**Dynamic Soil Properties**

**Draft - Analysis of Alternatives (AoA) Document**

**–**

**Olympia ART**

**Technology Solutions Provider Inc. (TSPi)**

**January 8, 2021**

**Draft**

Contents

[1) Document Objective 5](#_Toc61955649)

[2) Background 5](#_Toc61955650)

[A. Soil Science and Resource Assessment Division (Soil Properties) 5](#_Toc61955651)

[B. Science & Technology Division: NRCS Soil Health Division (Soil Health) 5](#_Toc61955652)

[3) Business Need 6](#_Toc61955653)

[4) Desired Future State 6](#_Toc61955654)

[A. Capabilities 6](#_Toc61955655)

[5) Business Priority List Received 12-15-2020 8](#_Toc61955656)

[7) Proposed Future State – Short-Term 3-6 Months 10](#_Toc61955657)

[1. SQL Server 2019 10](#_Toc61955658)

[2. PostgreSQL with PostGIS 10](#_Toc61955659)

[8) Technology for Further Investigation – Long-Term 11](#_Toc61955660)

[9) Analytics Tools 12](#_Toc61955661)

[10) Data Management Maturity (DMM) 12](#_Toc61955662)

[11) Workflow Tracking 12](#_Toc61955663)

[12) Requirements & Evaluation Approach 13](#_Toc61955664)

[A. Comparison Method 13](#_Toc61955665)

[B. Scoring Method 14](#_Toc61955666)

[C. Products Included in Evaluation 15](#_Toc61955667)

[13) Analysis of Alternatives 15](#_Toc61955668)

[A. General Findings 15](#_Toc61955669)

[B. Products 15](#_Toc61955670)

[C. Alternative Scoring Summary 15](#_Toc61955671)

[D. Alternative 1: AWS Redshift 15](#_Toc61955672)

[E. Alternative 2: Microsoft Azure Synapse 15](#_Toc61955673)

[F. Alternative 3: Google BigQuery 15](#_Toc61955674)

[G. Alternative 4: Crunchydata PostgreSQL 15](#_Toc61955675)

[H. Alternative 5: Snowflake 15](#_Toc61955676)

[I. Alternative 6: MongoDB 15](#_Toc61955677)

[14) Alternative Comparison 15](#_Toc61955678)

[15) Recommendations 15](#_Toc61955679)

[16) Appendix 15](#_Toc61955680)

**Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Document Version History** | | | |
| **Version** | **Primary Author** | **Date** | **Notes/Comments** |
| Draft | H. Baumgartner, L. Bolleter | 12/07/2020 | Initial Draft |
| 0.5 | H. Baumgartner, L. Bolleter, D. Gahr, K. Mader | 12/22/2020 | Added AoA Portions for Platform, Tools, Data Management Maturity (DMM), and Workflow Tracking Document |
| 0.6 | H. Baumgartner, K. Mader | 12/28/2020 | Added content, questions |
| 0.7 | L. Bolleter | 1/6/2021 | Additions, clarifications, and edits |
|  |  |  |  |

# Document Objective

The Soil and Plant Science Division (SPSD) of the Natural Resources Conservation Service (NRCS) needs the ability to bring Soil data from multiple resources into a common collection hub. The hub is referred to as the Dynamic Soil Property (DSP) Hub. The DSP Hub will provide data for analysis for USDA and University scientists initially followed by public access for appropriate data and analytics.

In coordination with the Conservations Innovation Grants (CIG) project within the Science and Technology (Area, Division, ?) and the National Soil Survey Center (NSSC).

This analysis of alternatives (AoA) document provides background on SPSD’s needs, a summary of warehouse tool alternatives, a summary of how Technology Solutions Provider Inc. (TSPi) evaluated tool alternatives, and data collection tool recommendation for SPSD moving forward.

# Background

Background on the two divisions that that are the focus of this initiative.

## Soil Science and Resource Assessment Division (Soil Properties)

Within NRCS is the Soil Science and Resource Assessment (SSRA) Deputy Area that has operational control of the Soil and Plant Science Division, National Soil Survey Center (NSSC), Dave Hoover, Director. Within the NSSC are there are 6 branches. Those impacted for this initiative is the Soil Business Systems and Soil Science Research.

The SPSD products are available for internal, other federal agencies, and private or public partners. Their products are critical decision-support tools to conservationists, farmers, ranchers, producers, as well as federal, state, local, tribal, and private partners. Their products are the authoritative source for soil, plant, and ecological information for the United States, its territories, and international cooperators, making it the largest source of natural resource information in the United States.

## Science & Technology Division: NRCS Soil Health Division (Soil Health)

Over the last decades the critical importance of applying principles of soil biology and interactions between soil biological, physical, and chemical processes toward system-based soil health management, has gained significant recognition and understanding. Improving soil health on our nation’s agricultural lands will allow farmers and ranchers to simultaneously improve water quality, increase soil water availability, enhance resilience to extreme weather, enhance nutrient cycling, increase carbon sequestration, provide wildlife habitat (including pollinators), enhance rural economic opportunity, and meet the food production needs of a rapidly growing population on a shrinking available land base. Efforts to improve soil health will thus provide significant return on the nation’s conservation investment. In 2014, recognition and demand from stakeholders led to the initiation of the new NRCS Soil Health Division. The purpose of the new Soil Health Division is to incentivize and facilitate producers in implementing science-based, effective, economically viable soil health management systems on the nation’s diverse agricultural lands, in collaboration with partner organizations. Key goals of the Soil Health Division include providing advanced training to diverse audiences, and facilitating soil health assessment, farm/ranch scale soil health management planning adapted to local conditions, and assistance for on-the-ground implementation of soil health management systems

# Business Need

This initiative is to support the needed data collection for soil health or dynamic soil properties database for storage of dynamic soil properties which include CIG and EQIP funded related soil health field and laboratory metrics, operational outcomes, land use and management information that could be linked to current and past conservation practice standards.

This information will also provide needed data to provide for the compilation and analysis of effective conservation practices for soil health and conservation programs.

The Dynamic Soil Property initiative is to create a Dynamic Soil Properties (DSPs) data hub to include properties that change with land use and management on a human time scale. DSPs are often a measure of soil health, ecosystem change, and conservation evaluation. Successfully documenting DSPs requires novel data collection and analysis, and a data verification process. DSP projects combine disciplines and properties as well as collection hierarchies across space and time that do not easily fit in any existing data infrastructure.

The DSP data hub would store standalone information about the CIG On-farm Soil Health Demonstration Trail project context (and eventually on aggregation and extrapolation) as well as fetch and link to data in other databases, such as: soil properties (NASIS), ecological sites, states and processes (EDIT), and management information (CR-LMOD). Ingestion of cooperator data into the National Cooperative Soil Survey Repository will be facilitated through storage of soil health methods metadata and associated information.

Intended Audience are the users of a soil health and dynamic soil property database which include CIG Conservation Practice Database (CPD) users, NRCS employees and technical services providers, National Cooperative Soil Survey cooperators (e.g., universities, state and local governments, non-government organizations, and other federal agencies) and other interested individuals. Authentication will likely be required for some portion of intended audience. Some or all of the publicly accessible datasets will have the same user base and may or may not require a certain level of authentication and scrubbing the data.

# Desired Future State

## Capabilities

* 1. Database storing soil properties, location, field management information, soil properties and metadata over time.
     1. Data elements needed

1. Rationale from projects and data collected (why, what does it represent)
2. Multi-scale hierarchy of field, map unit and sample location information (spatial, temporal, and conceptual)
3. All other site identifiers available in NASIS
4. Calculated and derived fields to aid in aggregation and reporting
   * 1. Access for NRCS user data input and data curation
     2. Data structure that captures spatial and temporal elements (potentially ArcGIS server and novel databases)
     3. Manipulation and reporting capability (output)
     4. Soil properties management
     5. Big data capabilities
     6. Machine learning
     7. Supports statistical analysis

* 1. Links to existing corporate databases, tools, and applications:
     1. CIG CPD
     2. NASIS – specified data sets
     3. CR-LMOD – current and historical land use, management, and operation
     4. CD/NPAD- conservation plans and conservation practices
     5. EDIT – description of ecological sites, including state and transition models with application to multiple land uses
     6. SCAN or other regional and local soil climate and weather data
     7. Outcomes (under development with Soil Health Division)

1. Function measures
   * 1. Production/yield
2. Economic outcomes
   * 1. Inputs
     2. Profit
     3. Risk assessment?
     4. USGS National Geospatial Program datasets
     5. USGS National Hydrography datasets
     6. Multi-Resolution Land Characteristics (MRLC) Consortium datasets
     7. NASIS gSSURGO
     8. NASIS SSURGO/STATSGO2
     9. USGS Moderate Resolution Imaging Spectroradiometer (MODIS) datasets
     10. EarthEnv cloud cover datasets
     11. United States Census Bureau County data
     12. WorldClim datasets
     13. Oregon State University PRISM datasets
     14. University of Maryland Global Forest Change datasets
     15. American Association for the Advancement of Science Forest datasets

* 1. Data Management Maturity
     1. Define and document the science question to be answered
     2. Establish criteria for valid source data
     3. Identify and configure proper criteria and decision models to create authoritative data
     4. User interface to input data with choice lists that guide users to use the same terminology
        1. Links to existing corporate databases – choice lists

* 1. DSP and Soil Health products available through multiple interfaces, should include:
     1. Spatial and tabular data over time
     2. Soil Survey information about soil series, map units and properties specific to land use and management scenarios
     3. Spatial maps with predicted properties under variable land management scenarios
     4. Heatmaps
     5. Soil Health indicator potentials (both tabular and spatial)

* 1. Accessibility of data:
     1. Import data as authorized
     2. Query builder- so users do not need to know structured query language to extract data
     3. Structured query language – maintain flexibility to use structured query language
     4. Download/package datasets for R analysis
     5. Export data as authorized
     6. Deliver data to system
     7. Publish to externally facing system

* 1. Integrated version control for the data such as
     1. Version source data before any data cleanup that might be done
     2. Record methods used to create a soil property model
     3. Potentially multiple versions of draft data before reviews

Version of data reviewed and certified by authority

* 1. Ability to track and curate data with multiple levels of authority
     1. eAuthentication, roles and scopes will be required for some levels of DSP HUB users

# Business Priority List Received 12-15-2020

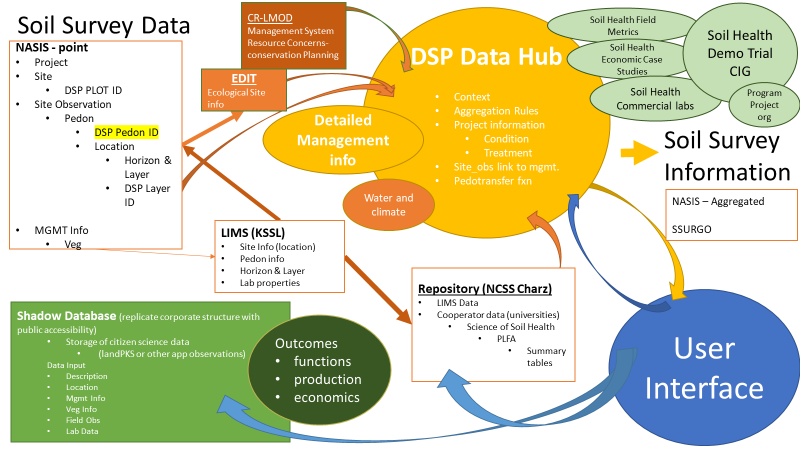
Technology Needs from soil scientists:

* PostgreSQL
* PostGIS
* R and RStudio)
* R Studio Server
* ArcGIS server
* Rules engine might have to be a google earth engine or something similar or more flexible.
* Platform to tie everything together (most important but still TBD)

The platform must allow for the data stewards to do the following:

1. Platform must support robust geospatial analysis
2. Data visualization needs to be included in every part of the process (loading, analysis, testing statistical models and algorithms)
3. Platform must be able to ingest/import raw data and allow for tagging and processing of data, adding additional data such as chain of custody, data source information, and other metadata. Some of this should be automated but need flexibility to do manually. Chain of custody should include a quality check(s) and approval(s) for upload.
4. Data Quality Control - ensure consumer expectations and criteria are met.
5. Platform must support developing statistics/algorithms/rules to perform analysis on existing reference data and raw data
6. Tool needs to support manual querying and filtering to curate and aggregate data
7. Tool needs to support automated data mining and machine learning
8. Ability to aggregate and analyze data for some simple quick summaries (counts, averages, quick stats) from the database
9. Data unification and joining from multiple data sets to create new datasets and performing spatial and tabular joins
10. Ability to create and attribute data products (outputs) sets that will be used in various models
11. Analysis tools that allow for complex statistical framework (building and developing new models)
12. The tool includes a “test mode” with data visualization (charts, graphs, maps) to develop and test statistical models and algorithms
13. Tool needs to support generating and providing different output types, including customization of outputs
14. Data curation for external users allows for automated decision-making on systems that an organization uses to manage its interactions with customers, employees and suppliers.
15. The tool should allow for users to perform the above without doing scripting
16. Options to use R Studio server and publish to shiny app
17. Options for Python and SQL
18. API services
19. Predictive modeling
20. Overview Diagram from Business

The DSP Hub is a data aggregation platform requested from NRCS SPSD. Their initial diagram:



*Diagram from: Idea Document – Dynamic Soil Properties (DSP) and Soil Health (SH)*

The DSP hub requires a database platform, a library of reference files, and data mart capabilities. Integral to the DSP hub is validation of data as authoritative along with documentation on the lineage of the data, information, and report.

# Proposed Future State – Short-Term 3-6 Months

For the three-to-six-month timeframe there are two options being considered:

### SQL Server 2019

* + Environment considerations
    - Strong support with DSO/PSO team, with need to get additional direction on cloud-based solution with migration strategy and feasibility
    - Currently loaded on a Soils Development Server provisioned with 2 TB of storage capacity
    - Need to consider an order of magnitude jump in storage needed over a few months as data sources are brought in and as data stewards start creating model versions, logic used to create the model is recorded, and to run statistical models
    - ESRI ArcGIS available to augment SQL Server spatial capabilities
      * Established SQL Server CI/CD pipeline process
  + Technical considerations
    - Extensive list of services and support including:
      * Data warehousing support and columnstore indexing
      * Machine Learning Services with support for Python and R
      * Analysis Services
      * Integration Services (ETL via SSIS)
      * Master Data Services
      * Data Quality Services
      * Time-series
      * Spatial datatypes
    - Security capabilities
    - Fault-tolerance capabilities
    - Management toolsets
    - Development toolsets
    - No cloud migration strategy “lift and shift” currently exists
      * To stay on SQL Server, the cloud option would need to be Azure Synapse with spatial extensions because it supports R Services (timeframe consideration: Azure and its applications on Azure are not yet platform ATO’d)

### PostgreSQL with PostGIS

Business has requested consideration of PostgreSQL with the accompanying PostGIS spatial package. PostgreSQL and PostGIS are open-source software (SW) applications. PostgreSQL and PostGIS have better spatial capabilities compared to SQL Server without ESRI ArcGIS. Additional detail on capabilities and differences with be added in future versions of the DSP Hub AoA.

The TSPi team met with the FPAC DSO/POS team about investigative work with PostgreSQL and PostGIS for the DSP Hub project. VMs can be provided, but the following cautions on the use of PostgreSQL with PostGIS in the DISC environment included:

1. Platform ATO complications and possible time extension
2. Security requires additional work and is a risk to the team and department developing the software
3. Patching is not provided
4. Support and internal expertise is not provided by DSO/PSO
5. Performance monitoring is not provided
6. Backups are not provided, including PostgreSQL w/ESRI
7. Approval of a non-mainstream software stack
8. SAFe implication of having working software in production
9. Maintenance – Beverly caught that I had not included this
10. Will need to establish a CI/CD pipeline process
11. Less important – industry adoption and cost

# Technology for Further Investigation – Long-Term

After the initial three-to-six-month period additional cloud options will be pursued including:

1. Major Cloud Service Provider (CSP) platforms:
   1. AWS Redshift with spatial extensions
      1. EDAPT to be investigated (USDA data lake) - Requires additional analysis to weigh pros/cons.
      2. Athena or Spectrum?
   2. Azure Synapse with spatial extensions (timeframe consideration: Azure and its applications on Azure are not yet completed the ATO process)
   3. Google BigQuery GIS – rounding out 3 main CSP offerings
2. Additional Hosted Solutions
   1. Crunchydata PostgreSQL – from Dylan Beaudette and Jason
      1. Strong geospatial capabilities with available inclusion of PostGIS
   2. Snowflake
      1. Geospatial needs to be thoroughly assessed
      2. Investigate if Athena would be complementary
   3. MongoDB
      1. Geospatial needs to be thoroughly assessed
      2. Investigate if Athena would be complementary

Benefits of a cloud platform may include:

* FPAC security recommends moving toward AWS
* Potential for instant sandboxes
* Expandable resources (faster and likely less expensive to add instances as needed)
* Lower built-in system administration cost (to be documented)

Considerations:

* There is a need to consider DISC to CSP network bandwidth capacity. If DSP Hub needs a lot of communication to databases such as NASIS and NPAD, moving to a CSP may need to wait for those databases to also move to a CSP.
* CI/CD, source control
* Inter-system communication
* Compatibility of data types with ESRI/ArcGIS
* Potential performance issues with DISC <-> AWS or other cloud solution pipeline

# Analytics Tools

Analytics work is typically accomplished with multiple tools available to data engineers and analytics developers. A core set of tools, available in USDA, are:

* 1. RStudio and RStudio Server
  2. Python
  3. Tableau

The analytics data flow process will need to encompass any analytics tool adapted. Determining which tools should be available to DSP Hub users is solely up to the soil scientists using the DSP Hub. The 3 tools listed above are a fundamental set of first available tools. Additional tools can be added.

# Data Management Maturity (DMM)

DMM relies heavily on people and processes. As the DSP Hub project progresses, processes will be developed iteratively. Templates from prior work with systems including PRS, IDEA, and CARS will be reviewed. DMM is a component of workflow tracking. A separate section of this document will cover workflow tracking.

Informatica products and licensing has been purchased by FPAC. It is understood that work has started with the Informatica Data Catalog product. Additional investigation and work on FPAC DMM efforts with Informatica tools will be pursued.

# Workflow Tracking

Current USDA environment tools that already have ATO and can be used for development of a workflow tracking application include:

* Pega can embed REACT or .NET Core, which could be useful for user admin configurations for standardizing soil data methodologies, options etc. similar to configuration/survey-questionnaires. Teams have developed libraries to comply with USDA digital style recommendations. Typically, Pega has lower licensing costs than Salesforce. It has recently been used with Microsoft SQL Server.
* Salesforce
* ServiceNow

Two workflows are currently required for the DSP Hub projects.

* 1. Data engineering - this workflow is typically more straight forward and includes:
     1. Data set(s) identified
     2. Data set(s) aggregated
     3. Data cleaned
     4. Data quality assurance
     5. Initial or subsequent analysis
     6. Preparation of analytics products for review (documentation, establish scale and use case limitations, data steward approvals, etc.)
     7. Submission to review process to gain approval for release as authoritative analytics product
     8. Review analytics product against established criteria within an established business process
     9. Approve or send back to any of steps above for additional QA/analysis
     10. Receive notifications
     11. Set reminders
     12. If approved, publish analytics product through appropriate outlet

Data engineering capabilities are included in the analytics tools of RStudio, Python, and Tableau covered in the Analytics Tools section of this document. Additional tools used also include SQL and Microsoft Excel.

* 1. The analytics process is considered a more creative and flexible process. That said documentation of data sets used, tools used, analytics used, and reproducibility of the model or analytic developed is important. Tools such as R Markdown, R Notebooks, and Python’s Jupyter Notebook cover some of the functionality identified and needed.

# Requirements & Evaluation Approach

## Comparison Method

To find the optimal technical solution for the DSP Hub needs, TSPi and RMS collected functional requirements and use cases. Functional requirements are the features of a system that directly support users in accomplishing their work. TSPi used the functional requirements, along with common non-functional requirements, to develop a requirements questionnaire. This questionnaire included questions to determine how well each alternative met functional and non-functional requirements. TSPi and the DSP Hub team then assigned a relative weight of importance to each requirement/question.

The weight of 1=Lowest and 5=Highest was derived from the following criteria:

* Must Have (Yes/No). This means if the product does not meet this requirement or use case, it will not be considered.
* Priority (H=High, M=Medium, L=Low). For example, a Low would carry less weight than a Medium when evaluating the product.
  + Any requirements or use cases that were considered by the group as “Must Have” and “High Priority” are given a 5 for weight.

This comparison method considers all evaluation criteria and ensures that the most critical criteria is given the highest priority. Table below provides the requirement category, the associated market research requirement question, and their assigned weights.

Table 1: Org Structure Requirement Questions and Relative Weight

|  |  |  |
| --- | --- | --- |
| **Requirement Category** | **Market Research Requirement Question** | **Relative Weight of Importance** |
| Accessibility | Does the product meet Web Content Accessibility Guidelines (WCAG) 2.0 Level A and Level AA Success Criteria (508 Compliance)? |  |
| Does the application provide integration points for external applications to interact with the product? REST APIs, other? |  |
|  |  |
|  |  |
|  |  |
| Authorization | Does the product support SSO (Single Sign-On) such as SAML or OAuth? |  |
| Does the application support role-based access? What user roles does the system support? |  |
| Can the application call a webservice to retrieve a user's assigned roles after authentication is completed? |  |
| Additional |  |  |
|  |  |
|  |  |

## 

## Scoring Method

A score was assigned to each listed requirement/question. If the product fully met the requirement out of the box, two (2) points were awarded. If the product partially (or limited, with customization) supported the requirement, one (1) point was awarded. If the product was not able to support the requirement, zero (0) points were awarded. If for some reason the product met the requirement, and demonstrated additional value for no additional cost, then three (3) points were awarded. The score for each requirement is calculated based on the points multiplied by the weight for that requirement. See table below for the example of scoring method:

Table 2: Org Structure Tool Scoring Method - Example

| Weight of Requirement  (1 thru 5) | Product Meets Requirement | Points Awarded | Score |
| --- | --- | --- | --- |
| 5 | Yes, plus extra value at no extra cost | 3 points | 15 |
| 5 | Yes | 2 points | 10 |
| 5 | Limited | 1 point | 5 |
| 5 | No | 0 points | 0 |

A product that fully meets the requirements (scoring 2 on all questions) would receive a total weighted score of 264 points.

## Products Included in Evaluation

# Analysis of Alternatives

## General Findings

## Products

## Alternative Scoring Summary

## Alternative 1: AWS Redshift

## Alternative 2: Microsoft Azure Synapse

## Alternative 3: Google BigQuery

## Alternative 4: Crunchydata PostgreSQL

## Alternative 5: Snowflake

## Alternative 6: MongoDB

# Alternative Comparison

# Recommendations

# Appendix