Processing in IAGOS

Context of Processing in IAGOS

Summary of IAGOS requirements for Processing

Detailed requirements

1. Data processing desiderata: input

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

- **Typologies**. tabular data
- **Volume**. per flight 1Mo
- **Velocity**. human operation (3 days for validation / 2-6 months for calibration)
- **Variety**. homogeneous

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

iii. Please provide concrete examples of data. example of NASA Ames provided

2. Data processing desiderata: analytics

i. Computing needs quantification:

- **How many processes do you need to execute?** One per level of data
- **How much time does each process take/should take?** For a flight : Few seconds for levels 0 to 2. Level 4: 3 hours
- **To what extent processing is or can be done in parallel?** Flight granularity

ii. Process implementation:

- **What do you use in terms of:**
  - **Programming languages?** Java, Python, Fortran
  - **Platform (hardware, software)?** Linux, opensource softwares
  - **Specific software requirements?**
- **What standards need to be supported (e.g. WPS) for each of the above?** none
- **Is there a possibility to inject proprietary/user defined algorithms/processes for each of the above?** no
- **Do you use a sandbox to test and tune the algorithm/process for each of the above?** no

iii. Do you use batch or interactive processing? yes

iv. Do you use a monitoring console? Nagios for hardware management. We plan to use nifi for dataflow management

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.)**
- **Do you reuse sub-processes across processes?** Black box so far

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

3. **Data processing desiderata: output**

i. **What data are produced? Please provide:**

- **Typologies** tabular
- **Volume** L2+L4 = 10Mo per flight
- **Velocity** L2: 2-6 months for calibration (L4 produced automatically when data level is changed)
- **Variety** homogeneous

ii. **How are analytics outcomes made available?** Available on download but no web-based workspace

4. **Statistical questions**

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

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) no

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)? no

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)? aircraft data (flight granularity)

iv. Typically how many datums approximately are there? One each 4 seconds during flight

7. **Statistical data analysis** NA

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

ii. Is it desired that there is some sort of outlier/anomaly assessment?

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?

Formalities (who & when)

Go-between	Yin Chen
RI representative	Damien Boularnger<damien.boulanger@obs-mip.fr> is the Manager of the IAGOS Information system
Period of requirements collection	Nov-Dec 2015
Status	Completed