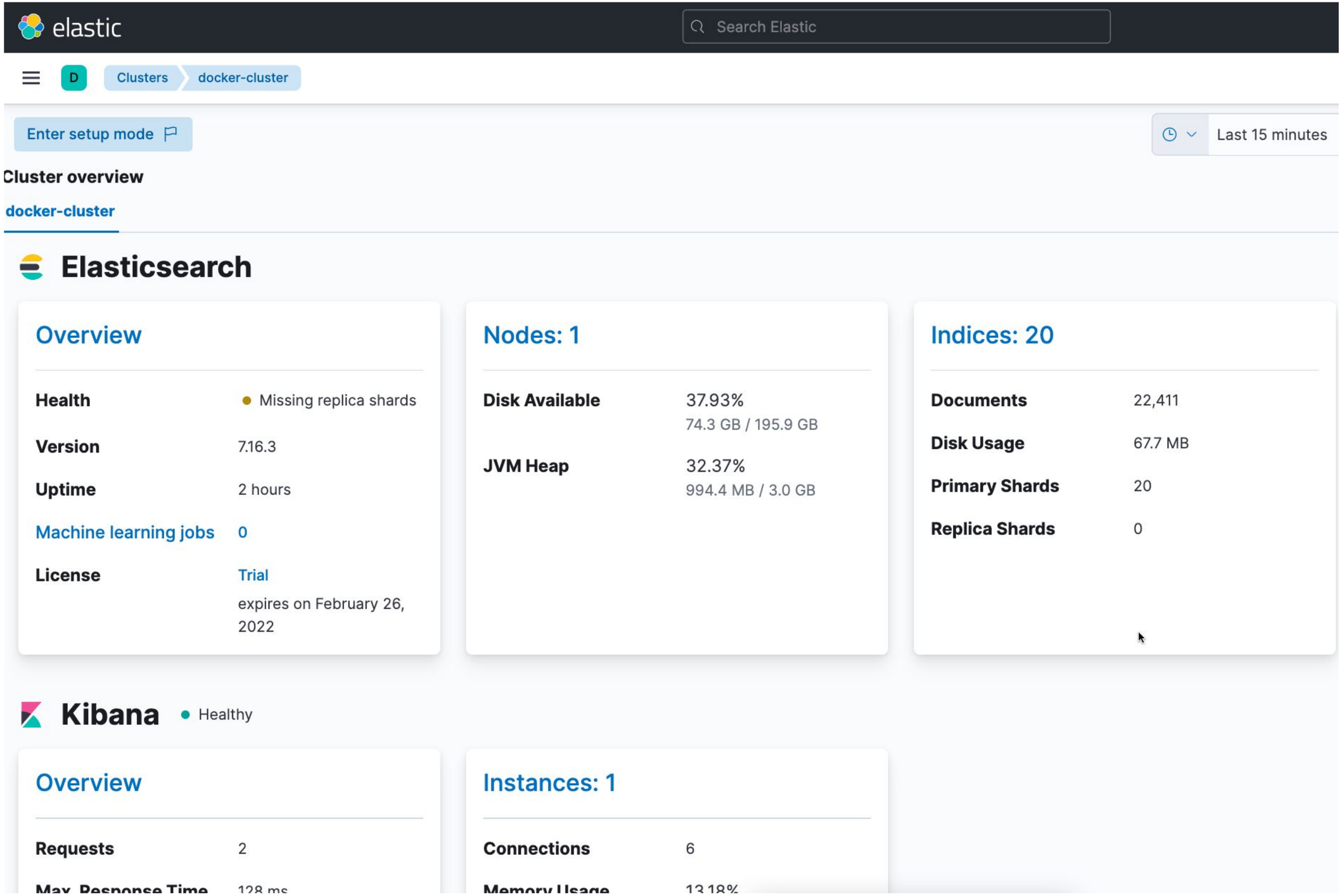
Top values of top1.keyword	Count of records
projector	61
hard disc, hard disk, fixed disk	31
oil filter	6
motor scooter, scooter	3
shoe shop, shoe-shop, shoe store	3

GPU Temp F



ML Labels




elastic
Search Elastic
D
Clusters
docker-cluster
Enter setup mode
Last 15 minutes
Cluster overview
docker-cluster
Elasticsearch
Overview
Health
Missing replica shards
Version
7.16.3
Uptime
2 hours
Machine learning jobs
0
License
Trial
expires on February 26, 2022
Nodes: 1
Disk Available
37.93%
74.3 GB / 195.9 GB
JVM Heap
32.37%
994.4 MB / 3.0 GB
Indices: 20
Documents
22,411
Disk Usage
67.7 MB
Primary Shards
20
Replica Shards
0
Kibana
Healthy
Overview
Requests
2
Instances: 1
Connections
6