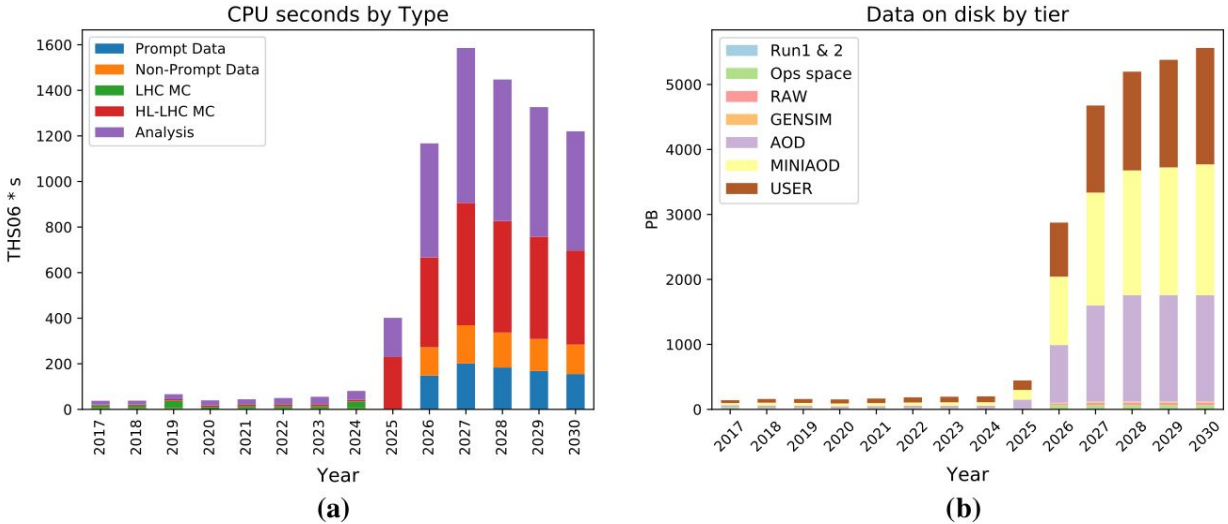
Analysis Platforms, Analysis Facilities

Linking data access and analysis with computing resources

Looking into the future

- Relevant question in scientific communities: FAIR data and analysis facilities
- How to expand the current ESCAPE infrastructure to serve analysis reproducibility ?
- *Example: CMS estimated CPU (a) and disk space (b) resources required into the HL- LHC era, using the current computing model with parameters projected for the next 12 years [1]*



[1] E Sexton-Kennedy 2018 J. Phys.: Conf. Ser. ***1085*** 022006 https://iopscience.iop.org/article/10.1088/1742-6596/1085/2/022006

Layers ...

TOP

bottom

- Analysis software
- Analysis framework
  - facilitates the interaction between the analysis software and the common reduced data format
- User interface
  - Jupyterhub / Binderhub
- Batch infrastructure
  - PBS, Slurm, LSF, and HTCondor
  - submit node to the batch system may not be the same node as the interactive node for users on the cluster
- Storage infrastructure
  - Software to transform data formats
  - Data delivery system (XRootD, HTTP..) ➜ accessed directly by the distributed analysis application or via the transformation layer
  - Distributed file system layer (HDFS, CephFS, EOS)

[2] Analysis Cyberinfrastructure: Challenges and Opportunities. K. Lannon, P. Brenner, M. Hildreth, K. Hurtado Anampa, A. Malta Rodrigues, K. Mohrman, D. Thain, B. Tovar https://arxiv.org/abs/2203.08811

Discussion

1. Technology
2. Scientific analysis model – needs
3. Variability: multi-threading, parallel processing, caching
4. Multi-user, multi-tenant
5. Collaborations
6. Examples

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?
2. What analysis MODELS should a platform support? Use cases demand different things!

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?
2. What analysis MODELS should a platform support? Use cases demand different things!
3. VARIABILITY and multi-threaded processing?
    a. When would you use caching and when would you not? What type of caching?
    b. How to extend to multi-node / multi-datacentre? *e.g. running same pipeline on 2 datasets (SKA and LOFAR) in different locations and combine results*
    c. How to efficiently stream processing of data on a remote storage?
    d. How to optimise parallel batch analyses (HTC)?
    e. How to handle intermediate results? *e.g. in Radio Astronomy moving back and forth is not optimal*

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?
2. What analysis MODELS should a platform support? Use cases demand different things!
3. VARIABILITY and multi-threaded processing?
    a. When would you use caching and when would you not? What type of caching?
    b. How to extend to multi-node / multi-datacentre? *e.g. running same pipeline on 2 datasets (SKA and LOFAR) in different locations and combine results*
    c. How to efficiently stream processing of data on a remote storage?
    d. How to optimise parallel batch analyses (HTC)?
    e. How to handle intermediate results? *e.g. in Radio Astronomy moving back and forth is not optimal*
4. How to scale to MULTI-USER, MULTI-TENANT systems?
    a. Security, public/private data, data/notebook sharing, data provenance, citability, etc.
    b. DataLake is somewhat in the middle: I can not really hide my data from other tenants, or make it publically available to “Jane Doe”

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?
2. What analysis MODELS should a platform support? Use cases demand different things!
3. VARIABILITY and multi-threaded processing?
    a. When would you use caching and when would you not? What type of caching?
    b. How to extend to multi-node / multi-datacentre? *e.g. running same pipeline on 2 datasets (SKA and LOFAR) in different locations and combine results*
    c. How to efficiently stream processing of data on a remote storage?
    d. How to optimise parallel batch analyses (HTC)?
    e. How to handle intermediate results? *e.g. in Radio Astronomy moving back and forth is not optimal*
4. How to scale to MULTI-USER, MULTI-TENANT systems?
    a. Security, public/private data, data/notebook sharing, data provenance, citability, etc.
    b. DataLake is somewhat in the middle: I can not really hide my data from other tenants, or make it publically available to “Jane Doe”
5. Extending the current ESCAPE infrastructure for EOSC-Future: COLLABORATION?
    a. What is a good process to understand what infrastructure is needed?
    b. Who could provide computational resources?
    c. Is it better to have the analysis platform as an add-on to tiers already existing or start everything on cloud resources?

Time for discussion !

1. What type of TECHNOLOGIES should we take on board in future projects?
2. What analysis MODELS should a platform support? Use cases demand different things!
3. VARIABILITY and multi-threaded processing?
    a. When would you use caching and when would you not? What type of caching?
    b. How to extend to multi-node / multi-datacentre? *e.g. running same pipeline on 2 datasets (SKA and LOFAR) in different locations and combine results*
    c. How to efficiently stream processing of data on a remote storage?
    d. How to optimise parallel batch analyses (HTC)?
    e. How to handle intermediate results? *e.g. in Radio Astronomy moving back and forth is not optimal*
4. How to scale to MULTI-USER, MULTI-TENANT systems?
    a. Security, public/private data, data/notebook sharing, data provenance, citability, etc.
    b. DataLake is somewhat in the middle: I can not really hide my data from other tenants, or make it publically available to “Jane Doe”
5. Extending the current ESCAPE infrastructure for EOSC-Future: COLLABORATION?
    a. What is a good process to understand what infrastructure is needed?
    b. Who could provide computational resources?
    c. Is it better to have the analysis platform as an add-on to tiers already existing or start everything on cloud resources?
6. What are some EXAMPLES of other analysis platforms that you know are successful?