

Deprecated Workflow Generator

This page is **no longer maintained**. We have released a new *Workflow Generator* as part of the [WorkflowHub Project](#):

Workflow Generator

To facilitate evaluation of workflow algorithms and systems on a range of workflow sizes, we have developed a set of synthetic workflow generators. These generators use the information gathered from actual executions of scientific workflows on the Grid as well as our understanding of the processes behind these workflows to generate realistic, synthetic workflows resembling those used by real world scientific applications.

The code used to generate all of the synthetic workflows below, and many others, is available from the GitHub repository. The java workflow generator sometimes generates negative task runtimes, so watch out for that.

Simulator

WorkflowSim can be used to simulate the workflows generated by the Workflow Generator.

Traces

Traces and execution logs from real workflows are available here: [here](#), [here](#), and [here](#). Data sets like these were used to parameterize the Workflow Generator.

Synthetic Workflows

Pegasus Workflows

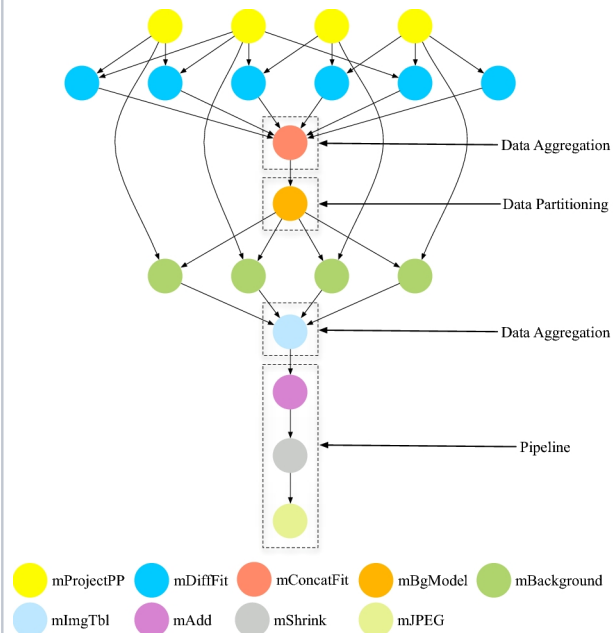
These workflows come from a [paper by Bharathi, et al. \[2\]](#). There is [another paper with more information about the workflows by Juve, et al. \[3\]](#).

[A large collection of DAXes similar to the ones listed below is available here](#). Note that it is about 375 MB.

Workflow Type	Example	DAX
---------------	---------	-----

Montage

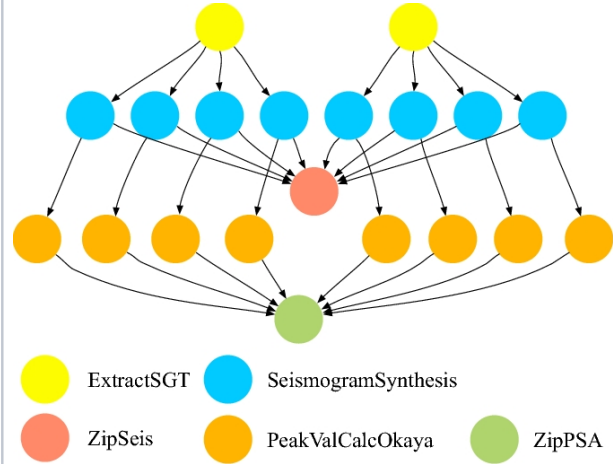
The Montage application created by NASA/IPAC stitches together multiple input images to create custom mosaics of the sky.



25 Node DAX
50 Node DAX
100 Node DAX
1000 Node DAX

CyberShake

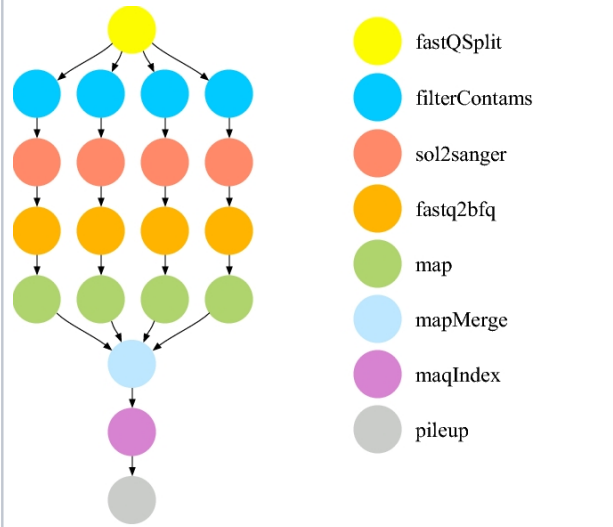
The CyberShake workflow is used by the Southern California Earthquake Center to characterize earthquake hazards in a region.



30 Node DAX
50 Node DAX
100 Node DAX
1000 Node DAX

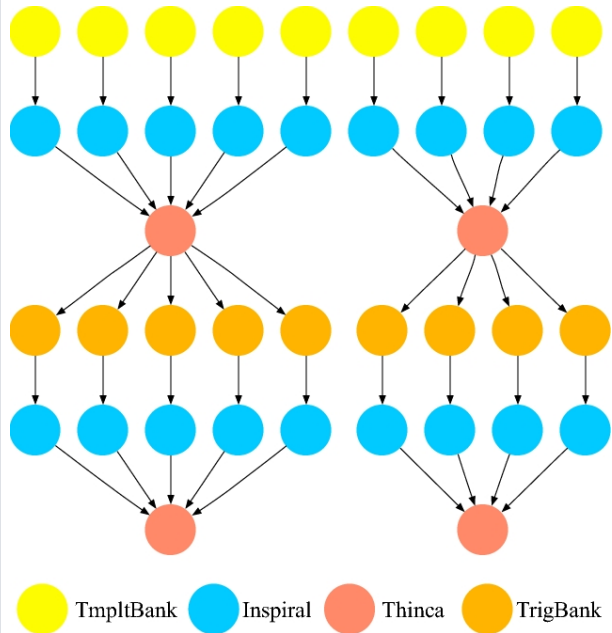
Epigenomics

The epigenomics workflow created by the USC Epigenome Center and the Pegasus Team is used to automate various operations in genome sequence processing.



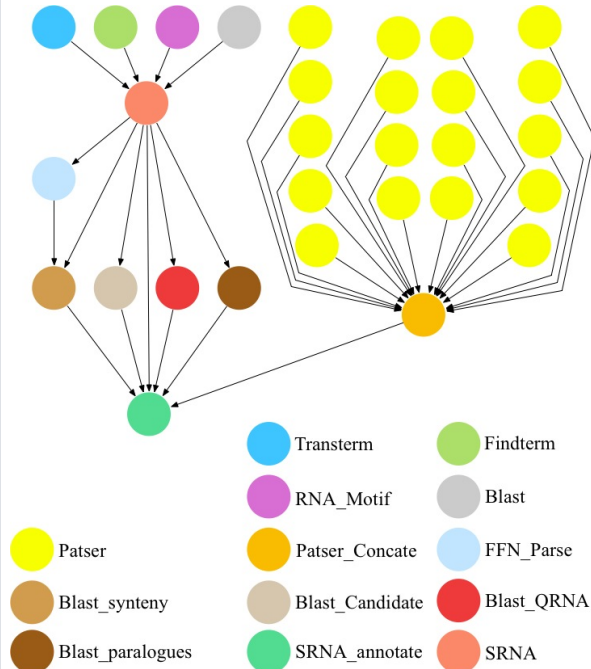
24 Node DAX
46 Node DAX
100 Node DAX
997 Node DAX

LIGO Inspiral Analysis
 LIGO's Inspiral Analysis workflow is used to generate and analyze gravitational waveforms from data collected during the coalescing of compact binary systems.



30 Node DAX
 50 Node DAX
 100 Node DAX
 1000 Node DAX

SIPHT
 The SIPHT workflow, from the bioinformatics project at Harvard, is used to automate the search for untranslated RNAs (sRNAs) for bacterial replicons in the NCBI database.



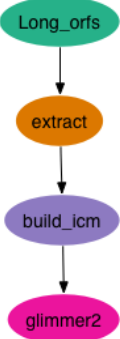
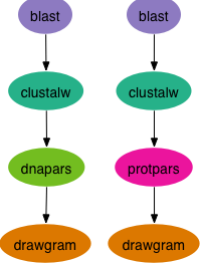
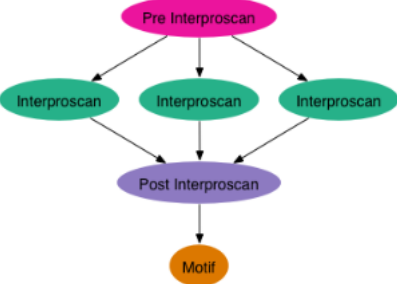

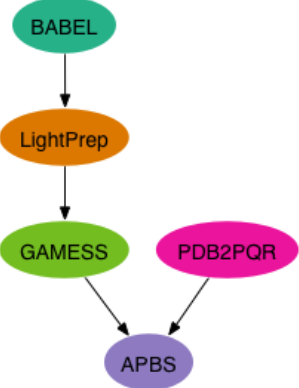
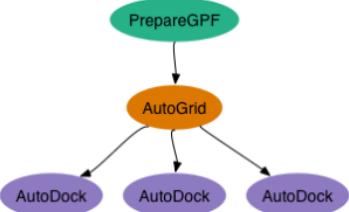
30 Node DAX
 60 Node DAX
 100 Node DAX
 1000 Node DAX

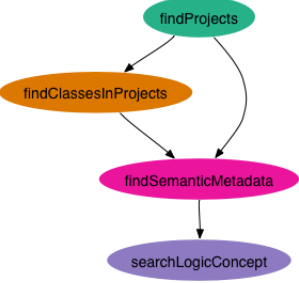
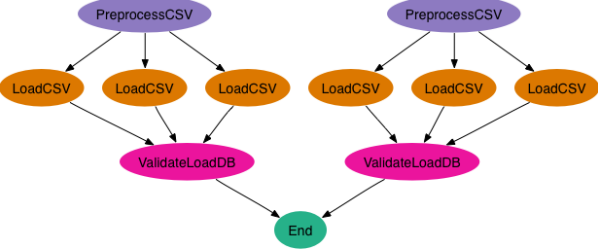
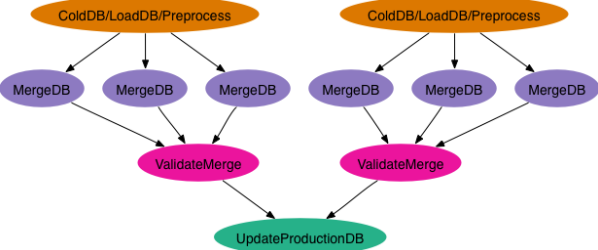
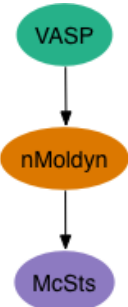
Ramakrishnan and Gannon Workflows

These workflows come from a [report](#) by Ramakrishnan and Gannon [3].

Workflow Type	Figure in Report	Example	DAX
---------------	------------------	---------	-----

LEAD Mesoscale Meteorology	Figure 1	<pre> graph TD TP[TerrainPreProcessor] --> LB[LateralBoundaryInterpolator] TP --> 3D[3DInterpolator] Wf[WrfStatic] --> LB LB --> ARPS[ARPS2WRF] 3D --> ARPS ARPS --> WRF[WRF] </pre>	leadmm.xml
LEAD ARPS Data Analysis System	Figure 2	<pre> graph TD TP[TerrainPreProcessor] --> LB[LateralBoundaryInterpolator] TP --> ADAS[ADASInterpolator] Wf[WrfStatic] --> LB LB --> ARPS[ARPS2WRF] ADAS --> ARPS ARPS --> WRF[WRF] </pre>	leadadas.xml
LEAD Data Mining Workflow	Figure 3	<pre> graph TD SD[StormDetection] --> RA[RemoveAttributes] RA --> SC[SpatialClustering] </pre>	leaddm.xml
Storm Surge SCOOP Workflow	Figure 4	<pre> graph TD A1[Adcirc] --> PP[PostProcessing] A2[Adcirc] --> PP A3[Adcirc] --> PP </pre>	scoop_small.xml scoop_medium.xml scoop_large.xml
Floodplain Mapping	Figure 5	<pre> graph TD A1[Adcirc] --> SWANON[SWAN Outer North] A1 --> SWANSO[SWAN Outer South] A1 --> SWANNI[SWAN Inner North] A1 --> SWANIS[SWAN Inner South] W[WaveWatchIII] --> SWANON W --> SWANSO W --> SWANNI W --> SWANIS SWANNI --> A2[Adcirc] SWANIS --> A2 </pre>	floodplain.xml

Glimmer	Figure 6	 <pre> graph TD A([Long_orfs]) --> B([extract]) B --> C([build_icm]) C --> D([glimmer2]) </pre>	glimmer.xml
Gene2Life	Figure 7	 <pre> graph TD A1([blast]) --> B1([clustalw]) B1 --> C1([dnapars]) C1 --> D1([drawgram]) A2([blast]) --> B2([clustalw]) B2 --> C2([protpars]) C2 --> D2([drawgram]) </pre>	gene2life.xml
Motif Network	Figure 8	 <pre> graph TD A([Pre Interproscan]) --> B1([Interproscan]) A --> B2([Interproscan]) A --> B3([Interproscan]) B1 --> C([Post Interproscan]) B2 --> C B3 --> C C --> D([Motif]) </pre>	motif_small.xml motif_medium.xml motif_large.xml
MEME-MAST	Figure 9	 <pre> graph TD A([MEME]) --> B([MAST]) </pre>	mememast.xml
Molecular Sciences	Figure 10	 <pre> graph TD A([BABEL]) --> B([LightPrep]) B --> C([GAMESS]) B --> D([PDB2PQR]) C --> E([APBS]) D --> E </pre>	molsci.xml
Avian Flu	Figure 11	 <pre> graph TD A([PrepareGPF]) --> B([AutoGrid]) B --> C1([AutoDock]) B --> C2([AutoDock]) B --> C3([AutoDock]) </pre>	avianflu_small.xml avianflu_medium.xml avianflu_large.xml

caDSR	Figure 12	 <pre> graph TD findProjects([findProjects]) --> findClassesInProjects([findClassesInProjects]) findProjects --> findSemanticMetadata([findSemanticMetadata]) findClassesInProjects --> findSemanticMetadata findSemanticMetadata --> searchLogicConcept([searchLogicConcept]) </pre>	cadsr.xml
Pan-STARRS Load	Figure 13	 <pre> graph TD PreprocessCSV1([PreprocessCSV]) --> LoadCSV1([LoadCSV]) PreprocessCSV1 --> LoadCSV2([LoadCSV]) PreprocessCSV1 --> LoadCSV3([LoadCSV]) LoadCSV1 --> ValidateLoadDB1([ValidateLoadDB]) LoadCSV2 --> ValidateLoadDB1 LoadCSV3 --> ValidateLoadDB1 PreprocessCSV2([PreprocessCSV]) --> LoadCSV4([LoadCSV]) PreprocessCSV2 --> LoadCSV5([LoadCSV]) PreprocessCSV2 --> LoadCSV6([LoadCSV]) LoadCSV4 --> ValidateLoadDB2([ValidateLoadDB]) LoadCSV5 --> ValidateLoadDB2 LoadCSV6 --> ValidateLoadDB2 ValidateLoadDB1 --> End([End]) ValidateLoadDB2 --> End </pre>	psload_small.xml psload_medium.xml psload_large.xml
Pan-STARRS Merge	Figure 14	 <pre> graph TD ColdDB1([ColdDB/LoadDB/Preprocess]) --> MergeDB1([MergeDB]) ColdDB1 --> MergeDB2([MergeDB]) ColdDB1 --> MergeDB3([MergeDB]) MergeDB1 --> ValidateMerge1([ValidateMerge]) MergeDB2 --> ValidateMerge1 MergeDB3 --> ValidateMerge1 ColdDB2([ColdDB/LoadDB/Preprocess]) --> MergeDB4([MergeDB]) ColdDB2 --> MergeDB5([MergeDB]) ColdDB2 --> MergeDB6([MergeDB]) MergeDB4 --> ValidateMerge2([ValidateMerge]) MergeDB5 --> ValidateMerge2 MergeDB6 --> ValidateMerge2 ValidateMerge1 --> UpdateProductionDB([UpdateProductionDB]) ValidateMerge2 --> UpdateProductionDB </pre>	psmerge_small.xml psmerge_medium.xml psmerge_large.xml
McStas	Figure 15	 <pre> graph TD VASP([VASP]) --> nMoldyn([nMoldyn]) nMoldyn --> McStas([McStas]) </pre>	mcstas.xml

[1] R. F. da Silva, W. Chen, G. Juve, K. Vahi, E. Deelman. Community Resources for Enabling Research in Distributed Scientific Workflows. 10th IEEE International Conference on e-Science (eScience 2014)

[2] S. Bharathi, A. Chervenak, E. Deelman, G. Mehta, M.-H. Su, and K. Vahi, "Characterization of Scientific Workflows", 3rd Workshop on Workflows in Support of Large Scale Science (WORKS 08), 2008.

[3] Gideon Juve, Ann Chervenak, Ewa Deelman, Shishir Bharathi, Gaurang Mehta, and Karan Vahi, "Characterizing and Profiling Scientific Workflows", *Future Generation Computer Systems*, 29:3, pp. 682–692, March 2013.

[4] L. Ramakrishnan and D. Gannon, "A Survey of Distributed Workflow Characteristics and Resource Requirements", Indiana University Technical Report TR671, 2008.