
Gongjing Cao, Lei Dou, Quinn Hart, Bertram Ludaescher

UC Davis
Google Cloud Computing in Kepler
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</td>
<td>dedicated server</td>
</tr>
<tr>
<td></td>
<td>web page/browser</td>
<td>mail, polling, callback</td>
</tr>
</tbody>
</table>
Actors in Kepler/G-Pack

- **Authorization**
  - 1\textsuperscript{st} 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.

---

**Refinery-Copy2**

- **From:** kepler.ucd@gmail.com
- **To:** viewonly2010@gmail.com
- **Date:** Fri, Feb 11, 2011 at 9:06 AM
- **Subject:** refinery-copy2
- **Mailed by:** doclist.bounces.google.com

I've shared a document with you:

refinery-copy2

https://spreadsheets.google.com/ccc?key=0AgT1sEQRg

It's not an attachment -- it's stored online at Google Docs.
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.
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
Kepler is an open source, scientific workflow platform. It provides domain scientists with an easy-to-use yet powerful system across a broad range of scientific and engineering disciplines.

New feature: leverage Google cloud
Google cloud provides a powerful platform for online data storage, sharing, edit, computing and visualization, and relieves people from repetitive work and tedious maintenance.

Integration of Google cloud services in Kepler

- “Outsourcing” computing Kepler task to Google cloud
  - Kepler actors: access Google services through API
  - Orchestrate various tasks accessing Google cloud services together with “none-google” tasks as an complete data analysis pipeline.
  - Kepler workflow engine

The Google VizDatasource API lets you access multiple sources of structured data. It could be static data, or from a database, or cvs file etc. Google Visualization API enables you to expose your own data, stored on any data-store that is connected to the web, as a Visualization compliant datasource. Thus you can create reports and dashboards as well as analyze and display your data through the wealth of available visualization applications. The Google Visualization API also provides a platform that can be used to create, share and reuse visualizations written by the developer community at large.

To integrate google cloud services in kepler, for one side, we “outsourcing” the Kepler tasks to Google resources by develop many google oriented kepler actors to access google cloud services through corresponding API. For another side, based on these actors, the Kepler workflow engine would effectively orchestra these google tasks together with none-google tasks as an intact data analysis pipeline.
Sometimes scientific workflows require human interaction. It could be dynamic branching based on scientists’ runtime decision. A simple case is a workflow asking for users' grant or denial to certain Google services. Semi-automatic data curation.
This actor makes a request to a servlet which is implemented with Google VizDataSource API.

The Google VizDatasource API is used to convert multiple sources of data into GoogleVizDatasource. Sources could be static data, from a database, or csv file etc. So it enables you to expose your own data as a Visualization datasource online. Thus, besides Kepler can access it through this actor, you can also create reports and dashboards as well as analyze and display your data through the wealth of available visualization applications Google provides. Google also provides a platform that can be used to create, share and reuse visualizations applications.
The response we saw from the servlet is in a format of JSON String. Then VizDatasource acts by passing it to Kepler tokens.
We notice there is already a mail actor in Ptolemy, which can send email through SMPT by asking username/passwd. Sending email through OAuth is the new feature this new actor has.
Kuration package is actually developed under comad director. Comad make it easy to handle hierachical data. Comad is highly adaptable to new functions and data. Moreover comad provides great support for provenance tracking. We could see more in the incoming video.

The normal kepler actor we’ve introduced previously could be encapsulated as a comad actor in a way similar to composite actor. Besides, we also develop the following comad actors for data curation purpose, including actors for data clustering, duplicates fuse and test of specified condition which is useful for data boundary inspection.

Functionally, the conditionTester equals to expression actor. It’s an application of expression actor in comad director.
Evapotranspiration (ET) is a term used to describe the sum of evaporation and plant transpiration from the Earth's surface to atmosphere.

In this workflow it is calculated from data that were collected from weather stations of California irrigation management information system.
This workflow is used to demonstrate how the copy share and auditor actors are used.

Originally there is a template which contain computing models and graphs to show price, cost, profit information for biofuel refineries.

When it comes that many owners want to use this model to assist them make decisions, and prices could change with time.

This workflow allows this template to be copied and shared to a user, then the new user may use the existing information in the spreadsheet and edit them, like gas price, feedstock price, feedstock amount to find best cost/profit point to help them make decision for their refinery.
Due to historical reason, incomplete data standard and data complexity, many literally different specimen records are actually the same. How to find out these duplicates and fuse them into one record is a big issue in biodiversity research area. Usually such curation can't be finished fully automatically. Human participation is necessary. So we develop this SpecimenRecordMerge workflow to demonstrates how specimen data with duplicates are fused and curated semi-automatically.

This workflow demonstrates how specimen data with duplicates are fused and curated semi-automatically. The duplicate set are identified and fused automatically at the first phase and then it will be presented to curators. Finally the curation decision of curators will be collected once the curation is submitted.
In scientific area, change is normal, especially for the area of data curation. It’s common to change the curation process to correct the problems found from the previous curation process. Workflow technology provides great support to meet the requirement of such dynamic features.

This workflow makes the following function improvement based on SpecimenRecordMerge workflow. It’s demonstrated it’s very easy to reconfigure the workflow to adapt to new functions.

1. Add SpecimenRecCleaner actor to clean the source data by removing "bad" record whose collector is "et al."
2. Reconfigure the Clustering actor by using a clustering function implementing fuzzy match method to cluster data against collector (name) field. Therefore the records whose collector are "E. L. Morris" and "E. Morris" could be put into one duplicate set.