Collaborative Digital Library Services in a Cloud

Kurt Maly[1], Harris Wu[2], Mohammad Zubair[1], Milena Mektesheva[1]
Department of Computer Science, Old Dominion University, Norfolk, VA, USA

maly@cs.odu.edu, hwu@odu.edu, zubair@cs.odu.edu
1) Department of Computer Science, Old Dominion University, Norfolk, VA, USA
e-mail: maly@cs.odu.edu; zubair@cs.odu.edu, mmekt001@odu.edu
2) Department Information Technology and Decision Sciences, Old Dominion University, Norfolk, VA, USA
e-mail: hwu@odu.edu

***Abstract*— We have developed a web-based system that allows users to collaboratively organize large online image collections according to an evolving faceted classification schema. One of the major issues identified in the early deployment and evaluation in a university setting is the scalability of the system on traditional server implementations. Traditional computing cannot support ever-increasing number of users, documents, schema objects, schema history, and automated classification processes without difficult, expensive and time consuming resource reconfiguration. To address this problem we are proposing to move our system on a cloud-based Microsoft Windows Azure platform exposing it to users as a collaborative cloud service. Cloud computing will enable the Facet System to scale virtually unlimited. In this paper, we describe the architectural design for deploying our facet-based system and early prototypical modules as well as the work in progress implementing it on the Windows Azure platform.**

***Keywords*-faceted digital library; windows azure; cloud computing**

I. INTRODUCTION

We have developed a web-based system that allows users to collaboratively organize large online multimedia collections into an evolving faceted classification [1; 2; 3]. The system includes backend algorithms that systematically enrich the classification and automatically classify documents [4; 5; 6]. Evaluation of the prototype system, hereby referred to as the Facet System, shows promise [7; 8]. One of the major issues identified is the scalability of the system on traditional server implementations. Traditional computing cannot support ever-increasing number of users, documents, schema objects, schema history, and automated classification processes without difficult, expensive and time consuming resource reconfiguration. To address this problem we are proposing to move our system on a cloud-based Microsoft Windows Azure platform [14], exposing it to users as a collaborative cloud service. Cloud computing will enable the Facet System to scale virtually unlimited. There are a number of cloud services available in the commercial market but, as an academic institution, we will propose to use the time being made available on Microsoft Azure by a joint RFP by the National Science Foundation and Microsoft. Evaluating a social classification service on Windows Azure will allow us answer research questions specific to large-scale deployment of social systems that harness and cultivate collective intelligence. In section 2, we provide a review of the existing facet based system with an emphasis on the compute intensive nature of some of the features.. In section 3, we describe more specifically the evaluations we have performed thqat demonstrates the existence of severe scalability issues. Section 4 gives our design for deploying the facet system onto an azure architecture. In the final section we lay out the future work tasks.

II. BACKGROUND

In this section, we briefly review the existing facet based system. Fig. 1 and Fig. 2 show the browsing and classification screens in the Facet System, through which users can browse and collaboratively evolve a multi-faceted classification. In addition to a global classification of a large collection, we are adding a personal schema feature to the Facet System. In Fig. 2, clicking on the "person" icon above the faceted classification (the 1st of the 3 icons) will display a personal schema. The personal schema allows user to have a personal, persistent, idiosyncratic view of the collection [9]. In addition, most likely a user is interested in only a subset of the collection, when the collection is very large. Furthermore, a category in a personal schema can be a dynamic link (like a "shortcut" in Windows File Explorer) to a category in the global collection, which is constantly evolved by the community.

Fig. 3 shows the classification screen with both global and personal schemas. The back-end algorithms utilize the metadata in personal schemas for enrichment (construction, pruning, etc.) of global schema and automated classifications. Except for shortcuts, maintenance of facets and categories in a personal hierarchy requires separate storage in the database. To reduce user efforts in classifying documents into her personal hierarchy, a user can instruct the system to automatically classify or recommend documents from the overall collection into the personal schema.

When automated classification is enabled for the personal hierarchy (in user preference settings), the backend algorithms take significant amount of computing resources for each additional user. Furthermore, our system supports schema history – which allows users to examine global or personal schema at any given point in time.

 ISBN: 978-1-61208-105-2

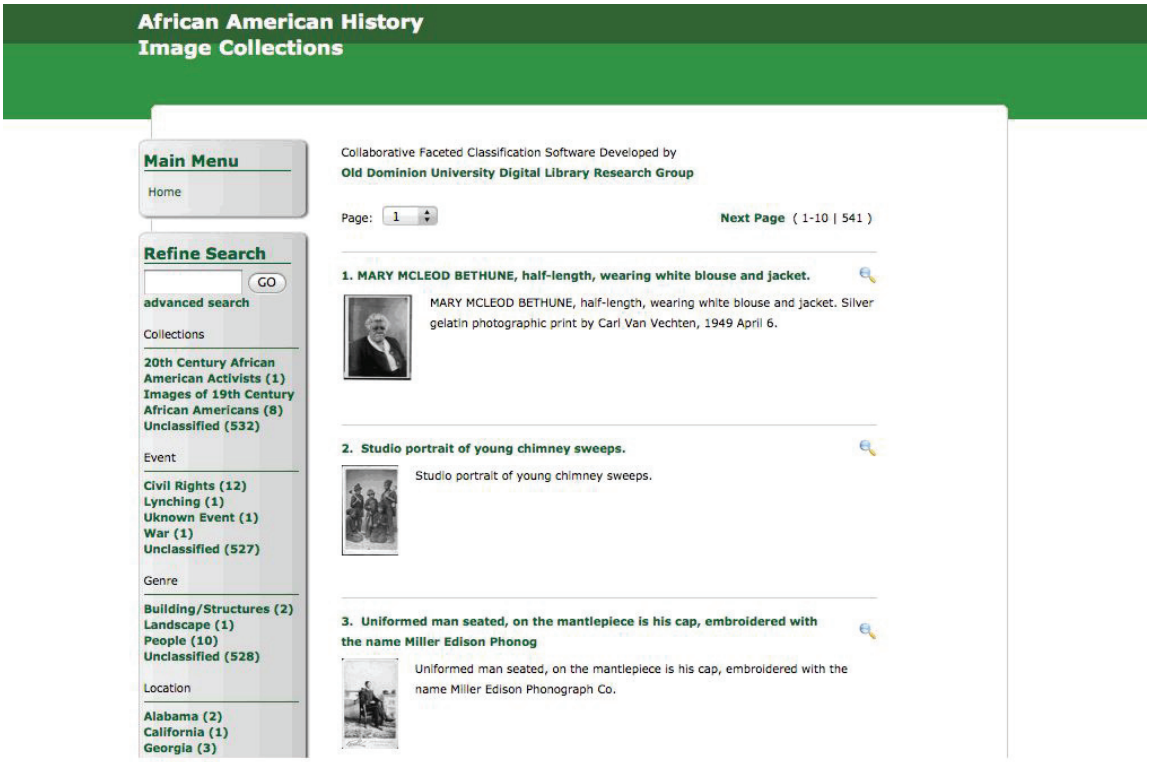

Figure 1. Browsing screen of the Facet System.

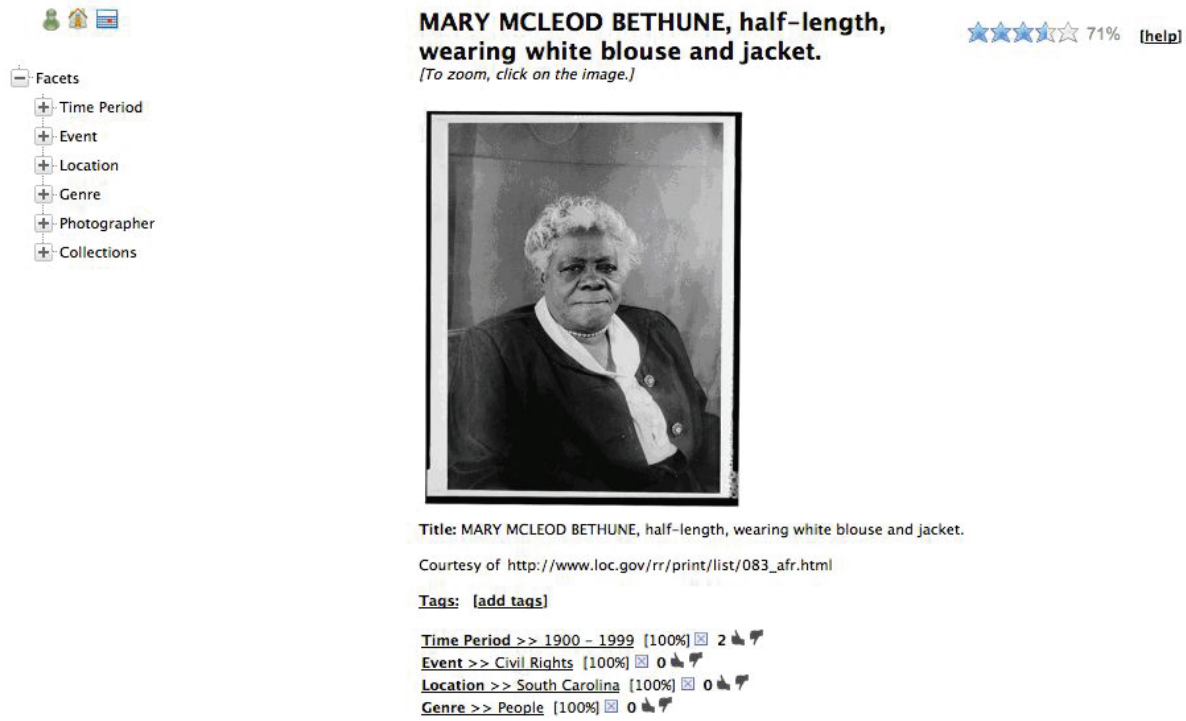

Figure 2. Classification screen of the Facet System.

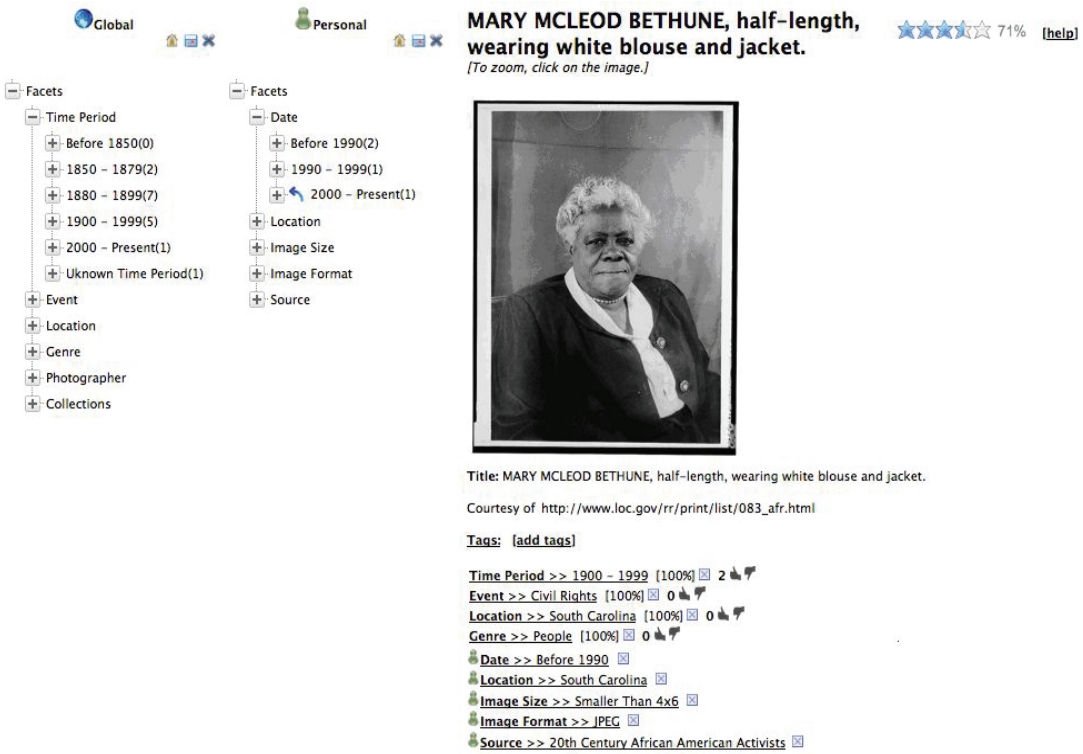

Figure 3. Personal and Global schemas in the Facet System.

While one of the users' favorite features based on user evaluation, the History feature requires significant database storage. Just as data has a time dimension in a data warehouse, the metadata has a time dimension in the Facet database. Therefore we refer to the metadata storage as a Metadata warehouse. In Fig. 2 and Fig. 3, clicking on the Calendar icon will allow users to see the schema at a given point in time. The History feature shows users the evolution of the schema and the trends in the community. Without the History feature, users may have difficulty finding documents at the "place" where they used to be, as the schemas evolve over time [10].

III. EVALUATION AND SCALING ISSUE

We have evaluated the Facet System for over a year with over 300 students at the Old Dominion University and the University of Delaware, and will continue the evaluation in collaboration with additional universities. We have developed image harvesting programs that can ingest thousands of public domain images per day on a given topic [11]. We have tested the system by simulating a large number of users. The scaling issue proves to be a critical factor in expanding the evaluation and deploying our system for public use in a multimedia document repository.

Traditional computing cannot support ever-increasing number of users, associated personal schemas, schema history logging, schema enrichment, and automated classification processes. With traditional computing, resources are typically configured rigidly with respect to both hardware and software (including licenses) to handle expected usage for a fairly short time horizon. Scaling up the configuration is a non-trivial effort and in many case literally impossible within the infrastructure of the supporting organization due to interdependencies. Cloud computing, on the other hand, enables our system to scale to virtually unlimited number of users. Enabling a collaborative document organization system to support virtually unlimited users may lead to a breakthrough in the way that electronic documents are organized. Our long-term vision is that this cloud-based document-organization approach may go beyond organizing an online multimedia collection to organizing knowledge bases in a large enterprise or a global research community. Users can store and organize documents in a computing cloud instead of on their desktop computers or departmental file servers. Besides maintaining virtual, personal "file systems", users can collaboratively evolve community-wide document collections. The cloud not only eliminates the storage limitation of desktop computers and traditional file servers, but also reduces duplicate storage and allows for value-added services such as document version controls (as in Microsoft SharePoint).

We have explored Microsoft Windows Azure and the development fabric, and see great potential of deploying the Facet system on Windows Azure. As the primary purpose of our Facet system is to organize documents including multimedia, we feel that the Microsoft platform will facilitate adoption of the Facet system by research

 ISBN: 978-1-61208-105-2

communities, industry enterprises and the society. While the cloud computing infrastructure has been developed for several years, social and organizational support of cloud computing paradigms is still lacking. Deploying our system on Microsoft Windows Azure will open the door of evaluating the system to large enterprise environments in addition to online communities. Using Microsoft Windows Azure explores the potential of integrating our system with the Azure cloud storage, an upcoming popular choice for enterprise file sharing, and Microsoft SharePoint service, a dominating player in today's document management market.

In the following section, we present the preliminary design of how our system is being deployed on the Microsoft Windows Azure cloud. We believe that migrating our system from an open-source stack to Microsoft Windows Azure will provide valuable lessons to other similar efforts in the research community.

IV. CLOUD DEPLOYMENT ARCHITECTURE

Our system is currently implemented as an open source extension of Joomla [15], a popular content management system built on LAMP stack: Linux [16], Apache[17], MySQL[18], and PHP[19]. We are now deploying the system along with PHP and MySQL on Windows Azure. To achieve scalability, we are moving metadata storage from MySQL to SQL Azure, and document storage from the Web server's file system to Windows Azure Storage. To minimize code changes and demonstrate the modularity of our system, we are continuing to use MySQL to support the core user interface features provided by Joomla!, such as authentication and menu management. Our system has a modular design to facilitate integration with other content management systems. If time permits, we will explore the potential of integrating our system with Microsoft SharePoint services on the Azure platform. Such integration would eliminate the need for MySQL.

The virtual machines in Windows Azure take either Web roles or Worker roles. The Web Role instances run our user-facing Facet System, which is programmed in PHP. The Worker Role instances run the MySQL database and back-end schema enrichment and classification programs in Java. Both Web Role and Worker Role instances connect to SQL Azure to access the metadata warehouse, and utilize Windows Azure Storage for multimedia file storage. Fig. 4 provides a general overview of the virtual machines (VMs) in the Azure cloud, adapted from a Microsoft white paper [12]. Windows Azure Storage and SQL Azure are part of the cloud utilized by the Web Role and Worker Role instances.

In our deployment there are different Web Roles and Work Roles. There are two Web Roles: FacetUI and FacetAdmin. The FacetUI instances serve the end-user interface to the Facet system. The FacetAdmin role contains administrative tools (such as PHPMyAdmin) that administer the database and caches. There are three Work Roles: MySQL, MemCached, and FacetBackend. The MySQL instances host the MySQL database that supports user authentication, menu management and other core-Joomla features. Master-slave configuration of MySQL databases

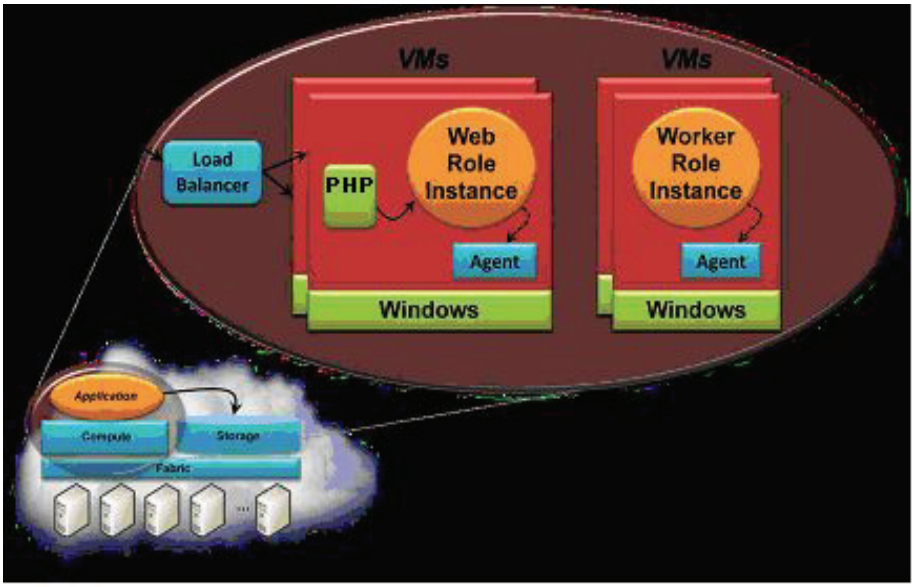

Figure 4. Overview of the Azure Cloud where the Facet System will be deployed.

 ISBN: 978-1-61208-105-2

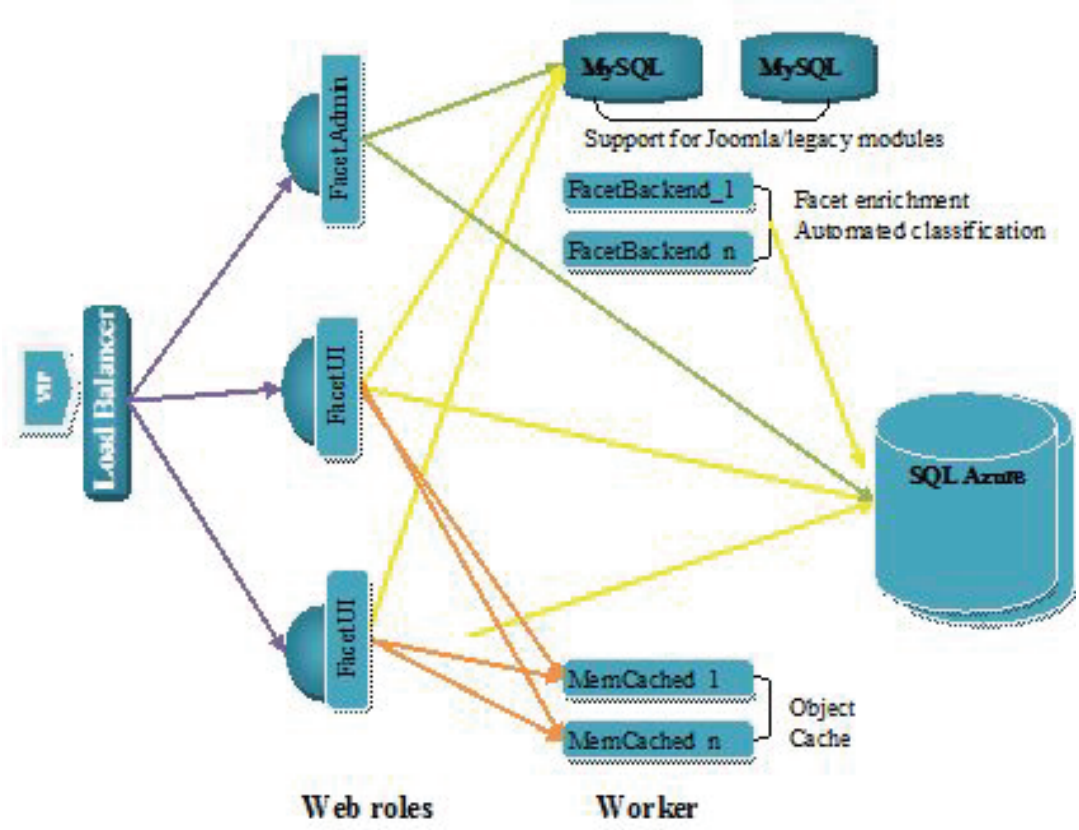

Figure 5. Architecture of proposed deployment on Windows Azure.

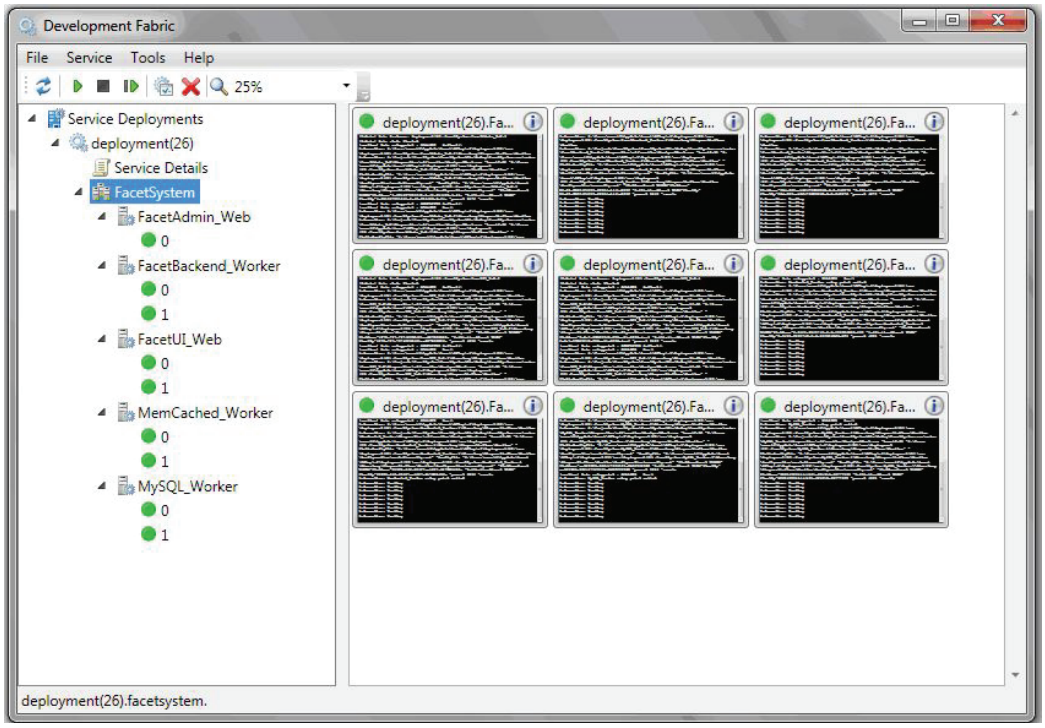

Figure 6. Deployment of Facet System on Azure Development

provides added reliability. The MemCached instances host Memcached, a popular distributed object cache system. A FacetUI instance attempts to read data from the object cache first. If the object does not exist in the cache, the FacetUI instance reads from the database, and load the object into the cache. Many common user requests, such as request to common menu items and the global metadata, are being served by Memcached to lower the database load and improve the performance. The FacetBackend role contains systematic schema enrichment and automated classification algorithms, which operate with data stored in SQL Azure. Deployment of these Web and Worker Roles will utilize Microsoft Windows Azure Solution Accelerators for PHP, MySQL, and Memcached [13]. Fig. 5 shows the architecture for the cloud deployment.

We spent limited concept-proving efforts in exploring the Azure development fabric. Fig. 6 shows the deployment fabric configured with different Web and Worker instances.

V. FUTURE WORK

On the user-oriented side of the system, we will address issues that come with the large scale, such as how to manage a large number of personal schemas and historic views of metadata. Among other adjustments, the user interface needs to be made stateless for load balancing and cloud deployment. On the back-end, we will address scalability issues of schema enrichment (mainly clustering and association mining) and automated classification (current implemented using Support Vector Machine) algorithms. We will evaluate various aspects of system functionality, including both user interface and backend algorithms. In parallel with code changes, we will develop a large test bed (an expansion of the current collection of African-American history images) that allows us to test the scalability of the system. As we gain experience with the large-scale deployment on Microsoft Windows Azure, we will explore design alternatives and make improvements to the system architecture, user interface and backend algorithms.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0713290.

REFERENCES

[1] G. Arnaout, K. Maly, M. Mektesheva, H. Wu, and M. Zubair, “Exploring Historical Image Collections with Collaborative Faceted Classification,” Proceedings of the Digital Humanities 2008, pp. 44-47, Oulu, Finland, June 2008.

[2] H. Wu, K. Maly, and M. Zubair, “Collaborative Faceted Classification System,” the 17th Annual Workshop on Information Technologies and Systems (WITS’07), Montreal, Canada, December 2007, pp. 237-248.

[3] H. Wu, M. Zubair, and K. Maly, “Collaborative Classification of Growing Collections with Evolving Facets,” Proceedings of the ACM 17th Conference on Hypertext and Hypermedia (Hyptertext’07), pp.167-170, Manchester, UK, September 2007.

[4] K. Maly, H. Wu, M. Zubair, and V. Antonov, “Automated Support for a Collaborative System to Organize a Collection using Facets,“ 13th International Conference on Electronic Publishing, Milan, Italy, June 2009, pp. 187-203.

[5] H. Wu, M. Zubair, and K. Maly, “Harvesting Social Knowledge from Folksonomies,” Proceedings of the ACM 17th Conference on Hypertext and Hypermedia, Odense, Demark, 2006, pp. 110-115.

[6] H. Wu, K. Maly, and M. Zubair, “Maintaining and Evolving a Taxonomy with Social Tagging,” INFORMS, Washington, D.C., 2008.

[7] H. Wu, K. Maly, and M. Zubair, “Supporting Multi-Criteria Decision Making through Collaborative Faceted Classification,” 20th International Conference on Multi Criteria Decision Making, Chengdu, China, 2009, CD-ROM.

[8] K. Maly, M. Zubair, and H. Wu, “A Collaborative Faceted Categorization System – User Interactions,” 14th International Conference on Electronic Publishing, Helsinki, Finland, 2010, CD-ROM.

[9] H. Wu and M.D. Gordon, “From Social Tagging to Social Hierarchies: Sharing Deeper Structural Knowledge in Web 2.0,” Communications of the Association for Information Systems Vol. 24, Article 45, 2009.

[10] H. Wu and M. Gordon, “Collaborative Structuring: Organizing Document Repositories Effectively and Efficiently,” Communications of the ACM 50, July 2007, pp. 86-91.

[11] L. Fu, K. Maly, M. Zubair, and H. Wu, “Building Dynamic Image Collections from Internet,” Digital Humanities 2010, 2010, CD-ROM.

[12] D. Chappell, “Introducing the Windows Azure Platform,” http://www.microsoft.com/windowsazure/whitepapers/, 2009.

[13] M. Srivastava, and T. Shanbhag, “Developing PHP and MySQL Applications with Windows Azure,” Microsoft Professional Developers Conference, 2009.

[14] Microsoft Windows Azure platform. http://www.microsoft.com/windowsazure (Last accessed 2010)

[15] Joomla. www.joomla.org (Last accessed 08/16/2010)

[16] Linux. http://www.linux.org/ (Last accessed 08/16/2010)

[17] Apache. http://www.apache.org/ (Last accessed 08/16/2010)

[18] MySQL. http://www.mysql.com/ (Last accessed 08/16/2010)

[19] PHP. http://www.php.net/ (Last accessed 08/16/2010)

 ISBN: 978-1-61208-105-2