Joint Research Centre (JRC)

EOS as storage back-end for JRC data science

Armin Burger
Franck Eyraud
Pier Valerio Tognoli
Marco Scavazzon

European Commission - Joint Research Centre
Directorate I Competences, Unit I.3 Text and Data Mining

Joint Research Centre

EOS Workshop, CERN, 3-5 Feb 2020

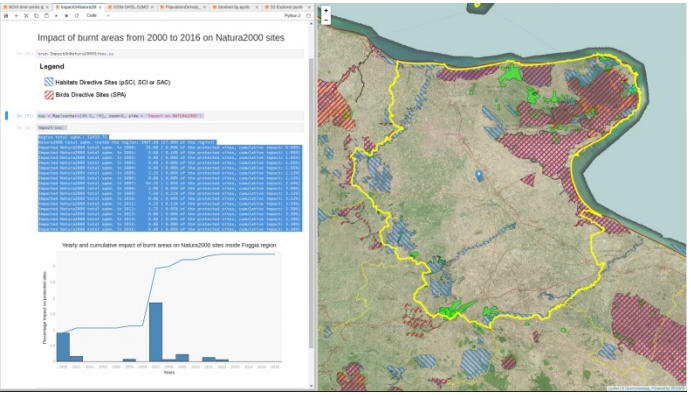
JRC Earth Observation Data and Processing Platform (JEODPP)

JEODPP
The JRC Big Data Platform

Versatile platform bringing the users to the data

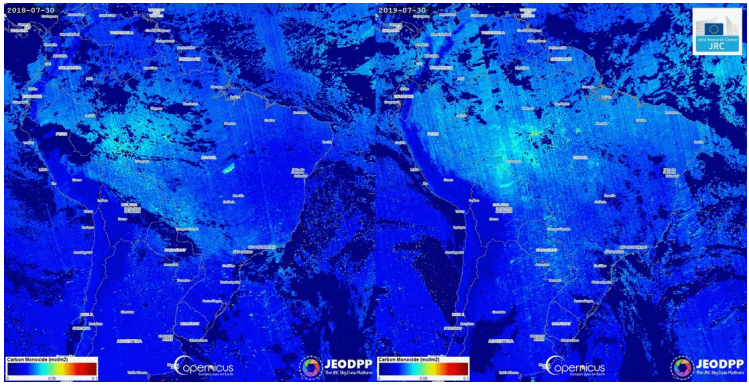
Examples of exploratory analyses and interactive visualization

JEODPP
The JRC Big Data Platform



Impact of forest fires on Natura 2000 sites

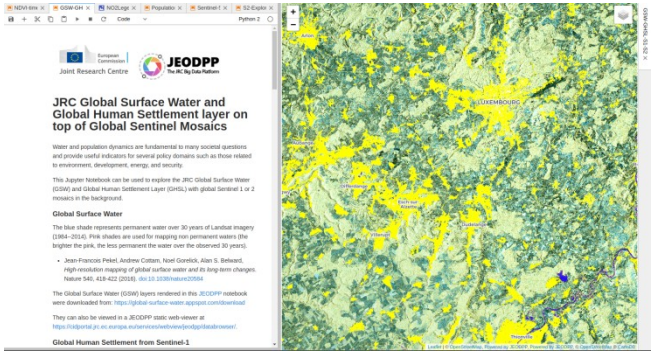
Deforestation (time lapse)



CO emissions (time lapse 08/18 vs 08/19)

Ship traffic (heat maps)

Deforestation (split map 2000 vs 2017)



Interactive combination of information layers

Joint Research Centre

European Commission

EOS instance at JRC – set-up

- In production since mid 2016,
  set up with support by CERN EOS team
- Hardware configuration:
  - *1 or 2 JBOD's per FST*
  - *JBOD's of 24x6 TB and 24x10 TB disks*
  - *2x10 Gbps Ethernet interconnection per FST*
- Software:
  - *FST/MGM with CentOS 7*
  - *Citrine release, v4.5.15/17*
  - *QuarkDB 4.1*

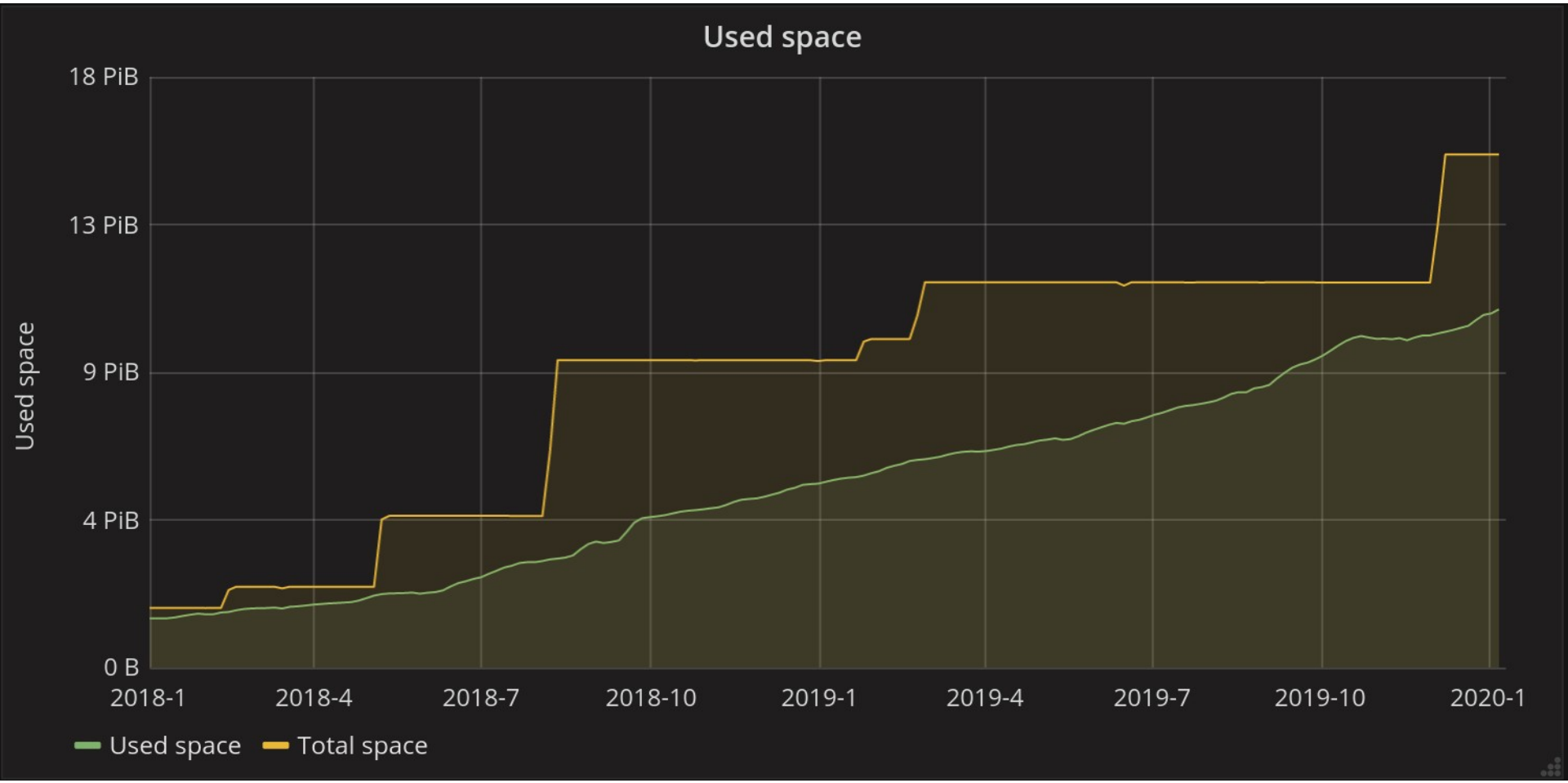
EOS instance at JRC – set-up [2]

- Using replica 2
- 15.5 PiB gross capacity deployed
  → currently space used: 75%
  → 850M files & 175M directories
- Kerberos for authentication
- Using FUSE & FUSEX client for file access

EOS data access

- Direct access (r/w) via FUSE(X) mounts from
  - *Batch processing environments*
  - *Web-based remote desktops (guacamole)*
  - *Jupyter notebook environments*
- CIFS read-only access via SAMBA gateway server from desktop PC's
- Some EOS group areas linked to NextCloud via “external storage” functionality
- FTPS via gateway FTP server for selected projects

EOS storage evolution last 2 years



Changes and incidents in 2019

- Migration of namespace to QuarkDB
- Extended use of FUSEX client
- Reached the 28 bits limit for total number of created directories
- QuarkDB node hardware failure

Migration of namespace to QuarkDB

- To anticipate memory exhaustion
- ~1 day downtime (included FST version upgrade)
- Few config issues in the process, quickly fixed
- Start-up is now immediate (1s)
- Memory consumption controlled
- Browsing lots of files is a bit slower than before, but got improved
- Slower to detect if files are existing or not
  => mainly GDAL library making lots of ***`stat`*** *requests*

Extended use of FUSEX client

- Moved many services and nodes to use FUSEX
- Better OS integration
- Stable in normal use and controlled environments
- Drawbacks:
  - *Directory listing limited to 32k entries*
  - *Tend to crash slightly more often than FUSE, especially in uncontrolled environments (user terminal access, processing nodes, ...)*

28 bits limit for directory counter

- Due to protocol limitation in FUSEX communication
- Could not use directory with ID>0xfffffff in FUSEX, so no new directory was usable
- Means EOS instance could not create more than 268M directories during its lifetime
- Limit was known, code was ready, a hot fix was provided to apply the change in the MGM & FST's

  → v4.5.15+ and **`EOS_USE_NEW_INODES=1`**

QuarkDB node hardware failure

- RAID controller failed, QDB went on working
- Temporary crash when node went back from maintenance, few minutes downtime
- Bugfix and some *fsync* mechanism added in QuarkDB 4.1 to address such issues

EOS – limitations in a few use cases

- Cannot be used for MPI shared logging due to caching issues
  => requires additional storage type
- Jupyter notebooks: user home environment requires locking for sqlite files
- Arbitrary failures writing files via FTPS gateway
- CIFS access via SAMBA not supporting secondary group membership for ACL's
- Long-running processes (several days) can be interrupted due to crash of FUSE(X) client

EOS instance at JRC - some issues

- FUSE and FUSEX client
  - *Still irregularly file read failures, difficult to reproduce*
  - *Some remaining unexplained client crashes; some stack traces have been provided to EOS development team*
  - *Mixed FUSE/FUSEX cooperation not fully reliable*
    → files are not always up to date
  - *Looks like upgrade to v4.5.15+ improved situation*
    e.g. only 6 jobs out of 30k failed due to i/o errors
- FSCK stat disabled due to too many files
  → waiting for new version
- Balancing also gives problem with too many files
  → MGM blocks when new balancing is needed

EOS – the strong points

- Virtually unlimited data size and namespace entries (thanks to QDB back-end)
- Usage of heterogeneous commodity hardware
- Easy to expand capacity
- Robust architecture that handles well incidents
- Cluster processing performance
- Support from developers

Wish-list for features/improvements

- Waiting new FSCK in stable branch
- Better FUSE/FUSEX collaboration
- Increased stability of FUSEX client (might already be the case now)
- Increase 32k directory listing limit
- Exporter for Prometheus monitoring
- Samba gateway functionality also supporting ACL's related to secondary user groups
- Extended ACL tags, like `!r !w !x`

EOS evolution and plans for 2020

- Extension of capacity
  - *additional 2 PB gross in preparation*
  - *move to 60 disk JBOD's*
    → more cost-efficient
    → saving rack space
- Extending EOS test instance for more in-depth testing of functionality and usage
- Possible test of CERNBox

Future Generation Computer Systems
Volume 81, April 2018, Pages 30-40

A versatile data-intensive computing platform for information retrieval from big geospatial data

P. Soille, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, V. Vasilev

Show more

https://doi.org/10.1016/j.future.2017.11.007

Get rights and content

Thank you for your attention!

Big Data Analytics project
Unit I.3 Text and Data Mining Unit
Directorate I Competences

Joint Research Centre

European Commission