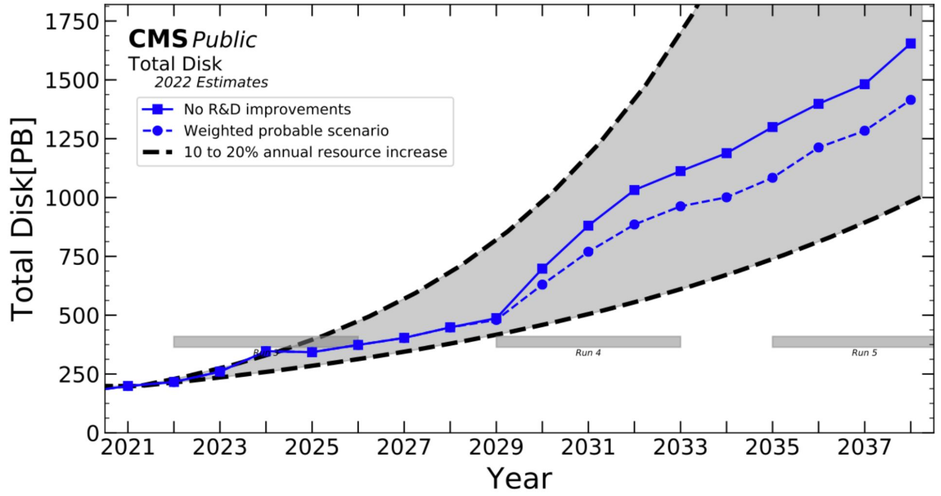
Coffea-casa analysis facility

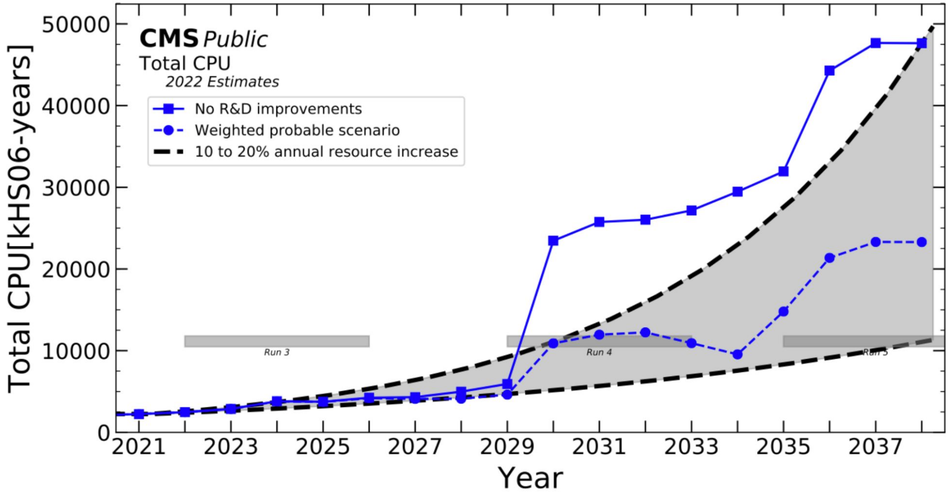
Sam Albin, Ken Bloom, Oksana Shadura,
Garhan Attebury, Carl Lundstedt, John Thiltges
[University of Nebraska, Lincoln, USA]
Brian Bockelman
[Morgridge Institute, Madison, USA]

International Symposium on Grids & Clouds (ISGC) 2023

Preparing High Energy Physics (HEP) Analysis Facilities for HL-LHC

- Preparing for HL-LHC era starting in 2029
  - Two general-purpose detectors, **ATLAS and the Compact Muon Solenoid (CMS), are expected to generate up to an exabyte each year =** about 1000 times as much as the data used to discover the Higgs boson in 2012
- **Community will need to come up with new computing model and integrate existing resources**





HEP Analysis facilities: what it is?

HEP Analysis Facilities are usually used for end-user analysis

- **People, software / services, and hardware** used to support analysis activities for an experiment.
- **Services which** includes:
  - Access to experimental data products.
  - Storage space for per-group or per-user data (often ntuples).
  - Access to significant computing resources.
- **Physics software**: ROOT and the growing Python-based ecosystem.
- **Computing hardware** (currently) looks like most of our computing facilities: CPUs and disks, maybe GPUs

B.Bockelman, Analysis Facilities for the HL-LHC, Snowmass

"How is this different than an existing facility, e.g. USCMS LPC @ Fermilab?"

Faster analysis and better end user experience

HEP Analysis facilities: requirements

- **Interactivity:** need to support both interactive analysis and batch mode
- **Low latency data access:** expected to facilitate low latency data access to achieve best performance (via various data access patterns)
- **Reusability:** should support extraction of user defined experiment data formats to migrate them onto laptops, desktops, workstations at home institutions or at home
- **Easy Deployment:** AF services expected to be deployed with industry standard platforms such as Kubernetes to facilitate easy deployment within a Tier-1s/Tier-2s/Tier-3s (easily co-locating next to existing resources)

Building blocks used for designing AFs

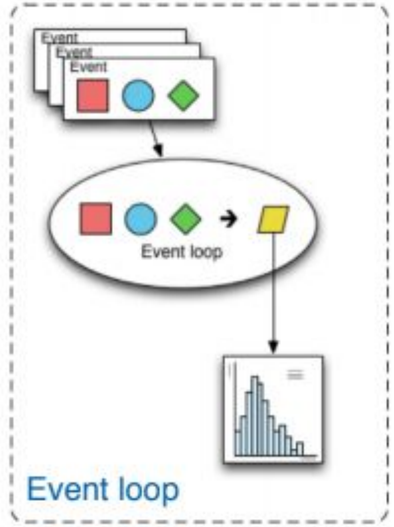
- Modern authentication (AIM/OIDC), tokens, macaroons
- Efficient data delivery and data management technologies
- Columnar analysis and support new pythonic ecosystem
- Modern deployment and integration techniques
- Support for object storage
- Efficient data caching solutions
- Easy integration with existing HPC resources

Analysis facilities: columnar analysis frameworks

Traditionally physicist have done **row-wise analysis**

– Process one physics event at a time serially

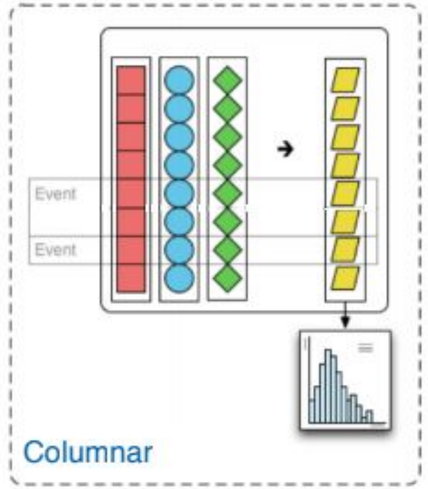
– Used in scripts and batch jobs (and now also in notebooks)



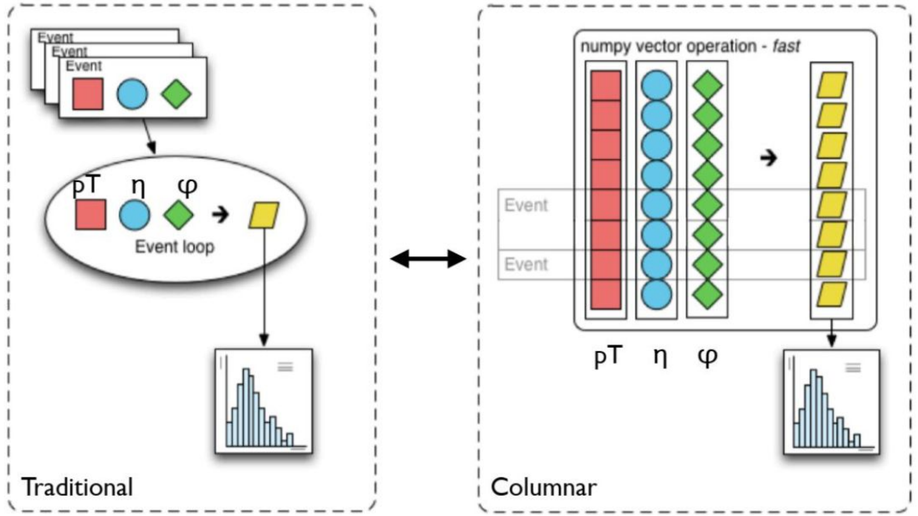
Trending towards **column-wise (tidy/big data) analysis**

– Low-latency access to CPU resources

– Fast access to a significant amount of disk



Building blocks: columnar analysis frameworks (coffea analysis framework)



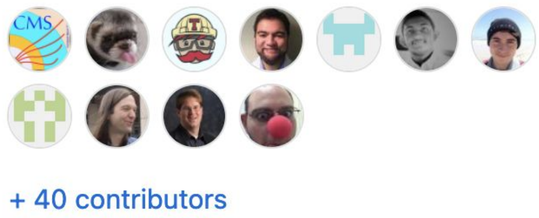
New columnar data analysis concepts!

User just needs to define a high-level wrapper around user analysis code: **the coffea processor** and coffea framework will take care of everything incl. ***scaling-out***

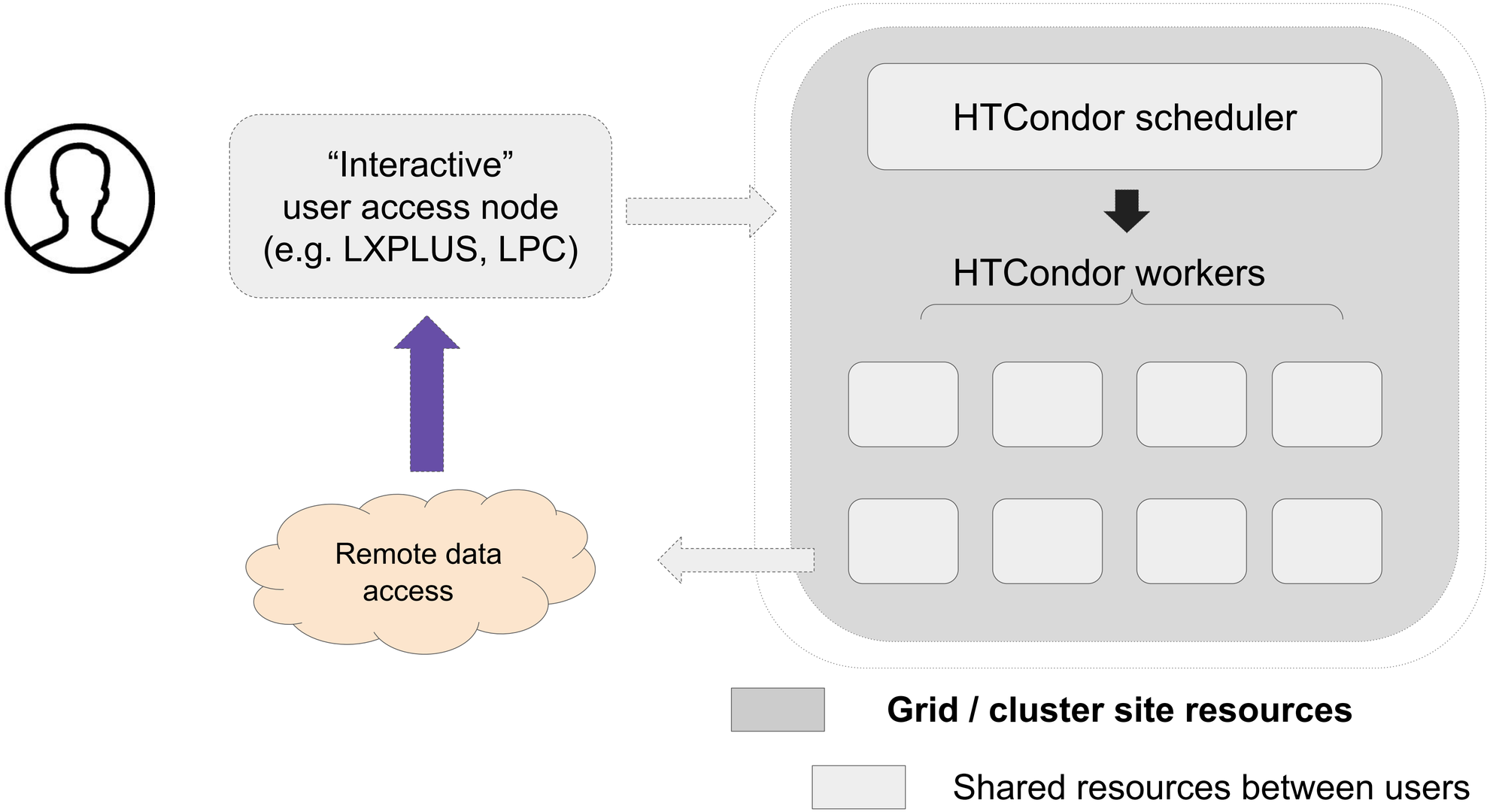
Distributed executors!

Smith, N., Gray, L., Cremonesi, M., Jayatilaka, B., Gutsche, O., Hall, A., ... & Pivarski, J. (2020). Coffea columnar object framework for effective analysis. In *EPJ Web of Conferences* (Vol. 245, p. 06012). EDP Sciences.

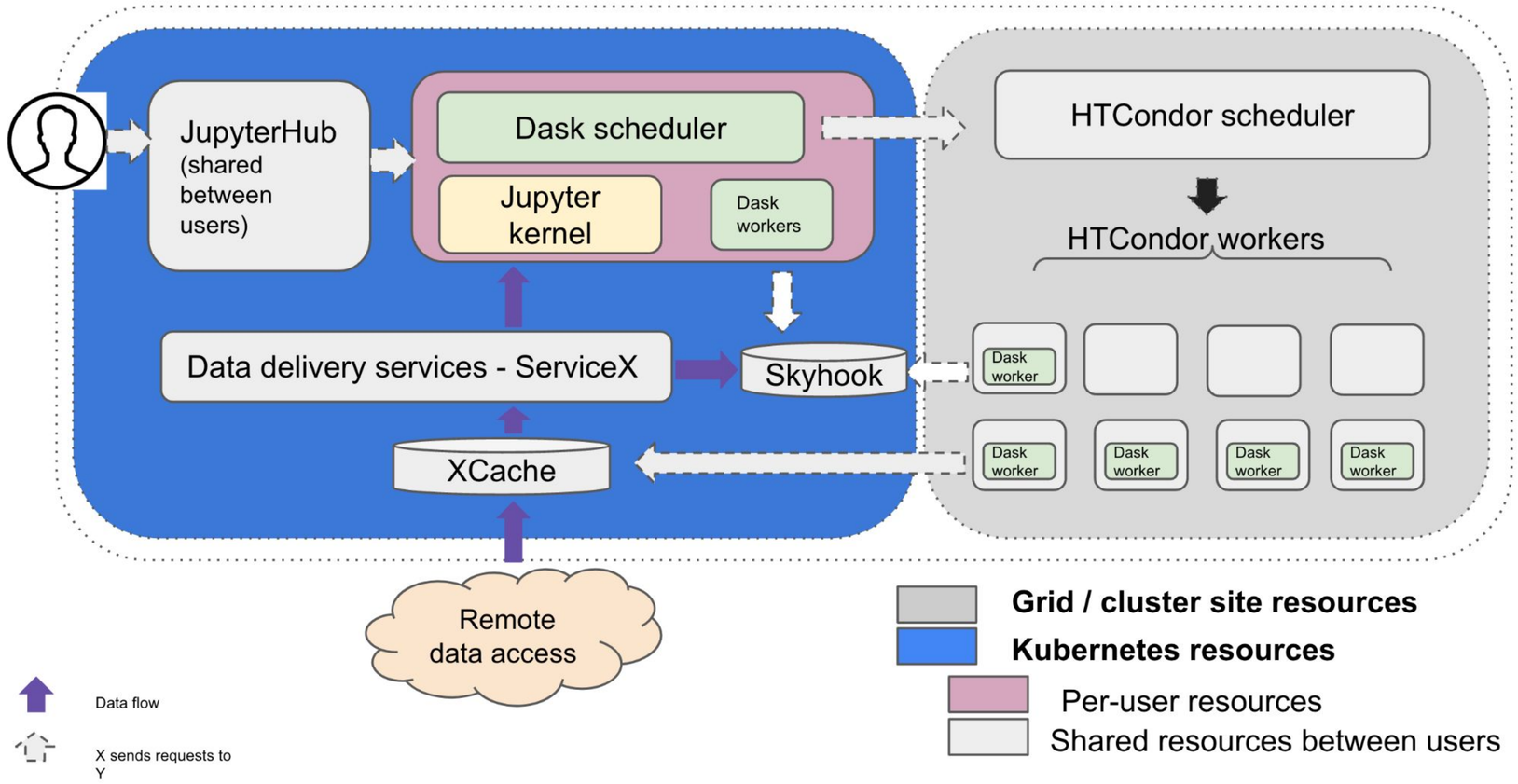
https://github.com/CoffeaTeam/coffea



Simplified diagram of hypothetical Analysis Facility currently used by user



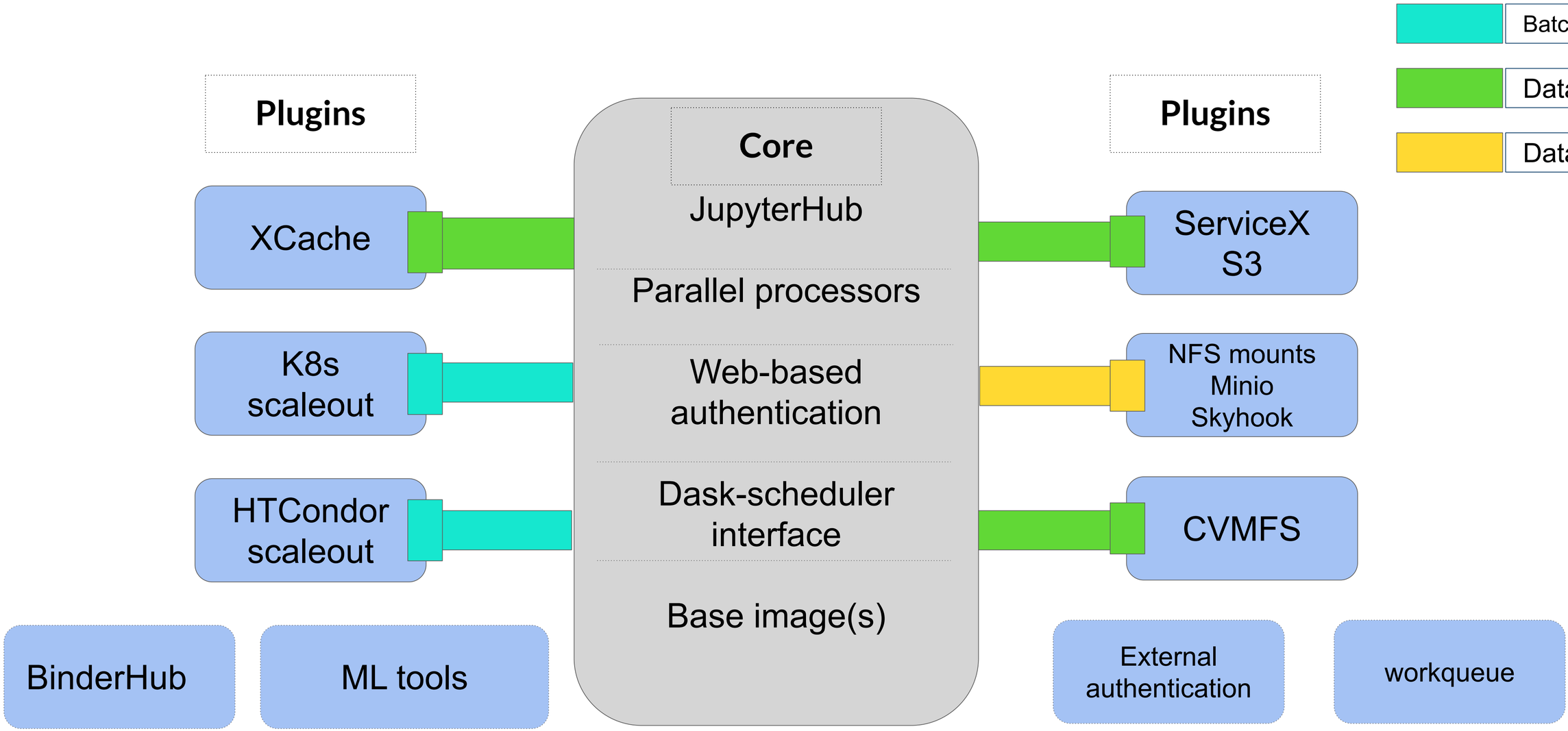
Coffea-casa Analysis Facility



Coffea-casa facility @ UNL is co-located at US.CMS Tier-2 at University Nebraska-Lincoln and other instance is co-located at US.ATLAS Tier-3 at University UChicago

[Coffea-casa team]

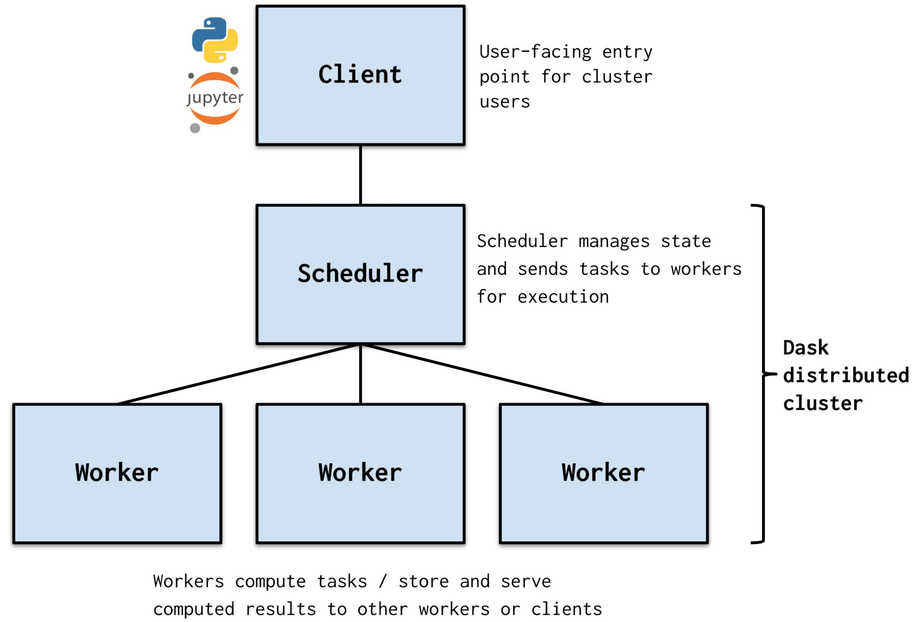
Coffea-casa AF components



[Coffea-casa team]

Building blocks: Dask

- Dask provides a task-management computational framework in Python (based on the manager-worker paradigm)
- Integrates with HPC clusters, running a variety of schedulers including SLURM, LSF, SGE and HTCondor via *“dask-jobqueue”*
- **This allows us to create a user-level interactive system via queueing up in the batch system**



Dask can be used inside Jupyter or you can simply launch it through Jupyter and connect directly from your laptop

Building blocks: Batch scale-out

- ***CoffeaCasaCluster:* extending HTCondorCluster integration for Dask**
  - To handle the customizations needed for the Coffea-casa environment, we developed the *CoffeaCasaCluster* object, an extension of ***Dask-jobqueue***'s *HTCondorCluster* object.
  - *CoffeaCasaCluster* ensures the Dask worker starts with the appropriate Docker container in the HTCondor batch system with appropriate configurations and with the firewall ports configured correctly.
- **Looking into new backends:**
  - *Workqueue* (http://ccl.cse.nd.edu/software/workqueue/)

Building blocks: AuthN & AuthZ

- **Authentication inside the system is independent of grid credentials**
- Coffea-casa facility uses **OpenID Connect (OIDC)**
  - CMS, ATLAS IAM for experiment specific facilities
  - CILogon with COmanage for Opendata facility
- **Token authentication**
  - Pre-generated token for authentication with HTCondor, required for Dask scale-out to the larger resources
  - Pre-generated data access token for authentication with a local XRootD server (XCache)
  - R&D on enabling tokens for ServiceX data delivery service
- **Other credentials**
  - Generated X.509 credentials (including a CA, host certificate, and user certificate) for use in Dask for TLS as well for user communication to Dask scheduler endpoint
    - Security: TLS enabled communication between workers and scheduler by default

Building blocks: orchestration

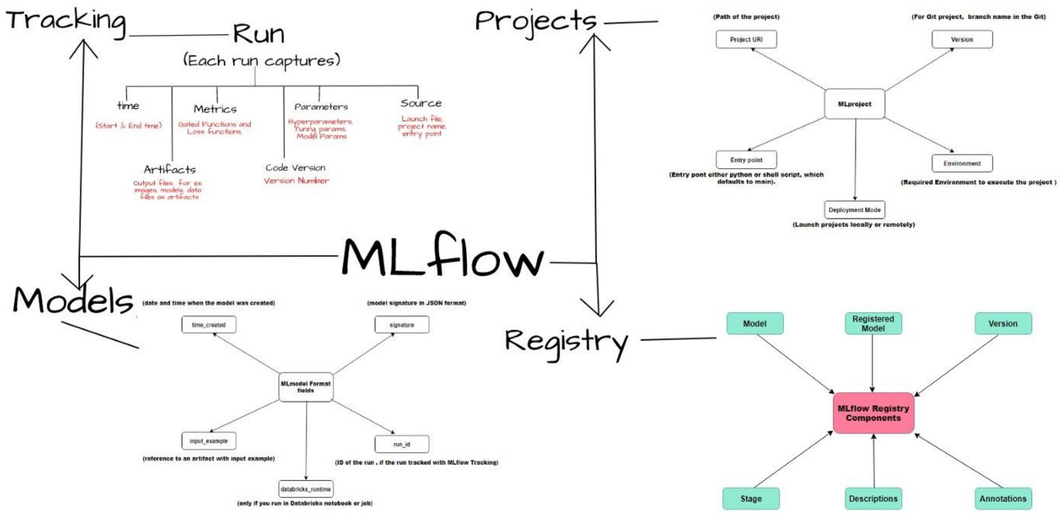
For users:

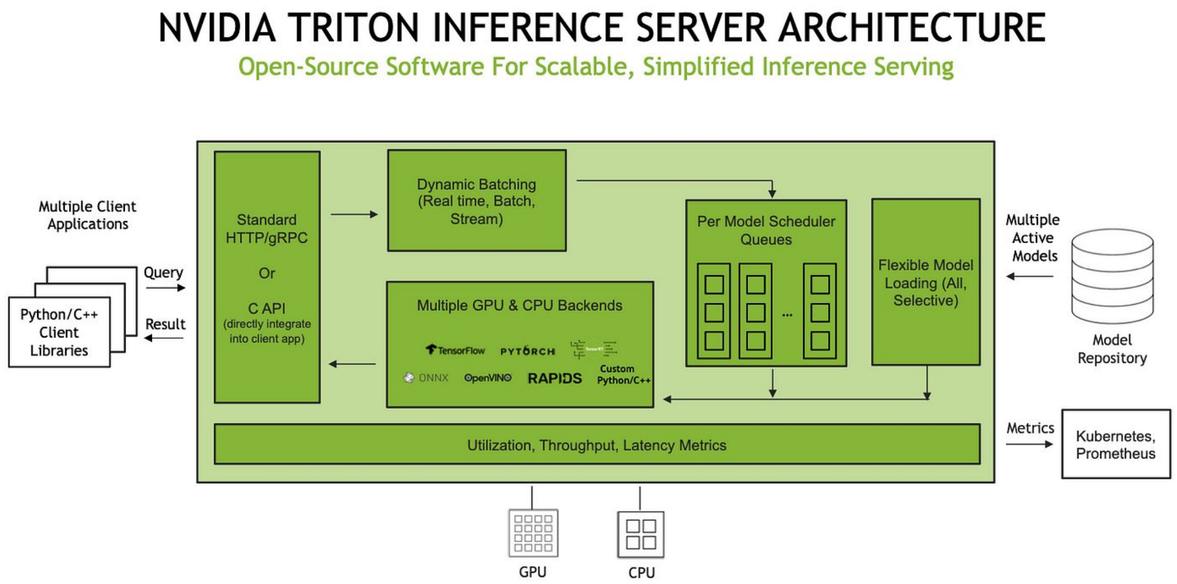
- Highly customized **“analysis” Docker container(s)**
- **Planning to add Binderhub support**
  - It will allow users to share reproducible interactive computing environments from their code repositories at coffea-casa

For developers:

- All features are **incorporated into a Helm chart** (Kubernetes packaging format)

Building blocks: machine learning services and tools





MLflow

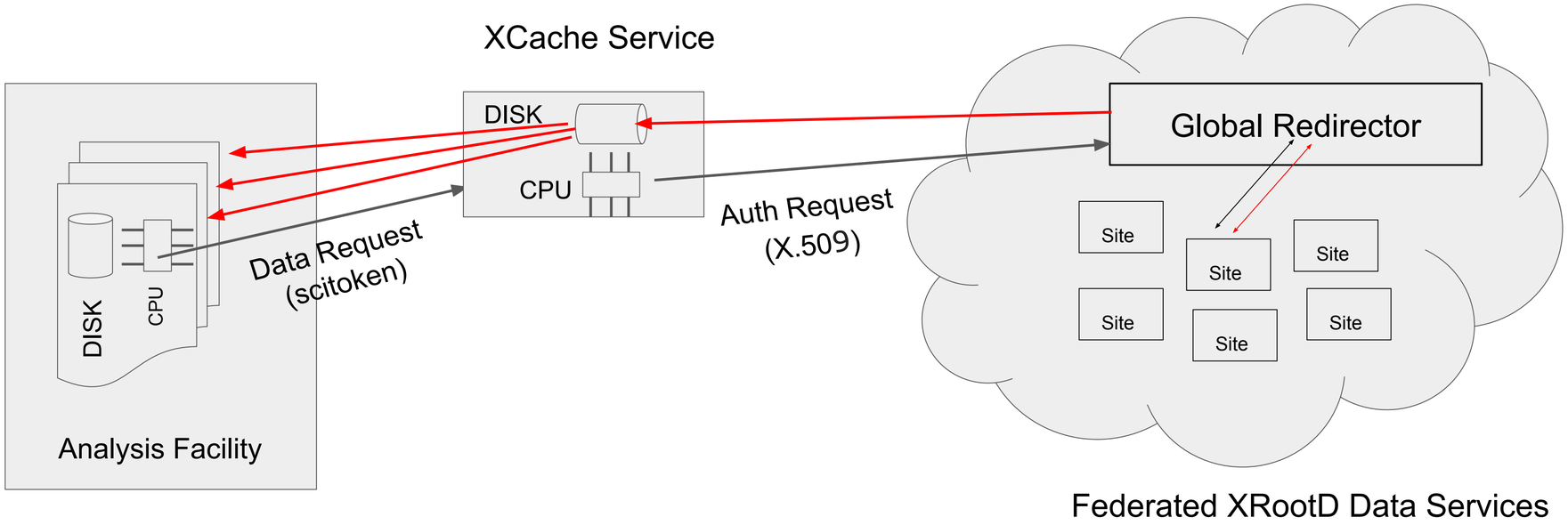
- Provide a central store to manage models (their versions and stage transitions)
- Allow packaging and re-deploying models
- Allows easily to reproduce code
- Provides easy tracking of ML experiments

Triton Inference Server

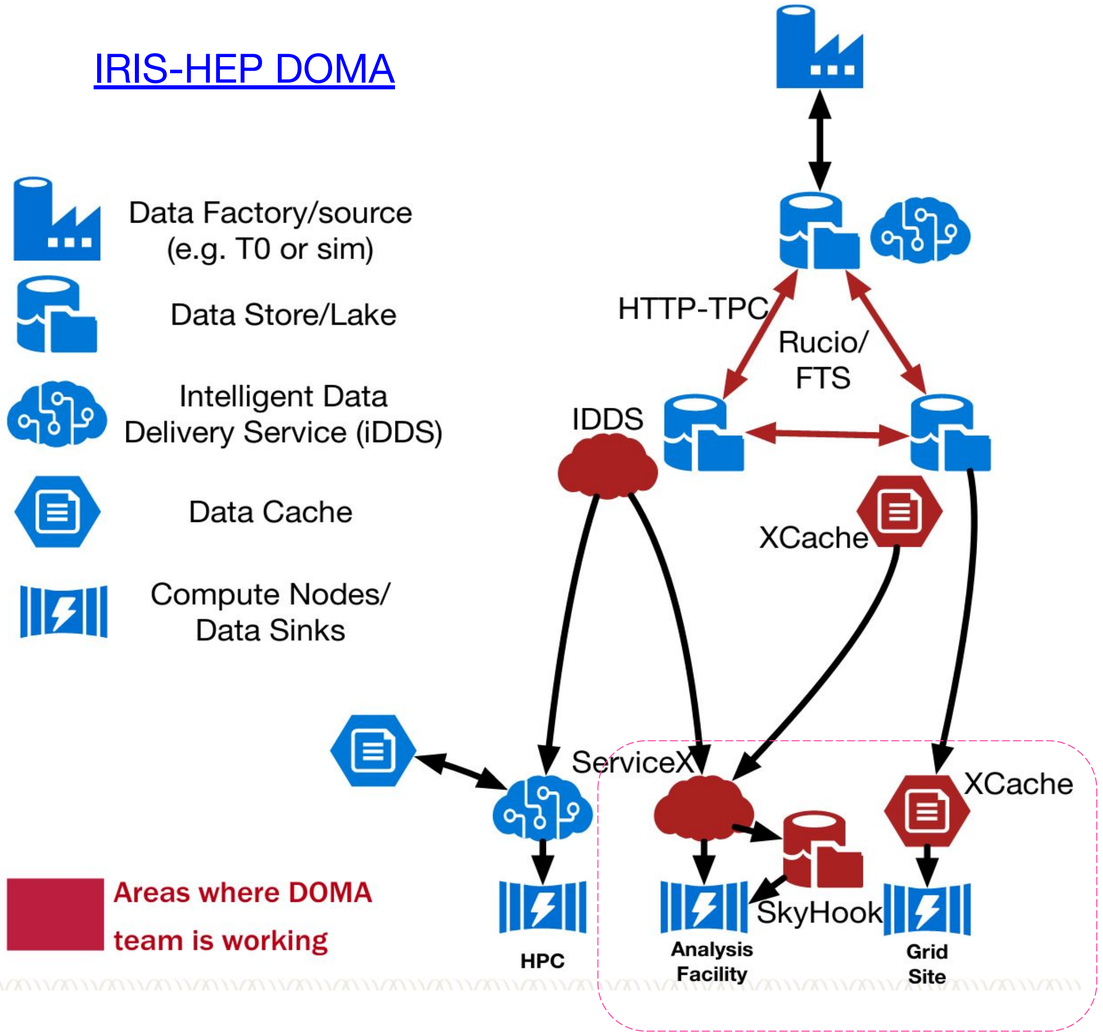
- **Support for various deep-learning (DL) frameworks**
- **Simultaneous execution**—Triton can run multiple instances of a model, or multiple models, concurrently, either on multiple GPUs or on a single GPU.
- **Dynamic scheduling and batching**
- **Model ensembles**

Building blocks: data access (e.g. XCache)

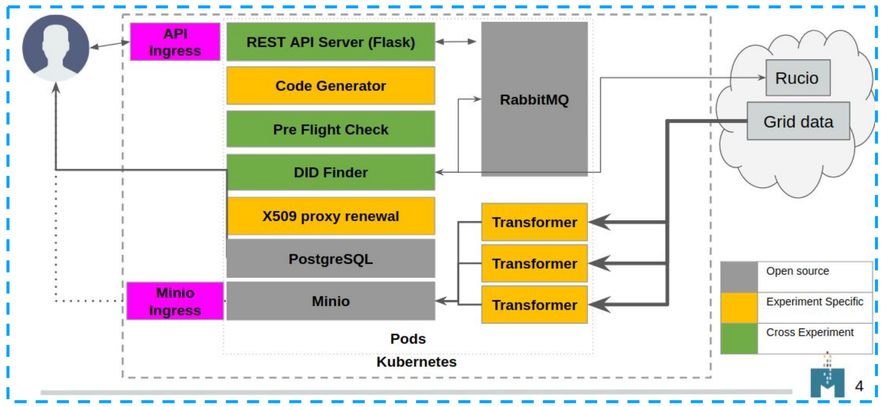
- For speeding up data access Nebraska Tier-2 hosts an **XCache service with 90TB of cache space**
- Access data hosted by an HEP experiment:
  - *no GSI credential within the facility,* **the auto-generated data access scitoken can be used to authenticate with an proxy service based on XRootD/XCache**



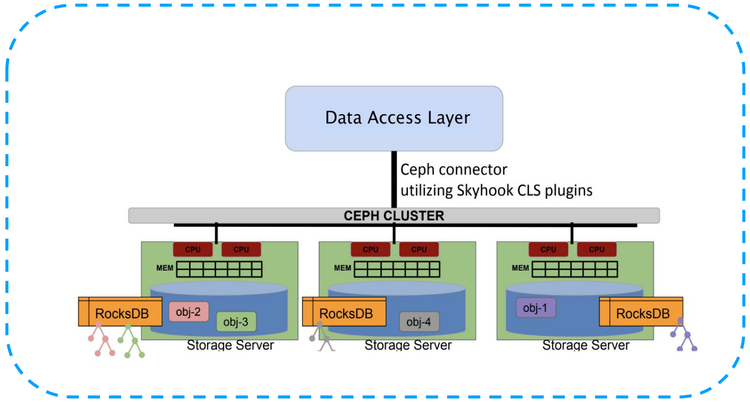
Building blocks:
Data Organisation and Management and Access (DOMA)



Coffea-casa



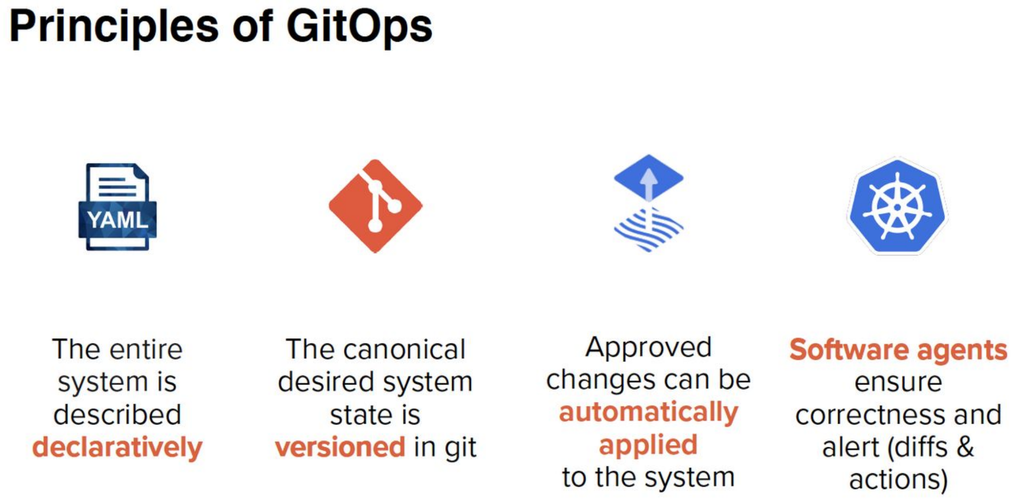
Servicex - *data extraction and delivery delivery service for columnar analysis*

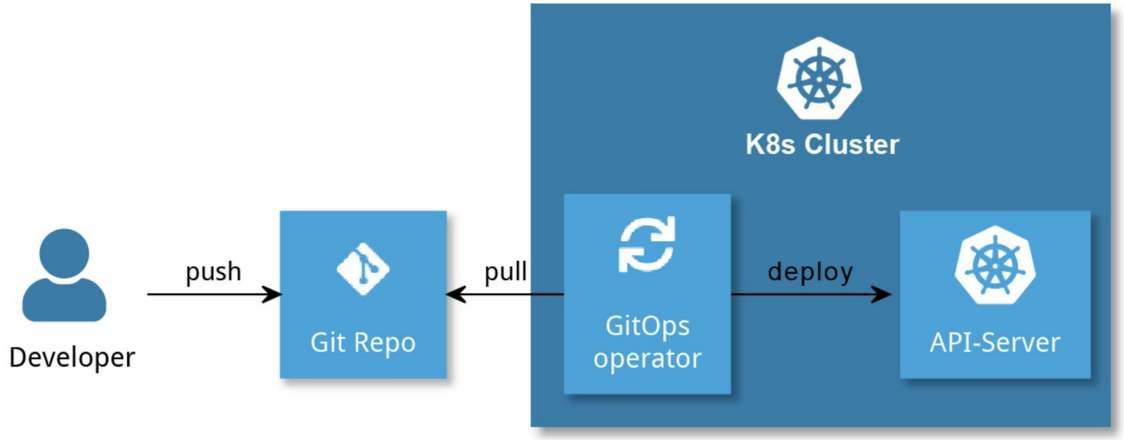


Skyhook - *extension of the Ceph for the scalable storage of tables and for offloading common data management operations*

New development strategies - GitOps

- **GitOps** defined as a model for operating Kubernetes clusters or cloud-native applications (e.g. coffea-casa AF)
- ***Concept: infrastructure-as-a-code***
  - Allow for rapid collaboration, better quality control, and automation (CD/CI).
  - AF is easily handled via a collaborative group of administrators in a deterministic manner.
  - Allows easily packages the core infrastructure (e.g., removing the site-specific passwords and secret keys) as a Helm chart.



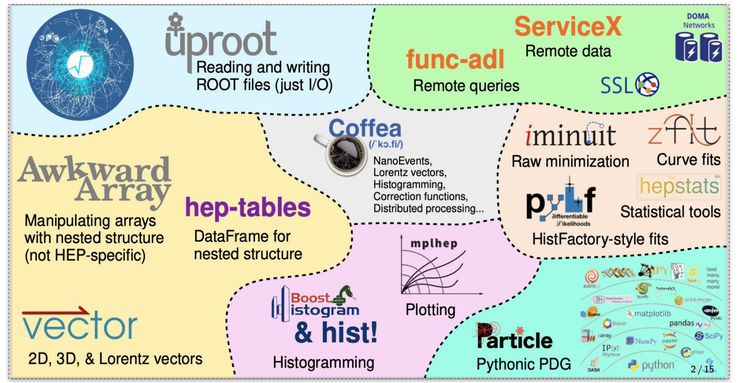


Analysis Grand Challenge

Motivation:

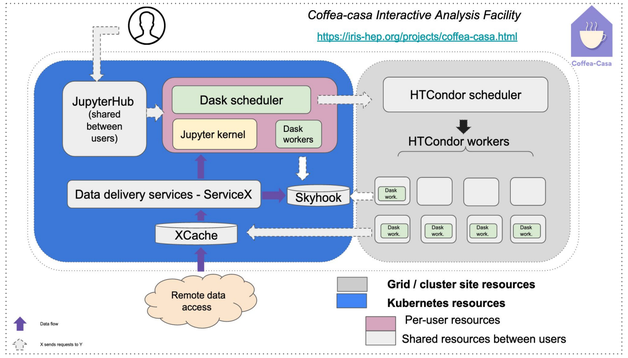
- Allow coping with HL-LHC data sizes by rethinking data pipeline
  - Evaluating the new Python analysis ecosystem and integrating a differentiable analysis pipeline
- Provide flexible, easy-to-use, low latency analysis facilities

Analysis Tools



Analysis Facilities

(coffea-casa AF or any other facility matching tech.requirements)



Execution of AGC analysis benchmark

Check the talk from A. Held "Physics analysis workflows and pipelines for the HL-LHC" on Thursday

Analysis Grand Challenge **will be conducted during 2021–2023, leaving enough time for tuning software tools and services developed as a part of the IRIS-HEP ecosystem before the start-up of the HL-LHC and** ***organized together with the US LHC Operations programs, the LHC experiments and other partners.***

Thank you!
Q&A