

WP5 - Interactive Data Analysis WG Report

Stelios Voutsinas
University of Edinburgh

Summary of Goals of WG

- Based on original Project plan for WP5:
 - ESAP should implement a set of tools that allow interactivity and visualization.
 - Data Visualization is one of the core services to be provided.
 - A framework to find, deploy and configure those tools on ESAP is necessary.
 - Data exploration of large data sets requires interaction with large computing facilities. Jupyter is a possible framework to offer interactive visualization capabilities, together with remote graphical access services.
 - Users should be able to select from an existing list of Workflows (Notebooks) and either download, or deploy on available facilities



What we are **not** building

- We are not building one central ESAP JupyterHub service, to be used by all users

What we **are** building

- Project specific Prototype JupyterHub Services to be integrated as test cases for the Platform
- A federation (network) of Compute facilities that can be discovered using ESAP
- A tool for discovering Workflows for different science use cases
- A way for a user to spawn an Interactive analysis environment with a known Workflow and dataset in an automated way

(* All of this is subject to change!)

User Interface Flow



Logged in as: stv

Find Compute Facilities:

Or Choose from a list:

- JHub Edinburgh
- Spark Cluster Edi
- JHub SKA

Your data: Dataset #1

Logged in as: stv

Showing Results for: "JHub"

- JHub Edinburgh
- Spark Cluster Edi
- JHub SKA

Additional Info for Selection for "JHub Edinburgh":

Available Interpreters:

- VO Access
- ML Libs
- Python3.6

Your Data: Dataset #1

Logged in as: stv

Find Workflows:

Or Choose from a list:

- Notebooks
- Workflows
- Containers
- ML Notebook #1
- My Awesome Notebook
- VO Data Access Note

Your Data: Dataset #1

Your Compute: JHub Edinburgh

Logged in as: stv

Checkout:

Your Data: Dataset #1

Your Compute: JHub Edinburgh

Your Workflow: ML Notebook #1



Description of example usage

- User searches using a keyword-search for known Compute (IDA) facilities.
- User discovers & adds a Dataset to their Shopping Cart.
- User searches and adds a Workflow (Notebook) to their Cart.
- User checks out cart, after which they are redirected to the IDA environment, with data and workflow preloaded (if possible)



Plans - Phase 1

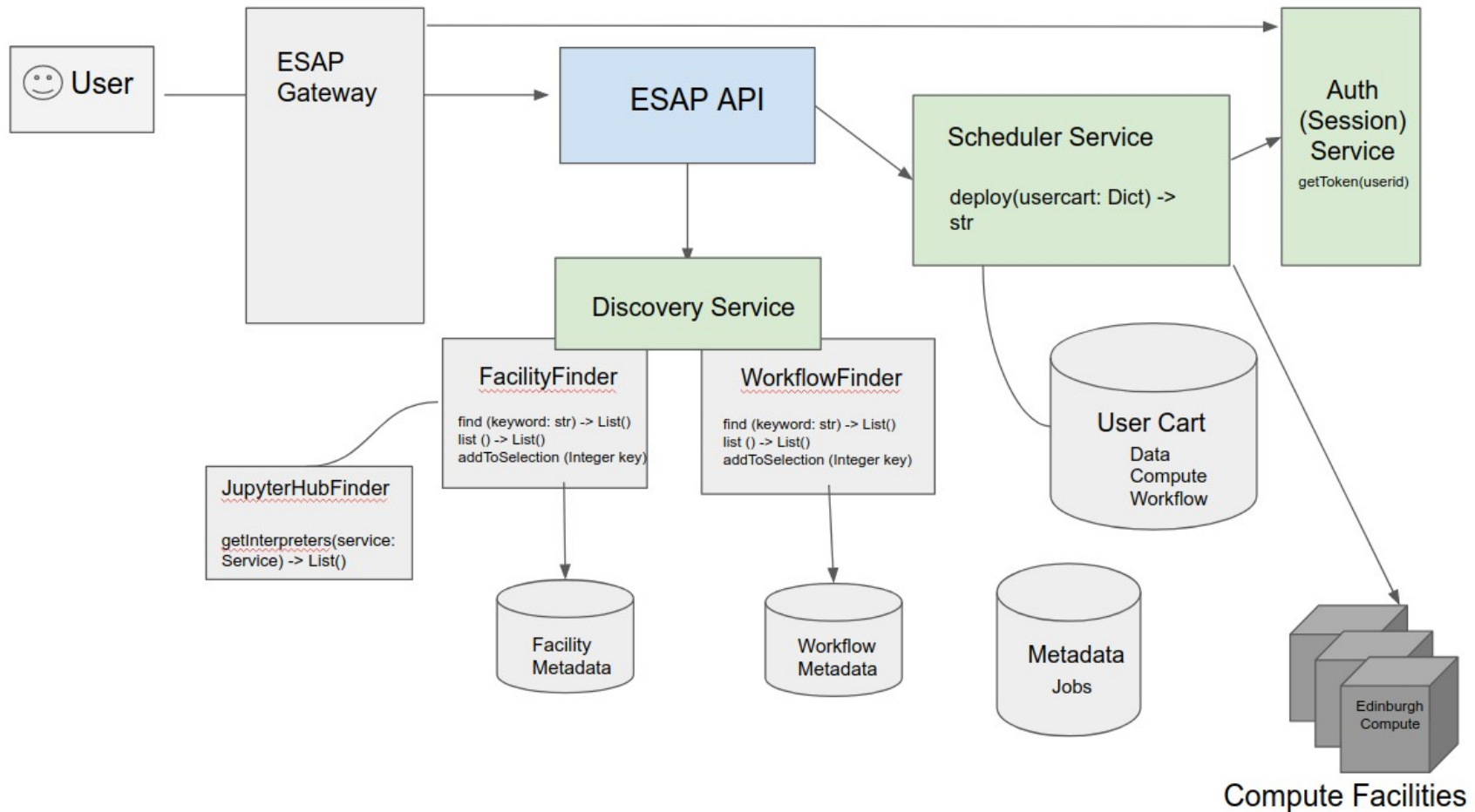
- Enable the described user flow, with JupyterHub & Notebooks as the Compute & Workflows accordingly
 - * to the extent that it is possible, otherwise start with simpler functionality
- Enable Notebook discovery, using public repositories as a first use case
- Allow users to preselect from a known list of JHub facilities
- Automate as much as possible using SSO/Tokens & REST APIs

Progress



- System Design, UI Design & Task Definitions created for Phase 1
- Weekly meetings to discuss and iterate over development
- Some initial code changes committed to enable some basic REST Services
 - Using stub data for now
- A few prototype JupyterHub services in place and ready to be integrated
- Investigation into JupyterHub REST API to see what can & can't be done

System Design



Task & Issue Tracking

The screenshot displays the ASTRON SDC Issues page. The left sidebar shows the navigation menu with 'Issues' selected, indicating 15 items. The main content area shows a list of issues with the following details:

Issue Title	Status	Created	Updated	Labels
Implement Scheduler Service	Open 15	#16 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Define & Develop JupyterHub finding service	Closed 1	#15 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Define & Develop Facility finding service	All 16	#14 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Implement Workflow finding service		#13 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Document if and how to spawn a Jupyter notebook with additional metadata information in the environment		#12 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Document and provide examples of querying the REST API of a JupyterHub for metadata on interpreters		#11 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Document metadata extraction options from Docker hub Repos.		#10 - opened 2 months ago by Stelios	updated 2 months ago	IDA
Implement "UserProfile" model in accounts app		#9 - opened 2 months ago by mever	updated 1 month ago	AAI, Doing

Progress

Django REST framework

Search Facilities

OPTIONS

GET ▾

Get a list of facilities that match a keyword search

If no keyword provided, return all facilities

examples:

/esap-api/ida/facilities/search?keyword=SKA

GET /esap-api/ida/facilities/search

HTTP 200 OK

Allow: GET, HEAD, OPTIONS

Content-Type: application/json

Vary: Accept

```
{
  "description": "ESAP API Gateway",
  "version": "ASTRON - version 21 aug 2020",
  "requested_page": "1",
  "requested_page_size": null,
  "default_page_size": 50,
  "max_page_size": 500,
  "count": 2,
  "pages": 1,
  "links": {
    "next": null,
    "previous": null
  },
  "results": [
    {
      "id": 1,
      "name": "ROE Edinburgh Jupyterhub",
      "description": "JupyterHub instance at the ROE. Available for ROE Users",
      "url": "https://jupyterhub.roe.ac.uk/"
    },
    {
      "id": 2,
      "name": "SKA Jupyterhub Service",
      "description": "SKA JupyterHub instance",
      "url": "http://127.0.0.1"
    }
  ]
}
```



Plans - Phase 2

- Expand Functionality to more than JupyterHub Services
 - Virtual Machines
 - Zeppelin / Spark
 - Other
- Investigate how/if we can use Rosetta as a Scheduler Microservice
- Implement/allow registration of Compute Facilities
- Define required libraries needed at Jhub facilities to enable automated staging



Open Questions & Considerations

- What can be done with the JHub REST API to enable staging and automation of the getting the workflow & data to the compute facility?
- How do we handle AuthN/AuthZ for this?
 - Do we need an account with admin privileges?
 - Can we reuse token of authenticated user in IAM for Jhub service if its also IAM protected?
- Where will we store Workflows (Notebooks) from which we generate the list in the UI?

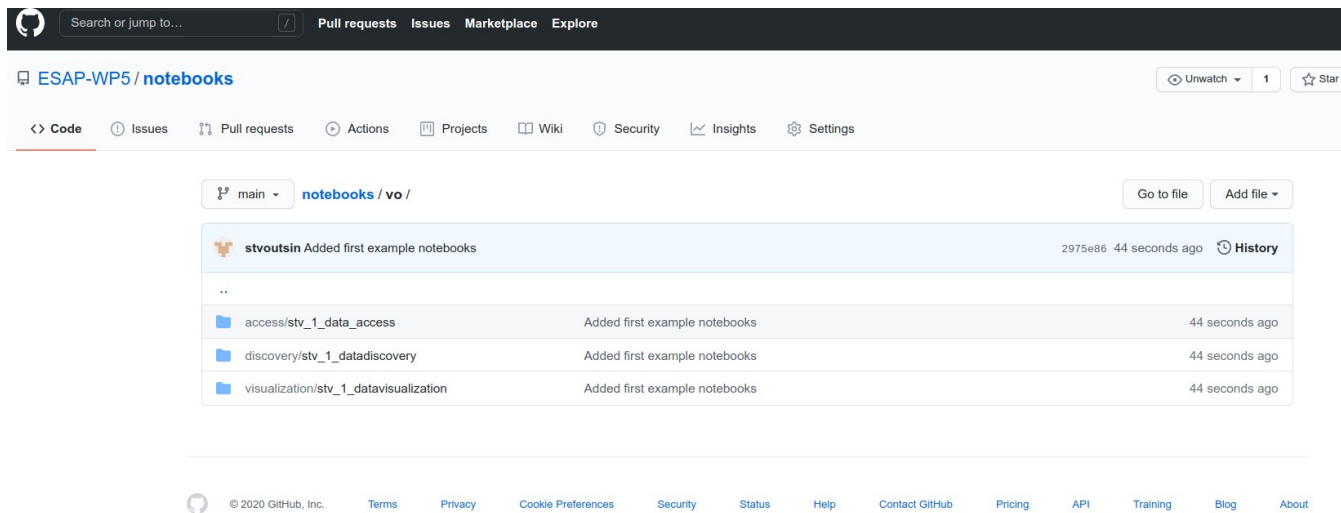


Open Questions & Considerations

- Do all Compute Facilities appear for all users? Are they registered and specific to a user? Public vs Private? And how are they registered?
- How do we collect metadata for a Compute Facility?
 - I.e. JupyterHub: List of Servers? Kernels? Libraries
 - What about VM compute? What kind of information do we display to users? Is there common metadata we can collect and display?
- How does data get staged at the Compute Facility?
 - An idea is to use rucio/vo clients to fetch the data, add them to a notebook cell which we automatically upload and run using the REST API to Jhub .

Where will we store Workflows (Notebooks) from which we generate the list in the UI?

- An initial idea is to collect these example notebooks in a public (Github) repository
- As a first step for our prototype we've setup a public repo:
 - <https://github.com/ESAP-WP5/notebooks>



The screenshot shows the GitHub interface for the repository 'ESAP-WP5/notebooks'. The main content area displays a commit by user 'stvoutsin' with the message 'Added first example notebooks'. The commit includes three files:

File Path	Commit Message	Time
access/stv_1_data_access	Added first example notebooks	44 seconds ago
discovery/stv_1_datadiscovery	Added first example notebooks	44 seconds ago
visualization/stv_1_datavisualization	Added first example notebooks	44 seconds ago

What can be done with the JHub REST API to enable staging and automation of the getting the workflow & data to the compute facility?

- Our current plan involves making calls to the REST API to:
 - Create the user's selected notebook
 - Insert a cell to import the selected dataset
 - Rucio / VO tools
 - Run the first cells so that data staging process is initialized

Links

- ESAP Gateway & UI Repositories:
 - <https://git.astron.nl/astron-sdc/esap-api-gateway>
<https://git.astron.nl/astron-sdc/esap-api-gateway/-/tree/esap-gateway-ida>
<https://git.astron.nl/astron-sdc/esap-gui/>
- ESAP Notebooks Repository:
 - <https://github.com/ESAP-WP5/notebooks>
- User Interface Design Collab Document:
 - <https://docs.google.com/presentation/d/18uxa9njdknJw73Vt2jUij78ot60JQfWwvSubOBA3KLA/>



Questions?