University of Basel

Data Analysis for Improving High Performance Computing Operations and Research

A EUCOR Seed Money Project

Aurélien Cavelan, Florina M. Ciorba, October 3 2019

Challenges

The goal is to improve the research and operations activities of NEMO at University of Freiburg, sciCORE at University of Basel and at University of Strasbourg.

- Collect HPC logs
- Ensure that the data follows the FAIR (findable, accessible, interoperable, and reusable) data principles
- Legal compliance with EU, CH, G, F data and privacy protection laws
- Data analytics

The outcome will be solutions for improving the HPC operations and research of three Eucor HPC centers, and satisfy the data protection and privacy requirements.

Monitoring

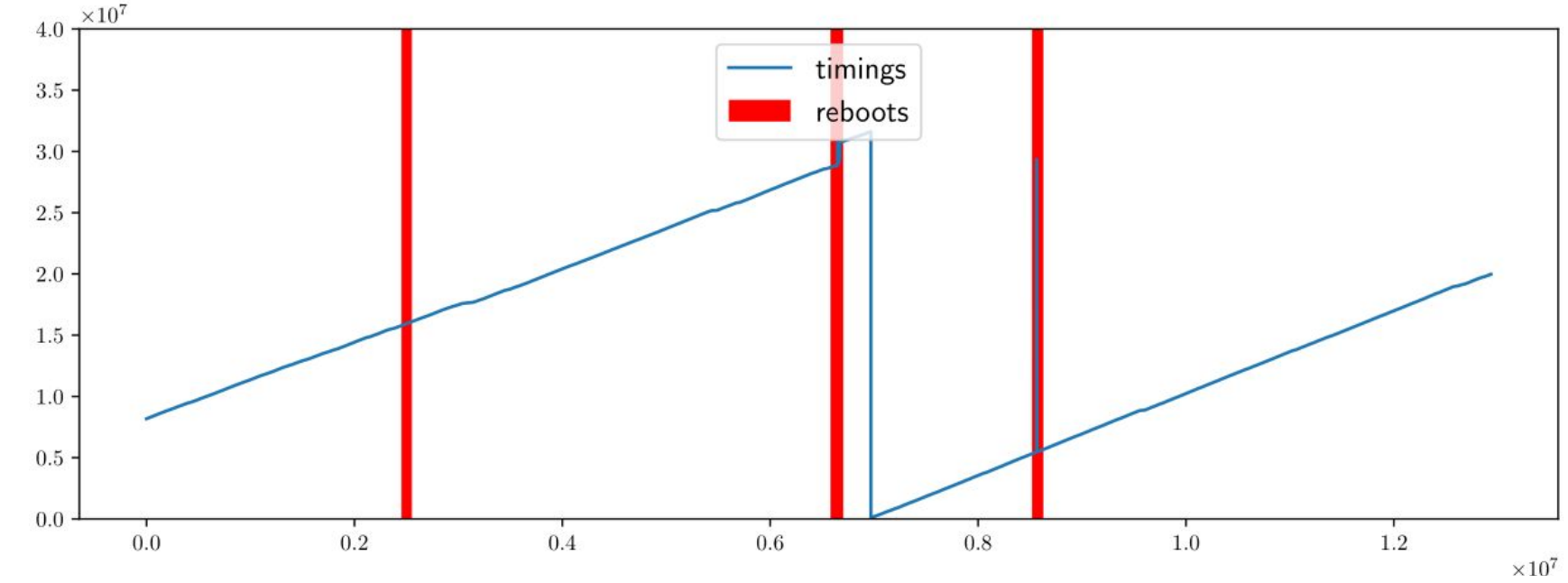
- At **sciCORE**: monitoring since April 2018
- 500GB of (compressed) data
- Estimated > 10TB uncompressed data

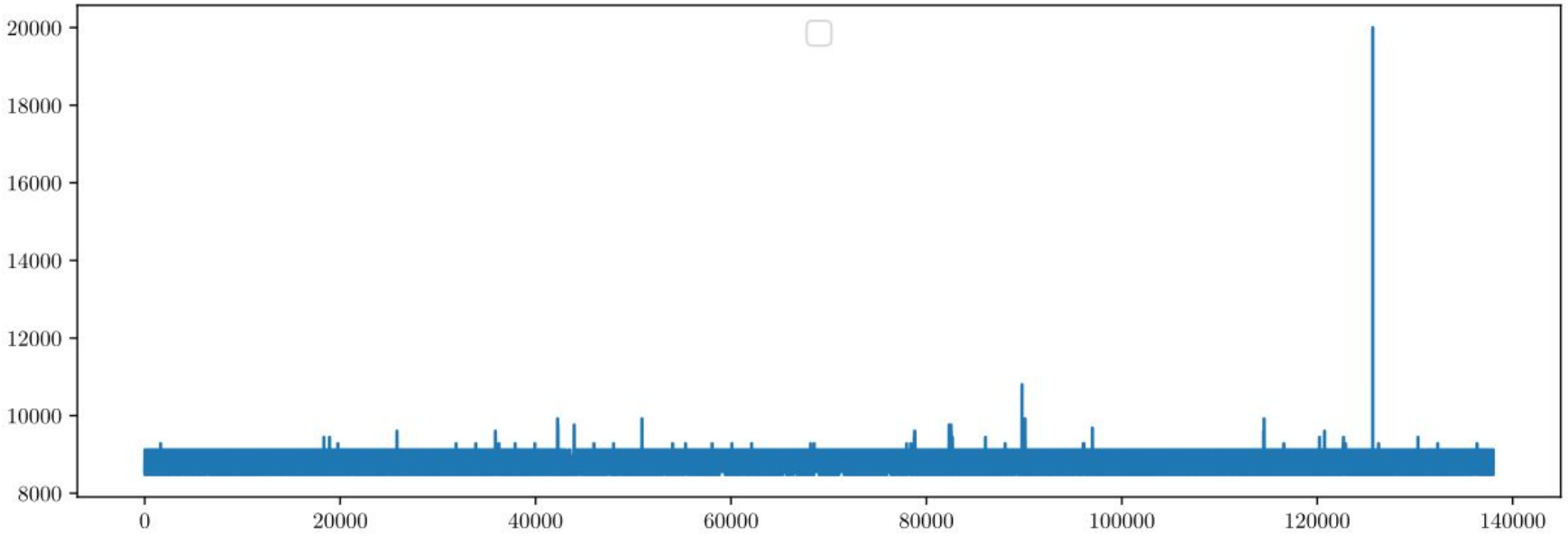
```
[cavelan@login10 logstore]$ ls
bc2-gateway04  gpfs-monitor    lii07  lii28     messages-20180408  messages-20180903  sgi22  shi13  shi34  shi55              usi03                  usi115  usi136  usi57
bc2-gateway05  Lenovo_Support  lii08  lmi01     messages-20180415  messages-20180909  sgi23  shi14  shi35  shi56              usi04                  usi116  usi137  usi58
bc2-gateway06  lhi01           lii09  lmi02     messages-20180423  messages-20180917  sgi24  shi15  shi36  shi57              usi05                  usi117  usi138  usi59
bc2-linux10    lhi02           lii10  lmi03     messages-20180429  messages-20180923  sgi25  shi16  shi37  shi58              usi06                  usi118  usi139  usi60
bc2-linux11    lhi03           lii11  lmi04     messages-20180507  messages-20180930  sgi26  shi17  shi38  shi59              usi07                  usi119  usi140  usi91
bc2-tsgpfs03   lhi04           lii12  lmi05     messages-20180513  messages-20181008  sgi27  shi18  shi39  shi60              usi08                  usi120  usi41   worker01
bc2-tsgpfs04   lhi05           lii13  lmi06     messages-20180521  messages-20181014  sgi28  shi19  shi40  shi61              usi101                 usi121  usi42   worker02
ems01          lhi06           lii14  lmi07     messages-20180527  messages-20181022  sgi29  shi20  shi41  shi62              usi102                 usi122  usi43   worker03
ems01x         lhi07           lii15  lmi08     messages-20180604  messages-20181028  sgi30  shi21  shi42  shi63              usi103                 usi123  usi44   worker04
ess01          lhi08           lii16  login10   messages-20180610  messages-20181105  shi01  shi22  shi43  shi64              usi104                 usi124  usi45
ess01x         lhi09           lii17  login11   messages-20180617  messages-20181111  shi02  shi23  shi44  shi65              usi105                 usi125  usi46
ess02          lhi10           lii18  login12   messages-20180625  messages-20181118  shi03  shi24  shi45  shi66              usi106                 usi126  usi47
ess03          lhi11           lii19  login13   messages-20180701  messages-20181126  shi04  shi25  shi46  shi67              usi107                 usi127  usi48
ess04          lhi12           lii20  login14   messages-20180709  messages-20181202  shi05  shi26  shi47  shi68              usi108                 usi128  usi49
ess05          lhi13           lii21  login16   messages-20180715  messages-20181210  shi06  shi27  shi48  shi71              usi108.cluster.bc2.ch  usi129  usi50
ess06          lhi14           lii22  login17   messages-20180722  messages-20181216  shi07  shi28  shi49  shi72              usi109                 usi130  usi51
ess07x         lii02           lii23  login18   messages-20180730  rhsm               shi08  shi29  shi50  smi01              usi110                 usi131  usi52
ess08x         lii03           lii24  login19   messages-20180805  sbi01              shi09  shi30  shi51  syslog01           usi111                 usi132  usi53
gateway01      lii04           lii25  login20   messages-20180812  service06          shi10  shi31  shi52  system-login11.log  usi112                 usi133  usi54
gateway02      lii05           lii26  mariadb   messages-20180820  sgi01              shi11  shi32  shi53  usi01              usi113                 usi134  usi55
gateway09      lii06           lii27  messages  messages-20180826  sgi21              shi12  shi33  shi54  usi02              usi114                 usi135  usi56
[cavelan@login10 logstore]$
```

Monitoring HPC Logs

- Data is collected on every node
- Periodically copied to a dedicated storage
- System logs (connection attempts, commands, deamons, services…)
- Sensor data (temperature, fan speed, CPU frequency, memory errors and many more depending on availability on the node)
- **Personal data** (name, access times, location, research, …)

```
Apr  4 16:45:56 login10 sshd[1597]: pam_sss(sshd:auth): authentication success; logname= uid=0 euid=0 tt
Apr  4 16:45:56 login10 sshd[1597]: Accepted password for eleliemy from 131.152.54.236 port 53726 ssh2
Apr  4 16:45:56 login10 sshd[1597]: pam_unix(sshd:session): session opened for user eleliemy by (uid=0)
Apr  4 16:54:39 login10 sshd[1597]: pam_unix(sshd:session): session closed for user eleliemy
```


×10^7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
timings
reboots
0.0
0.2
0.4
0.6
0.8
1.0
1.2
×10^7


20000
18000
16000
14000
12000
10000
8000
0
20000
40000
60000
80000
100000
120000
140000

FAIR Data

«Ensuring that the HPC monitoring data follows the findable, accessible, interoperable, and reusable (FAIR) data principles.»

Main challenges DA-HPC-OR:

- **Ensuring access to the data** by the project members (findable, accessible)
- **Achieving meaningful integration of the various types and format** (interoperable and reusable)
- Project members in three countries (Germany, Switzerland, France)
- Requires **legal compliance**

Which data protection law is applicable?

- **Applicability of the Swiss Cantonal Data Protection Law = Information and Data Protection Act of the Canton Basel-Stadt (IDG – Informations-und Datenschutzgesetz des Kantons Basel-Stadt)**

- Swiss legislation shall be compliant to EU-legislation: **GDPR** must be taken into account

When data protection law is applicable?

First principle (in EU, CH, G and F): The requirements of data protection regulation (G, CH and F) are taken into account only if the activity (-ies) in question concerns personal data.

➔ How to determine personal data?
➔ *Are logfiles of the HPC-project personal data?*

Personal data: Any information relating to an identified or identifiable person -> 3 criteria

```
Apr  4 16:45:56 login10 sshd[1597]: pam_sss(sshd:auth): authentication success; logname= uid=0 euid=0 tt
Apr  4 16:45:56 login10 sshd[1597]: Accepted password for eleliemy from 131.152.54.236 port 53726 ssh2
Apr  4 16:45:56 login10 sshd[1597]: pam_unix(sshd:session): session opened for user eleliemy by (uid=0)
Apr  4 16:54:39 login10 sshd[1597]: pam_unix(sshd:session): session closed for user eleliemy
```

Data protection law (EU, CH, G, F) provides facilitated conditions for the processing of personal data for scientific research purposes

Requirements:

- **Processing of personal data not related to specific persons:** knowledge/information from data does not specifically target one particular individual but refers to information from data as a whole

Make sure that monitoring of systemlogs, as long as it refers to personal data, targets only the purpose of analyzing and understanding the HPC systems (no permission to use personal data that will be discovered by the monitoring for different purposes).

- **Planned scientific research with a purpose** performed with scientific methods

Knowledge gained by monitoring and analyzing the systemlogs must be “new” and must be based on scientific methods.

- **Processing is in the context of public duties**

University has the duty to conduct research?

Which data protection requirements will be facilitated if data processing is for research purposes?

- **Alleviates purpose limitation:**

Allows further processing of existing personal data for the research

- **Alleviates consent**

In Basel-Stadt public bodies may process personal data **without consent** if the processing refers to a performance of a public duty. University = public duty to do research?

GDPR: consent will be basically needed even for scientific purposes but facilitated consent if it is not possible to fully identify the purpose of personal data processing for scientific research **at the time of data collection**

- **Alleviates storage duration**

Data needs to be stored for the purposes of scientific research.

GDPR: allows to prolong the storage exceeding the appropriate duration of processing individual data

- **Processing of sensitive personal data**

Swiss law: requires a strong explicit permission

GDPR: The basic prohibition of processing sensitive data provides an exception for scientific research

Core Problem (in EU, CH, D,F): duty to anonymize or pseudonymize personal data as soon as it is permitted by the research purpose.

- In Basel-Stadt: **anonymization from the beginning on is required**

At least pseudonymisation is necessary if the research purpose can be achieved, or prove that the research purpose cannot be achieved otherwise

GDPR: anonymization is required as soon as the research or statistical purpose allows, unless this conflicts with legitimate interests of the data subject.

- **No use of personal data for a person related purpose**

Once processed the personal data for research purposes it is not allowed to process these data again for a particular purpose relating to an individual person or transfer these data to another public body for allowing them to use this data for an particular purpose relating to an individual person.

- **Publishing of research results**

Basel-Stadt (Swiss cantonal law): publishing of results in a manner that identification of (the) person(s) is not possible

GDPR: publishing of results only if the person concerned gives her consent or if doing so is indispensable for the presentation of research findings on contemporary events.

De-identification and anonymization

[Box 2] Pseudonymisation and anonymisation: understanding the difference

Pseudonymisation entails substituting personally identifiable information (such as an individual's name) with a unique identifier that is not connected to their real-world identity, using techniques such as coding or hashing. However, if it is possible to re-identify the individual data subjects by reversing the pseudonymisation process, data protection obligations still apply. They cease to apply only when the data are fully and irreversibly anonymised.

Anonymisation involves techniques that can be used to convert personal data into anonymised data. Anonymisation is increasingly challenging because of the potential for re-identification.

Re-identification is the process of turning pseudonymised or anonymised data back into personal data by means of data matching or similar techniques.

De-identification and anonymization

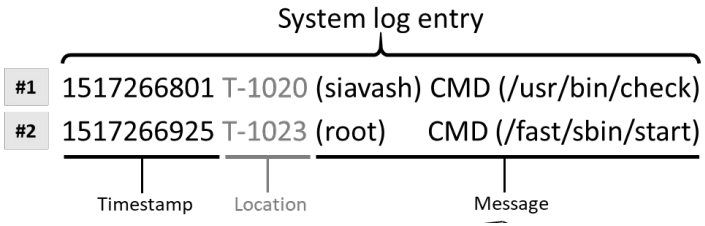
Example 4: Five most frequent event patterns and their frequency

```
43% (#USR_#) CMD (#PATH#)
 9% starting #DAEM#
 9% finished #DAEM#
 5% Received disconnect from #IPv4# disconnected by user
 5% pam_unix(sshd:session): session closed for user #USR_#
```

50 most frequent log event patterns derived from **90+ %** of all syslog entries.

[Adapted from: Opinion 05/2014 on Anonymization Techniques, Art. 29, Data Protection Working Party, 0829/14/EN, WP216.]
[Source: Assessing Data Usefulness for Failure Analysis in Anonymized System Logs, S. Ghiasvand, F. Ciorba, ISPDC 2018]

De-identification and anonymization



#	Message	Hash key	Category
1	(siavash) CMD (/usr/bin/check >/dev/null 2>&1)	a8848910	66dc2742
2	(florina) CMD (/usr/lib32/lm/lm1 1 1)	10a31145	66dc2742
3	(siavash) CMD (run-parts /etc/cron.hourly)	a6a420a6	66dc2742
4	starting 0anacron	47c6b01d	dd740712
5	Anacron started on 2018-01-30	bd94c195	e5a59462
6	Jobs will be executed sequentially	f1e7eac3	f1e7eac3
7	Normal exit (0 jobs run)	e46c1bdb	eac7924f
8	finished 0anacron	76690e70	a5803a8a
9	(siavash) CMD (/usr/lib32/lm/lm1 1 1)	bacc6097	66dc2742
10	(root) CMD (/usr/lib32/cl/cl2 1 1)	eefabc01	66dc2742
11	(root) CMD (/usr/lib64/lm/lm1 1 1)	4237ce2c	66dc2742
12	(siavash) CMD (/usr/bin/check >/dev/null 2>&1)	a8848910	66dc2742
13	(florina) CMD (/usr/bin/run >/dev/null 2>&1)	8470df87	66dc2742
14	(siavash) CMD (/usr/bin/exec >/dev/null 2>&1)	dd0e4a50	66dc2742
15	(siavash) CMD (run-parts /etc/cron.hourly)	a6a420a6	66dc2742
16	starting 0anacron	47c6b01d	dd740712
17	Anacron started on 2018-01-31	d414932d	e5a59462
18	Jobs will be executed sequentially	f1e7eac3	f1e7eac3
19	Normal exit (4 jobs run)	0c3b639c	eac7924f
20	finished 0anacron	76690e70	a5803a8a

[Adapted from: Opinion 05/2014 on Anonymization Techniques, Art. 29, Data Protection Working Party, 0829/14/EN, WP216.]
[Source: Assessing Data Usefulness for Failure Analysis in Anonymized System Logs, S. Ghiasvand, F. Ciorba, ISPDC 2018]

De-identification and anonymization

#	Message	Hash key	Category
1	(siavash) CMD (/usr/bin/check >/dev/null 2>&1)	a8848910	66dc2742
2	(florina) CMD (/usr/lib32/lm/lm1 1 1)	10a31145	66dc2742
3	(siavash) CMD (run-parts /etc/cron.hourly)	a6a420a6	66dc2742
4	starting 0anacron	47c6b01d	dd740712
5	Anacron started on 2018-01-30	bd94c195	e5a59462
6	Jobs will be executed sequentially	f1e7eac3	f1e7eac3
7	Normal exit (0 jobs run)	e46c1bdb	eac7924f
8	finished 0anacron	76690e70	a5803a8a
9	(siavash) CMD (/usr/lib32/lm/lm1 1 1)	bacc6097	66dc2742
10	(root) CMD (/usr/lib32/cl/cl2 1 1)	eefabc01	66dc2742
11	(root) CMD (/usr/lib64/lm/lm1 1 1)	4237ce2c	66dc2742
12	(siavash) CMD (/usr/bin/check >/dev/null 2>&1)	a8848910	66dc2742
13	(florina) CMD (/usr/bin/run >/dev/null 2>&1)	8470df87	66dc2742
14	(siavash) CMD (/usr/bin/exec >/dev/null 2>&1)	dd0e4a50	66dc2742
15	(siavash) CMD (run-parts /etc/cron.hourly)	a6a420a6	66dc2742
16	starting 0anacron	47c6b01d	dd740712
17	Anacron started on 2018-01-31	d414932d	e5a59462
18	Jobs will be executed sequentially	f1e7eac3	f1e7eac3
19	Normal exit (4 jobs run)	0c3b639c	eac7924f
20	finished 0anacron	76690e70	a5803a8a

[Adapted from: Opinion 05/2014 on Anonymization Techniques, Art. 29, Data Protection Working Party, 0829/14/EN, WP216.]
[Source: Assessing Data Usefulness for Failure Analysis in Anonymized System Logs, S. Ghiasvand, F. Ciorba, ISPDC 2018]

Summary and Next Steps

Monitoring

- Extract and classify the data (+**interoperability, +reusability**)
- Ensure that the data is FAIR (+**findable and +accessible**)
- Comply with CH, G, F, EU data and privacy protection lwas

De-identify and anonymize the data

- Required by law (with some exceptions)

Next: FAIRU Data Analysis

- Predict failures
- Identify misuse of the system