ESCAPE/DIOS: use case presentation from science projects

In the joint session of the WP2/WP5 workshop we would like to collect input from the science projects being part of ESCAPE. WP2 explored many technical solutions for the Data Lake in the fortnightly meetings of the last 3 months. WP5 held its first workshop starting defining the analysis platform. We ask for the WP2/WP5 joint session of the workshop each science project to prepare a 15 minutes presentation (followed by a 15 minutes discussion) explaining one (the most) representative use case that ESCAPE should be able to enable by the end of the project. In doing so, the presenter should consider the input we gathered in the last three months and try to explain which components/workflows would fit in a possible implementation, and which would need improvements/enhancements.

Through this process we should be able on wednesday to start sketching an architecture and an implementation. We also would have a concrete set of use cases to start with and we can then define metrics of success based on them.

To facilitate the process, through the presentation of the use cases, one should be able to get an answer to the following questions:

Data production

- Which is the estimated number of files produced per unit of time (per year/month/day, per run or per observation)
- Estimated file size of primary (raw) data?
- Is the primary data compressed afterwards? If so which ratio?
- Should this primary data be kept forever (archived)?

Data model

- General description of your computing and processing model, ie.
    - Where data is collected and stored? How many more copies you replicate and where (single site vs. distributed copies)?
    - Do you produce different data formats? which ones? how often do you produce them (raw vs. reconstructed vs. analysis-ready data)?
    - Do you have data lifecycles (hot/cold/archive/delete) defined and for which type of data?
    - Do you have processing campaigns and if so how often? And which data do they produce?
    - Do you reprocess the data periodically (ie. once a year) ?
    - Do you need searchable metadata? Do you need dynamic definition of metadata?

Data access and data processing

- Is there a need for (fast, quasi-online) data processing of the primary data before data is usable for analysis? If so, describe the steps.
- Which protocols do you use to transfer and access data? Or which tools do you plan to use?

- Based on the topology and size of your current or future computing infrastructures:
    - Do you see a benefit of a data caching layer for file re-usability?
    - Do you see a benefit of a read-ahead cache for latency hiding purposes?
    - Do you see a need for different quality of service (QoS) for storage (other than the classic disk and tape storage endpoints)?
- Do you plan or do you have an interest in having a file popularity management service?
- What is the typical data access pattern, ie.
    - Do you copy the data to the processing node and access it locally or open files remotely and read directly from the storage?
    - Do you consume full files (read from a to z) or just some random parts within the file?
    - Did you evaluate the impact of remote (rather than local) data access depending on RTT and bandwidth?
    - Do you process (or intend to process) data stored on tape and how do you plan to do that?
- Do you have a Workload Management System (WMS)? Does this WMS also orchestrate data movement? or you intend to have a different system to orchestrate data movement?
- Do you need some kind of CLIs, APIs and/or Web Interfaces?
- Do you have need for MPI jobs and are you considering the usage of HPC facilities?
- Data loss and recovery mechanisms:
    - What is the impact of data loss?
    - What is the impact of having temporary unavailable data?

Data access control:

- What is the policy for embargo and open access?
- Do you foresee anonymous users to access the data?
- What are the ACL (Access Control) needs from a global point of view for your community? what are the main use cases?
- Do users belong into groups with specific privileges?
- How flexible would your community be regarding different authentication/Authorization tools/schemes, ie. Open-ID, X509, Sci-Tokens, SAML etc.

AAI general questions

AAI expert/contact person:

It would be good to have a list of the AAI “experts” for each experiment. The AAI expert knows about the experiment computing model and can answer AAI-related questions

How many active end-users has your community?

- 10-100
- 100-1000

- > 1000

How do your users access your experiment computing and analysis tools/services? (multiple choices allowed)

- via a Web browser
- via a terminal session
- via a native application installed on their computer
- other

How do you currently manage user authentication at your computing and analysis tools/services? (multiple choices allowed)

- identity federation (EduGAIN, or similar)
- X.509 certificates
- application-specific credentials
- ssh keys
- other

How do your currently manage user registration and lifecycle management for the experiment?

Provide a description of how users gets registered into your experiment, how they are assigned privileges and list the tools that you're currently using.

How is authorization structured in your computing/analysis services? Do you rely on groups/roles to grant different access privileges to your users? Do you use other mechanisms?

Provide a description of how you structure your users and grant them different access privileges across your experiment applications/services and the tools that you're currently using.

How is authentication and authorization implemented for data access?

Provide a description of which protocols are used for accessing data (e.g., HTTP, FTP, other) and how do you implement currently authentication and authorization.