

Long Term Analysis Preservation

with

CERN Analysis Preservation

Markus Zimmermann

29.03.2017

# What is Analysis Preservation?

- Documenting an analysis to reproduce later
  - the approved plots
  - an analysis within ALICE
  - an analysis outside of ALICE
- Preserve the full analysis configuration
- Preserve the necessary software

# CAP - CERN Analysis Preservation

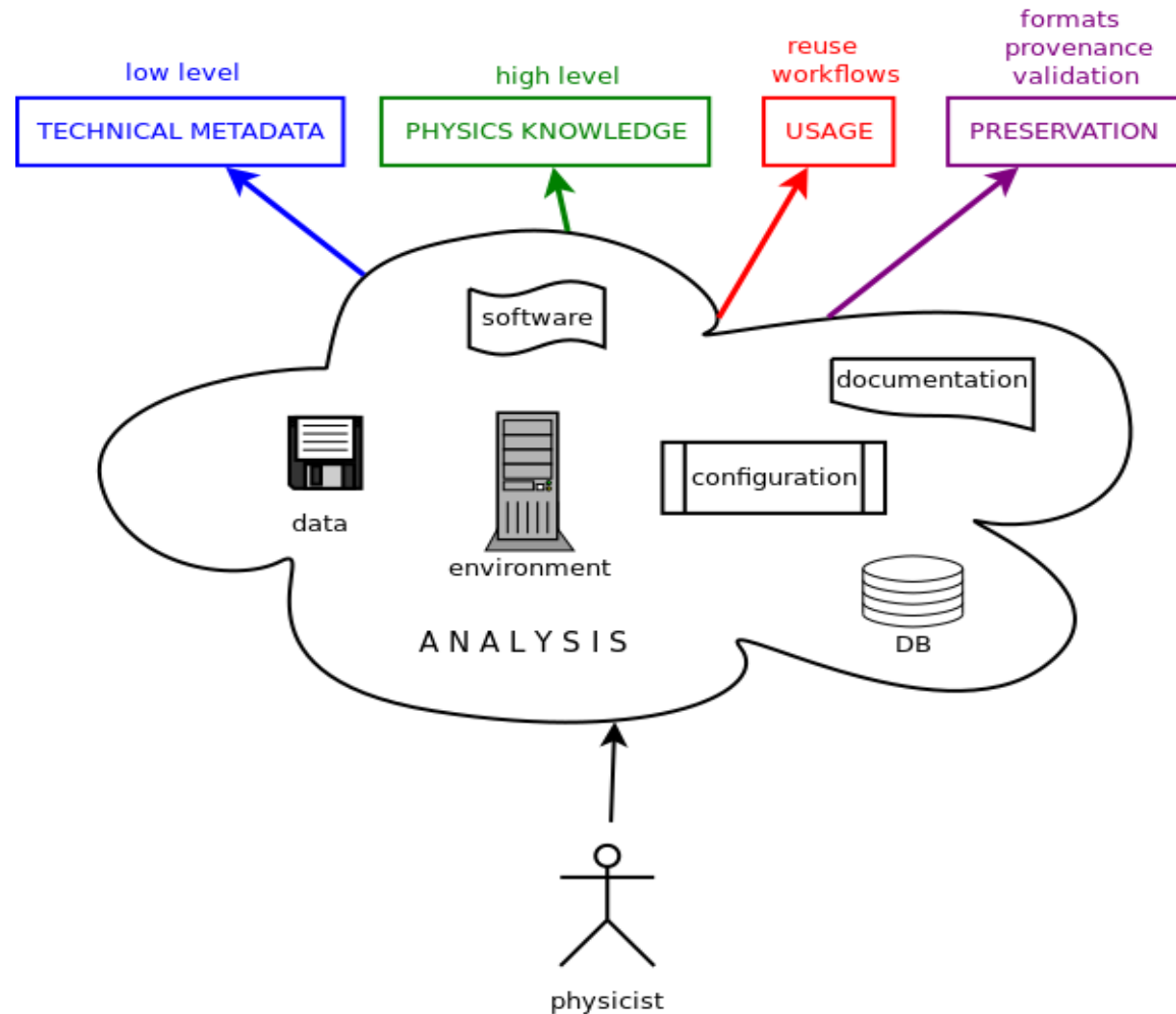
CAP efforts focus on three pillars:

- **describe** the data analysis processes
- **capture** the software
- **reuse**: re-instantiate the preserved analyses

# Describe

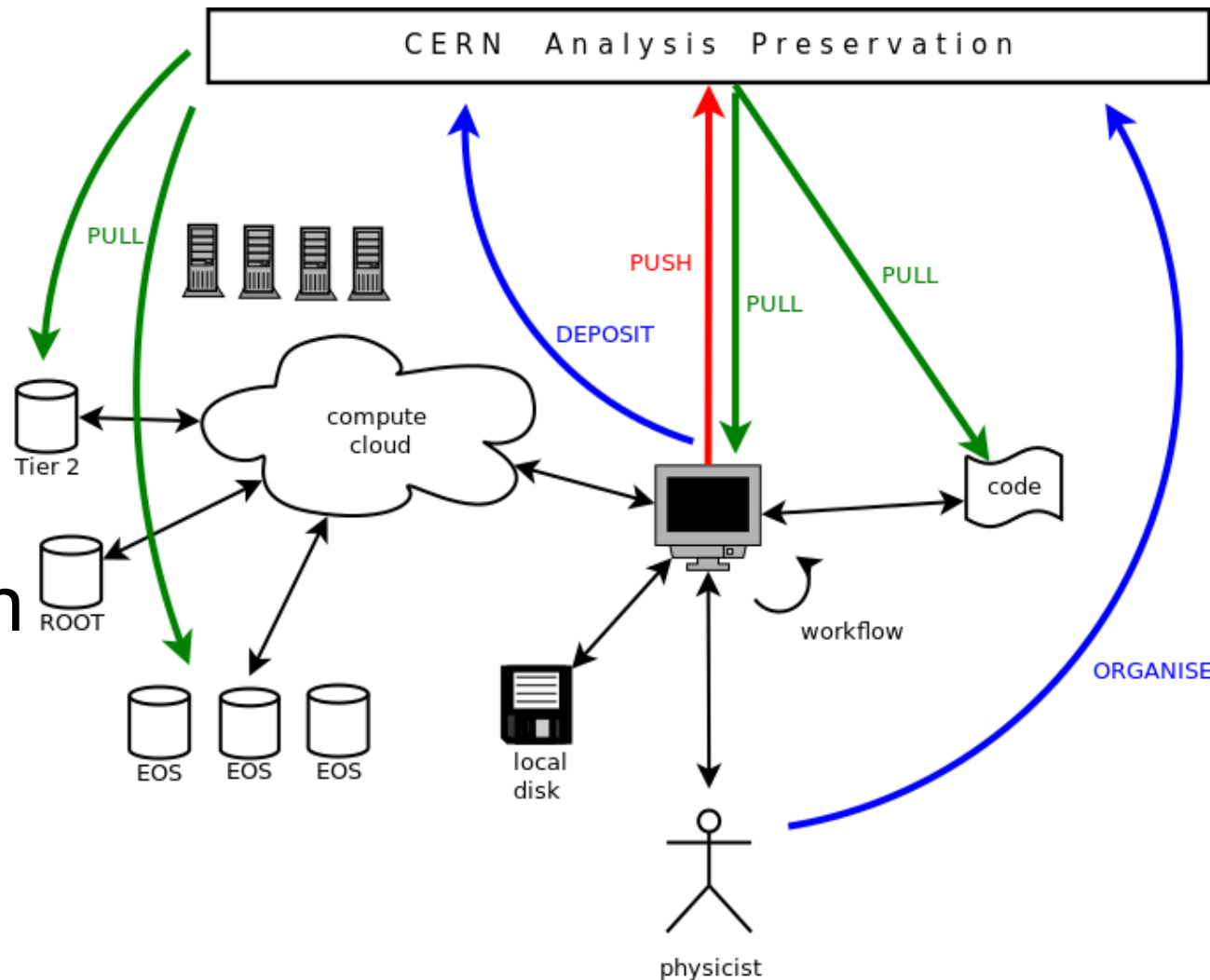
Create references between

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



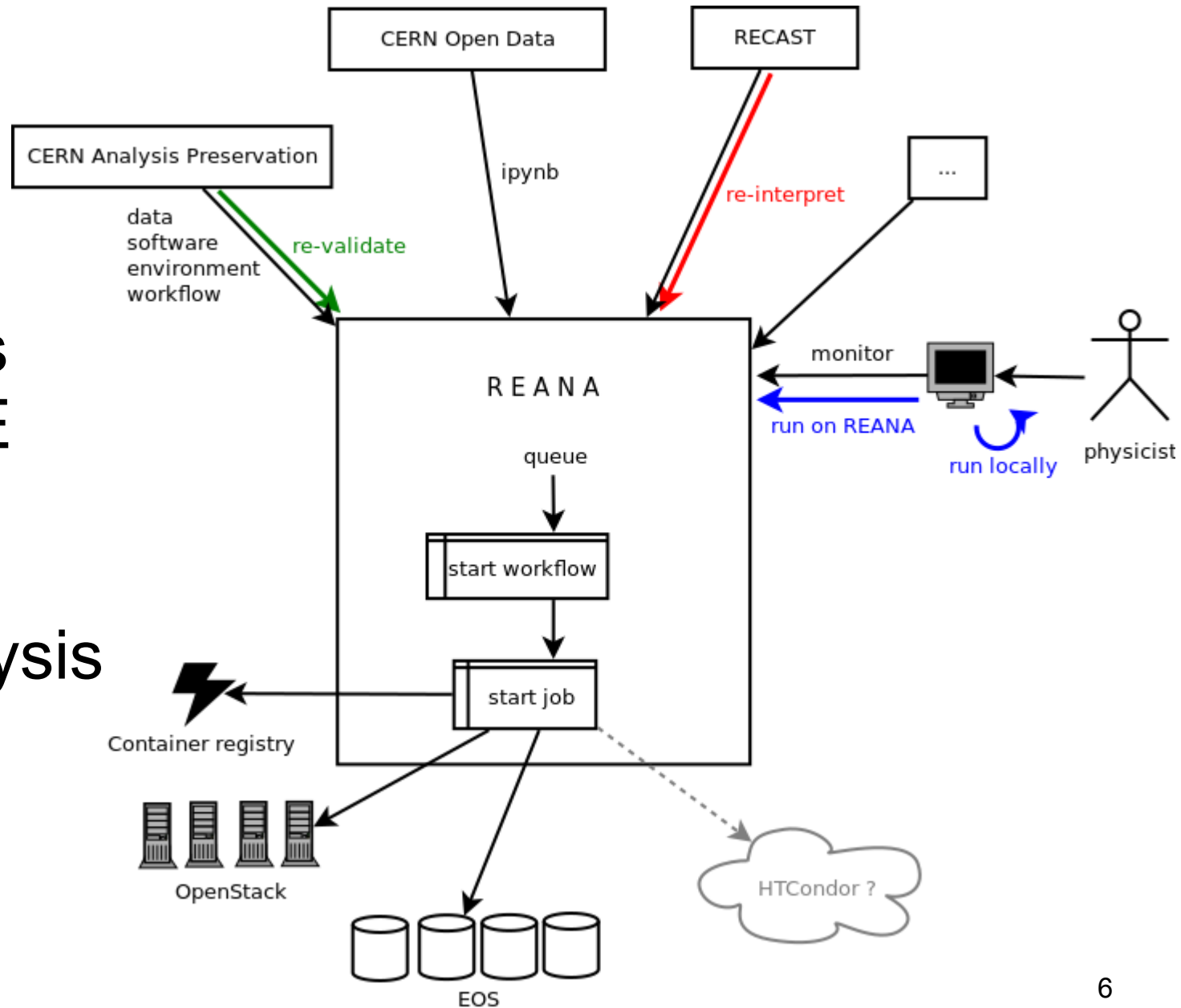
# Capture

- 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



# CAP

- <https://analysispreservation-dev.cern.ch>
- Only available from inside CERN

 ALICE  DEMO Create User Power

Home  
Shared Records  
Search

MY DEPOSITS

Shared  
Drafts

WORKING GROUPS

WG1  
WG2  
WG3

CREATE

ALICE Analysis

Hit  for shortcuts

Save as draft

Filter fields...

ANALYSIS TITLE ▶ N/A  
MAIN ANALYSIS ▶ N/A  
MC ANALYSIS ▶ N/A

Analysis Title  
E.g 2+1 correlations

MAIN ANALYSIS

Train ID  
E.g 1

Run ID  
E.g 120

Configuration Files  
E.g PWGZZ/Devel\_1/120\_20160219-2029/config

Wagon Names  
E.g TwoPlusOneCorrelation

Wagon Paths  
E.g PWGCF/Correlations/macros/twoplusone/AddTaskTwoPlusOne.C

Dataset  
E.g LHC11h\_AOD145\_nanoAOD

Reference Production  
E.g Derived Data: Devel\_1 (1), run 106 (26716)

Dataset AOD  
E.g nano AOD

# How to work with CAP

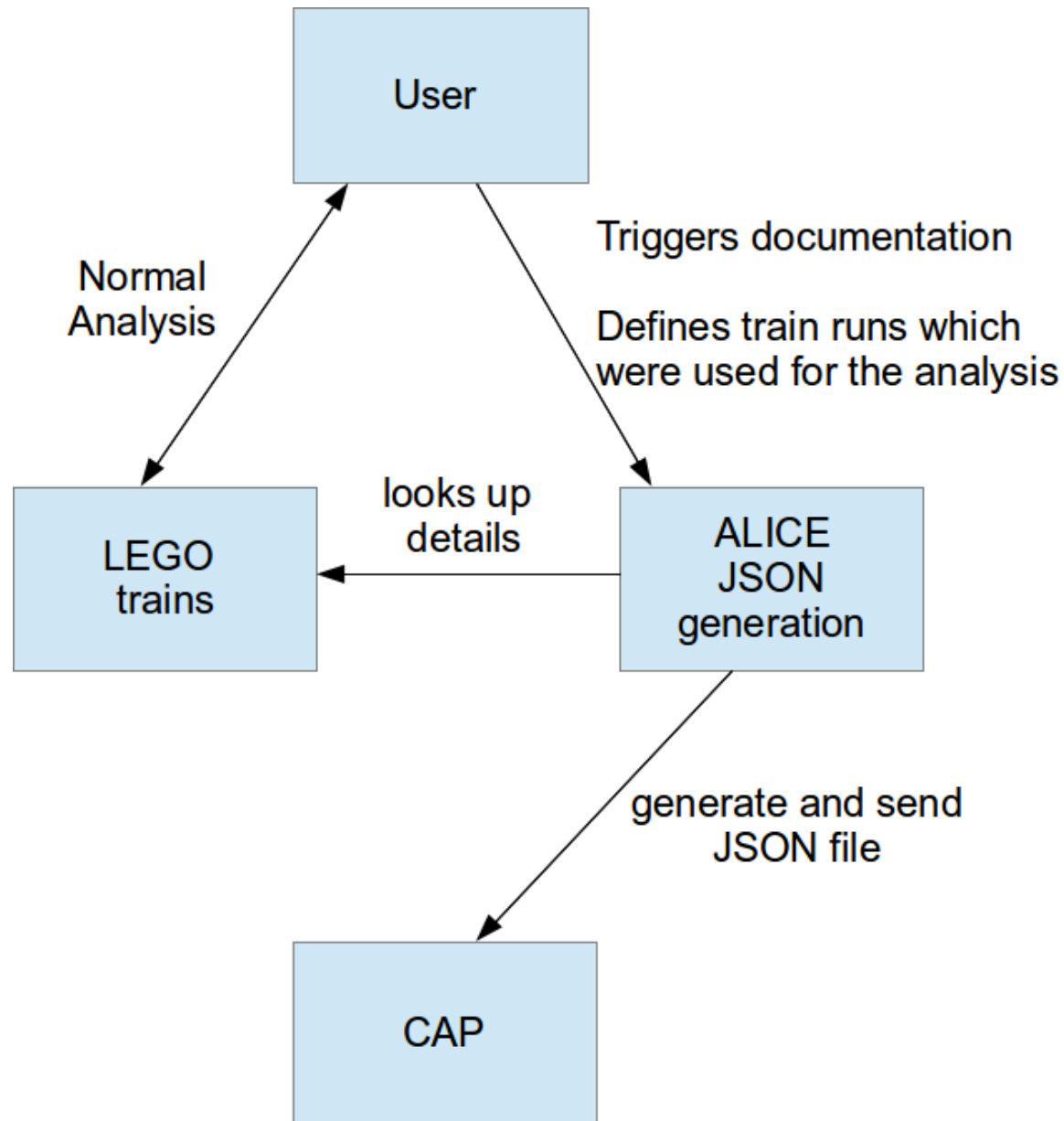
- Schema for the analysis preservation has to be predefined  
→ we suggest changes to the CAP developers and they do it
- every ALICE member can create a new analysis on CAP by
  - adding all information manually on the CAP web page
  - identifying an analysis on the trains
    - we create a JSON file and send it to CAP
    - CAP accesses our database and fills the missing fields automatically
- Work on CAP entry with multiple people (e-groups)
- Share a finished CAP entry with the whole collaboration



# How to work with CAP

- Schema for the analysis preservation has to be predefined  
→ we suggest changes to the CAP developers and they do it
- every ALICE member can create a new analysis on CAP by
  - adding all information manually on the CAP web page
  - identifying an analysis on the trains  
→ we create a JSON file and send it to CAP
  - CAP accesses our database and fills the missing fields automatically
- Work on CAP entry with multiple people (e-groups)
- Share a finished CAP entry with the whole collaboration

# JSON file generation



# JSON file from the LEGO trains

JSON file for

from train

MC analysis from train

Analysis title

run

run

```
{
  "title": "",
  "main analysis": {
    "train id": 1,
    "run id": 120,
    "configuration files": "http://alitrain.cern.ch/train-workdir/PWGZZ/Devel_1/120_20160219-2029/config",
    "wagon names": "TwoPlusOneCorrelation",
    "wagon paths": "$ALICE_PHYSICS/PWGCF/Correlations/macros/twoplusone/AddTaskTwoPlusOne.C",
    "dataset": "LHC11h_AOD145_nanoAOD",
    "reference production": "Derived Data: Devel_1 (1), run 106 (26716)",
    "dataset AOD": "nano AOD",
    "run numbers": [{}],
    "AliPhysics": "AliPhysics::vAN-20160219-1",
    "derived dataset": {
      "train id": 1,
      "run id": 106,
      "configuration files": "http://alitrain.cern.ch/train-workdir/PWGZZ/Devel_1/106_20160120-1236/config",
      "wagon names": "NanoAODFilter_Data_2plus1",
      "wagon paths": "$ALICE_PHYSICS/PWG/DevNanoAOD/AddTaskNanoAODFilter.C",
      "dataset": "LHC11h_AOD145_2",
      "reference production": "FILTER_Pb-Pb_145_LHC11h",
      "dataset AOD": "AOD production",
      "run numbers": [167915, 168115, 168460, 169035, 169238, 169859, 170228, 167920, 168310, 168464, 169091, 169411, 169923, 170230, 167985, 168311, 168467, 169094, 169415, 170027, 170268, 167987, 168322, 168511, 169138, 169417, 170081, 170269, 167988, 168325, 168512, 169144, 169835, 170155, 170270, 168069, 168341, 168514, 169145, 169837, 170159, 170306, 168076, 168342, 168777, 169148, 169838, 170163, 170308, 168105, 168361, 168826, 169156, 169846, 170193, 170309, 168107, 168362, 168988, 169160, 169855, 170203, 168108, 168458, 168992, 169167, 169858, 170204], {170593, 170572, 170388, 170387, 170315, 170313, 170312, 170311, 170309, 170308, 170306, 170270, 170269, 170268, 170230, 170228, 170207, 170204, 170193, 170163, 170159, 170155, 170091, 170089, 170088, 170085, 170084, 170083, 170081, 170040, 170027, 169965, 169923, 169859, 169858, 169855, 169846, 169838, 169837, 169835, 169591, 169590, 169588, 169587, 169586, 169557, 169555, 169554, 169553, 169550, 169515, 169512, 169506, 169504, 169498, 169475, 169420, 169419, 169418, 169417, 169415, 169411, 169238, 169167, 169160, 169156, 169148, 169145, 169144, 169138, 169099, 169094, 169091, 169045, 169044, 169040, 169035, 168992, 168988, 168826, 168777, 168514, 168512, 168511, 168467, 168464, 168460, 168458, 168362, 168361, 168342, 168341, 168325, 168322, 168311, 168310, 168115, 168108, 168107, 168105, 168076, 168069, 167988, 167987, 167985, 167920, 167915}, {167902, 167903, 167915, 167920, 167987, 167988, 168066, 168068, 168069, 168076, 168104, 168107, 168108, 168115, 168212, 168310, 168311, 168322, 168325, 168341, 168342, 168361, 168362, 168458, 168460, 168461, 168464, 168467, 168511, 168512, 168514, 168777, 168826, 168984, 168988, 168992, 169035, 169091, 169094, 169138, 169143, 169144, 169145, 169148, 169156, 169160, 169167, 169238, 169411, 169415, 169417, 169835, 169837, 169838, 169846, 169855, 169858, 169859, 169923, 169956, 170027, 170036, 170081}, {170040, 170083, 170084, 170085, 170088, 170089, 170091, 170155, 170159, 170163, 170193, 170203, 170204, 170207, 170228, 170230, 170268, 170269, 170270, 170306, 170308, 170309, 170311, 170312, 170313, 170315, 170387, 170388, 170572, 170593}, {169040, 169044, 169045, 169099, 169418, 169420, 169475, 169498, 169504, 169506, 169512, 169515, 169550, 169553, 169554, 169555, 169557, 169586, 169587, 169588, 169590, 169591, 170207}, {170311, 170312, 170313, 170315, 170387, 170388, 170572, 170593}],
      "AliPhysics": "AliPhysics::vAN-20160119-1",
    }
  }
}
```

# JSON file from the LEGO trains

- Send JSON file to the CAP servers and create entry
- CAP entry generation is triggered from the LEGO trains
- Not all fields can be filled automatically from the trains
  - Add information manually on CAP
  - Implement fields on a dedicated page within the LEGO trains and send them to CAP
- Have to implement the API to create and send the JSON file

# Why using CAP?

- dedicated long term preservation service
- CAP allows searching and grouping of analyses
- look up details in the LEGO trains
  - low amount of work by the user
- Rerunning an analysis by a third party (REANA)

# REANA

- REusable ANALyses
- Possibility to rerun an analyses without the ALICE infrastructure
- Workflow can be described in JSON
  - Use documentation from CAP?
- Compose analyses out of modules
  - rerun the train analysis
  - run macros to analyze the train results
  - Create plots from the analysis
- Support for multiple workflow engines
- Integrates CVMFS
- Runs on Docker containers

# REANA

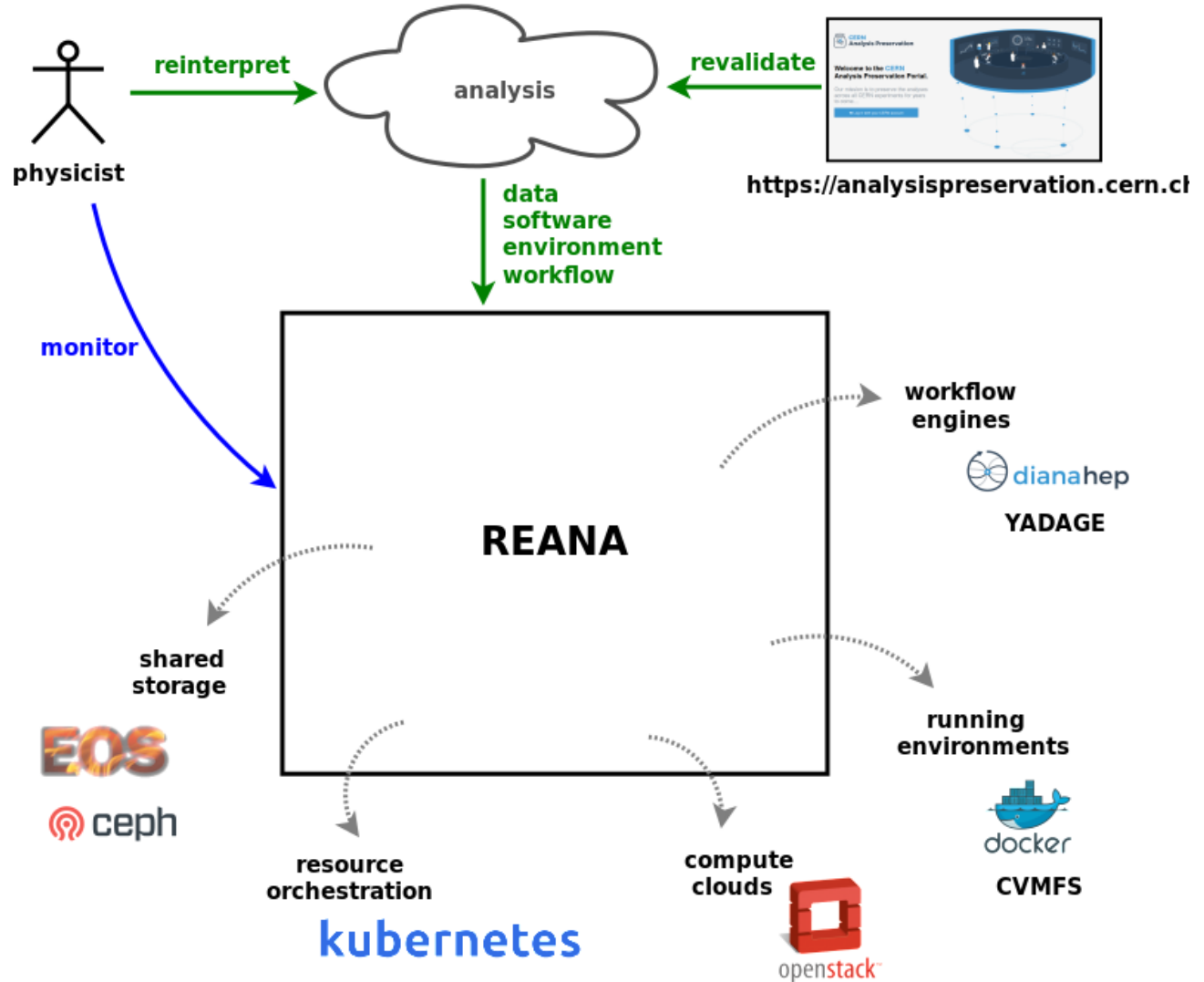
To use REANA provide:

- Data
- Software
- Environment
- Workflow

Can use this for:

- The train run
- Plot production after the train run

A REANA test run is planned with published data



# Summary & Outlook

- Introduced the CERN Analysis Preservation
  - Tool for long term analysis preservation
  - Entries can be created from the LEGO trains
    - low amount of work for the user
- Introduced REANA
  - Rerun analysis without ALICE infrastructure
  - Create approved plots and preserve the procedure
- Decide on the preservation schema
- Implement the API to send the JSON file to CAP
- Test run on REANA



**BACKUP**

# Information for the CAP Entry

- Train id, run id
- Configuration files
- Train wagon path (relative to AliPhysics)
- Code version
- Dataset
  - Dataset type
  - Run numbers
  - Reference train run (in case it is nano AOD)