PANACEA massive data tutorial

WP 3

Author	Marc Poch
Affiliation	UPF

Status	internal report
Version	1.1
Release	draft
Date	2012-01-10
History:	

Abstract: This document is a short tutorial for PANACEA web service providers and workflow designers.

Contents

1. Introduction

This document is a list of recommendations to handle massive data for Service Providers and workflow designers.

2. TIPS FOR WEB SERVICE PROVIDERS

2.1. TOMCAT

- **TOMCAT 6 or TOMCAT 7** http://tomcat.apache.org/
- **The Apache Tomcat Native**

  Tomcat native library using the apache portable runtime: you must use it to improve performance.

  Package name: **libtcnative-1**

  http://tomcat.apache.org/native-doc/

  On your logfile catalina.out you'll find a sentence like this before installing the library:

  INFO: The APR based Apache Tomcat Native library which allows optimal performance in production environments was not found on the java.library.path

2.2. SOAPLAB

- **SOAPLAB 2.3.2**
- You must use **Soaplab output size limit patch** http://myexperiment.elda.org/files/3

  Soaplab.properties configuration file (we use a 1k limit):

  results.sizelimit = 1000

  If you have a web service which for some reasons requires a bigger output direct data size you can set a different parameter value for that ws.

  E.g. the statistics_analysis.vocabulary_analysis web service requires a bigger limit:

  results.sizelimit = 1000
  statistics_analysis.vocabulary_analysis.results.sizelimit = 50000

- OPTIONAL: **Limit the web service usage**

  Limit the amount of requests to your web services. If you have a few web services you may consider migrating your ACD files to this simple and efficient system.

  http://myexperiment.elda.org/files/4

- Soaplab **important tmp folders (inside tomcat)**:
  - /temp/_R_/RESULTS
  - /temp/_R_/SANDBOX
  - Webapps/soaplab2-axis/results

2.3. TEMPORARY FILES

This is a very variable topic and it really depends on the server and the web service provider wishes.

Recommendations:

- Check your **HD space** (we check every 30 minutes)
- Check the tomcat and soaplab **logs size** (we check every 30 minutes)
- Check the **32K max. amount of folders in one directory** limit (for linux) (we check every 30 minutes)

  /temp/_R_/RESULTS

  /temp/_R_/SANDBOX

- Erase **old** files (we daily erase files older than 2 days)

  /temp/_R_/RESULTS

  /temp/_R_/SANDBOX

  Webapps/soaplab2-axis/results

3. TIPS FOR WORKFLOW DESIGNERS

3.1. TAVERNA

Taverna 2.3.0

3.1.1. Taverna parameters

In-memory: YES

Always try to run workflows with the in-memory option activated.

Provenance: only if you need it. Interesting for the first tests of a workflow.

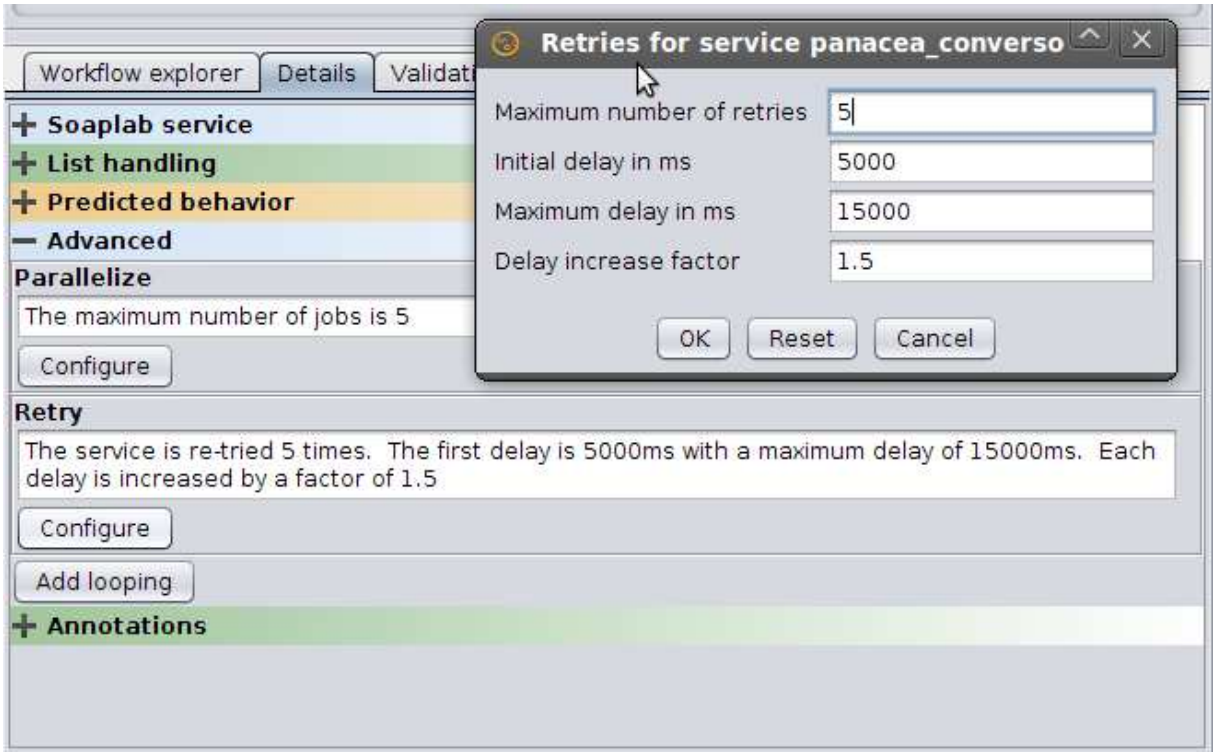
3.1.2. Workflows parameters

Parameters for Retries, parallelization and Polling can only be optimized by empirical observation. You must do some small tests!

RETRIES

Always use retries!

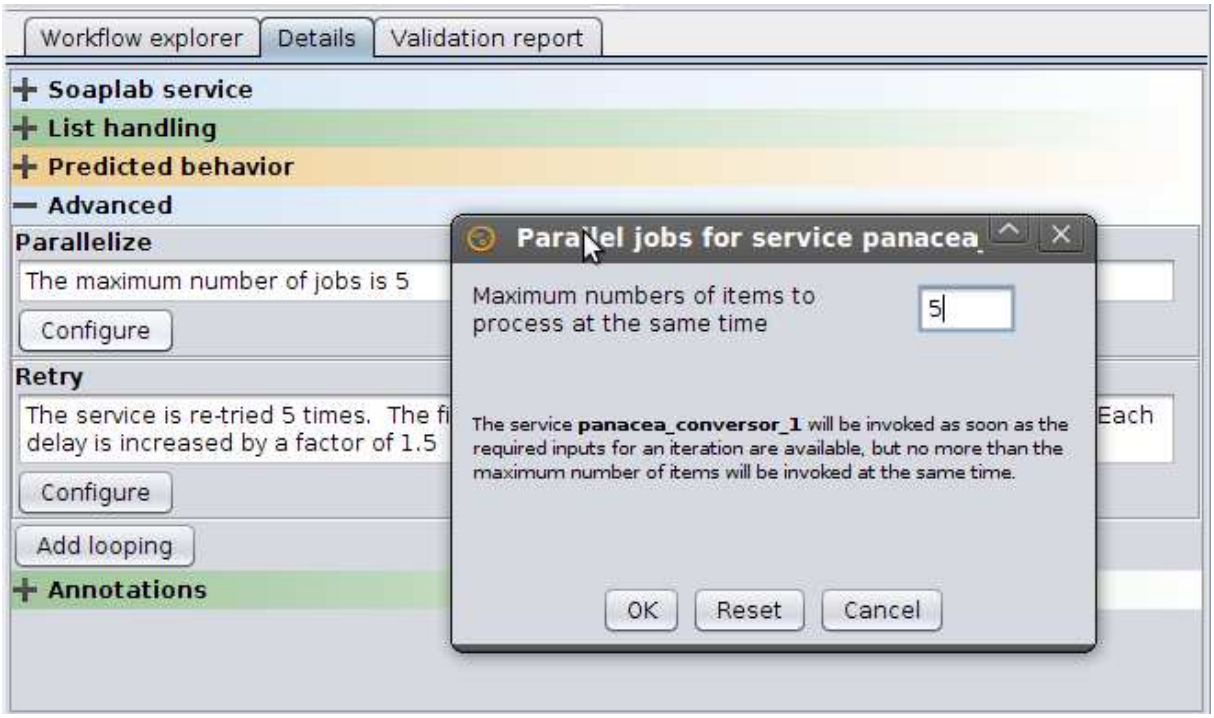
Watch the video: http://www.mygrid.org.uk/dev/wiki/display/taverna/Retries



PARALLELIZATION

Use carefully! Usually 3 and 5 for stress. It depends on the web service! Some may only handle 1 while other ws may accept more.

http://www.mygrid.org.uk/dev/wiki/display/taverna/Parallel+service+invocation



POLLING (only for Soaplab web services)

It is used to avoid timeouts when running long lasting executions in Soaplab web services.

If you use a big interval the system will be very slow and if you use a very small interval the network will be saturated.

http://www.mygrid.org.uk/dev/wiki/display/taverna/SoapLab