Processing in IS-ENES2

- Context of processing in IS-ENES2

Summary of IS-ENES2 requirements for processing

Detailed requirements

1. **Data processing desiderata: input**

   i. **What data are to be processed? What are their:**

   - **Typologies:** Hierarchical collection of data, are characterized by entries from controlled vocabularies.
   - **Volume:** Very high volume data, size of individual files ranges from mega to gigabytes, but processing is normally done at collection level involving multi terabyte input collections.
   - **Velocity:** Low velocity – data collections are growing (in a controlled manner) and new versions of existing data products are available in the data federation
   - **Variety:** Very low – data is based on highly structured data items (well defined binary data types, representing multi-dimensional date entities, e.g. netCDF). Data entities are organized in well structured hierarchies (structured according to time, variables, project characterization etc.).

   ii. **How is the data made available to the analytics phase? By file, by web (stream/protocol), etc.**

Normally the data is made available to analytics based on a local or mounted file system. A separate data-import step is responsible for filling up the input data pool.

   iii. **Please provide concrete examples of data.**

Temperature and precipitation according to various scenarios, generated by different climate models. Compute statistics to compare characteristics of the different climate models or climate indices – characterizing the individual climate model performance.

2. **Data processing desiderata: analytics**

   i. **Computing needs quantification:**

   - **How many processes do you need to execute?** Highly dependent on use case. Computing is more I/O bound then processor bound.
   - **How much time does each process take/should take?** Also very use case dependent, some multi-model analytics may run for days on a small cluster – others for minutes – as before: time is more dependent on data access characteristics.
   - **To what extent processing is or can be done in parallel?**

Most processing would benefit from a parallel map reduce phase, where first distributed data near pre-processing is done, reducing the amount of data to be transferred. Thereafter more complex, shared disk/memory parallel analytics is done on the parts from the map-reduce phase.

Some analysis use cases can benefit from shared memory and distributed memory parallelism to accelerate time to solution. Also to notice: some analysis phase are well suited for parallel approach (such as one process per model for example)

   ii. **Process implementation:**

   - **What do you use in terms of:**
     - **Programming languages?** Python, R, C, C++, Fortran
     - **Platform (hardware, software)?** Linux clusters, mostly open source software basis
     - **Specific software requirements?**
   - **What standards need to be supported (e.g. WPS) for each of the above?** Data near processing for ENES/ESGF sites is based on the OGC WPS standard.
   - **Is there a possibility to inject proprietary/user defined algorithms/processes for each of the above?** Yes – by contributing to open source data analytics software projects of various kinds (UV-CDAT, birdhouse, ESMValTool,etc).
   - **Do you use a sandbox to test and tune the algorithm/process for each of the above?** Yes – concrete test procedure is project dependent.

   iii. **Do you use batch or interactive processing?** Both.

   iv. **Do you use a monitoring console?** Yes.

   v. **Do you use a black box or a workflow for processing?**

   - **If you use a workflow for processing, could you indicate which one (e.g. Taverna, Kepler, proprietary, etc.)**

The choice of workflow engine analysis is project or framework specific, e.g. proprietary workflow wrappers, dispel4py.

- **Do you reuse sub-processes across processes?** Analysis project dependent, mostly not.

vi. Please provide concrete examples of processes to be supported/currently in use;

Simple: Subsetting of data, mean etc. statistics, downscaling of data, interpolation of data, climate indices calculation (ENSO, NAO, PDO, etc).

-Complex: vegetation modeling, geographical mosquito dispersal

3. Data processing desiderata: output

i. What data are produced? Please provide:

- **Typologies:** Various: netCDF files, graphics, text, logs
- **Volume:** Various: normally orders smaller then input data
- **Velocity:** High – depending on analysis activity
- **Variety:** High

ii. How are analytics outcomes made available? Different means: only per researcher or research group on file system, some outputs are published in catalogues and accessible via web and python notebook for example.

4. Statistical questions

i. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?

Data is collected according to the requirements and pre-defined characteristics defined for climate model intercomparison projects.

5. Will questions/hypotheses be generated or refined (broadened or narrowed in scope) after the data has been collected? (N.B. Such activity would not be good statistical practice)

The requirements and characteristics are refined after every round of model intercomparison projects to improve the next round and to be react on the new possibilities new technical infrastructures provide (e.g.improved processing power to support larger ensembles and finer resolution in models).

6. Statistical data

i. Does the question involve analysing the responses of a single set of data (univariate) to other predictor variables or are there multiple response data (bi or multivariate data)? Depending on analysis activity.

ii. Is the data continuous or discrete? Discrete.

iii. Is the data bounded in some form (i.e. what is the possible range of the data)?

Data represents several hundreds physical quantities (temperature, precipitation, wind speed, etc.) and in that sense are bound by physical laws.

iv. Typically how many datums approximately are there?

Data are stored on grid point covering the entire Earth System influencing the climate (atmosphere, ocean, sea ice, land…). So there are several thousands of data.

7. Statistical data analysis

i. Is it desired to work within a statistics or data mining paradigm? (N.B. the two can and indeed should overlap!)

Statistics are very important in climatic analysis as we are looking after robust and significant signal.

ii. Is it desired that there is some sort of outlier/anomaly assessment? Yes – but on a Petabyte scale difficult to achieve.

iii. Are you interested in a statistical approach which rejects null hypotheses (frequentist) or generates probable belief in a hypothesis (Bayesian approach) or do you have no real preference?

Needs more details. Yes a priori. The range of scientific analysis done using the data of our RI is very large. But those complex analyses are usually done within the scientific teams not by the RI itself.

Formalities (who & when)

Go-between	Yin Chen
RI representative	Sylvie Joussaume <sylvie.joussaume@lsce.ipsl.fr> Francesca Guglielmo <francesca.guglielmo@lsce.ipsl.fr>
Period of requirements collection	Oct -Nov 2015
Status	Completed