Stratosphere

A multilevel data-driven social-network for cloud computing

Miller, Joshua
Claxton, Spencer
Mukora, Alice
Griffis, Eric

18 June 2013

Abstract

Conventional social networking tools are fine for some things, but to our knowledge there has not yet been a social networking tool driven by data and research for scientists. OSDC offers an environment with big advantages to big data researchers. In our opinion the OSDC offers a unique opportunity for scientists to form relationships among peers that share research interests centered around the data analyzing power that it provides. In this proposal we discuss *Stratosphere*, a social network extension of the current OSDC console that provides data researchers with new opportunities to identify and connect to peers with similar research interests. Stratosphere is a collection of data-oriented tools that connects cloud researchers with similar data consumption patterns. Our methods for doing this include data access record aggregation and presentation, peer authentication of user generated data, and making users and their research visible to other users. Looking toward the future, post-implementation of Stratosphere, we conclude by proposing that Stratosphere could also facilitate a paradigm for quantifying scientific collaboration in the domain of scientific computing.

Contents

1 Introduction

By adding a social component to the OSDC experience, we aim to provide data researchers with a unique opportunity to identify and connect to peers with similar interests and data consumption habits. Beyond the obvious benefits of conventional social networking tools like Facebook for fostering collaboration, OSDC users also stand benefit from the fact that their interests are encoded in their data usage patterns.

This proposal proceeds as follows. First we motivate the Stratosphere design by clearly defining what we perceive to be the insufficiency of popular social networks for fostering data intensive research collaboration relationships, in light of the unique opportunities offered by the OSDC. We also here highlight what Stratosphere adds to the OSDC in terms of how efficiently data is handled and curated. We then proceed to outline our vision for the initial Stratosphere design. Finally we end by detailing the necessary mechanisms required to implement our vision for Stratosphere. We conclude with a glance toward the possibilities that Stratosphere offers OSDC in the future.

Why Stratosphere

Stratosphere attempts to leverage implicit interconnectivity between scientists using the OSDC system. Currently, researchers are typically connected to each other a priori, by institution or through conferences or prior collaborations. The core concept of scientific cloud computing in an open setting is the promotion of collaboration and multifaceted approaches to data analysis. Users are currently cut off from peer collaboration on research projects conducted via the cloud. By distributing the responsibility of project documentation and making users visible to one another, the OSDC would be promoting direct interactions between researchers from different projects that express interest in the same datasets. We suspect that connections forged within the context of the OSDC and Stratosphere are more likely to lead to successful collaborations because such connections are derived from explicit matching within the interest and problem domains.

We also foresee that by keeping track of the data usage of users, Stratosphere will enable the OSDC administration to be more sensitive to the data-oriented needs of its users. For example, Stratosphere would allow the OSDC to have more insight into choosing what data to keep more or less up to date on the cloud based on what data is more or less relevant to users. Along the same lines, when peer-to-peer authentication is a reality on the OSDC, Stratosphere provides a way for users to monitor who is working with their data and how often.

2 What is Stratosphere

Stratosphere is a data-driven social networking platform that connects researchers in related groups and disciplines. Stratosphere is composed of two layers: the user interaction layer and the data interaction layer. The user interaction layer is a superstructure designed with social networking in mind to create a social interdisciplinary network centered around datasets. The data interaction layer handles user access to the data. Stratosphere provides a peer approval system for file permissions of user submitted data and informs users of who is using which data and with what intent.

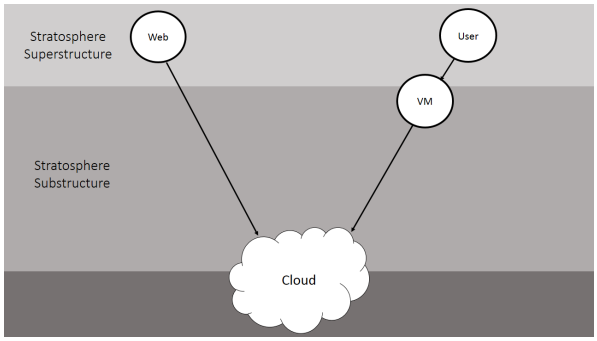

Figure 1: Stratosphere structure

3 How Stratosphere Works

A Data-driven Network

The main infrastructure of Stratosphere is *data-driven*. This mean that all interaction between users and the rest of the network is mediated

through the data. One way we propose to accomplish this is by logging user-data interaction in a metadata descriptor attached to each data set. For example, when a user accesses a particular dataset, rate counters specific to each individual user could be incremented. Using these rates, we could present lists of the most active users for each dataset to other users who express interest in the same datasets. Furthermore, data frequently used by an individual would be represented as specializations on that users profile on Stratosphere front end. A metadata entry might include a username, the most recent access time, and an access tally. Other relevant user-supplied information like project descriptions and contact information would be stored similarly.

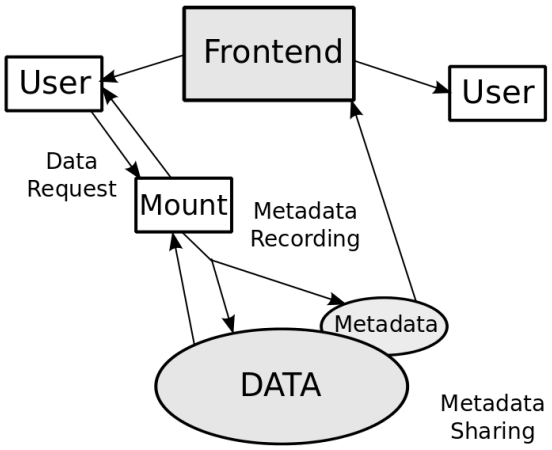

Figure 2: Process of recording user access metadata

Peer Authentication

This social-interactive system would also provide a framework for the moderation of user permissions. When users add data to the cloud they are able to assign ownership to the data set, with the default being public ownership. Publicly owned data sets require no permissions to access. Should the creator of a data set request ownership, then he,she, or selected other users take the responsibility of being an administrative user. Users are then granted access to the data set by the administrative user.

User Visibility and User-to-User Interaction

The main goal of Stratosphere is to provide transparency through the current OSDC system. To do this we propose that users be given public profiles that display the data they are working with and their contact information. However, we are well aware that some actions need not be recorded in a user's data access. For this reason, we propose the use of defaults which lean towards full record keeping and visibility, with the option to remove such visibility. The primary way in which users will access these public profiles will be, again, data-centered. That is, users will have access to other users by seeing their the data-access records in the metadata. Thus, the data becomes the primary link between researchers. It is this ability of the metadata to describe the data in terms of users access that makes the Stratosphere a truly data-driven social network. In this way, users are able to see directly what peers are doing with the data they themselves are accessing. An implementation of user profiles would simply entail an addition to the current OSDC web front end that displays user info pulled from a database containing user activity statistics.

User-to-Data Interactions

The user-to-data interaction is the most challenging aspect of this project. Required tools include implementations for

- file monitoring for user access records
- peer authenticated permissions

File Monitoring

Possibilities for monitoring the frequency of data access include tools such as *inotify*, or direct patching of the file mounting, or other proprietary/open software. Preliminary testing of accepted tools such as inotify seem improbable to use. Starting the service ran at approximately 16 GB/s for an 80GB section of the open data. This appears to be non-scalable (70 hours to index the entire cloud). Therefore, we propose a patch to the mounting system such that each new instance of a file access increments the metadata use counter. The main

assumption is that the users who access the data the most have the highest interest and experience with the particular data set.

Metadata Creation

We propose to include metadata based on who accesses the data, how often they access the data, and for what purpose they are accessing the data. In order to maintain privacy, the user is not required, but encouraged to assign a purpose comment or project name to each data set accessed. This data set descriptor persists for each user until changed. When the data is accessed, it is the job of the file mounting system to intercept and record the request. Our initial plan for a prototype implementation of such a tool was to mount a FUSE driven file system to a WebDAV share on a web server. File access request would then be intercepted and logged by the server. The user information contained in the server log would then be dropped into the metadata record. The main issue with this implementation is its scalability to an OSDC system with many users accessing a large number of very large datasets.

Permissions

The task of creating peer authenticated permissions is a far more challenging problem. Currently, the OSDC cloud operates on basic Unix group permissions, and due to the binary grouping of open data and protected data, this works for now. However, we would be forced to store the entire combination of each user inclusion in each research permission group. A file system with *Dropbox*-style permissions is certainly preferable. However, we currently have no recommendation for how such a file system would be implemented so that it is big-data scalable.

4 Conclusion and the Future

We believe that, in light of the unusual opportunities made possible by the OSDC infrastructure, conventional social networks are poorly positioned to target opportunities for data intensive research collaboration between scientists. To tap this new resource we proposed Stratosphere, a social networking platform that we believe rings true to the spirit of the OSDC's scientific cloud computing effort. We envision that with Stratosphere, the OSDC will become a microcosm for quantifiable collaboration between scientists using computational methods over big-data. That is, with Stratosphere, we will be able to gather data on how scientists interact with each other through the data that they are analyzing. We could thereby quantify in some way the collaboration that takes place between scientists, e.g. looking at the amount of analysis performed on a dataset before and after implementation of Stratosphere on the OSDC.