
Gongjing Cao, Lei Dou, Quinn Hart, Bertram Ludaescher

UC Davis
Interactive Scientific Workflows

- Requirements for human interaction in scientific applications
  - dynamic branching based on scientists’ runtime decision
  - semi-automatic data curation

- Category of human interaction

<table>
<thead>
<tr>
<th></th>
<th>Synchronous</th>
<th>Asynchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>time cost</td>
<td>short time, instantly</td>
<td>unknown</td>
</tr>
<tr>
<td>people involved in interaction</td>
<td>workflow executor</td>
<td>people other than workflow executor</td>
</tr>
<tr>
<td>workflow blocking</td>
<td>yes</td>
<td>not necessarily</td>
</tr>
<tr>
<td>implementation approaches</td>
<td>graphical window, web page/browser</td>
<td>dedicated server, mail, polling, callback</td>
</tr>
</tbody>
</table>
Google Cloud Computing in Kepler
Actors in Kepler/G-Pack

- **Authorization**
  - 1st step to acquire access to Google services

- **Spreadsheet Operations**
  - Various manipulations on Google Spreadsheet, like copy, share, import, export, query, audit.

- **Data Analysis**
  - Various operations especially for data curation purpose, like duplicates identification and fuse, data boundary inspection.

- **Data Access**
  - Google visualization datasource actor allows SQL-like access to Google cloud data

- **Mail Service**
  - MailSender Actor supports sending email through SMTP with UserName/Password or OAuth token/secret.
OAuthAuthorizer Actor

1. Requests token with OAuth
2. Responds with unauthorized request token
3. Invokes Web browser with auth request URL
4. Requests authorization for request token
5. Redirects user to Access Consent page
6. Logs in & grants denies access
7. Redirects with verification token
8. Feeds verification token to installed app
9. Requests exchange for OAuth token&secret
10. Responds with OAuth token&secret
11. Requests data with OAuth token&secret
12. Responds with requested data
Spreadsheet Operation Actors

<table>
<thead>
<tr>
<th>Actors</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importer</td>
<td>import data to a spreadsheet</td>
</tr>
<tr>
<td>Exporter</td>
<td>export data from a spreadsheet</td>
</tr>
<tr>
<td>Copy</td>
<td>copy a spreadsheet from a template</td>
</tr>
<tr>
<td>Share</td>
<td>share the spreadsheet with another user</td>
</tr>
<tr>
<td>Query</td>
<td>query data from the spreadsheet</td>
</tr>
<tr>
<td>Auditor PollingQuery</td>
<td>allow human interaction during the execution of the workflow</td>
</tr>
</tbody>
</table>
Data Access
VisualizationDataSource Actor

- Get the response from a servlet which is implemented with Google visualization datasource.

  e.g. servlet:
  
  http://comet.cs.ucdavis.edu:8080/CimisVis/sql
  
  Table: daily
  
  query: select * where d_date > date'2005-01-09' and d_date < date'2005-01-20'
Data Access
VisualizationDataSource Actor

- Actor parses JSON string to Kepler token
- array of RecordToken
MailSender Actor

- Gmail and other mail server supporting SMTP protocol
  - It supports sending email through SMTP with UserName/Password or OAuth token/secret.
Data Analysis Actors (COMAD actors)

<table>
<thead>
<tr>
<th>Actors</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>Do clustering on a list of RecordToken by applying specified function for specified field.</td>
</tr>
<tr>
<td>DataFuser</td>
<td>Fuse a list of RecordToken with specified function</td>
</tr>
<tr>
<td>ConditionTester</td>
<td>Test whether specified condition is satisfied or not.</td>
</tr>
</tbody>
</table>
Evapotranspiration Workflow

Spreadsheet is used as data storage and calculation tool during this demo. ETOs are calculated from CIMIS weather station data according to different models.
Biofuel Refinery Workflow

This workflow allows users to copy a spreadsheet from a template, then use the existing data or edit the spreadsheet to help them make decisions in building a biofuel refinery.
SpecimenRecordMerge Workflow

This workflow demonstrates how specimen data with duplicates are fused and curated semi-automatically.

1. Firstly the specimen records are imported from a file in csv format.
2. Secondly the duplicated sets of specimen records are identified through specific clustering function and for each duplicate set a fused record is generated.
3. Thirdly the original specimen records and the fused record are imported into a google spreadsheet for human curation. Meanwhile the corresponding curators are informed of such curation request.
4. Finally, the human curation result is collected once it’s finished by the curator.
AdvancedSpecimenRecordMerge Workflow

This workflow makes the following improvement of SpecimenRecordMerge workflow. It’s demonstrated how easy it is to reconfigure COMAD workflow to adapt to new functionalities.

1. Insert SpecimenRecCleaner actor to clean source data by removing "bad" specimen records with unclear collector of "et al."
2. Reconfigure Clustering actor to cluster specimen records against collector field with fuzzy match method. Therefore the specimen records collected at the same time and the same location but by different collector of "E. L. Morris" and "E. Morris" could be identified as duplicates.
Thanks & Question