HEC Yeah!! How Priceline Uses HEC to Ingest 4TB of Data Every Day?

Jagadeesh Motamarri | Senior Software Engineer
Mukund N Murthy | Software Engineer

priceline

October 2018

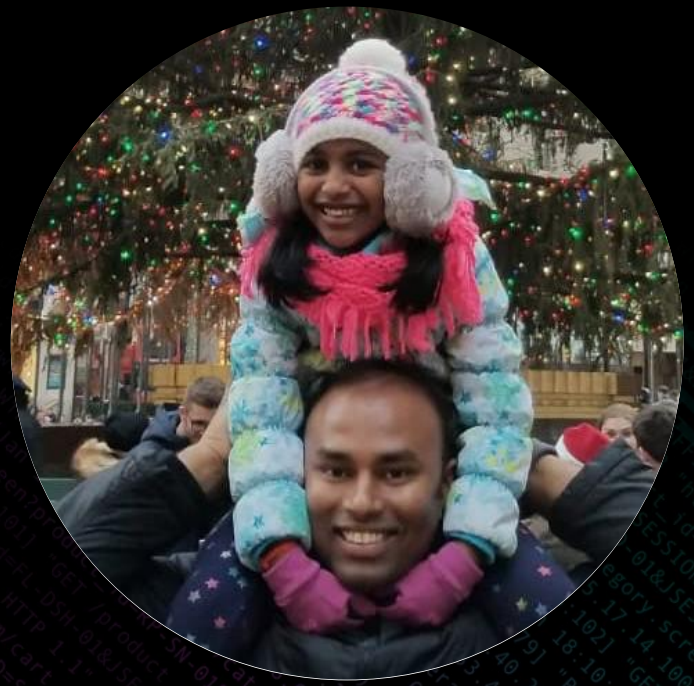
Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

splunk> .conf18

About us



JAGADEESH MOTAMARRI

Sr. Software Engineer, priceline

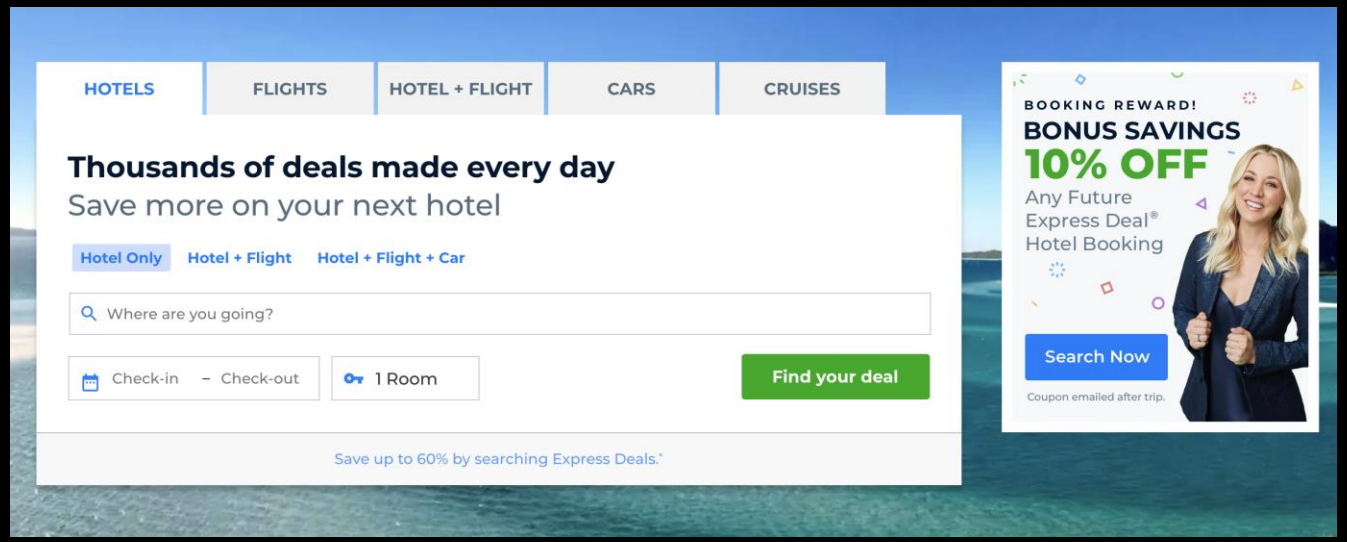
MUKUND N MURTHY

Software Engineer, priceline

About Priceline

priceline is part of Booking Holdings, the world leader in online travel & related services.

priceline offers more ways to save and more deals than anyone else in travel.



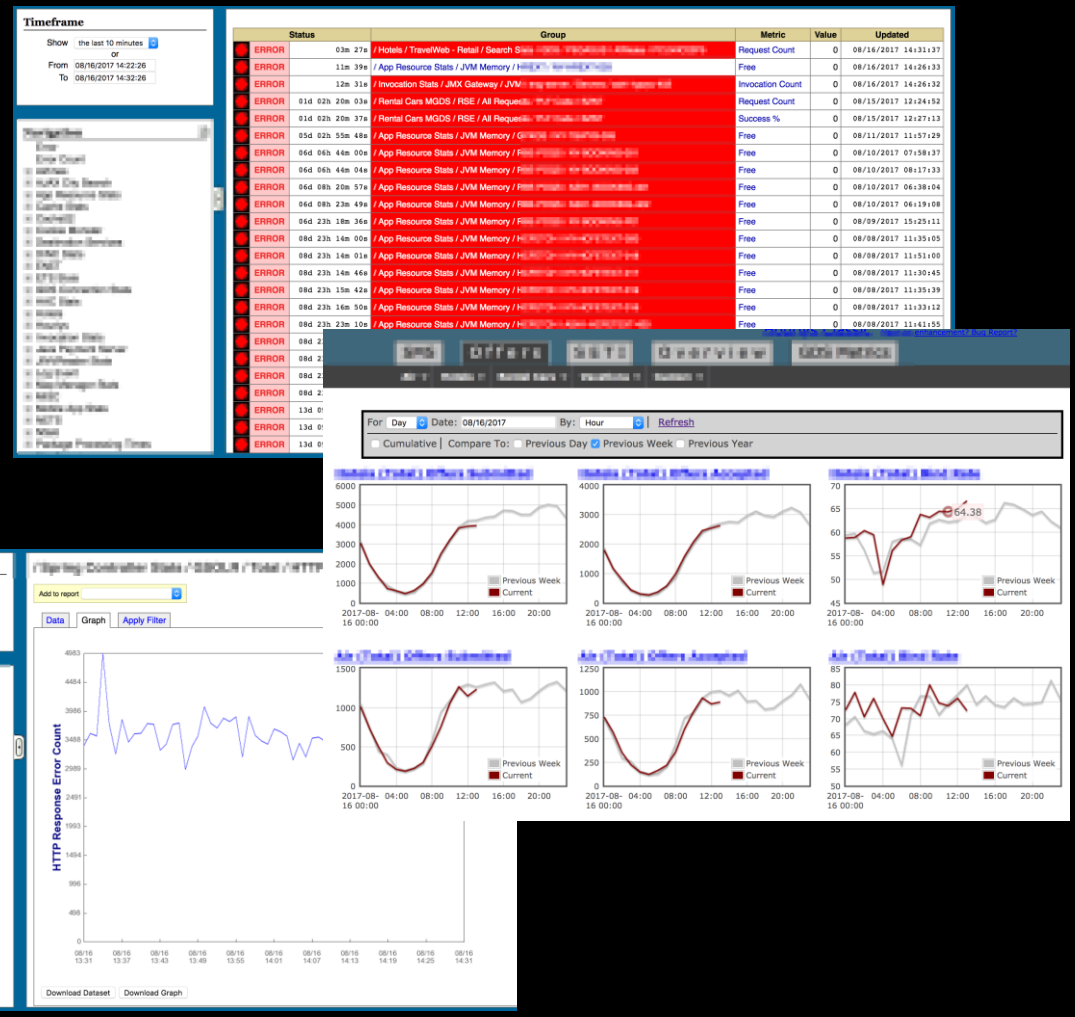
priceline Booking.com KAYAK agoda Rentalcars.com OpenTable®

How we got here…

The problem we set out to solve

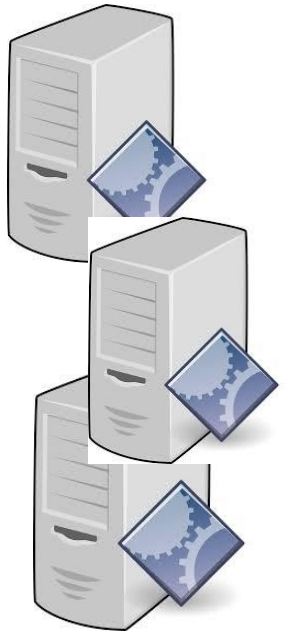
- We had a collection of bespoke monitoring systems, evolved over 18+ years; **weren't investing** enough to get the full value from them.

- Separate systems meant that we sometimes had **difficulty seeing data** across applications or application layers in the same context



Legacy Architecture

Custom Dashboards



Application Servers

Custom Alerting System

Kibana Dashboards

priceline

splunk> .conf18

So, What the HEC?

Narrowing the field

- Splunk is a fully features, powerful platform for collecting, searching, monitoring and analyzing machine data.
  - **Rich ecosystem of 3rd party apps**

    Crowdstrike, F5 BIG-IP, Cisco, *nix, Windows, Palo Alto, Catchpoint, AWS, GCP
  - **Flexible data ingestion architecture**

    HTTP/REST (**HEC**), log scraping, dedicated apps

Stats

Beast mode ON

125+

# of Applications

~4.5TB

Daily Ingestion

200+

Unique Sourcetypes

250K+

Queries per day

HEC Yeah!!

The driving force

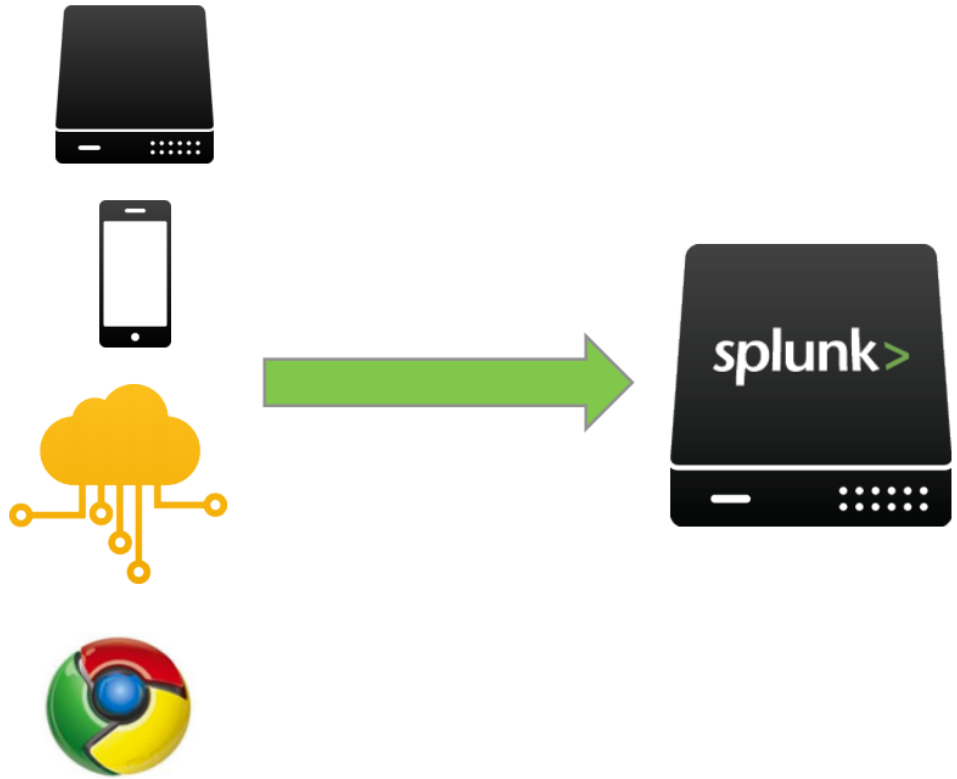
priceline

splunk> .conf18

HEC Yeah!!

What is HEC?

- Token based API for collection of events
- Send events directly from anywhere
- Easy to Configure and works out of the box
- Easy to Secure
- Highly Performant ,scalable and available

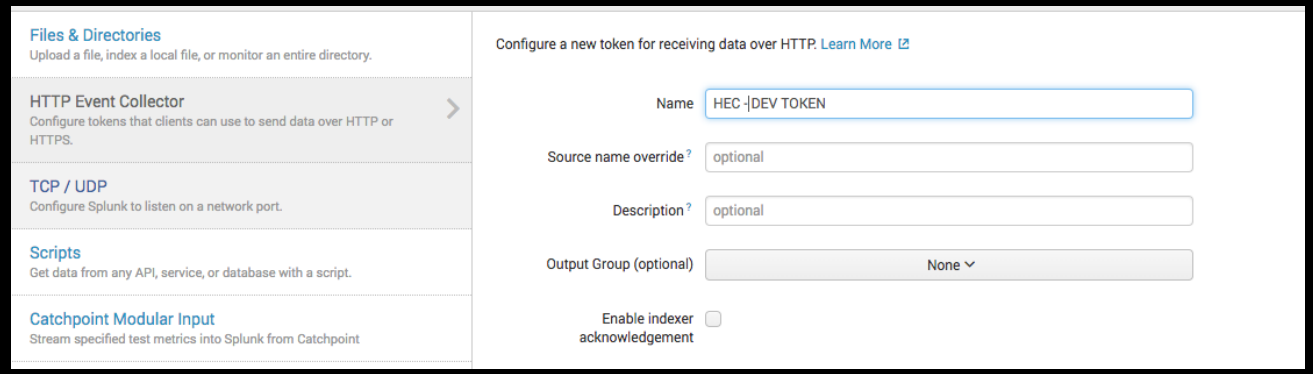


Source: https://www.slideshare.net/Splunk/splunk-http-event-collector/2

HEC Yeah

Set-up and Usage

- Enable HTTP Event Collector
- Create/Get a token
- Send events to Splunk using the token
  - Use HTTP Directly
  - Use logging libraries



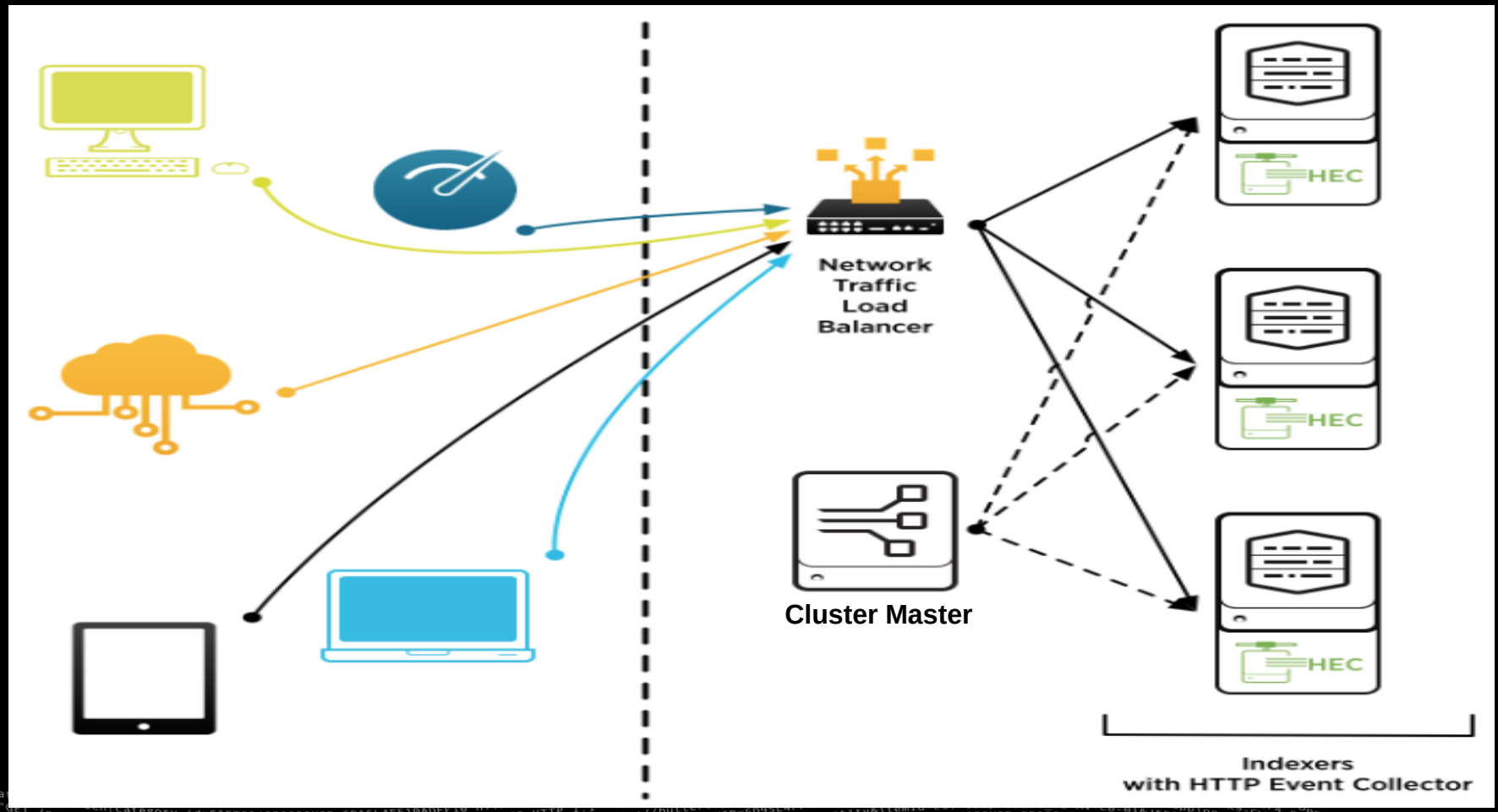
Sending Data

```
curl -v http://localhost:8088/services/collector
-H "Authorization: Splunk 9F7F64FC-8E3F-4D85-
B7F3-F6EC5B71ED1B" -d
'{"event":{"uid":"hrottenberg","action","login"}}
'
```

HEC Yeah

Distributed Deployment

- Adopted Pattern # **Traffic load balancer, no Heavy Forwarder, pool of indexers, using cluster master**



Priceline Data Collection Platform

Core Modules

Data Collection

Data Processing / Ingestion

Dynamic Configuration

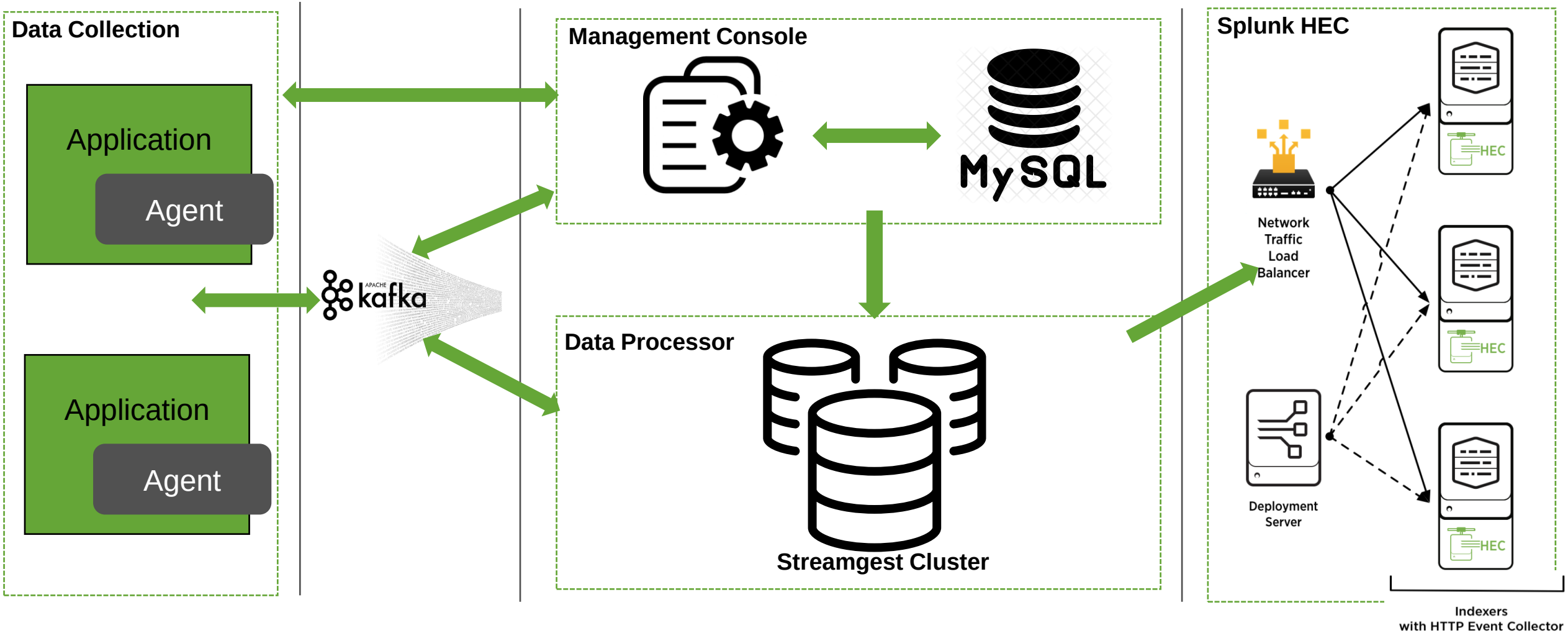
Management Console

priceline

splunk> .conf18

Priceline Data Collection Platform

High Level Architecture



priceline

splunk> .conf18

Management Console

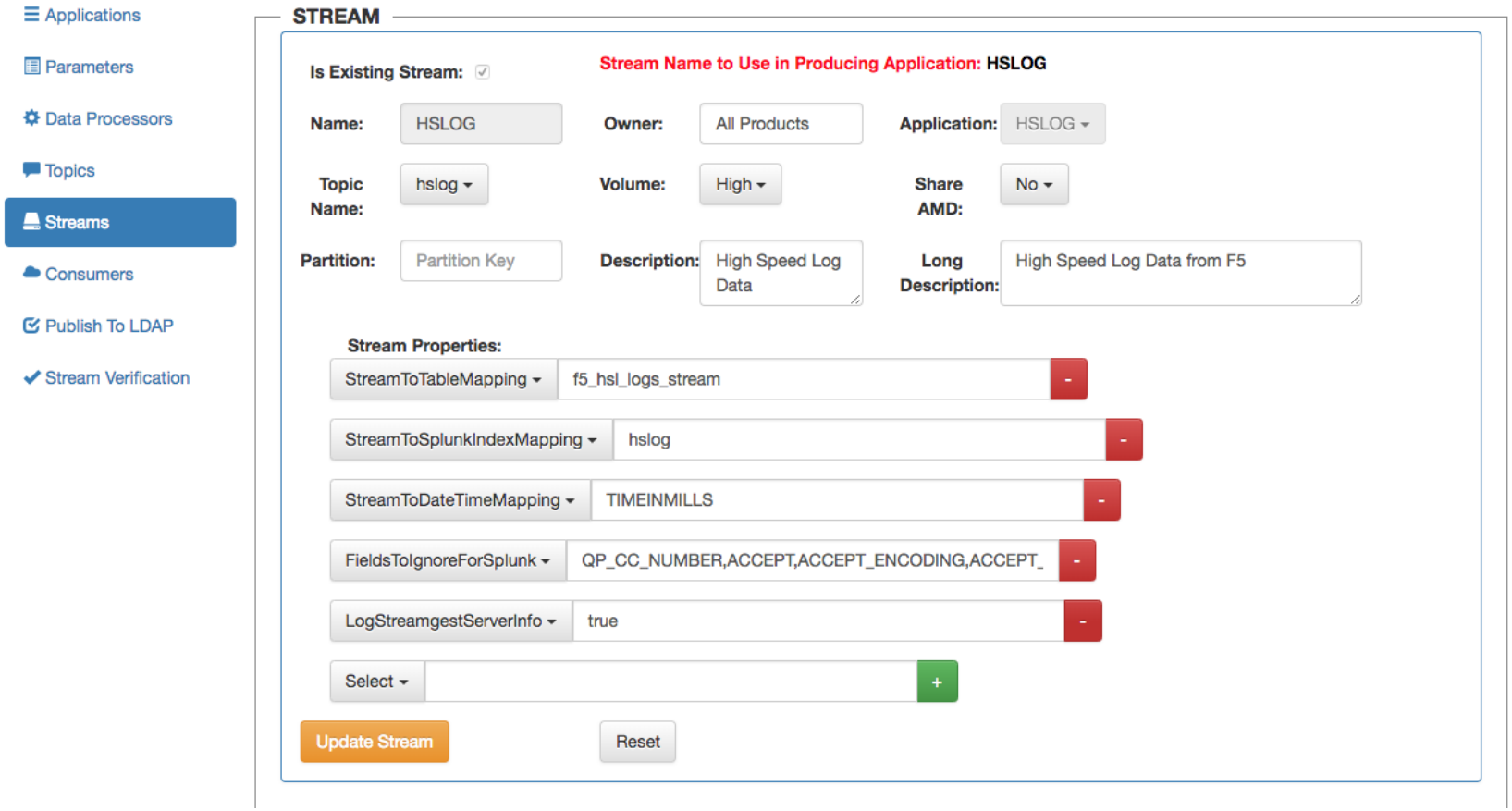
Config System for managing meta-data

priceline

splunk> .conf18

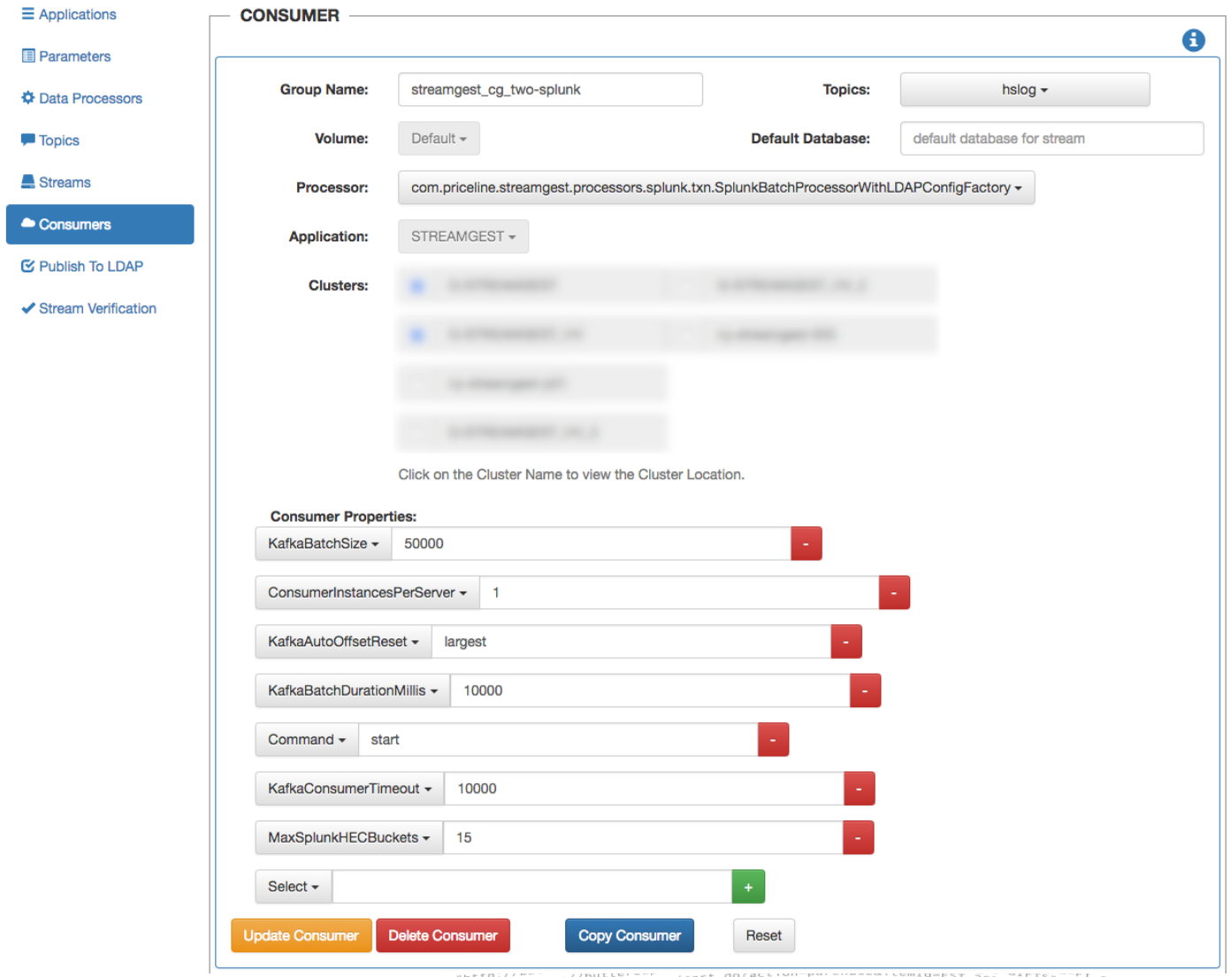
Management Console

Priceline's home grown application



Management Console

Priceline's home grown application



priceline

splunk> .conf18

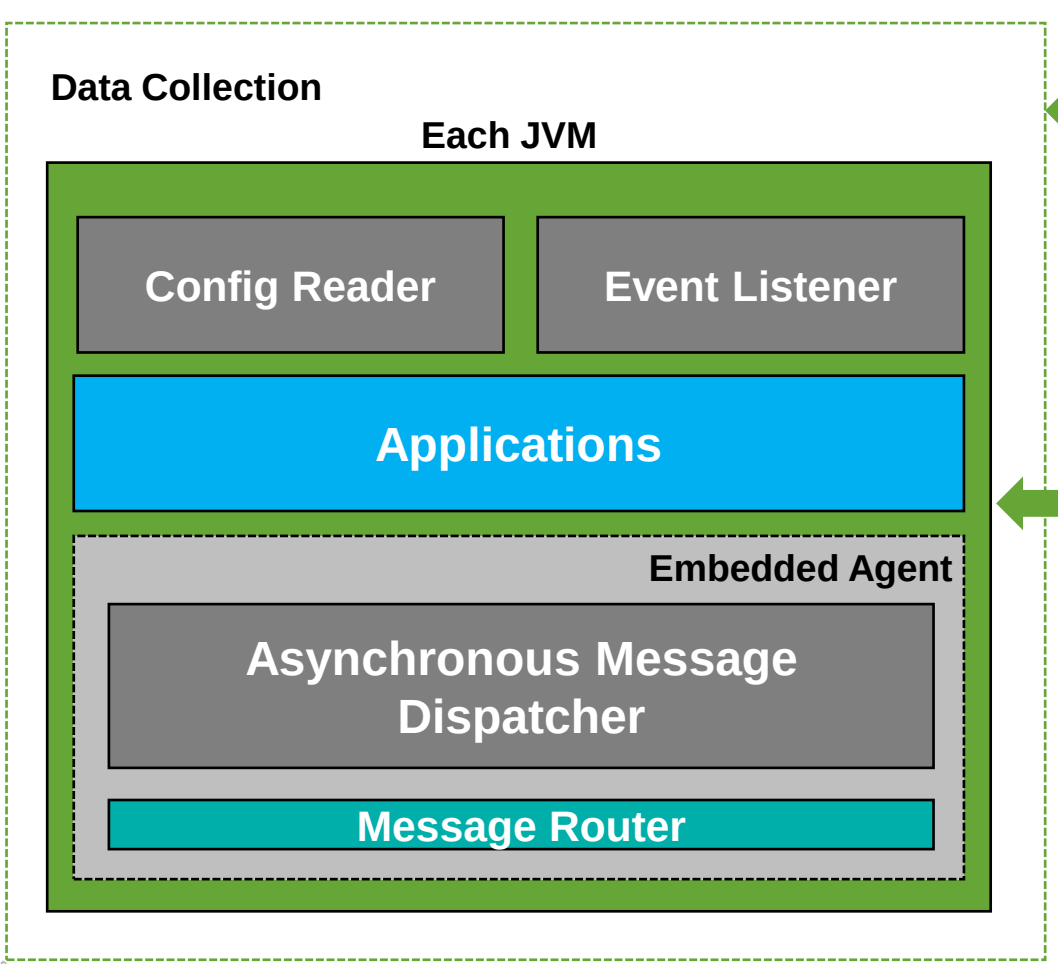
Data Collection

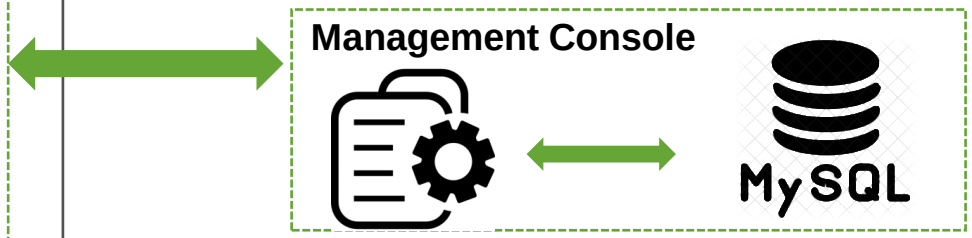
Embedded JVM Agent and REST

priceline

splunk> .conf18

Data Collection – Embedded JVM Agent





- Dispatches asynchronously
- Non blocking data collection
- Data dependent routing to different kafka partitions
- Handles back pressure

Data Processor

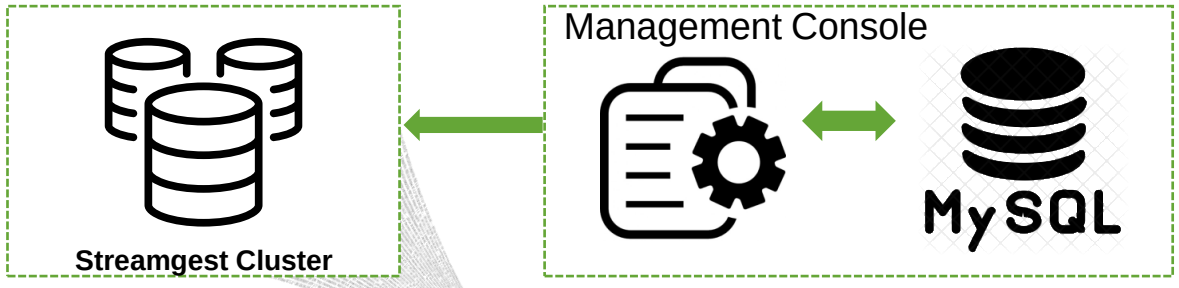
Streaming + Ingestion = Streamgest

priceline

splunk> .conf18

Data Processor

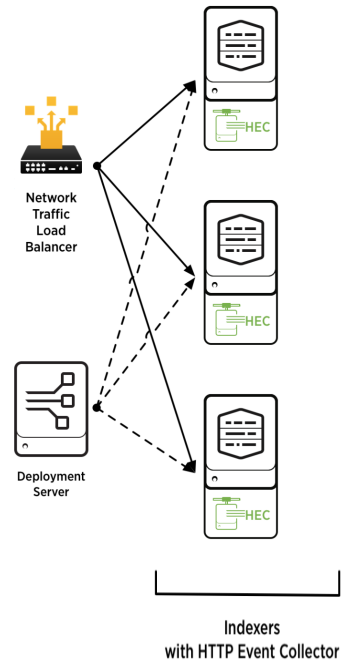
Streamgest – Priceline's home grown application



Single Streamgest Node

Consumers | **Plugins** | **Event Batches** | **HTTP Handlers**

- **Message Consumers**
  - Read messages from Kafka topics
  - Assign metadata from the Config Management System
  - Parse, munge and route events through the processing pipeline

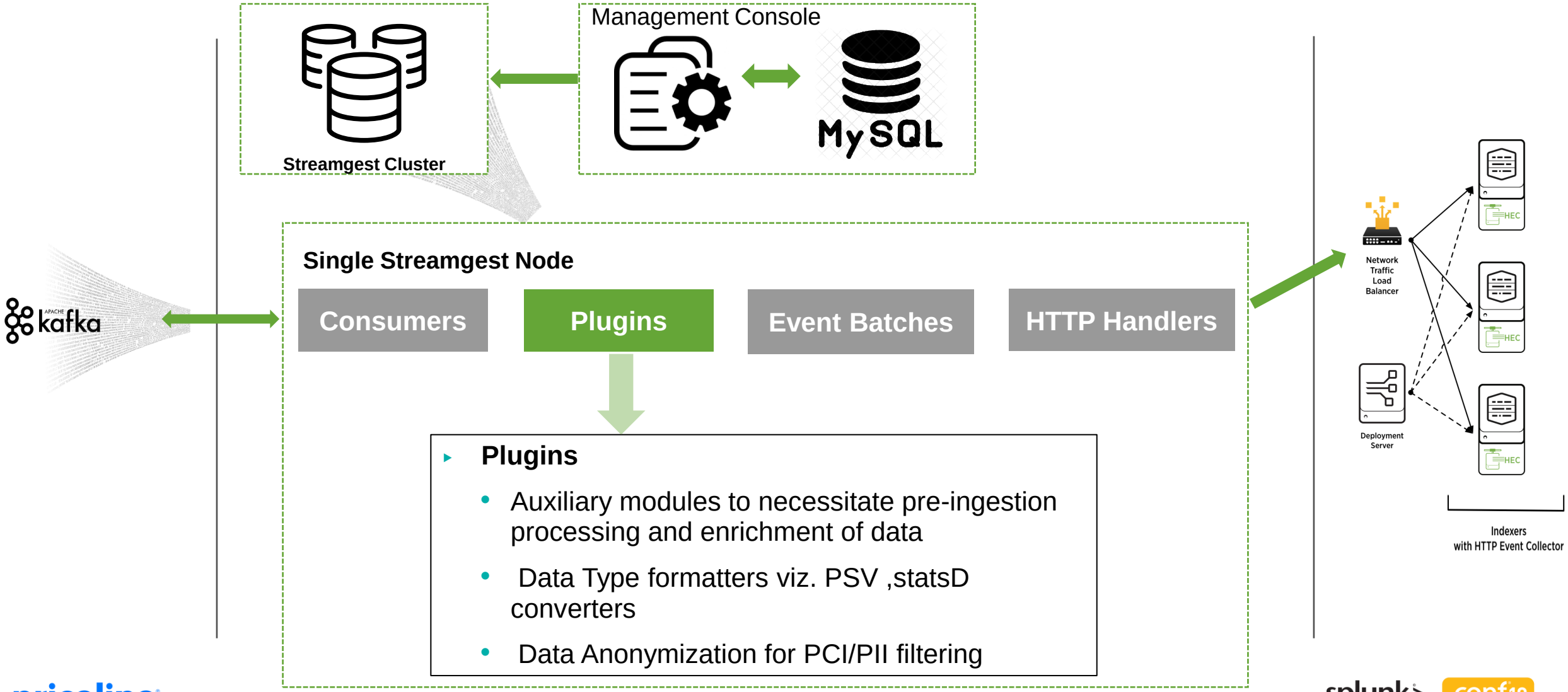


priceline

splunk> .conf18

Data Processor

Streamgest – Priceline's home grown application

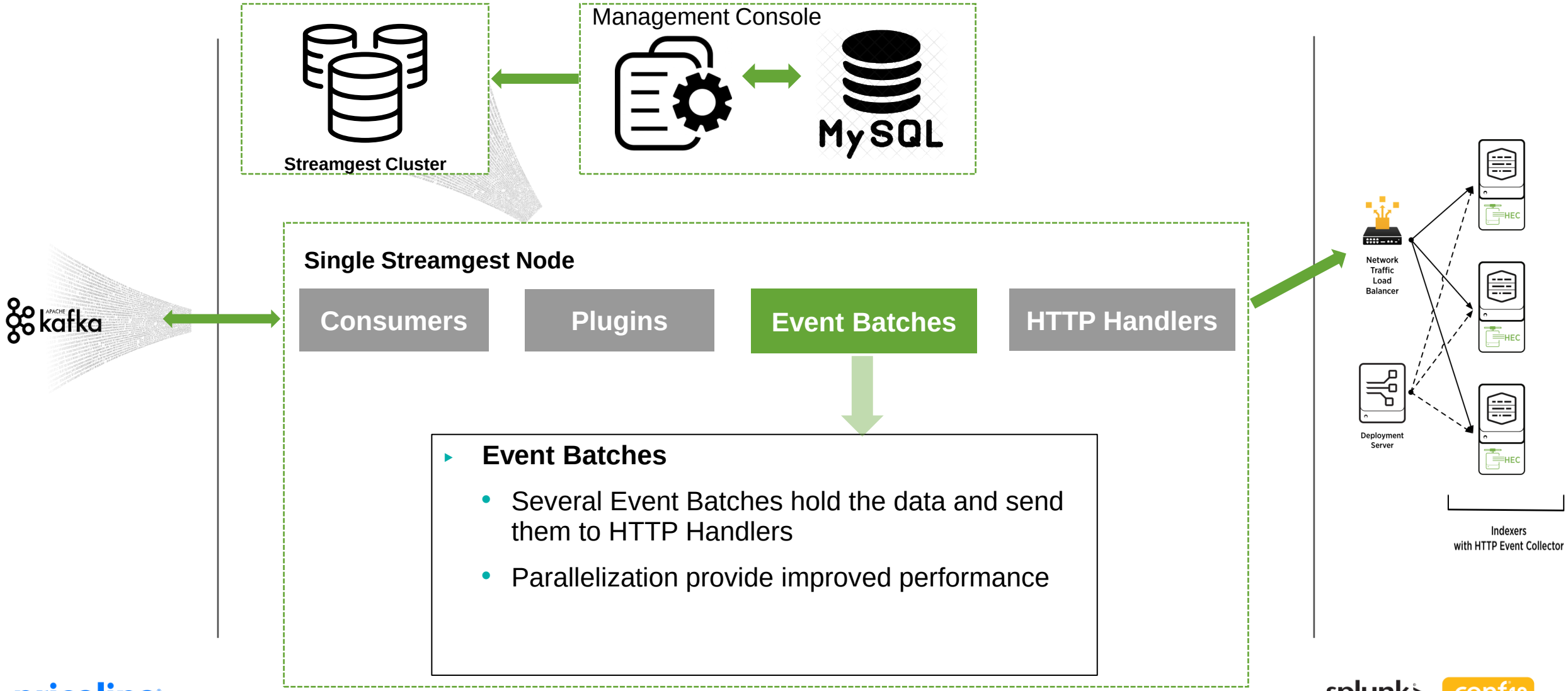


priceline

splunk> .conf18

Data Processor

Streamgest – Priceline's home grown application



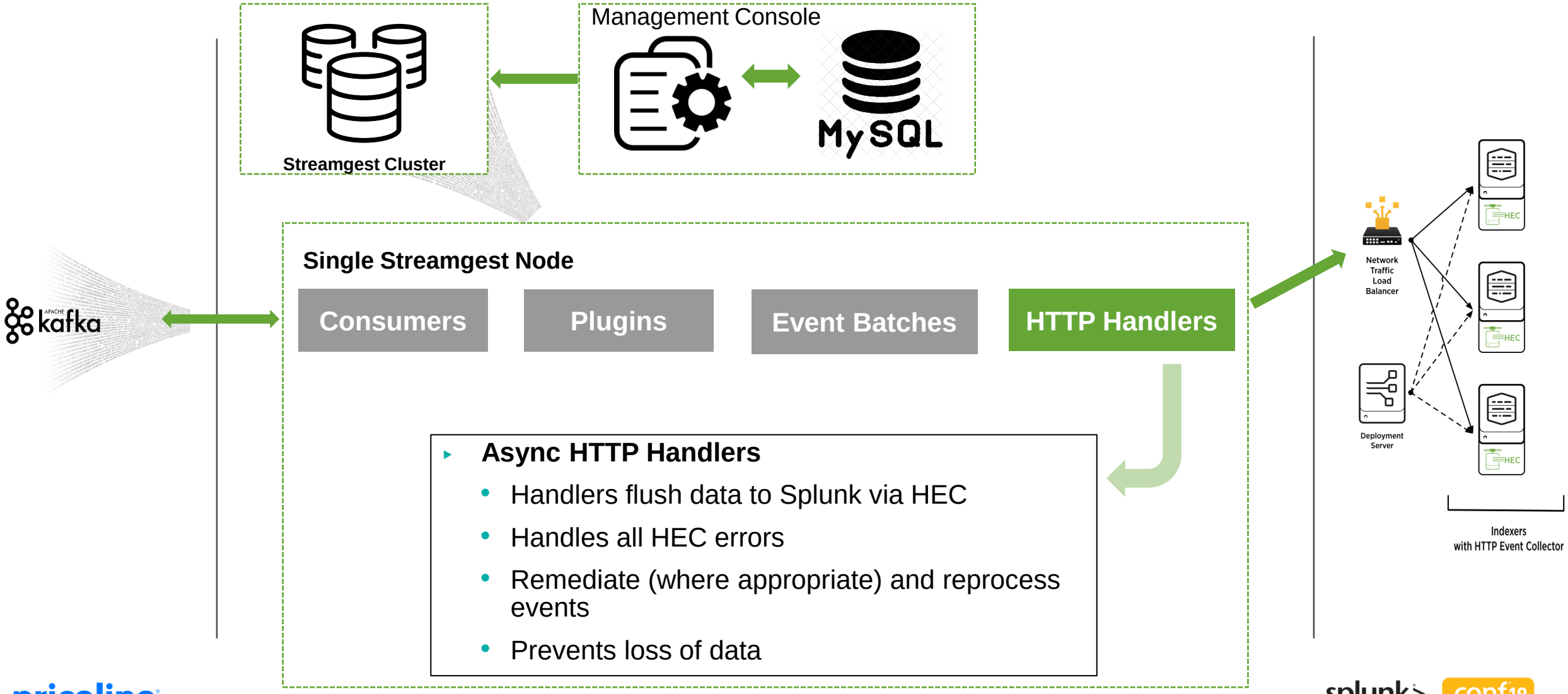
- **Event Batches**
  - Several Event Batches hold the data and send them to HTTP Handlers
  - Parallelization provide improved performance

priceline

splunk> .conf18

Data Processor

Streamgest – Priceline's home grown application



- **Async HTTP Handlers**
  - Handlers flush data to Splunk via HEC
  - Handles all HEC errors
  - Remediate (where appropriate) and reprocess events
  - Prevents loss of data

priceline

splunk> .conf18

Data Processor (Monitoring)

Streamgest – Priceline's home grown application

VA - Message Rate vs Consumption Rate

Click on events to drilldown

DATA_CENTER	CONSUMER_GROUP	Message Rate	Consumption Rate
VA-STREAM	streamgest_cg_eleven-splunk	8590/s	8672/s
VA-STREAM	streamgest_cg_htl_hrapi-splunk	5030/s	5001/s
VA-STREAM	streamgest_cg_three-splunk	4011/s	4027/s
VA-STREAM	streamgest_cg_six-splunk	3581/s	3625/s
VA-STREAM	streamgest_cg_common_one-splunk	3474/s	3487/s
VA-STREAM	streamgest_cg_ten-splunk	3165/s	3133/s

Lag Threshold

> 5k

Current Streamgest Consumer Lag (Status=Alert)

GROUP	CLUSTER	TOTAL_LAG	values(TOPIC)
streamgest_cg_htl_hrapi-splunk	VA-STREAM	28224	bam_hrapi
streamgest_cg_seti_one-splunk	VA-STREAM	27929	ace-stats_splunk bam_setisvcs
streamgest_cg_three-splunk	NY-STREAM	27604	bam_hcfetch
streamgest_cg_eleven-splunk	VA-STREAM	25347	appmetrics bundlebook_data bundleprice bundlesearch_data hrcs_ops plconnect_ops

priceline

splunk> .conf18

Data Processor (Alerts)

Streamgest – Priceline's home grown application

Splunk APP 2:30 PM

Kafka Consumer Lag

This alert is triggered if Kafka Consumer Lag > 250K per consumer group!

Trigger History Search Results:

```
VA-STREAM streamgest_cg_common_logs-splunk  1478680
```

Splunk APP 11:50 AM

CPS-STREAMGEST-APPLNLOG-ERROR

This alert is triggered if number of error are more than 3 in Streamgest Servers

Trigger History Search Results:

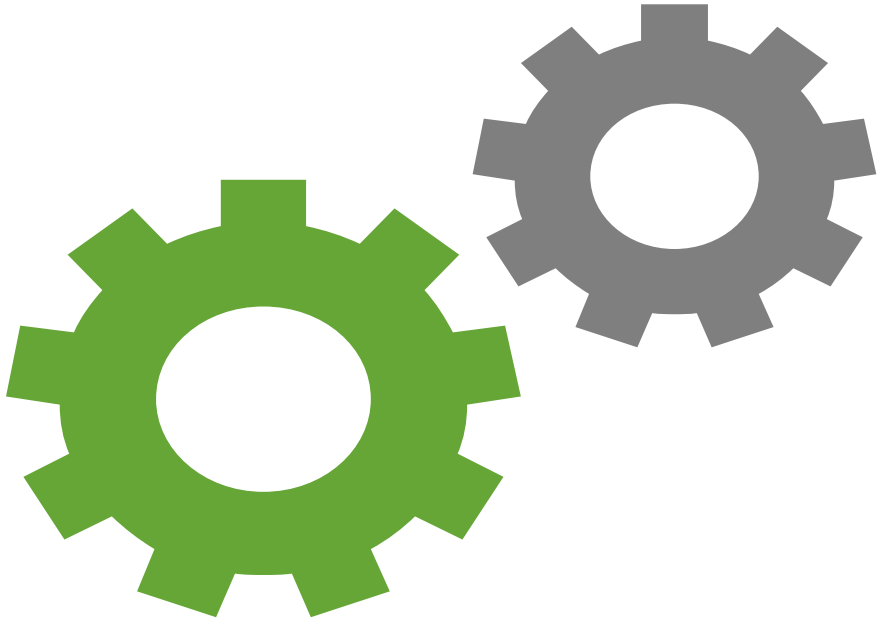
```
ny-[redacted]
```

Key Takeaways

Lessons Learned

priceline

HEC Tuning



Do. Defaults are good, but customization may be better. [limits.conf]

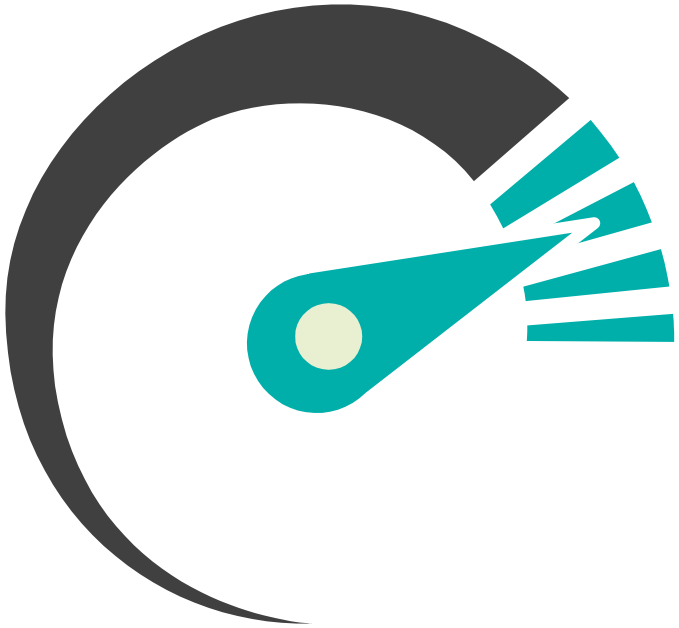
Example: `**max_content_length**` has been modified to suit our bandwidth requirement.

Key Takeaways

Lessons Learned

priceline

Client-side Tuning



Do. Send events in **batches** to maximize HEC performance

Key Takeaways

Lessons Learned

priceline

Client-side Tuning

Do. Keep-Alive allows HTTP clients to reuse the same connection for multiple batches.

Key Takeaways

Lessons Learned

Send data over HTTPS only when required. More performance while sending data over HTTP.

Key Takeaways

Lessons Learned

priceline

Handling Errors

Do. HEC errors should be handled gracefully and iteratively to prevent loss of data.

Q&A

Jagadeesh Motamarri | Senior Software Engineer
Mukund Narayana Murthy | Software Engineer
priceline

priceline

splunk> .conf18

Thank You

Don't forget to rate this session in the .conf18 mobile app

priceline®