Portable Data Management Cloud for Field Science

Yuma Matsui, Aaron Gidding, Thomas E. Levy, Falko Kuester, Thomas A. DeFanti
California Institute for Telecommunications and Information Technology (Calit2)
University of California, San Diego
La Jolla, USA
e-mail: {yumatsui, agidding, tlevy, fkuester, tdefanti}@ucsd.edu

***Abstract*—A modern field science such as archaeology is heavily data-driven using various kinds of state-of-the-art measurement instruments. It requires sophisticated computer infrastructure to manage large amounts of heterogeneous data. The concept of cloud computing provides a flexible cyber infrastructure for large-scale data management, which is being deployed at university campuses. A problem unique to field research is that researchers often work at remote field sites with limited computer and network resources. For a data management system that has to work in the campus cloud and under vastly different field conditions, portability of computer infrastructure and common data access methods are essential requirements. This paper explores the portability of cloud infrastructure and illustrates the portable data management system that we used in a recent archaeological expedition.**

Keywords-cloud portability; IaaS; data management; cloud storage; database

I. INTRODUCTION

Data-intensive scientific research, in which large amounts of digital data are collected and analyzed for new discoveries, is becoming mainstream in many fields of science. At the California Institute for Telecommunications and Information Technology (Calit2), we deploy a cyber-archaeology methodology to deal with the persistent data avalanche [1]. For example, we collect massive data amounts through diagnostic and analytical imaging including geospatially referenced, ground and airborne, multi-spectral imagery, combined with topography information acquired via high-resolution laser scanning (LIDAR). We also record a huge quantity of metadata for archaeological sites and artifacts, including data for artifacts' shapes, spectrums, temporal information, inventory, and so forth. We are developing a data management system to handle these [2].

To utilize these large-scale data, we deploy a data life cycle depicted in Fig. 1. The first step is to collect digital data in the field, followed by data organization and annotation, which requires sophisticated computer infrastructure to streamline the workflow and continuous stream of collected data. The next steps are data visualization and analysis to extract information from the co-located, organized data assets, and develop or refine hypotheses before results are fed back into the cyber-archaeology loop.

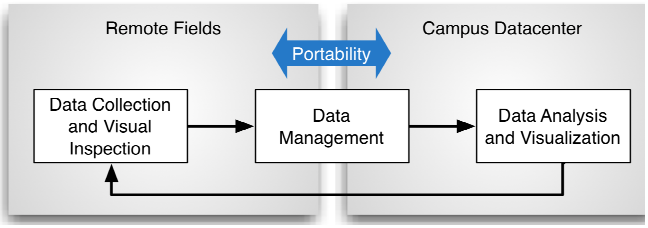

Figure 1. Life cycle of field research data in cyber-archaeology.

In terms of field science activities, there are two spatially separated, tightly coupled locations. One is the campus that has a rich computing infrastructure with high-speed network, and the other is an isolated field site where data are being collected in a comparably IT starved setting. Despite of this digital divide, data have to be transported between the "dirt archaeology" and the "digital archaeology" site in the field and on-campus, respectively. To bridge this gap, we need a mobile data management system equipped with abstracted program and data hosting infrastructure and means of data access.

II. PORTABILITY OF DATA MANAGEMENT CLOUD

A. *Program and data portability*

According to the NIST definition of the cloud [3], the service models of the cloud are classified as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS hides details of computing infrastructure from users and provides them granular control at the operating system level. Ideally, IaaS-based services would be portable in nature because an entire computer environment is virtualized and independent of the hardware. An open-source IaaS software stack such as OpenStack can handle standard virtual disk formats and thus realizes interoperable portability of virtualized computers.

On the other hand, PaaS serves as an application platform. It is difficult to achieve portability of PaaS-based applications because each PaaS has different underlying application frameworks and middleware. Still, an open-source PaaS software stack like Cloud Foundry is emerging, and it is equipped with open-source de facto standard application frameworks and databases.

In contrast, SaaS is a specific web application on the Internet. In SaaS, users cannot touch underlying programs, and their data can be hardly exported. Therefore, SaaS applications are not considered portable.

Our system is hosted on private campus cloud, and we can choose a suitable cloud service model for our use case that requires portability. We have adopted the IaaS model to take advantage of its fully controllable virtualized environment.

B. *Data access*

The goal of our system is to streamline data collection, data management, analysis, and visualization for cyber-archaeology.

In the data management process, we handle two types of data: geospatial and artifact. Geospatial data include data of three-dimensional points and lines from total stations, three-dimensional point cloud data from LIDAR, and stereo aerial photographs. Artifact data involve inventory information, archaeological metadata, and measurement data such as spectrum data from XRF or FTIR instruments. We also record artifact shapes with digital cameras and 3D scanners.

Representation of these data is in two forms. Raw measurement data such as LIDAR, XRF, and FTIR data are generated as files in their unique formats. These data should be kept as the original files for later analyses and visualizations. On the other hand, artifact inventory data, artifact metadata, and total station data are structured data. It is suitable that file-based data are stored in cloud object storage, and structured data are stored in a database.

We need a unified way to access these data in order to make them available to analysis or visualization applications. We adopt REST web services as a data access method because of the simple interoperability. Regarding cloud storage, there are some attempts to achieve vendor-free federated storage such as RACS [4] or MetaStorage [5], and they use Amazon S3 style API as a standard. We also utilize S3 style REST API to access file objects and XML/JSON REST API to access structured data.

III. PRELIMINARY SYSTEM

We have taken the IaaS approach to design a portable data management system. Fig. 2 illustrates the components of our system. On our campus, the system consists of data management servers and visualization facilities. The servers include a front-end web application server and a database server, which are running on virtualized IaaS cloud environment, and large-scale object storage. The visualization facilities are the immersive virtual reality environment (StarCAVE) or the huge tiled displays (HiPerSpace) at Calit2 to visualize and analyze a large volume of data. The servers make the data public with the web service API. In remote fields, we have some measurement instruments and data management servers in a small scale. The servers exist in a closed local network and work as a central data repository to archive all the data.

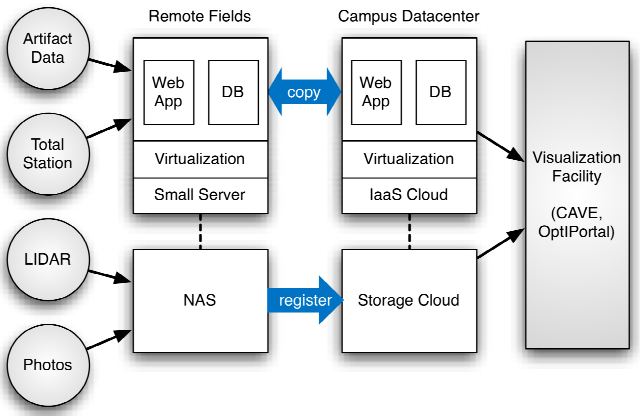

Figure 2. System components of cyber-archaeology infrastructure.

The workflow at a field site is as follows.

- Various data are collected in excavations.
- Structured data are put into a database through the web application.
- Raw file data are stored in network-attached storage.

When we go back to our campus, we perform the following operations.

- Data and programs from fields are moved to a campus cloud infrastructure.
- Data analyses and visualizations are executed with the collected data.

Data synchronization between separate environments is essential for cloud portability. In our system, a web application server and database server are in virtualized environments, and thus one-way synchronization from a field to a campus is easily achieved by copying virtual disk images. When we need granular control over data synchronization, we can directly use the database API. Concerning synchronization of file objects, we need to register all files temporarily stored in a NAS to the storage cloud. Once files are registered, the storage generates URL for each file, and thereafter the data can be uniquely accessed with the URLs. In the final analysis and visualization phase, web-based or stand-alone analytical or visualization programs will pull structured data from the database and raw files from the storage cloud with the established web service API.

IV. CONCLUSION AND FUTURE WORK

We are building a cyber-archaeology data management infrastructure that is usable both in field sites and on our campus. It is equipped with IaaS-based virtualized hosting environments and data access web services. We used the preliminary system at an excavation in 2011, and it worked as expected. Integration of the system with analysis and visualization is still in progress and is a major objective in our future work.

REFERENCES

[1] V. Petrovic, A. Gidding, T. Wypych, F. Kuester, T. DeFanti, and T. Levy, "Dealing with Archaeology's Data Avalanche," Computer, vol. 44, no. 7, pp. 56-60, Jul. 2011.

[2] A. Gidding, Y. Matsui, T. E. Levy, T. DeFanti, and F. Kuester, "e-Science and the Archaeological Frontier," in 2011 IEEE Seventh International Conference on eScience, 2011, pp. 166-172.

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology Special Publication 800-145, 2011

[4] H. Abu-Libdeh, L. Princehouse, and H. Weatherspoon, "RACS: A Case for Cloud Storage Diversity," in Proceedings of the 1st ACM symposium on Cloud computing, 2010, pp. 229-239.

[5] D. Bermbach, M. Klems, S. Tai, and M. Menzel, "MetaStorage: A Federated Cloud Storage System to Manage Consistency-Latency Tradeoffs," in 2011 IEEE 4th International Conference on Cloud Computing, 2011, pp. 452-459.