2013 IEEE 16th International Conference on Computational Science and Engineering

SAFAL: A MapReduce Spatio-temporal Analyzer for UNAVCO FTP Logs

Kathleen Hodgkinson
Plate Boundary Observatory
UNAVCO
Boulder, CO 80301, USA
Email: hodgkin@unavco.org

Abdelmounaam Rezgui
Dept. of Computer Science & Engineering
New Mexico Tech
Socorro, NM 87801, USA
Email: rezgui@cs.nmt.edu

Abstract— **UNAVCO is a National Science Foundation (NSF) funded consortium that facilitates geoscience research and education using geodesy. It is responsible for the collection, archiving and distribution of data from GPS sites installed in every continent of the world. In addition to GPS data, UNAVCO collects borehole seismic, strainmeter, meteorological, and digital imagery data. One of UNAVCO's largest programs is the Plate Boundary Observatory (PBO), the geodetic component of the NSF funded Earthscope program. PBO consists of over 1100 continuous GPS sites plus 80 borehole strain and seismic sites. In this paper, we present SAFAL, a Spatio-temporal Analyzer of FTP Access Logs collected by UNAVCO's data center. We developed SAFAL using Hadoop/MapReduce. The motivation for this work was to build an efficient system able to quickly identify trends in GPS data usage. The system is able to processes millions of lines of data in minutes. It supports queries such as: (i) what is the most downloaded GPS site, (ii) who is downloading the data most, or (iii) what periods of data are of greatest interest. Answers to these and similar queries are useful for planning network growth, allocating Web resources, and tracking hot topics in geoscience research. They also may be extremely useful to help UNAVCO illuminate dark data.**

Keywords-component; MapReduce; Hadoop; Web usage mining; GPS sites; FTP access logs.

I. INTRODUCTION

UNAVCO (University Navstar Consortium) [1] is a National Science Foundation (NSF) funded consortium that facilitates geoscience research and education using geodesy. It is responsible for the collection, archiving and distribution of data from GPS sites installed in every continent of the world. In addition to GPS data, UNAVCO collects borehole seismic, strainmeter, meteorological, and digital imagery data. One of UNAVCO's largest programs is the Plate Boundary Observatory (PBO), the geodetic component of the NSF funded Earthscope program. PBO consists of over 1100 continuous GPS sites plus 80 borehole strain and seismic sites (Figure 1). UNAVCO makes data obtained from these GPS sites available to users worldwide for access and download. An important objective is to quickly identify access trends to these data, i.e., be able to efficiently answer queries such as: (i) what is the most downloaded GPS site, (ii) who is downloading the data most, or (iii) what periods of data are of greatest interest. Answers to these and similar queries are useful for planning network growth, allocating Web resources, and tracking hot topics in geoscience research. They also may be extremely useful to help UNAVCO illuminate dark data, i.e., make scientists aware of useful data that they did not know exist before. These queries must be evaluated over data that is extremely large and continuously growing.

In this paper, we focus on analyzing a particular type of data, namely, FTP access logs. We present SAFAL, a **S**patio-temporal **A**nalyzer of **F**TP **A**ccess **L**ogs collected by UNAVCO's data center. SAFAL is a system for Web usage mining developed using Hadoop/MapReduce. As it is designed with focus on scalability, it is able to analyze millions of lines of spatio-temporal data in just a few minutes.

Analyzing FTP access logs is a type of Web usage mining which is the process of determining trends from access logs, e.g., which Web pages in a Web site are accessed most, when most for the accesses are made, the durations of these accesses, etc. Web usage mining is one of three branches of Web mining along with Web structure mining and Web content mining [4,5]. Analyzing FTP access logs reveals useful information about data users, the remote hosts, the file size, and the name of files downloaded. Patterns could then be identified and analyzed.

This paper is organized as follows. We first present the process of data collection and storage at UNAVCO. Section III explains the format of UNAVCO's FTP access logs. In Section IV, we outline UNAVCO FTP directory structure and available data sets. In Section V, we give an overview of the system's architecture and describe how MapReduce is used to process the FTP log files. In Section VI, we present a set of experiments that we ran using SAFAL. Section VII is the paper's conclusion and Section VIII discusses future work.

II. DATA COLLECTION

As GPS data reach the UNAVCO Data Center, they are stored in an anonymous FTP site for users to download. Data

[1] http://www.unavco.org

978-0-7695-5096-1/13 $31.00 © 2013 IEEE
DOI 10.1109/CSE.2013.157

are provided in raw and processed formats. Raw data are typically stored in a year, day of year and site directory structure. Processed data are stored as one file per site. By parsing the file path in the FTP log and inspecting the file name, one can tell what type of data the user is downloading and from which GPS site. An idea may then be formed of the most popular sites across the network and which data types (raw or processed) are being accessed. The logs could also give information as to whether data downloads tend to be routine or event driven, e.g., there is a significant increase in downloads after an earthquake or volcanic eruption.

Figure 1. UNAVCO GPS Sites

Combining the data download information with the knowledge of the site locations will allow UNAVCO to determine if there are regional hotspots, e.g., whether the data from sites installed along a particular file zone are being heavily accessed. This information is useful for planning future network growth. Bandwidth could be increased for highly used sites or the network densified in a region if scientists are keen to get measurements from that area. It could help in operations and maintenance planning as highly used sites could be prioritized for site maintenance. Conversely, if certain sites are rarely being used, this might trigger an investigation of whether the data latency from those sites is too low or the errors associated with the measurements from those sites too large to make the data useful. Each of these examples would warrant attention from an engineer or an examination of the data quality by an analyst. Over the longer term it would be useful to track the growth in interest in certain sites or regions, as this may be an indication of the topics in which the earth science community is interested.

III. UNAVCO FTP ACCESS LOGS

UNAVCO uses the xferlog format to store FTP log files. An xferlog file contains the transfer logging information from the UNAVCO FTP data server. A typical UNAVCO xferlog collected over 24 hours consists of about 180,000 lines and is about 21 MB. Each line has 18 fields and has the following format:

```
DDD MMM dd hh:mm:ss YYYY transfer_time
remote_host file_size filename
transfer_type special_action_flag
direction access_mode username
service_name authentication_method
authenticated_user_id completion_status
```

Where:

- `DDD MMM dd hh:mm:ss YYYY` represents the date and time the record was made
- `transfer_time` is the transfer time in seconds
- `remote_host` is the name of the remote host
- `file_size` is the size of the transferred file in bytes
- `filename` is the name of the file downloaded
- `transfer_type` is a character indicating the type of transfer (a for ASCII, b for binary)
- `special_action_flag` indicates if any special action was taken (C = compression, U = uncompressed, T = file was tarred, '-' means no action was taken)
- `direction` represents the direction of transfer (o=outgoing, i=incoming, and d=deleted)
- `access_mode` represents the method by which the user is logged in (a=anonymous, g=passworded guest, r=a local authenticated user)
- `username` is the local username
- `service_name` is the name of the service being invoked
- `authentication_method` is the method of authentication used (0=none, 1=RFC931 authentication)
- `authenticated_user_id` is the user id returned by the authentication method (* means an authenticated user id was not available)
- `completion_status` is a single character indicating the status of the transfer (c=complete, i=incomplete).

For example the following line:

```
Mon Jan 07 00:00:08 2013 1 192.52.65.100
391771
/ap/ftp/pub/rinex/obs/2012/185/kosm1850.
12d.Z b - o r ftp ftp 1 * c
```

means that on Monday January 7th, 2013 at 00:00:08 a researcher working from remote host 192.52.65.100 downloaded the binary file kosm1850.12d.Z via FTP. The file was 391771 bytes and took one second to download. No

special action was taken during download, the direction was outgoing and the user was logged in as a local authenticated user with a local username "ftp". The file was downloaded via FTP, RFC931 authentication was used but no authenticated user id was available. The file download was complete.

The filename and file path in the FTP log is rich in information. By parsing it, the GPS site can be determined along with the type of data (e.g., raw, processed), the sample rate, the time period, the measurements span, etc. Once the GPS site name is known, the coordinates of the site can be found from other sources and a map view animation of downloads can be created. Examination of the time window may give an indication of the event being studied. For example, it may span an earthquake, volcanic eruption or other geophysical event. The time lag between the event occurring and the data being downloaded could also give some insight as to how rapidly data users respond after an event.

IV. UNAVCO FTP Directory Structure

UNAVCO GPS data can be accessed by the scientific community via the `ftp://data-out.unavco.org/pub` directory. Each GPS site (Figure 1) is uniquely identified by a 4 character alphanumeric code. The data are available in various formats and sample rates. The type of data accessed can be determined from the path name in the FTP access log. Data are available in the GPS receiver's native format (here termed "raw"), in Receiver Independent Exchange (RINEX), and in BINary EXchange (BINEX) formats which are the translation of the native formats into standardized exchangeable formats. These data are usually made available in UTC-day or hour-long files and are sorted by year, day of year and site name. Processed data, here termed products, are available for each site usually as one file per site that grows with time or network wide solutions. Data users familiar with the FTP server structure can download data directly from the FTP site. Others may choose to identify the data they require via an interactive Web page. In this case, the data are placed in a special area on the FTP server.

A full description of the FTP directory structure is given at: `http://facility.unavco.org/data/ftp.html`. Tables 1 through 4 provide an overview of the FTP directory structure where *yyyy* refers to the year, *YY* the last two digits of the year, *ddd* to the day of year, and *site* to the GPS 4-charcter code. Data with sample intervals of 1s or less are considered high rate GPS and held in the high rate directory and sorted according to sample frequency (F-Hz). Non-high rate GPS usually has a sample interval of 15 seconds. The near real time (nrt) data are held in the nrt directory.

TABLE I. Low Frequency (Raw, Rinex and Binex) Data

Directory path	File name	Description
/L1/rinex/yyyy/ddd	siteddd0.YY[don].Z	Compressed RINEX files, hatakana obs (d), obs (o), navigation(n) file
	siteddd0.S	teqc qc report file
/raw/yyyy/ddd	siteyyyyMMDDHHMMa.tnn	Raw file, sample interval > 1 s, nn refers to session number.
/rinex/obs/yyyy/ddd	siteyyyy0.YY.[do].Z	Observation files
/rinex/met/yyyy/ddd	siteyyyy0.YY.m.Z	Meteorological (met) files
/rinex/nav/yyyy/ddd	siteyyyy0.YY.n.Z	Navigation files
/rinex/qc/yyyy/ddd		QC files

TABLE II. High Rate (Raw, Rinex and Binex) Data

Directory path	Filename	Description
/highrate/F-Hz/binex/yyyy/ddd/site	siteyyyyMMDD.bnx	BINEX file, UTC day-long files
/highrate/F-Hz/raw/yyyy/ddd/site	siteyyyyMMDDHH00b.t00	Raw file, hour-long files
/highrate/F-Hz/rinex/yyyy/ddd/site	sitedddh.yyX.Z	Hour-long rinex files, h refers to the hour of the day and is labeled a through x. The X character refers to the type of data in the file and can be (d, n, m, or o)
/hourly/rinex/yyyy/ddd/site	sitedddh.yyX.Z	(as above)
/hourly/binex/yyyy/ddd/site	siteyyyyMMDDHH00b.t00	Raw BINEX hour-long files
/nrt/rinex/yyyy/ddd/site	sitedddh.yyX.Z	(as above)

TABLE III. PROCESSED DATA

Directory path	Filename	Description
/products/position/site	Site.inst.type.pos Site.inst.type.pos	*inst* refers to the group that prepared the data, *type* refers to the type of solution. The files are available in position (pos) format or csv format. Files appended to each day.
/products/position	inst.type.pos.tar.gz	tar file containing positions for entire network
/products/sinex/WWWW	instWWWWd.type.snx	Network wide solution, stored under GPS week (WWWW) in sinex format
/products/velocity/	inst.type.vel	Network-wide velocity files
/products/events	Pbo_*description*.evt Pbo_*description*.ps	Data sets created for specific events. *Description* refers to the event.
/products/troposphere/yyyy/ddd/	instWWWWd.yyyyMMDD.a.met.gz	Tropospheric data.

TABLE IV. METADATA AND PREPARED DATA SETS

Directory path	Filename	Description
/logs	sitelog.txt	ASCII metadata file
/pickup	varies	Miscellaneous data
/dai	varies	Files selected through the DAI interface are placed here

V. ARCHITECTURE

Figure 2 shows an overview of SAFAL's architecture and its interactions with the UNAVCO's FTP server and with MapReduce/Hadoop. Apache Hadoop is an open source Java based framework that enables distributed processing of large data sets across multiple clusters of machines [2]. Hadoop is designed for scalable, distributed, data-intensive computing and can handle petabytes of data across thousands of machines (nodes). Several large companies have used Hadoop in data-intensive applications. Examples include Google, Facebook, eBay, and Twitter. It is the ideal tool for processing large data sets in virtualized data centers such as Amazon's Elastic Compute Cloud (EC2). The basic idea in Hadoop is to split massive data sets into many small data sets that may be processed in parallel on many nodes in one or several clusters. Hadoop consists of three subprojects: a common utilities package, the Hadoop Distributed File System (HDFS), and MapReduce which is a programming model developed initially by Google [1] to process and analyze many petabytes of data quickly.

The underlying premise of MapReduce is that the task can be cast as a key-value problem and processing is divided into two functions: mapping and reducing. The user defines a mapper function where the input information is analyzed and a key value pair created. This feeds into a reducer function, (also defined by the user) that merges the output from the mapper function. When the MapReduce function is called from the user's main program the process can be summarized in a series of steps [1]:

A master node splits the input data set into several pieces and distributes them to worker nodes. Each worker node, called a mapper node, processes the small data set. It parses the data and passes the information to the user's mapper function to create key-value pairs. When the task is complete, the results are written to local memory and the mapper node lets the master know the location of the results. The master node then alerts another worker node, a reducer, to the location of the results. The reducer node sorts the data and passes each unique key and its associated values to the user's reducer function. The results from the reducers are merged. When all the mapping and reducing tasks are completed, the MapReduce call returns to the main code.

The master node distributes the work across the cluster, monitors progress and takes action if any node fails. The mapping and reducing nodes work independently of each other which means the work can be done in parallel and be distributed across multiple nodes. Duplication of work across nodes provides a safeguard against node failure.

In this paper, the objective is to analyze information in the UNAVCO FTP log, the information that is being sought by the analyst is used to build the key and the number of times that key occurs in the log file becomes the output value. For example, if the analyst wants to know which GPS site has had the most downloads the key becomes the site name. The mappers parse lines in the log file and extract the site name as the key. That key and the value of 1 get sent to the reducers which count how many times that key is seen. The final output is a list of GPS site names and the number

of times each name appears in the log file. Using MapReduce allows a quick analysis of millions of lines of data.

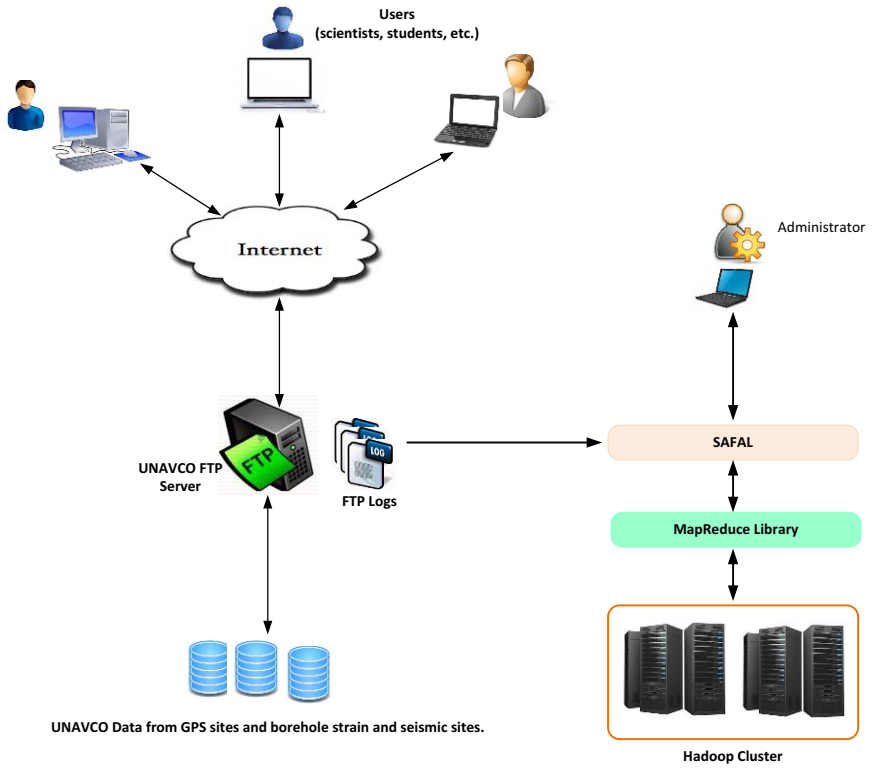

Figure 2. Architecture of SAFAL and its interactions with the FTP server and MapReduce/Hadoop.

VI. EXPERIMENTS

The input data that we considered in this paper were from the FTP access log file recorded by UNAVCO from January 1st to January 30th 2013. The log file contained about 5 million lines and was about 0.5 GB in size.

In this section, we investigate different aspects of data usage: the IPs that most visit the FTP site, the most popular GPS sites, the time taken for downloads, variations in the number of downloads over a month, and dates the downloaded data were recorded. For each aspect of usage, a key is built from the information in the log line based on the information sought.

The version of Hadoop used in the implementation was hadoop-0.20.2-cdh3u6 (03-Apr-2013). The mappers and reducers were written in Perl.

A. IPs Accessing the UNAVCO FTP site

To find out which IPs most accesses the network, the IP number becomes the key passed from the mapper to the reducer. Since GPS sites for which data are recorded in hour-long files must be accessed 24 times more than a site that has one day-long file, the key also contained information that reflected if the IP was downloading an hour-long file. We down weighted files recorded in hourly chunks by a factor of 24. Through the month of January 2013 the FTP area was accessed by 1338 unique IPs. The number of downloads from specific IPs drops off exponentially on a semi log plot (Figure 3). The IPs with the largest number of downloads were from Menlo Park, CA, Westford, MA and Irvine, CA. Using Hadoop, the 5 million line file was processed in 2.2 minutes.

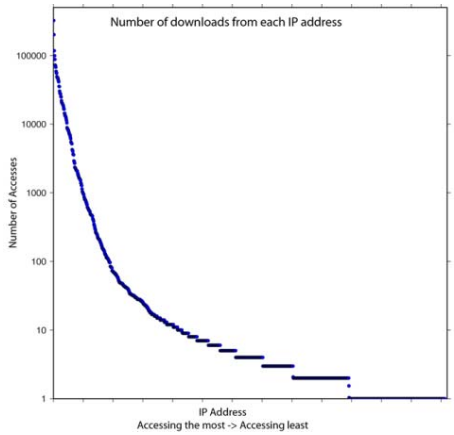

Figure 3. Number of downloads from each IP address.

B. *Most Downloaded GPS Sites*

To determine the most downloaded sites, we looked at the number of times each GPS site had files downloaded from it. We limited this to RINEX observation data, processed position and log (metadata) files. Again, if the file accessed was an hour-long file, the count was down weighted by a factor of 24. The key in this case was the GPS site name with the type of file appended, e.g., P403_position or P403_log etc. Data were downloaded from 2873 GPS sites in January 2013.

Figure 4 shows the results ordered from sites with the most to sites with the least downloads. The sites with the most processed data downloads were a cluster of sites in Yellowstone. About 100 log files were downloaded across the network each day. The most downloaded RINEX files were from sites in the Papua New Guinea (sa42), the Caribbean (airs) and a site at Stanford University, CA (SLAC).

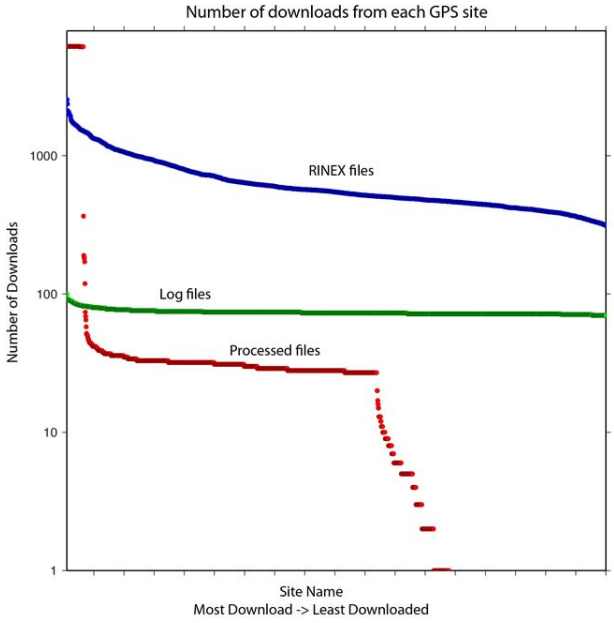

Figure 4. Number of downloads from each GPS site

C. *Download Time*

To analyze the time taken for downloads, the time in seconds in the log file becomes the key. The hour files were separated out and the time counted separately. 99.8% of the hour-long files and 83% of all other type of files were downloaded in 1 second or less (Figure 5).

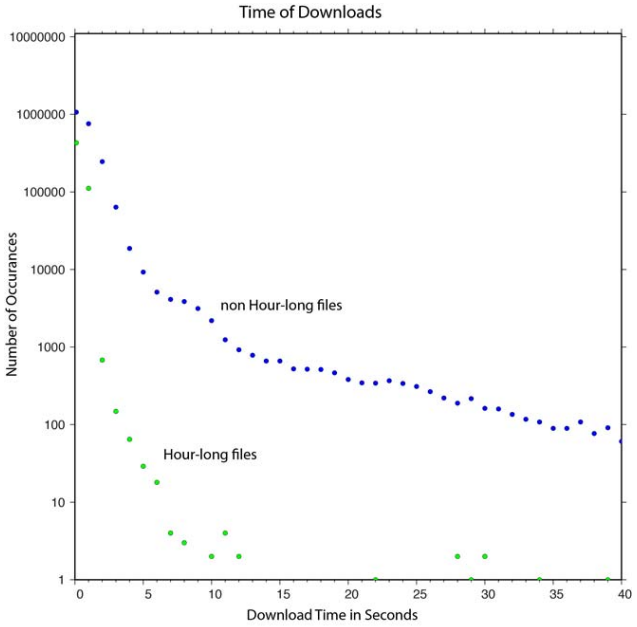

Figure 5. Download time

D. *Variation In Number Of Downloads Through The Month Of January 2013*

Here, we examined the number of downloads each day. The files were divided into RINEX observation files, processed position files and log files. The key in this case was the day of the month with the file type appended, e.g., Jan_01_log or Jan_01_position. There is a strong weekly periodicity in the downloads of RINEX files with maximum downloads on Thursdays and minimum downloads on Sundays (Figure 6). The lowest number of downloads was on January 1st, a public holiday. While the number of processed position files remains fairly constant through the month, there is some structure to the log file downloads after January 12th when the trend resembles that of the RINEX downloads.

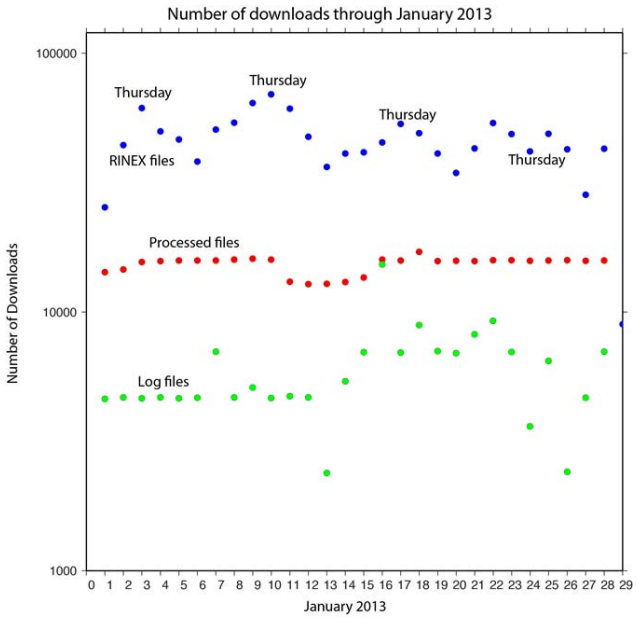

Figure 6. Variation in the number of downloads

E. *Dates Data Were Recorded*

The pathname of the RINEX observation files contains the day on which the data were recorded. To examine the dates users were downloading data from, the year and day number were used as the key, e.g., 2009_033 refers to day 33 of 2009. Hour-long files were down weighted by a factor of 24. The number of downloads for data increases exponentially on a semi-log y-axis for data recorded after January 1 2008 (Figure 7). There was interest in data collected pre-2008 with ~10 downloads from each day between 2000 and 2008. Data recorded in 2003 at GPS site Thule in Greenland was downloaded several times. The earliest data download was from January 1 1992.

F. *Discussion*

The analysis of FTP logs files performed using SAFAL and presented in this section revealed valuable information about the users' research interests. It is possible to determine the IPs that are accessing the site most. These heavy data users could be targeted as researchers from which to gain guidance on network growth and future proposals. Knowledge of which GPS sites are most accessed allows UNAVCO to determine which sites are of most interest to the scientific community. It also highlights which sites are being underutilized. Analysis of the download times gives information on the ease with which researchers are downloading the data they are interested in. Further investigation of the times at which download times are greater than a few seconds could help to trouble shoot problems with the FTP system. Analysis of patterns in data downloading shows that researchers seem to favor Thursdays for downloads. This type of knowledge could help in allocating web resources. In analyzing the dates of data downloaded, we find that the greatest number of downloads are of data collected from 2008 onwards. However, although the numbers are small compared to current data, researchers are downloading data collected before 2008. This indicates that it is worthwhile to keep historical data online.

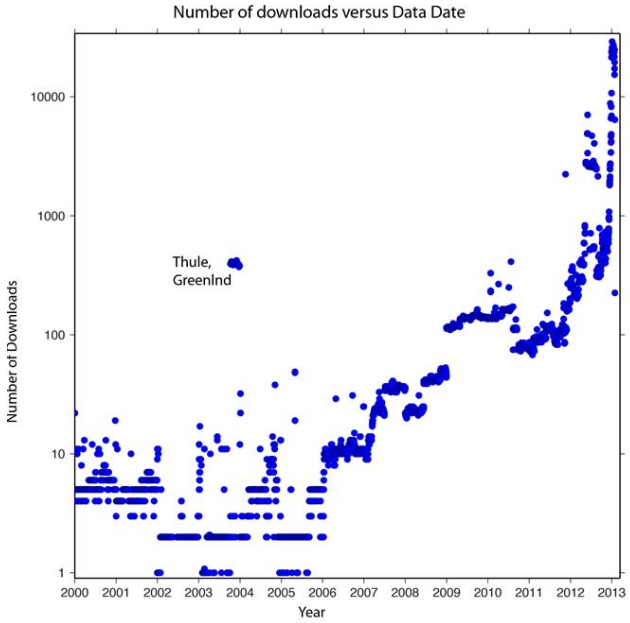

Figure 7. Number of downloads vs. data date

VII. CONCLUSION

Analyzing UNAVCO's FTP log access files could reveal important patterns that are very useful for planning network growth, allocating Web resources, and tracking hot topics in geoscience research. These logs contain massive amounts of data and are continuously growing. As a result, traditional ways of analyzing FTP logs do not scale. To address this issue, we developed SAFAL, a Spatio-temporal Analyzer of FTP Access Logs collected by UNAVCO's data center. The system was developed using MapReduce/Hadoop. We conducted several experiments to evaluate SAFAL. The system was able to analyze millions of lines of FTP access logs very efficiently. We find that MapReduce/Hadoop is an excellent tool for analyzing FTP logs efficiently. The speed with which the files can be analyzed suggests that it could be used to create near-real time maps and analysis of researchers' data downloads.

VIII. FUTURE WORK

We plan to extend SAFAL with a recommendation capability whereby the system suggests relevant data products to scientists based on the analysis of existing FTP access logs. Similar work has been done to investigate whether shoppers interests as captured in web search data could be exploited to generate commercial recommendations for customers. For example, in [3], Jagabathula et al. present a methodology to forecast future commercial intents of shoppers by building relationships between search queries as recorded in web logs. They could then suggest to shoppers

products they were not originally looking for. This type of analysis could be combined with the fast FTP log analysis outlined in this paper to suggest to researchers data products they did not know existed but might be useful for their research. By analyzing the download habits of groups of IPs (researchers), creating connections between groups of sites that are commonly downloaded, identifying geophysical events and their associated data sets, recommendations could be made to a researcher on data sets they may not have known about but that might be useful to them. For example, someone who has downloaded a lot of GPS data from Greenland may be interested to know we have added new GPS sites to that network or that another scientist has installed high rate meteorological instruments across the region. Someone analyzing data along the San Jacinto fault zone may not realize UNAVCO has several pore pressure sensors in the region that might be useful to them. Combining the rapid analysis of FTP logs enabled by SAFAL with advanced recommendation techniques (such as those presented in [3] could be a step towards shedding light on dark data.

REFERENCES

[1] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. Sixth Symposium on Operating System Design and Implementation, OSDI, San Francisco, CA, December, 2004, pp. 137-150.

[2] Apach Hadoop, `http://hadoop.apache.org`.

[3] Srikanth Jagabathula, Nina Mishra, and Sreenivas Gollapudi, Shopping for Products You Don't Know You Need. In Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM), ACM, New York, NY, USA, 2011, pp. 705-714.

[4] Florent Masseglia, Pascal Poncelet, and Maguelonne Teisseire, Web Usage Mining: How to Efficiently Manage New Transactions and New Clients, PKDD, 2000, pp. 530-535.

[5] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations Newsletter, 2000, 1(2):12-23.