USGS Data Mesh PoC Architecture (2024-01-22)

Data Mesh Component 1: Intra-domain STAC Catalog & JupyterHub server ('Walled Garden')

Deployment

Terraform script ☁️
- Allocates walled garden compute, storage, security groups.
- 🖼 Phase 3

Domain Initialization ☁️
- Via Dockerfile RUN commands
- Initial setup/ingest of static resources
- Regester this domain with the central catalog of catalogs
- 🖼 Phase 2

Server Components (Docker composed together)

Domain STAC Catalog 🏁
- implemented via Pystac/static file or pgSTAC/Postgres 🛢
- 🖼 Phase 1

STAC API ▥
- provides RESTful interface to domain STAC catalog
- implemented via Python/FastAPI Docker container
- 🖼 Phase 1

STAC Browser GUI ▥
- GUI front-end querying local STAC REST API
- 🖼 Phase 2

JupyterHub ☁️
- Run OpenTopo Jupyter notebooks
- long-term: Databricks, or Pangeo Cloud Federation JupyterHub Kubernetes
- 🖼 Phase 3

GSA PoC Customization of STAC API: RegisterDataProduct()

Baseline API endpoint definition:
- STAC API Extension Endpoint 'Transaction'
- Supports HTTP methods POST, PUT, PATCH, DELETE
- Input body: 'STAC Item' GeoJSON Feature augmented with 'Foreign Members'

GSA Customization 1: Self-reported ID 🆔
OAuth2/IAM/AD/ID.me (Stub for PoC)

GSA Customization 2: Provenance ▥
- Enforce existence of 'Link' object of relation type='derived_from'
- could be upstream data service endpoint URL, filename, REST API call, GUI-based map service, etc

GSA Customization 3: Data Product Info ▥
- Jupyter notebook filename/URL (register new notebook now?)
- Source code has to be referenceable
- Plain language description
- return value type of data product (Stub for PoC)

Legend

Four Pillars of Data Mesh:
- 🏁 Domain-Oriented Decentralized Data Ownership
- ▥ Data as a Product
- ☁️ Self-service Data Infrastructure as a Platform
- 🆔 Federated Computational Governance

GSA PoC Feature Milestone Phases:
1. 🖼 February 2024
2. 🖼 March 2024
3. 🖼 April 2024

Hover over/click boxes for links to definitions.