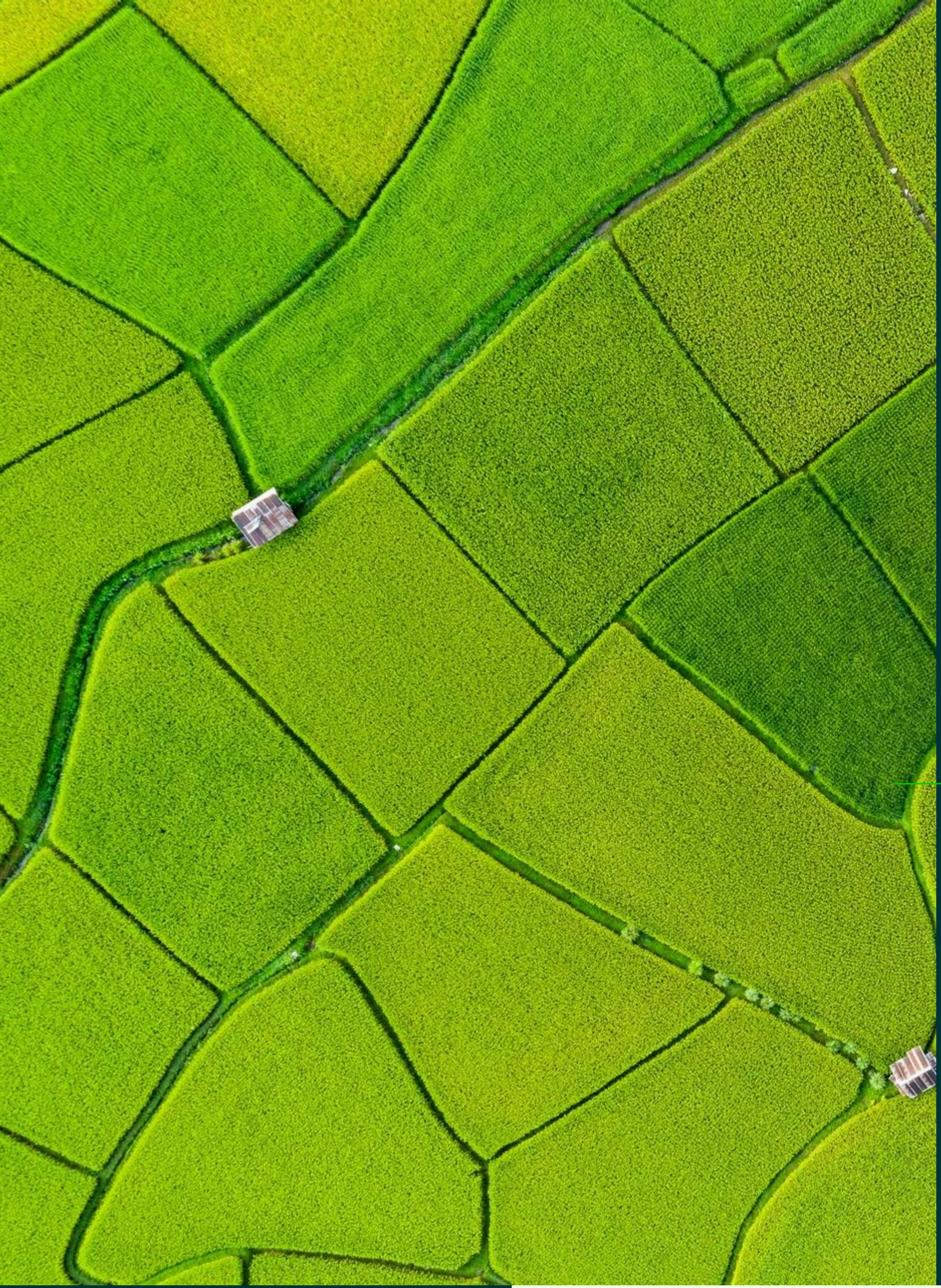


AUDITOR experience @ INFN-CNAF

27th May 2025

Speaker: Alessandro Pascolini

alessandro.pascolini@cnaf.infn.it

ecosister

Finanziato dall'Unione europea
NextGenerationEU

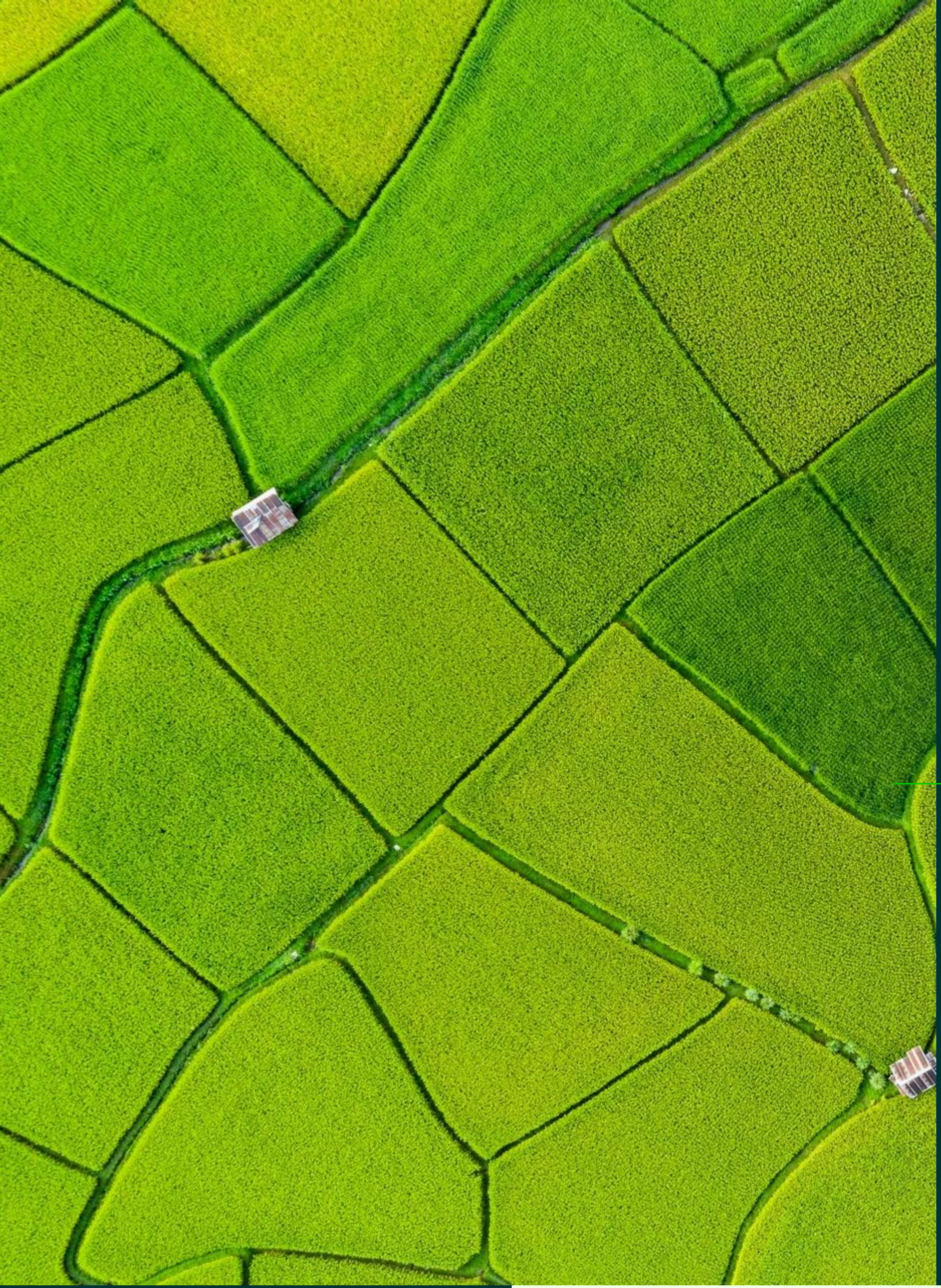
Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER
ATTRATTIVITÀ RICERCA TERRITORIO

INFN
CNAF

Overview



Introduction

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER
ATTRATTIVITÀ RICERCA TERRITORIO

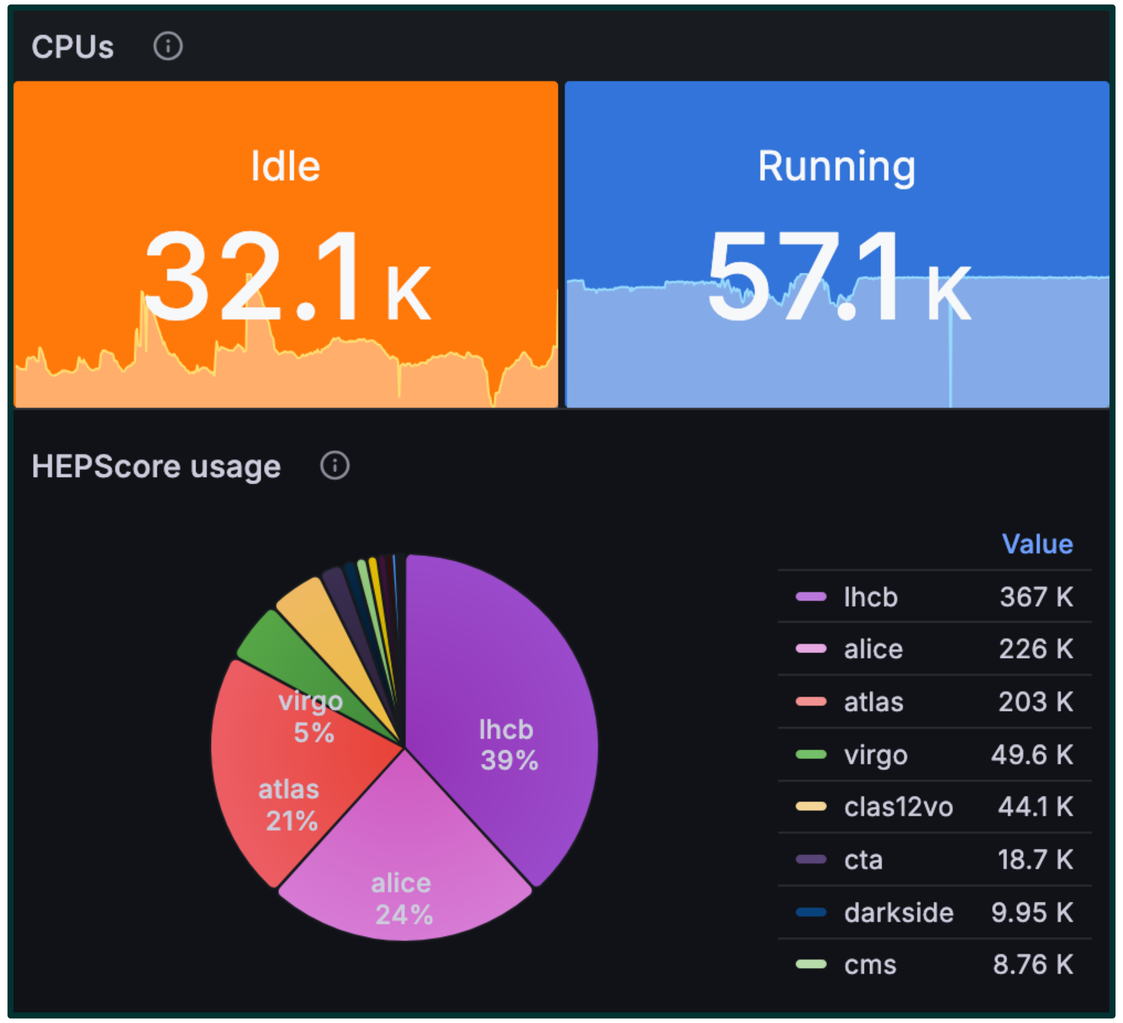
INFN
CNAF

INFN-CNAF

- CNAF hosts **2 WLCG sites:**
  - **INFN-T1** → the Italian Tier1
  - **INFN-CNAF-LHCB** → the only Italian LHCb Tier2

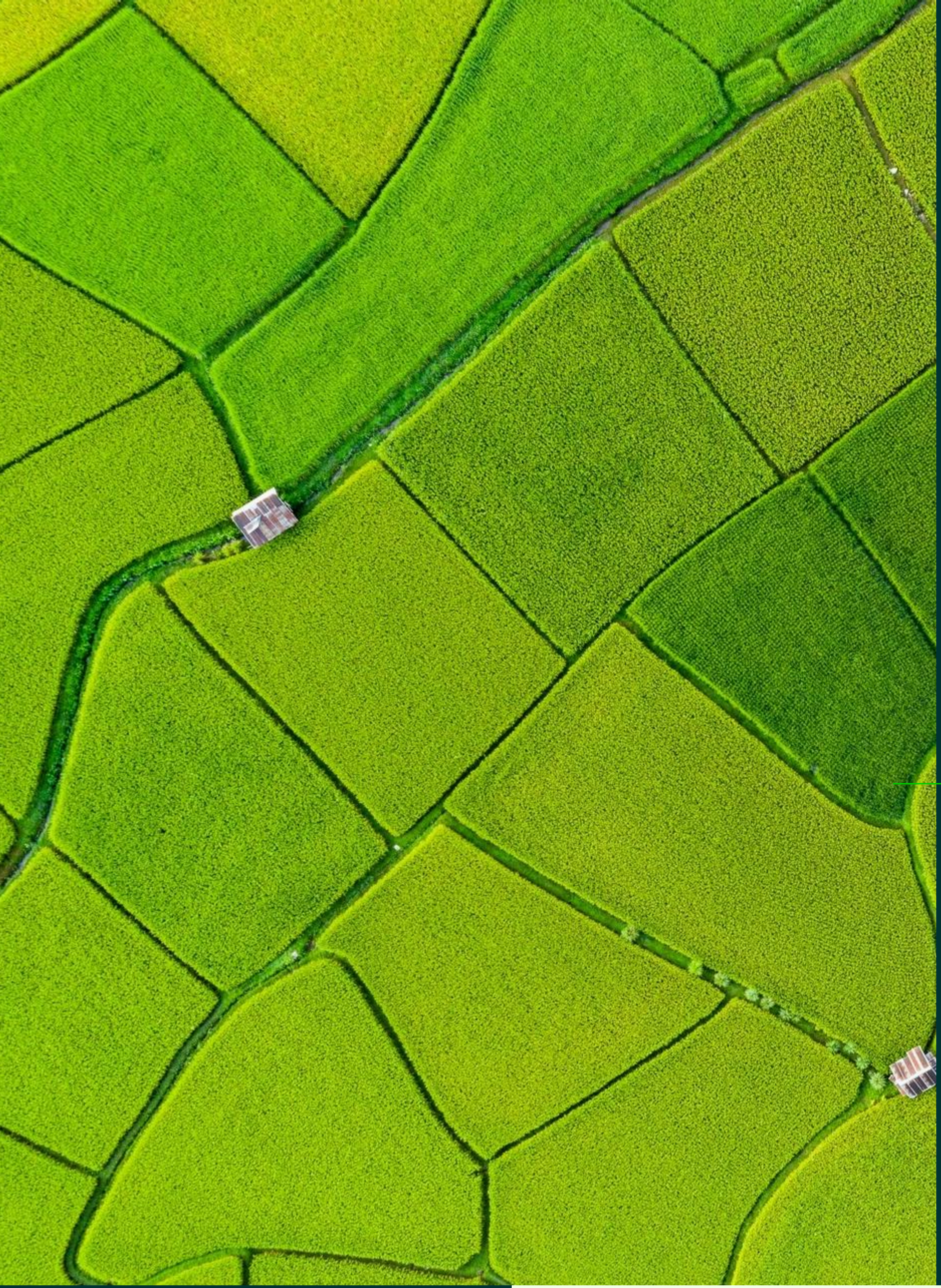
- CPU resources provisioned via a **multi VO HTCondor** batch system:
  - **6 Tier1 CEs** → ce0[1-6]-htc.cr.cnaf.infn.it
  - **1 Tier2 CE** → ce01-lhcb.cr.cnaf.infn.it
  - **1 local Schedd**
  - **~60k available cores**

- More than **50 pledged VOs** to be accounted
- Total pledge: **~850 kHS23**



Accounting Issue!

HTCondor doesn't provide native accounting fit for WLCG



Current Accounting infrastructure @CNAF

ecosister

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

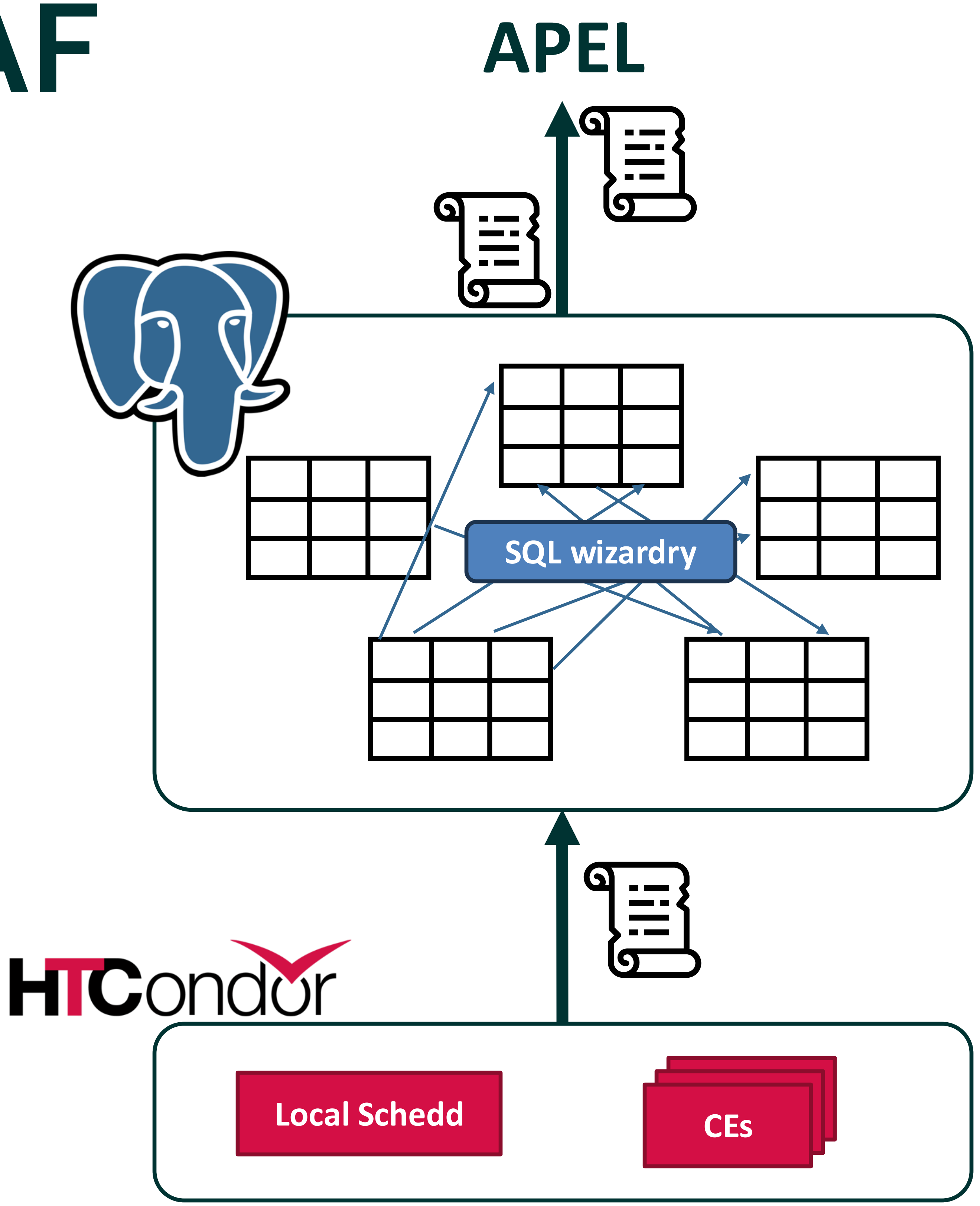
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER
ATTRATTIVITÀ RICERCA TERRITORIO

INFN
CNAF

Accounting setup @CNAF

- The legacy infrastructure **lacks documentation**
- The current setup is heavily based on **old, custom scripts** to:
  - **Collect** accounting data from HTCondor Schedd/CEs
  - **Store** data into a Postgres DB
  - **Query** accounting data to get an **APEL-like format**
  - **Send** accounting data to EGI



Accounting setup @CNAF

This accounting workflow has many issues:

- **No standard solution**
- **Old and messy** DB management
- Many components **may randomly break**
- **Prone to errors**
- **Difficult to maintain**

It all works until it doesn't anymore...

Resource Centre INFN-T1 — Elapsed time * Number of Processors (hours) by VO and Month (All VOs)

VO	Apr 2024	May 2024	Jun 2024	Jul 2024	Aug 2024	Sep 2024	Oct 2024	No
039e2956-2e56-44c6-987c-25b240430d89	0	0	0	0	0	0	0	
08ca855e-d715-410e-a6ff-ad77306e1763	150	550	426	373	0	0	0	
09c53d33-7e74-48d3-9758-0eac675fba48	0	0	0	0	0	4	6	
0b25e12f-48e3-4891-a791-25aeea28128b	0	0	0	0	0	0	0	
0f6e9a97-88ad-4c27-ad02-b67da7ac6f73	0	0	0	0	0	0	0	
130d098c-3d90-4b26-b695-6e0f7534df4e	0	0	0	0	0	0	0	
14a55eee-8e34-484d-9d20-4d71e783a581	0	0	0	0	0	0	0	
16d13b4b-da7f-40bf-8490-b43bca42e905	0	0	0	0	0	0	0	
1ad9770e-353a-4b1a-8529-e2135fca8081	0	0	0	0	0	0	0	
2238c828-1040-4246-9411-89864218c75c	0	0	0	0	0	0	4	0
22a0eb75-33cc-4d4f-b087-5236e509ac9c	0	0	0	0	0	0	0	0
2c5255ba-9480-4815-aea0-88159f6602b7	0	0	0	0	0	0	0	0
447f3a36-a5a3-46e0-8c11-c1d37a8212cd	0	0	0	0	0	0	0	0

```
acct=# \d
                    List of relations
 Schema |          Name           |   Type   |  Owner
--------+-------------------------+----------+----------
 public | acct_valid_202304       | table    | acct
 public | aggrcbjobs              | table    | acct
 public | allwn                   | table    | acct
 public | allwn_id_seq            | sequence | acct
 public | apelgridjob             | view     | acct
 public | apelhtjob               | view     | acct
 public | apelnew                 | table    | acct
 public | atlacc                  | table    | acct
 public | blahjob                 | table    | acct
 public | blahjob_2015            | table    | acct
 public | blahjob_2016            | table    | acct
 public | blahjob_2017            | table    | acct
 public | blahjob_2018            | table    | acct
 public | blahjob_hist            | table    | acct
 public | blahjob_id_seq          | sequence | acct
 public | cancellami              | table    | acct
 public | cinecahpc               | table    | acct
 public | cinecahpc_id_seq        | sequence | acct
 public | daily_aggr              | table    | accguy
 public | daily_aggr_id_seq       | sequence | accguy
 public | daily_summary           | table    | accguy
 public | daily_summary_id_seq    | sequence | accguy
 public | dayjob                  | table    | acct
 public | dnsub                   | table    | acct
 public | dnsub_id_seq            | sequence | acct
 public | duphunt                 | table    | acct
 public | duphunt_id_seq          | sequence | acct
 public | efficiency              | table    | postgres
 public | gridjob                 | table    | acct
 public | gridjob_2015            | table    | acct
 public | gridjob_2016            | table    | acct
 public | gridjob_2017            | table    | acct
 public | gridjob_2018            | table    | acct
 public | gridjob_hist            | table    | acct
 public | gridjob_sent            | table    | acct
 public | gridjob_test            | table    | acct
 public | groupexp                | table    | acct
 public | hlrtid_seq              | sequence | acct
 public | hostjobs                | table    | acct
 public | htcapeljob              | view     | acct
 public | htj_feb6                | table    | acct
 public | htjob                   | table    | acct
```

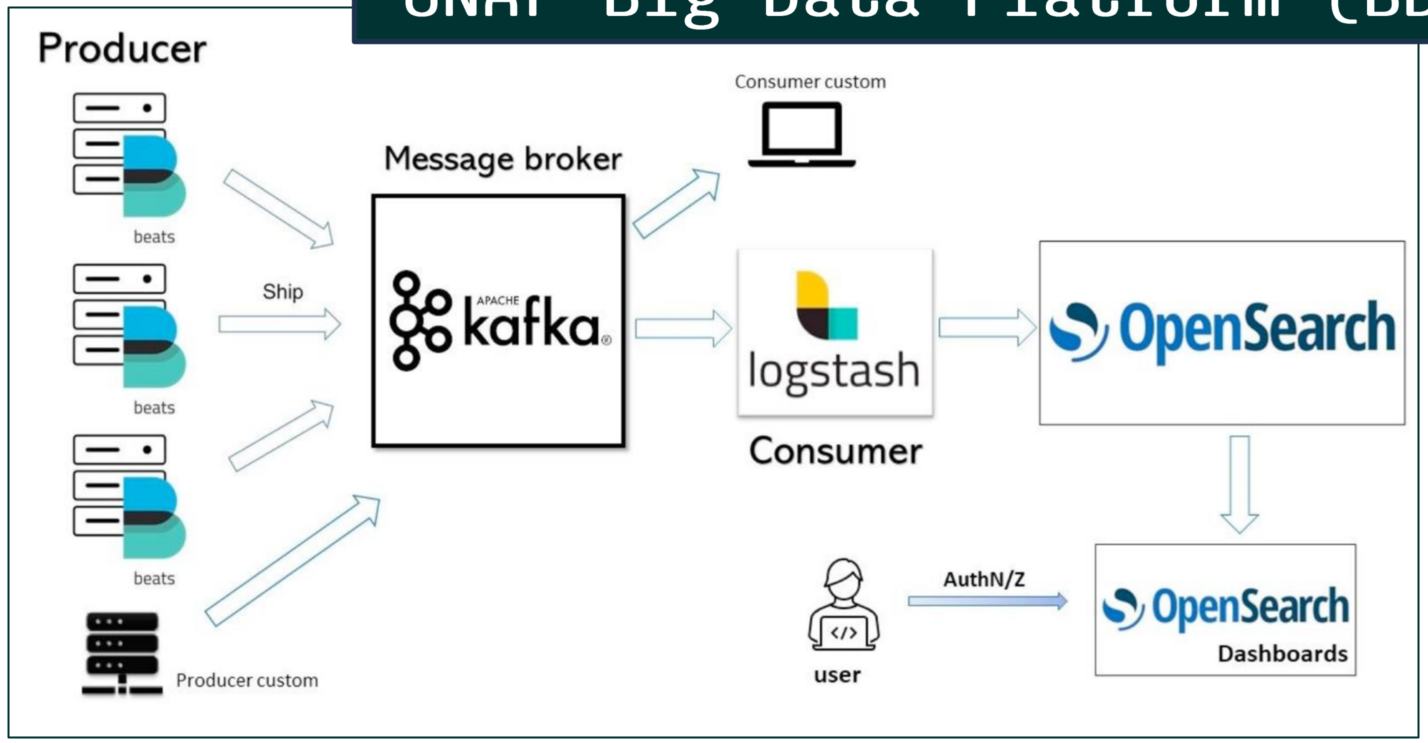
Alternatives

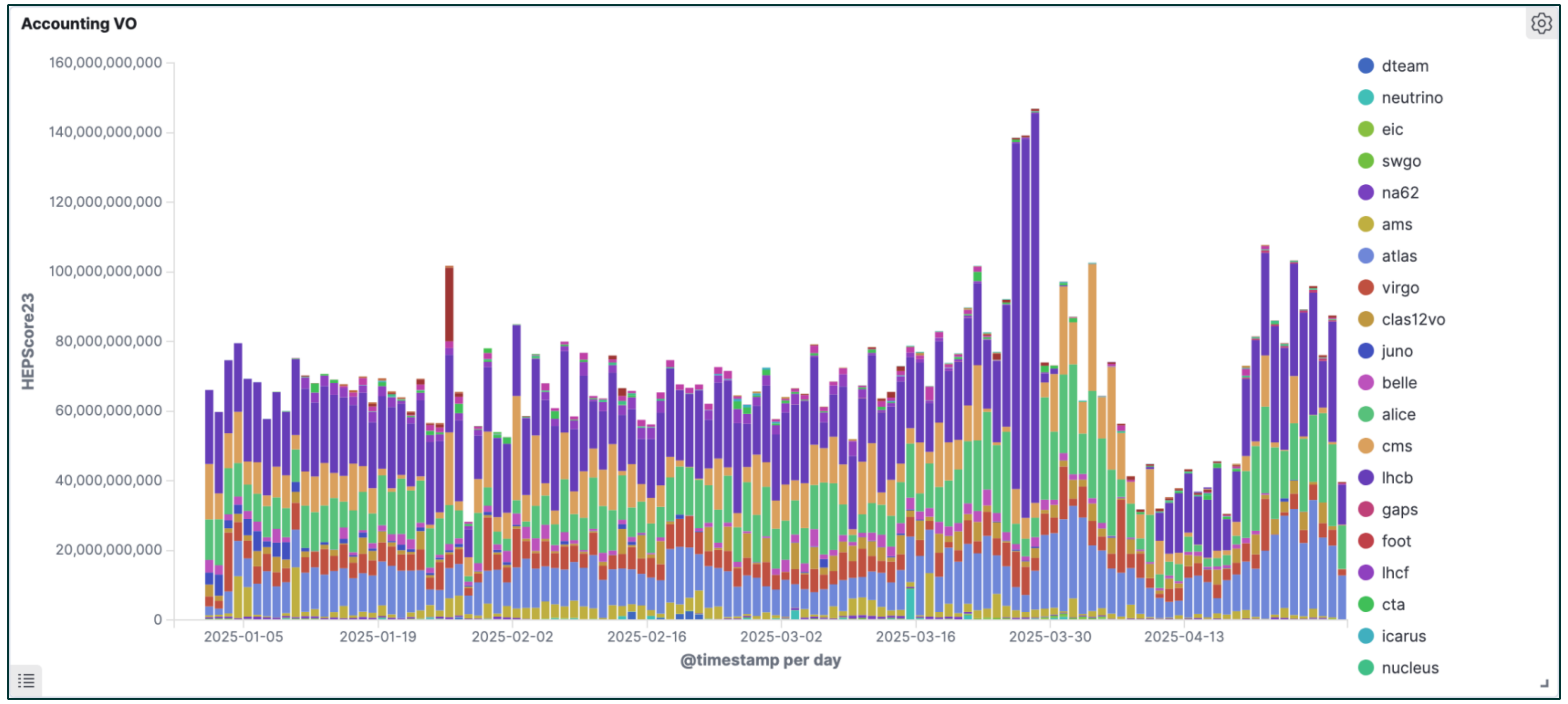
In the past years we started working on new accounting based on standard solutions:

- HTCondor provides **condor_adstash** to send job history to Elasticsearch
  → may store **more info than needed** for our use-case
- Using the ELK stack to filter only accounting information
  → CNAF already has the infrastructure

Promising approach but needs lots of manpower and time

CNAF Big Data Platform (BDP)

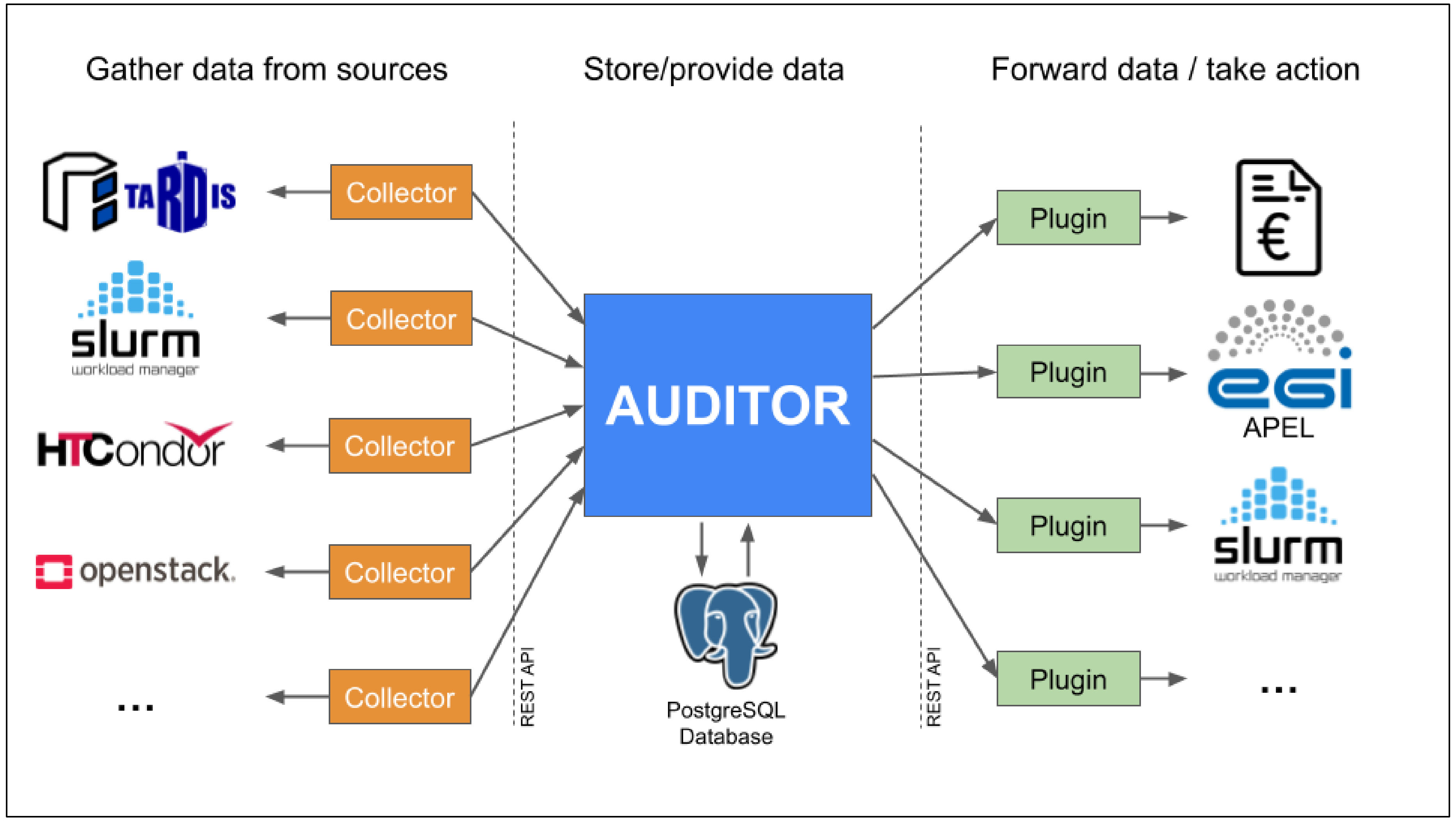


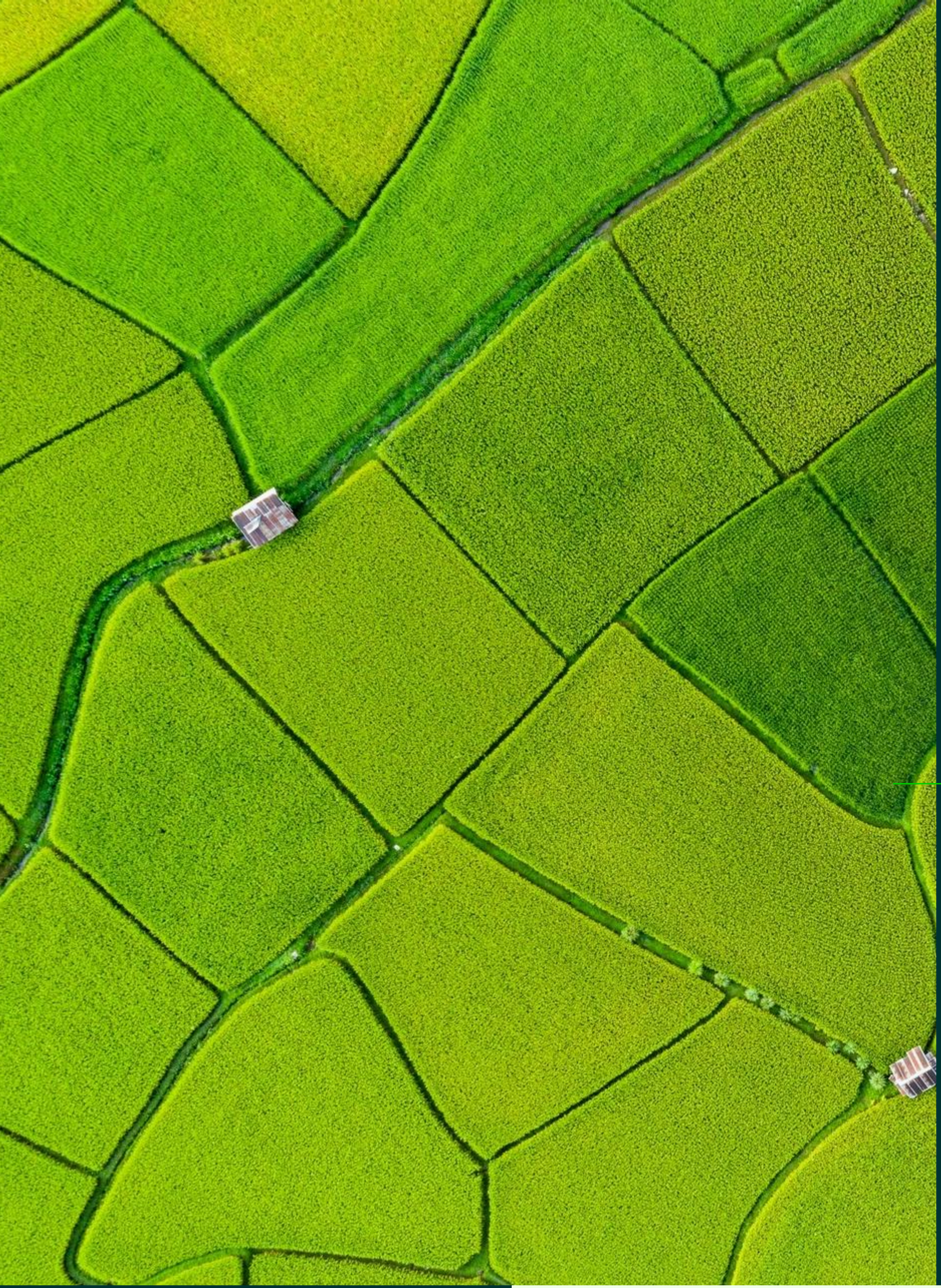


AUDITOR

AUDITOR provides a **valid alternative** to our current accounting infrastructure:

- Stores data in PostgreSQL DB
  → **we already have an instance in place** from the current system
- Provides **collectors for HTCondor** jobs
- Can **aggregate and send** data to APEL
- Easily configurable
  → **few yaml files** to have a **working setup**
- **Provides API** to interact with the AUDITOR Server





AUDITOR

ecosister

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER
ATTRATTIVITÀ RICERCA TERRITORIO

INFN
CNAF

ecosister

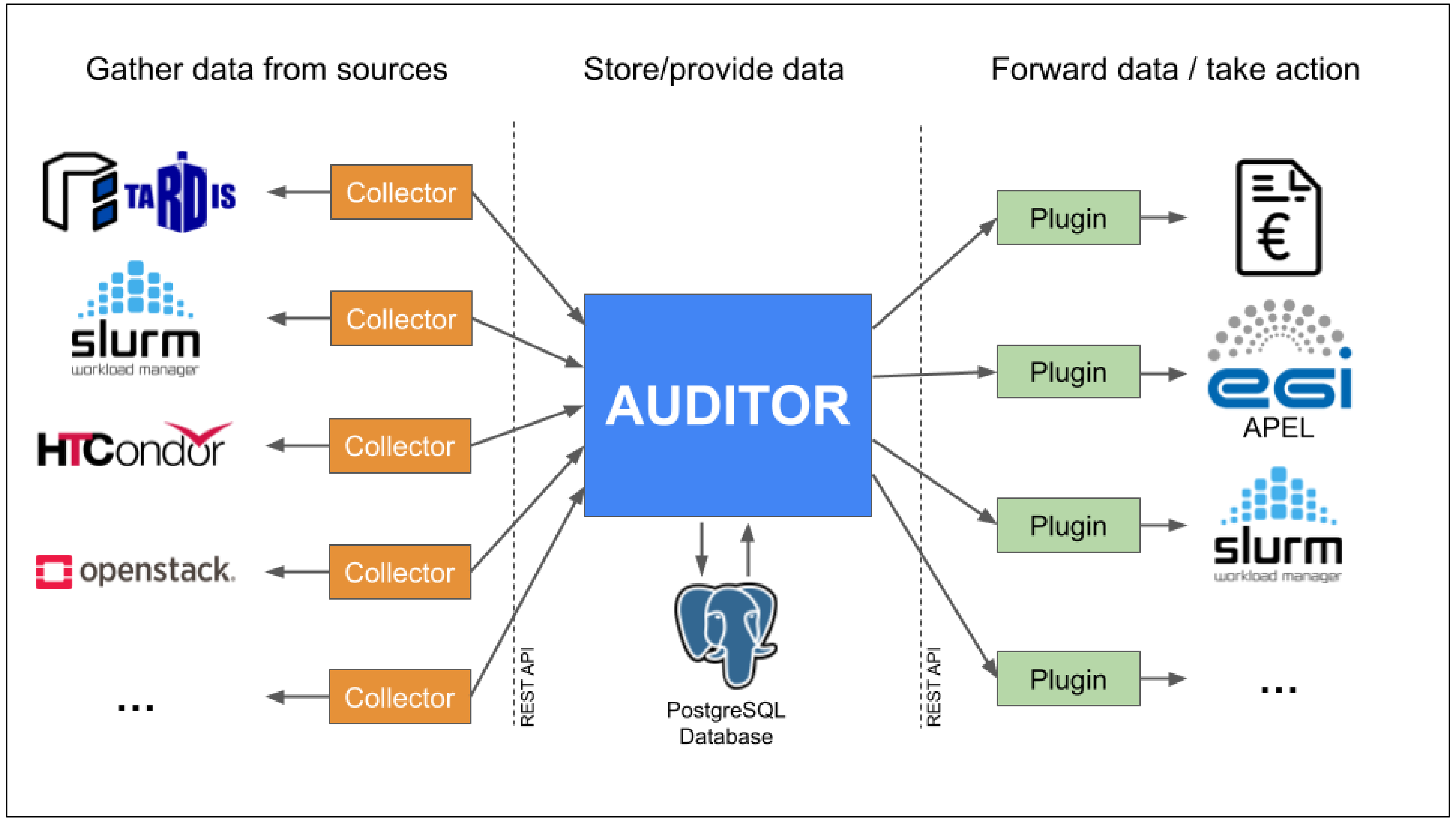
27 / 05 / 2025 - Workshop CCR

AUDITOR structure

- Collectors
  → query the resources to be accounted (e.g. HTCondor/SLURM batch-systems)

- AUDITOR Server
  → handles accounting data sent by collectors
  → interacts with the Accounting DB
  → exposes APIs

- Plugins
  → gets accounting data from the AUDITOR server
  → can send data to consumers (e.g. APEL)

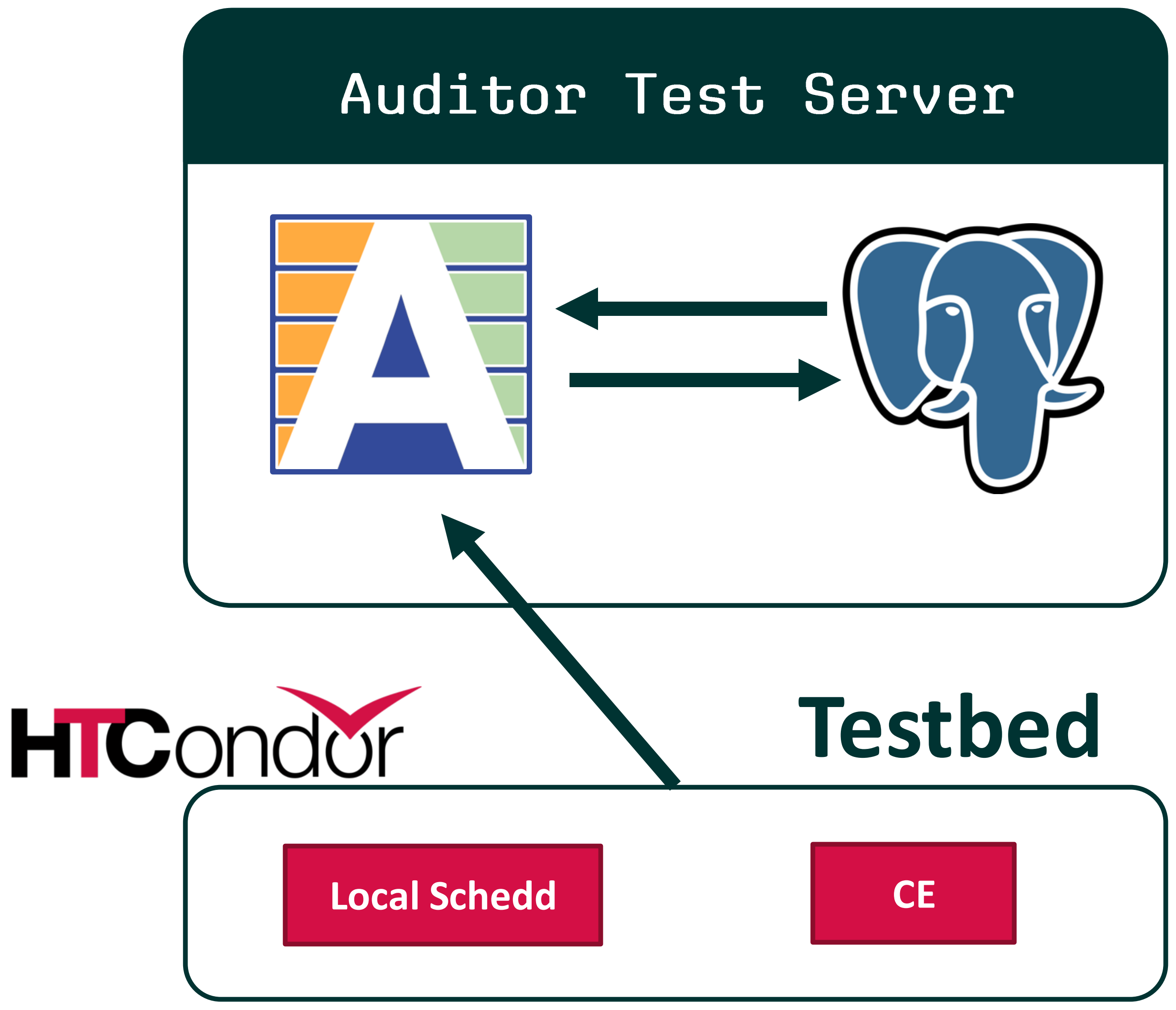


AUDITOR tests

Preliminary Setup:

- Pre-production system with local PostgreSQL
- Collectors installed on HTCondor Testbed

Needed to understand how to tune our config files and HTCondor Cluster configuration



AUDITOR tests

Tuning of HTCondor Cluster

- Each job should have some attributes to allow accounting:
  - **HEPScore23/core** ratio of the machine where the job is executed
  - **Site name**
  - **VO name**

- All this information is assigned to the job at submission time via some **HTCondor Job Routers***

Auditor HTCondor Collector configuration

- The collector needs to have references to all the job attributes previously mentioned

htcondor-collector.yaml

```yaml
meta:
  user:
    key: Owner
    matches: ^(.+)$
  group:
    key: t1_voname
    matches: ^(.*)$ # match  the first part of the Acctgroup
  submithost:
    key: "GlobalJobId"
    matches: ^(.*)#\d+.\d+#\d+$
    # As this regex contains a group,
    # the value for 'submithost' is set to the matching group.

  # For `site` the first match is used.
  site:
    - name: "site1"  # This entry
      key: "SITE"
    - name: "UNDEF"  # If no match is found, site is set to "UNDEF"


components:
  - name: "Cores"
    key: "CpusProvisioned"
    scores:
    - name: "HEPSPEC"
      key: "MATCH_t1_hs06_ratio"
    - name: "HEPscore23"
      key: "MATCH_t1_hs23_ratio"
  - name: "Memory"
    key: "MemoryProvisioned"
  - name: "UserCPU"
    key: "RemoteUserCpu"
```

The "key" field should refer to the <u>job attribute</u>

*https://htcondor.readthedocs.io/en/latest/grid-computing/job-router.html

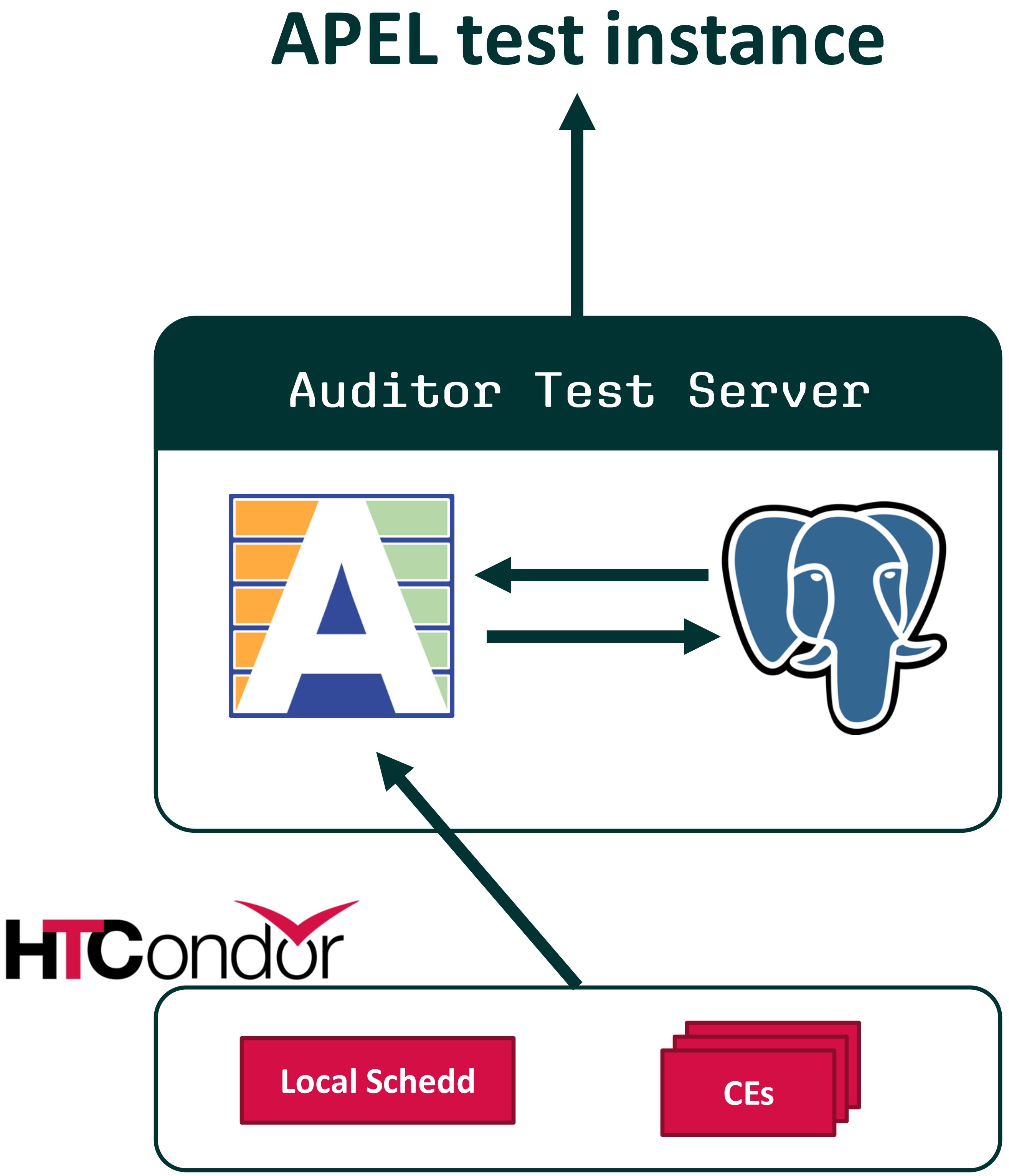
AUDITOR tests

Pre-prod Setup:

- Pre-production host with local PostgreSQL
- Collectors installed on **HTCondor production cluster**
- Sending data to APEL test infrastructure
  → currently **finalizing the configuration** of APEL sender plugin

```
[2025-05-26 19:42:52] INFO     Total numbers reported by the plugin:
Site: INFN-T1
Month: 5
Year: 2025
NumberOfJobs: 709852
WallDuration: 20234841573
NormalisedWallDuration: 367381348965
CpuDuration: 14668225115
NormalisedCpuDuration: 260990707649

 (//opt/auditor_apel_plugin/venv/lib/python3.9/site-packages/auditor_ape
[2025-05-26 19:42:52] INFO     Getting records for site INFN-CNAF-LHCB w
ckages/auditor_apel_plugin/core.py at line 45)
[2025-05-26 19:42:53] INFO     Total numbers reported by the plugin:
Site: INFN-CNAF-LHCB
Month: 5
Year: 2025
NumberOfJobs: 3713
WallDuration: 215605284
NormalisedWallDuration: 3828187766
```



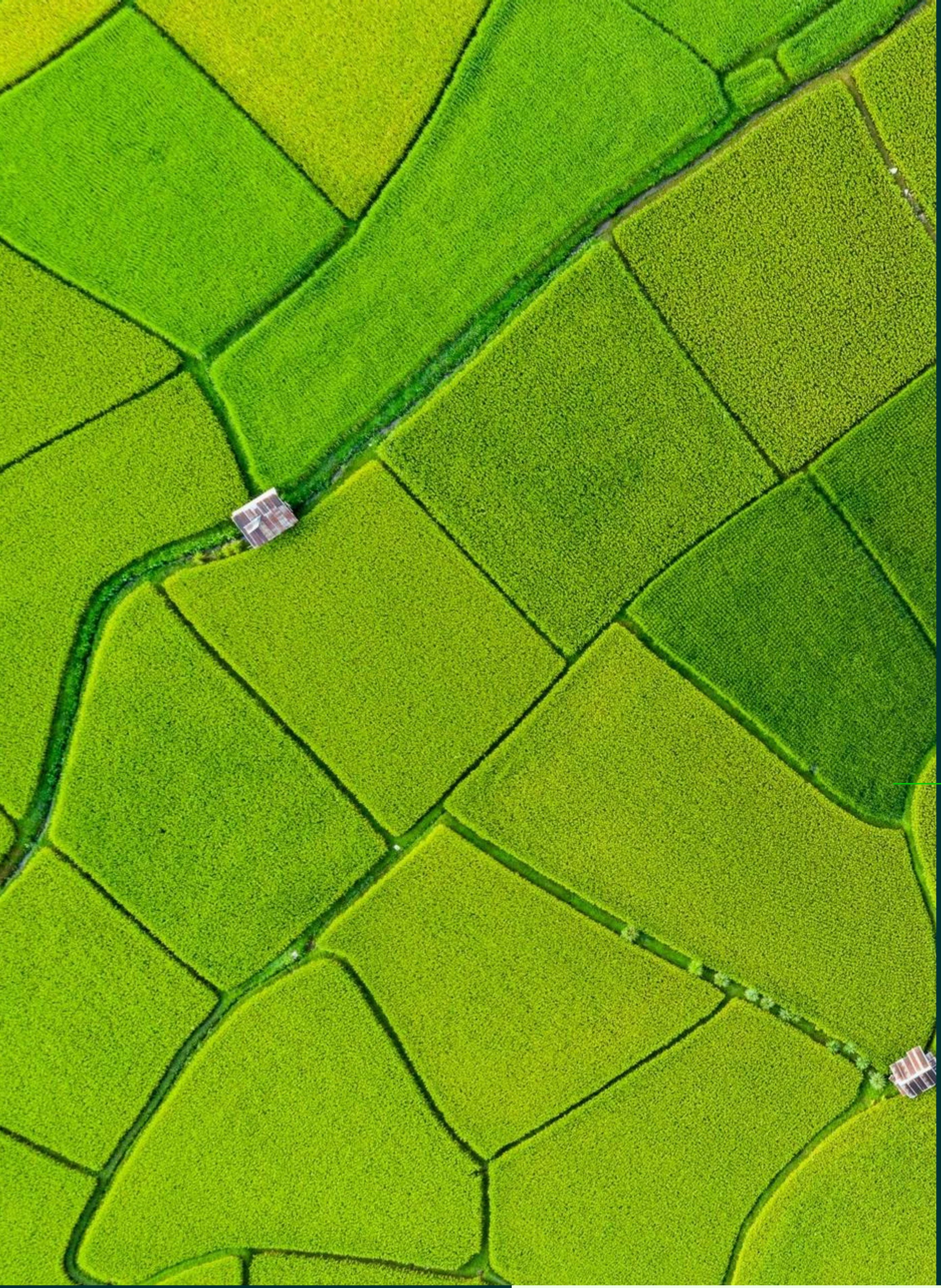
Provisioning

- All hosts at CNAF are managed via **Foreman** and **Puppet**
- To introduce the new accounting system, we need to integrate it within our provisioning system
  → write a new puppet module

Puppet module

• Supports different roles:
  • Collector
  • Server
  • Plugin (APEL)
• Manages config files via templates
• Manages systemd units for AUDITOR server

All installation is managed <u>via RPMs</u>
- Asked via AUDITOR users mailing list
- Promptly added in AUDITOR v0.8 release



AUDITOR and WLCG

ecosister

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER
ATTRATTIVITÀ RICERCA TERRITORIO

INFN
CNAF

AUDITOR in WLCG

Many sites are testing AUDITOR

- KIT
  → already in production
- CERN
- CNAF
- DESY
- Wuppertal

WLCG Workshop contributions
https://indico.cern.ch/event/1484669/timetable/

AUDITOR rpms available in WLCG repository

2:20 PM **AUDITOR. First results of the assessment by the WLCG sites** 1h 30m Auditorium P. Lehmann, Buildi...

AUDITOR (Accounting Data Handling Toolbox for Opportunistic Resources) is a flexible and extensible accounting system designed to support a wide range of use cases and infrastructures. Its integration with APEL enables it to function as a generic component within the WLCG accounting infrastructure, tracking the usage of various types of site computing resources. Several WLCG sites have evaluated AUDITOR, and the outcomes of these assessments will be reviewed and discussed.

Speakers: Alessandro Pascolini (Universita e INFN, Bologna (IT)), Alexander Raphael Kleinemuhl (Bergische Universitaet Wuppertal (DE)), Maria Alandes Pradillo (CERN), Max Fischer (Karlsruhe Institute of Technology), Michael Boehler (University of Freiburg (DE)), Michael Boehler (Deutsches Elektronen-Synchrotron (DE)), Thomas Hartmann (Deutsches Elektronen-Synchrotron (DE))

AUDITORintro.pdf

KIT input 15m

Speaker: Dr Max Kühn (Karlsruhe Institute of Technology)

KIT_MK_AUDITOR.pdf

CERN input 15m

Speaker: Maria Alandes Pradillo (CERN)

cern-auditor-pilot-wl...

CNAF input 15m

Speaker: Alessandro Pascolini

AUDITOR@CNAF.pdf

DESY input 15m

Speaker: Thomas Hartmann (Deutsches Elektronen-Synchrotron (DE))

desy-hh_auditor.pdf

Wuppertal input 15m

Speaker: Alexander Raphael Kleinemuhl (Bergische Universitaet Wuppertal (DE))

auditor_wuppertal.p...

Input from the developers and summary 15m

Speaker: Michael Boehler (University of Freiburg (DE))

2025-05-07_WLCG-...

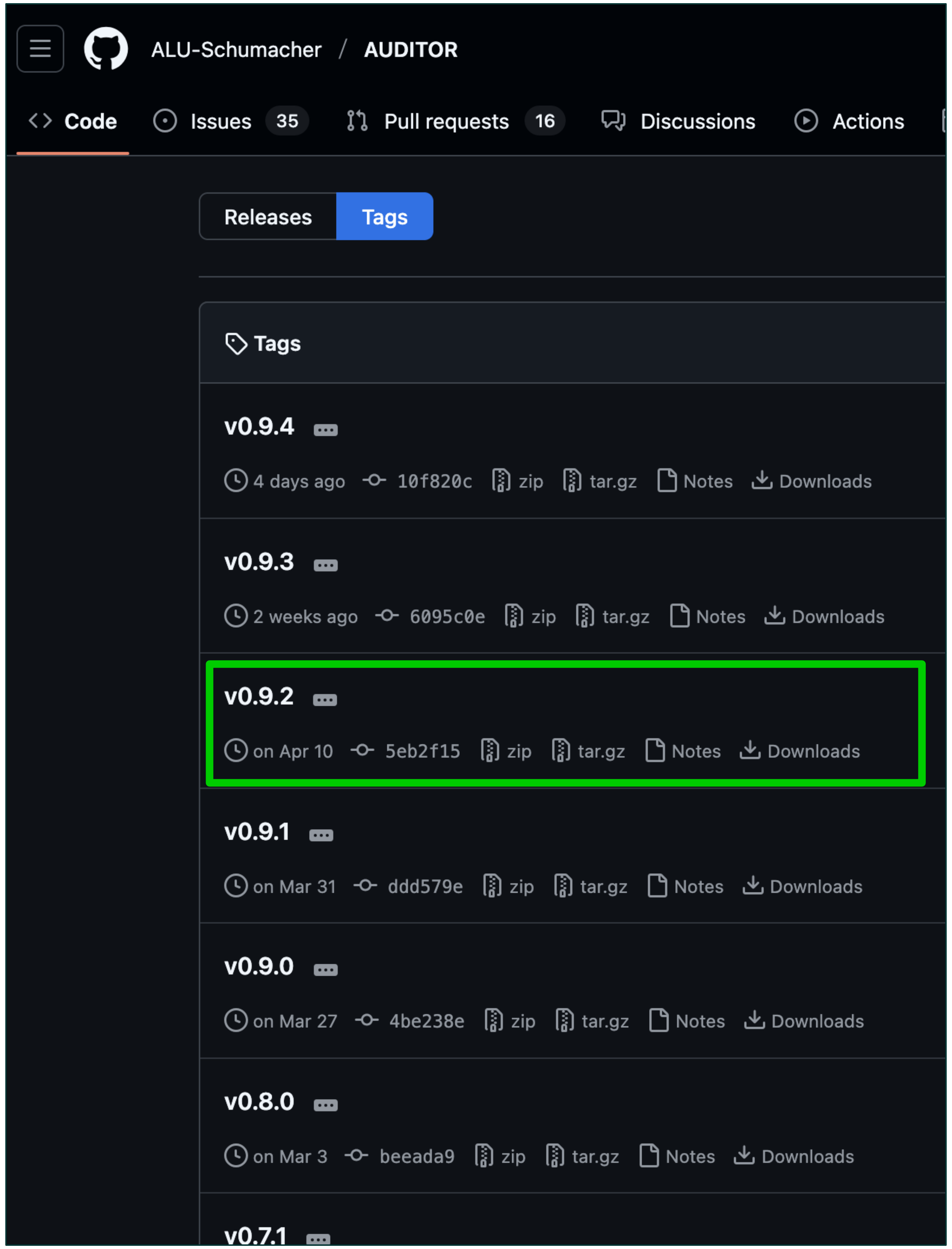
AUDITOR Activity

Currently in a fine tuning phase

- Feedback from all the sites testing the software
- Many releases to improve:
  - RPMs structure
  - DB size
  - Memory usage
  - New functionalities

All releases are fully operative
→ KIT is **still running v0.6.8** in production

Currently at CNAF we are testing v0.9.2



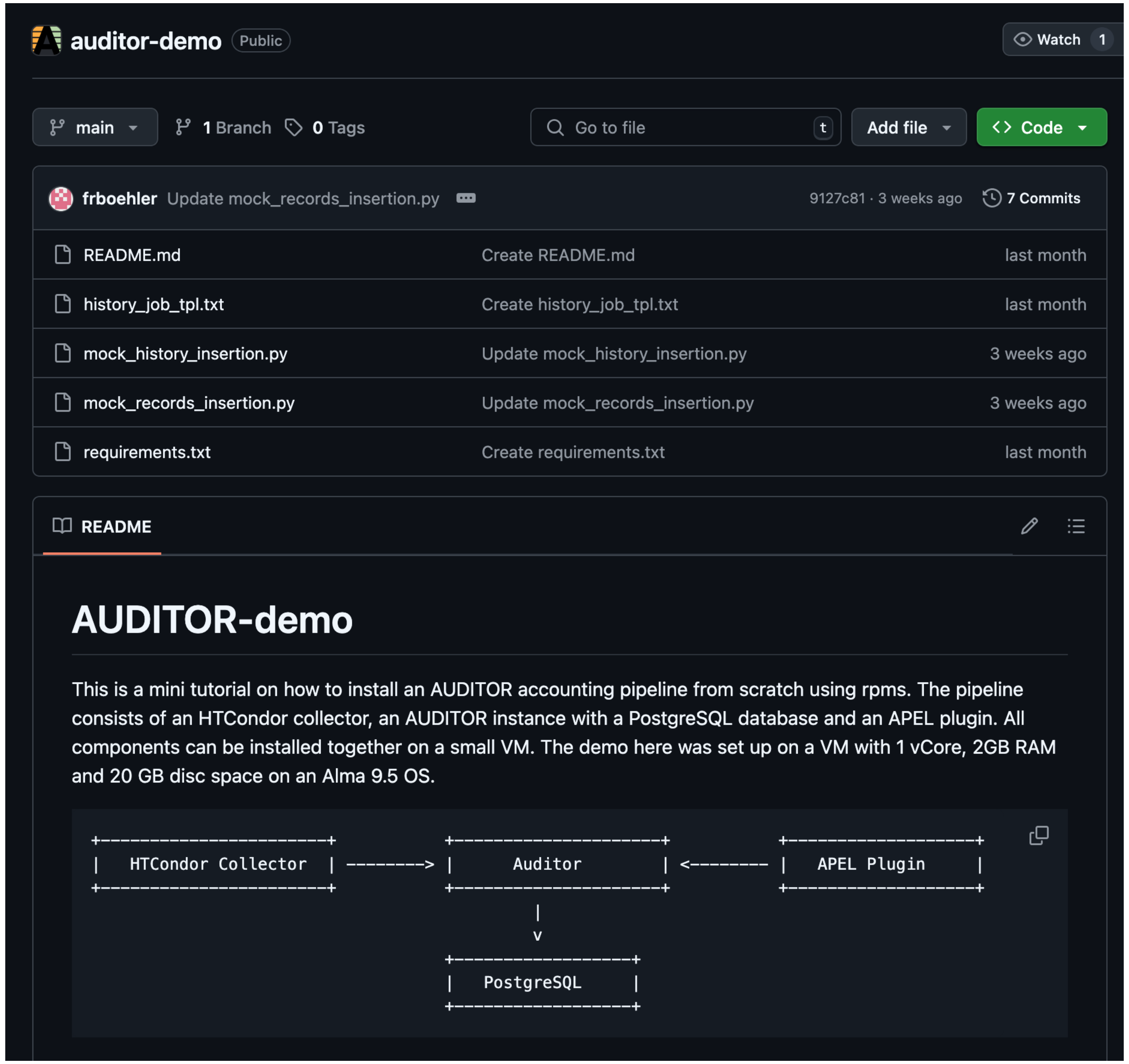
AUDITOR New Testbed Feature

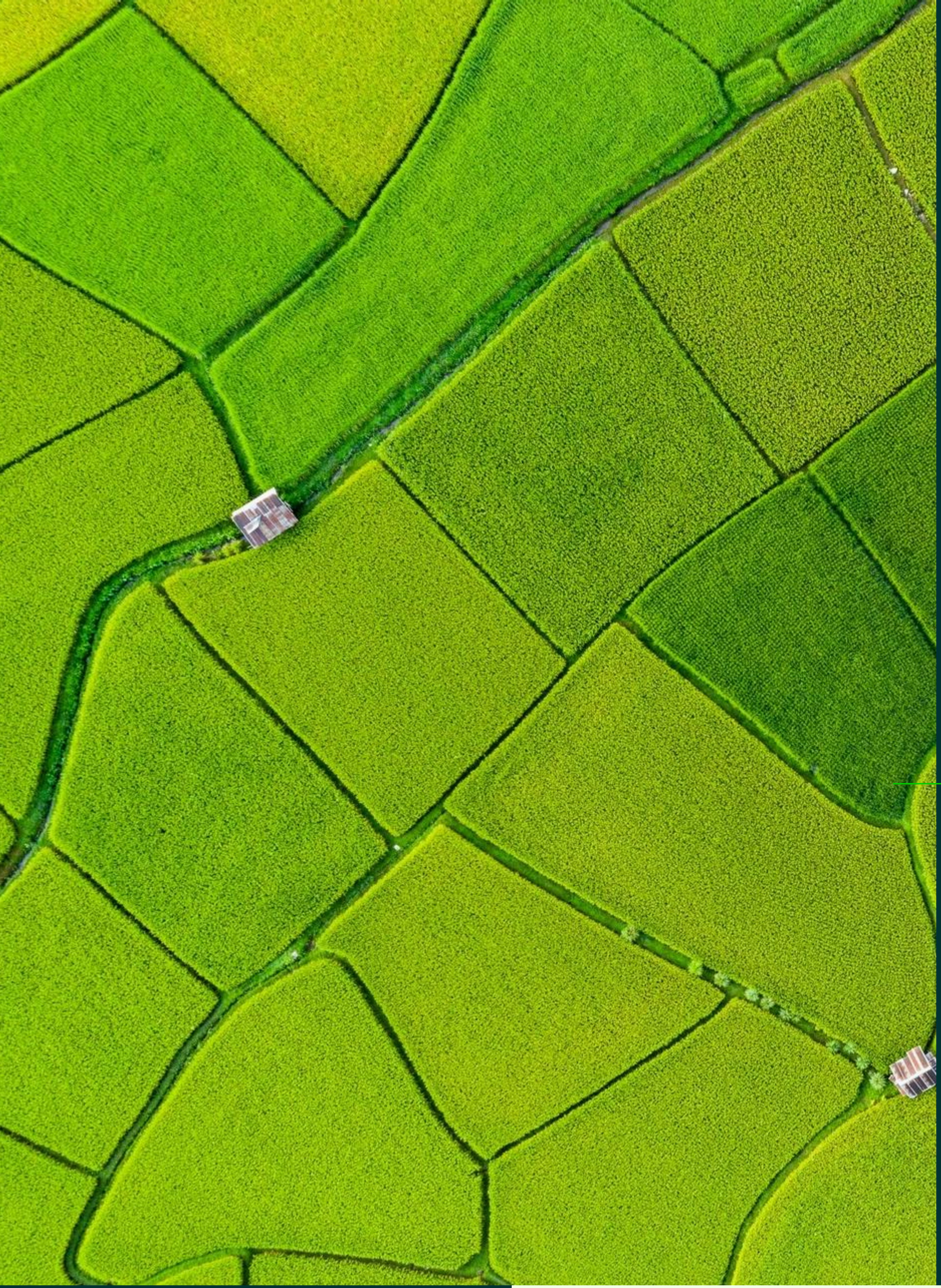
AUDITOR provides instruction to set up a testbed

- Simulates an HTCondor Cluster
- Sets up:
  - HTCondor collector
  - AUDITOR server + DB
  - APEL plugin

Can be useful to test if the solution may fit your requirements

https://github.com/ALU-Schumacher/auditor-demo





Future Perspectives

ecosister

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

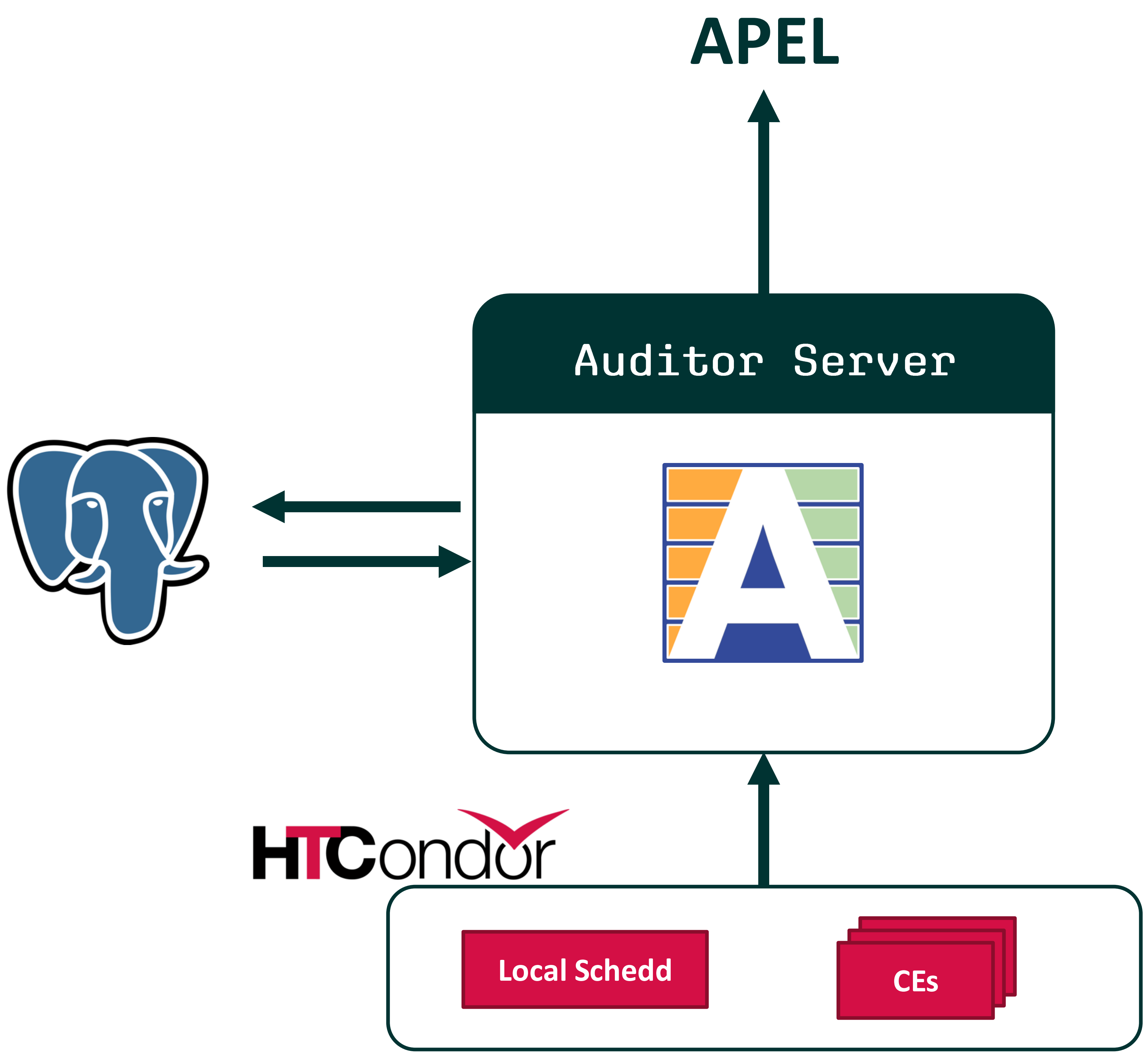
ART-ER

INFN CNAF

Future Perspectives

- We currently are in **pre-prod phase**
- Planning to **move to production** as soon as all the tests are completed
  → **switching** from local DB **to production PostgreSQL**
  → sending accounting metrics to **production APEL**

- Study AUDITOR APIs to get some accounting dashboards for internal use



More use-cases

kubernetes

openstack

Cloud Services @CNAF

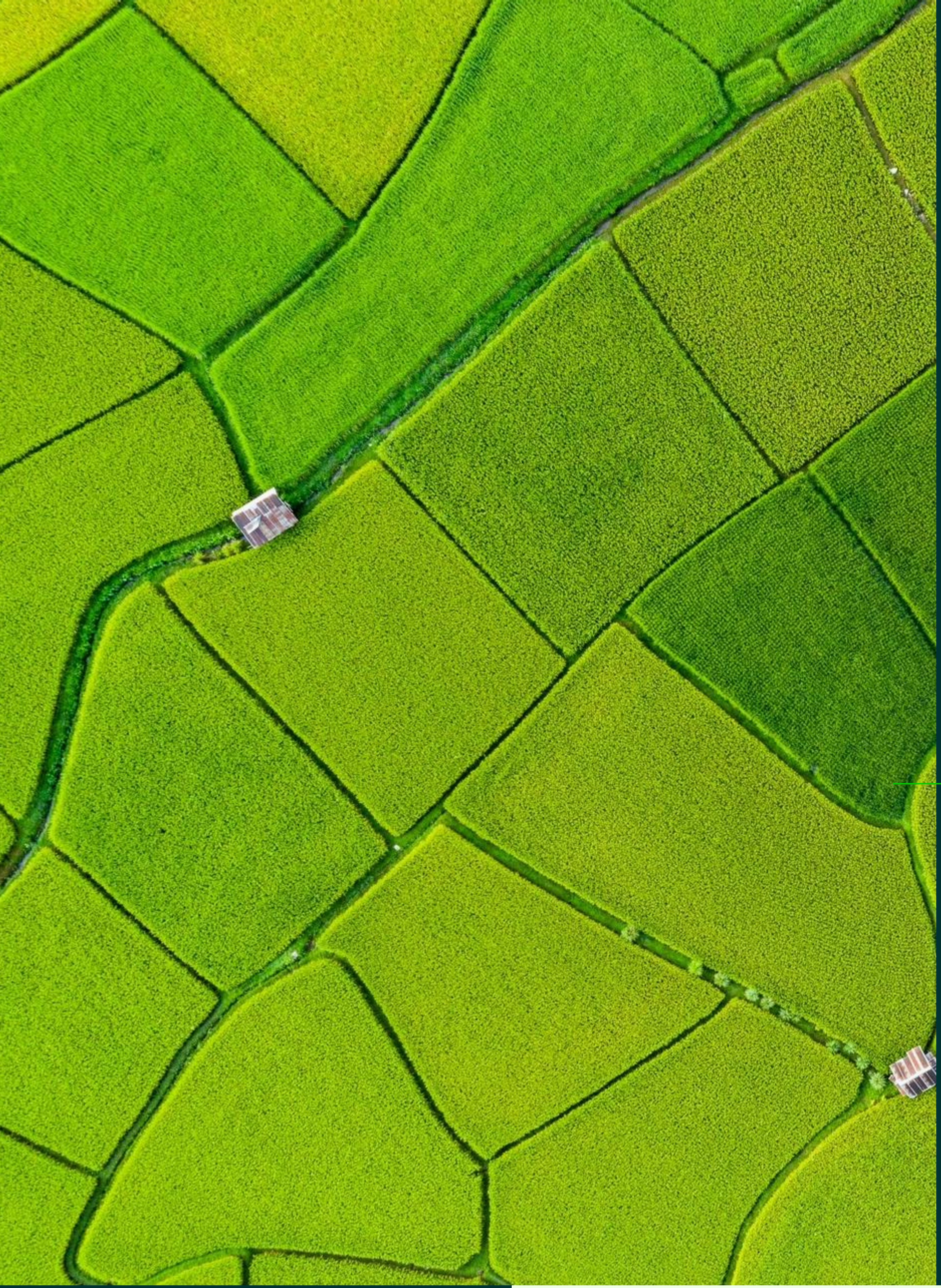
- Kubernetes clusters
- OpenStack instances

  → no AUDITOR collector right now ☹

INFN Cloud project

- Heavily based on K8s and OpenStack as well
- Currently using its own accounting system based on APEL

INFN CLOUD



Thanks!

ecosister

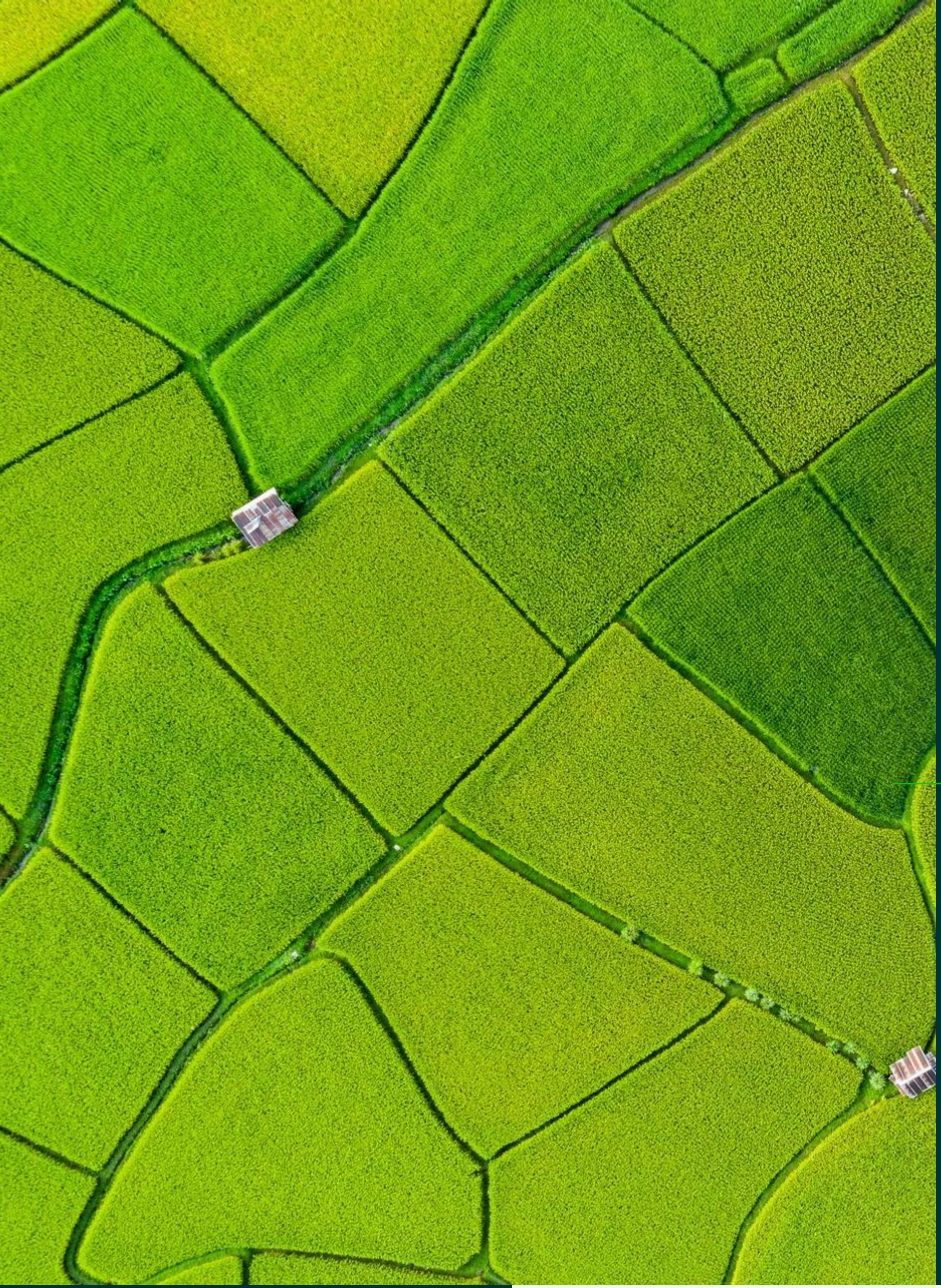
Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ART-ER

INFN CNAF



Backup

ecosister

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

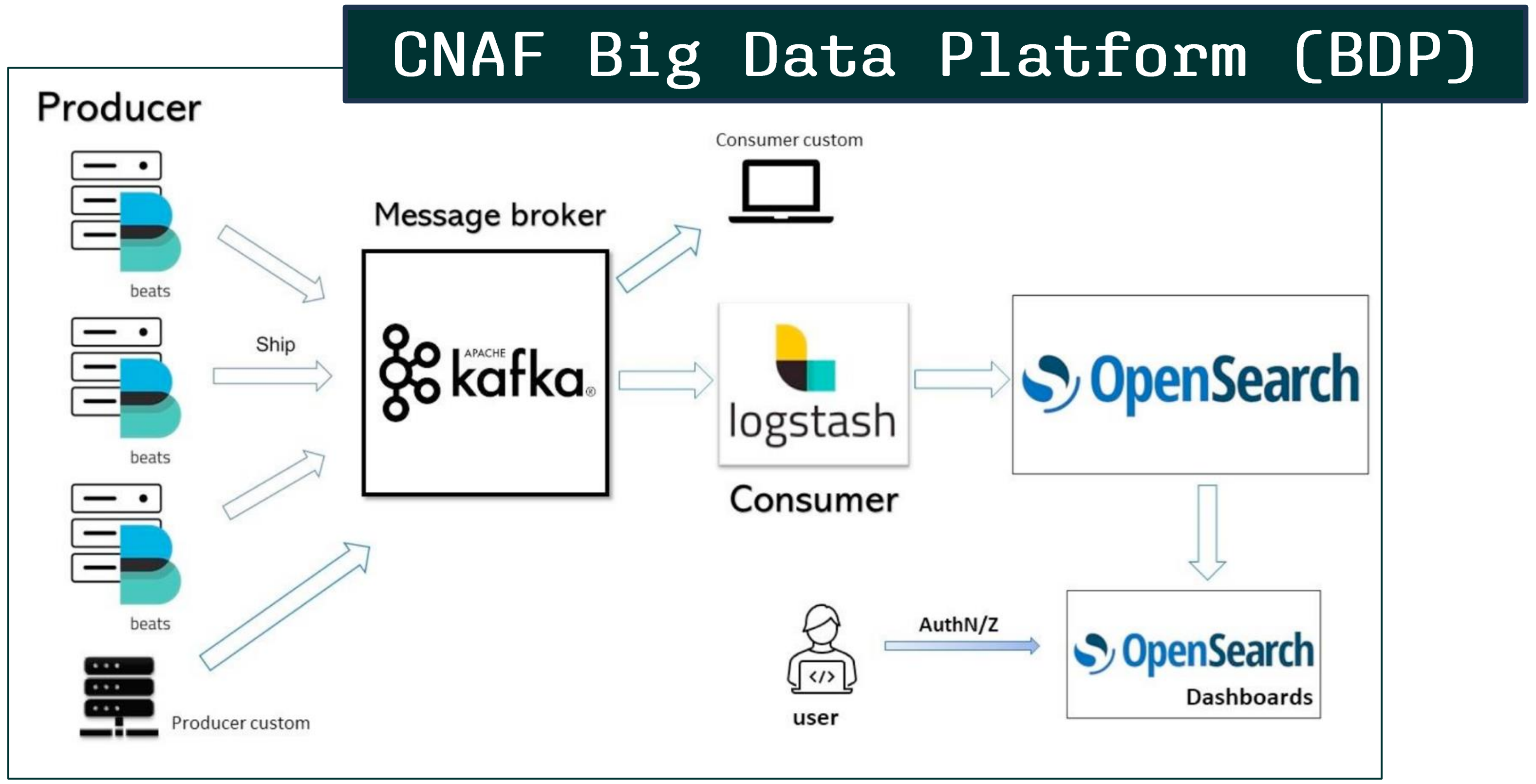
ART-ER ATTRATTIVITÀ RICERCA TERRITORIO

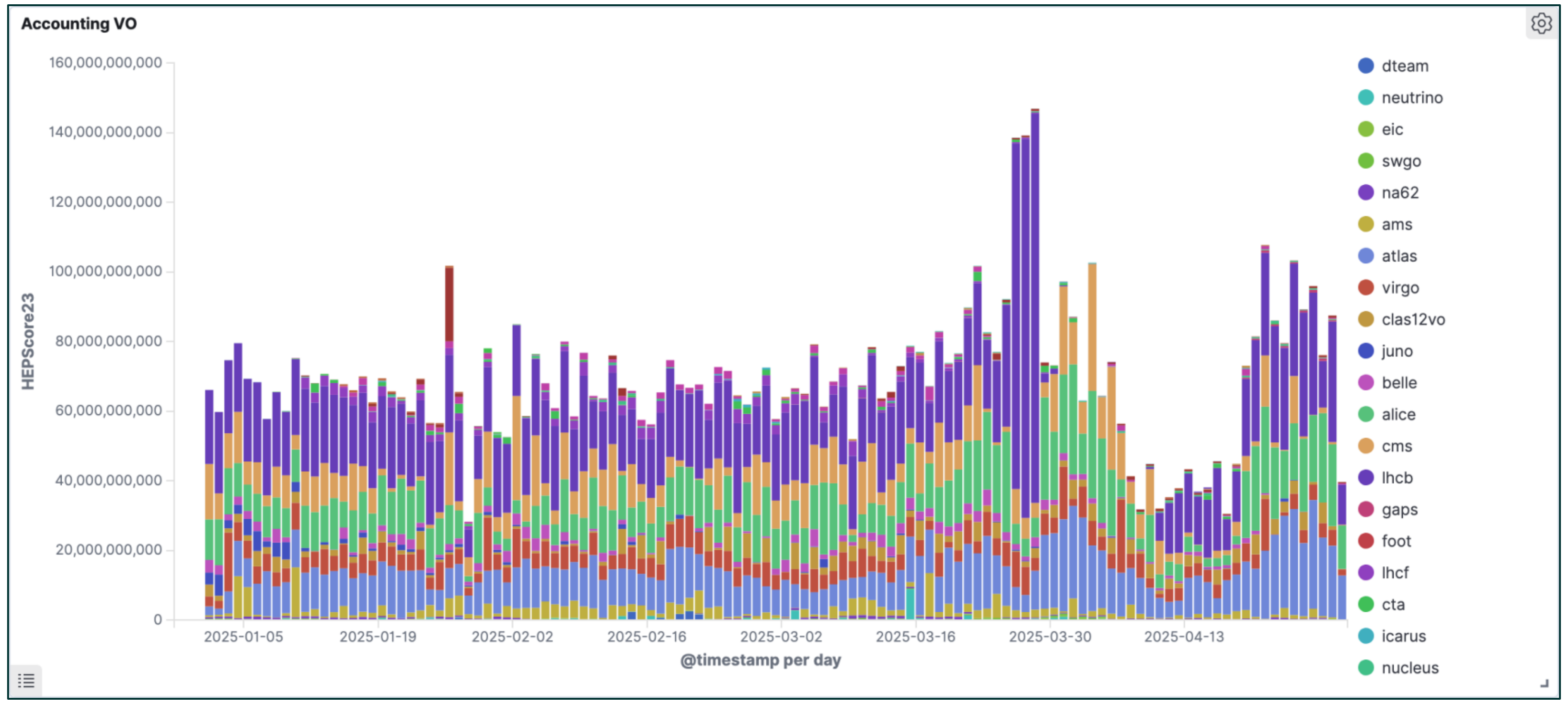
INFN CNAF

27 / 05 / 2025 – Workshop CCR

BDP accounting tests

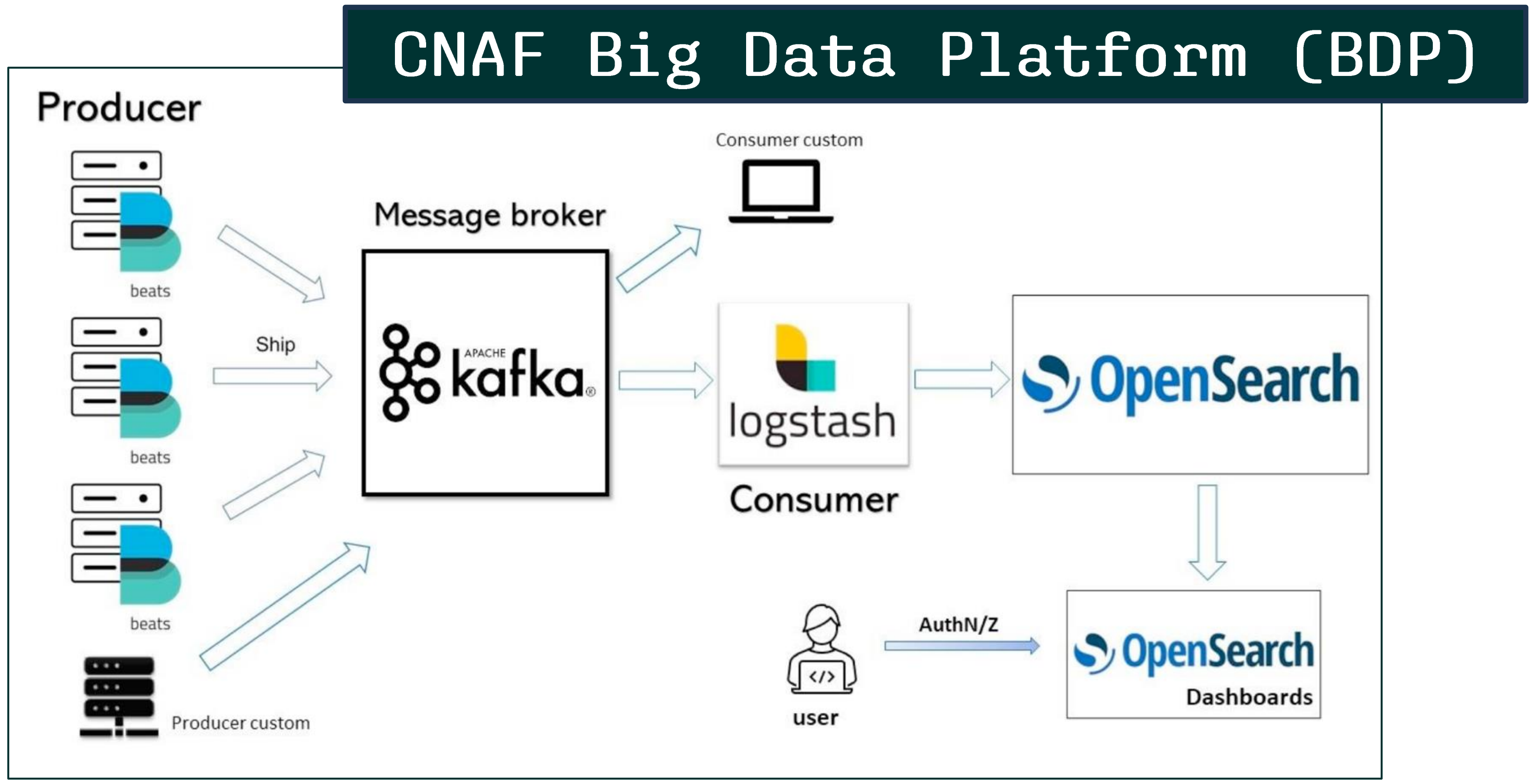
- Each Schedd/CE works as a «producer»
  → **Filebeat** reads HTCondor history files
- History information are sent to **Kafka**
- **Logstash** reads info from **HTCondor topic**
  → parsing of history data to get **accounting info**
- **OpenSearch** collects info to provide **dashboards**

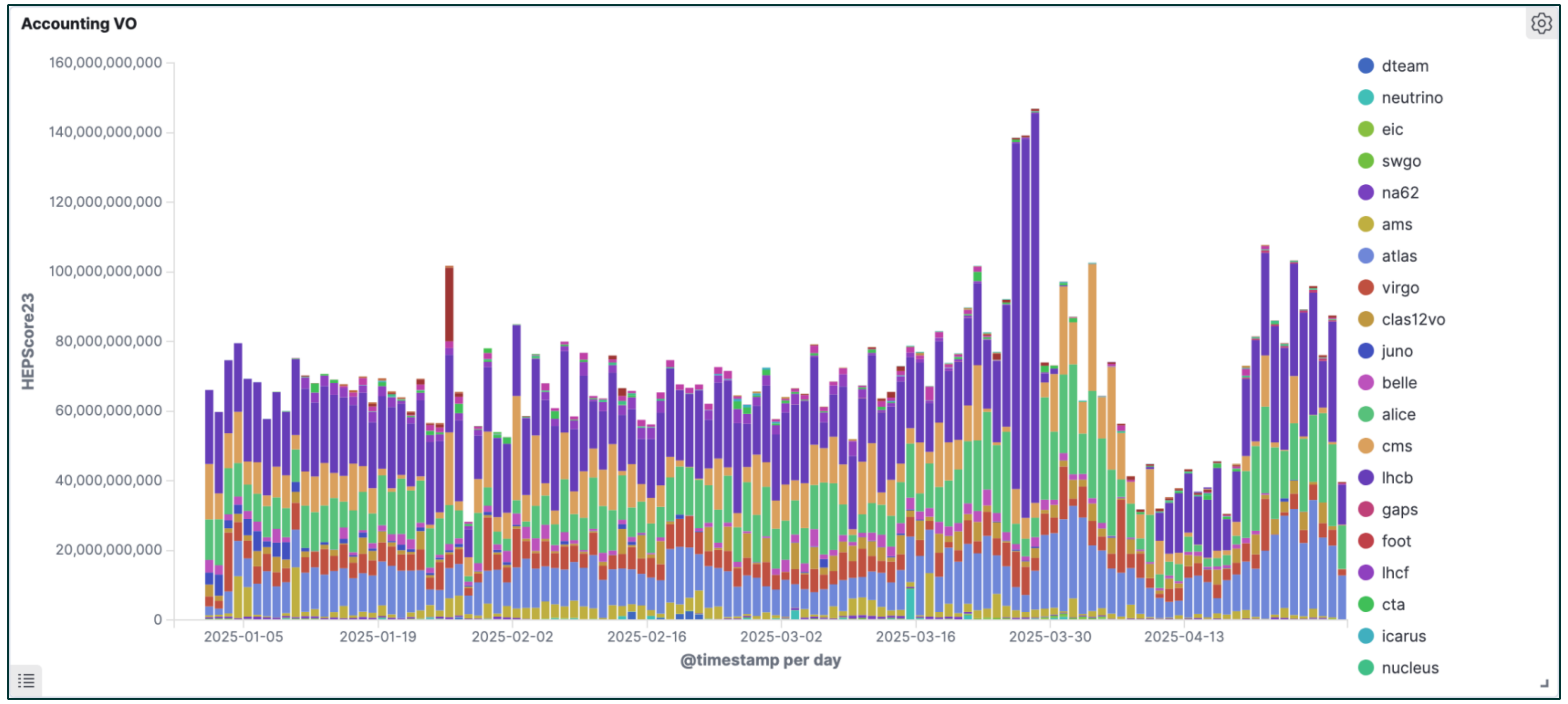




BDP in the future

- Useful to store old jobs info
- Planning on saving some **raw history info***
  to the BDP platform
  → ease debugging
  → history of HTCondor usage





* HTCondor attributes **tend to change**, the Logstash parsing is quite sensitive
→ keeping RAW info prevents things to break

HTCondor Accounting

Tuning of HTCondor Cluster

- HEPScore23/core ratio of the machine where the job is executed
- Site name
- VO name

vomap.conf

```
* agata vo.agata.org
* ams.prod ams02.cern.ch
* ams.analysis ams02.cern.ch
* cta.analysis vo.cta.in2p3.fr
* cta.prod vo.cta.in2p3.fr
* dampe vo.dampe.org
* darkside vo.darkside.org
* hyperk hyperk.org
* icarus icarus-exp.org
* km3 km3net.org
gridpp.ac.uk
e.org
```

sn.conf

```
##### Assign Voname from mapfile
use FEATURE:SetJobAttrFromUserMap(t1_voname,acctgroup,T1SetVOname, $(T1_SHARED_SCRIPT_DIR)/vomap.conf)

JOB_TRANSFORM_NAMES = $(JOB_TRANSFORM_NAMES) set_voname t1_accounting

##### Assign HS06, HS23 and sitename to job
JOB_TRANSFORM_t1_accounting @=end
  SET    t1_wn_hepscore      "$$(t1_wn_hepscore:0)"
  SET    t1_wn_hs06          "$$(t1_wn_hs06:0)"
  SET    t1_hs06_ratio       "$$(t1_hs06_ratio:0)"
  SET    t1_hs23_ratio       "$$(t1_hs23_ratio:0)"
  SET    TotalCpus           "$$(TotalCpus:16)"
  SET    SITE                "INFN-T1"
@end

JOB_TRANSFORM_set_voname @=end
  REQUIREMENTS  MY.t1_voname =?= undefined
  SET    t1_voname           split(MY.AcctGroup,".")[0]
@end
```

AUDITOR entries

- AUDITOR is based on PostgresDB but has a **peculiar DB structure**
- Queries can be done in an easier way via APIs
  https://alu-schumacher.github.io/AUDITOR/pyauditor/latest/

```
auditor=> select * from auditor_accounting order by id desc limit 2;
-[ RECORD 1 ]---------------------------------------------------------------------------------------------------------------------------------------------------------------
id         | 712991
record_id  | htcondor-ce04-htc.cr.cnaf.infn.it#3829732.0#1748274425
meta       | {"site": ["INFN-T1"], "user": ["atlasprd004"], "group": ["atlas"], "submithost": ["ce04-htc.cr.cnaf.infn.it"]}
components | [{"name": "Cores", "amount": 1, "scores": [{"name": "HEPSPEC", "value": 11.38888888888889}, {"name": "HEPScore23", "value": 11.38888888888889}]},
 {"name": "Memory", "amount": 2048, "scores": []}, {"name": "TotalCPU", "amount": 5985, "scores": []}]
start_time | 2025-05-26 17:47:10+02
stop_time  | 2025-05-26 19:26:55+02
runtime    | 5985
updated_at | 2025-05-26 19:35:21.865371+02
-[ RECORD 2 ]---------------------------------------------------------------------------------------------------------------------------------------------------------------
id         | 712990
record_id  | htcondor-ce04-htc.cr.cnaf.infn.it#3830310.0#1748279844
meta       | {"site": ["INFN-T1"], "user": ["atlasprd004"], "group": ["atlas"], "submithost": ["ce04-htc.cr.cnaf.infn.it"]}
components | [{"name": "Cores", "amount": 1, "scores": [{"name": "HEPSPEC", "value": 25.08928571428572}, {"name": "HEPScore23", "value": 25.08928571428572}]},
 {"name": "Memory", "amount": 2048, "scores": []}, {"name": "TotalCPU", "amount": 518, "scores": []}]
start_time | 2025-05-26 19:17:25+02
stop_time  | 2025-05-26 19:26:03+02
runtime    | 518
updated_at | 2025-05-26 19:35:21.749038+02
```

WLCG and Accounting

AUDITOR has been a nice upgrade from previous accounting solutions

Many other issues are still present:

- **HS06/23 accounting**
  → no clear way of accounting CPU power
  → experiments don't have a way of knowing how much HS are using, only #cores

- **VO names**
  → with **x509 proxies** the VO-name was reported inside the proxy
  → JWT don't have the same attributes, APEL and HTCondor teams are working on this*
  → right now each site is handling this in their own way

- Not easy interaction with EGI
  → not really clear how accounting dashboards work...

*https://indico.cern.ch/event/1484669/contributions/6463528/attachments/3062635/5416144/TDack-WLCGWShop-May25-TokenAccounting.pdf