Analysis and data preservation initiative in ALICE

ALICE Tier-1/Tier-2 Workshop

Markus Zimmermann
05.05.2017

Outline

- Analysis Preservation
  - CAP: CERN Analysis Preservation
  - Which information should be preserved?
  - How to extract these information
- Data Preservation
  - Storage on Open Data
  - Re-analysis on REANA

What is Analysis Preservation?

- Documenting an analysis to reproduce
  - approved results by the collaboration
  - an analysis with the possibility to modify the procedure
  - an analysis by a third party outside ALICE
- Preserve beyond the ALICE lifetime
  - full analysis configuration
  - necessary software

CAP - CERN Analysis Preservation

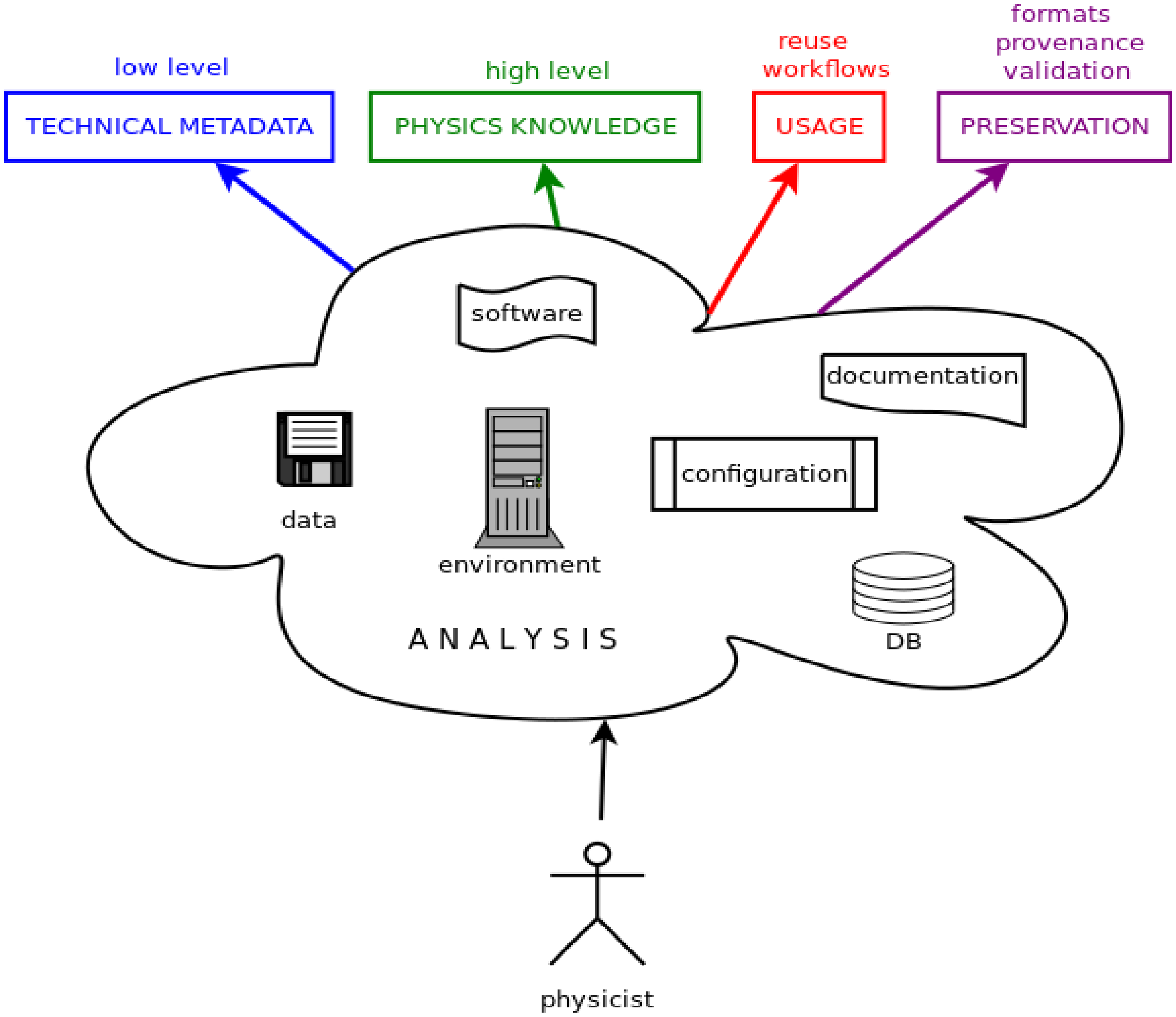
- Long term analysis preservation system at CERN
- Test system which works only within the CERN network
- Production system has still some bugs

- CAP efforts focus on three pillars:
- **Describe** the data analysis process
- **Capture** the sofware
- **Reuse**: re-instantiate the preserved analysis

Describe

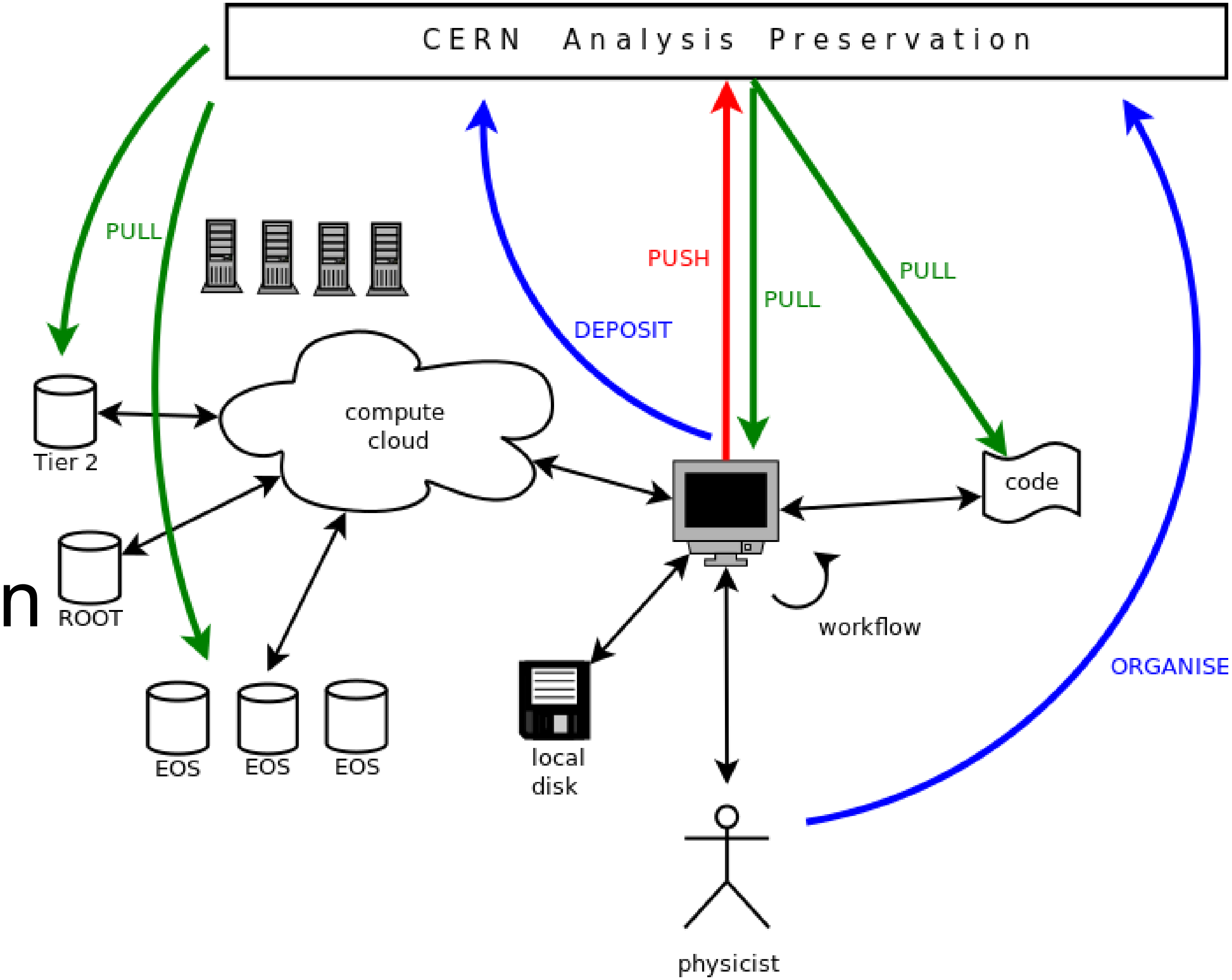
Create references between

- •used dataset
- •computing infrastructure
- •code in AliPhysics
- •Analysis code configuration
- •analysis note
- •train runs on the LEGO trains
- •paper publication



Capture

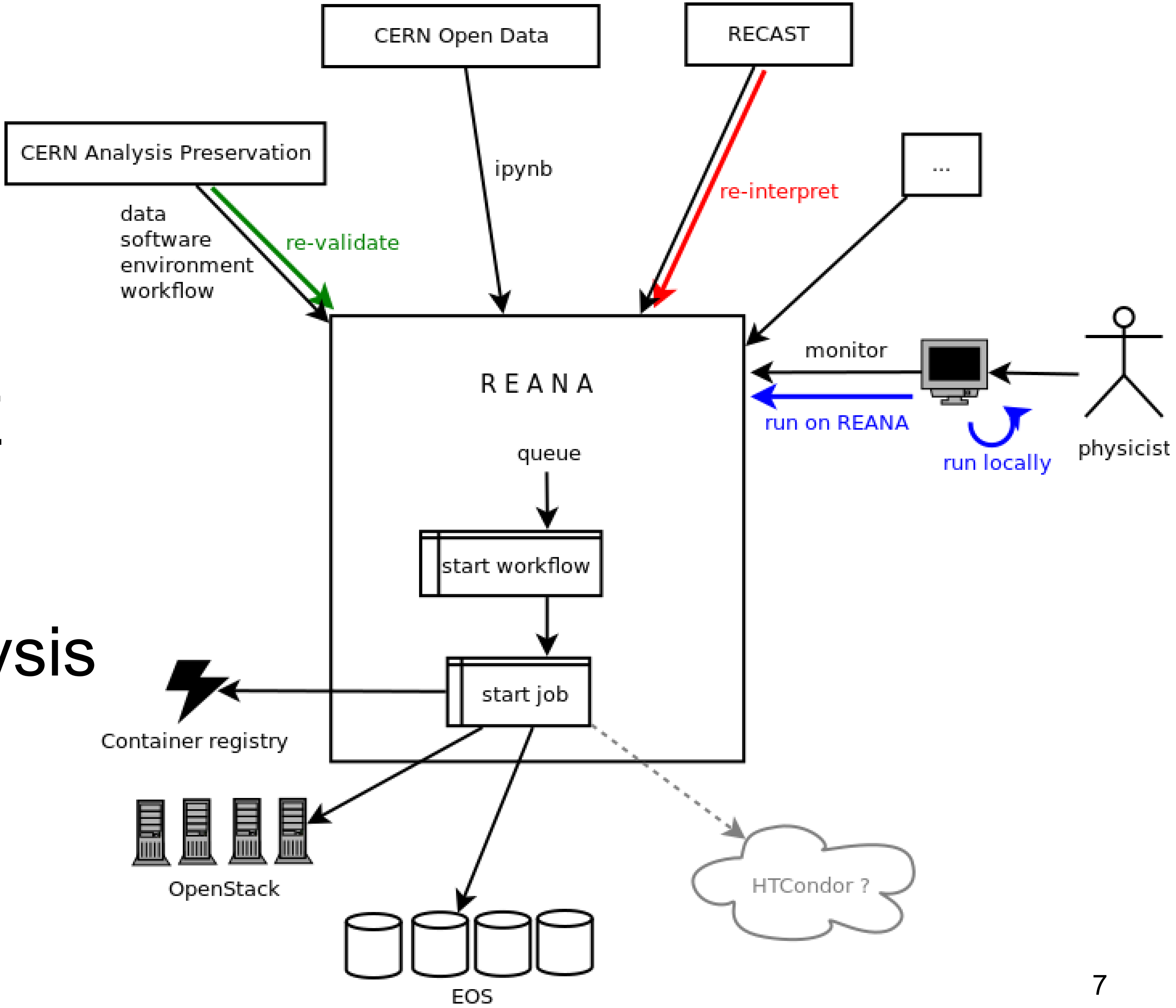
- ensure all code is in AliPhysics
- preserve
  - train configuration
  - local macros
  - dataset definition
  - analysis note



Reuse

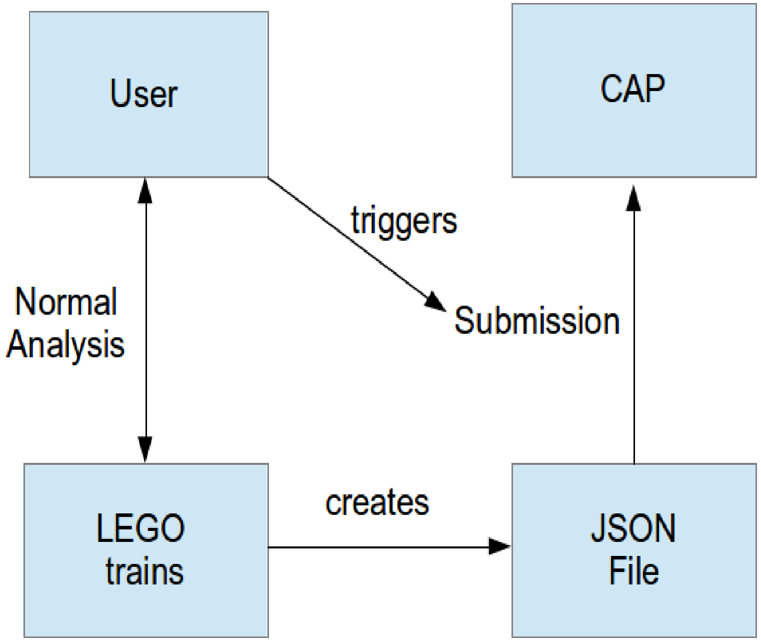
- Inside ALICE
  - Rerun trains
- Outside ALICE
  - REANA

- Preserve analysis steps after the trains



How to work with CAP

- LEGO trains create JSON file for each train run
  - Transfer to CAP has to be triggered by the user or the conference committee
- Additional information can be added afterwards on CAP manually
- If LEGO trains are not used, the full entry has to be generated manually



- Work on a CAP entry with multiple people (e-groups)
- Share finished entry with the whole collaboration

Information to Preserve

- Used dataset
  - Identifier in RCT
  - Run numbers
- Computing infrastructure
  - ALICE analysis configuration
- Analysis code
  - AliPhysics code on Github
  - AddTask in AliPhysics
  - Code configuration
  - LEGO train run
- Link to documentation/publications
  - ALICE analysis note
  - Journal reference

Information from the LEGO trains

Information has to be added manually

Information to Preserve

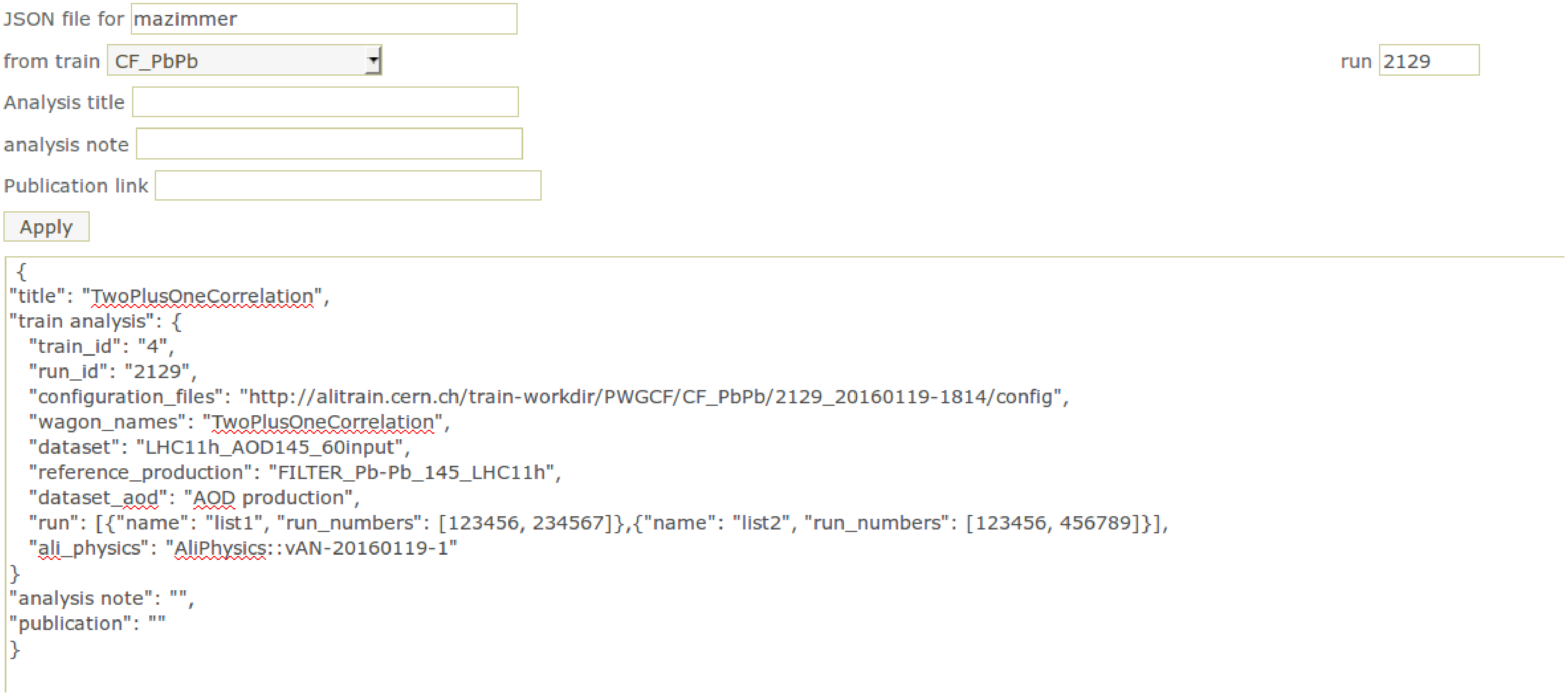
- Used dataset
  - Identifier in RCT
  - Run numbers
- Computing infrastructure
  - ALICE analysis configuration
- Analysis code
  - AliPhysics code on Github
  - AddTask in AliPhysics
  - Code configuration
  - LEGO train run
- Link to documentation/publications
  - ALICE analysis note
  - Journal reference

RCT and Github repository
have to be preserved separately

Information from the
LEGO trains

Information has to be
added manually

JSON File from the LEGO Trains



JSON file for mazimmer

from train CF_PbPb

run 2129

Analysis title

analysis note

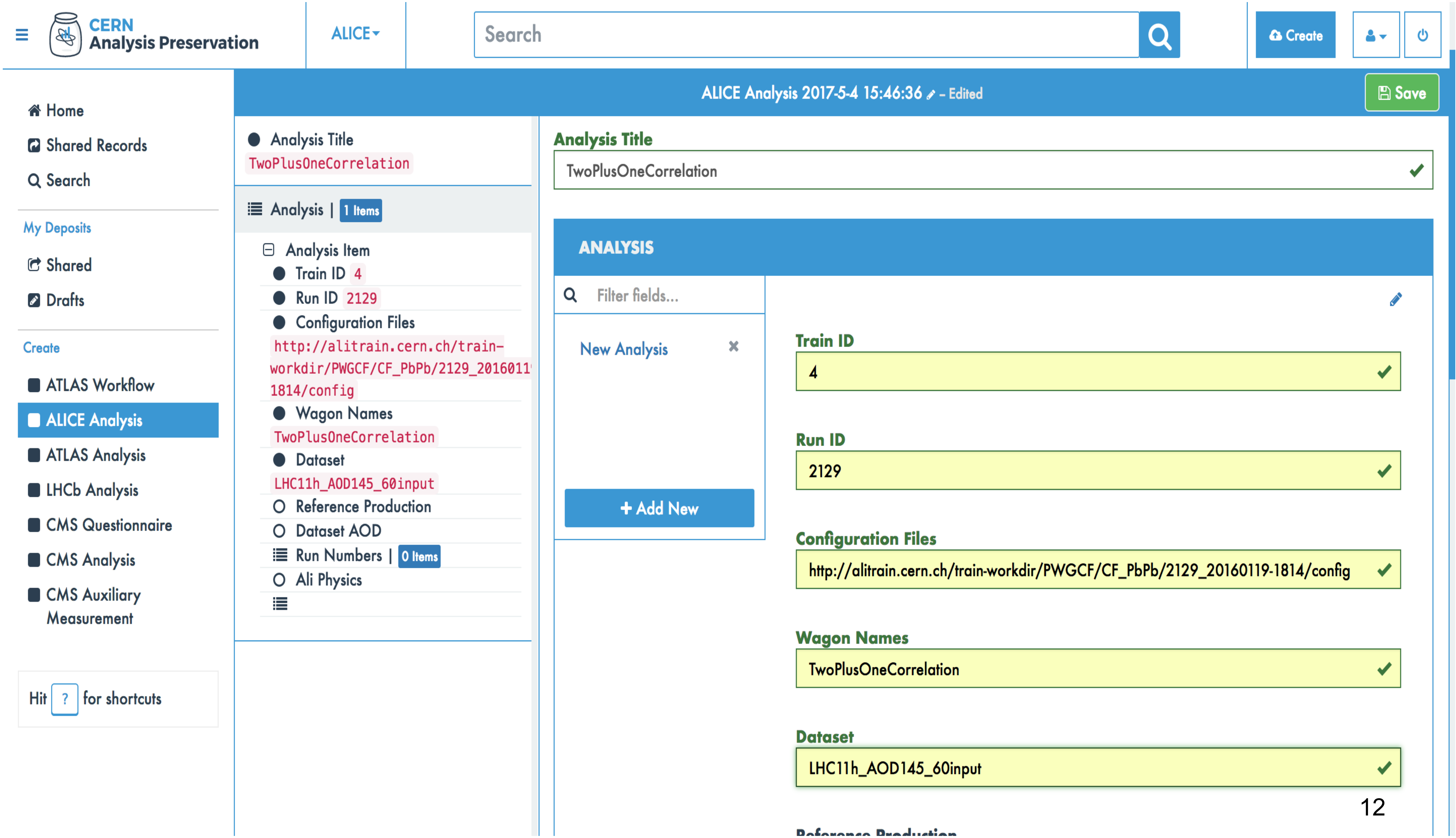
Publication link

```
 {
"title": "TwoPlusOneCorrelation",
"train analysis": {
  "train_id": "4",
  "run_id": "2129",
  "configuration_files": "http://alitrain.cern.ch/train-workdir/PWGCF/CF_PbPb/2129_20160119-1814/config",
  "wagon_names": "TwoPlusOneCorrelation",
  "dataset": "LHC11h_AOD145_60input",
  "reference_production": "FILTER_Pb-Pb_145_LHC11h",
  "dataset_aod": "AOD production",
  "run": [{"name": "list1", "run_numbers": [123456, 234567]},{"name": "list2", "run_numbers": [123456, 456789]}],
  "ali_physics": "AliPhysics::vAN-20160119-1"
}
"analysis note": "",
"publication": ""
}
```

- Describes train analysis
- Changes possible in the CAP web page
- Add local macros in the CAP web page

CAP

- https://analysispreservation.cern.ch
- Some issues with the availability outside CERN



ALICE Analysis 2017-5-4 15:46:36 – Edited

● Analysis Title
TwoPlusOneCorrelation

- Train ID 4
- Run ID 2129
- Configuration Files http://alitrain.cern.ch/train-workdir/PWGCF/CF_PbPb/2129_2016011 1814/config
- Wagon Names TwoPlusOneCorrelation
- Dataset LHC11h_AOD145_60input
- ○ Reference Production
- ○ Dataset AOD
- Run Numbers | 0 Items
- ○ Ali Physics

Analysis Title
TwoPlusOneCorrelation ✓

Train ID
4 ✓

Run ID
2129 ✓

Configuration Files
http://alitrain.cern.ch/train-workdir/PWGCF/CF_PbPb/2129_20160119-1814/config ✓

Wagon Names
TwoPlusOneCorrelation ✓

Dataset
LHC11h_AOD145_60input ✓

Reference Production

Why using CAP?

- Long term preservation service
  - Maintenance provided by CERN IT
  - Lifetime beyond ALICE lifetime
- Searching and grouping of analyses
- Option to upload local files from the users
- Entries can be automatically created
  - LEGO trains can provide most information
  - Convenient web page to fill up additional information
  - Manual insertion necessary if LEGO trains are not used
- Option to rerun the analysis with REANA

Data Preservation

ALICE Data Preservation Strategy

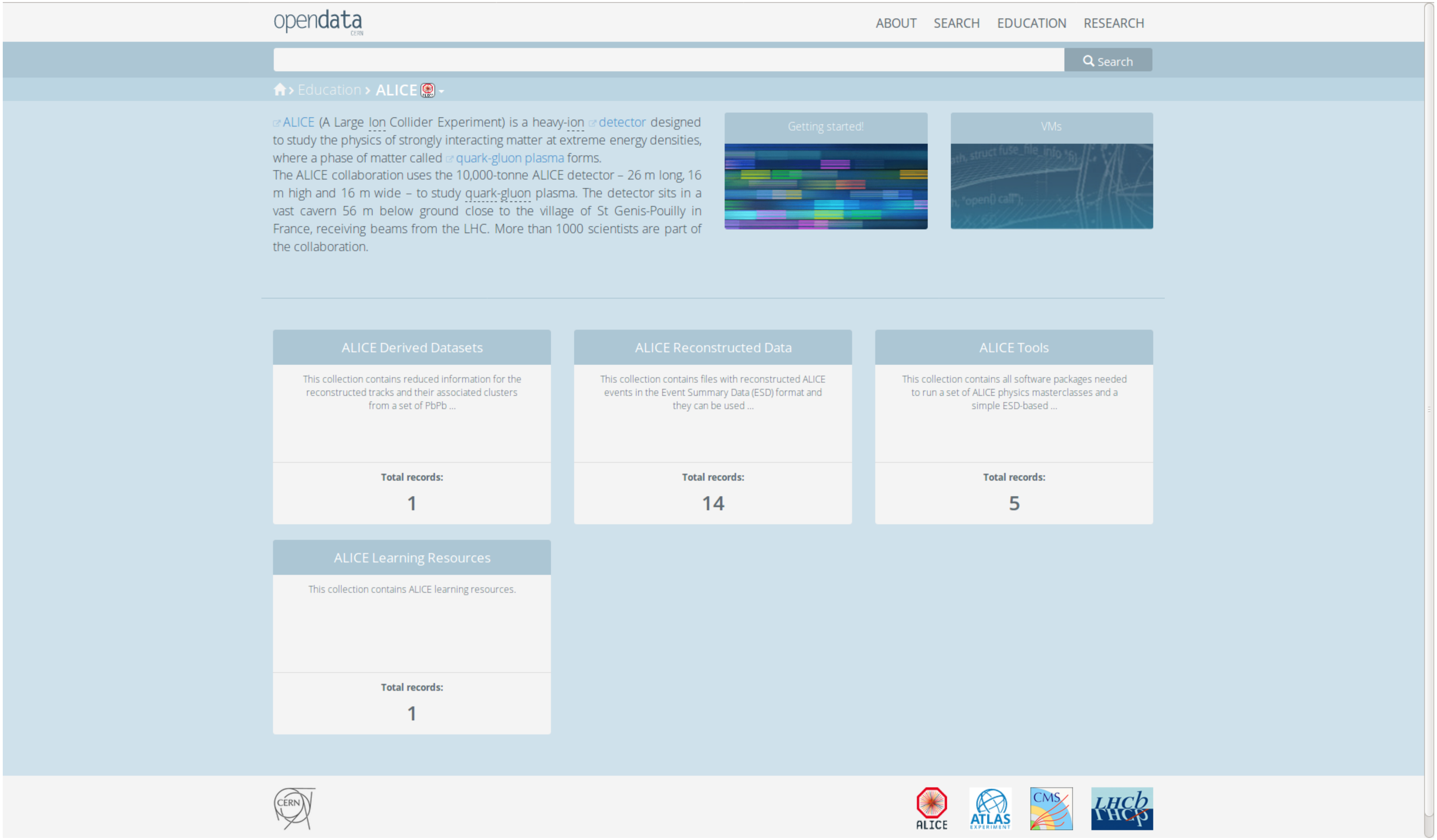
- Purpose of data preservation
  - Preserve data and software inside ALICE
  - Sharing data with the larger scientific community
  - Give access to reduced datasets to the general public for educational and outreach activities
- Preserved data is only meaningful in combination with the software to analyze it
- Publish AOD data and MC truth
  - 10% of the data after 5 years
  - 100% of the data after 10 years
- For long term preservation
  - Use Open Data, web portal provided by CERN IT

Open Data

- CERN IT platform to share data and software with the public
- Currently published ALICE data
  - 14 reconstructed ESD datasets (Minimum Bias interactions)
  - LHC10b pp collisions (a few files for Masterclasses)
  - LHC10c pp collisions $27 \cdot 10^6$ ($400 \cdot 10^6$)
  - LHC10h PbPb collisions $2.9 \cdot 10^6$ ($53 \cdot 10^6$)
    - Runs 139038, 139173, 139437, 139438, 139465
    - Only some files from Run 138275 ($2.5 \cdot 10^6$) are uploaded for Masterclasses
- Current storage capacity of 50 TB is donated by IT
  - 8 TB in use
  - 42 TB are still free

Open Data

- http://opendata.cern.ch/research/ALICE



ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms.
The ALICE collaboration uses the 10,000-tonne ALICE detector – 26 m long, 16 m high and 16 m wide – to study quark-gluon plasma. The detector sits in a vast cavern 56 m below ground close to the village of St Genis-Pouilly in France, receiving beams from the LHC. More than 1000 scientists are part of the collaboration.

ALICE Derived Datasets

This collection contains reduced information for the reconstructed tracks and their associated clusters from a set of PbPb ...

Total records: 1

ALICE Reconstructed Data

This collection contains files with reconstructed ALICE events in the Event Summary Data (ESD) format and they can be used ...

Total records: 14

ALICE Tools

This collection contains all software packages needed to run a set of ALICE physics masterclasses and a simple ESD-based ...

Total records: 5

ALICE Learning Resources

This collection contains ALICE learning resources.

Total records: 1

Publish Data from 2011

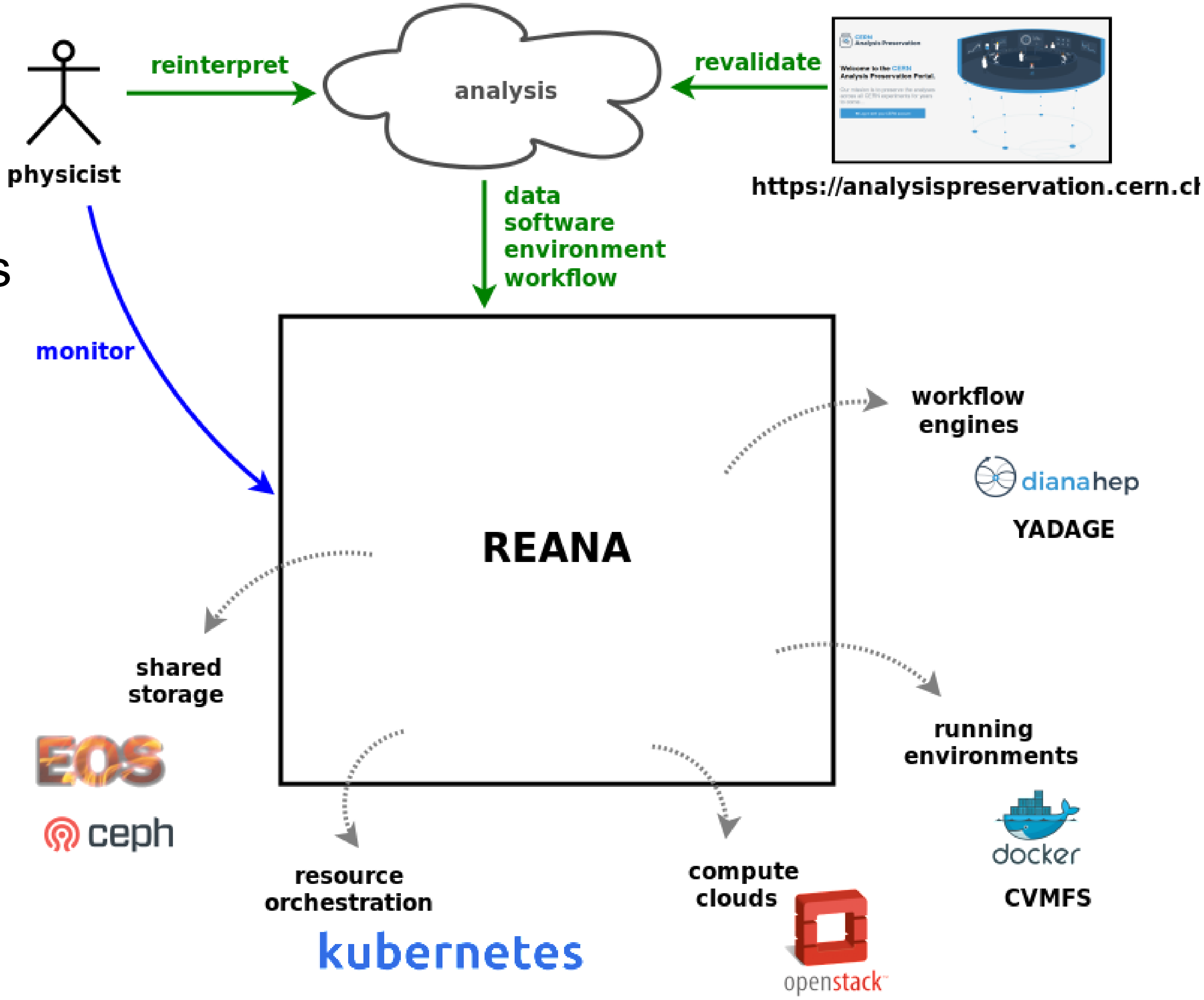
- Criteria for the data to be published
  - Good global quality
  - No detector missing
- LHC11 PbPb collisions
  - LHC11h PbPb collisions $11.3 \cdot 10^6$ minimum bias events
  - Publish $1.1 \cdot 10^6$ events, e.g. with these run numbers:
    - 168464, 168512, 168115, 168311, 169855, 168342, 169838, 168826, 169846, 168108, 169411, 167920, 167987, 168107, 168511, 169417, 168467, 169035, 168361, 169094, 169099, 170040, 169588
    - Estimated disk space: ESD 130 TB
      AOD 48 TB

REANA

- REusable ANAlysis
- Possibility to rerun ALICE analysis without ALICE infrastructure
  - first test runs are ongoing
  - use input from CAP
  - run code within docker container
- Analysis is composed out of separate modules
  - Different train wagon analyses
  - Post-processing of the analyses
- Possibility to Integrate CVMFS

REANA

- To use REANA provide
  - Data: open data
  - Software: CVMFS
  - Environment: LEGO trains
  - Workflow: user defintion
- Can be used for
  - The train run
  - Plot production with local macros
- A test run is planned with docker containers



physicist

reinterpret

analysis

revalidate

https://analysispreservation.cern.cl

data
software
environment
workflow

monitor

REANA

workflow engines

dianahep

YADAGE

shared storage

EOS

ceph

running environments

docker

CVMFS

resource orchestration

kubernetes

compute clouds

openstack

Summary & Outlook

- CERN Analysis Preservation
  - Tool for long term analysis preservation
  - Each LEGO train automatically generates a file to fill CAP
- Data Preservation with Open Data
  - Have 42 TB of free disk space to publish 2011 PbPb data
  - Find more storage capacity for the other datasets
  - Do test run for future data publications
- REANA
  - RERUN analysis without ALICE infrastructure
  - Preserve procedure to create approved plots
  - REANA test run is ongoing

BACKUP

Open Data

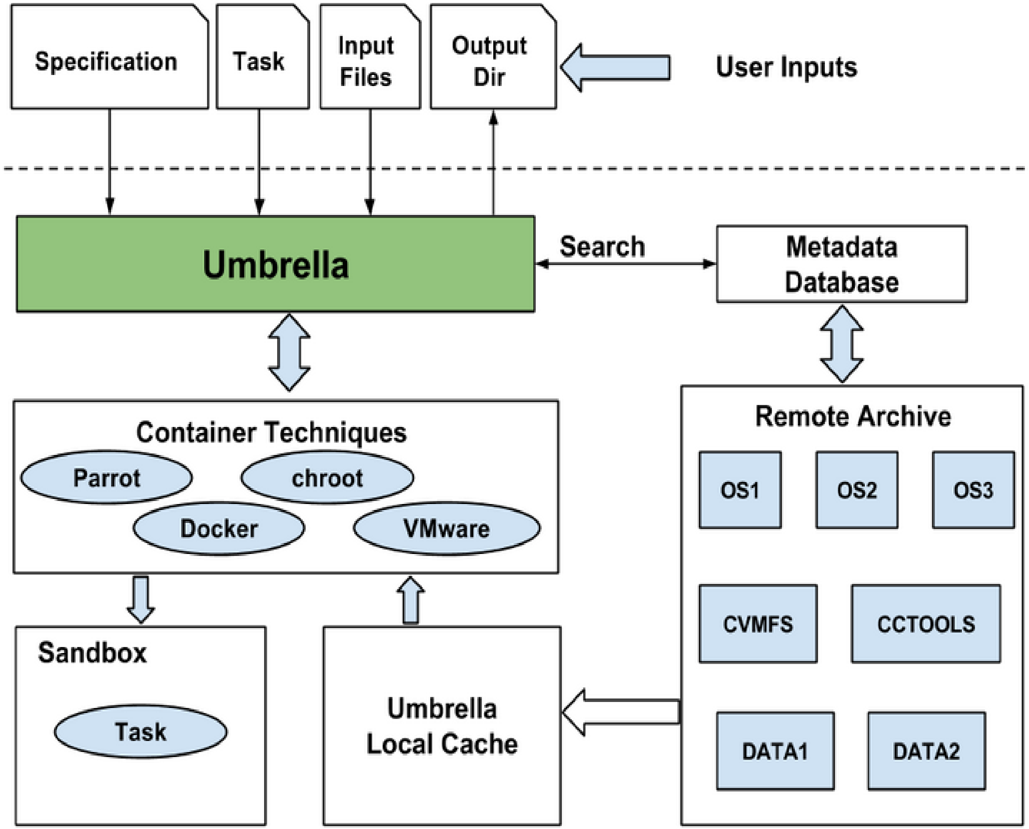
- CERN Platform to share data with the public
- Currently used to publish ALICE data
  - 14 reconstructed ESD datasets
  - LHC10b pp collisions 0.5GB (Master classes)
  - LHC10c pp collisions 1.4TB
  - LHC10h PbPb collisions 4.6TB
- Option to publish more from 2011?

Responsibility for the published Data

- ALICE data is released under Creative Commons CCO waiver
  - Re-use under the responsibility of the final user
- Publications from non-members must contain
  - Acknowledgement: “data was collected by ALICE”
  - Disclaimer: “no responsibility is taken by the ALICE collaboration for the results published here”

Umbrella

- Framework to run an analysis independent of the system architecture
- We provide working code within umbrella
- Umbrella guarantees compatibility in the future
- Input are the LEGO train files


Specification
Task
Input Files
Output Dir
User Inputs
Umbrella
Search
Metadata Database
Container Techniques
Parrot
chroot
Docker
VMware
Remote Archive
OS1
OS2
OS3
CVMFS
CCTOOLS
DATA1
DATA2
Sandbox
Task
Umbrella Local Cache