Approach Architecture Design for Provenance Layer in Opensource Cloud Computing

N. S. Gowri Ganesh¹,* and A. Rajiv Kannan²

¹Centre for Development of Advanced Computing, Chennai 600 113, India.
²Department of CSE, K.S.R. College of Engineering, Bapatla 522 101, India.
e-mail: nsgganesh@cdac.in; gowriganesh@gmail.com

Abstract. Provenance is a metadata that describes where and why an object came to be in its present state. There are many security issues in the applications that uses provenance. Major factor is that the application data and the provenance does not share the same access control. Cloud Computing allows the users to host, develop and use the application through the internet. The services of the cloud computing does not have any provision separately for storing provenance. Few works had dealt with the automatic provenance collection in Amazon web Services environment. In this paper it is proposed to have the approach architecture design for the inclusion of the provenance layer in the opensource cloud computing environment.

Keywords: Provenance, Open source, Cloud computing.

1. Introduction

Cloud computing is the emerging technology that allows the convenience for the application hosters to host their application without having to invest largely in the infrastructure, application developers to develop the application of wide variety of platform and users to use the application in the internet without constraining to a particular platform. The service providers are rolling out many services such as AmazonWeb Services (AWS), Google’s AppEngine, Microsoft’s Windows Azure, and Rackspace’s Cloud Servers and Cloud Sites. Also these providers offer various types of services such as computation, storage and networking Amazon webservice offers computational service called as Elastic compute cloud (EC2), SimpleDB as storage service, Simple Query Service and Amazon networking service that connects the other service [1].

Provenance is the french word (proveneir) means origin, source, history of objects ownership or location of the object. In Computer Science terminology it is the information about the objects and the action taken by/on that particular object [2]. It is metadata that is stored in a database either manually or automatically. Provenance aware storage Systems (PASS) records the provenance automatically in a database system and maintain over time [7].

Provenance differs [3] from the data for which it is generated in the following ways

It is the meta-data represented by directed acyclic graph(DAG) in which each node represents an entity and the edge is the causal relationship.

Though it records the history of change of data, the provenance itself is immutable.

In paper [4], the authors had discussed that provenance is important as it is vital in understanding, validating and sharing a certain product as well as the applications and programs used to derive it. It also insists about the important of the provenance that clouds are the future playground for escience research to track the process and reproducibility of results.

2. Related Work

There are many works related to Provenance and Cloud computing which is discussed in twofolds. Firstly, how the provenance is collected in a standalone and distributed environment. Secondly, it deals with the various implementation using opensource technologies and the provenance in cloud computing.

*Corresponding author
2.1 Provenance

The paper [5] discusses about the automatic provenance collection called as Provenance Aware Storage System (PASS). PASS is a linux kernel modified for automatic provenance collection, a kernel level port for Berkely DB for indexing and a stackable system layer called PASTA. Provenance is represented in name/value pairs.

The provenance [6] can be observed provenance and disclosed provenance systems. Observed provenance is that it observes the events of the process whenever it creates or modifies an object. In Disclosed provenance system, users or process provide the system, using the provenance system as storage and query engine. Some of the disclosed provenance systems are manual provenance systems or annotations and specified provenance systems.

In this [7] paper it suggests that operating system is responsible for provenance collection and storage for the provenance management.

There are other approaches for provenance collection such as job provenance [9], virtual data grid [10] and Taverna [11] which embed provenance collection in workflow engines.

An initiation of provenance standardization is conceived by at the IPAW’06 conference among the participants which resulted in series of first provenance challenge, second provenance challenge and third provenance challenge workshops. As a consequence Open Provenance Model was drafted to achieve interoperability in the provenance challenge series. W3C incubator on provenance has identified usecases and requirements for provenance on the web. Distributed PASS [2] had attempted for the automatic provenance collection in distributed environment.

2.2 Cloud Computing

Many players are involved in provision of different services under Cloud computing such as Elastic Computing cloud (EC2), Google App Engine 2, Salesforce.com, Microsoft Azure platform. In many of the cloud computing platforms, users are unaware of status of the data. This insists that users will be interested to know what happens to the data object and the process that works on it. The provenance will be the solution for such requirement. The same concept is insisted in this paper [4] in which provenance is termed as the solution for issues like tracking data production among service providers Eucalyptus, XCP and Open Nebula are some of the opensource cloud computing solutions.

Following are the utility and needs of the provenance for the cloud [8]

- Debug Experiment results
- Detect and avoid faulty data propagation
- Improving text base results

This paper [8] presented the design and implementation of storage of provenance with the existing Amazon web Services. Generally cloud services are Object store services, Database service and messaging services. Three protocols is designed for storing provenance, one as standalone cloud store, two as cloud store with database and third as cloud store with database and messaging services.

3. Background: Opensource Cloud Computing

3.1 Eucalyptus

Eucalyptus enables the creation of on-premise private clouds, with no requirements for retooling the organization’s existing IT infrastructure or need to introduce specialized hardware. It provides an EC2-compatible cloud computing platform and S3-compatible cloud storage platform. Eucalyptus has become very popular and is seen as one of the key open source cloud platforms. Since Eucalyptus makes its services available through EC2/S3 compatible APIs, the client tools written for AWS can be used with Eucalyptus as well. Walrus acts as the front end for the storage subsystem.

3.2 Open nebula

OpenNebula is an open-source toolkit to easily build any type of cloud: private,public and hybrid. OpenNebula has been designed to be integrated with any networking and storage solution and so to fit into any existing data center. With OpenNebula you can transform your data center into a flexible and agile virtual infrastructure which dynamically adapts to the changing demands of the service workload. OpenNebula orchestrates storage, network and virtualization technologies to enable the dynamic placement of multi-tier services (groups of interconnected virtual machines) on
distributed infrastructures, combining both data center resources and remote cloud resources, according to allocation policies.

### 3.3 Nimbus

Nimbus is a set of open source tools that together provide an Infrastructure-as-a-Service” (IaaS) cloud computing solution. Our mission is to evolve the infrastructure with emphasis on the needs of science, but many non-scientific use cases are supported as well. Nimbus allows a client to lease remote resources by deploying virtual machines (VMs) on those resources and configuring them to represent an environment desired by the user. It was formerly known as the “Virtual Workspace Service” (VWS) but the “workspace service” is technically just one the components in the software collection.

### 4. Provenance in Opensource Cloud Computing

Each solution of the open source cloud computing presents different vision in cloud architecture and implementation. This paper proposes the introduction of provenance layer in the existing architecture of Eucalyptus. The architecture can be described as in the figure below.

#### 4.1 Provenance layer

Where Provenance layer resides Walrus Storage provides simple storage service using REST and SOAP API that stores machine images, snapshots. Provenance layer resides within the walrus storage layer and the storage controller.

Walrus supports Simple storage services(S3) S3 services as specified below to mention few services that are supported by Walrus in Eucalyptus.

- Read and delete objects containing from 1 byte to 5 terabytes of data each. The number of objects you can store is unlimited.

- Each object is stored in a bucket and retrieved via a unique, developer-assigned key. Uses standards-based REST and SOAP interfaces designed to work with any Internet-development toolkit.

- Components of Provenance layer Provenance layer which is part of Walrus storage can contain the components-an interface to access objects, File system and storage system which is provenance aware.

- Interface to access objects Client side application can initiate actions such as read or write objects from and to the storage system through the web interfaces or RESTFUL or SOAP web services.

- File System File System is of provenance aware to recognize the each objects in the storage system. Both the provenance and the files are stored in cache in the file system during the operations between File open to File close when the files are in use. Storage System Both the provenance and files are stored as objects in the Storage system as per the S3 storage system specifications.

- Sequence for storing provenance and data Read the provenance file and convert into the attribute value pairs as required by walrus.
What Provenance data in the cloud? The provenance data in the cloud could be the customer identification number, data that is transferred between the virtual machines and Physical Machines.

5. Conclusion

The architecture design proposed will provide the enhancement for the existing cloud computing environment with additional benefits of provenance such as security, data integrity which is an essential components for the cloud.

Acknowledgments

The author is employed in Centre for Development of Advanced Computing under DeITy, Government of India. Author bow fully thank the organization for providing the immense support to perform the research work.

References