Towards Use And Reuse Driven Big Data Management

Zhiwu Xie[1], Yinlin Chen[1], Julie Speer[1], Tyler Walters[1], Pablo A Tarazaga[2], and Mary Kasarda[2]
[1]University Libraries and [2]Department of Mechanical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, USA
{zhiwuxie, ylchen, jspeer, tyler.walters, ptarazag, maryk}@vt.edu

ABSTRACT
We propose a use and reuse driven big data management approach that fuses the data repository and data processing capabilities in a co-located, public cloud. It answers to the urgent data management needs from the growing number of researchers who don't fit in the big science/small science dichotomy. This approach will allow researchers to more easily use, manage, and collaborate around big data sets, as well as give librarians the opportunity to work alongside the researchers to preserve and curate data while it is still fresh and being actively used. This also provides the technological foundation to foster a sharing culture more aligned with the open source software development paradigm than the lone-wolf, gift-exchanging small science sharing or the top-down, highly structured big science sharing. To materialize this vision, we provide a system architecture consisting of a scalable digital repository system coupled with the co-located cloud storage and cloud computing, as well as a job scheduler and a deployment management system. Motivated by Virginia Tech's Goodwin Hall instrumentation project, we implemented and evaluated a prototype. The results show not only sufficient capacities for this particular case, but also near perfect linear storage and data processing scalabilities under moderately high workload.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital libraries – *collection, dissemination, systems issues*.

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Distributed systems, performance evaluation*.

Keywords
Big data; digital library; cloud computing; digital repository; smart infrastructure; sensor data

1. INTRODUCTION
What can the digital libraries community contribute to tame the data deluge? In terms of the conceptual framework, infrastructure, and implementation, the answers vary from the optimistic "just read and implement the OAIS specification" [6] [13], the less encouraging "can't do" at the institutional level [23] because it "takes big organization" [26], to the cautious "knowledge infrastructures are not yet in place" [7]. We resonate more with the cautious note, especially its assessment that focusing on archiving inactive data "limits the application of digital libraries for scientific data management" [7].

Indeed, the OAIS Reference Model [17] may form a tunnel vision for data repositories and reduce them to a niche far less relevant to researchers than self-serving librarians and archivists. No matter how thorough we document and how detailed we describe the data and their usage context [34] [24], mummifying live data out of their natural habitats of analysis to be preserved in an isolated vault can significantly diminish their value. This is particularly evident in the big data management scenario, where making sense of data requires extensive technology maneuvers and infrastructure support. If an archival repository does not include appropriate capabilities to perform analytics tasks and directly answer data-intensive science questions, the researchers will find the repository impractical, giving data producers less incentive to hand over fresh and hot data therefore making the repository even less useful. To break away from this paradox, we need to adopt a use and reuse driven approach, in which the digital library is a pluggable component of the research infrastructure. It then follows that the data preservation and curation will result as a by-product of the research process, not its ultimate goal or end product.

In this paper we focus on the high volume and high velocity aspects of big data, which pose different challenges from those caused by high variety and high veracity. As of this writing, typical institutional repository implicitly or explicitly limits the unit data deposit to an arbitrarily low size, e.g., 10GB [23] or 20GB [18]. This certainly is far below the volume now churned out from the laboratories and observatories. The "big" aspect should therefore be gauged against the data capacity currently handled by a typical digital library.

It is easy to dismiss the capacity concern as purely technical therefore trivial. After all, given sufficient funding, any existing repository can build up its storage capacity and raise the bar. But it will soon become clear that simply expanding the storage capacity does not solve the problem. The use and reuse pattern starts to change when the data volume reaches a threshold. Beyond that, researchers cannot easily move the data from the

storage to a remote analytics environment therefore need sufficient computing capacity close to where the data are stored. If provisioned with new technologies such as virtualization, big data analytics, and cloud computing, this new usage pattern will then open up opportunities for organizational, cultural, and social changes surrounding big data sharing and reuse. Unfortunately, existing archive-centric repositories are not well prepared for these changes.

Prior work [15], [6], [7], [13] tend to associate big data with big science, which is also characterized with big organization and big budget. We intentionally avoid the big/small science dichotomy for two reasons. First, the data volume separating the big and small science is shifting up very fast that many big science data management challenges are now also confronting small research teams. For example, the much-revered 1000 Genomes project [43] produced 200 TB of data from 2008 to 2012. The Sloan Digital Sky Survey [46] produced about 130TB of raw and derivative data over 8 years in phase I and II [38]. In contrast, the sensors installed in Virginia Tech's Goodwin Hall alone can collect as much data in shorter period of time, and we expect the data acquisition lasts much longer. While motivating this research, the Goodwin Hall project was started by two faculty members, a small lab, and until now is mostly internally funded.

Second, managing big data does not have to begin with a big organization and a big budget. Ramping up big data management with limited budget and in small team settings can transcend both big and small science. Organization structure wise, starting from a bottom-up approach will be more appropriate, where user communities naturally form and self-organize around the data out of their own needs, use cases, and perspectives. It is also important to build systems that allow diverse communities to contribute human, technology, and financial resources in a self-motivated, ad hoc, and on-demand manner. The approach proposed in this paper will foster collaborations significantly different from both the top-down, tightly managed big science model and the lone-wolf, gift-exchanging small science model [44].

Motivated by the Goodwin Hall project, we propose a big data management approach driven primarily by data use and reuse. We then build an implementation prototype in the cloud and evaluate its performance. The results show perfectly linear scalability under moderately high workload, indicating that the public cloud can be a viable big data sharing and management platform.

This paper is structured as follows. After describing the Goodwin Hall project, we analyze its requirements and show why OAIS reference model is insufficient for its data management needs. We then make the case for building a use and reuse driven big data management infrastructure in the public cloud and describe its system model. The next few sections address its implementation prototype and evaluation, discuss the "soft" impacts of this approach, and conclude the paper with reviewing and comparing related work.

2. BACKGROUND

The main motivation of this research is the need for a big data management system primarily intended for, but not limited to, the Virginia Tech Goodwin Hall sensor data.

Virginia Tech's Smart Infrastructure Laboratory (VT-SIL) is building a full-scale living laboratory in the newly opened 160,000-square-foot Goodwin Hall. VT-SIL is finishing mounting over 240 vibration-monitoring accelerometers in more than 130 strategic locations throughout the building, as well as planned for hundreds of temperature, flow, and other sensors to be installed in the near future. Upon completion, Goodwin Hall will be the world's most instrumented building for vibrations and will generate more than 60TB of sensor data per year.

The Goodwin Hall instrumentation differs from similar projects [41] [22] in that from the very beginning it was designed as a multi-purpose living laboratory instead of just for seismology and structural monitoring. Higher density of sensor mounts were directly wielded to the structural beams during the building construction instead of that as an afterthought, and the multi-dimensional accelerometers are strategically positioned and are sufficiently sensitive to detect human movements in the building [14] [39]. This opens up opportunities for multi- and cross-disciplinary exploration and discovery.

VT-SIL will utilize the collected data to improve the design, monitoring, and daily operation of civil and mechanical infrastructure as well as to investigate how humans interact with the built environment. In collaboration with Virginia Tech Libraries, VT-SIL also intends to open up much of the data to the public through both live streaming and a data repository, the latter being the main motivation of this research. The objective is to encourage exploratory researches and foster an open and inclusive community of researchers and educators in a myriad of disciplines. As of this writing, VT-SIL has engaged researchers from many universities, covering a broad range of disciplines including civil, construction, mechanical, electrical, environmental, industrial, safety, systems engineering and mathematics, computer science, and even visual and performing arts, all interested to explore how to use the Goodwin Hall sensor data.

The research reported in this paper intends to apply digital libraries methods and techniques, assist researchers in data management, and foster sustainable user communities around them. Although seismology data repositories exist, they usually mandate data access methods, processing tools, and have very limited search options that mostly gear towards seismology researches. In contrast, we purposefully avoid prescribing and limiting what the data shall be used for and how they are used. In the next section we will examine our goals against OAIS, a popular data management model, to identify gaps.

3. OAIS CONSIDERED INADEQUATE

The OAIS Reference Model [17] is widely recognized to have provided a conceptual framework and common vocabulary to digital preservation problems. Many document-centric digital repository software claim to be OAIS-compliant, indicating its usefulness as an abstraction for archival functions, workflows, and organizations. Some would even suggest that implementing the OAIS model is sufficient for managing scientific data [6][13]. However, after examining the Goodwin Hall project use cases, we find the OAIS model inadequate for this purpose.

3.1 OAIS Environment

The OAIS environment divides the organizational and functional roles of the information producer, the archive, and the information consumer in an overly specific and sequential manner. This artificially elevates the archive from a facilitating infrastructural piece that should have been hidden in the background to an

autonomous, potentially self-serving, and even intrusive middleware.

In reality, typical researchers don't produce data solely for others to consume, therefore the first and foremost information consumers are usually the information producers themselves. These users already understand the data, don't necessarily have an immediate need for long-term preservation, but may need infrastructural help to store, process data, and make sense of them. Studies have shown that these researchers are highly sensitive to the first-use right when considering sharing [42][44]. This is particularly true for big data sets since they are both more costly to collect and potentially more valuable to science. But an archive in the OAIS sense does not have much to contribute to the first-use cycle therefore is unlikely to be engaged.

If we are to follow the OAIS model, in order for the data ingestion to happen, the archive needs to negotiate a binding submission agreement with the producers outlining the legal responsibilities between the two parties, implying once the data are out of their hands, the producers will have little control. The obvious consequence of this zero-sum strategy is that, even if the producers are still willing to share their data through the archive, the sharing would not happen until the producers have exploited the data to the maximum, which is usually fairly late in the research cycle. Because moving large data sets from where they originally reside to an archiving environment is understandably more difficult [38], it may take even longer for the archive to begin sharing the data. Extrapolating to research data the findings linking the paper recency to citation rate [35][28], we reasonably expect lower reuse rate for older data sets. Moreover, since researchers collaborate outside of the archive anyways while data are still fresh and active, such sharing would not be captured, improved, and augmented by the archive; further diminishing its potential value.

The model is also problematic on the consumer side, in that it needs to predict the unpredictable designated communities in the future. Besides the original data producers and their collaborators, unconventional, cross-disciplinary, and explorative researchers constitute important sources as data consumers. Indeed, one of the most exciting aspects of the data-intensive science is to break the artificial disciplinary boundaries, ask unconventional questions, and shed new light on old data. Enforcing disciplinary consensuses as well as biases in the archive, as many disciplinary data repositories currently do, may serve to build up the unwanted boundary.

3.2 OAIS Information Model

The OAIS information model is based on the assumption that the archive should primarily act as a storage facility that receives from the producer and disseminates to the consumer variably packaged information in discrete transmissions. The actual information consumption happens outside of the archive; therefore it is of little concern to it. This assumption starts to break down in big data management scenarios.

First, making sense of big data depends heavily on technologies and infrastructure. Big data sets cannot be easily analyzed with personal computers. No matter how complete and detailed the Representation Information is, without access to big data analytics infrastructure such as appropriately configured computer clusters, the Dissemination Information Package (DIP) is of little help to consumers.

Second, OAIS archives typically disseminate information through download links, but moving big data sets in and out of data centers is error-prone and time-consuming. It is not a secret that shipping hard drives is still the preferred option to move data when the size reaches a certain threshold. Disseminating information through user downloads is not likely to scale well.

These technology obstacles can easily paralyze an OAIS archive for big data management. Moreover, the OAIS information model overemphasizes describing and understanding information than putting the information into actual use. After all, we learn about fire by getting burnt, not by reading manuals and documentations describing its chemical compounds and temperature measure. The OAIS approach puts unnecessary burdens on both the producers and the consumers to produce and consume Representation Information. For example, how can a well-documented, but hardly actionable, big data DIP help an explorative researcher who needs to filter 60TB of vibration data against a specific wave pattern?

The above pattern filtering use case also brings out an important big data usage pattern not sufficiently addressed by the OAIS model. A document-centric archive can easily perform full-text indexing against all the content then use the resulting indexes as Access Aids to Information Packages. The assumption is that the indexing only needs to be performed once as long as the content has not changed. However this assumption does not hold for research data sets made up of digits instead of natural language vocabularies [12]. There is no good way to pre-index them for filtering queries yet to be invented in the future. Different from the specified OAIS data flow, such access queries need to be routed back to Archival Storage and trigger heavy data analytics computing jobs. A preservation and storage focused archive is not typically equipped with such capabilities.

3.3 The Use and Reuse Driven Approach

Now it should be clear that we are proposing a different data management approach than the OAIS model. This approach is more aligned with the DCC Curation Lifecycle Model [16], with the focus that the use and reuse should be the driving force. Metaphorically speaking, a data management system should not function like an antique store where people offload their attic findings and scavengers dig for hidden treasure. Instead, it should be a lively workshop equipped with powerful tools to handle big data sets as the raw materials. The workers then collaboratively build more sophisticated and specialized tools to cut, polish, and assemble the raw materials and various intermediate products in ways that make sense to their own needs, and along the way trade both their products and tools. In this workshop, data preservation and curation is like the stocking crew who cleans up the spaces and puts various materials and tools back to clearly labeled and easily accessible shelves. The purpose is to facilitate more efficient production, not to compete for the best-stocked shelves. A similar metaphor is "treating data like software" [40].

4. CORE REQUIREMENTS

The use and reuse driven data management approach described in section 3 gravitates towards a digital repository coupled with big data storage and processing capabilities. It should serve as both a data archive and a low-barrier big data analytics platform. In this section we extract its core capacity, performance and functional requirements and provide justifications.

4.1 Data Storage

Large, affordable, high-performance, highly available, and easily accessible long-term storage is the foremost core requirement to manage big data. We estimate the raw sensor data from Goodwin Hall alone will be accumulated at the rate of about 60 TB per

year. The data acquisition system runs continuously with no pause in between. To detect long-term structural changes, we must compare data spanning long enough periods, e.g., more than 5 years. When running analytics, we will need to access large amounts of data but prefer not to wait for very long. We will also need to open some of the data for public access therefore the storage system must not be constrained behind a security perimeter. These form the minimum storage requirements that must be met by any viable data management system. Since intermediate data will need temporary and in many occasions also long-term storage, the volume requirement will surely far exceed the minimum estimate.

Although the storage pricing keeps dropping, building mass storage systems still goes beyond the budget of most academic libraries. Before we receive sufficient funding to build our own storage system, we should consider renting from commercial storage vendors or applying for storage grants from institutional or even national computing infrastructure. However, we must carefully evaluate their pricing, performance, availability, and accessibility. Many storage grants are not meant for long-term use and must be cleared out in a few years.

4.2 Data Processing

Servers running archival repositories do not usually require high CPU or memory because once ingested the content does not need to be constantly and heavily processed. This may change for our system because it also serves as the data analytics platform for users. Depending on the analysis, some may even require extremely powerful servers. We differentiate these scenarios into horizontally and vertically scalable analysis tasks. Clustering many lower-end computing nodes can solve the former, but the latter must use high-end computers. Most shallow analytics tasks belong to the former and many even fall into the so-called "embarrassingly parallelizable" category. We should fully support horizontal scaling, but may choose to limit the support for vertical scaling, mainly because high-end machines are much harder to come by and would otherwise limit our infrastructure options and flexibility.

4.3 Links Between Storage and Processing

The links between the data storage and data processing must be fast, reliable, and scalable enough to sustain large amount of data movements. To understand the link scalability, consider moving *n* equally sized files from the storage to *n* different computing nodes. If the time required to complete this move is about the same as that to move a single file from the storage to the computing node, then the links scale very well. Such scalability is usually achieved by replicating and/or sharding the data among different storage nodes. Moreover, the physical bandwidth between the storage and processing nodes becomes a significant bottleneck when the size of the data reaches a threshold. Co-locating both type of nodes to the same data center can effectively break the blockage.

4.4 Data Repository

The repository may not need to physically store data, but can link to external data storage. Nevertheless it should be very fast, flexible, and scalable. Traditional repositories make the assumption that their workloads are largely READ dominant, but this may differ in our case, because ingesting large datasets and their derivatives may cause very high WRITE workload. The repository will need to scale well for both workloads. Preferably it would also support flexible metadata schemas or even linked data to accommodate multi-disciplinary data use and reuse.

4.5 Data Analytics

Although data curation may need some data analysis capability, the main purpose to build data analysis into our system is to empower researchers to directly answer science questions from the data. The barrier to perform analysis should be as low as possible. We must not make assumptions on how researchers ask questions and perform analysis, including what tools they'll be using. Although a few VT-SIL graduate students use MATLAB to filter data, we must not extrapolate this to all users and make MATLAB a mandatory tool to access and analyze the data. We should not even assume the computing platform, the operating system and its distribution, the programming language, the compiled libraries and their version numbers.

4.6 Reusable Data Analytics Deployment

Achieving 4.5 used to be very difficult. But with the advancement of virtualization techniques, analysis tools can now be developed and deployed through virtual machines without losing much usability and performance. As a result, analysis tools and processes built by one researcher may be easily replicated, replayed, and improved by another researcher. Our system should be able to support reuses of this type.

4.7 Data Reuse

Beside the tools built to analyze the data, the results of the analysis may also be deposited back to the repository as derivatives, and linked to both the tools and deployment used to perform the analysis as well as the researcher who developed them.

4.8 On-Demand Scalability

Much of the data analysis workload may not persist over time but occur in an ad hoc and on-demand manner. If we build the system based on the maximum possible load then most of the system resources may be wasted during the idle time. Ideally our system should be built on a shared infrastructure such that we can scale system resources up and down based on the workload.

4.9 Sustainability

Making use of large scale IT infrastructure certainly is expensive. We must consider how the system can sustain itself both technically and financially. What if the technologies used to build the system becomes obsolete? Is it possible to share the financial burden with the data users?

5. THE CASE FOR PUBLIC CLOUD

We have proposed a rather ambitious vision that depends heavily on IT infrastructure. Libraries do not usually consider themselves as being in the business of building infrastructure, especially at the individual institution level. However this does not prevent us from leveraging the existing infrastructure to architect data management systems. In this section we discuss the options we have and make the case for building the system in the public cloud.

5.1 Proprietary or Local Infrastructure

A small number of research projects may be fortunate enough to have sufficient funding to build proprietary big data management infrastructure. Assuming such an infrastructure satisfies the capacity and performance requirements, what are downsides? Cost/benefits balance can be the most significant one. If built by the peak load capacity, by not sharing infrastructure these projects will waste a lot of resources during low tides. Sustainability will

also become problematic because IT infrastructure ages rather fast therefore needs sustained funding.

Many research institutions also have campus-wide high-performance computing (HPC) infrastructure, typically consisting of clusters with hundreds of nodes plus some scratch storage space. Massive, long-term storage has not been incorporated in most such infrastructure, and the institutional boundary may pose a major obstacle for sharing and collaboration. Most such HPC infrastructure also suffers similar weaknesses described in the next section.

5.2 National HPC Infrastructure

Many big science projects have taken advantage of regional or even national HPC infrastructure to build their data management solutions. Why can't projects like the Goodwin Hall instrumentation do the same? Although we indeed intend to apply for such grants, we also notice a number of potential problems.

First, although new capacities are continuously added, in general the national HPC infrastructure is fairly crowded. It has been reported that some supercomputers have reached extremely high usage rate that even high-profile existing projects had to limit usage and look for external computing resources such as Amazon cloud [27]. Long-term storage is far from sufficient to support large numbers of big data projects. As of this writing, within XSEDE only the Texas Advanced Computing Center (TACC) is capable of allocating sufficient long-term storage space for the Goodwin Hall sensor data from its RANCH long-term storage system [46]. Even with more advanced data transfer tools [11], moving data across XSEDE still presents a major challenge [3]. If we are to avoid moving data far from the storage, TACC's Lonestar and Stampede become the only viable resources for the associated data processing and reuse. Furthermore, HPC services may experience frequent interruptions and overall do not match the quality of service promised by their commercial counterparts [3].

Second, the HPC resources tend to be allocated and used in a very rigid way. Very few supercomputers run virtualizations and allow users to customize operating systems and software. In order to take advantage of the computing resources, users have to adapt existing analysis code and software to different supercomputer environments. Deployments also differ from one environment to another therefore are generally not portable.

The HPC resources funding model is also rather apathetic to small teams and explorative researches. To gain access to national HPC resources the users must write grant proposal to gatekeeping committees to justify the request. But big data projects from small team, at infancy and researchers at the exploratory stage can hardly present well-defined use cases to compete with well-organized grand challenges. The situation is so severe that some would even allege, "there is an increasing divide" [9] between the Haves and the Have-Nots. Researchers in the latter group, while the majority in numbers, are more and more distanced from participating in the data intensive science due to the lack of access to big data resources and the required computing infrastructure.

5.3 Public Cloud

In contrast to the HPC infrastructure, public clouds have already provided a much larger pool of storage and computing resources, with much lower barriers for entry, and affordable pricing. Google, Microsoft, and Amazon are each running cloud services with about 1 million computing nodes located in a handful of mega data centers. Amazon EC2 spot prices remain low, indicating a low usage rate. Amazon S3 and Glacier were both estimated to have long surpassed exabyte order of storage. With large amount of co-located storage and computing nodes, moving big data sets out of the data center is no longer necessary.

Virtualization enables the users to run a wide range of software and code in the cloud without modification. Automated deployments may be shared between users like sharing code, and data can be acted upon without understanding every bit of details about the context. The cloud elasticity makes it possible to only pay for the resources consumed and return the excesses back to the shared pool. More importantly, a valid credit card is the only requirement to gain access. Combined with a nimble usage model, large public clouds are becoming more and more attractive to researchers with big data management needs. Recognizing this trend, even NSF has recommended in its CIFS21 Vision and Strategic Plan to balance traditional HPC services with "the growing number and capabilities of cloud systems and services" [31].

What about the handful of smaller, research oriented cloud providers such as the Open Science Data Cloud and CloudLab? Although these providers retain all the technical benefits of the cloud computing, for our purpose they have two serious limitations. First, the total size of the cloud is too small for us to benefit from economies of scale. While the Goodwin Hall sensor data set seems large for these smaller clouds, it is rather tiny in big ponds such as Google, Amazon, or Microsoft cloud. Second, these cloud providers retain the HPC funding model that allocates resources via grant application and review, therefore discriminates against open-ended, collaborative projects like ours.

Balancing the above options, it seems that large, openly shared public cloud such as Amazon, Google, or Microsoft becomes the most viable infrastructure candidate for our project. We therefore choose Amazon cloud as a starting point and build an implementation prototype, as described in the next few sections.

6. SYSTEM ARCHITECTURE AND WORKFLOWS

In this section we describe the cloud-based system architecture and its typical workflows of the proposed data management system.

6.1 Architecture

As illustrated in Figure 1, the system consists of 4 modules: data acquisition, cloud storage, digital library, and data processing.

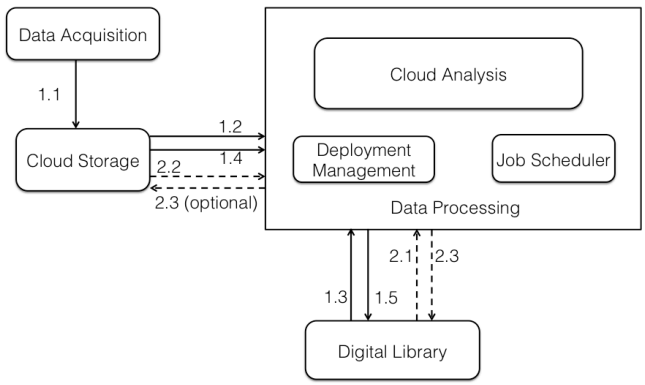

Figure 1. System Architecture and workflows

The data acquisition module collects data from distributed sensors, then calibrates, preprocesses, formats, and stores them in a local storage server. The Goodwin Hall accelerometer measurements are currently formatted in HDF5. The acquisition system also uploads the data to the cloud storage module in parallel. In case of any network interruption or upload error, the locally stored copy will be used to rectify the problem.

The cloud storage module provides the space to store both the raw sensor data uploaded from the data acquisition module and the derivative data resulting from various analysis and curation actions. Its reliability should be comparable to regular hard disc based file systems. The storage module should also be capable of triggering messages upon successful completion of typical file creation operations. We make the assumption that the storage and data processing modules are connected with reliable, high bandwidth, and highly scalable network links. This assumption will be evaluated in the next section.

The data analysis module is in turn made up of cloud analysis, deployment management, and job scheduler modules. The cloud analysis module holds the elastic virtual computing resources also called workers. It performs the user-defined data analysis tasks as well as system-defined data curation tasks. These tasks usually involve querying the digital library, finding out the cloud storage locations of various bitstreams, moving them via the scalable links to workers in the cloud analysis module, performing analysis, and if appropriate, depositing the results back to the digital library and the cloud storage. In order for this to happen, we need the deployment management module to deploy the workers and scale them depending on the size of the job. In many occasions we also need the job scheduler module to allocate computing jobs to workers. We will also evaluate in the next section how scalable the data processing module is.

The digital library module is a typical repository system with the exception it does not directly store all bitstreams in itself. Many bitstreams, especially those that constitute large data sets, are stored in the cloud storage with only the access links stored in the library. The library also holds the metadata, the software codes used for analysis and those to deploy analysis in the cloud analysis module. Relations between the data, the analysis and deployment codes, and users are also recorded in the library in forms of metadata. The library also provides web interfaces for user management and various user interactions with the data management system, such as performing data analyses and downloads.

In this architecture, the cloud storage and the cloud analysis modules must be co-located in the same data center.

6.2 Workflows

We now describe two typical workflows: data ingestion and data analysis.

6.2.1 Ingestion

The ingestion workflow starts from the data acquisition module. As shown in Figure 1 in solid lines, once 1.1) a data file is uploaded to the cloud storage, 1.2) the storage sends a message to the deployment management module notifying the upload completion. The deployment management module then initiates a data ingestion deployment, which 1.3) takes stock deployment script from the library to deploy workers, and schedules an ingestion job. When this job is popped from the job queue, either a new worker is created for this job or an existing free worker is assigned to this job. The worker starts to run the ingestion code, and 1.4) copies the file from the cloud storage to the worker and extracts the necessary metadata, then 1.5) creates a suitable container object in the library, attaches the corresponding bitstream link and extracted metadata. If many files are uploaded to the cloud storage in parallel, the job queue may grow longer, at which point more workers will be created to ingest files until the job queue shortens and workers become idle. Then the idle workers will be shut down one after another until the job queue is empty.

6.2.2 Analysis

The analysis workflow starts from the user triggered digital library action, as shown in Figure 1 in dotted lines. The library 2.1) initiates a data processing request, the deployment management module takes analysis deployment script from the library to deploy workers and schedules analysis jobs. When an analysis job is popped from the job queue, the assigned worker 2.2) copies corresponding file from the cloud storage to the worker to run the analysis job. Upon completion, 2.3) the worker returns the results back to the library. Optionally, if the user chooses to deposit the analysis results back to the repository, the worker moves the derivative bitstreams to the cloud storage and creates in the library an object container, the associated bitstream link, and metadata.

7. PROTOTYPE AND EVALUATION

The proposed system and its architecture would immediately collapse if the following two assumptions, even if reasonable in theory, do not hold up in reality. First, moving data between co-located storage and computing nodes should be fast and scale well. Second, embarrassingly parallelizable data analysis workload should scale well even if it involves previously assumed data movements. In order to gain high confidence before advancing to the next development phase, we build an evaluation prototype in the Amazon cloud based on the Goodwin Hall sensor data.

7.1 Prototype

As shown in Figure 2, the implemented prototype is largely based on the system architecture described in section 6.

We choose the newly released Fedora Repository version 4.0.0 [10] as the digital library, hoping its new features including improved performance, linked data platform support, clustering capability, and embedded fixity check etc. will form a solid foundation for future developments. We assume the typical digital library and archival functions can be successfully implemented in this prototype, given that a subset of this prototype is almost identical to what APTrust has already implemented, based on Fedora 3, Hydra, and Amazon S3 [1], as one of the five primary preservation nodes in the Digital Preservation Network (DPN).

We use Amazon S3 as the cloud storage module and Amazon EC2 as the cloud analysis module. We manually deploy EC2 instances as both workers and the library, and use Amazon SQS as the job scheduler. All Amazon services are deployed in the Amazon US East region.

We wrote Python codes that extract metadata from HDF5 files, perform simple mathematical operations such as calculating the maximum, minimum, mean, and median values from the data file, split and merge files, and draw wave charts from them, as well as call Fedora 4 APIs to create containers and attach bitstreams and metadata to them. We then deploy these code to 1-16 EC2 instances as workers to perform the evaluation tests. Figure 3

shows a small segment of a wave chart drawn from the accelerometer measurements from one signal channel.

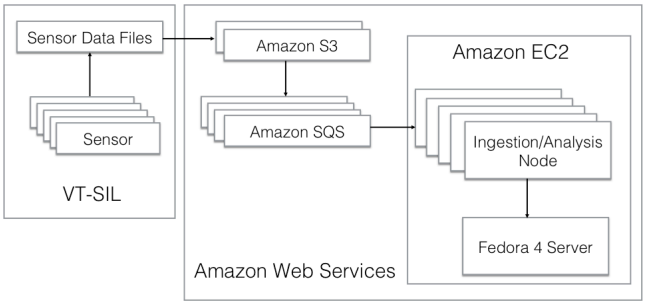

Figure 2. Evaluation prototype

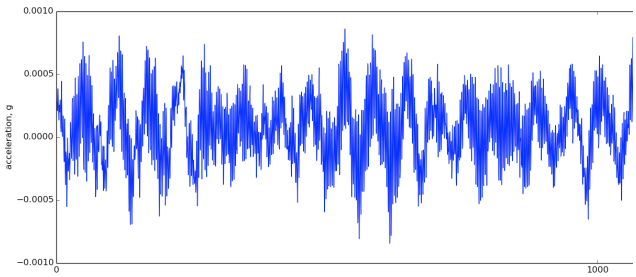

Figure 3. Accelerometer measurements

7.2 Evaluation

We evaluate the prototype using 24 hours of accelerometer measurements from 12 signal channels, with a total data size of about 130 GB. The actual data volume will be much higher than this because there are many more channels and other sensor types. We then evaluate the system against three different test cases: 1) simple data ingestion without data copying into the library, 2) simple data analysis with additional metadata write back to the library, and 3) a more CPU, I/O, and network intensive job. The third job first splits the data file into 6-second segments, then creates visualization from each segment, and finally deposits all created images, 172,800 in total, back to the library as a derivative data set and creates relations between them and the measurement data set. We run these three cases under the same system settings and the number of ingestion/analysis computing nodes, or workers, increases from 1 to 2, 4, 8, and finally 16.

For the sake of simplicity, our implementation took the following shortcuts without affecting the validity and the accuracy of the evaluation.

First, instead of fully developing the deployment module, we manually deploy the workers. We have not developed any web interfaces for these tests and other user interactions. All tests are completed with command line.

Second, we combine the data analysis test cases 2 and 3 with the simple ingestion test case 1 to create three similar test cases so that we can compare them on the same footing. In test case 1, we move the data from S3 to the workers, extract metadata, then create a container in Fedora 4 and insert the data items and metadata we extracted. In test case 2, we do all the above plus calculate the minimum, maximum, mean, and median values for each file then write them back to the library as the data item metadata. In test case 3, we do everything in test case 1 plus splitting the data, drawing 172800 image files, and depositing them back to the library.

Third, we bypass the upload stage of the workflow to save time. This is acceptable because we are not testing the Amazon upload bandwidth for external users. Instead, we copy the sensor data from one S3 bucket to another to emulate the upload completion.

Fourth, to ensure the library does not constitute a performance bottleneck in the evaluation, we tested and chose a rather powerful EC2 instance type, r3.8xlarge, for the Fedora 4 server. We choose m3.large for the worker nodes.

7.3 Results

The experimental results are presented in this section. The experiments were repeated for multiple times to ensure the results are reproducible. It takes about 2 weeks in total to complete all the experiments.

Table 1 lists the average time used to copy one single file from the cloud storage to a worker. A full day's vibration data collected from 12 channels are split into 972 files in more or less equal size. When multiple workers are deployed, they can process these sensor data files in parallel. Because the average time spent to move each file from S3 to EC2 does not change significantly with the job complexity or the number of workers, we can reasonably expect data movement speeds up linearly with the number of workers. The first assumption has been shown to hold up well.

Table 1. Average time in second spent to copy a file from Amazon S3 to Amazon EC2, both in the US East Region

	Number of Workers				
	1	2	4	8	16
1	0.2150	0.2525	0.2665	0.2045	0.2077
2	0.2186	0.2134	0.2675	0.2192	0.2073
3	0.2257	0.2412	0.2233	0.2086	0.2063

Table 2 lists the time spent to complete three different test cases using different number of workers. The results clearly show that in all three test cases, if we double the number of workers, the job gets done in approximately half the time previously required. The data processing scales linearly, therefore the second assumption is also shown to be satisfactory.

Table 2. Time in second spent to complete the test case

	Number of Workers				
	1	2	4	8	16
1	213.77	105.54	52.74	25.80	12.93
2	625.92	320.71	157.03	77.13	38.27
3	81564.42	40619.44	20284.53	10113.05	5059.34

8. DISCUSSIONS

Although much still needs to be done to fully realize the vision of use and reuse driven data management, the evaluations presented in section 7 have clearly demonstrated the technical feasibility to

manage big data in the cloud. Since technology changes inevitably carry organizational, cultural, and social consequences, in this section we briefly discuss what potential "soft" impact this approach may bring out.

8.1 Finance

The elephant in the room is how to fund the cloud operation in the long run [5]. The underlying concern is that the cloud may not be the most economic option.

If only considering the storage cost, a prior comparison argues the cloud preservation costs more than buying own servers [37], which is further corroborated by a few anecdotal reports from university IT departments who built their own mass storage facilities. These, however, have not taken into consideration the need for elastic computing capabilities co-located with the storage facility. We leave the more detailed cost analysis as important work in the near future, particularly when more realistic data reuse scenarios are implemented.

In addition to the monetary cost, we must also consider the opportunistic cost not included in these calculations. As explained before, time is of the essence for many research data management projects. If fresh data are not properly managed and sufficiently use and reused to answer science questions, they will age quickly, and their value will dilute and vanish. In most cases we do not have the budget, time, and expertise to build massive data centers, without which big data management cannot be done. The opportunistic cost is indeed what drives many small start-ups as well as large and mature IT organizations to the public clouds.

The flexible cloud billing approach also carries its own challenges and opportunities. Universities, especially the land grant, public accountable institutions, often frown upon novel billing methods. Unless the recurring charge is on a grant or kept to a rather small minimum, we are repeatedly advised by our controller's office to get an annual contract with Amazon for our pay-as-you-go cloud expenses.

The positive side of the flexible billing and deployment is that now we do not have to run a data management system fully funded by ourselves. At least we have the option not to cover the user-generated data usage and analytics costs, which is only fair. We may also adjust the operational costs on demand. For example, if very few people are using a certain data set for an extended period of time, we may move it to a cheaper, but less accessible storage tier, spin it off to user communities who strongly believe its long-term value, or eventually remove it altogether.

8.2 Organization

The use and reuse driven data management approach calls for a slightly different organizational structure from what is currently in place. We need to put the researchers' needs in higher priority and even adjust job responsibilities to reflect this change of mindset. Virginia Tech Libraries has created and filled data consultant positions co-funded with the academic colleges and departments.

The library IT department also adjusted with the technology changes. More system administrators are learning virtualization and cloud deployment, and one system administrator was converted to a systems engineer position created to support the cloud operation.

8.3 Culture

Researches have identified two distinctive data sharing culture among researchers: the top-down, highly organized sharing culture mandated by big science projects, and the lone-wolf, gift-exchanging culture predominating the small science projects. It is curious to us why the ad hoc sharing culture commonly seen in the open source software developments has not seeped into research data sharing. After all many influential software projects flourish in this culture and many developers are themselves researchers. We speculate the reward mechanism currently in place for data sharing may need some more tweaking.

Many researchers worry their data may be used for publications without properly acknowledgments. This rarely bothers the open source development, where each line of published code can be traced back and attributed to individual developers. For big data sets, it is hard enough to take the data to a different environment to be useful, therefore our approach of providing analytics capabilities with a library alongside with the data may actually create the overarching environment where every tiny bit of contribution can be recorded.

The "viral" open source license has also contributed to the success of the open source movement. It may be worth trying to enforce similar licenses in our system that demands the data user not only acknowledge the data producer, but also open up their own derivative work under the same terms and conditions.

8.4 Society

As the poor man's data center, the computing cloud has been attributed to the democratization of science [4][11]. By adding the digital library piece, we hope the increasing number of data-intensive researchers not fitting in the big science/small science dichotomy can also find their niche and flourish.

9. RELATED WORK

Equipping digital libraries with superior processing capabilities is not a new idea. Simulation libraries like HUBZero [30] and SimDL [25] would be futile if without powerful computing resources to run the models. HUBZero can tap into national HPC centers through the grid. SimDL is backed by a local cluster and co-located mass storage facility. Similarly, the SCAPE project outfits a large image repository with a local cluster running Hadoop in order to quickly validate and extract features from large number of archived JP2 images [19].

However, as we point out in section 5, these infrastructure choices may not be the best fit for big data management. A recent study clearly illustrates the limitations of using HPC for data management. It reports the experience running file format identification tool DROID on TACC's Stampede to extract metadata from a 4.3 TB dataset. A single Stampede node can reduce the processing time from 2 days to 6 hours, but moving data from where they are stored to Stampede, both of which are on the University of Texas Austin campus, took 28 hours [3]. In contrast, if using our 16-worker test results as a reference, we may be able to copy the same amount of data from S3 to EC2 in 7 minutes or even shorter if we use more workers. Moreover, in order to run DROID on Stampede, the researcher has to modify and rebuild DROID software.

HPC centers around the world are routinely occupied by big science teams, especially high-energy physicists. Their take on long-term data preservation [20][21] is worth comparing with the librarians' view. To these physicists, data is not just a package

that can be stashed away. Instead, "data analysis capabilities must be preserved"[2], e.g., in the form of deployable virtual machines or even better, recipes to recreate these machines [20] [21]. [2] gives an example on preserving 4PB of CDF data from Fermilab to INFN-CNAF in Italy. Besides replicating data, CNAF also offers grid-computing resources to run replicated Fermilab analysis tools on CERN supercomputers. Running virtual machines in private clouds is proposed as the key preservation strategy [20] [21] [36]. The approach proposed in this paper is more aligned with these scientists' view.

Cloud computing slowly gains traction in digital libraries. Duracloud [8] and APTrust [1] are two examples in which cloud storage is used for as long-term preservation. [45] describes migrating CiteSeerX to a private cloud, although the system shares crawled data in very specific, limited ways and does not allow users to perform customized analysis against the data.

If paring with Amazon cloud services, Amazon's own public data sets [33] may be used in very flexible ways. But except for a searchable catalog page and links to each data set, no other digital library function appears to exist. Data submission is a manual process via a contact email. Data sets are made available in various ways. Some stored in S3, some must be mounted as EBS blocks or EC2 images. Without a digital library, users lack an integrated environment to collaborate and share tools as well as derivative data.

10. CONCLUSIONS

Motivated by big data management needs arising from a small team, we work towards integrating digital libraries and big data analytics in the cloud. We developed a prototype that has been shown to largely satisfy our requirements. It scales linearly therefore warrants more extensive use in the future. We also discuss the "soft" impact of this approach.

Despite the cloud's vast popularity among researchers and the industry, its full recognition and adoption by the digital libraries community is surprisingly long and strenuous. The community has raised many legitimate concerns, ranging from the cost, the vendor lock-in, and some unacceptable terms of use. Balancing these concerns with the cloud's obvious advantages and values should be beneficial.

11. REFERENCES

[1] Academic Preservation Trust: *http://aptrust.org/.* Accessed: 2015-01-23.

[2] Amerio, S., Chiarelli, L., dell' Agnello, L., Girolamo, D.D., Gregori, D., Pezzi, M., Prosperini, A., Ricci, P., Rosso, F. and Zani, S. 2014. Long Term Data Preservation for CDF at INFN-CNAF. *Journal of Physics: Conference Series*. 513, 4 (Jun. 2014), 042011.

[3] Arora, R., Esteva, M. and Trelogan, J. 2014. Leveraging High Performance Computing for Managing Large and Evolving Data Collections. *International Journal of Digital Curation*. 9, 2 (Oct. 2014), 17–27.

[4] Barga, R., Gannon, D. and Reed, D. 2011. The Client and the Cloud: Democratizing Research Computing. *IEEE Internet Computing*. 15, 1 (Jan. 2011), 72–75.

[5] Berman, F. and Cerf, V. 2013. Who Will Pay for Public Access to Research Data? *Science*. 341, 6146 (Aug. 2013), 616–617.

[6] Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S. and Matthews, B. 2013. Data Management and Preservation Planning for Big Science. *International Journal of Digital Curation*

[7] Borgman, C.L., Darch, P.T., Sands, A.E., Wallis, J.C. and Traweek, S. 2014. The ups and downs of knowledge infrastructures in science: Implications for data management. *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (Sep. 2014), 257–266.

[8] DuraCloud: http://www.duracloud.org/. Accessed: 2015-01-23.

[9] Farcas, C., Balac, N. and Ohno-Machado, L. 2013. Biomedical CyberInfrastructure Challenges. *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery* (New York, NY, USA, 2013), 6:1–6:4.

[10] Fedora Repository: http://fedorarepository.org/. Accessed: 2015-01-23.

[11] Foster, I. 2011. Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *IEEE Internet Computing*. 15, 3 (May 2011), 70–73.

[12] Friedrich, T. and Kempf, A.O. 2014. Making research data findable in digital libraries: A layered model for user-oriented indexing of survey data. (Sep. 2014), 53–56.

[13] Gray, N., Carozzi, T. and Woan, G. 2012. Managing Research Data in Big Science. University of Glasgow. arXiv:1207.3923. (Jul. 2012).

[14] Hamilton, J.M., Joyce, B.S., Kasarda, M.E. and Tarazaga, P.A. 2014. Characterization of Human Motion Through Floor Vibration. *Dynamics of Civil Structures*, Volume 4. F.N. Catbas, ed. Springer International Publishing. 163–170.

[15] Heidorn, P.B. 2008. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*. 57, 2 (2008), 280–299.

[16] Higgins, S. 2008. The DCC curation lifecycle model. *International Journal of Digital Curation*. 3, 1 (2008), 134–140.

[17] ISO 14721:2003 2003. Open Archival Information System - Reference Model.

[18] Johns Hopkins University Data Management Services. 2014. Archiving Services We Offer: *http://dmp.data.jhu.edu/preserve-share-research-data/archiving-services-we-offer/.* Accessed: 2015-01-09.

[19] Jurik, B.A., Blekinge, A.A., Ferneke-Nielsen, R.B. and Moldrup-Dalum, P. 2014. Bridging the gap between real world repositories and Scalable Preservation Environments. *IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (Sep. 2014), 127–136.

[20] Kemp, Y. and Ozerov, D. 2012. Preparing experiments' software for long term analysis and data preservation. *Journal of Physics: Conference Series*. 396, 6 (Dec. 2012), 062011.

[21] Kemp, Y., Strutz, M. and Hessling, H. 2012. A validation system for data preservation in HEP. *Journal of Physics: Conference Series*. 368, 1 (Jun. 2012), 012027.

[22] Kohler, M.D., Heaton, T.H. and Bradford, S.C. 2007. Propagating waves in the steel, moment-frame factor building recorded during earthquakes. *Bulletin of the Seismological Society of America*. 97, 4 (2007), 1334–1345.

[23] Krafft, D.B. 2014. The National Data Service: A Library Perspective. http://www.youtube.com/watch?v=_C2UIgRrcB4

[24] Lee, A.C. 2011. A framework for contextual information in digital collections. *Journal of Documentation*. 67, 1 (2011), 95–143.

[25] Leidig, J., Fox, E.A., Hall, K., Marathe, M. and Mortveit, H. 2011. SimDL: A Model Ontology Driven Digital Library for Simulation Systems. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (New York, NY, USA, 2011), 81–84.

[26] Lynch, C. 2008. Big data: How do your data grow? *Nature*. 455, 7209 (Sep. 2008), 28–29.

[27] Madduri, R.K., Dave, P., Sulakhe, D., Lacinski, L., Liu, B. and Foster, I.T. 2013. Experiences in Building a Next-generation Sequencing Analysis Service Using Galaxy, Globus Online and Amazon Web Service. *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery* (New York, NY, USA, 2013), 34:1–34:3.

[28] Martin, T., Ball, B., Karrer, B. and Newman, M.E.J. 2013. Coauthorship and citation patterns in the Physical Review. *Physical Review E*. 88, 1 (Jul. 2013).

[29] McKay, D. 2014. Bend me, shape me: A practical experience of repurposing research data. *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (Sep. 2014), 399–402.

[30] McLennan, M. and Kennell, R. 2010. HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Computing in Science Engineering*. 12, 2 (Mar. 2010), 48–53.

[31] NSF. 2012. *Cyberinfrastructure for 21st Century Science and Engineering (CIF21) Advanced Computing Infrastructure: Vision and Strategic Plan*. [Online]. Available: http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf. [Accessed: 05-Dec-2014]"

[32] Palmer, C.L., Weber, N.M. and Cragin, M.H. 2011. The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*. 48, 1 (2011), 1–10.

[33] Public Data Sets on AWS: http://aws.amazon.com/public-data-sets/. Accessed: 2015-01-23.

[34] PREMIS Editorial Committee. 2008. *PREMIS data dictionary for preservation metadata*, version 2.0.

[35] Redner, S. 2005. Citation statistics from 110 years of physical review. *Physics Today*. 58, 6 (Jun. 2005), 49–54.

[36] Resines, M.Z., Heikkila, S.S., Duellmann, D., Adde, G., Toebbicke, R., Hughes, J. and Wang, L. 2014. Evaluation of the Huawei UDS cloud storage system for CERN specific data. *Journal of Physics: Conference Series*. 513, 4 (Jun. 2014), 042024.

[37] Rosenthal, D.S.H. and Vargas, D.L. 2013. Distributed Digital Preservation in the Cloud. *International Journal of Digital Curation*. 8, 1 (Jun. 2013), 107–119.

[38] Sands, A.E., Borgman, C.L., Traweek, S. and Wynholds, L.A. 2014. We're Working On It: Transferring the Sloan Digital Sky Survey from Laboratory to Library. *International Journal of Digital Curation*. 9, 2 (Oct. 2014), 98–110.

[39] Schloemann, J., Malladi, S., Woolard, M., Hamilton, J. M., Buehrer, M., Tarazaga, P.A. 2015. Vibration Event Localization in an Instrumented Building, *IMAC XXXIII A Conference and Exposition on Structural Dynamics*, (Orlando, FL, 2015).

[40] Schopf, J.M. 2012. Treating Data Like Software: A Case for Production Quality Data. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY, USA, 2012), 153–156.

[41] Snieder, R. and Şafak, E. 2006. Extracting the building response using seismic interferometry: Theory and application to the Millikan Library in Pasadena, California. *Bulletin of the Seismological Society of America*. 96, 2 (2006), 586–598.

[42] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*. 6, 6 (Jun. 2011), e21101.

[43] The 1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature*. 467, 7319 (Oct. 2010), 1061–1073.

[44] Wallis, J.C., Rolando, E. and Borgman, C.L. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*. 8, 7 (Jul. 2013), e67332.

[45] Wu, Z., Wu, J., Khabsa, M., Williams, K., Chen, H.-H., Huang, W., Tuarob, S., Choudhury, S.R., Ororbia, A., Mitra, P. and Giles, C.L. 2014. Towards building a scholarly big data platform: Challenges, lessons and opportunities. (Sep. 2014), 117–126.

[46] XSEDE Storage: https://www.xsede.org/storage. Accessed: 2015-01-22.

[47] York, D.G., Adelman, J., Anderson Jr, J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J.A., Barkhouser, R., Bastian, S., Berman, E. and others 2000. The sloan digital sky survey: Technical summary. *The Astronomical Journal*. 120, 3 (2000), 1579