Processing in EMBRC

Context of processing in EMBRC / St Andrews

Questionnaire answers from EMBRC/St Andrews on Processing available at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/40

Summary of EMBRC / St Andrews requirements for Processing

Detailed requirements

1. Data processing desiderata: input
   a. What data are to be processed? What are their:
   > Typologies Varies
   > Volume Varies
   > Velocity Varies
   > Variety Varies

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

   c. Please provide concrete examples of data.
   > It varies a lot. There are also data protection issues.

2. Data processing desiderata: analytics
   a. Computing needs quantification:
   a.1 How many processes do you need to execute?
   a.2 How much time does each process take/should take?
   > Varies

   b. Process implementation:
   b.1 What do you use in terms of:
   > Programming languages varies
   > Platform varies
   > Specific software requirements varies

   c. Is there a possibility to inject proprietary/user defined algorithms/processes for each of the above?
   > Yes

   d. Do you use a sandbox to test and tune the algorithm/process for each of the above?
   > Yes

   f. Do you use batch or interactive processing?
   > Both

   g. Do you use a monitoring console?
   > It varies

   h. Please provide concrete examples of processes to be supported/currently in use;
   > It varies

3. Data processing desiderata: output
   a. What data are produced?
   > Mainly results of analysis

4. How are analytics outcomes made available?
   > By paper

5. Statistical questions
   a. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?
   > Varies

   b. 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)
   > Hopefully not

6. Statistical data
   a. 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)?
   > Varies

   b. Is the data continuous or discrete?
   > Varies

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

   d. Typically how many datums approximately are there?

> Can be millions

7. Statistical data analysis
   a. Is it desired to work within a statistics or data mining paradigm?
   > Mainly statistical

   b. Is it desired that there is some sort of outlier/anomaly assessment?
   > desirable

   c. 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 a no real preference
   > Both

Formalities (who & when)

Go-between	Cristina Adriana Alexandru
RI representative	Charles Paxton
Period of requirements collection	November 2015
Status	