Enabling Real-Time User Interaction For Decision Support: Experiences Extending A Local Agave Platform Metadata Service

Sean B. Cleveland
University of Hawai'i - System
Honolulu Hawai'i, USA
seanbc@hawaii.edu

Jared H. McLean
University of Hawai'i - Hilo
Hilo Hawai'i, USA
mcleanj@hawaii.edu

Kolja Rotzoll
University of Hawai'i- Water
Resources Research Center
Honolulu Hawai'i, USA
kolja@hawaii.edu

Scot K. Izuka
U.S Geological Survey, Pacific Islands
Water Science Center
Honolulu Hawai'i, USA
skizuka@usgs.gov

Gwen A. Jacobs
University of Hawai'i - System
Honolulu Hawai'i, USA
gwenj@hawaii.edu

ABSTRACT

The University of Hawai'i Information Technology Services Cyberinfrastructure team in partnership with the United States Geological Survey developed the Hawai'i groundwater recharge tool, a decision support tool, as part of the 'Ike Wai Gateway to support water sustainability research for the state of Hawai'i. To enable the development of the tool within the existing University of Hawai'i Agave platform that the 'Ike Wai Gateway is built on, the Metadata application programmatic interface (API) and Agave infrastructure needed to be extended to store and retrieve spatial data and handle a larger number of requests. The development team successfully added the functionality to the local University of Hawai'i Agave platform instance to support the storage of deserialized vector data and metadata that can be spatially queried across hundreds of thousands of records to enable decision support applications within a science gateway.

KEYWORDS

gateways, spatial, agave, decision support

ACM Reference Format:

Sean B. Cleveland, Jared H. McLean, Kolja Rotzoll, Scot K. Izuka, and Gwen A. Jacobs. 2019. Enabling Real-Time User Interaction For Decision Support: Experiences Extending A Local Agave Platform Metadata Service. In *Practice and Experience in Advanced Research Computing (PEARC '19)*, July 28-August 1, 2019, Chicago, IL, USA. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3332186.3332252>

1 INTRODUCTION

The primary source of freshwater to Hawai'i's groundwater systems is recharge. Recharge primarily originates as precipitation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PEARC '19, July 28-August 1, 2019, Chicago, IL, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-7227-5/19/07...\$15.00

<https://doi.org/10.1145/3332186.3332252>

In Hawai'i, most precipitation is in the form of rain, with some contribution from fog and snow. Only a fraction of the water from precipitation contributes to recharge; the remainder runs off to the ocean by way of streams, or returns to the atmosphere through evapotranspiration (evaporation and transpiration by plants). Changes in precipitation caused by climate changes can affect recharge rates. Changes in land cover can also affect recharge because different types of vegetation and land uses have different rates of evapotranspiration. Irrigation of crops, parks, and golf courses can enhance recharge. In urban areas, leaking water mains and seepage from septic systems can also contribute to recharge (Fig 1).

Accurate estimates of recharge are critical for groundwater-resource management and planning. The University of Hawai'i (UH) Information Technology Services (ITS) Cyberinfrastructure (CI) team in partnership with the United States Geological Survey (USGS), as part of the state of Hawai'i Established Program to Stimulate Competitive Research (EPSCoR) 'Ike Wai project, were tasked with developing a groundwater recharge decision support tool that allows users to evaluate the effects of land-cover and climate changes on groundwater recharge (Fig 2) (<https://github.com/ikewai/HI-GRE>). Recharge data for the decision support tool are computed using the USGS "soil water-balance" program SWB version 2.0[1]. The application provides a baseline estimate of recharge representing recent conditions of precipitation (1978-2008 average) and land cover (2010).

The pilot application is currently limited to the island of O'ahu, but has been designed to be expandable so that other islands can be added in the future. Users can change land cover and rainfall conditions to evaluate the effects on groundwater recharge. Results are displayed as on-screen maps, and users are provided options to generate interpretive graphics or export results in various formats (Fig 2B).

To support the development of the tool within the existing UH Agave platform that supports the 'Ike Wai project's science gateway[2], the Metadata application programmatic interface (API) and infrastructure needed to be extended to store and retrieve spatial data and handle additional request volume. The development team successfully added the functionality to support the storage of deserialized vector data that can be spatially queried across hundreds

Figure 1: Soil Water-Balance (SWB) Model Diagram for Recharge Calculation. SWB accounts for water entering, leaving, and being stored within the soil on a daily basis. Most groundwater recharge originates as precipitation. Irrigation contributes additional water in some areas (the net effect on groundwater resources may be positive or negative, depending on whether the irrigation water comes from groundwater or surface water). Some of the water runs off the land surface or evaporates. The remainder infiltrates the soil, where it is subject to uptake by plants through transpiration (the combined effect of evaporation and transpiration is referred to as "evapotranspiration"). In the soil water-balance computations of SWB, groundwater recharge occurs when more water infiltrates the soil than can be held in the soil given its water-storing capacity, antecedent water content, and losses through evapotranspiration. In some areas, additional water is contributed by leaks from ponds, reservoirs, water mains, sewer lines, and cesspools; in the SWB simulations, these contributions are considered "direct recharge" that bypasses the evapotranspiration processes in the plant root zone.

of thousands of records. This enables decision support applications for the state of Hawai'i EPSCoR 'Ike Wai project, that focuses on water sustainability in Hawai'i.

This paper will present the experiences in developing and implementing these modifications to the UH Agave platform's Metadata service and configuration of infrastructure to enable water recharge decision support.

2 BACKGROUND

In order to address the diverse needs of statewide community stakeholders and the research aims of the Hawai'i EPSCoR project, an open source water science gateway was developed at the University of Hawai'i (UH), called the 'Ike Wai gateway. In Hawai'i, 'Ike means knowledge and Wai means water. The 'Ike Wai gateway aims to provide centralized web-based user interfaces and application programming interfaces (APIs) supporting multi-domain data management, computation, analysis, and visualization tools to support reproducible science, modeling, data discovery, and decision support for the Hawai'i EPSCoR 'Ike Wai research team, water managers, decision makers, and the wider Hawai'i water community in order to answer questions related to water sustainability.

To deliver a solution to the array of challenges in the project, Agave [3,4,5] was chosen as the platform to build the 'Ike Wai gateway due to the wide range of supported functionality that enable supporting the above needs and UH's partnership with the Texas Advanced Compute Center (TACC) that supports it. Agave is an open source, platform-as-a-service for hybrid cloud computing and data management that is powering a number of current community science gateways like the iPlant Collaborative [6,7], now CyVerse (<http://www.cyverse.org/>), which provides life scientists with powerful computational infrastructure to handle huge datasets and complex analyses that enable data-driven discovery. Agave has been in production as the middleware that supports CyVerse with a proven track record. Therefore, Agave was adopted as the middleware API layer for the 'Ike Wai gateway [4].

2.1 Agave Platform

The Agave platform is a multi-tenant, cloud-native distributed system. All services within the platform run as Docker containers, orchestrated as a single microservice architecture. The platform can be viewed as three logical tiers: platform services, science APIs, and support services (Fig 3).

Platform Services contain services providing identity and API management, client registration, tenant administration services, and documentation. These services are where all API requests are routed first and then based on authentication and authorization validations traffic can then be routed to the appropriate science APIs or rejected.

Science APIs contain the primary Science as a Service (SaaS) functionality used to power science gateways such as data and job management, application publishing, notifications, etc. This tier is further divided into a front-end collection of loosely coupled services, microservices, and back-end service workers. The microservices serve the user HTTP requests to the REST APIs, and service workers handle processing of asynchronous requests such as data movement and application publishing.

Support services include databases, message queues, caches, object stores, service discovery, and notification relays. These services are where all platform data are persisted and provenance is tracked.

Agave uses "tenants" to represent a group of users, applications, and entitlements. Each instance of Agave can support one or more tenants. Individual tiers and microservices can be replicated, configured, and scaled in support of a specific tenant, or shared to better leverage a single, consolidated resource footprint.

Figure 2: Hawai'i Groundwater Recharge Tool User Interface. A.) The initial map view representing the current land cover usage for the island of O'ahu with a yellow box representing a change in default land-covers to a High Intensity development without storm drains.. B.) Recharge visualization map of O'ahu with details pane showing baseline vs user changes (the box) and the recalculation of the island recharge.

Since 2016, the UH-ITS Cyberinfrastructure team has managed a full Agave deployment within the University of Hawai'i data center. The UH Agave deployment has been used to support researchers computational workflows and has been specifically leveraged to build the 'Ike Wai gateway that support the Hawai'i EPSCoR project. By adopting the Agave platform, UH has successfully deployed a science gateway that ties data and advanced computing resources together to support domains including hydrology, microbiology, geochemistry, geophysics, economics and humanities with computational and modeling workflows delivered in a user-friendly web-based interface as well as REST based APIs.

3 THE CHALLENGE AND APPROACH

Agave offers almost everything necessary for all the use cases in the 'Ike Wai project out of the box. However, when it came to one of the decision-support-use cases as well as a desire to annotate and query data spatially, there was a gap. This gap revolved around the specific ability to tag information with spatial coordinates in a way that spatial-based queries could be used to retrieve data, for example returning values contained within a search bounding box. Further, there was a need to have access to database-type features for storing data values in addition to the metadata and files service supported by the Agave platform. However, in this instance Agave already provides a Mongo database that could be leveraged for holding data as JavaScript Object Notation (JSON) documents, an open and both human and machine-readable standard that facilitates data interchange on the modern web, using the Metadata API.

Unfortunately, our use case of storing hundreds of thousands of recharge values and retrieving them with spatial queries was not a use case that the Metadata API had been developed to address.

As the saying goes "there are many ways to skin a kiwi" and three choices presented themselves as solutions to providing the additional functionality: 1) modifying the existing APIs to support the needed features; 2) develop additional APIs; 3) leverage Agave as it was and work around the issues. The challenge with modifying the existing APIs was the lack of expertise in our local team with the existing code base so the time investment would have been substantial. Developing our own API might have gone quicker based on local API expertise but would still require a significant amount of additional testing and validation in addition to design planning and security reviews. In the end, we went for the option that would produce the desired outcome leveraging existing Agave infrastructure and code we already could trust to behave in a certain way, was secured and vetted by a large group of developers and users, and could be contributed back to the larger science gateway community.

4 ENABLING SPATIAL SUPPORT

Agave's Metadata service utilizes MongoDB as the database for storing metadata JSON documents in a Mongo collection called "Metadata." MongoDB offers geospatial support for data stored as GeoJSON objects, a format for encoding a variety of geographic data structures with JSON. In order to leverage this capability for Agave we needed to do two things. First a MongoDB field that

Figure 3: Agave Platform Components. The Agave platform services (green) accept and process user REST requests by authenticating and validating permissions and then proxying to the Science APIs (blue). The Science APIs perform the various platform tasks related to the users workflow and return results back to the user via the proxy. The Support services (yellow) are accessed by both the Platform and Science API services and provide persistence of platform information and metadata.

```

"value": {
  "longitude": -157.96443908403094,
  "recharge_scenario1": [55.581, null, 47.211, null, 27.945, 42.479, 40.092, 40.785, 15.709, 6.004,
    43.76, 31.521, 63.997, 527.976, 47.192, 78.119, 46.646, 41.904, 39.827, null, 75.462, 44.468, 44.146,
    45.188, 46.895, 454.979, 38.908, 33.51, 35.847, 0.0],
  "latitude": 21.509228955006815,
  "y": 304,
  "name": "dataset12042018",
  "recharge_scenario0": [61.031, null, 52.667, null, 33.259, 48.325, 46.277, 46.414, 17.97, 6.868,
    49.259, 37.144, 63.997, 527.976, 53.319, 84.51, 51.498, 47.384, 45.586, null, 81.795, 50.443, 50.14,
    50.909, 52.348, 454.979, 45.02, 38.858, 41.72, 0.0],
  "x": 443,
  "loc": {
    "coordinates": [-157.96443908403094, 21.509228955006815],
    "type": "Point"
  }
}
[A] }
https://agavetenant_domain/meta/v2/data?q={"$and":[{"value.name":"dataset12042018"},{"value.loc":
  {"$geoWithin":{"$geometry":{"type":"Polygon","coordinates":[[[-157.964776,21.508892],[-
  157.964776,21.509565],[-157.964047,21.509565],[-157.964047,21.508892],[-
  157.964776,21.508892]]]]}}]}&limit=10000&offset=0
[B]

```

Figure 4: UH Agave Metadata API spatial examples. A.) Example JSON Metadata API recharge data schema using the spatial 'value.loc' field. B.) Example Metadata API bounding box GeoJSON spatial query.

would contain the information needed to be established and standardized upon for storing any GeoJSON information for a piece of metadata/data that would be used as part of a query. Second, a spatial index needed to be created within the Agave MongoDB "Metadata" collection that uses the established field, thus allowing MongoDB to support complex geospatial queries across all JSON documents that contain the field. We also chose to use MongoDB's "2dsphere" index to support queries on an earth-like sphere which also supports indexing both GeoJSON object and coordinates for querying.

For the Agave platform instance at UH we standardized on a JSON document field identifier of "value.loc", "loc" being short for location (Fig 4A). The "value" field is what Agave uses to store user-defined data within a Metadata document, so all user-defined fields are child fields of "value." With these straightforward modifications to the Agave MongoDB the existing Metadata API functions and queries executed with no further changes necessary. The queries for searching on spatial Metadata conform to standard MongoDB spatial queries and could easily be utilized in the Agave Metadata API REST GET requests (Fig 4B).

5 SERIALIZING AND STORING MODEL OUTPUTS IN AGAVE

To support the groundwater recharge decision support tool we needed to allow a user to select regions of the island of O'ahu and modify how that region was used by defining its land cover type

for a possibility of 29 types. Once the land cover type was changed for a region the corresponding recharge values for the updated area needed to be used to recalculate the island's recharge. The USGS had pre-computed recharge data using the SWB model for each land cover type for a grid of 688x882 cells that are 75m x 75m each and cover all of O'ahu. These model recharge values are calculated under two climate scenarios, so the results are 29 land cover recharge values for each scenario which is 58 total values for a grid cell. The USGS model outputs the data in a NetCDF format, so essentially 58 matrices of recharge values with latitude and longitude data in a separate matrix. The way we enabled this application to accomplish accessing this information with Agave was by storing the model output data within the Metadata database and querying for the appropriate grid cells with spatial queries. In order to do that we built an ingestor application (https://github.com/UH-CI/landuse_ingestor) that would transform the NetCDF model output to JSON documents for each cell that hold spatial information, (x, y) indices and land-cover recharge values for each scenario (Fig 2A)

The ingestor application was written in Python and leverages the Agave REST Metadata APIs to create the Metadata documents. Initial implementations of the ingestor were serial which upon testing were projected to take over a week to insert the 606,816 JSON Metadata documents, as each request takes close to 1 second to process. Naturally, parallelizing the insertion process was necessary if the data were to be updated in a timely fashion. Thus, a parallel version, using the Python multiprocessing library, that could thread to the desired scale was implemented. Using the University of Hawai'i High Performance Computing (UH HPC) infrastructure the ingestor could be effectively threaded to the maximum core count of 20 on compute nodes with dual Intel Xeon E5-2680v2 'Ivy Bridge' 10-core, 2.8GHz processors. This resulted in 330 insertions/minute and a completion time of 30 hours per dataset ingestion.

Although, dataset updates were to occur infrequently, better performance was still desired so adding parameters to offset dataset row processing so that the ingestor could be parallelized further by dividing up the dataset and leveraging a job array proved straightforward and a simple solution. This implementation worked up to a point and then the entire UH Agave Metadata API service would become overwhelmed and stop accepting requests, resulting in errors in data insertion and effectively stopping usage of the API. It turns out that we were effectively performing a denial of service attack on our APIs when we went over 33 threads pushing requests as fast as possible at the UH Agave Metadata API. So for our basic Agave platform setup we could successfully run 30 threads for the ingestor and keep the Metadata API up and responding and processing requests. That brought down processing time to 20 hours per dataset ingestion.

Better performance was still desired as well as service reliability. So the next step explored was adding more capacity for handling the Metadata creation request. The base UH Agave deployment has a single copy of each science API microservice, a docker container, running in production. It was a relatively straightforward task to add an additional service since the Agave Deployer management tool that UH uses to deploy and administrate our Agave instances supports high availability of each service out of the box, meaning two services can easily be active for each API. The Apache proxy

service that sits in front of all the science APIs (Fig 3) is configured with ports for each set of redundant services. In order to add our second Metadata API we modified docker compose files to include an additional service and the appropriate port routing that the Apache proxy service configuration was looking for in a second Metadata API microservice. With these relatively minor updates we were able to add an additional container service to handle requests in a manner that would be sustainable to manage with the current UH Agave management strategy[5].

With the addition of the second Metadata API microservice we were able to ramp up the job array to 3 jobs at 20 threads each for a 60 thread total resulting in a processing time of 10 hour dataset ingestion completion time. The addition of the second Metadata API microservice also improved the performance of the decision support application response when fetching hundreds of thousands of grid cell documents, which was an excellent side effect.

6 OTHER GOTCHAS

During the course of this work, we ran into some issues that required some workarounds when making such heavy and "off label" use of the Agave Metadata API service. One issue that the Agave development team predicted we would run into was the way that permissions were implemented. The Metadata service appends document universally unique identifiers (UUIDs) to queries for non-administrative users to ensure users only access metadata that they have permissions to retrieve. When we do crazy things like insert 600k+ records we run into things like exceeding MongoDBs maximum BSON document size when 600k UUIDs are used in a query. Our workaround for this was simple, a specific application account was used for this data and it's queries do not append metadata document UUIDs. This account has limited queries and access within the Agave API manager to the other APIs and REST functions. Another permission related issue we ran across was during all the testing of ingesting datasets multiple times; we would insert millions of JSON documents and eventually see the system become extremely unresponsive. Cleanup by removal of unnecessary JSON documents would not fix the issue. We eventually discovered that the performance had nothing to do with MongoDB having millions of records but rather that we had created millions of MySQL permission records, as we were adding permissions for our gateway administrators. Removal of the MySQL permission records would return performance. To ensure this wouldn't be a future issue we didn't add explicit permissions for the other gateway administration account for the recharge datasets.

7 FUTURE WORK

The current implementation of the software is a correctly working product, however, we feel that additional improvements would allow our decision-support features to scale and perform better for our stakeholders and potentially the broader community. We plan to develop APIs that support the bulk insertion of JSON data objects with integrated document permission to enhance Agave's existing APIs to better support this use case. Also, we are looking to adopt Kubernetes to manage the science API docker containers so that scaling individual services, like the Metadata API, becomes a simple configuration change.

8 CONCLUSION

The Agave platform is a robust framework for supporting science and gateways and can be successfully extended. The UH Agave platform showcases the ability to deliver decision support features with Agave that include spatially indexed and queryable data and handle the insertion and queries of hundreds of thousands of JSON documents based on the modifications we have implemented.

9 ACKNOWLEDGMENTS

This work was made possible by grant funding from the following: RII Track 1: 'Ike Wai Securing Hawai'i's Water Future NSF OIA #1557349, Collaborative Research: SS2-SSI: The Agave Platform: An Open Science-As-A-Service Cloud Platform for Reproducible Science NSF ACI #1450413

REFERENCES

- [1] Westenbroek, S.M.; Engott, J.A.; Kelson, V.A.; and Hunt, R.J.; 2018, SWB Version 2.0: A soil-water-balance code for estimating net infiltration and other water-budget components: U.S. Geological Survey Techniques and Methods, book 6, chap. A59, 118 p., <https://doi.org/10.3133/tm6A59>.
- [2] Cleveland, S.B.; Geis, J.; Jacobs, G.A.; 2018. "The 'Ike Wai Gateway-A Science Gateway For The Water Future of Hawai'i" Proceedings of Science Gateways 2018, Austin TX, USA September 2018 (SGC18) DOI:<https://doi.org/10.6084/m9.figshare.7152464.v2>
- [3] Dooley, R. et al. 2012. "Software-as-a-service: the iPlant foundation API." 5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS) (2012).
- [4] Dooley, R; Fonner, J; Jacobs, G.A; Brandt, S. (2018): "The Agave Platform: An Open Science-As-A-Service Cloud Platform for Reproducible Science." figshare. Poster. <https://doi.org/10.6084/m9.figshare.4675765.v2>
- [5] Cleveland, S.B.; Dooley, R.; Perry, D.; Stubbs, J.; Fonner, J.M.; Jacobs, G.A.; 2018. "Building Science Gateway Infrastructure in the Middle of the Pacific and Beyond: Experiences using the Agave Deployer and Agave Platform to Build Science Gateways" Proceedings of the Practice and Experience in Advance Research Computing conference, Pittsburgh, Pennsylvania USA, July 2018 (PEARC18),. <https://doi.org/10.1145/3219104.3219151>
- [6] Goff, S.A. et al. 2011. "The iPlant Collaborative: Cyberinfrastructure for Plant Biology." *Frontiers in Plant Science*. 2, (2011).
- [7] Merchant, N. et al. 2016. "The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences." *PLOS Biol.* 14, 1 (Jan. 2016), e1002342.