WP6. TECHNICAL WORKS
D6.14B

EST DATA CENTER CONCEPT

P. Caligari (KIS), N. Bello González (KIS), S. Berdyugina (KIS), P. Kehusmaa (KIS)

EST-DC Preliminary Requirements

1. **Data Policy** (decision needed before starting designing!)

- Data **needs a License!** (e.g.: https://creativecommons.org)
  - CC0 1.0 (public domain) — use for metadata
  - CC BY 4.0 (credit, share & adapt, commercial usage allowed) — use for freely accessible data
  - The more restrictive a license the more problematic it becomes to combine with data from other sources

- Ownership?
  - EST / Consortium?
  - PI / Observers?
- Copyright?
- Embargoes?
  - How long by default? — SDC: 1 year
  - Longer if PhD involved? — SDC: 2 years
  - What about 3rd party campaigns?
  - What about technical campaigns, ad-hoc campaigns, exceptions from the default?

EST-DC Preliminary Requirements

2. Governance

- Requirements management — TOGAF
  - Tightly bound to scientific requirements
- Project management framework — PMBOK
  - Define Objectives
  - Resource, constraints and risk management
- Service Management — ITIL
- General Governance and Management of IT — COBIT
- Security — ISO 27000
- DevOps — GitLab based, CI/CD

3. Data

- Metadata Standard — SOLARNET
- Data format — FITS

Data Rates

EST Instrument	Data rate with present-day technology	Data rate with upcoming new technology
IFS IR 1083-1565nm	3 GB/s	40 GB/s
TIS R 680-1000nm	4 GB/s	6 GB/s
IFS R 680-1000nm	5 GB/s	6 GB/s
IFS V 500-680nm	30 GB/s	55 GB/s
TIS V 500-680nm	4 GB/s	6 GB/s
IFS B 390-500nm	30 GB/s	55 GB/s
TIS B 390-500nm	4 GB/s	6 GB/s
	80 GB/s (= 1,2 PB/d)	**174 GB/s (=2,5 PB/d)**

- Instruments (Quintero Noda et al. 2022)
  - 4x IFS
  - 3x TIS
- 100 days / year @ 4 hours / day
- Between 120 PB/y & 250 PB/y (≅ data volume of LHC in 2022!)
- Replication of 2,5 PB to European mainland in 1 day requires 400 Gbit/s leased line

Data Types and Processing

Data Levels

- **Level 0:** raw data & calibration and parameter files
- **Level 1:** calibrated science ready data (Speckle or similar goes here?)
- **Level 2:**
- **Level 3:**

Data Processing

- Hybrid infrastructure
  - Virtualize whenever possible (VMware, k8s, …), possibility to outsource to public cloud
  - Containerized software, BinderHub
  - Requires central public software repository

Data Center Topology, Dataflow & Compute Resources

- **Primary DC**
- All data (L0, L1, Ln)
- Data calibration (L1) & high level data products (Ln)
- Main on-premise computing power goes here (CPU, GPU, cloud computing
- Infrastructure: DM, central Services GitLab, Webpages, …

client

Ln

- **Secondary DC**
- Mirror of all data (L0, L1, Ln)
- Complementary data products (Ln)
- Some on-premise computing power (complementary)

client

(Ln)

(Ln)

L0 & L1

- **Edge DC** (partner)
- Subset of data (instrument)
- Data calibration (L1) & data products (Ln)
- Limited computing-power needed

client

L0 & L1

L0

- **On-site DC**
- Raw data (L0)
- Limited data calibration (L1) & previews
- Little compute power (2030: previews by ML)

Data Access

- **Authentication & Authorization Infrastructure**
  - Distributed multitenant user management
  - Common AAI

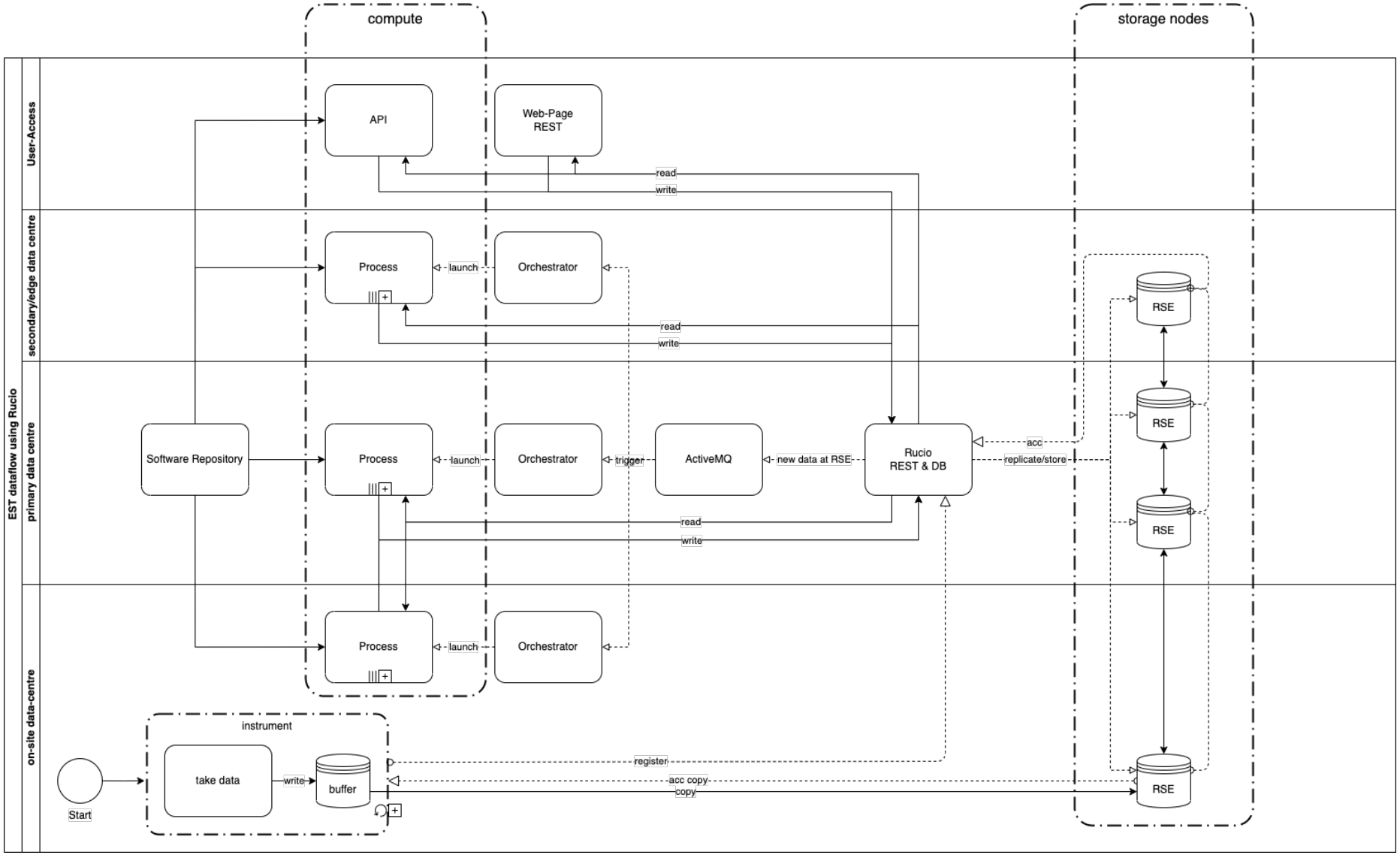
adopt an existing Grid Infrastructure, consider environment (ESCAPE)

- **User access**
  - Project webpage (tools, how-to, …)
  - Searchable webpage to data in the archive with previews
  - Integrate with similar efforts (SOLARNET SVO, ESCAPE, ESOC, …)
  - Helpdesk
- **Programmatic**
  - RESTfull interface of the archives's search webpage
  - API (Python, integrate with SunPy & IDL)

Data Management & Lifecycle: Rucio

- **Rucio**
  - Open-source data management framework actively developed at CERN
  - Rule- and interest-based data lifetime and replication
  - Inteligent resource and distance aware data distribution
  - Flat WORM like namespace
  - Used in ESCAPE, SKA & CTA
  - Wide support for storage solutions (dCache, xRootD, Ceph, …) & protocols (WebDAV, GridFTP, NFS, S3, …)
  - Wide support for storage types: HDDs, SSDs, tape, cloud
- **Problems**
  - Poor support for custom metadata (SK working on that)
  - No support for embargoes
  - One DC-wide central Rucio instance

Dataflow with Rucio



Cost estimates (HDD Storage)

Personel