Ziggy, a Portable, Scalable Infrastructure for Science Data Processing Pipelines

Peter Tenenbaum[1], Bill Wohler[1], Jon Jenkins[2], Yohei Shinozuka[3], Jennifer Dungan[2], Ian Brosnan[2], Chris Henze[2], Mark Rose[4], and Andrew Michaelis[2]

[1]SETI Institute
[2]NASA Ames Research Center
[3]Bay Area Environmental Research Institute Sonoma
[4]PSGS / NASA Ames Research Ctr

November 24, 2022

Abstract

We describe Ziggy, an infrastructure for pipelines that process large volumes of science data. Ziggy is based on the pipeline infrastructure software that was developed to process flight data for the Kepler and TESS exoplanet missions. In this latter capacity, multiple terabytes of data are processed every month. Ziggy provides execution control, logging, exception management, marshaling, and persistence, and data accountability record management for user-defined sequences of processing steps. Users define a pipeline via a set of XML files that specify the order in which processing algorithms are applied (including optional branching, in which one step is followed by multiple algorithms that run simultaneously), inputs, outputs, and any instrument models or control parameters that are required for each step. Ziggy supports heterogeneous pipelines: each processing algorithm can be in any supported language, and each step can run locally on a server or remotely on a supercomputer or cloud computing facility. Ziggy is sufficiently lightweight to run on a laptop and sufficiently robust to run on a supercomputer; builds on Mac OS X and Linux are supported. Ziggy is currently in use as the pipeline infrastructure tool for reprocessing the full data volume of the EO-1/Hyperion mission data and is a candidate for use in the upcoming Surface Biology and Geology (SBG) mission of the Earth System Observatory (ESO). Ziggy contains no proprietary or sensitive/controlled software or algorithms, and approval for its release as a NASA Open Source Software Project is underway.

Ziggy, a Portable, Scalable Infrastructure for Science Data Processing Pipelines

Peter Tenenbaum[1,2], Bill Wohler[1,2], Jon Jenkins[2], Yohei Shinozuka[3,2], Jennifer L. Dungan[2],
Ian G. Brosnan[2], Chris Henze[2], Mark Rose[2], Andrew Michaelis[2]

[1]SETI Institute, Mountain View, CA USA [2]NASA Ames Research Center, Moffett Field, CA USA [3]Bay Area Environmental Research Institute, Moffett Field, CA USA

Abstract:

We describe Ziggy, an infrastructure for pipelines that process large volumes of science data. Ziggy is based on the pipeline infrastructure software that was developed to process flight data for the *Kepler* and TESS exoplanet missions. In this latter capacity, multiple terabytes of data are processed every month. Ziggy provides execution control, logging, exception management, marshaling, and persistence, and data accountability record management for user-defined sequences of processing steps.

Users define a pipeline via a set of XML files that specify the order in which processing algorithms are applied (including optional branching, in which one step is followed by multiple algorithms that run simultaneously), inputs, outputs, and any instrument models or control parameters that are required for each step. Ziggy supports heterogeneous pipelines: each processing algorithm can be in any supported language, and each step can run locally on a server or remotely on a supercomputer or cloud computing facility.

Ziggy is sufficiently lightweight to run on a laptop and sufficiently robust to run on a supercomputer; builds on Mac OS X and Linux are supported. Ziggy is currently in use as the pipeline infrastructure tool for reprocessing the full data volume of the EO-1/Hyperion mission data and is a candidate for use in the upcoming Surface Biology and Geology (SBG) mission of the Earth System Observatory (ESO). Ziggy contains no proprietary or sensitive/controlled software or algorithms, and approval for its release as a NASA Open Source Software Project is underway.

Motivation:

CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... *MIGHT* NOT BE.
I GUESS THAT'S SOMETH-
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

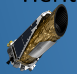
xkcd.com. Used under terms of Creative Commons License 2.5.

Science data pipelines need to do a lot more than science:

- Logging
- Execution flow
- Execution monitoring
- And much more!
- Data accountability
- Configuration management
- Error handling

Ziggy handles all the not-science and lets the scientists get on with the science!

Heritage:



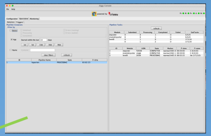
Used for science processing of the *Kepler* mission [1], where scientists got on with discovering 2879 confirmed exoplanets and 2528 exoplanet candidates!

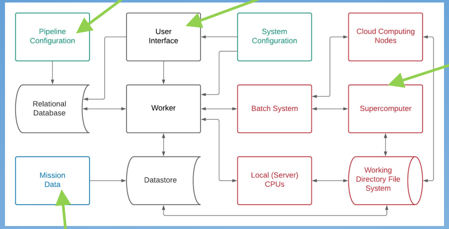
[1] Todd C. Klaus *et. Al.*, "The *Kepler* Science Operations Center pipeline framework," Proceedings of the SPIE 7740, 774017 (2010).

Used for science processing of the TESS mission, multi-TB/month data rate and discovery of 172 confirmed exoplanets and 3125 exoplanet candidates!

Under the Hood:

The mission defines the pipeline: algorithms, execution order, instrument models, etc.

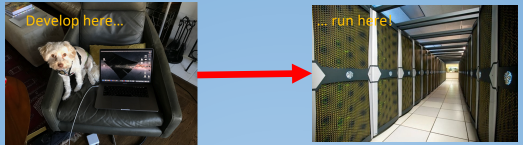




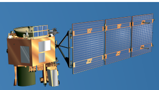
Ziggy supports all scales from local processing on a server to running on petascale systems. Supports hybrid (cloud / HPC) solutions!

Data, instrument models, etc., can use any desired format. Ziggy supports "keep-up" processing (just process new data) and reprocessing (do everything).

Ziggy is lightweight:

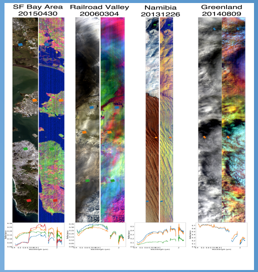


EO-1 Hyperion Reprocessing:



Ziggy is the pipeline controller for reprocessing of the 55 TB archive of data from the EO-1 Hyperion instrument to Level 2. Two benefits:

- First-ever uniform reprocessing of the full Hyperion dataset, which will be made available to the community.
- Provides experience and testbed for processing of future Earth System Observatory (ESO) datasets.



First-pass L1R processing of all scenes complete (Python translation of original IDL)

First-pass L2 processing using ISOFIT started

Open-source Science Initiative: Ziggy supports hybrid HPC and cloud solutions, and will soon be available open-source on NASA's GitHub site!