ODC SOW

To Do

- //www.smartsheet.com/how-write-statement-work-any-industry

Introduction

- USGS/EROS fully supports the ODC Initiative
  - EROS values its partnership with GA on this initiative

  This document defines the joint research specifically between GA and USGS/EROS and benefits the larger science and ODC communities.

Objectives/Purpose

The Big WHY:

> In a nutshell scientists just want to do science!

The old model of: - specifying new equipment; and waiting for it to be deployed - purchasing or deploying scientific tools - filtering and downloading data - sub-setting and further filtering data

Has been disrupted/replaced by a cloud centric approach: - Infrastructure is stood up from code (IaaS) (Infrastructure as Code) - Data has been processed to standard scientific levels - Data is at your fingertips - Cloud tools (like ODC) are present and free to use - Scientists focus on science

> Also; There is a goal of 20 by 20 Cubes in the larger ODC context.

Scope of Work

- This document covers the technical scope of this work only. All political agreements are the responsibility of the governing bodies of each agency.
- The primary scope of this work is to create a Data Cube instance in the cloud that exploits s3 object stored ARD data.

The research for this task will involve the following disciplines at a minimum: 1. Project Management 2. System Administration 3. Cloud Engineering 4. Software Development/Engineering 5. Data Base Administration 6. Data Science

Limiting the Target Use Case

To be effective we recommend limiting the target use case to:

1. Hayden Island (Portland Oregon) ## Spaghetti,
2. US ARD Data
    - Specifically h03v03 ## 2350 scene/tiles
    - target bucket (as of this writing) ga-odc-eros-co3-west.
        - still need to code a distributed COG generator with terraform.
3. AWS cloud provider
    - us-west-2 (Oregon) region # its all about the Oregon
    - ga-aws-usgs - account
    -
4. Data Format Constraints
    - Cloud Optimized Geotiff - This is the **focus**.
        - Comparisons to tiff, geotiff and tarred geotiff may be documented
5. Data Base Constraint
    - PostgreSQL (SQLAlchemy)
    - Run on a small EC2 instance; for first demonstration
    - as with all factors in this project, this could scale up in many ways.
        - Should use Amazon Relational Database Service (RDS) – AWS in round 2/phase 2

Expected Outcomes

Cloud Technology Momentum

1. The cloud ship has sailed; come sail away with me
2. Private industry is already using satellite data in interesting ways
3. USGS/EROS compute and storage roles are almost all embarassingly parallel
4. The AWS cloud services are a good match and amazingly cost effective for this class of compute/storage problems.

GDAL is the key to application and data synergies

How to best optimize data in s3

1. For widest scientific use (end user)
2. For image processing; As a low cost ubiquitous data cache

Develop the data and application roadmap baseline

1. Refine and accelerate the schedule for pets —> cows migration

Reusable Technology Recipes for Cloud Data Exploitation

1. Reusable Infrastructure as code
2. Validate the use of Docker containers as the primary application building block
    - Docker containers still need to be accepted by USGS security team.

Deliverables

S3 Findings Paper and Presentation - (February 14, 2018)

Deadline Approaching

- This work needs to be completed by January 31, 2018 to be displayed on February 14, 2018.

1. Data Storage of one ARD tile (full temporal legacy)
2. Project and Task Plans
3. Data Cube Instances
    - AWS us-west-2
    - private libvirt instances
4. Prototype code for indexing from ARD.xml[2350] to PostgreSQL
5. Demonstration of water though time over the Hayden Island area of Oregon
    - Highlights Landsat strengths
    - Highlights ODC strengths
    - Determination of s3/cloud viability for ODC and Landsat

Terraform low cost container service and simple orchestration model

Success Metrics

Needs some thought and discussion

1. Lines of code
2. Budget adherence
    - Creative cost containment in the cloud
3. Number of demonstrations
4. Number of scenes
    - stored
    - processed per minute
5. Number of meetings and conferences
6. Number of hours on the timesheet
7. Number of AWS services exploited or avoided

Constraints and Risks

Constraints

1. Limited access to project sponsors
2. Limited access to experts at EROS
3. Limited size of AWS instances and mostly transient use
4. Large technology gamut and limited experience in certain technologies
5. So many ideas; so little time
    - Decision fatigue
    - Direction fatigue
6. The status quo
    - Big highly integrated applications (we love our pets)
    - Competing priorities

Mitigations

1. Limit the scope; phase the project
    - Use instincts and filters to determine priorities
    - Manage the queue
    - Favor technologies with easily exploitable abstractions
2. Follow project management rigor
    - Validate priorities with key stakeholder(s)
3. Target all communication to the reader; evaluate the effectiveness
    - for example executive summary level for stakeholders; technical detail among the technologists
4. Schedule monthly briefings
5. Allow for some autonomy and self directed teams
6. Build some contingency lighter weight demos - jupyter notebook - synthetic simulations squared
7. Limit tools and embrace the cloud collaborative paradigm
8. Develop decision filters that help streamline technology and direction choices
    - Find the common denominators
    - Stay agile

Roles and Responsibilities

This is where we define who does what.

GA (Australia)

1. All things Data Cube Code Related
    - Indexing the data

- Demonstrating the user interface
- Testing water over time (Hayden Island)

USGS/EROS (Tony)

1. All things infrastructure and data storage
    - Create and populate buckets
    - Create and tear down infrastructure
    - Document techniques and results; improve automation

Related Work

1. Level1 Cloud Data Storage
    - Commercial use and exploitation of level1 data
    - Collection 1
    - Related white papers on cloud storage