A Big Data Platform for heterogeneous data collection and analysis in large-scale data centers

S. Rossi Tisbeni

CNAF is the national center of INFN (Italian Institute for Nuclear Physics) for Research and Development on Information and Communication Technologies

• Focus on **Software development and distributed systems**: development of Grid middleware for WLCG and of CLOUD technologies (coordination of the distributed INFN Cloud infrastructure)

• Hosts the Italian **Tier-1** data center for HEP experiments
  
  ◦ Provides resources, support and services needed for data storage, distribution and processing
  
  36PB of data on disk storage, 86PB of data on tape storage, 400k HS06

• Hosts the **INFN-TTLab** that through Technology Transfer activities provides industrial partners with resources and know-how
Context (1/2)

INFN Tier-1 datacenter

- Monitoring system based on Sensu, InfluxDB, Grafana
  - General datacenter KPIs
  - Specific KPIs for batch system, storage, network and cloud services
  - Specific dashboards for critical services based on Elastic stack

INFN Cloud infrastructure across several sites

- Monitoring system based on Zabbix, ELK, Grafana
  - Cloud infrastructure monitoring and accounting
  - Tenant VMs and applications monitoring
Context (2/2)

• IoTwins EU project to create a big data platform for optimized and replicable industrial and facility management models
  ◦ Methodology for the monitoring infrastructure reuse and deployment in new and different contexts
  ◦ Focus on monitoring of IoT/Edge/Cloud integrated infrastructure
  ◦ Predictive analysis of fault and support on troubleshooting
  ◦ https://iotwins.eu

• EPIC Cloud (for applications with high security requirements)
  ◦ Alarm system based on Sensu/Uchiwa/Slack
  ◦ Monitoring system based on Sensu/Grafana/InfluxDB/Prometheus
  ◦ In the future a system for monitoring security events and enrich these data with a remote threat intelligence source will be probably needed
Big Data Platform / Goals

- Centralize **collection** of logs, metrics and data from **heterogeneous** data sources and present them in a **flexible and actionable** format
- Harmonize the current **monitoring** infrastructures in production at the center
- Provide different storage solution with proper retention policies based on security and privacy requirements
- Provide policy-based, **authenticated** access to the data
- Provide sysadmins and users with tools for **information discovery** from system data and monitoring
- Develop a reliable, extensible, scalable and manageable platform for collection and analysis of **Big Data** to be offered as a service
A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
Ingestion & Aggregation

- **Filebeat** for log ingestion
  - Tail on file with variable structure and multiline support
  - Can be deployed as puppet module on production machines
  - Rsyslog already in use for critical service

- **Fluentd** for aggregation, format, and filtering
  - Small footprint, easy to install, open source
  - JSON native
  - Flexible in refactor and filtering
  - Supports various sources and outputs via Plugins

Local persistent cache with buffer

Guarantee At Least Once delivery to the broker
Broker

Apache Kafka, topic based publish-subscribe data distribution system

• Decouple data ingestion and consumption
• Allow broadcasting data to multiple subscribers simultaneously
• Allow for fault tolerance and high performance

Proposed setup:

• 3 brokers with 3 replicas for each partition
• Retention set to 3 days
Architecture / Ingestion

A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
Storage

The **Elastic Stack** acts as Short-Term Storage database for log files

- **Logstash:**
  - Routing data from Kafka to the Elastic Index
  - Wide use of filters to parse information from logs
  - At Least Once buffered delivery

- **Elasticsearch:**
  - Performs indexing of data
  - Stores logs as JSON objects
  - Expose SQL like interface for queries

- **Kibana:**
  - Discovery and visualization through dashboard

---

PoS(ISGC2019)027
Storage

MinIO as Long-Term Object storage solution

• Disaggregate storage and computing resources
• Cloud native architecture, can be federated with other clusters
• High-performance, bucket-based object storage
• Policy based, OIDC Authenticated access through INDIGO IAM
• S3 and REST API for data access

Integrated with storage technologies from INFN Cloud
Architecture / Consumers

Broker

Topic

Topic

Topic

Routing

Routing

Elastic Index

Object Storage
Batch Analysis

• **Jupyter**
  ◦ Provides a web-based frontend for users to access the data
  ◦ Open Source
  ◦ Supports multiple programming languages (Python, Scala, Java)
  ◦ Simultaneous access to host from multiple users
  ◦ Integration with OpenID connect through INDIGO IAM

• **Spark:**
  ◦ Supports high-performance batch analysis with map/reduce paradigm
  ◦ Access to diverse data sources, including time-series database and s3 Object storage
  ◦ Run natively on cloud in orchestrated clusters
  ◦ Supports Python, Scala and R
  ◦ Powers dataframe handling and machine learning libraries
Orchestration

Kubernetes: modern orchestrator for containerized services

- Manages the deployment of JupyterLab single-user containers (spawned by JupyterHub)
- Allows to allocate different virtual resources (CPUs, RAM) to each user
- Allows each user to spawn containerized environment through JupyterHub
- Use sandbox containers as users workspace
- Manages the deployment of Spark executors with fine control on the number of executors and resources for each executor (CPU cores, RAM)
- ReplicaSet guarantee: a new pod is automatically instantiated on the next available node with almost zero downtime if a cluster node goes down
Architecture / Batch Analysis

A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
## Sample Workflow

### Time | Hostname | Name | Raw | HTTP Code | Path | WD Method
---|---|---|---|---|---|---
Mar 19 07:56:30 | ds-816 | storm-webdav-server-access | odC.physics_Main.PhysCont_DA00.t0pro22_v81_a21296.20201123.log.23644713.017584.log.tgz HTTP/1.1 | 207 | DELETE | /atlas
Mar 19 07:56:30 | ds-816 | storm-webdav-server-access | 2001:1458:381:0:0:0:100:32f -- | -- | 200 | GET | /well-known/oauth-authorization-server HTTP/1.1
Mar 19 07:56:30 | ds-816 | storm-webdav-server-access | 188.183.82.95 -- | -- | 204 | DELETE | /atlas/atlasscratchdisk/rucio/user/sgurudas/d aio/0.0/sgurudas.23648449_013534.output.root HTTP/1.1
Sample Workflow

Request per WebDAV endpoints

Methods requested to WebDAV

23/03/2021
Sample Workflow

Connect to MinIO

Read Logs

```python
[1]: conf = SparkConf().set("spark.executor.core", "2")
    spark = SparkSession.builder.getOrCreate()
    sc = spark.context.getOrCreate(conf)
    sc

[3]: SparkContext

Spark UI

Version

Master

AppName

v3.1.1

k8s://https://kubernetes:443

 pyspark-shell

[3]: df = spark.read.json("s3a://storage-en/dr-s-916/2021/03/28/storm-webdev-server-*.json")
    df.toPandas()
```

<table>
<thead>
<tr>
<th>@timestamp</th>
<th>@version</th>
<th>hostname</th>
<th>http_code</th>
<th>name</th>
<th>path</th>
<th>raw</th>
<th>throughput</th>
<th>topic</th>
<th>wd_method</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021-03-20T07:13:39.6852</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(auth/tokens)</td>
<td>2001458:301:cd:0:0:0:0:0:100:32</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>POST</td>
</tr>
<tr>
<td>2021-03-20T07:14:56.627</td>
<td>1</td>
<td>ds-916</td>
<td>207</td>
<td>storm-webdev-server-access</td>
<td>(atlas/atlasdataak)</td>
<td>2001458:301:cd:0:0:0:0:0:100:32</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>PROPFIND</td>
</tr>
<tr>
<td>2021-03-20T07:14:56.627</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(status)</td>
<td>153:1:0:1:0:0:0:0:100:32</td>
<td>2.978876+06</td>
<td>storage-test-err</td>
<td>GET</td>
</tr>
<tr>
<td>2021-03-20T07:13:19:1077</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(well-known)/auth-authorization-server</td>
<td>2001458:301:cd:0:0:0:0:0:100:32</td>
<td>8.366667+04</td>
<td>storage-test-err</td>
<td>GET</td>
</tr>
<tr>
<td>2021-03-20T07:13:27:4962</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(well-known)/auth-authorization-server</td>
<td>2001458:301:cd:0:0:0:0:0:100:32</td>
<td>6.2000000+04</td>
<td>storage-test-err</td>
<td>GET</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2021-03-20T07:07:08:6742</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(atlas/atlasdataak)</td>
<td>2001458:301:cd:0:0:0:0:0:100:313</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>HEAD</td>
</tr>
<tr>
<td>2021-03-20T07:07:07:3327</td>
<td>1</td>
<td>ds-916</td>
<td>404</td>
<td>storm-webdev-server-access</td>
<td>(atlas/atlasdataak)</td>
<td>2001458:301:cd:0:0:0:0:0:100:32</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>HEAD</td>
</tr>
<tr>
<td>2021-03-20T07:03:06:58:2552</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(auth/tokens)</td>
<td>2001458:301:cd:0:0:0:0:0:100:305</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>POST</td>
</tr>
<tr>
<td>2021-03-20T07:03:06:58:4562</td>
<td>1</td>
<td>ds-916</td>
<td>200</td>
<td>storm-webdev-server-access</td>
<td>(auth/tokens)</td>
<td>2001458:301:cd:0:0:0:0:0:100:305</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>POST</td>
</tr>
<tr>
<td>2021-03-20T07:03:06:58:5382</td>
<td>1</td>
<td>ds-916</td>
<td>404</td>
<td>storm-webdev-server-access</td>
<td>(atlas/atlasdataak)</td>
<td>2001458:301:cd:0:0:0:0:0:100:305</td>
<td>NaN</td>
<td>storage-test-err</td>
<td>PROPFIND</td>
</tr>
<tr>
<td>201546 rows x 30 columns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Next Steps / Anomaly Detection use case

In the context of the IoTwins project, replicate Anomaly Detection use case from CINECA datacenter on CNAF data

• Parse unstructured log data to find possible correlations with anomalies
• Perform clusterization to identify clusters on the base of keywords
• Use the result of the clusterization to perform semi-supervised anomaly detection
• Perform the analysis in streaming to label the new entries as they are included in the dataset

This analysis would greatly benefit from the Big Data Platform
Next Steps / Streaming Analysis

A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
Next Steps / Metrics Integration

- Current:
  - Sensu
  - influxdb
  - Grafana
  - Kafka

- Future:
Next Steps / Thingsboard

ThingsBoard

Device

Rule Engine

Realtime Dashboard

Kafka Plugin

Broker
A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021
Log file’s structure

- Log files range from heavily structured to free flowing and descriptive

```
0 12/01 03:48:42 : [# 1105 lifetime=18:25:00] S [OK:604700,F:74281,E:0,m:0.000,M:612.382,Avg:0.136] A [OK:61070,F:0,E:0,m:0.000]
1 12/01 03:49:42 : [# 1106 lifetime=18:26:00] S [OK:605056,F:74301,E:0,m:0.000,M:612.382,Avg:0.137] A [OK:61111,F:0,E:0,m:0.000]
2 12/01 03:50:42 : [# 1107 lifetime=18:27:00] S [OK:605398,F:74324,E:0,m:0.000,M:612.382,Avg:0.137] A [OK:61152,F:0,E:0,m:0.000]
3 12/01 03:51:42 : [# 1108 lifetime=18:28:00] S [OK:605896,F:74346,E:0,m:0.000,M:612.382,Avg:0.137] A [OK:61194,F:0,E:0,m:0.000]
...```

```
1410 12/02 03:17:42 : [# 2514 lifetime=41:54:00] S [OK:1341442,F:157428,E:0,m:0.000,M:612.382,Avg:0.126] A [OK:133474,F:0,E:0,m:0.000]
```

```
0 00:00:00.140 - ERROR [xmlrpc-5916] - srmRm: File does not exist
1 00:00:00.144 - INFO [xmlrpc-5916] - srmRm: user </DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614260/CN=Robot:
2 00:00:00.144 - INFO [Timer-4] - ADVANCED PICKER: dispatching 15 requests.
3 00:00:00.447 - INFO [pool-4-thread-33] - srmPrepareToPut: user</DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=61
...```

```
4136257 23:59:56.076 - INFO [GPFSQuotaSubmitter-1] - Submitting GPFS quota info computation for vfs rooted at /storage/gpfs_atlas
```
Data Format

JSON object:

• Mandatory fields for source and name of the data
  ◦ hostname of the client that sent the event
  ◦ name of the data produced (i.e. name of the program that generated the log, or ident)

• Optional payload based on the data type and source
  ◦ Type of data generated (i.e. log, metric...) for routing
  ◦ Log message or value for metrics and sensor
  ◦ Topic for routing the message to the correct output
  ◦ Timestamp of the event generation

• Supports nested field for complex data structure
A BIG DATA PLATFORM FOR HETEROGENEOUS DATA COLLECTION AND ANALYSIS IN LARGE-SCALE DATA CENTERS - ISGC 2021