DOMA “Challenges and Prototypes”

Objective: Data Challenges and Data Lake prototypes driving the 2021 work plan

The WLCG/HSF workshop on storage (November 2020) exposed the needs of connecting the DOMA R&D concepts, ideas and technologies with the reality of the experiments and the sites providing resources. There is a diversity of opinions in the community, but enough quorum to start prototypes. There is no push for unique solutions, the R&D activities purpose is to explore new computing models and expand the current possibilities for sites and federations needing to adapt to new scenarios. The goal for next year is to set up a prototype infrastructure and identify relevant and known workflows to run on this prototype. These workflows can be stripped down to metrics/numbers to make the comparison between today's requirements and the future HL-LHC scenario. Like this we can have a comparison of what and how we do today and compare it to what and how we could do it in +5 years making use and evaluating the impact of the new concepts, ideas and technologies.

From the HL-LHC review [doc] (next iteration of the review will be in about a year)

- Implementing a computing model in the (political) reality
- Data lake approach and load on parties involved [...] it has to be evaluated [...] whether this scenario can be implemented in all major participating countries
- Recommendation: [...] include in its programme of work specific tasks and their associated milestones for quantifying the potential reductions in equipment costs and manpower savings resulting from the data caching & streaming strategy for delivering data, as well as from the usage of different kinds of hardware according to the expected usage by the experiments.
- Recommendation: [...] make sure that sites expected to be at the core of the foreseen WLCG data delivery network are engaged and play an active role in the project so that their inputs are taken into account early on in the design and testing phases of the system

Proposal for a Kick-off

- **Infrastructure:**
  - Defining two **Data Lakes prototypes** (US and Europe). A single DOMA Rucio instance? One per experiment?
    - Experiments define their data lakes view at the RUCIO level, based on a subset of sites (storage endpoints) that gives resources to some or all the experiments. Understand early QoS mappings.
    - Provide enough resources to measure realistic performances using standard workflows: ~1PB and ~1k cores.
      - Some EU sites already volunteered, US initiative advancing.
  - Interested sites/federations providing state-less storage (**caching**) being able to get data from the lake
  - Interested sites/federations accessing data remotely (storage-less sites)
  - External resources. Leverage data ingress/processing/egress from/to the Data Lakes on (at least one) **HPC** center, site opportunistic resources or **cloud** provider.
  - Understand how **Analysis Facilities** would integrate with Data Management systems and eventually with Data Lakes. Good to have some AF participation in the near future.

- **Activities:**
  - **Data challenge for production processing** (10PB/day initiative). ATLAS and CMS will take 0.5 exabyte of RAW data. The challenge is to process an exabyte ”as quickly as possible” after the end of the annual data taking.

        - Document the system wide implications on the infrastructure for different scenarios. This has to assess the impact on archives, staging areas, networks on the data sources and consuming sites of all types (processing centers, HPC systems and (commercial) Clouds)/
        - Follow-up on the scalability of the orchestration tools: Rucio, FTS.
        - Tape retrieval bandwidth measurements at T1s
            - With ongoing data taking and without
    - **Implementation of high-level workflows on the Data Lake prototype:** identify and agree on a few high-level workflows, for example: Reconstruction, Derivation and Analysis. For Derivation and Analysis especially the impact of the new compact analysis data formats has to be understood. Where and how will mini/nano AODs be accessed? What is the RUCIO's role in this (move the produced m&n-AODs to an Analysis Facility and then leave it to the AF?)
        - Identify which are the requirements now and which are the ones for the HL-LHC scenario
            - Storage needs (real volume, versions, etc.), bandwidth needs, iops, data lifecycle (permanent vs. non-permanent storage), content delivery and caching usage, etc.
        - Run the required processing to produce analysis-like datasets, e.g. AOD, min-AOD or nano-AOD
        - Exercise file workflow (QoS transition) after the data is produced, data is moved to a different QoS
        - Distribution of the analysis-like datasets to the processing sites, data caches and Analysis Facilities

- **Evaluation**: Based on the defined activities and the infrastructure, collect metrics and see how they compare with the current infrastructure: overall **time for completion**, **storage** used, **bandwidth** used, CPU **efficiency**, **operation** burden, etc.
    - Storage-less sites, remote processing vs. local download, latency impact, feedback from *local* scientific communities. Latency hiding mechanisms on the WNs. What is the impact of an increase in the number of cores on the WN scratch space?
    - State-less storage sites (streaming caches), latency impact, operation of caches, cache usability, feedback from *local* scientific communities
    - External resources. Are we able to deploy a content delivery (caching) service at the "edge" of these sites? Is this better than remote access? security hurdles?
    - Storage operations at the "data-processing-oriented" facilities:
        - Time spent by sysadmins on managing a stateful storage (DPM/dCache/…) vs. a possible stateless storage (XCache)
        - Time spent by the experiment data management team to "cure storage namespaces" abroad vs. leveraging/managing caches
        -

Data "needs"	Metrics and measurables	Datalake pros/cons
-Data volume (?) -Percentage of Data on disk vs Archive -How many sites contribute? -Which sites contribute: T1s, big T2s, all? -How often reco is done? (continuous activity during data taking, sporadic during LSs) -...more...	-Bandwidth per job? -File size? -Processing pattern: suitable for caching (sequential) or sparse reads? -IOPS needs (to evaluate latency impact, QoS benefits,...) -...more..	-How could caches help in processing (engage more sites, cpus, clouds, hpcs, reducing latency, etc ? -How could caches help on diminishing stateful data storage needs? -How could file workflows/QoS help in defining data lifecycles (files auto-transition after some time to cheaper storage/archive)? - Penalties for extra data streaming?
