Stratosphere

A multilevel data-based forum for cloud computing

Miller, Joshua
Claxton, Spencer
Mukora, Alice
Fleming, Zac

18 June 2013

Abstract

Stratosphere is a data-based collection of tools for connecting cloud researchers amongst disciplines. The method for doing this includes data access records, peer authentication of user generated data, and making users and their research visible to other users.

We are attempting to sell an idea here. We have no proof of concept save the precedent of other social networks and the benefits they offer.

Contents

1 What is Stratosphere

Stratosphere is a data-based social-network-like system with the goal of connecting researchers in different 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 the data. The data interaction layer handles user access to data. Stratosphere provides a peer approval system for file permissions of user submitted data and informs the users of who is using the data and with what intent.

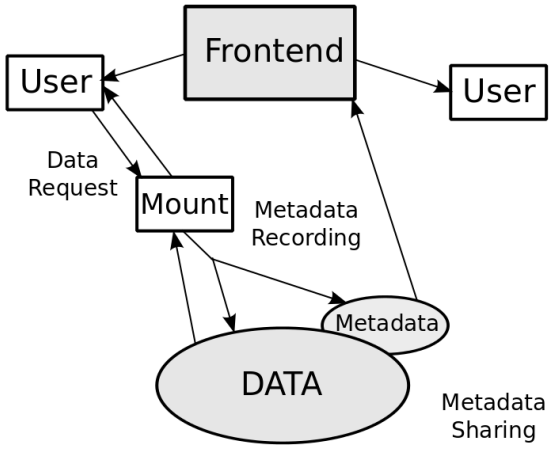

Figure 1: Process of recording user access metadata

Why Stratosphere

Stratosphere attempts to address the problem of inter-connectivity between scientists using the OSDC system. Currently, researchers are only connected to each other a priori, by institution or by prior collaboration. 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 the users visible to one another, the OSDC would promote direct interaction with researchers from different projects who desire to work on the same data set.

A Data-based network

The main infrastructure of Stratosphere is data based. This means that user access is documented in a metadata descriptor attached to each data set. For example, user access to the dataset might tally a rate counter specific to that individual user. Using this rate, we would be able to present a list of top users for each dataset to other users interested in the data. Furthermore, data frequently used by an individual would be represented as specializations on the Stratosphere front end.

Peer-Authentication

This social-interactive system aids 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.

Visibility and Privacy

The main goal of Stratosphere is to provide transparency through the cloud. 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.

2 Implementation Challenges

User-to-Data Interactions

The user to data interaction is the most challenging aspect of this project. The tools that need to be implemented for this include the implementation of

- 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 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. We have provided a proof of concept implementation (using XML metadata) of a file system mount adapted from **fusedev** for this purpose.

The metadata includes username, last access, number of accesses, project, project description, and contact.

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.

User-to-User Interaction

The main aspect of interactions between users in the Stratosphere is the ability of the metadata to describe itself in terms of users. Users are able to see what peers are doing with the data of interest and are further able to get to a full research profile of the users that are working with the same data. To make this possible we propose the implementation of user profiles that are accessible from and linked to the data that the user has accessed. The biggest challenge here would be integrating such a social network with the current OSDC interface (the Tukey portal).