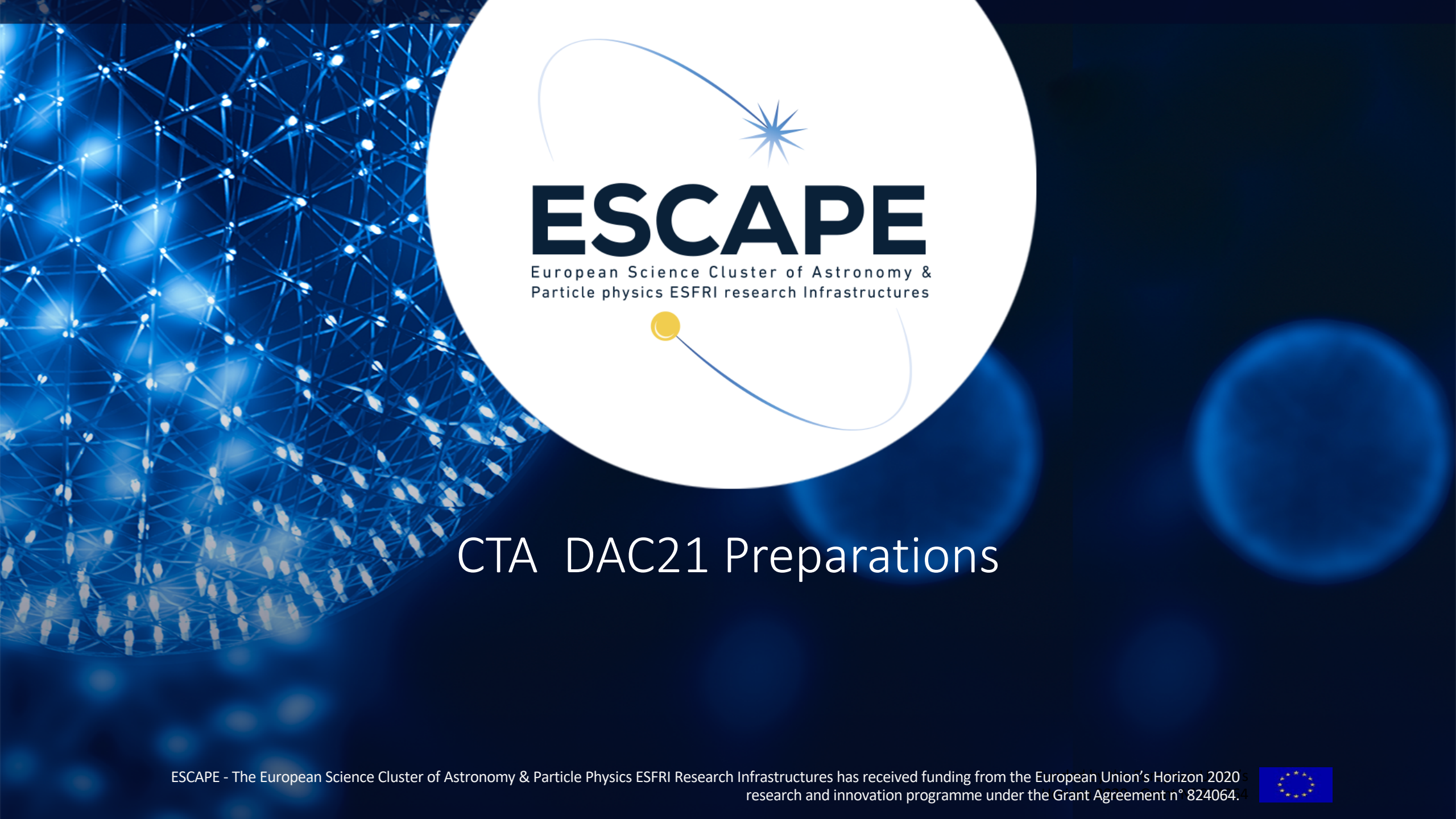


CTA DAC21 Preparations

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

People who do the work

- LAPP:
  - Frederic Gillardo
  - Berkay Turk
- PIC:
  - Agustin Bruzzese
  - Jordi Delgado
  - Gonzalo Merino
- External:
  - Luisa Arrabito (in2p3)
  - Nektarios Benekos & Karl Kosack (CTAO)
- CTAO/ESCAPE:
  - Nadine Neyroud, Matthias Füßling & Gareth Hughes

Funded by the European Union's Horizon 2020 - Grant N° 824064

Use Cases

Several CTA Use Case Document and redmine
For DAC21 we will focus on 2 + 1 of them

Name	1. Long haul ingestion and replication
ID	CTA001
Goal/Aim	Ingestion of CTA data from a remote site (RSE at th the data lake, transfer and replication in off-site RSI of the data at origin
Workflow	• (Fake/simulated/reference) data is injected into th (CTA Array Site). • Data is then automatically transferred from the re a RUCIO instance (CTA data lake instance) • Data is replicated to a separate site within the CT copy) • A copy of the data is transferred onto tape. • A check (e.g. checksum) is made that the data is **verification that the data is really stored on tape** **temporary staging cache of tape)** • The data at the origin is removed. • Data is findable (file can be retrieved using the ap in the ESCAPE data lake
Definitions	CTA Array Site = La Palma CTA data lake instance = CTA Rucio-based instanc
Requirements	Simulated/Reference data. CTA-RUCIO instance at PIC. CTA-RUCIO Storage Element on La Palma. Access to tape management system. Appropriate monitoring tools associated to CTA-RU
People Involved	Agustin, Jordi, Gonzalo, Frederic, Berkay, Nadine
Work Packages	WP2
Success	Data is successfully transferred, and file deleted on Data transfer was monitored. Metadata is stored. Data can be discovered using the ESCAPE or CTA

Name	2. Data Reprocessing
ID	CTA002
Goal/Aim	The ability to reprocess all raw data (DL0) to higher (DL3) leve
Workflow	• Raw (DL0) data is identified on tape (obsid or time range) via **metadata** e.g. using getMetaData method https://tinyurl.com/rnyyvab5 • Data volume is calculated. • Data is staged from tape (COLD) storage to temporary disk (• Data is reprocessed using CTA pipeline software via the wor management system (WMS, based on DIRAC) using a cache fly, transient data products • Final data products (DL3) are verified. • Cache and temporary data is cleared. • Ingest the resulting new DL3 data into the datalake. • Update the corresponding metadata.
Definitions	DL0 data = data level 0 (raw data) DL3 data = data level 3 (science-ready data) WMS = workload management system (based on DIRAC) CTA pipeline software = software for the reconstruction, calibra quality monitoring, simulation
Requirements	CTA pipeline software Simulated/Reference data (DL0). Tape storage. CTA WMS instance & interface. ESAP interface?
People involved	Agustin, Jordi, Gonzalo, Frederic, Berkay, Nadine, Luisa, Gare
Work Packages	WP2 WP3 WP5
Success	The data (DL3) are findable in the ESCAPE data lake (CTA ins updated to a new version.
Things to test	Processing time. Large-scale data production. Human interface.
Risks	RSEs ready in time Tape drive access PIC and DESY Reference data access

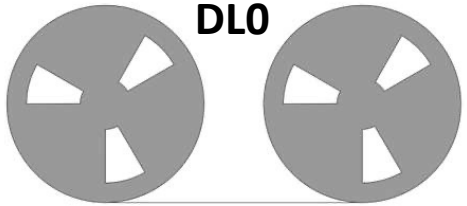
Name	4a Analysis of (simulated) CTA science data by a Principal Investigator (PI) [Interactive]
ID	CTA004a
Goal/Aim	A project PI is able to login to the ESAP and find and analyse the data from their proposal
Workflow	1. User logs in to the ESAP and is identified as a CTA project PI 2. Search for Data in the datalake a. Search for (simulated) CTA DL3 level data by project ID. b. Select data from search results or select all 3. Search for corresponding IRF (instrument response function) for the data selected 4. Search for corresponding metadata, log files etc .... 5. The data can now be analysed in interactive mode, in batch mode or downloaded 6. Interactive mode using Jupyterhub (see CTA005)
Definitions	PI = Principal Investigator IRF = Instrument Response Function
Requirements	CTA Science Tools Simulated / Reference data Instrument Response Function Fine grained IAM login with permissions Access to the interactive mode of ESAP Datalake metadata
People	Gareth, Axel, Cosimo
Work Packages	WP2 WP3 WP5
Success	High level products e.g. skymap, lightcurve, SED, ...
Things to test	Search for data belonging to someone else. Provenance.
Impact	Ability to search for and analyse embargoed data on the datalake.

Funded by the European Union's Horizon 2020 - Grant N° 824064

Setup

DL0

DL0



DL3

Multiple Sites
>2

DL3-6

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Setup

DL0

DL0

DL3

DL3-6

Multiple Sites
>2

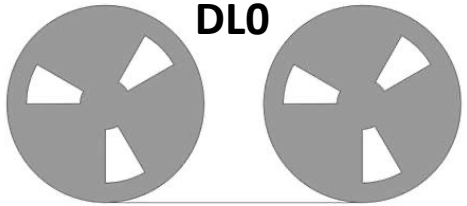
Funded by the European Union's Horizon 2020 - Grant N° 824064

Setup

DL0

UC1: Long-haul ingestion & replication

DL0



DL3

UC2: Data Reprocessing

DIRAC THE INTERWARE

Multiple Sites >2

jupyter

DL3-6

UC4a: Analysis

Funded by the European Union's Horizon 2020 - Grant N° 824064

Data & Technology being used

- Data:
  - Latest CTA simulation production used in the construction phase (DL0)
  - ~100TB
- Software:
  - RUCIO v1.23.15
  - FTS v3.10.1, gridFTP, dCache, HTTP-TPC
  - CTADIRAC r7v2
  - ctaPipe (DL0 -> DL1 only atm)
  - gammapy v0.18.2
- Infrastructure:
  - CTA-RUCIO instance @ PIC
  - CTA-DIRAC test cluster @ in2p3
  - ESCAPE-RUCIO

Funded by the European Union's Horizon 2020 - Grant N° 824064

Use Case 1: Long-haul ingestion and replication

Workflow:

1. Simulated data is injected into the datalake at on-site (La Palma)
2. Data is then automatically transferred from on-site to the datalake using a CTA/PIC-RUCIO instance
3. A copy of the data is transferred onto tape. Data is replicated to a separate site within the datalake instance (second copy)
4. A check is made that the data is preserved
5. The data at the origin is marked for deletion
6. Data is findable (file can be retrieved using the appropriate query)

Tests (cumulative): 10s TBs

- automatic detection of the data ingested on-site (La Palma) and transferred to off-site (PIC)
- RUCIO triggers the deletion of the files from the origin (La Palma)
- increase the size of the datasets & add monitoring
- increase the size of the datasets to full one night of observation
- automatically detect and transfer files produced from observation at la Palma to the tape storage at PIC
- introducing rules to perform multiple copies of a single datasets

Success:

a. Data is successfully transferred, replicated, and file deleted on the origin RSE.

b. Data transfer was monitored.

c. Data can be discovered using the ESCAPE or CTA-RUCIO instance

Funded by the European Union's Horizon 2020 - Grant N° 824064

Use Case 2: Data Reprocessing

Workflow:

1. Raw (DL0) data is identified on tape via metadata (e.g. obsid or time range)
2. Data volume is calculated
3. Data is staged from tape storage to temporary disk storage
4. Data is reprocessed using CTA pipeline software via the workload management system (WMS, based on DIRAC) using a cache area for on-the fly, transient data products
5. Final data products (DL1) are verified.
6. Cache and temporary data is cleared.
7. Ingest the resulting new DL1 data into the datalake.
8. Update the corresponding metadata

Tests: ~100TB

- reprocessing of DL0 data trigger from hot storage
  - *Pull DL0 from hot storage*
  - *Process the DL0 using ctaPipe (DL0 to DL1)*
  - *Push DL1 on Hot storage*
- reprocessing of DL0 data trigger from cold storage (tape)
  - *Move DL0 from tape to staging area*
  - *Process the DL0 using ctaPipe (DL0 to DL1)*
  - *Push DL1 on Hot storage*

Success:

a. Use of RUCIO and DIRAC integration

b. The data (DL1) are findable in the datalake (CTA instance) and are updated to a new version

Funded by the European Union's Horizon 2020 - Grant N° 824064

Use Case 4a: Interactive Analysis

Workflow:

1. User logs in to the ESAP and is identified as a CTA project PI
2. Search for Data in the datalake
   a) Search for (simulated) CTA DL3 level data on the ESCAPE-RUCIO instance by project ID
   b) Select data from search results or select all
3. Search for corresponding IRF (instrument response function) for the data selected
4. Search for corresponding metadata, log files etc
5. The data can now be analysed in interactive mode using Jupyterhub

Tests:

- Gammapy enviroment can be deployed using ESAP
  - Final onboarding can be simulated using Zenodo sandbox / gitlab version
- Data can be discovered using ESAP
- The data can then be used by the software
- Data can be analysed to produce higher-level data

Success

a. High level products produced from data on the datalake

Funded by the European Union's Horizon 2020 - Grant N° 824064

Notes & Todo

- UC1 Long-haul transfer and replication:
  - Infrastructure:
    - CTA-RUCIO instance running on k8s @PIC including tape, limited access at the moment
    - agreement from partners on La Palma to use connection and computing faciities, RSE visiable from PIC and testing this week
    - Mock tests this week
  - Data: Requesting CTA Simulations ongoing
  - Software: RUCIO, Grafana

- UC2 Data Reprocessing:
  - Infrastructure:
    - CTA-RUCIO & computing @PIC – see above
    - CTADIRAC test cluster
    - First hurdle of a generic user that can use both to be tested soon
  - Data: Requesting CTA Simulations ongoing
  - Software: RUCIO-DIRAC (work ongoing) & CTApipe (DL0-1)

- UC4a Data Analysis:
  - Infrastructure: ESAP
  - Data: No request just simulate and add to ESCAPE-RUCIO
  - Software: gammapy not yet fully onboarded (but can be added to ESAP by hand), example notebook

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Conclusion

- Use cases and tests outlined mirror the proposed setup of CTA
  - >1 nights worth of data taking
  - ~1 years worth of reprocessing
  - Online analysis from science platform
- Transfering simulated data from Nothern array site will hopefully yield valuable lessons
- The ability to integrate tape and a WMS (eg DIRAC) are already CTA requirements – important test for RUCIO
- Whilst missing some small steps, these use cases and tests represent ingestion of data to (re)processing to analysis

Funded by the European Union's Horizon 2020 - Grant N° 824064