



Integrating High Resolution Solar Physics

Big Data storage for ground-based solar data

The KIS Science Data Centre (SDC) Case – Part III

3rd SOLARNET Forum Meeting for Telescopes and Databases

15 Nov 2021

S. Berdyugina (SDC PI), P. Caligari (SDC Head), N. Bello González (SDC Project Scientist), P. Kehusmaa (SDC Manager and System Architect), C. Schaffer (System Architect & DevOps) & the KIS SDC Team



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824135.

Outline

- Purpose of this contribution (NBG)
- The KIS Science Data Centre Case – Introduction (NBG)
- Big Data at KIS SDC – Storage and management of large data volumes (P. Caligari, SDC Head)
- Big Data at EST Data Centre – A consortium effort (S. Berdyugina, SDC PI)

Purpose

Deliverable D2.22: *Report on Big Data storage possibilities*

Lead: KIS | Due date: month 36

WP2.2.6 Big-data storage. Typically scientific institutes have hosted their own data. As data volumes to be stored grow and the market for cloud solutions develops, there is reason to reconsider traditional solutions. The possibilities of in-house vs. existing (public or commercial) clouds large-scale data storage for solar physics will be explored and recommendations written. We will consider to store data in the framework of the European Open Science Cloud (EOSC).

As many other activities within SOLARNET, this task is part of developing a concept for data storage and processing for EST

The KIS Science Data Centre Case – Introduction

We are currently developing at KIS the Science Data Centre for calibration, storage, curation, archiving and dissemination of data from

- GRIS-slit & GRIS-IFU (in collaboration with M. Collados/IAC), LARS, BBI, Hellride(2022+) and LEAP (2022+) instruments and ChroTel at the solar observatories in OT (Tenerife)
- DKIST Level 1 data in collaboration with the DKIST DC (agreements in preparation) and possibly Level 0 data in the future

SDC main focus is also the development of new diagnosis tools (e.g., stochastic analysis of fluctuations in physical parameters), data science (e.g., research on statistical properties from solar data all over the GRIS archive 2014-2019) and other high-level data products (e.g., M-E VFISV (Borrero et al. 2011) inversions run over all the GRIS data archive)

SDC Funds: 1.4Mio/y for 2021-2024 and beyond after evaluation (PI Berdyugina)

22/11/2021

The KIS Science Data Centre Case – Introduction

SDC Cooperation with EU & DE projects



- Coordination and participation in the WP5 *Towards a European Science Data Centre*
- Participation in the Virtual Access Programme with the SDC archive



- Participation on WP4 on integrating ground-based solar data in the Astronomical VO
- Participation on the ESFRI Science Analysis Platform (WP5) including high-level ground-based solar data products



- Representing the solar community in the PUNCH4NFDI consortium of the particle, astro-, astroparticle, hadron and nuclear physics community in Germany

The KIS Science Data Centre Case

Today we are launching the SDC platform as a service for solar community

<https://sdc.leibniz-kis.de/en>

During this year, we have been building a solid infrastructure foundation for future needs meaning flexibility and scalability for large data amounts, software development and computing power

For the next coming years we aim to bring in new analytic tools and functionalities

We warmly welcome your feedback and suggestions on how to improve the platform to better serve the solar community

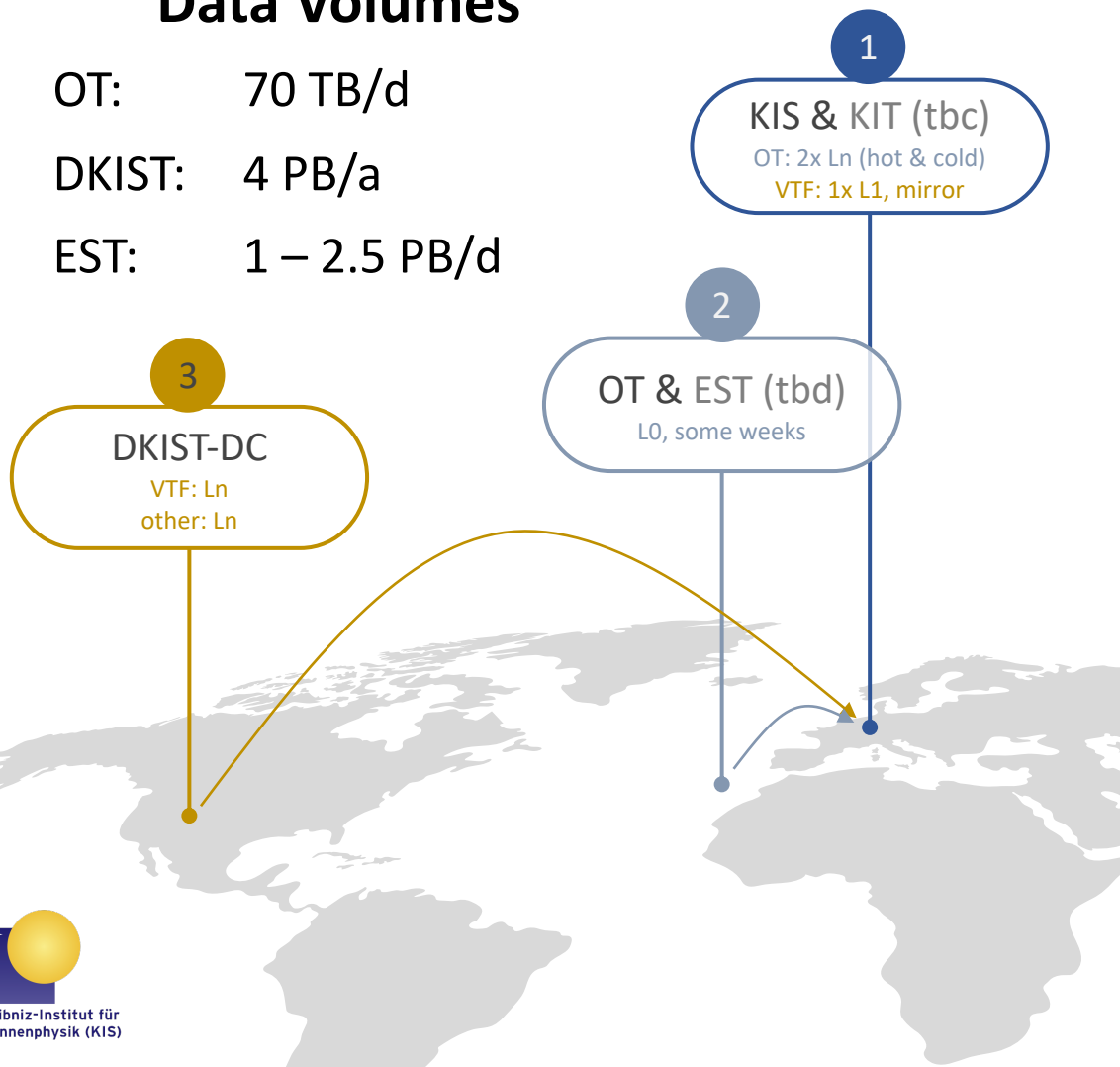
Big Data at KIS SDC

Data Volumes

OT: 70 TB/d

DKIST: 4 PB/a

EST: 1 – 2.5 PB/d

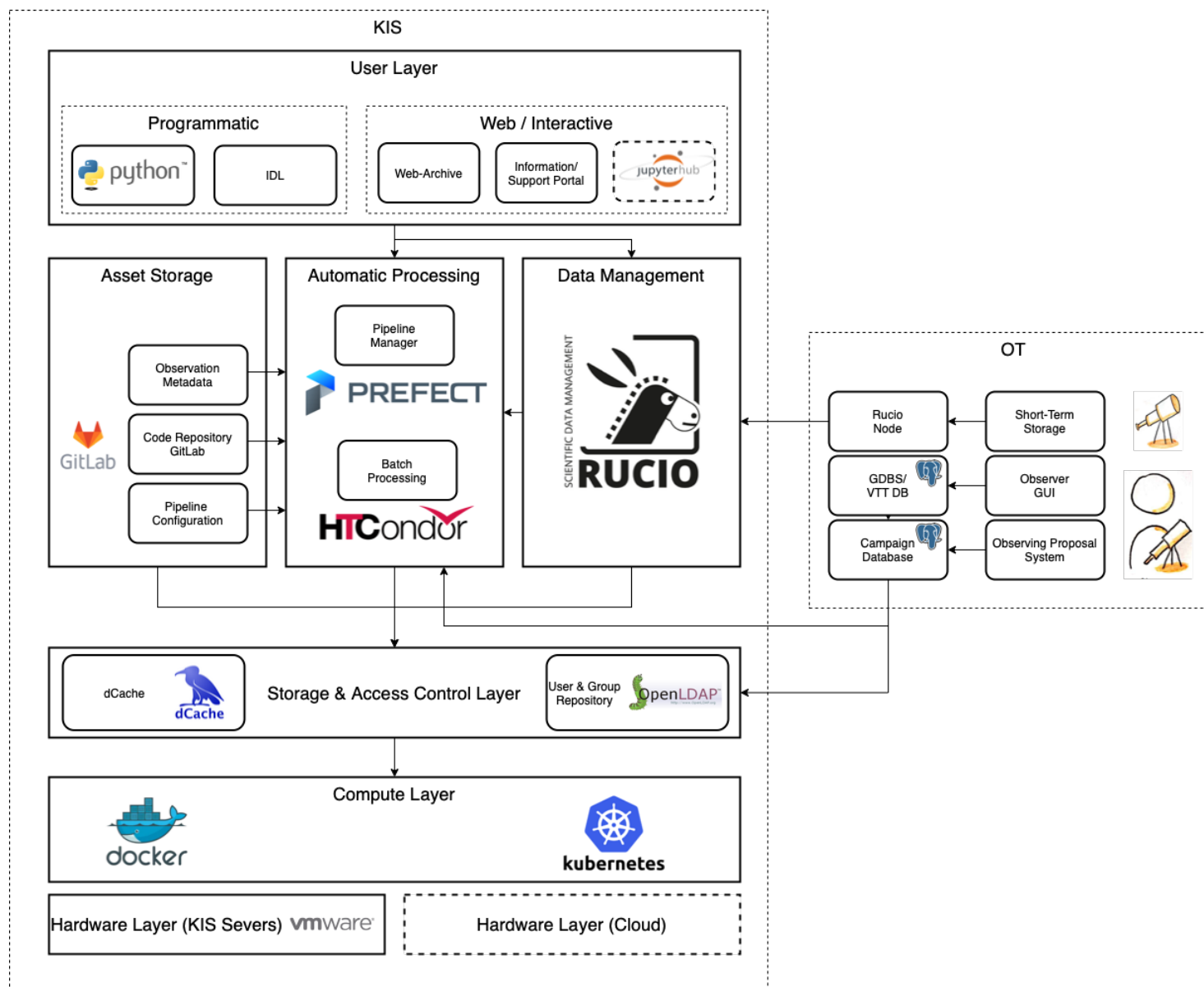


Requires:

- Multi site (edge storage & computation)
- Tiered solution, including
 - Tape libraries
 - Cloud S3 storage
- Flexible replication and data lifetimes

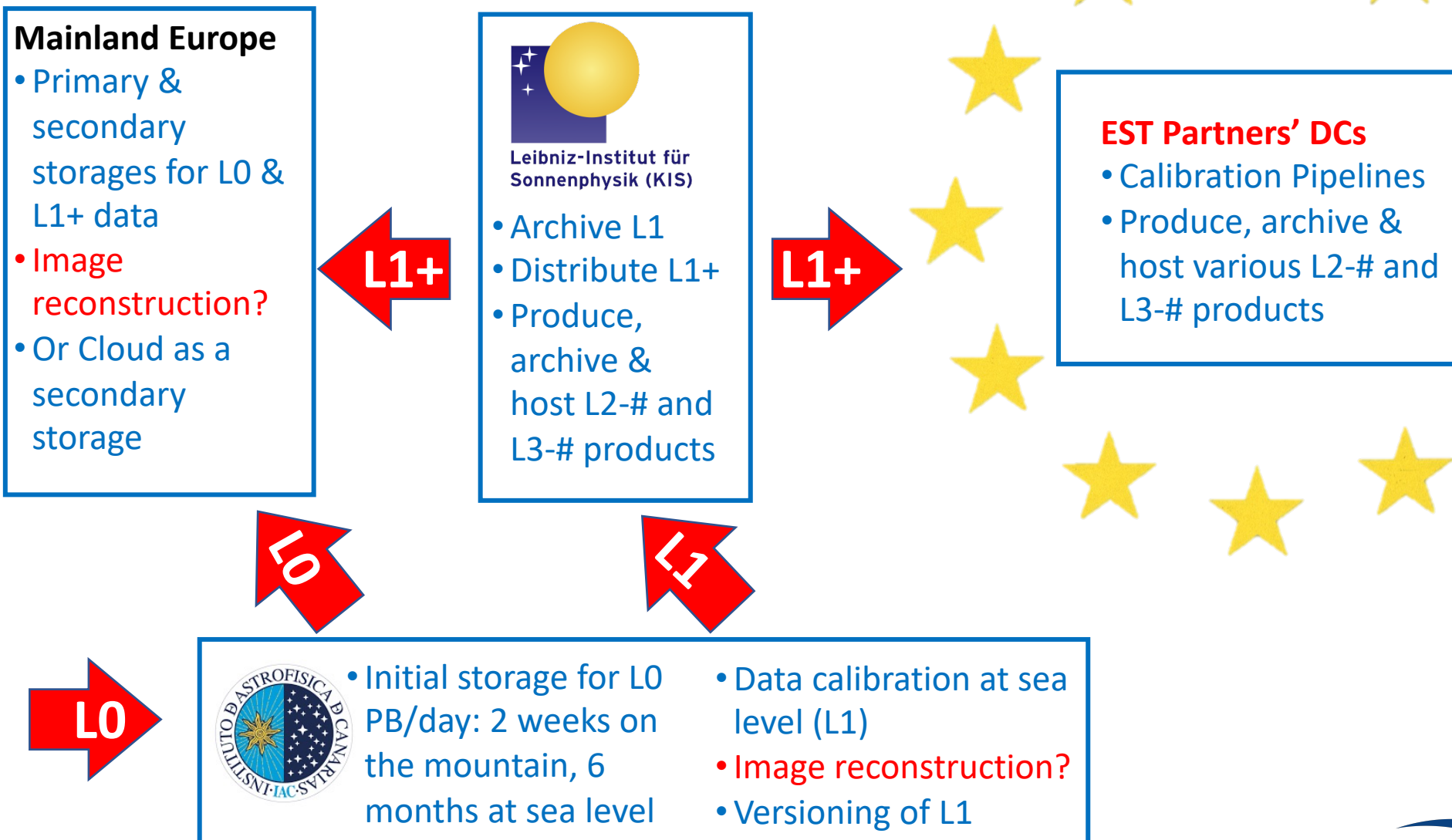
Answer:

- RUCIO & dCache
- Problem: No (native) embargoes



- Scales to multiple sites
 - OT already is an edge (in a sense)
 - Could have more sites similar to KIS
- One data lake managed by Rucio
- Standardized Software & Pipelines
- Open Source!
- Requirements are similar throughout astronomy
- Metadata Standards!

EST-DC vision: Distributed DC



EST DC Consortium

- Science requirements (storage amount, access speed, calibration time, etc)
- Governance, roles & responsibilities of partners
- Data management
 - Primary & secondary storages in mainland Europe (where?)
 - Data transfer (speed, cost, etc)
 - Life-cycle of data
- Software management
- Calibration pipelines
 - Definition of calibration for each instrument
 - Where is image reconstruction done? How much computing power is needed?
- High-level data
 - L2-# & L3-# data definition & production
- DC construction and operation costs at partner institutes