# Using ML clustering tools to improve data transfer management operations

L Rinaldi[1], L Clissa[2], L Morganti[3]
[1)]Bologna University and INFN, [2)]INFN-Bologna, [3)]INFN-CNAF

The **GRID** computing paradigms adopted by the main HEP experiments are based on the **distribution** of experimental data on computer resources located all over the world

All the **data transfer** processes are tracked in **log files** produced by the various services involved and such log files represent a source of **information** which is largely **underutilized**

An approach based on **unsupervised ML** techniques is used to automatically process information stored in **log files** with the aim of grouping the error messages and speed up the procedures for **detecting errors** and **solving problems**

## Clustering pipeline

**Step 1: text pre-processing**

- Lowercase transformation and punctuation/stopword stripping
- Tokenization
- URL split

**Step 2: vectorization**

- Transformation of the pre-processed text into numeric information to map each message to a point in a vectorial subspace (embedding)
- *word2vec* language model adopted

**Step 3: clustering**

- k-means++ algorithm: intuitive approach and good performance in a wide range of applications
- The number of clusters $k$ is selected at each clustering stage based on a grid search and geometrical criteria:
  - Within cluster Sum of Squared Errors (elbow method)

$$\mathrm{WSSE}(\mathrm{dist}, k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \mathrm{dist}\left(x_i - \bar{x}_j\right)$$

$x_i$ generic data point
$\bar{x}_j$ centroid of a generic cluster $C_j$

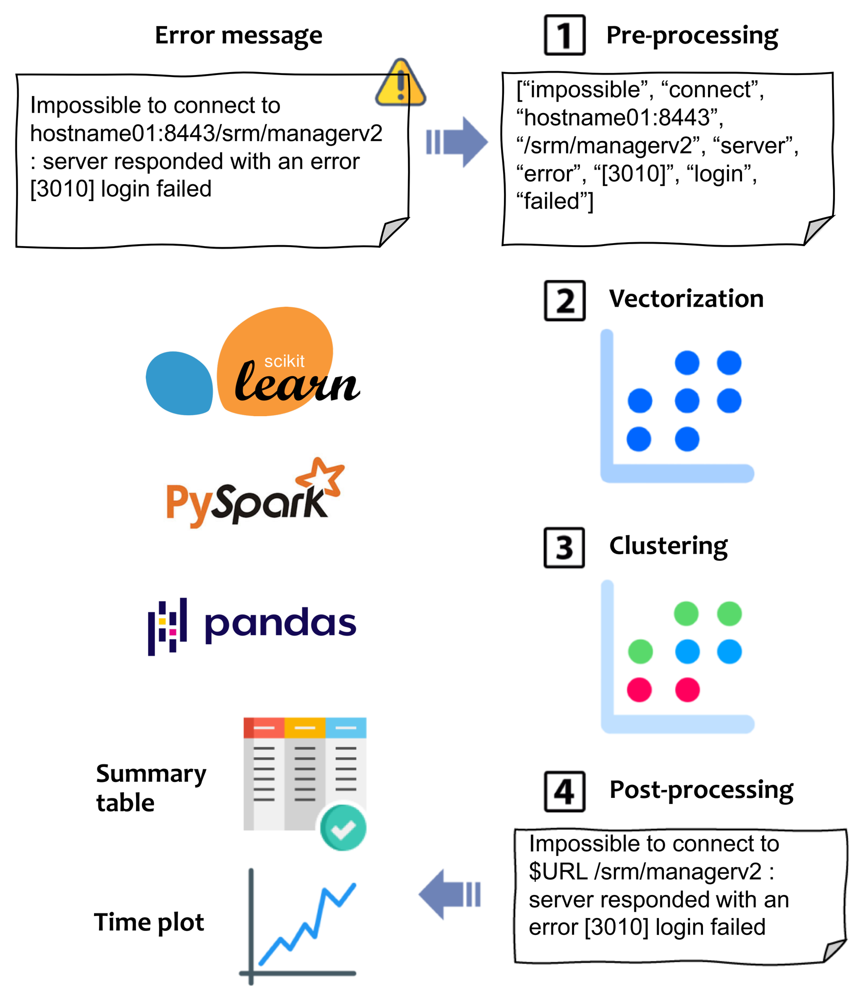
  - Average Silhouette Width (max)

$$\mathrm{ASW}(\mathrm{dist}, k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - \bar{a}_i}{\max\left(\bar{a}_i, b_i\right)}$$

$\bar{a}_i$ distance of $x_i$ from same cluster points
$b_i$ distance of $x_i$ from other clusters points

**Step 4: post-processing**

- Organize results according to the use case




## Analysis of File Transfer Service (FTS) errors

| ID | cluster size | # strings | # patterns | Top 3 message | n | % | source rcsite | destination rcsite |
|---|---|---|---|---|---|---|---|---|
| 0 | 819465 | 117 | 14 | destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not verify credential globus_gsi_callback_module the certificate has been revoked serial number = 1 (0xffffffffffff | 85545 | 10.44% | Site-1 | Site-4 |
| | | | | destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not verify credential globus_gsi_callback_module the certificate has been revoked serial number = 1 (0xffffffffffff | 84453 | 10.31% | Site-2 | Site-4 |
| | | | | destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not verify credential globus_gsi_callback_module the certificate has been revoked serial number = 1 (0xffffffffffff | 77410 | 9.45% | Site-3 | Site-4 |
| 6 | 9673 | 347 | 60 | source srm_get_turl srm-ifce err connection timed out err [se][statusofgetrequest][etimedout] \$URL /srm/managerv2 user timeout over | 1838 | 19.00% | Site-22 | Site-46 |
| | | | | transfer globus_ftp_client the server responded with an error 421 service busy connection limit exceeded please try again later closing control connection | 522 | 5.40% | Site-33 | Site-47 |
| | | | | transfer globus_ftp_client the server responded with an error 421 service busy connection limit exceeded please try again later closing control connection | 300 | 3.10% | Site-29 | Site-47 |
| 3 | 34183 | 1568 | 1537 | error reported from srm_ifce 2 [se][ls][srm_invalid_path] no such file or directory | 13118 | 38.38% | Site-12 | Site-35 |
| | | | | error reported from srm_ifce 2 [se][ls][srm_invalid_path] no such file or directory | 9333 | 27.30% | Site-12 | Site-17 |
| | | | | error reported from srm_ifce 2 [se][ls][srm_invalid_path] no such file or directory | 1707 | 4.99% | Site-12 | Site-22 |

**Cluster Description** (tabular format)

Analysis of the error messages in FTS log files (1 day)

The three most frequent triplets of <pattern>-<source>-<destination> reported in descending order for each cluster

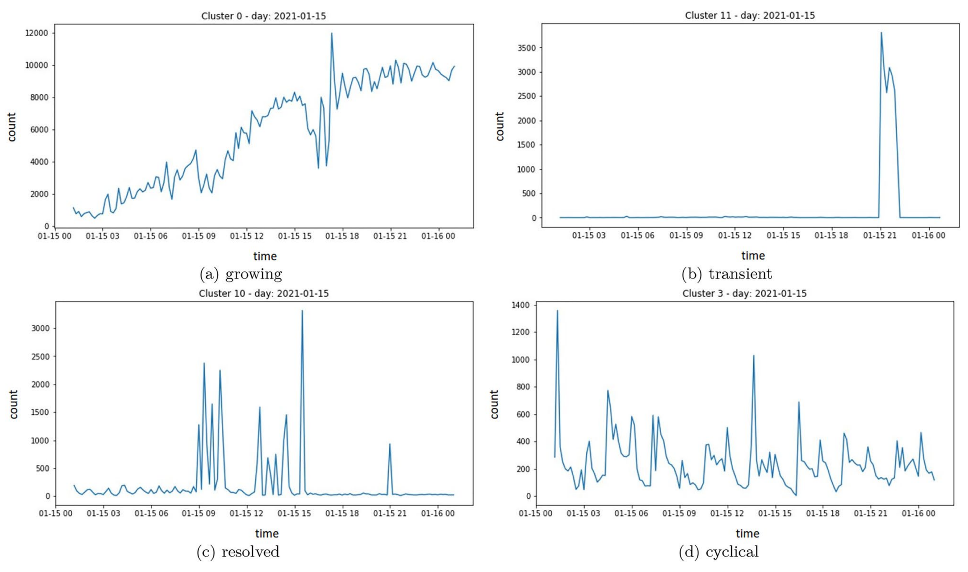
Precious insights for spotting:

- what type of errors and where they occur
- amount of errors, both absolute and relative to the error group

### Time Series

Temporal trend of the number of errors generated by each cluster

- escalating or cyclical failures → require immediate actions
- transient or in resolution




| N. clusters | ASW | WSSE | Perfect match | Fuzzy match | Partial match | False positives | False negatives |
|---|---|---|---|---|---|---|---|
| 15 | 0.89 | 17107 | 7 | 3 | 2 | 3 | 1 |

Summary of the cross-check between clusters and incidents reported in GGUS. Most of the groups discovered are linked to reported issues, with only 3 false positives and 1 false negative

Extensive testing using incidents reported in **GGUS** as a benchmark (17 days window)

- overlapping between discovered clusters and the reported issues

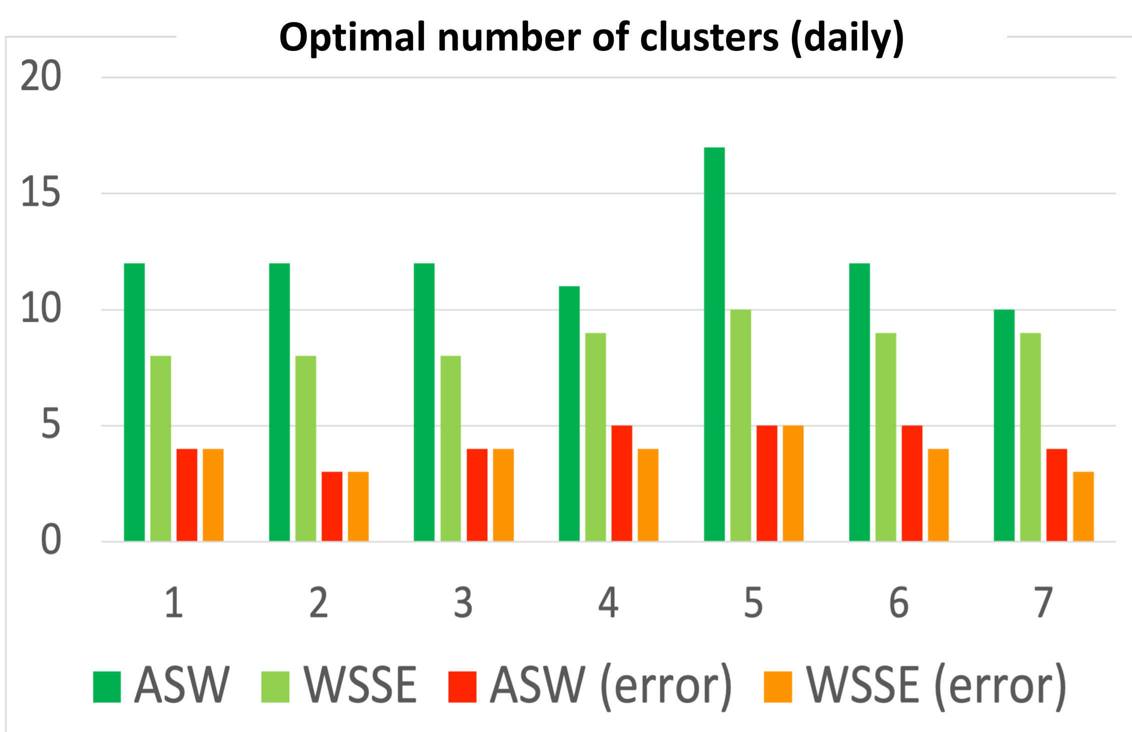
*Computing and Software for Big Science (2022) 6:16* https://doi.org/10.1007/s41781-022-00089-z

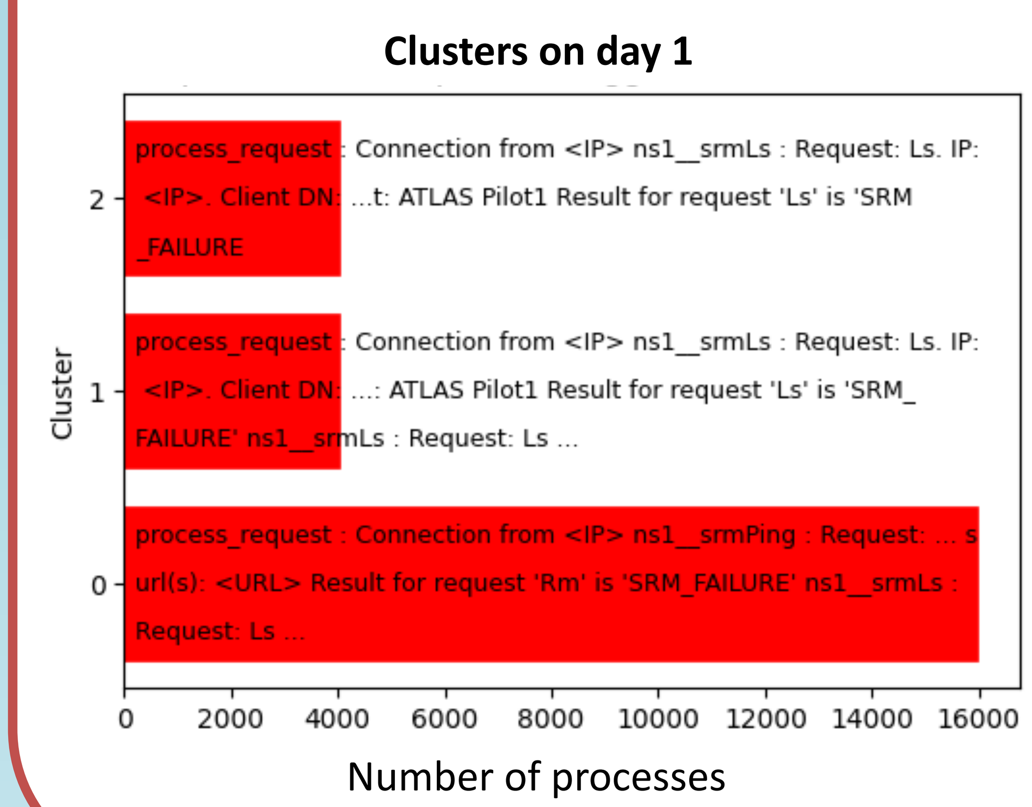
## Analysis of StoRM log files

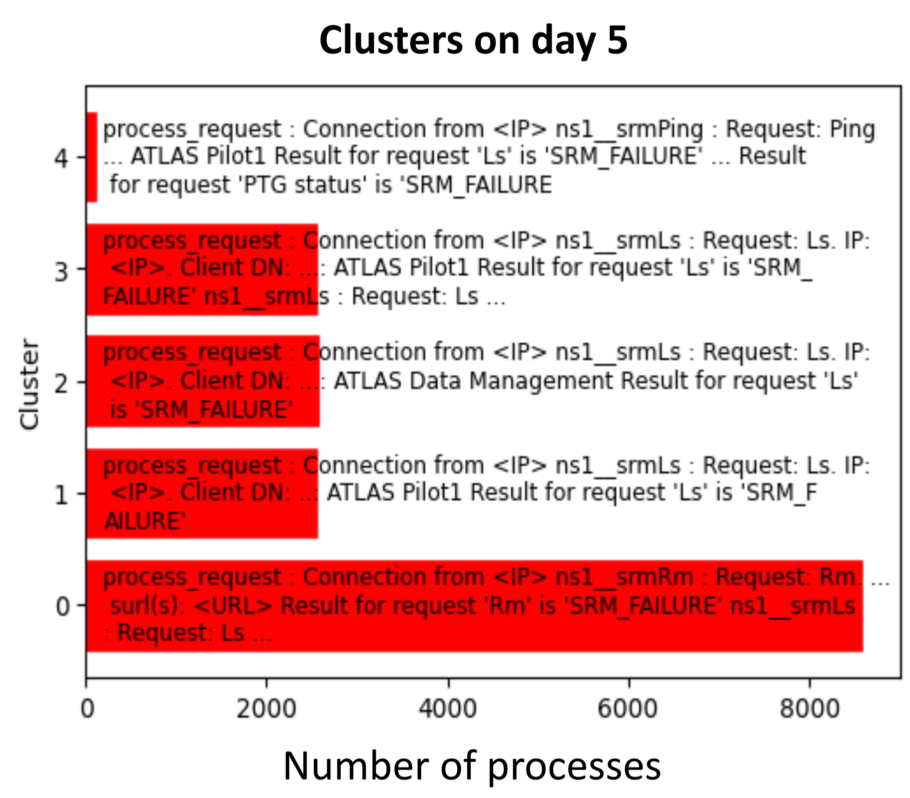
Analysis performed on a 7 days interval Dataset (daily log files) from INFN-CNAF StoRM system

Messages of a single log file are grouped by StoRM process (separately for explicit *error* messages) and then passed to the clustering pipeline

Higher values of the optimal number of clusters may give a hint of an anomaly (day 5 in this example)




The variation of the daily number of clusters and clustered error messages could give more information on the cause of the problem

*CHEP2023, 8th-12th May 2023, Norfolk, USA*
Contribution ID: 468 - Contacts: lorenzo.rinaldi@unibo.it