CWL AIRFLOW PIPELINE MANAGER AS A BACKEND FOR BIOWARDROBE EPIGENOMICS DATA ANALYSIS PLATFORM

Andrey Kartashov
Michael Kotliar
Dr. Barski lab.
BIOWARDROBE
AN INTEGRATED PLATFORM FOR ANALYSIS OF EPIGENOMICS AND TRANSCRIPTOMICS DATA

So easy, even a full professor can do data analysis!

Kartashov & Barski Genome Biology, 2015
BIOWARDROBE PROBLEMS

• Inefficient server utilization
• Cannot restart a failed step - just the whole pipeline
• Adding new pipelines is unreasonably difficult
• Updates to host system often brake pipeline dependencies – poor reproducibility
• Poor versioning
• No standard way to record metadata
COMMON WORKFLOW LANGUAGE

• Common declarative format made workflows portable between platforms
• The same tool image across different environment with containers (e.g. Docker) makes analysis reproducible
• Graphical workflow assembly (e.g. with Rabix Composer) from common components – convenient for biologists
• Ability to record metadata and version control
• Provides the opportunity to add platform specific futures ("sbg:x", "sd:upstream")
PIPELINE MANAGER REQUIREMENTS

• Strong community
• Easy installation and support
• Graphical user interface
• Python is a plus
• Scalable: Laptop/Server/Cluster/Cloud
• Famous 😊
GOOGLE CLOUD COMPOSER

CLOUD COMPOSER BETA

A fully managed workflow orchestration service built on Apache Airflow

Create a Hybrid & Multi-Cloud Environment

Cloud Composer gives you the ability to connect your pipeline through a single orchestration tool whether your workflow lives on-premises, in multiple clouds, or fully within GCP. The ability to author, schedule, and monitor your workflows in a unified manner means you can break down the silos in your environment and focus less on infrastructure.

Open Source at its Core

Cloud Composer is built on Apache Airflow, the popular open source orchestration tool. This open source project, which Google is contributing back into, provides freedom from lock in for customers as well as integration with a broad number of platforms, which will only expand as the Airflow community grows.
CWL-AIRFLOW STRUCTURAL SCHEME

- jobs
- workflows
- output folder

CWL-Airflow

job
{ "file": "/fastq", "param": 5 }

CWLStepOperator

CWLJobDispatcher

CWLStepOperator

CWLStepOperator

CWLStepOperator

CWLJobGatherer

Airflow scheduler
<table>
<thead>
<tr>
<th>DAG</th>
<th>Schedule</th>
<th>Owner</th>
<th>Recent Tasks</th>
<th>Last Run</th>
<th>DAG Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>biowardrobe-force-run</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td>2018-02-19 12:10</td>
<td>1, 3</td>
</tr>
<tr>
<td>biowardrobe_download</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td>2018-02-21 21:20</td>
<td>11</td>
</tr>
<tr>
<td>biowardrobe_download_trigger</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>biowardrobe_plugins</td>
<td>On</td>
<td>Airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chipseq-pe</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chipseq-se</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>plot-dna</td>
<td>On</td>
<td>Airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>plot-rna</td>
<td>On</td>
<td>Airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-pe</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-pe-dutp</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-pe-dutp-mitochondrial</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-se</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-se-dutp</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maseq-se-dutp-mitochondrial</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>satscript</td>
<td>On</td>
<td>Airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>super-enhancer</td>
<td>On</td>
<td>Airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trim-chipseq-pe</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trim-chipseq-se</td>
<td>On</td>
<td>airflow</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xenbase-maseq-pe</td>
<td>On</td>
<td>kot4or</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xenbase-maseq-se</td>
<td>On</td>
<td>kot4or</td>
<td>▼</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BIOWARDROBE SOLVED

✓ Efficient server utilization
✓ Restart/stop/pause at any stage of a pipeline
✓ CWL workflows can be easily added
✓ Updates to host system do not break workflows
VISIT US FOR A DEMO SESSION