Experiences with Running Data Extraction Application using UNICORE

18. Juni 2013 | Lara Flörke, Mathilde Romberg

Mitglied der Helmholtz-Gemeinschaft

Outline

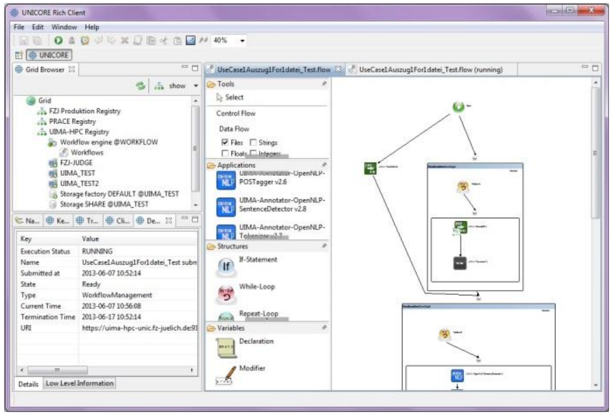
- Introduction
- The Use Case
- Observed Advantages and Restrictions
- Conclusion

Introduction

- UIMA-HPC:
  - BMBF funded research project
  - Collaboration partners
    - Fraunhofer SCAI
    - Taros Chemicals GmbH
    - Scapos AG
    - Forschungszentrum Jülich GmbH
  - Aims to realize an HPC-based solution for the automated analysis of multi-modal pharmaco-chemical document databases

Introduction (cont.)

- UIMA-HPC:
  - Several applications perform different annotations
  - Analysis applications embedded in UIMA (Unstructured Information Management Architecture)
  - UNICORE workflows for annotation process on HPC-systems



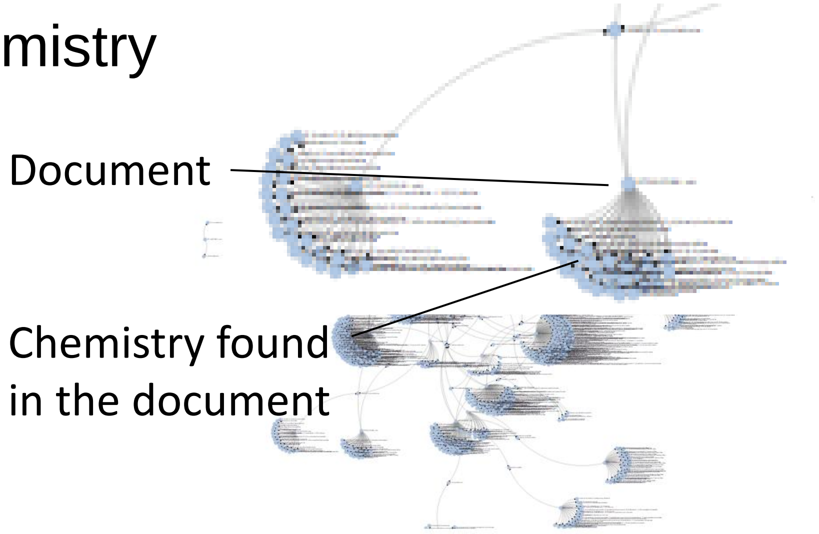
Key	Value
Execution Status	RUNNING
Name	UseCase1Auszug1For1datei_Test subm
Submitted at	2013-06-07 10:52:14
State	Ready
Type	WorkflowManagement
Current Time	2013-06-07 10:56:08
Termination Time	2013-06-17 10:52:14
URI	https://uima-hpc-unic.fz-juelich.de:91

  - Aim: shortest time to solution

    → Benefits as well as restrictions through UNICORE

The Use Case

- Analysis of chemical patents
  - Available as text or pdf files
- Communication over special data structure
- Applications for Natural Language Processing
- Applications to recognize chemistry
- Different outputs:
  - CSV file
  - RDF for Triple Store
  - BRAT



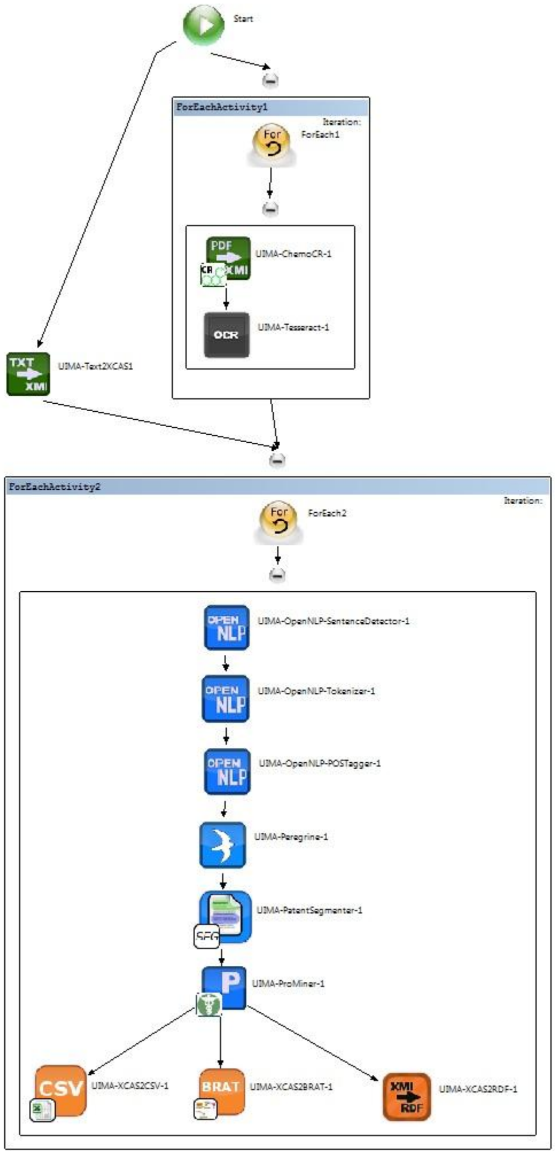
Document

Chemistry found in the document

The Use Case (cont.)

UNICORE workflow with twelve applications

- Preprocessing applications
- Second for-loop with chemical analysis applications



Observed Advantages and Restrictions

Tests performed:

- 60 txt and 60 pdf files, in total 3 GB

1. All input files in one tar-file, so only one job per application
   - Filetransfer with BFT: 5h and 50 minutes
   - Filetransfer with UFTP: 13 minutes

→ UNICORE provides UFTP, but abolish/reduce of file transfer time preferable

Observed Advantages and Restrictions (cont.)

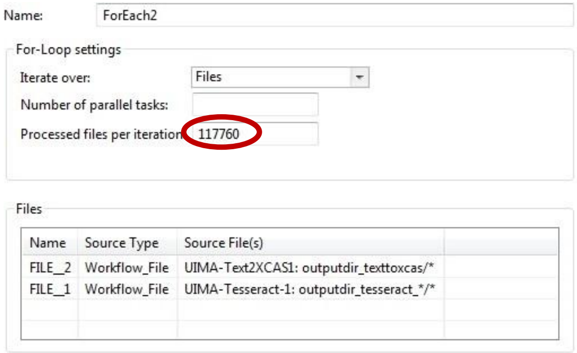
2. Input files are divided into different tar files
    - Parallelism reduces time to solution
    - Filetransfer with BFT: total: 6h and 23 minutes
    - Filetransfer with UFTP: total: 53 minutes

→ Problem: packing and unpacking of tar-files wastes 5-7 minutes

Observed Advantages and Restrictions (cont.)

3. Unpacked input data,
   for loop specified with datasize
   - Parallelism reduces time to solution
   - Filetransfer with BFT:
     total: 11h and 52 minutes
   - Filetransfer with UFTP:
     total: 13h and 53 minutes

→ Possibility to determine total datasize
   → Equally distribution of the input



Name: ForEach2

For-Loop settings

Iterate over: Files

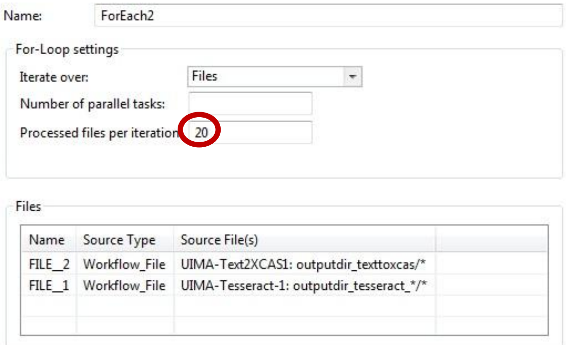
Number of parallel tasks:

Processed files per iteration: 117760

Files

Name	Source Type	Source File(s)
FILE_2	Workflow_File	UIMA-Text2XCAS1: outputdir_texttoxcas/*
FILE_1	Workflow_File	UIMA-Tesseract-1: outputdir_tesseract_*/*

For loop with a specified datasize as input for the jobs.



Name: ForEach2

For-Loop settings

Iterate over: Files

Number of parallel tasks:

Processed files per iteration: 20

Files

Name	Source Type	Source File(s)
FILE_2	Workflow_File	UIMA-Text2XCAS1: outputdir_texttoxcas/*
FILE_1	Workflow_File	UIMA-Tesseract-1: outputdir_tesseract_*/*

For loop with a specified file number as input for the jobs.

Conclusion

- Advantages:
  - UFTP or BFT for the transport
  - For loop with file number or datasize
- Restrictions:
  - File transfers waste time
- Necessary features:
  - Efficient transfer for large number of files (without tar)
  - Prevent unnecessary file transfers
  - Determine datasize in for loop

Acknowledgement

UIMA-HPC is funded by the German Ministry of Education and Research (BMBF) under grant id 01IH11012A-D.

Thanks to Bernd Schuller and Michael Rambadt for their support.

Tank you for your attention!

Do you have any questions?