ESCAPE WP5

Technical Work Review

Stelios Voutsinas
University of Edinburgh

REPRODUCIBLE WORKFLOWS

- JupyterHub Service Currently Running
  - Additional JupyterHub Service in Development on IRIS Cloud.
    - Jupyter Notebooks spawning Containers (Kubernetes)
    - DockerSpawner, SwarmSpawner
- Zeppelin Prototype in Development as well
- Cloud Hosted solutions? (GoogleCollab, mybinder)

REPRODUCIBLE WORKFLOWS

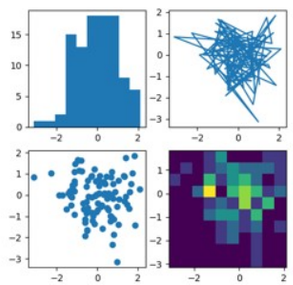
- Jupyter Notebooks
  - Discovery
    - VO Registry (pyvo)
  - Data Access
    - Astropy (Astroquery), Pyvo
  - Visualization
    - Plots using Matplotlib, Plotly, Bokeh
    - Interactive Hierarchical Sky view using AladinLite

Technology Summary

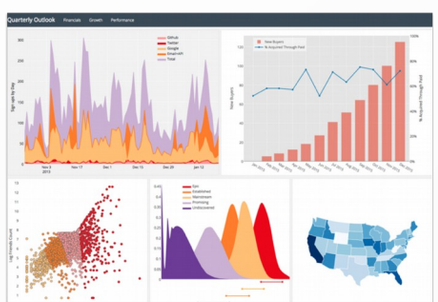
- PyVO
  - PyVO provides APIs for lots of VO protocols. (TAP, Registry, SIA)
  - It’s glue between astropy and python in general and the astronomical data services in the VO.
- Astropy
  - A common core package for Astronomy in Python developed as a community project.
  - https://github.com/astropy/astropy
- Aladin
  - Java-based Tool for interactive sky visualization
  - AladinLite: Javascript widget version of Aladin
  - IPyAladin: A bridge between Jupyter and Aladin Lite, enabling interactive sky visualization in IPython notebooks.
- Hips (Hierarchical Progressive Survey)
  - Hierarchical tiling mechanism which allows one to access, visualize and browse seamlessly image, catalogue and cube data.
  - http://www.ivoa.net/documents/HiPS/

Technology Summary: Jupyter Plotting Tools

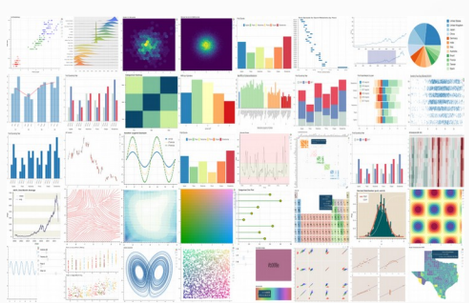
- Matplotlib
  - Python 2D plotting library
  - Simple, but widely used
  - Static or Interactive plots



- Plotly
  - Over 30 chart types, including scientific charts, 3D graphs etc.
  - Better looking graphs
  - Greater level of interactivity
  - Can be slower to render



- Bokeh
  - “Elegant, concise construction of versatile graphics”
  - High-performance interactivity over very large or streaming datasets.
  - Have not experimented with it yet



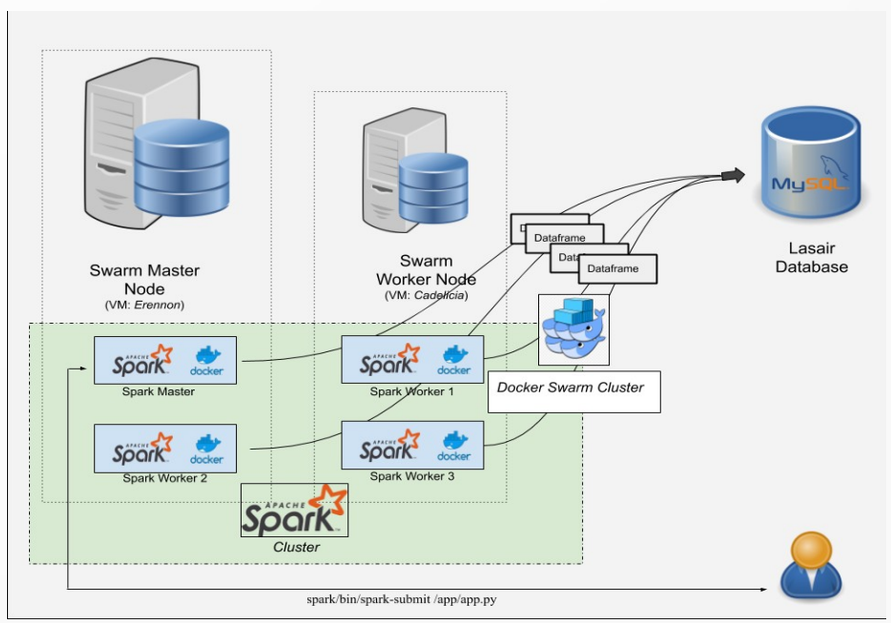
- Others (bqplot..etc)

REPRODUCIBLE WORKFLOWS: Next steps

- Currently building Docker Containers with all required (VO & other) libraries
- Notes & Automated Deployment scripts for running JupyterHub Service

BATCH PROCESSING

- Experiments with SPARK
  - PySpark on JupyterHub
  - Docker Swarm to setup Cluster
  - Spark Nodes as Docker Containers



BATCH PROCESSING:Next steps

- Setup Spark on Hadoop cluster
  - Yarn Deployment mode
  - **Cloudera**, Hortonworks, MapR
- Create reproducible deployment using:
  - Ansible
    - Configuration management, application deployment, orchestration and provisioning
  - Docker