SAND2019-8623PE

Europy: Jupyter Integration with DataSEA

Doug McGeehan, Missouri University of Science and Technology, Ph.D. Computer Science (Graduation: December 2019)
Adam Bowers, University of Missouri-Columbia, Ph.D. Computer Science (Graduation: December 2019)
Kelsie Box, University of New Mexico, B.S. Computer Science (Graduation: May 2020)
Alan Shen, University of New Mexico, B.S. Computer Science (Graduation: May 2021)
Levi Davis, UW Seattle, B.S. Bioengineering & ACMS (Graduation: May 2020)
Org 2664, Manager: Brian Post
Mentors: William DeRaad, Stephen Jackson, Michael Muggler, Malachi Tolman, Vincent VanGemert

Abstract

For DataSEA users that want to analyze system test data in a Jupyter Notebook, Europy enables users to launch a Jupyter notebook from a DataSEA record and have the time-series test data and metadata associated with the record immediately available to them. Europy enables engineers to quickly look at and manipulate data from a web browser without the need for Excel files or a complicated local Jupyter setup.

Introduction

- Current procedures for analyzing system test data is cumbersome
  - Searching and downloading *dataframes* (time-series test data)
  - Performing statistical analysis of dataframes
  - End product: comprehensive system health report
  - Provenance record: who did what to the data?
- DataSEA and SEDS web applications target search and download
- Statistical analysis is not well addressed
  - Analysis is currently performed on external apps (Matlab, SAS)
  - Setup takes a long time to access data in these systems
  - Committing changes and recording provenance requires tedious data entry
- Jupyter Notebooks: a web-based tool for statisticians and analysts
  - Provides a platform for in-browser data analysis
  - Statistical analysis through Python (e.g. scipy, scikit-learn)
  - Export analysis into multimedia formats (e.g. HTML, PDF)
- Europy: a bi-directional coupling between Jupyter and DataSEA
  - Easily download dataframes into Python pandas dataframes
  - Push data changes and provenance entries from Jupyter back to DataSEA
  - Utilize tools to minimize developer resources:
    - Swagger: creates a Python library that wraps DataSEA's API
    - Docker: future-proof system for easier deployment and scalability

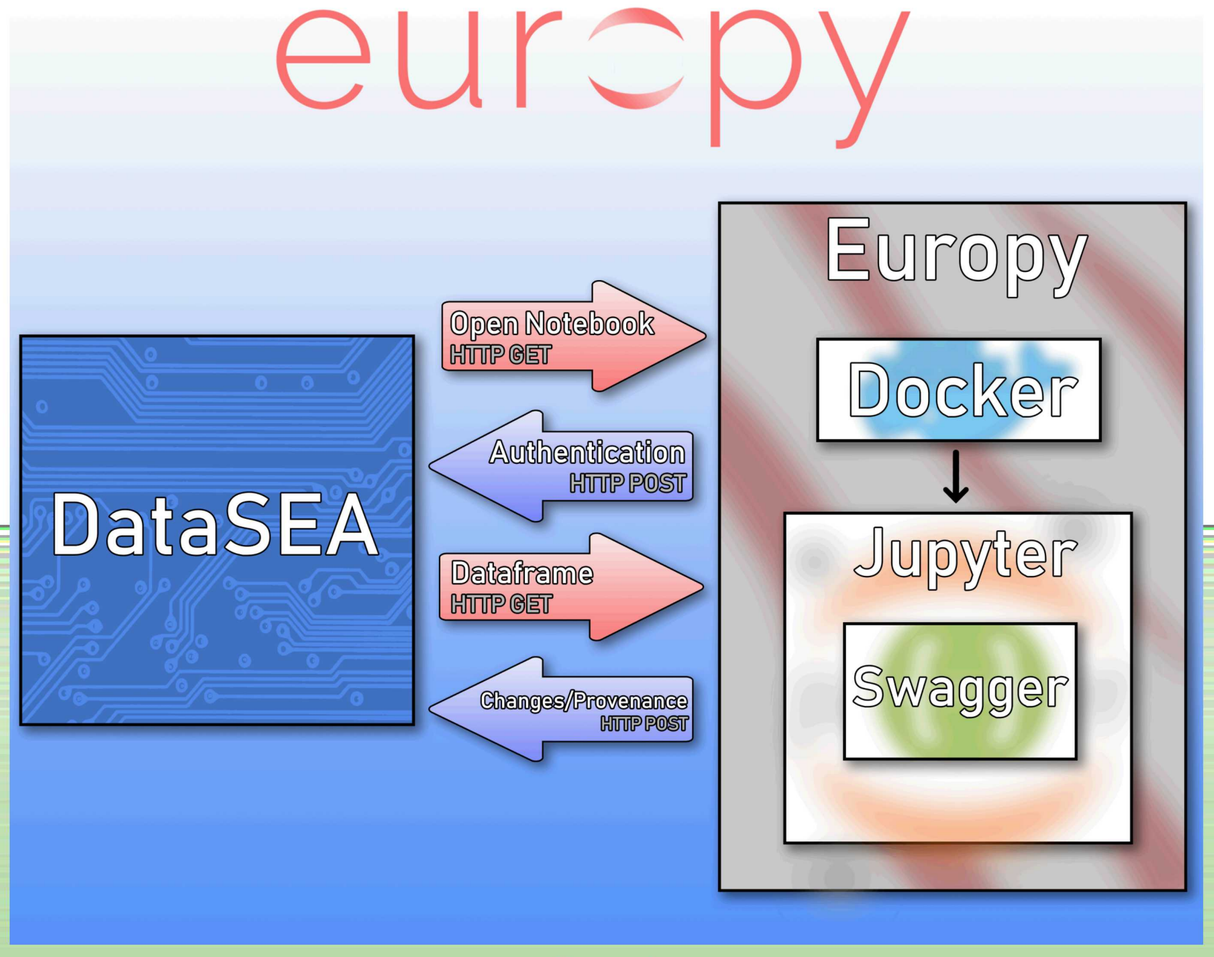

Figure 1
Process visualization of navigating from DataSEA to Europy to visualize a record's data within a Jupyter notebook.

Figure 2
Technologies employed by Europy: Jupyter Notebook server, Python 3.7, Swagger 3, Docker

System Process

1. Button is pressed in DataSEA to open a Jupyter notebook with a specific search query, dynamic record, or dataframe
2. Docker spins up a Jupyter instance for the requested data
3. Credentials are given to the Jupyter notebook by the user and sent to DataSEA for authentication
4. Once authenticated, DataSEA returns the requested data
   - Europy uses Swagger to talk to the DataSEA API.
5. User conducts their analysis using Python statistical analysis tools within the notebook
   e.g. pandas, scipy, scikit-learn, matplotlib
6. Europy pushes changes and updates provenance to DataSEA

Implemented Use Cases

Single dataframe access:
Load time-series data and let users process with Python tools
Single record/multi dataframe access:
Load a single dynamic record, all associated metadata, and dataframes for cross-dataframe analysis
Search results access:
Load all dynamic records matching a MongoDB search query and permit user to perform cross-record analysis
Saving / committing data changes:
Push back analysis notes and metadata changes to DataSEA as a new provenance revision on the given record(s)

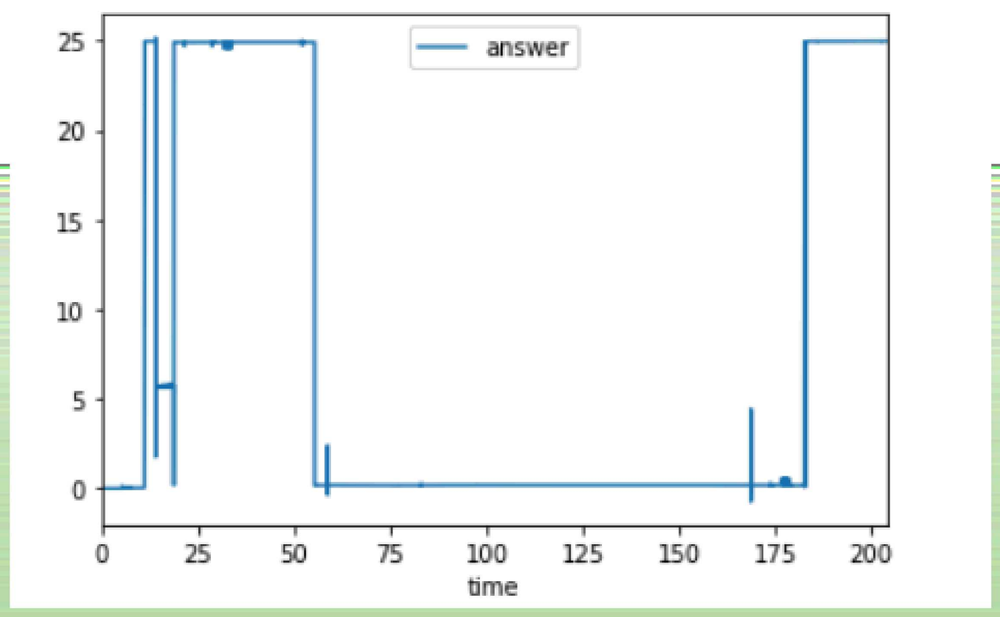

Figure 3
Example Jupyter notebook visualization of a dataframe using pandas and matplotlib

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____