CERN Analysis Preservation Status Update

Tibor Šimko
CERN IT

9th LHCb Computing Workshop · 15–19 May 2017

CERN Analysis Preservation

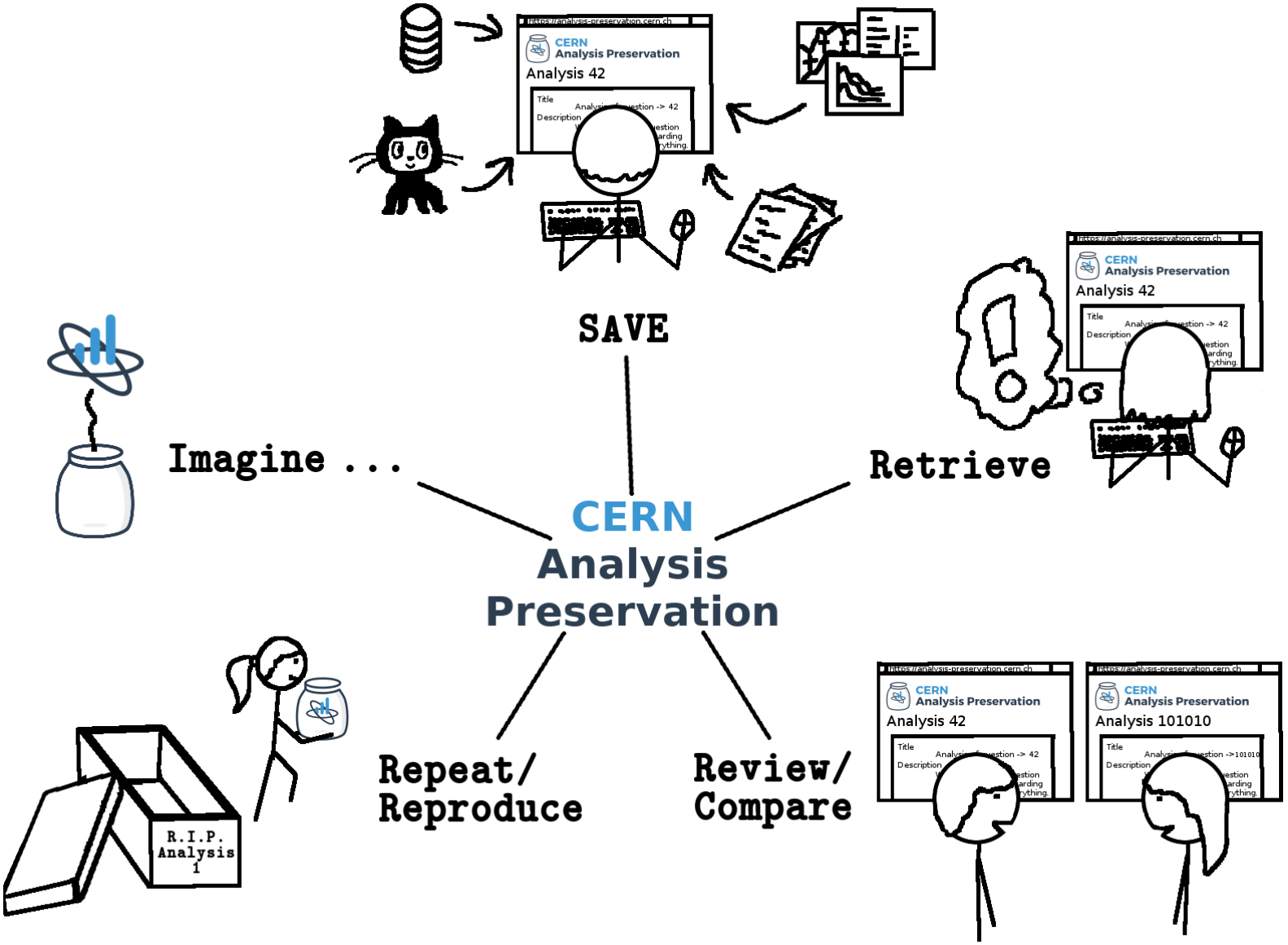
- A platform for **preserving knowledge** and **assets** of an individual physics analysis.
- Capturing the elements needed to **understand** and **rerun** an analysis even several years later:

  - ✔ data
  - ✔ software
  - ✔ environment
  - ✔ workflow
  - ✔ context
  - ✔ documentation

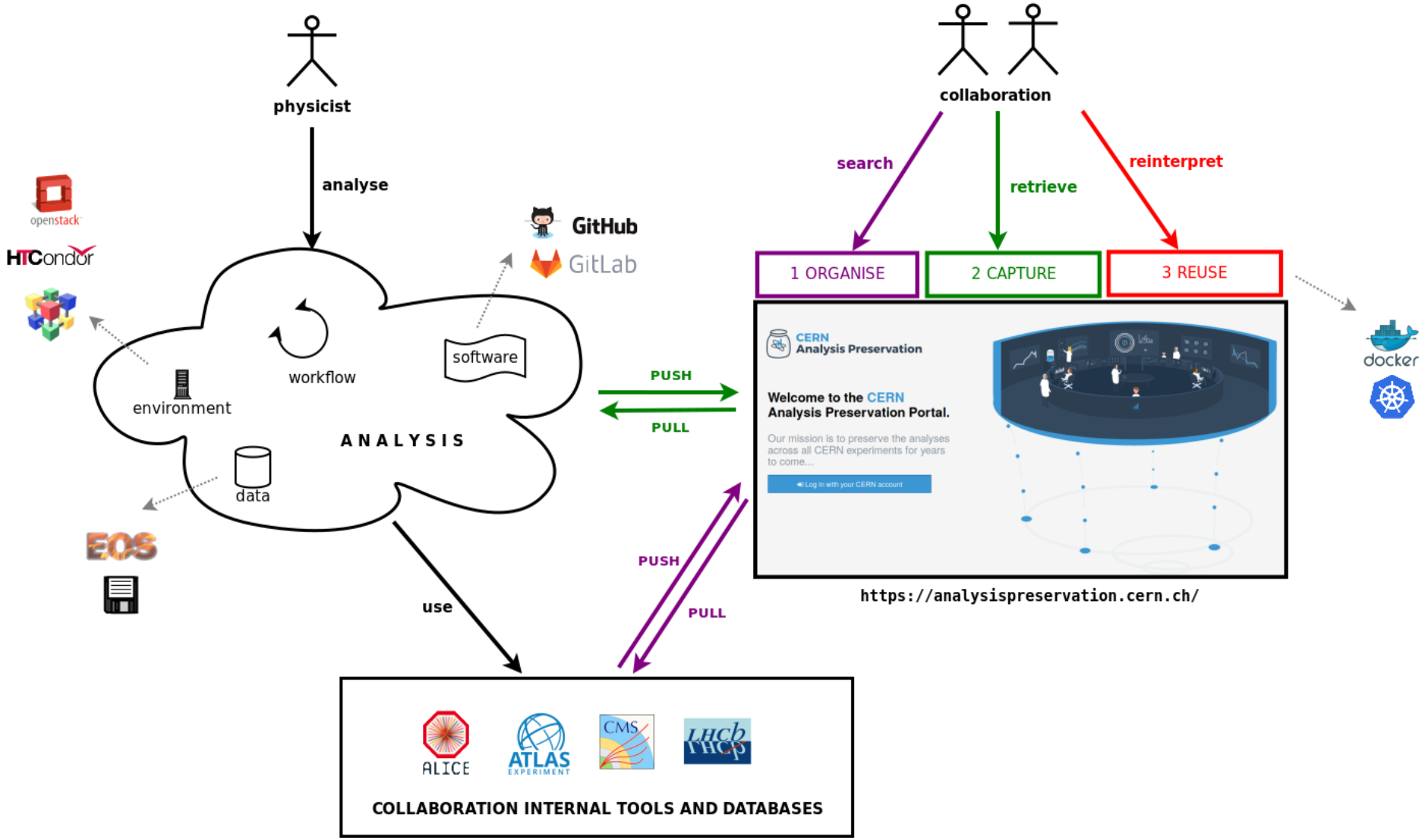
- Advanced **search** for high-level physics information
- Applying standard **collaboration access restrictions**

Developed by CERN IT and CERN SIS in close collaboration with LHC experiments

Use cases


CERN
Analysis Preservation
Analysis 42
SAVE
Imagine ...
Retrieve
CERN
Analysis
Preservation
Repeat/
Reproduce
Review/
Compare
R.I.P.
Analysis
1
Analysis 101010

System overview

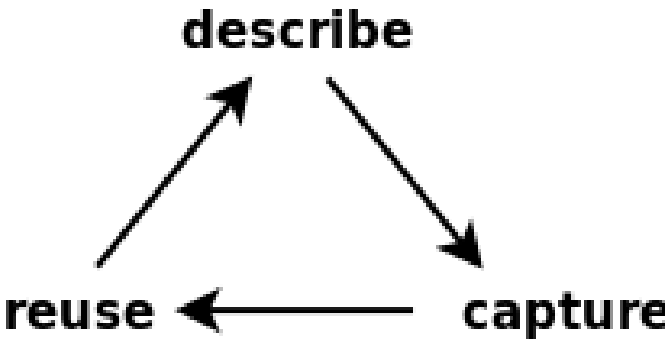


Three pillars

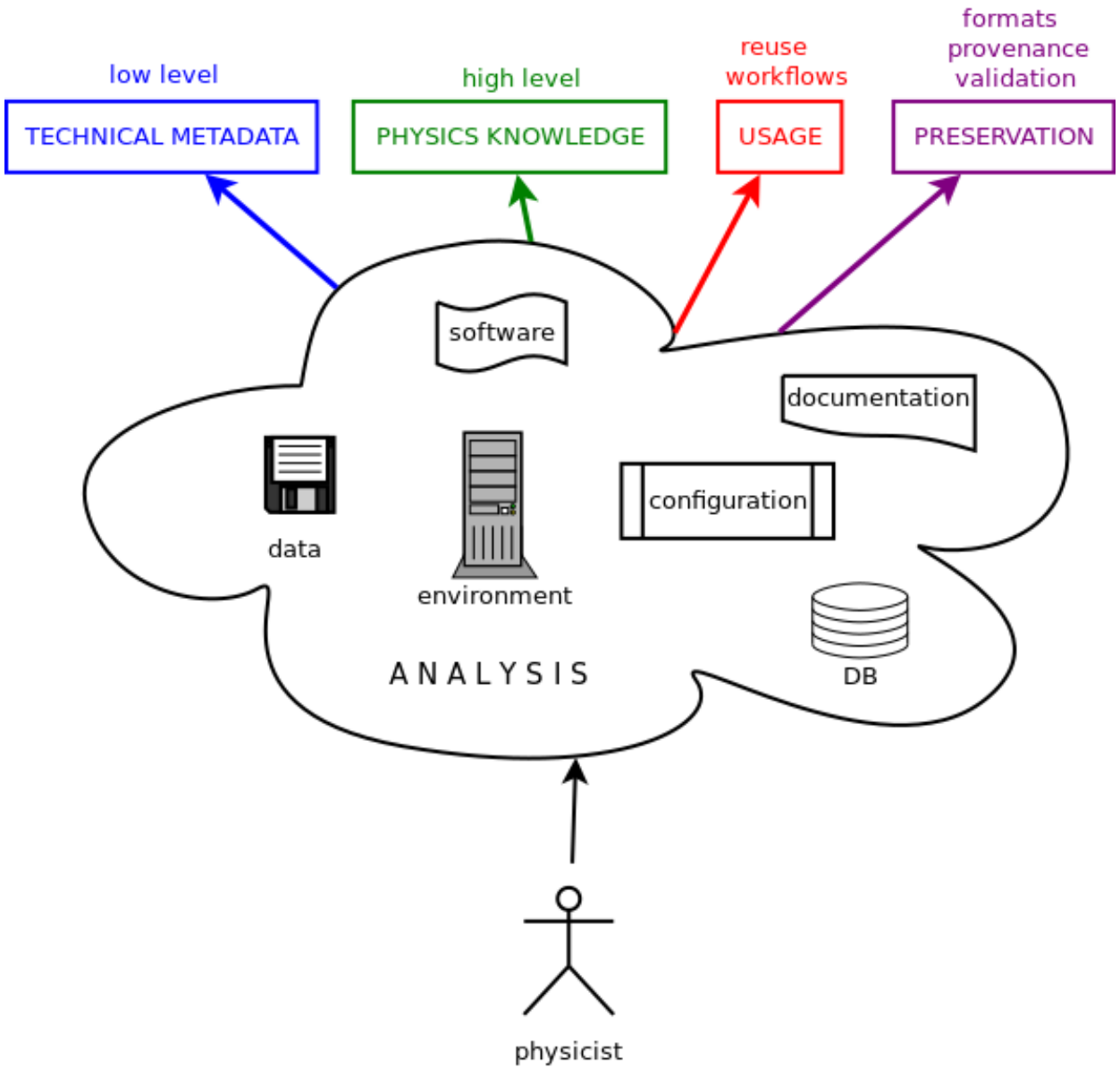
1. **Describe**
   Knowledge modelling
   Analysis description

2. **Capture**
   Push: deposit via API
   Pull: ingest via grabbing

3. **Reuse**
   Runnable components
   Reinstantiate analyses on cloud



1. Describing an analysis

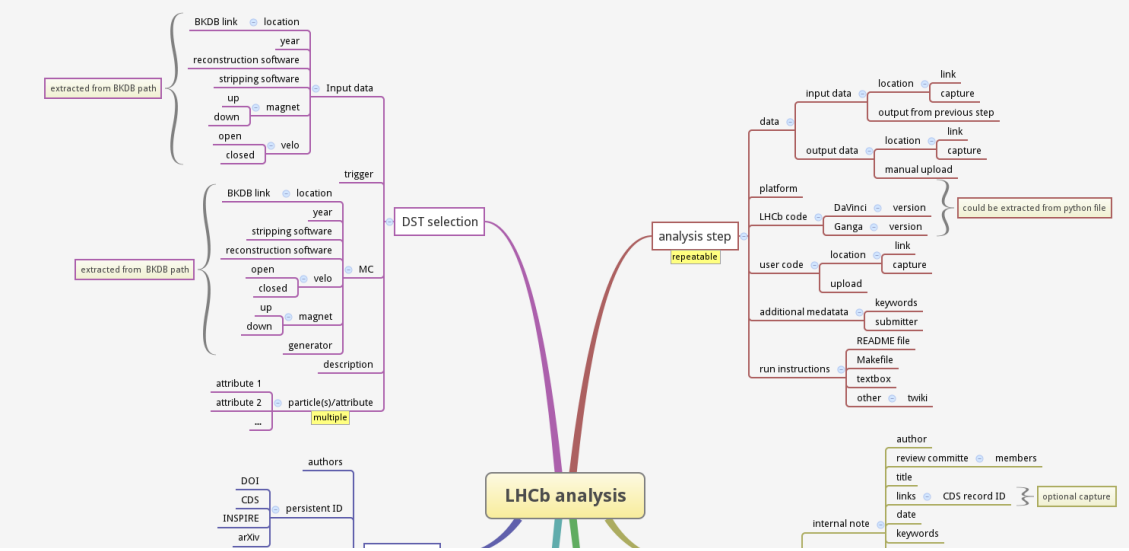


Knowledge representation

- **rare cross-discipline standards** (W3C DCAT)

```
"primary_dataset": [
  {
    "@type": "dcat:Dataset",
    "title": "/Mu/Run2010B-Apr21ReReco-v1/AOD",
    "description": "Mu primary dataset in AOD format from RunB of 2010",
    "licence": "CC0 waiver",
    "issued": "2011-04-26 11:32:43",
    "modified": "2011-05-02 21:22:30",
    [...]
```

- **domain-specific knowledge modelling**



Demo: LHCb production info

CERN Analysis Preservation | LHCb | Search | Create

LHCb Analysis 17/05/2017, 08:19:57 | Save

- Basic Information | 8 (3 req)
  - Analysis Name
  - Measurement
  - Proponents
  - Status
  - Reviewers
  - Review eGroup
  - Working Group
  - Keywords
- DST selection | 2
  - Code | 3
    - Application
    - Platform
    - User code | 0 Items
  - Production information | 2
    - Collision Data | 0 Items
    - MC Data | 0 Items
- Analysis Steps | 1 Items
- Additional Resources | 4

PRODUCTION INFORMATION

COLLISION DATA + Add New

Bookkeeping path

E.g. sim://LHCb/Collision12/Beam4000GeV-VeloClosed-MagDown/RealData/Reco14/Stripping20/90000000 (Full stream | Autofill

Processing Pass

E.g. Reco15a-Stripping22b

RECONSTRUCTION SOFTWARE

Name

E.g. Brunel Reco

Version

E.g. 13

“Describe” pillar status

- developed set of JSON schemas for LHC physics analyses
- `lhcb-v0.0.1.json`

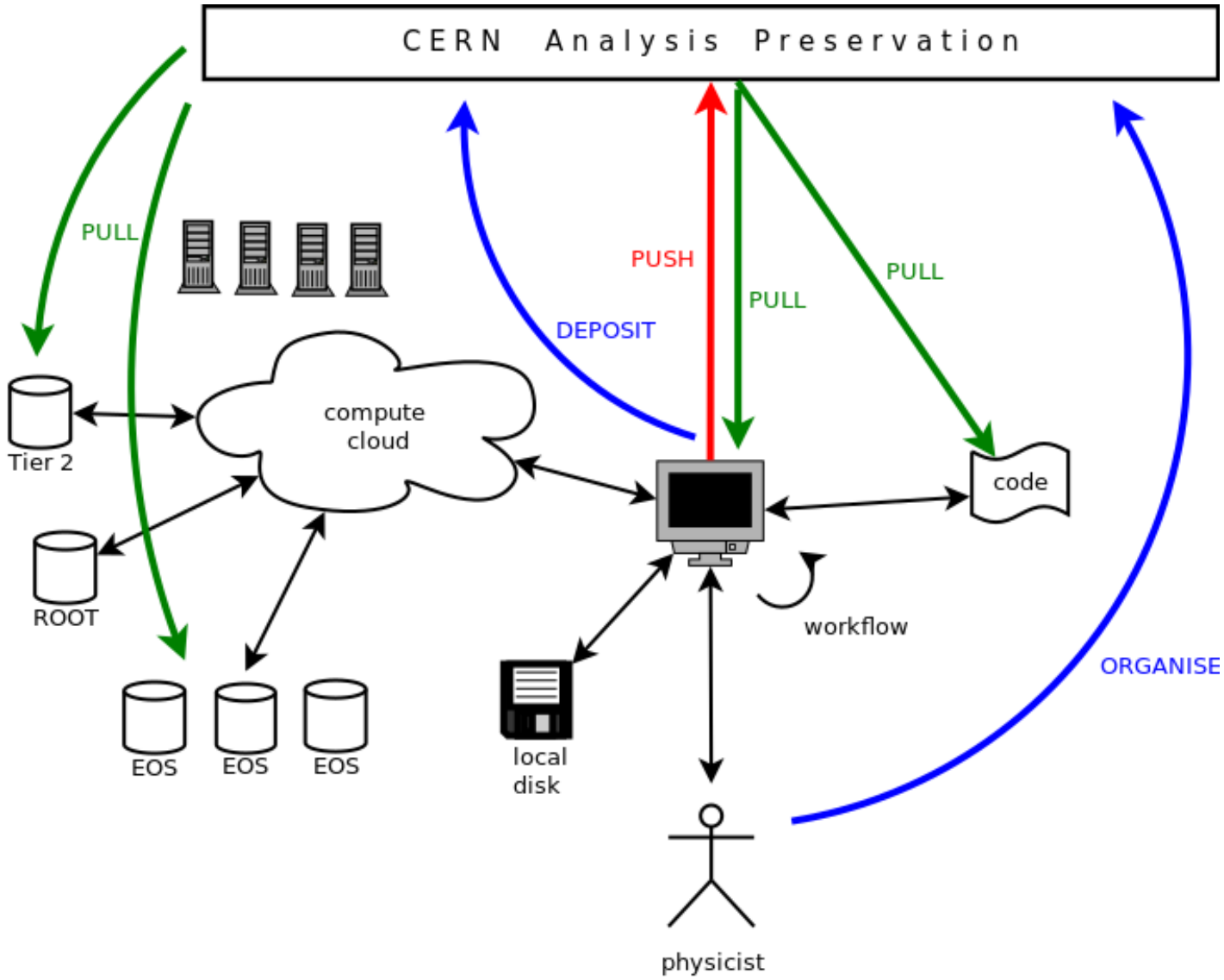
```
[...]
"dst_selection": {
  "properties": {
    "code": {
      "properties": {
        "lhcb_code": {
          "title": "Application",
          "type": "string"
        },
        "platform": {
          "title": "Platform",
          "type": "string"
        },
        "user_code": {
          "items": {
[...]
```

- ongoing tests with physicists studying feasibility

“Describe” pillar next steps

- **wider testing** on various types of LHCb physics analyses
  - different working groups?
- attaching **additional information** of interest for **reusable** knowledge preservation
  - stripping conditions?
  - particle properties?
- support for **schema migration**
  - several schemas co-existing on a system
  - `lhcb-v0.0.1.json` $\rightsquigarrow$ `lhcb-v2.3.1.json`
- **intelligent search** across captured information to ease discovery
  - enhanced faceted search
  - query language to search for e.g. loose opposite sign muons and certain p_T and η cuts?

2. Capturing an analysis



Capturing analysis assets

- capturing **datafiles** from various sources:
  - local storage
  - institute network storage
  - WLCG Tier 2 site

  via various **protocols**:
  - HTTP
  - XRootD
- capturing code from various software **repositories**:
  - Git
  - SVN
- capturing **additional information** from various sources:
  - collaboration information databases
  - TWiki
  - SharePoint

Taking consistent snapshot of information at a certain time

Demo: LHCb DaVinci script

CERN Analysis Preservation | LHCb | Search | Create

LHCb Analysis 17/05/2017, 08:19:57 -- Edited | Save

- Basic Information | 8 (3 req)
  - Analysis Name
  - Measurement
  - Proponents
  - Status
  - Reviewers
  - Review eGroup
  - Working Group
  - Keywords
- DST selection | 2
  - Code | 3
    - Application
    - Platform
    - User code | 0 Items
  - Production information | 2
    - Collision Data | 1 Items
      - Collision Data Item
        - Bookkeeping path
        - Processing Pass
        - Reconstruction software | 2
          - Name
          - Version
        - Stripping software | 2
          - Name
          - Version
        - Trigger | 1 Items

Year

E.g 2015

DAVINCI SCRIPT

Please select the preferred way to upload a file

- (•) External URL
- ( ) File Upload

URL

Preserve files from URL?

- (•) Yes
- ( ) No

E.g. https://git@github.com:johndoe/myrepo.git

“Capture” pillar status

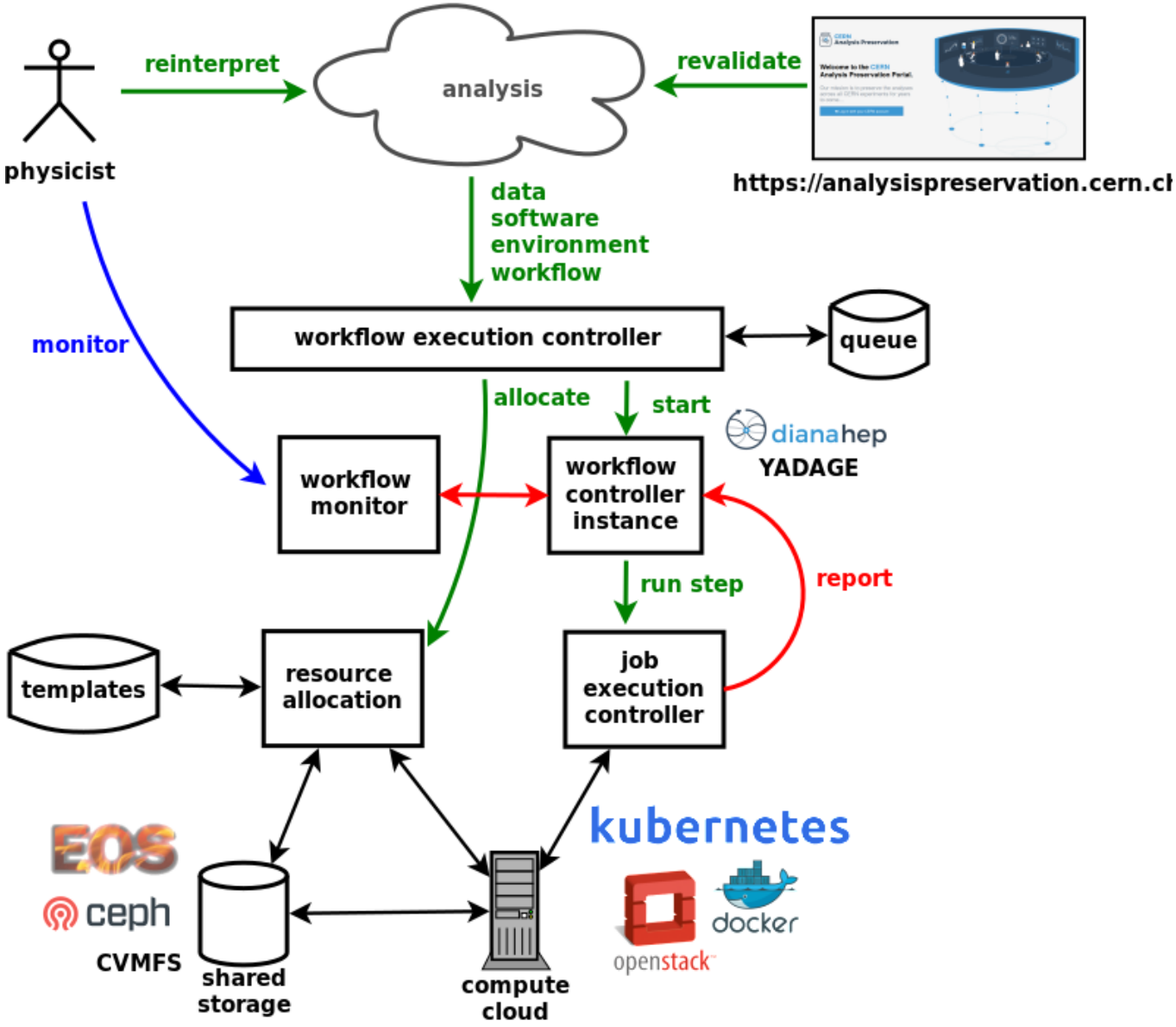
- **robust multi-server architecture**
  - high-availability: service duplication
  - Puppet
  - DEV $\rightarrow$ QA $\rightarrow$ PROD
- **collaboration-restricted access control**
  - CERN SSO and e-groups
  - OAuth
- **auto-complete** from **LHCb internal databases**
  - working group database
  - publication database
- **asset capture**
  - small files over HTTP protocol
  - push captured assets to EOS

“Capture” pillar next steps

- **dataset capture**
  - working on XRootD-based large-file capture
  - indicating status of background ingestion
- **software capture**
  - Git
  - SVN still needed?
- **information capture**
  - TWiki?
  - more?
- develop **rich** CERN Analysis Preservation **client**
  - plug into your analysis pipelines

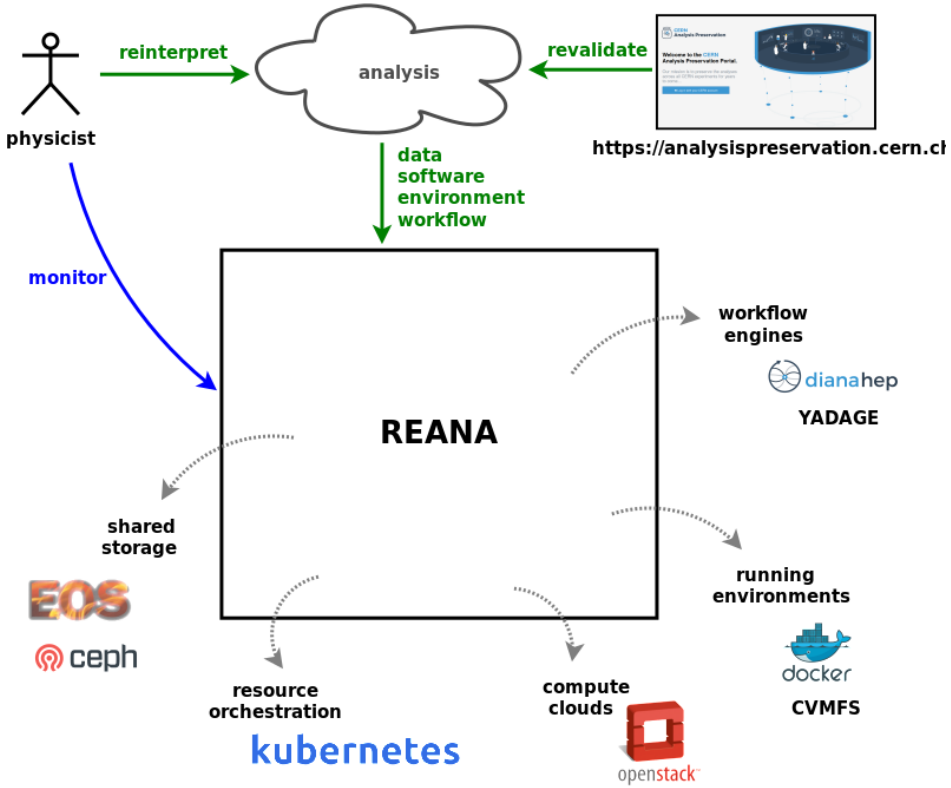
```
$ cap store ...
$ cap share ...
```

3. Reusing an analysis



REANA = RE usable ANA lyses

- a system to **instantiate** preserved analysis **on the cloud**
  - `https://reanahub.io`
- supporting **multiple scenarios**
  - multiple computing clouds
    → CERN OpenStack
  - multiple running environments
    → Docker
  - multiple resource orchestration
    → Kubernetes
  - multiple workflow engines
    → Yadage
  - multiple shared storage systems
    → Ceph, EOS

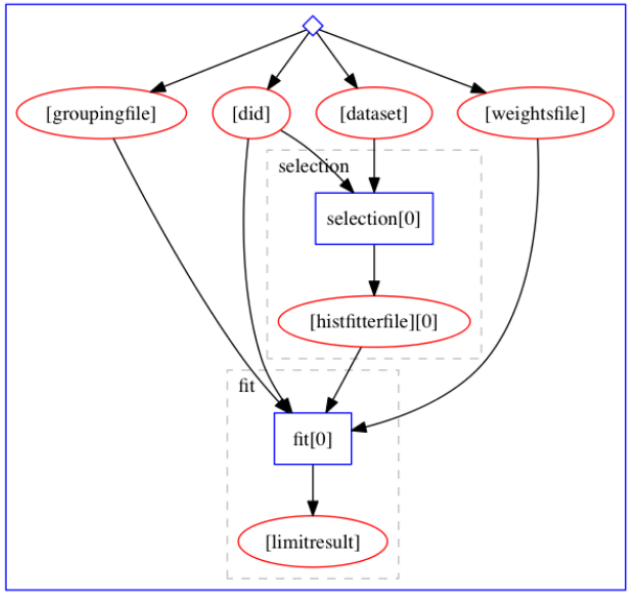


- close **collaboration** with DASPOS and recast

Demo: Reusable analysis pilot

- case study: ATLAS multi-B-jet analysis
- case study: LHCb Lb2LcD0K analysis

```yaml
stages:
  - name: selection
    dependencies: ['init']
    scheduler:
      scheduler_type: singlestep-stage
      parameters:
        dataset: {stages: init, output: dataset, unwrap: true}
        submitdir: '{workdir}/submitdir'
        outputprefix: '{workdir}/histfitter.root'
        did: {stages: init, output: did, unwrap: true}
      step: {$ref: 'selscript.yml#'}
  - name: fit
    dependencies: ['selection']
    scheduler:
      scheduler_type: singlestep-stage
      parameters:
        bkgtree: 'root://eosuser.cern.ch///eos/project/r/recast/Bkg_2.4.15-2-0_merged.root'
        datatree: 'root://eosuser.cern.ch///eos/project/r/recast/Data_2.4.15-2-0.root'
        outputjson: '{workdir}/fitoutput.json'
        selectionoutput: {stages: selection, output: histfitterfile, unwrap: true}
        weightsfile: {stages: init, output: weightsfile, unwrap: true}
        did: {stages: init, output: did, unwrap: true}
      step: {$ref: 'fitscript.yml#'}
```



Lukas Heinrich `http://github.com/diana-hep/yadage`

“Reuse” pillar status

- developer-oriented **internal release** of REANA
  - run on local `minikube` cluster
  - run on CERN Cloud infrastructure
- two basic **usage examples**
  - “hello world”
  - Jupyter notebook
- several **HEP analysis** examples
  - first ATLAS examples
  - first LHCb example
  - see Lukas’s next talk

“Reuse” pillar next steps

- extending **features**
  - better user monitoring
  - easier result publishing
- testing more **real-life analysis examples**
  - ALICE post-LEGO-train analyses (ROOT macros)
  - more scenarios from ATLAS, CMS, LHCb
- setting up central **REANA server** at CERN
  - used by services (CAP, COD, Zenodo, . . .)
  - used by physicists?
- developing **rich REANA client**

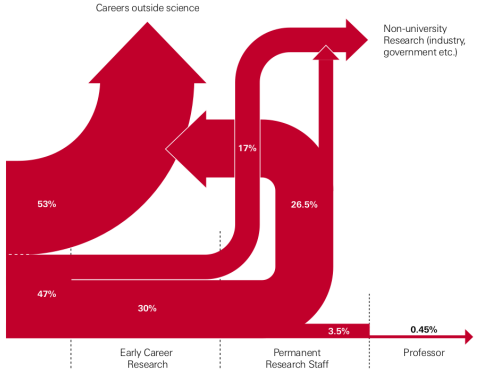
```
$ reana-client prepare myanalysis.yaml
$ reana-client run myanalysis
$ reana-client logs myanalysis
```

- support for **more backends**
  - other container technologies? Singularity? Umbrella?
  - other workflow engines? Snakemake? Luigi?

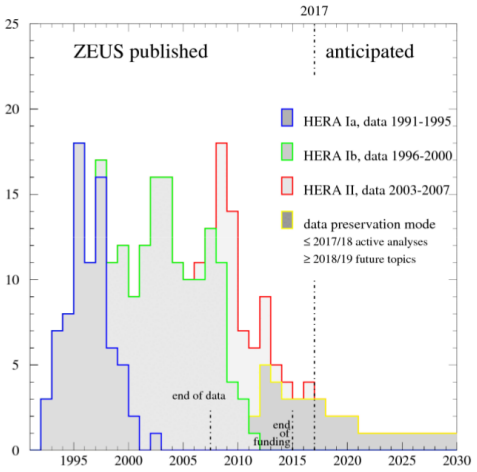
Pragmatic focus

- **publish or perish**
  - time devoted to preservation = time taken away from the next paper?
  - “preservation” platform $\rightsquigarrow$ “live” platform
- **cultural change**
  - adopt “runnable READMEs”
  - encapsulate running environment
  - use structured workflows
- scientific **benefit vs cost** of preservation
  - Achim Geiser’s study of ZEUS publishing history and long-term preservation efforts: $\sim$10% more papers for $<$1% of total cost (of which $\sim$90% during active phase)

Achim Geiser `https://indico.cern.ch/event/588219`



career after PhD
THE ROYAL SOCIETY



CERN Analysis Preservation

CERN Analysis Preservation

- http://analysispreservation.cern.ch
- http://github.com/cernanalysispreservation

REANA

- http://reanahub.io
- http://github.com/reanahub
- http://twitter.com/reanahub

CERN IT H. Hirvonsalo, D. Rodríguez, T. Šimko **CERN SIS** S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, A. Lavasa, A. Mattmann, I. Tsanaktsidis, A. Trzcinska **ALICE** M. Gheata, C. Grigoras, M. Zimmermann **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher **CMS** A. Calderon, A. Geiser, A. Huffman, K. Lassila-Perini, T. McCauley, A. Rao, A. Rodriguez Marrero **LHCb** S. Amerio, B. Couturier, S. Neubert, A. Trisovic **CERN CernVM** J. Blomer **CERN Kubernetes** R. Rocha **CERN EOS** L. Mascetti **DASPOS** M. Hildreth, H. Meng, D. Thain, A. Vyushkov **DPHEP** J. Shiers