Predictive@ENI WP4/INFN

Activities and progress

Alessandro Costantini, Antonio Falabella, Elisabetta Ronchieri
INFN-CNAF

alessandro.costantini<at>cnaf.infn.it
elisabetta.ronchieri<at>cnaf.infn.it
antonio.falabella<at>cnaf.infn.it

The Problem to be addressed

- Demonstrate the usability of modern AI-based techniques in the realization of systems for
  - Predictive maintenance
  - Modeling of the interdependencies between systems in complex industrial apparatuses.

WP4 Objectives (M 0-24)

Objectives

- ***Anomalies Detection Approaches for Generated Data***
- ***Ingestion of Heterogeneous data for Anomalies Detection***

Focus on Open-source technologies and solutions

Human resources

- Alessandro Costantini (1 PM)
- Elisabetta Ronchieri (3 PM)
- Antonio Falabella (1 PM)
- Staff (1 PM)
- One year of Assegno di Ricerca (Postdoc)

HW/SW requirements from ICSC:

- INFN: 1 server with 2 A100 equivalent GPUs for 24 months

Milestones

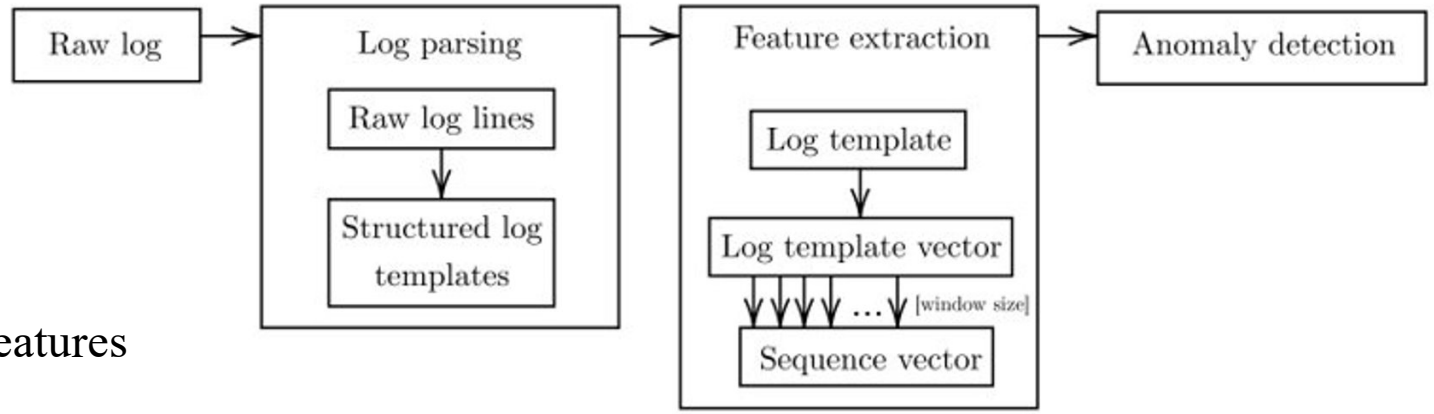
- M1-INFN (**synced with MS8-Jun24**): Selection of the approaches and related algorithms for analysis. PoC of the data ingestion platform to be used for the project purposes
- M2-INFN (**synced with MS10-Aug25**): PoC results and comparison with the different approaches proposed by the partners

Anomalies Detection Approaches for Generated Data

Use case: **Identify anomalies in service log files with Natural Language Processing solutions**

Activities:

- Turn unstructure data into structure data
- Preprocess service messages
- Build log files corpora
- Build a dictionary of anomalies
- Extract the most interesting Ngram-based features
- Label each message in anomalous or not
- Cluster service messages with topic modeling techniques and unsupervised machine learning

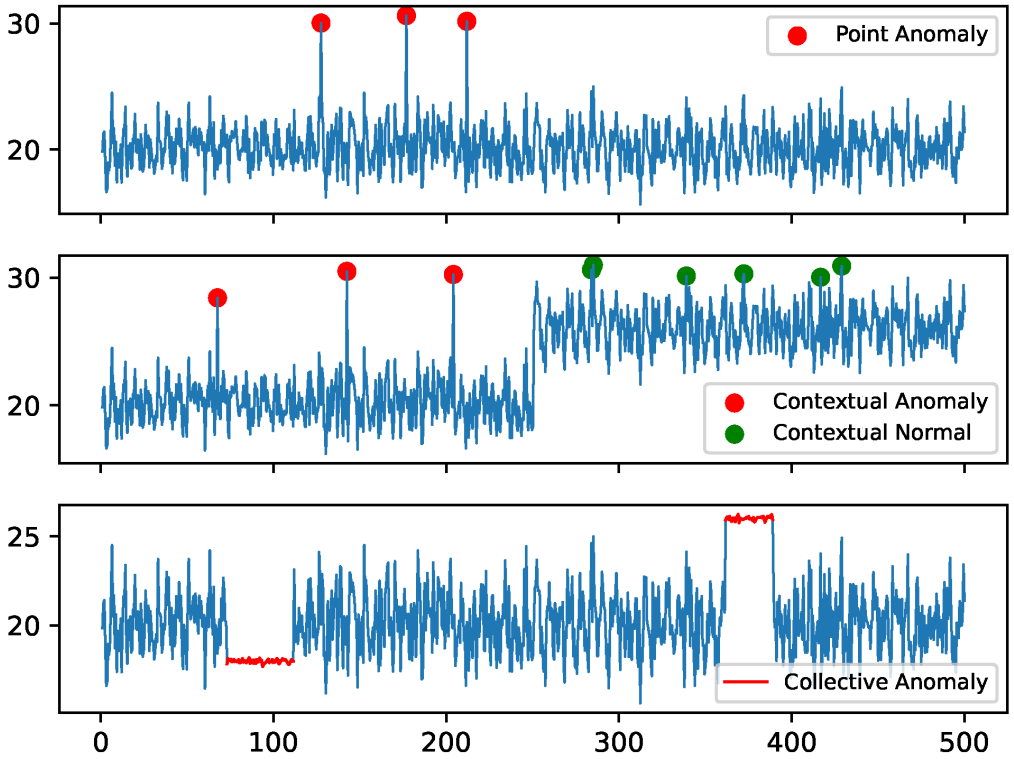


Anomalies Detection Approaches for Generated Data

Use case: **Identify anomalies in monitoring physical machine metric data**

Activities:

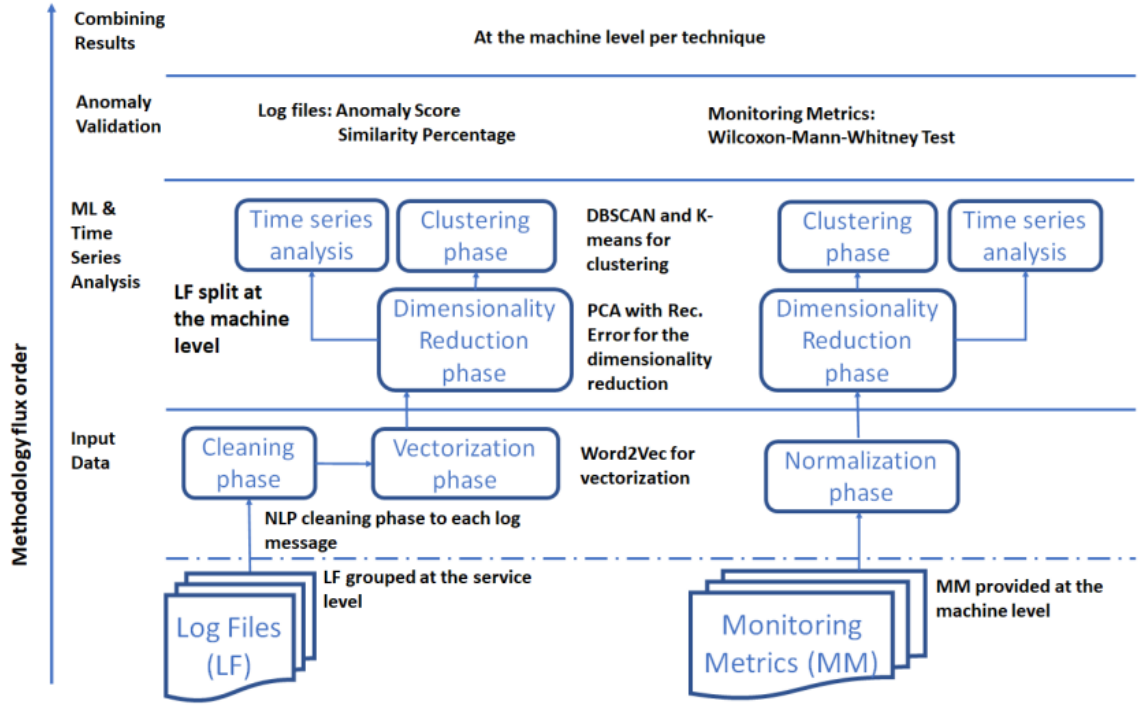
- Explore available data with time series, considering thresholds whenever possible
- Correlate variables to exclude redundant ones
- Identify anomaly slots
- Use Bayesian Optimization approach – a strategy for global optimization of expensive-to-evaluate functions - to predictive maintenance on imbalance data
- Use the JumpStarter solution - a multivariate time series anomaly detection approach – to compute anomaly score and label the various observations



Anomalies Detection Approaches for Generated Data

Use case: Identify anomaly pattern in heterogeneous data covering service log files and machine metrics

Activities:



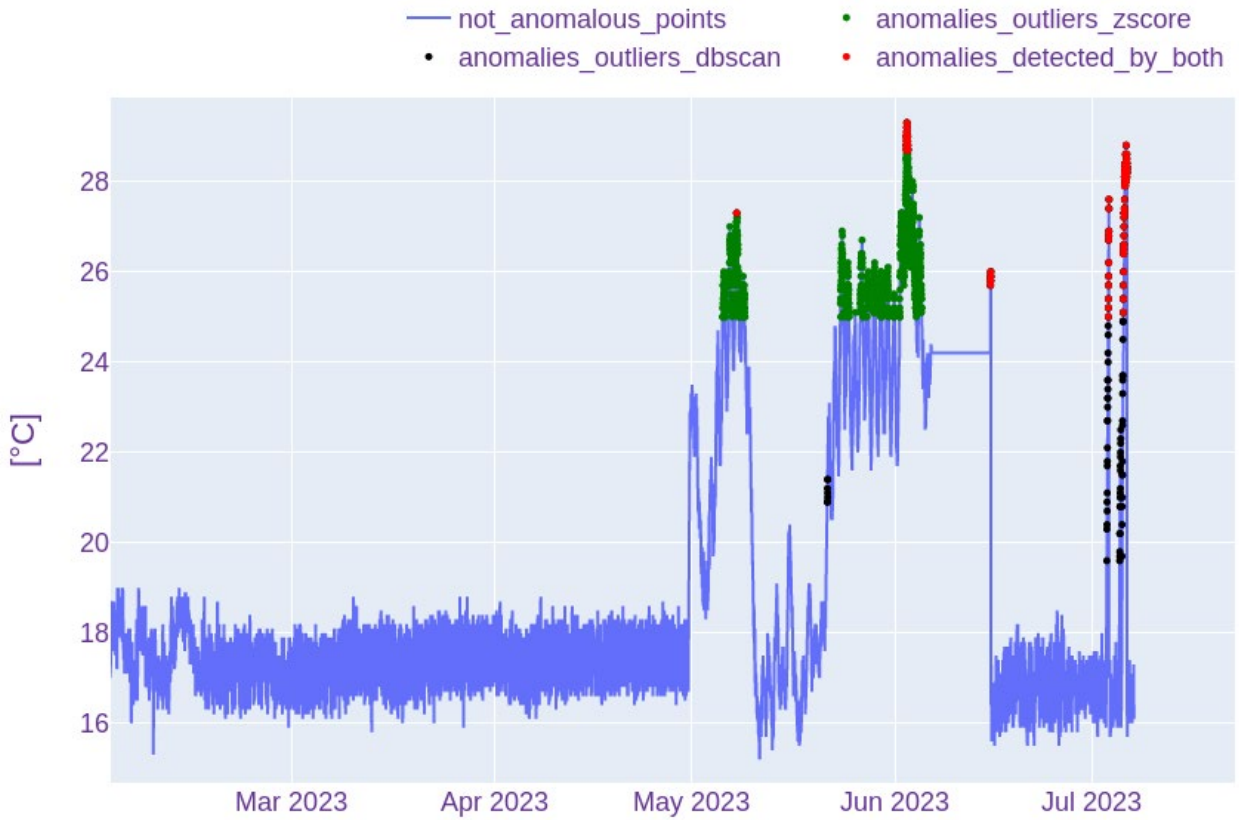
Bottom-up Anomaly Detection Approach with Log Files and Monitoring Metrics.

Anomalies Detection Approaches for Generated Data

Use case: **Identify anomalies in monitoring electrical plant and cooling system data**

Activities:

- Preprocess data
- Reduce data dimension with PCA
- Identify anomalies with DBSCAN
- Identify anomalies with z-score
- Investigate Graph Neural Network with multivariate time series



Ingestion of Heterogeneous data for Anomalies Detection

Use cases

Support different **data producers** → Kafka allows for different tools or customize using libraries in different languages

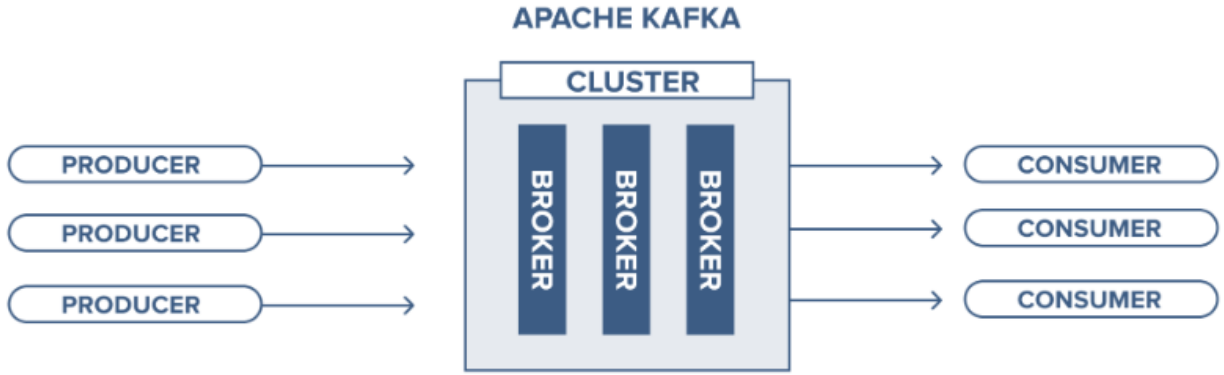
```
from kafka import KafkaProducer
import json
import time
from bson import json_util
import random
import time


producer = KafkaProducer(
     bootstrap_servers=['cnlog-
kafka01.cr.cnaf.infn.it:9192'],
     security_protocol="SASL_SSL",
     sasl_plain_username='test',
     sasl_plain_password='test',
     sasl_mechanism='PLAIN',
     ssl_cafile='./ca.pem',
)
```

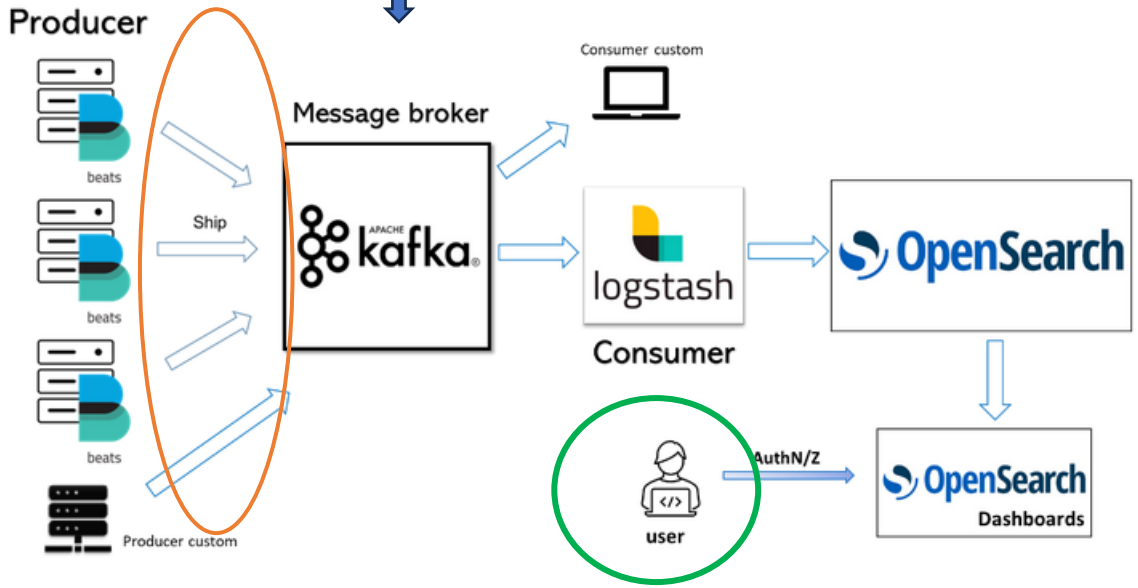
```
topicName = 'testTopic'


while True:
  ts = time.time()
  data = { 'tag ': 'blah',
        'name' : 'python_consumer',
        'index' : 1,
        'number' : random.randint(0,9),
        'timestamp' : ts
        }
  ack = producer.send(topicName, json.dumps(data,
default=json_util.default).encode('utf-8'))
  metadata = ack.get()
  print(data['number'])
  print(data['timestamp'])
  print(metadata)
  time.sleep(3)
```

Ingestion of Heterogeneous data for Anomalies Detection

Use cases

- Kafka allows for **data segregation** : **topic**
  - Replication
  - Retention
  - Partition
- Data from producer to kafka cluster **encrypted using SSL**
- **Specific ACL** for producers and consumers
- **OIDC AutheN/AuthZ** to get access to the data
  - Data consumed by single consumers/consumer groups





Ingestion of Heterogeneous data for Anomalies Detection

Use and adoption

Log analysis using Opensearch

- Log data organized in indexes