Software Design Proposal
Scientific Data Management System

Alex Fremier
Associate Professor
University of Idaho College of Natural Resources

Colby Blair
Computer Science Undergraduate
University of Idaho Computer Science Department

November 21st, 2011

Contents

List of Figures

1 Project Summary

In today's scientific environment, there is a growing focus on data storage and processing. Some of the largest problems scientists face are related to the data deluge, and the inability to conceptalize problem and solutions from large tracts of data. The result is many attempts by scientific professionals to design software, leading to many bad software designs. The vast amounts of expensively collected data never get processe [**?**], because the software designs tale so much maintenance. Data is often duplicated, lost, and never makes it from collection to analysis, because of the large scale lack of system designs to handle the data.

The proposal is for a project that helps solve these design problems for many scientists. The **PHAT/PORTAL** project is a data intermediary. It is a local web interface that allows scientists to collect, query, organize, and share data with other researchers. It allows users to filter data and prepare it for analysis. It also helps user visualize data graphically, to help the conceptualize, organize, and refine data, before processing it. This project proposes using PTAG data from the Columbia Basin to test design, but should be exdendable to many different users with very different data sets.

1.1 Background

In the Columbia Basin today, millions of federal dollars are spent on PTAG system [**?**] to collect spatial use of fish. Yet, the project results from most research are far from concrete. Despite the lack of understandable results, import decisions that affect the local environment and economy have to be made. Decisions like whether or not to spill water [**?**].

The Columbia Basin is just one small example. Bioinformatics is another data intesive study that generates more data that it can process. Its estimated that less than 10% of the data collected by bioinformatic researchers at the University of Idaho actually makes it through proccessing [**?**]. The rest has to be filtered as best as can be managed, and the low value data trimmed out. The problem with 10% data use, is that not just the fat, but the meat and bone has to be cut away. Either significantly more data must be analyzed, or significantly less should be collected. Or it goes to waste.

1.2 Problem Statement

One of the biggest problems researchers in the Columbia Basin have is getting their data somewhere meaningful. Central databases like PTAGIS offer a central storage and clients to push data there, but they don't offer useful tools for managing data. Once the data is pushed, it is hard to access and manipulate. Researchers are often in remote locations, and have low bandwidth connections.

Many researchers have significant data management tasks before even thinking about pushing data to the cloud. Once they are ready, they need seemless ways to synchronize data to and from the cloud. They also may want to query their data, and filter it into small subsets. Most researchers don't have time to learn new programming language or interfaces. They need a data management tool that has a user interface that is familiar to use cases they already understand. Once they have created a data subset, they will want to share it, save it, copy it, and compare it with other data sets.

1.3 Objectives

The proposed project will tackle the data deluge. The data management tool will be a web-like GUI that can be installed locally on the clients' computers. The tool will contain a expandable meta data server that can be connected to others, for a **decentralized** data storage cloud. It will contain an interface to an existing or custom **networking protocol** that will allow for many different data formats to be synchronized across the cloud. The project will also have graphic interfaces for **seamless data management**.

The decentralized cloud model allows some serious advantages to researchers. They can see exactly what their data looks like, before they submit it. They can filter out bad data before it consumes bandwidth, and can retract undesirable data from the cloud, even after synchronizing it. It also frees them from central storage service management and fees, and **encourages internode and inter-research communication**. The tool can be setup on more available hosts, however, and can be used for the benefits of the centralized cloud model. Each client can decide what model suites them best.

The network protocol will allow incremental synchronization of data from host to host, even in less reliable environments. The tool will create an **outreach** from researchers in high availability areas to those in low availability, low bandwidth areas, and back. The network protocol will have support for major data formats (i.e. SQL, CSV, more), and allow users to send incremental pieces of the data. In the case of very large data sets or low bandwidth, the reciever of the data can still use it. Whether the data takes more time to transfer, or never completes.

The seamless interface will allow users to sort their data needing only basic knowledge of computers. Users will be able their mouse to select datasets and apply filters to them. The tool will allow them to query local or remote data, or to dynamically join both. Queries will create data subsets that users can bring to a workspace. The tool will be able to sort data however the user likes, on the fly. It will then be able to graph the ranges in the subset in any way the user wants to resort them.

Once the user has manipulated their data subset to their satisfaction, they will be able to save it in the meta data server's format, or in a set of major data formats. The tool will also have analysis modules that will allow them to run analysis on the local host. These modules will be extensible to running analysis jobs on other designated compute hosts, like workstations, clusters, or even Amazon's EC2. The tool will come with basic functionality for analysis modules, but will be extensible to the heavier compute options.

2 Project Description

2.1 Agenda

2.1.1 Algorithms / Programming

Figure 1: Some great figure

2.2 Methodology

2.2.1 Project Software

TODO: This project will be written in Ruby on Rails or some other great MVC, will install on Windows, Linux, and Mac, blah blah blah. It will be downloadable to customers via a private github, Apple's App Store, or some other overpriced method. Consulting to set it up won't exist because the whole thing installs itself, and everything is so simple.

3 Customer

4 Statement of Qualifications

5 Conclusion

6 Bibliography

References Cited

[1] Foster, James. *Visualizing Human Microbiome Ecosystems*. University of Idaho: Computer Science Colloquium, December 7th 2010. Seminar.