STATEMENT OF WORK

CHS III
DATA LAKE

A-4.3 *Sample* Task Order #3: Data Lake

This is a sample Task Order for evaluation of the proposal. The response to this Data Lake Order is expected to demonstrate a representative plan that provides a description of efficient and innovative technical support for the USGS Cloud vision and migrate current USGS data sets into the Cloud environment. Additionally, USGS is encouraging innovative ways to demonstrate usage and access mechanisms for those available data sets. Integration of these data sets and supporting applications support local, regional, nationwide, and global science use cases and in all cases provide support to natural resource managers and in some cases support health and safety for our nation. Many decisions makers rely on these USGS systems to monitor hazards and time critical decisions.

The response should consider:

- A task plan for this task order to included (at a minimum) approach, scope, schedule, staffing (using titles from the skill matrix) by month and the basis-of-estimate. It is requested that a schedule primarily focused on major milestones and key deliverables be provided in soft copy using Microsoft Project.
- A representative sample Monthly Status Report as described in Task Order Objectives, which includes performance metrics.
- This task order is labor hours.

Background

The USGS has a need to make available information from scientific and business operations while improving data storage costs, consolidating reporting, and meeting federal mandates for scientific information. A set of business information systems for reporting needs and data calls have been identified as priorities in addition to various on-premise Trusted Digital Repositories containing scientific information. Scientific information, in particular, has a need to be shared external in addition to internally to meet the needs of partner agencies and academic institutions. To help meet this need the USGS recommends creation of a new Data Lake solution with analytical reporting capability that can integrate and broker this information appropriately to many customers and applications.

Assumptions

- All development will be done in the CHS environment.
- All development work will be completed no more than 12 months from the start of the project.

Requirements

Priority 1 -- Data Ingestion

- Sources:
    - Basis+
        - Basis+ is an on-premise Oracle system containing project management and budget data

  - IPDS
    - IPDS is an on-premise SQL Server/ SharePoint based system containing information about scientific products released by the USGS
  - ScienceBase
    - ScienceBase is a primarily on-premise PostgreSQL/ Elastic Search based system having a robust restful JSON API containing information about scientific data products. The system also stores scientific data products in on-premise and cloud storage.
  - FBA
    - FBA is a SQL Server based system containing information about facilities assets and related budgeting information.
  - Active Directory
    - Active Directory is not a traditional database, however data can be queries via PowerShell and LDAP. Active Directory contains information on staff such as employee IDs, username, office location, etc.
  - Various On-Premise Trusted Digital Repositories
    - Several scientific projects and applications currently have on-premise TDRs to preserve and protect scientific data assets.
- Subsets of data from each system will be ingested into the USGS cloud environment either as a replicated database (varying type) or data files (varying type) and stored in the cloud. A common automated solution should be provided for all systems reducing duplication of effort for each system to manage separate environments. Methods will be determined by what is feasible to meet the 12 month period of performance.
- Ingestion pipelines will be limited to existing services available in USGS Cloud or compliant with deployment within USGS Cloud.
- Source databases will not be replicated in full for any system. Some ETL processes will be designed to extract views or summarized tables from source systems as needed.

Priority 2 -- Data Curation and Cataloging

- Data will have a documented governance plan to define ingest schedules and methods and any ETL processes. Systems will define the governance rules, however the cloud solution must provide for multiple data management team roles and development (In Review) and production (approved) stagging of data.
- Data will be cataloged to keep track of the inventory of data assets available in a Web UI that is easily searchable and can provide data provenance information. Automation and integration with Priority 1 – Data Ingestion solution is preferable to manual updates of information.

Priority 3 – Trusted Digital Repository for Scientific data products

- The USGS has federal requirements to release scientific information. These requirements and goals are outlined in the USGS Public Access Plan.
- USGS has a growing number of "big data" scientific products stored in various on-premise TDR's that could benefit from cloud storage, however any cloud solution must meet Trusted Digital Repository requirements.

Priority 4 -- Data as a Service

- The data should be accessible by the USGS Tableau service and other custom USGS developed applications.

- Native connectors or extensions such as a lightweight Restful API is required to access the solution programmatically.
- Documentation will be provided for the API if an API is provided.
- A Web UI interface to access data is also preferred, but not required

Priority 5 -- Long Term Maintenance

- Feasibility and scope of resources needed for long term (10+ year) maintenance and support shall be considered and described, including life cycle or longevity of architecture components used within the solutions.