# Don’t Leave Money On The Table!

## How to tap into machine data for observability and business analytics

Karun Subramanian
IT Operations Expert

www.karunsubramanian.com

# About the Presenter

- 20+ Years of experience in Systems and Network Administration, Software Development and Monitoring & Observability
- Passionate about Machine Data Analytics at Scale
- Focused on modernizing IT Operations
- Splunk Certified Architect

# What will you learn in this session?

- Identify machine data in your org (Hint: It’s lot more than logs)
- The Hidden values in machine data
- Architectural patterns to collect, ingest and index Machine data
- Real world examples on how organizations are tapping into Machine data
- Developing a Machine data strategy

# Machine Data

# What is Machine Data?

Digital exhaust produced by any device in the Network

**Events**

A state change; an occurrence of something

**Application Logs**

Typically diagnostic information, including traces

**Metrics**

Measurement of a property

# Machine data answers “What”, “Where” and “Why” of the reality of a System

# Machine data is everywhere

- Authentication
- Audit
- Middleware
- OS
- OS Performance
- Network device
- Network packets
- Web Server
- Sensors
- IoT Devices
- Database
- Messaging Systems
- CI/CD
- Automation programs
- Mail Server
- LDAP Server
- Active Directory
- Containers
- Kubernetes/Container Orchestration
- Applications
- API
- Event viewer
- Mobile devices
- Call Detail records

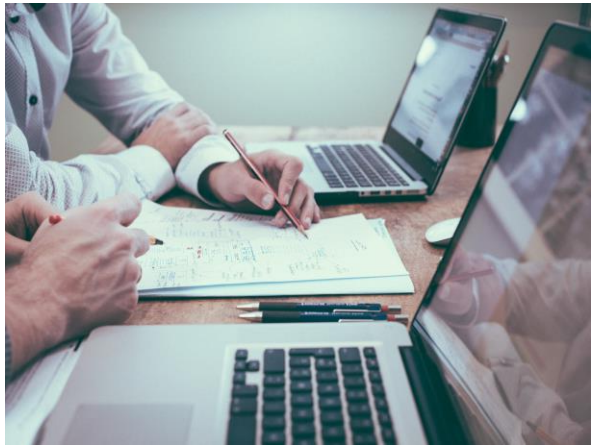
# What can you do with it ?


IT Operations/Monitoring

A spike in 500 internal server errors

Security/SIEM

A spoofing attack



Business analytics

How many repeat customers in the past month?

# Why is it hard to reap benefits from Machine Data?

Fast

Millions of records/sec

Huge

Multiple tera bytes per day

Mostly Unstructured

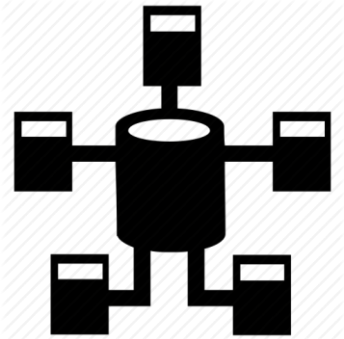
Logs/Traces

(Distributed)$^2$

A formidable challenge

Fun fact: IDC predicts the annual data generated will be 175 Zetta Bytes by 2025. (175 Billion Terabytes. Go figure)

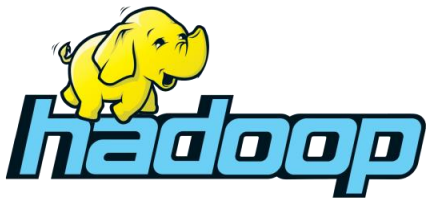
# Why Traditional Datastores won't cut it?



Data Warehouse

Complex, long process to get data in (ETL or ELT)

Not suitable for search and monitoring use case




Hadoop/Hbase

Not a low-latency system.

Complex data retrieval and processing. Need of an efficient MapReduce job

RDBMS

Machine data is primarily time-series. RDBMS is not suited for time-series data. Scalability becomes a bottleneck.

Give everyone the data analysis capabilities; not just the Data scientists.

# How does it look like?

**Apache Web Server Access Log**

192.168.198.92 - - [22/Dec/2002:23:08:37 -0400] "GET / HTTP/1.1" 200 6394 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1...)" "-"
192.168.198.92 - - [22/Dec/2002:23:08:38 -0400] "GET /images/logo.gif HTTP/1.1" 200 807 www.yahoo.com "http://www.some.com/" "Mozilla/4.0 (compatible; MSIE 6...)" "-"
192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET /news/sports.html HTTP/1.1" 200 3500 www.yahoo.com "http://www.some.com/" "Mozilla/4.0 (compatible; MSIE ...)" "-"
192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET /favicon.ico HTTP/1.1" 404 1997 www.yahoo.com "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv:1.7.3)..." "-"

**Linx PAM log**

Jul 7 10:51:24 srbarriga su(pam_unix)[14592]: session opened for user test2 by (uid=10101)
Jul 7 10:52:14 srbarriga sshd(pam_unix)[17365]: session opened for user test by (uid=508)
Nov 17 21:41:22 localhost su[8060]: (pam_unix) session opened for user root by (uid=0)
Nov 11 22:46:29 localhost vsftpd: pam_unix(vsftpd:auth): authentication failure; logname= uid=0 euid=0 tty= ruser= rhost=1.2.3.4

**Linux /var/log/messages**

Aug 16 22:49:37 tiger /bsd: uid 1000 on /var/www/logs: file system full

**Cisco pix firewall logs**

Sep 7 06:25:28 PIXName %PIX-6-302013: Built inbound TCP connection 141968 for db:10.0.0.1/60749 (10.0.0.1/60749) to NP Identity Ifc: 10.0.0.2/22 (10.0.0.2/22)
Sep 7 06:25:28 PIXName %PIX-7-710002: TCP access permitted from 10.0.0.1/60749 to db:10.0.0.2/ssh
Sep 7 06:26:20 PIXName %PIX-5-304001: 203.87.123.139 Accessed URL 10.0.0.10:/Home/index.cfm
Sep 7 06:26:20 PIXName %PIX-5-304001: 203.87.123.139 Accessed URL 10.0.0.10:/aboutus/volunteers.cfm

**SSHD log**

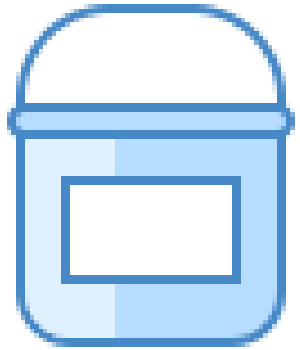
Aug 1 18:27:45 knight sshd[20325]: Illegal user test from 218.49.183.17
Aug 1 18:27:46 knight sshd[20325]: Failed password for illegal user test from 218.49.183.17 port 48849 ssh2
Aug 1 18:27:46 knight sshd[20325]: error: Could not get shadow information for NOUSER
Aug 1 18:27:48 knight sshd[20327]: Illegal user guest from 218.49.183.17
Aug 1 18:27:49 knight sshd[20327]: Failed password for illegal user guest from 218.49.183.17 port 49090 ssh2

Source: https://ossec-docs.readthedocs.io

# Architecture

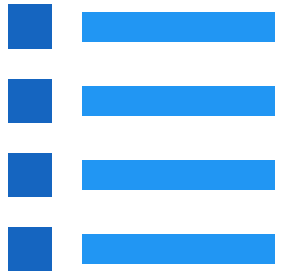
# Considerations

Search and Visualize (need of an inverted index)



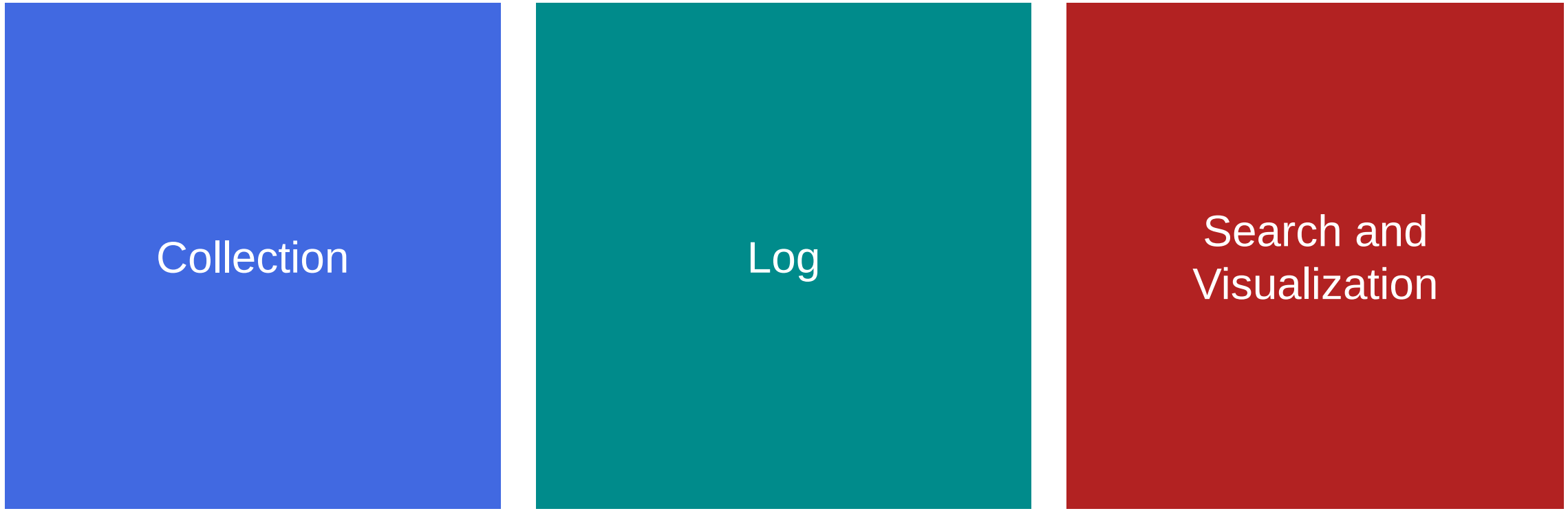
Time bucketing

Near real-time

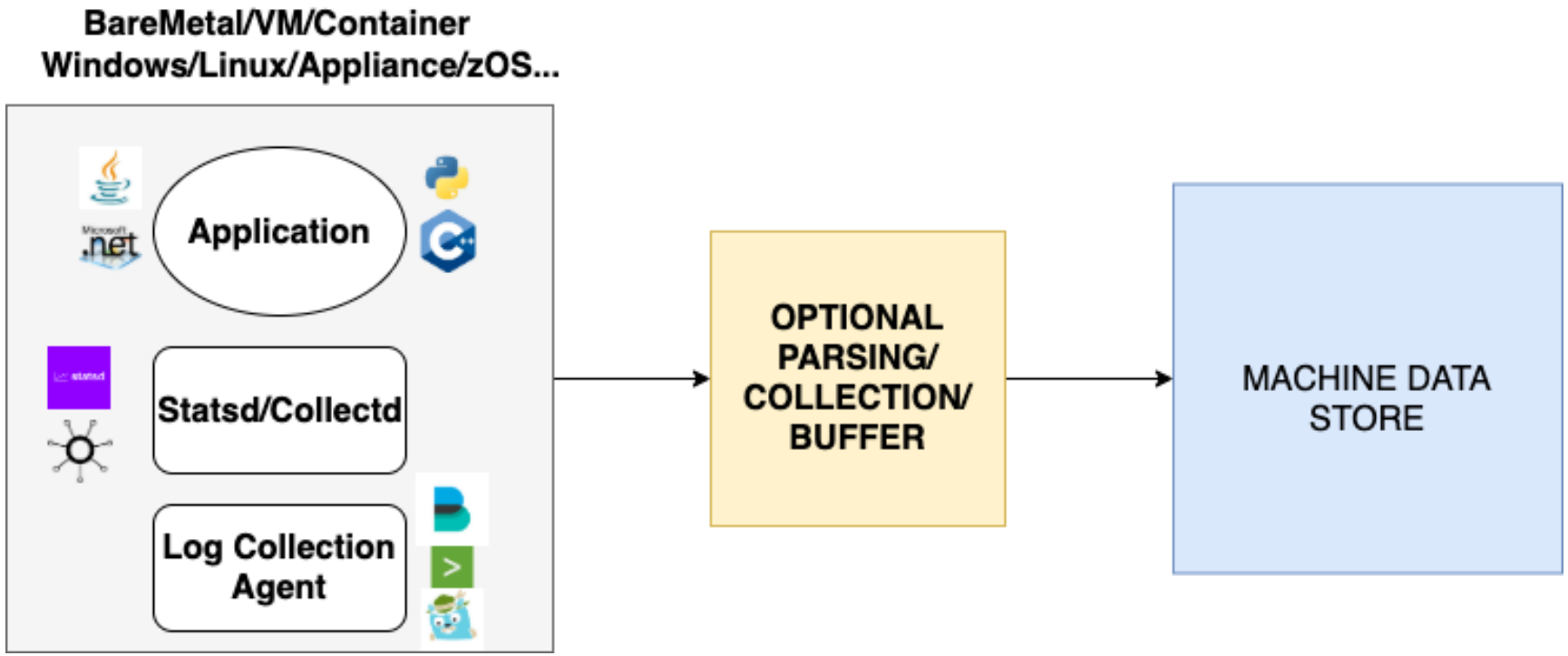


Index Events, Metrics and Logs

# Building Blocks



# Collection: Agent Based



# Collection: Agent Based

- Agents collect data and push to backend. In most cases, this is the most effective method
- Generally low footprint

Examples:

  - collectd/statsd
  - APM agents
  - Log collection agents (Beats,Splunk Universal Forwarder)

- Tricky in Cloud environments

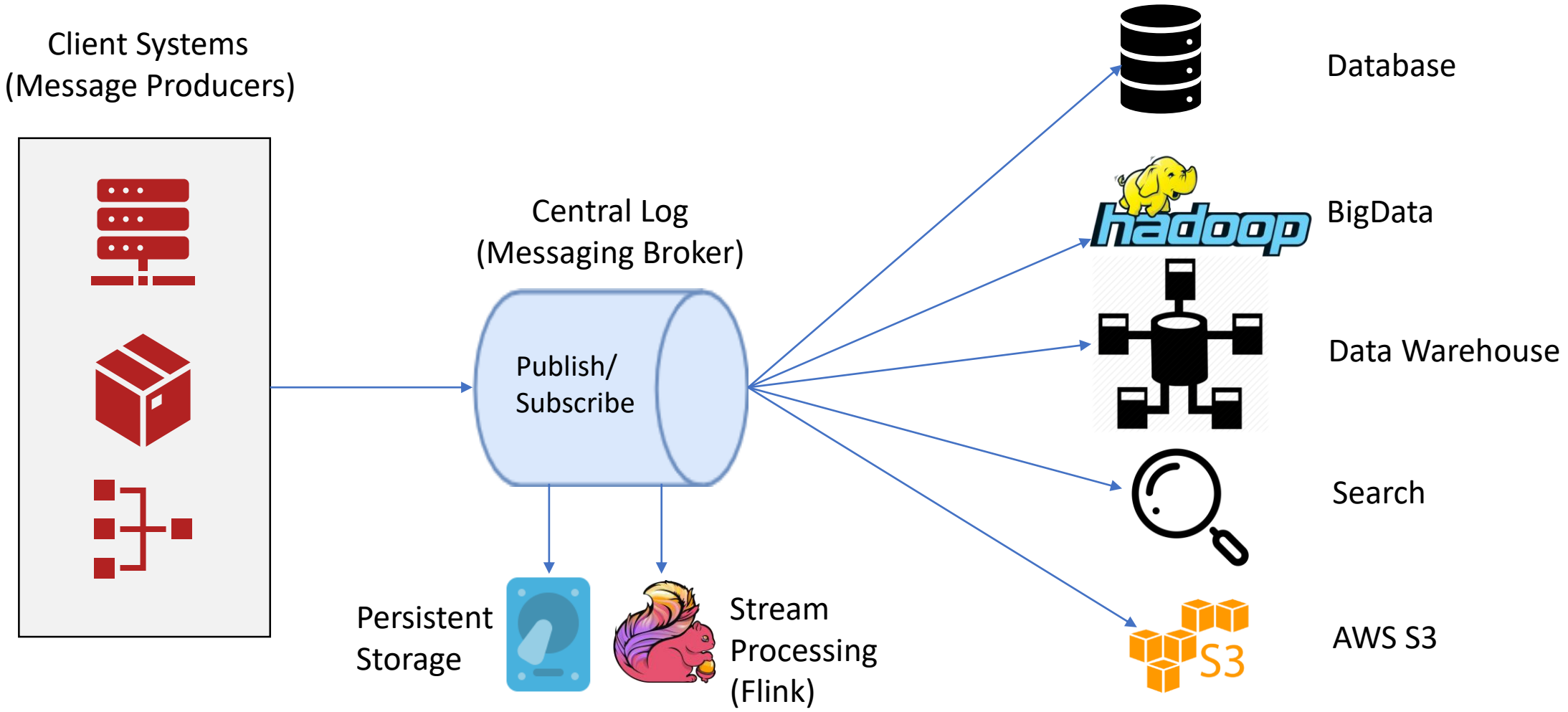
# Collection: Agentless

- Pull mechanism discouraged
- Push from application. Code changes required in some cases
- HTTP POST
- Kafka producer
- Open Tracing (A specification. Some implementations like Jaeger use Agents)

# Collecting in the Cloud

- Inherently difficult due to the ephemeral nature of the containers
- Docker/Kubernetes documentation is NOT clear when it comes to application logs
- Use Agentless mechanisms (HTTP, kafka producer) for application logs
- Use native mechanisms (Fluentd) for Container logs

# LOG Middleware



# LOG: Why a messaging middleware?

- Separation of subscriber and producer
- Buffering
- Speed of processing
- Retention
- Stream processing

# The Kafka difference


Speed

Can easily achieve 2 Million messages/sec


Data Persistence

Configurable retention

(Default 7 days)


Scales Linearly

Partitioning log helps in scaling linearly.

Messaging is not new. But never before a messaging system was created with this speed and scalability

# Search and Visualization using Timeseries data

- Need of a tool that maintains an inverted index (not much different from traditional search engines.
- A tool that crunches both unstructured text and metrics data
- Need to be able to produce rich visualization
- Examples: Solr, Elastic Search, Splunk

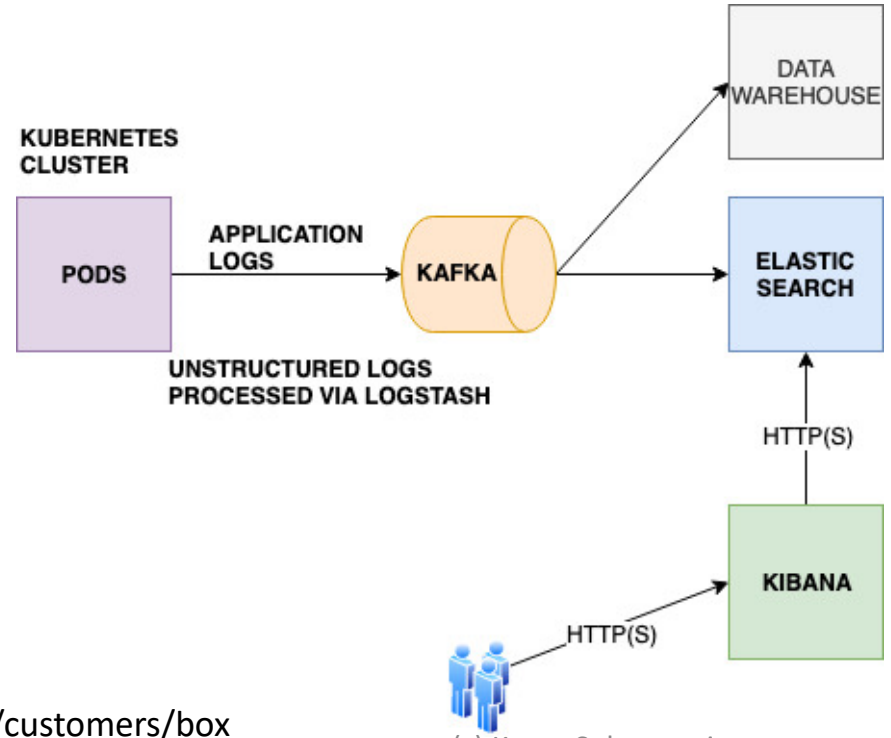
# Case Studies

# BOX

Cloud Storage Provider

Use case: Observability using Machine Data (Application and Operational Logs)

20 TB/day ingestion, 180 billion documents, 190TB total size




Source : https://www.*elastic*.co/customers/box

# Carnival Cruise Lines

World’s Largest Cruise Line

Use case: Observability using Machine Data (Application and Operational Logs), Security

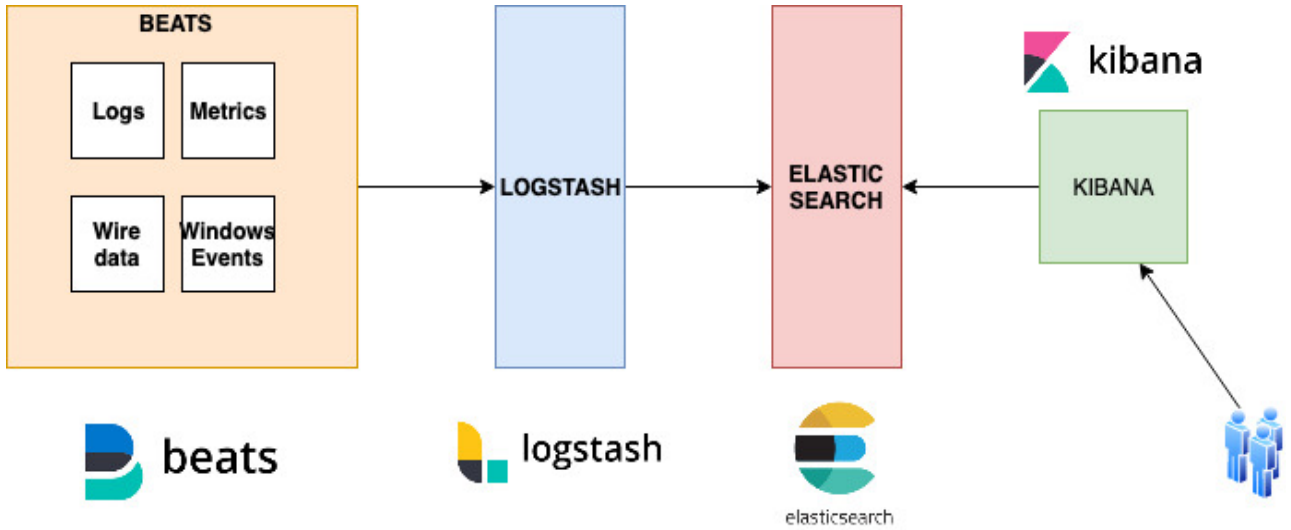
Data Sources: Applications, Satellites, Shipboard systems, Connected devices

Consolidates data from all the ships and corporate offices around the world

Source : https://www.splunk.com/en_us/customers/success-stories/carnival.html

# Harel Insurance & Financial Services

- One of Israel's largest insurance groups
- Use Case: IT Operations
- 25 Billion documents, 14.5 TB Total data size




*Source: https://www.elastic.co/customers/harel-insurance-and-financial-services*

# Machine Data Strategy

# Execution

- Establish an on-boarding process
- LOG (Kafka) the central component
- Dev team owns the content & structure of data
- Search and Visualize Platform
- Attack OS metrics first, if applicable

Next Gen IT Ops: Stream processing Machine data

To reap benefit from Machine Data, you must be able to collect, index, correlate and analyze in near real-time

# Questions?