An unconventional approach to the BigData projects: self-services data with Kafka, Elastic and Dremio

Stefano Pampaloni, CEO @Seacom srl
Typical High Level Architecture

Real-time Data Ingestion
Typical High Level Architecture

Real-time Data Ingestion

Stream Processing

Storage
Typical High Level Architecture

- Real-time Data Ingestion
- Stream Processing
- Storage
- Data Publishing / Visualization
Beats
Log Files
Metrics
Wire Data
your{beat}

kafka

Datstore
Web APIs
Social
Sensors
What’s Apache Kafka?

• Pub/Sub Logic - Messaging, Done Right
• STORE - Hadoop, Made Fast
• ETL & Data Integration as a Platform
• Stream processing
Stream Processing
Apache Kafka: A distributed streaming platform
STREAMING PLATFORM

- Apps
- Search
- Streaming Processing
- DWH - Hadoop
- Monitoring
- Real-time Analytics
- KAFKA
Big data was
The more the better

Stream data is
The faster the better

Stream data can be
big or fast (Lambda)

Stream data will be
big AND fast (Kappa)

Apache Kafka is the enabling technology of this transition
Beats

- Log Files
- Metrics
- Wire Data
  your(beat)

kafka

- Database
- Web APIs
- Social
- Sensors

#RedHatOSD
Beats

Log Files
Metrics
Wire Data
your{beat}

kafka

Datashare
Web APIs
Social

Custom UI
How many clusters do you count?

- Apache Kafka
- Storm, Flink, Spark Streaming, Ignite, Akka Streams, Apex, ...
- HDFS, NFS, Ceph, GlusterFS, Lustre, ...
- NoSQL (Cassandra, HBase, Couchbase, MongoDB, ...) or Elasticsearch, Solr, ...
Simplicity is the ultimate sophistication

Apache Kafka
Distributed Streaming Platform

Publish & Subscribe
to streams of data like a messaging system

Store
streams of data safely in a distributed replicated cluster

Process
streams of data efficiently and in real-time

Node.js
Beats
Log Files
Metrics
Wire Data
your{beat}

Elasticsearch
Master Nodes (3)
Ingest Nodes (X)
Data Nodes – Hot (X)
Data Notes – Warm (X)

Kibana
Instances (X)

kafka

Social
Datastore
Web APIs

Custom UI
Beats Log Files

Wire Data your\{beat\}

Elasticsearch

Master Nodes (3)
Ingest Nodes (X)
Data Nodes – Hot (X)
Data Notes – Warm (X)

Kibana

Instances (X)

Logstash

Nodes (X)

kafka

Datastore
Web APIs
Social

Elastic Stack

100% open source
No enterprise edition
Open Source Log Management

The Elastic Stack (sometimes known as the ELK Stack) is the most popular open source logging platform. Here's why.
Da strumento a Soluzione

Logging

#ESinADay2018

#RedHatOSD
Da strumento a Soluzione

Grey Lawson Style Sofa
$799.00 ★★★★★

Eggshell Mid Century Modern Sofa
$795.00 ★★★★

Classic Contemporary Loveseat Sofa
$499 ★★★

Charcoal Contemporary Loveseat Sofa
$499 ★★

Tan Contemporary Mid Century Sofa
$599 ★

#ESinADay2018
#RedHatOSD
Da strumento a Soluzione

PREPARE -> PROTECT -> PRIVACY PROCESSES

Data Mapping -> Privacy Impact Assessment -> Data Role & Legal Basis Confirmation -> Data Retention Planning -> Vendor and Sub-processor Review -> Security Measures -> Data Subject Rights Processes -> Data Transfer Processes -> Incident Response & Notification

TECHNICAL AND ORGANIZATIONAL MEASURES FOR SECURING DATA

Da strumento a Soluzione
Beats
Log Files
Metrics
Wire Data
your{beat}

Elasticsearch
Master Nodes (3)
Ingest Nodes (X)
Data Nodes – Hot (X)
Data Notes – Warm (X)

Dremio

Kibana

Logstash
Nodes (X)

D3

Custom UI

Seacom

#RedHatOSD
Your analysts and data scientists are hungry for data
Data is a massive engineering project today

- Custom ETL
- Fragile transforms
- Slow moving
Data is a massive engineering project today

Data Staging

- High overhead
- Proprietary lock in
- Custom ETL
- Fragile transforms
- Slow moving

Data Warehouse

- $$$

Tools:
- Tableau
- Power BI
- Qlik
- Excel
- Python
- R
- Spark
- SQL

dremio
Data is a massive engineering project today

- Data sprawl
- Governance issues
- Slow to update
- $$$
- High overhead
- Proprietary lock in
- Custom ETL
- Fragile transforms
- Slow moving

Cubes, BI Extracts & Aggregation Tables

Data Warehouse

Data Staging

SEACOM

#RedHatOSD
Data Catalog
Interactive data discovery, enterprise and personal data assets

Data Curation
Wrangle, prepare, enrich, join any source without making copies of your data

Data Acceleration
OLAP and ad hoc queries at interactive speed, without cubes or BI extracts

SQL Execution Engine
Apache Arrow-based columnar engine operates on any source: RDBMS, NoSQL, S3, Hadoop
These popular first projects are easy wins

- **Data Lake**: Accelerate SQL on your Data Lake
- **Hybrid Lake**: Join across Data Lake + other systems
- **Self-Service Data**: Across all your systems, on-prem + cloud

---

SEACOM

#RedHatOSD
Beats
Log Files
Metrics
Wire Data
your{beat}
Beats
Log Files
Metrics
Wire Data
your{beat}

Elasticsearch
Master Nodes (3)
Ingest Nodes (X)
Data Nodes – Hot (X)
Data Notes – Warm (X)

Logstash
Nodes (X)

Kibana
Instances (X)

Dremio

kafka

Social
Sensors

Web APIs
Datastore

#RedHatOSD
GRAZIE PER L’ATTENZIONE

Stefano Pampaloni, CEO @Seacom srl