What software packages and standardized languages or Application Programming Interfaces (APIs) are current or likely future dependencies for your relevant research and development activities?

Work in service: The data stewardship tool needs to assist users by facilitating data exchanges and confirming data accuracy, relevance, and reliance via uniform processes. *To manage in service, not in control.* The data stewardship tool must allow each project and their principal investigators to own how the data is ingested, created, cured, and loaded. Once that process is complete for each project, not done by a central authority, then the data stewardship tool should facilitate the exchange of data and results, while fostering a collaborative community. The data management tool should enable the ease of data exchanges between the projects but maintain the responsibility for data integrity, accuracy, accountability, and relevance on the principal investigators.

Cataloged data: Discover, describe, and load data across different projects, areas of interest, or business systems – including the Exascale Computing Project (ECP) hardware and data sets. Create a simple, common, and trusted view of data for business users and applications that are cleansed and deduped with integrated profiling tools. support numerical, text, and wildcard searches, making it simple to find existing records and avoid creating duplicates.

Role based access to data: Apply a security model that reflects the roles that users assume in the research and development of their projects. Make it easy for users to work with data and applications with minimized training requirements while ensuring that they have access to just those business assets necessary for their job. And audit all actions for later review with no coding required.

Low code web application: Develop web applications that integrate business logic and exploit data; and complement or extend data quality, business applications, master and application data management, and other processes – all with very little coding. Efficiently orchestrate processes with an embedded workflow engine so that data is actioned quickly.

Management and Oversight: Arm managers and data stewards with custom visualizations of automatically captured metrics, providing up-to-the-minute insights into the status, progress, and evidence of SLA compliance. Orchestrate | Create | Enforce

Curated data and business knowledge: Capture and preserve the business understanding of data and assets that are created to support a data initiative yet are typically lost at the conclusion — making them available for reuse to shorten time value for future initiatives.

Increase data reliability - reduce cost of bad data: Poor data quality is costing research dearly. Need to set universally agreed upon data guardrails and metrics for data owners to label and classify data readiness, validity, and veracity. The data steward must monitor these guardrails and enforce compliance and dramatic improvements in data quality.

What key capabilities are provided by these software packages?

- The integrity of day-to-day transactions: keeps track of all events that occur within the data stewardship program

- Executive decision making: enables executives to impose specific restrictions and compliance checkpoints that all projects must comply with

- Core master data attributes shared between multiple systems: uniformity and clear expectations should be part of the attributes for the data catalog

- Control the processes via a single source of truth: the data stewardship program should be a single source of truth and gather all projects, data, and results

- Empower business teams to take charge of their 'in-control' master data and implement their business rules to get data right the first time—and keep it right.

What key capabilities, which are not already present, do you anticipate requiring within the foreseeable future?

A data stewardship solution that requires no technical coding skills, that enables data owners to author data guardrails that embed the data steward rules, and provides a uniform process for all to follow. The uniformity will allow for better data sharing, collaboration, and validation of projects, data, and results.

Over what timeframe can you anticipate these requirements with high confidence?

- 18-36 months

What are the most significant foreseeable risks associated with these dependencies and what are your preferred mitigation strategies?

- Misusing the data and designing models that are not accurate given faulty data, while disseminating the information as accurate

- Not using the data without high level of confidence for reliability, nor a uniform process that applies to all projects and participants

- Not having access to data that could directly impact solutions or targeted results

- To mitigate, we will continue to search via open source libraries and databases

What strategies and technology do you employ, or intend to employ in the foreseeable future, to ensure the security and integrity of your software and its associated provenance metadata?

- **Maintain a single source of truth**: To ensure data security and integrity of the data and associated metadata, we employ a single source of truth. We avoid data silos that replicate or duplicate data, instead we maintain just one (1) copy of the data at a central location. The centralized data is then made available for research and development.

- **Proactive logs, monitoring and alerts**: Once a single source of truth (centralized storage location) is established, then we can maintain a logging and monitoring mechanism that is constantly scanning the use and exchange of data. Whenever unexpected behaviors are observed, the system automatically triggers a set of notifications and alerts to escalate the event and find a resolution. The logs are safely stored for posterior review of any incident.

What capabilities do you provide, or intend to provide in the foreseeable future, to assist users of your software with ensuring scientific reproducibility, recording the provenance of their work products, securing their information, protecting the privacy of others, and maintaining the integrity of their results?

- **Specific storage locations for each project:** Each user has a specific bucket storage location to record the provenance of their work products. The bucket is encrypted to ensure privacy and only specific users are given access.
- **Immutable Data:** The data stored into the bucket locations are immutable, meaning they cannot be changed or altered once created. By creating immutable records, we guarantee the integrity of the results.
- **Data Catalog:** Once the data is safely cataloged, then it can be placed, with the proper authorizations and pre-approvals, into a centralized location that distributes and shares the project data with any other project or principal investigators.

What infrastructure requirements do you have in order to productively develop state-of-the-art software for scientific and high-performance computing?

- **Cloud based infrastructure**: Our certified team has a series of scalable, highly available, and secure public cloud environments within AWS, Azure, Oracle, and GCP. We have a multi-cloud hybrid model that ensures compatibility with multiple systems and can serve a systems integrator for any user, application, or process. Our team also has experience working with private cloud infrastructures that use direct connections to data sources without the use of public internet domain.

- **On-Premises infrastructure**: Aside from public and private clouds, we have the ability to provide on-premise technologies for high-performance computing in the State of Florida.

These requirements might include access to testbed hardware, testing allocations on larger-scale resources, hosting for source-code repositories, documentation, and other collaboration tools. What are the key capabilities provided by this infrastructure that enables it to meet your needs?

- **Full Research and Development**: We provide each project with a specific loud-based space, sometimes referred to as a container sandbox, that can be customized to meet the specific environment requirements for each project. The container sandbox can be prepared with any type of resource, repository, collaboration tool, library, or framework. We provide different types of containers depending on their on-demand cloud-based functions, the elastic map-reduce (big-data) workloads to be performed, or the integration with machine learning tools lik Jupyter notebooks and integrated development environments.
- **Promotion to Production**: The sandbox once properly configured in development, can then be promoted to a certified environment and eventually to a production environment. Thus, beginning from a research and development sandbox, can then be promoted all the way to production. All from the same convenient location.

How much additional effort is needed to develop and maintain software packages for use by the wider community above the effort needed to develop and maintain software packages solely for use in specific research projects or for internal use? What tasks are the largest contributors to that additional effort? What are the largest non-monetary impediments to performing this additional work? How is any such additional effort currently funded? How does that funding compare to a level of funding needed to maximize impact?

- **Similar Efforts**: Given our scalable system, we can without much additional effort, maintain software packages that are used exclusively by a subset of the community or those used by the whole ecosystem.
- **Catalog of Services**: The process to identify and maintain a catalog of software tools and services used is automatically integrated into our solution, therefore each time a project develops specific tools or uses unique libraries or data sets, the catalog records all the events.

Once the project is ready, the catalog is shared, thus any community member will have access to the same resources used by the project.

Professional environment: What challenges do you face in recruiting and retaining talented professionals to develop software for scientific and high-performance computing?

- **Not enough candidates**: One of the biggest challenges we face is identifying talent. There is a high demand for software and data engineers, and not enough candidates. In order to mitigate this, we have decided to train internally and properly empower our staff to learn about new software tools and libraries.

What additional challenges exist in recruiting and retaining talented professionals from groups historically underrepresented in STEM and/or individuals from underserved communities?

- **Diversity challenges**: We are located in Florida and are proud of our Latino heritage, but it is even more difficult to identify STEM candidates from historically underrepresented communities. We have decided to train our non-experienced candidates. We much rather teach cloud technologies and software tools to underrepresented communities. Additionally with groups in Ft. Lauderdale and Miami, we have engaged with the community to identify talent and motivate future technologists.

As the complexity of the software ecosystem continues to increase, and number of stakeholders has grown, ASCR seeks to understand how it might encourage sustainable, resilient, and diversified funding and development models for the already-successful software within the ecosystem. What are the important characteristics and components of sustainable models for software for scientific and high-performance computing? What are key obstacles, impediments, or bottlenecks to the establishment and success of these models? What development practices and other factors tend to facilitate successful establishment of these models?

- **Bigger communities**: We foster new communities and inclusivity. Any sustainable model should be actively inviting new members to become more involved with initiatives, share data, share results, and even create buzz by advocating or challenging ideas, processes, and conclusions. By creating a bigger and more diverse community, we can encourage resilience and diversity.

- **Tech barriers**: Barriers need to be removed. At times, the data is difficult to access and even harder to identify, let alone classify for accuracy or reliability. We need to remove the barriers and open access with ease. Data catalogs help to increase the use of different systems and by indexing all projects, then easier opportunities for collaboration will emerge. Again breaking down barriers. Data catalogs accelerate the collaboration and dissemination process.

The section labeled Potential Scope, mentioned earlier in the RFI, outlines activities that ASCR currently anticipates potentially including in future programs stewarding the software ecosystem for scientific and high performance computing. Are there activities that should be added to, or removed from, this list? Are there specific requirements that should be associated with any of these activities to ensure their success and maximize their impact?

- **Work-In-Service**: Any steward solution must work for each project in service to the community. Control of the data, the processes, the reliability of the projects, results, and collaboration should be controlled by the projects, the participants, and the principal investigators. A data steward cannot be in control, a data steward must serve. By serving the community we mean eliminating barriers for exchanges, creating uniformity, enabling projects to share, cataloging the data, the services, and the projects.

- **Low code**: Any steward solution must be simple to use, even drag-and-drop type of interfaces, as opposed to a complex coding environment with unique syntax, rules, and libraries.

What do you anticipate will be effective models for management and oversight of the scientific and high performance-computing software ecosystem, and how would that management structure most-effectively interact with DOE and other stakeholders? In addition to DOE, who are the key stakeholders? How can the management structure coordinate with DOE user facilities and others to provide access to relevant testbed systems and other necessary infrastructure?

- **Management Infrastructure**:

  There needs to be a series of governance models to manage the data stewardship program.

  - At the highest level the plan must meet DOE key stakeholder requirements and compliance expectations. This high-level would focus on ensuring the data stewardship program meets the overall requirements and is accountable with key performance metrics.
  - At the second-next level, the data stewardship management team would create an infrastructure that is provided to all projects. Each project would follow a series of uniform processes and have the ability to use additional shared tools and services. These additional community items are easily provided to all projects.
  - Then, finally at the third-lowest management level, each project can be managed by the principal investigator who would ensure compliance with the overall data stewardship program requirements, but would also control the data and processes within their projects.

  The three (3) level management infrastructure ensures that the overall program meets the legal requirements, while at the same time providing each project the flexibility for research.

  Additional key stakeholders should be the LARGER community as a whole - meaning the data stewardship program should invite new members to use and exchange the data, results, and projects. The program should foster the use of the data with innovative and novel initiatives that expand the footprint and impact of the DOE mission

What kinds of metrics or criteria would be useful in measuring the success of software stewardship efforts in scientific and high-performance computing and its impact on your scientific fields or industries?

- **New Users**: The data stewardship program should onboard new users and track the growth of the community.

- **Data Shares**: To increase the accuracy and efficiency of the projects, data should be shared more between projects. By tracking the amount of shared information and identifying key data that seems to be extremely important, we can understand the health of the community and the effectiveness of the projects.

- **Project Collaborations**: A key metric needs to be the amount of collaboration happening between the projects, since that would be a clear indicator about the reliability of the projects, the accuracy of the data, and the health of the community.

- **Event Monitoring**: Proactive alerts and notifications regarding outlier behavior and logging for all events should ensure the resilience of the community and provide safeguards for unexpected events.

- **Scalability**: If done correctly, we would anticipate a growth in the use of resources, sharing of information, and community interactions, therefore a key metric needs to be guaranteeing the growth scalability of the data stewardship program.

- **Resilience**: Any community needs to be kept up and running. By tracking the resilience of the solution we can track the effectiveness of the projects.

What are key obstacles, impediments, or bottlenecks to progress by, and success of, future development of software for scientific and high performance computing? Are there other factors, issues, or opportunities, not addressed by the questions above, which should be considered in the context of stewardship of the ecosystem of software for scientific and high performance computing?

- **Pace of Innovation**: The data stewardship program must keep up with the different environment requirements, data exchanges, and collaboration initiatives to maintain relevance and effectiveness.

- **Red Tape**: Projects need to meet compliance requirements, however the data stewardship program must be flexible for each project to run their own unique scenarios with their own variables, tools and services.

As indicated in our communication of 18 November to Dr. Finkel, our focus when participating in any project is to first clearly define the problem at hand, establish simple and measurable objectives and then identify the areas of opportunity that will allow us to create the most efficient plan, inclusive of an action road map and deadlines.

Moving forward,

Carlos Gonzalez