Cloud Based Document Delivery Service for Emerging Markets

Sakhee Dheer
Asankhaya Sharma

July 2010

International Conference on Services in Emerging Markets (ICSEM) 2010

Cloud Based Document Delivery Service for Emerging Markets

Sakhee Dheer

Asankhaya Sharma

Abstract — In this paper we propose a novel framework for providing Document Delivery Service to educational institutions in emerging markets. Many of these educational institutions do not have the means to subscribe to the wide variety of the publications and scientific journals available (via IEEE, ACM and other digital libraries). We show that our proposed service is an easy to implement framework which leverages cloud technologies to enable access to expensive journals and publications to students of emerging markets. We explore and discuss a web proxy based implementation of our service using Windows Azure and how we can help reduce the cost of access for document delivery in emerging markets. Through this framework we can help almost 5 million students in India alone to get access to the vast repository of research findings and related publications.

Keywords- service design engineering; IT service, education, cloud technologies;

I. INTRODUCTION

With the increasing cost of possessing research journals, publications and academic articles, the challenge of how to balance financial costs and the availability of these documents for students is a daunting task for the educational institutions. It's very important to provide students in professional colleges with the latest discoveries and research data in their field of specialization. There are close to 9000 colleges in India [11] alone with very few of them having access to the digital libraries.

The institutional membership to IEEE/ACM digital libraries is very expensive. Most of the colleges in emerging markets have the bare minimum infrastructure and possessing an up to date digital library is a farfetched reality. In this paper we will be exploring a Cloud Based Document Delivery Service framework to make documents in well-known IEEE/ACM digital library economically available to these educational institutions. We will show how we can federate access for a bunch of institutions together using a cloud-based service. The Document Delivery Service owner will pay the cost of institutional membership of the ACM/IEEE document repositories while the participating institutions will pay for using the Document Delivery Service. This will help in reducing the cost for the educational institutes.

The imbalance between the demand of research articles, journals and publications and supply of the same due to economic constraints can be overcome by the framework proposed by us. Our framework will cover all concerns from selection of the articles (document selection), making the request to deliver the article (document request), retrieval of the articles requested for (document retrieval) and finally transferring the articles back to from where the request was made(document transfer). We show that by bringing together educational institutions we can solve the imbalance between demand and supply of research articles.

The paper is organized as follows, in section 2 we will deal with related work and background. In section 3 we will describe the generic framework of the service we propose, section 4 we will discuss the implementation in detail, section 5 will deal with results and discussion and finally we will end with our conclusions and future work.

II. BACKGROUND AND RELATED WORK

Classically document delivery is considered as a form of library service [1]. It is the transfer of document from one source to another. The sources may be two libraries or a library and an end user. Thus, document delivery can be broadly categorized as the process of providing document to an end user.

Most of the libraries have traditionally offered a Document Delivery Service [1]. They usually provide delivery services to researchers from other institutions on a request. The researchers may send a written request to the librarian giving full bibliographic details. This service is provided on cost basis. More recently we have seen them in their digital avatar as the subscription services provided by IEEE/ACM in form of IEEE Xplore and ACM digital library. The basic principles are the same just that now the document request and delivery is all electronic and over the internet.

Suchitra Patnaik [2] explains various facets of document delivery in a digital environment in an end to end scenario in an emerging market. This details the work which was to be done to digitize books, documents, manuscripts, doctoral

dissertations etc. However, the challenges and obstacles encountered to digitize libraries and provide easy access online are not the area of work that we are exploring. Their research mainly deals with creating digital libraries; however what we propose is providing easy and economic access to existing digital libraries (like ACM/IEEE's digital library).

Higher education plays a major role in the development of an emerging market and thus has been an active area of research. Microsoft released Windows Multipoint Server 2010 for schools in emerging markets that don't have the money to buy a PC for every student [3]. HP has launched their Multi-Seat Computing Solution for schools in emerging markets [3]. As far as we know still not much has been done in the field of higher education to enable students to access vast repositories of publications and journals online.

Microsoft Research India is conducting a number of research projects to tackle problems of developing countries. Research is being done to explore mobile centric internet usage and how low-income communities can benefit from them [4]. This is being extended to provide mobile phone enabled banking and payment [5]. These studies help us to understand the needs and requirements of students coming from low-income households. They throw light on the social and economic context of the emerging countries. Our work is motivated by the social and economic needs which are specific to an emerging market like India and may have being overlooked elsewhere.

These projects provide a good background for our proposed framework as they also target a specific section of population in the emerging regions, people those who possess mobiles [5]. Similarly, our research is also predominantly aimed at students in educational institutions who are not able to access academic journals due to economic constraints but still have means of access to a computer via the university computer center or a personal computer at home. It was helpful for us to keep these social economic models in mind when evolving our framework in the design phase. Another project by Microsoft Research [6] makes an attempt to understand the user interface (UI) and user experience (UX) design requirements for people in emerging markets [6]. However, their work is targeted at the rural segment, and it did not provide us much guidance on how to design for some student in higher education (say a private engineering college). We assume that the students studying in such institutions know how to use computers for basic tasks like word processing and web browsing.

A recent attempt has been made by Nithya Sambasivan et al [7], in the use of intermediated technology for developing communities. This research is notable for its effort in increasing the reach of technology in areas where it has been inaccessible due to lack of technology-operation skills, non-literacy and financial constraints etc. This research helped us understand context of economic concerns faced in such communities.

Windows Azure is a cloud computing platform that allows developers to build applications in an Internet-accessible virtual environment which is hosted at Microsoft data centers. We plan to leverage Window's Azure on demand application instance where we can host the web application in Microsoft's datacenters without having to worry about its supportability and maintenance.

The deployment basics of services in cloud as explained by Thomas, Arman and Herjorn [8] clearly explain how we can compose and host services on the cloud. Windows Azure is an obvious choice for its ease to use and interoperability. Azure has various storage options like tables, queues and blobs. The tables in Azure are non-relational and implemented as named entity pairs. On the other hand, SQL Azure is a relational database in cloud which extends the capacity of Windows Azure to include relational storage and implement some of the classical relational architectural patterns [10].

As far as we know, the Cloud Based Document Delivery Service for the students in emerging markets is the first attempt to explore the use of cloud technologies in this context. We propose to leverage cloud computing to make the Document Delivery Service website accessible to all the educational institutions without them having to care about its supportability in their environment in any way [11].

III. PROPOSED SERVICE

In this section we describe a formal framework for the Document Delivery Service. The key goal of this framework is to be able to use existing digital libraries (such as those of ACM/IEEE) at a low cost by using cloud technologies like Azure.

Figure 1

We begin with a description of various components involved in document delivery paradigm which are illustrated in the figure 1.

a. Requestors
b. Documents Sources and Types
c. Retrieval Process

A. Requestors

Our service is aimed at students of higher education institutions in emerging markets, so the requestors are these students. Any user of the service who wishes to access and download the document is the requestor.

B. *Document Sources and Types*

The document source for the current service framework is the existing digital libraries like IEEE, ACM etc…Going forward however, community or city libraries can develop their own digital repository which can be seamlessly on-boarded to the Document Delivery Service. The service is mainly targeted at students and the digital libraries we propose to integrate are libraries with academic documents, research journals and distinguished articles.

C. *Retrieval Process*

The retrieval process can further be divided into the below generic components:-

a. Document Selection

Students will be given access to the Document Delivery Service website, where they will be authenticated based on the username and password entered. Once the students are authenticated they will enter the link of the document requested.

The students will have to use the internet and available web browsers to search for the document they desire. Once they have the link they can directly request for it from the Document Delivery Service Website.

b. Request and Retrieval

Once a student has selected the document, the Document Delivery Service will request the digital library for the document referred by the link entered in the website. Digital libraries like IEEE and ACM already have the procedure of retrieval built in. The digital library retrieves the document and sends it back to the requestor.

The actual implementation of this service is based on Windows Azure. In the next section we will describe how a document will actually be fetched and delivered to the requestor using web roles and worker processes which are built in Azure.

IV. IMPLEMENTATION

We propose to implement the Document Delivery Service using cloud technologies. More specifically we compose the service using Azure Services Platform. Some constraints on the service design also stem from the fact that we need to be able to access already existing document repositories like those provided by the digital library of ACM/IEEE etc. It may not be clear at first why do we impose such restriction on our implementation. The following discussion will detail the reasons as to why our design is constrained by the already existing digital library services (such as those provided by IEEE/ACM).

One of the main motivations for use of our service is that it will provide easy access to a variety of publications to students of those institutions which have no means to provide such access. For most of the engineering and technology related areas the digital library provided by ACM/IEEE is the most comprehensive and well known. It is imperative that we as a first step at least enable the document delivery for those set of publications residing in these repositories. Both ACM and IEEE provide institutional membership for academic institutions (or a consortium of such institutions) which grants them unrestricted access to all the documents in the respective digital libraries. To make access matter simple the way this is enabled is by use of certain IP Addresses which the digital library recognizes as registered for unrestricted access and lets users download documents without any fee. As one can imagine this provides an easy and ubiquitous way for members to use the institutional subscription to access the library. Since all of them go through the common network proxy to whose address is typically given to ACM/IEEE while registering for institutional membership. This enforces a design constraint on our service composition. Since access to these document repositories is controlled using external IP address we need our service to somehow allow access federation across a group of educational institutions.

At first we can take the simplest approach and envision a network proxy in the cloud which can be used by students to connect through and get access to documents. This approach will work well within a single institution where everyone connects through the same network proxy. A problem which could arise in the implementation of this service could be the issue of privacy as more and more educational institutions are grouped together. All participants of the consortium may not be comfortable browsing the internet and downloading all data through a common network proxy just to access the Document Delivery Service. The alternative then would be to create a VPN to let the users come and access the service in a secure manner.

However, it is not clear how one could achieve such a thing (VPN/Network Proxy) in a truly PAAS (Platform as a Service) based platform like Windows Azure. We did not explore the possibility of a network proxy based implementation using a more IAAS (Infrastructure as a Service) like platform say Amazon EC2. Instead of that we actually choose to base our implementation using a web based proxy. When using a web based proxy many of the above issues are circumvented and we can provide the implementation using artifacts from the Azure Service Platform. At this point it would be appropriate to give a brief introduction of Windows Azure Platform and then we will show how we implement our service on Azure.

As shown in the Figure 2 the Azure Services Platform consists of three different services- Windows Azure, SQL Azure and the Windows Azure AppFabric. Each of these services provide different aspects of the platform. Windows Azure provides computation and hosting services, SQL Azure is a transactional database in the cloud and AppFabric provides services for connecting applications in cloud and on premise.

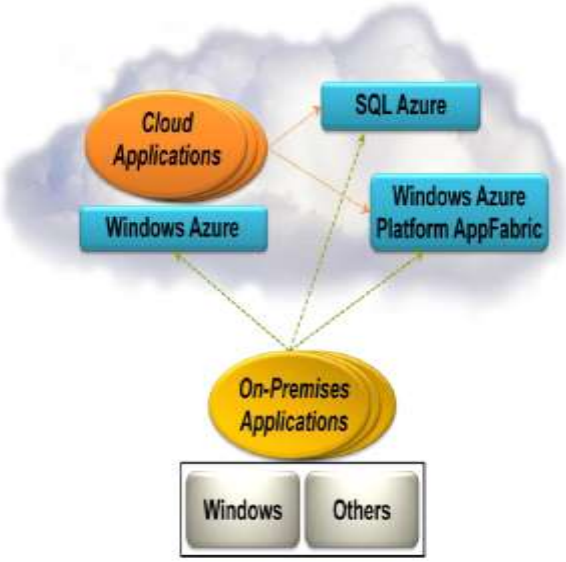


Figure 2

The compute service provides two kinds of instances – web role and worker role (as illustrated in Figure 3). A web role is essentially an IIS hosted Website or Web Service while a worker role provides means of background computation. Along with this we get three different kinds of storage – tables, queues and blobs. This storage is different from SQL Azure service which actually is a fully transactional database in cloud.

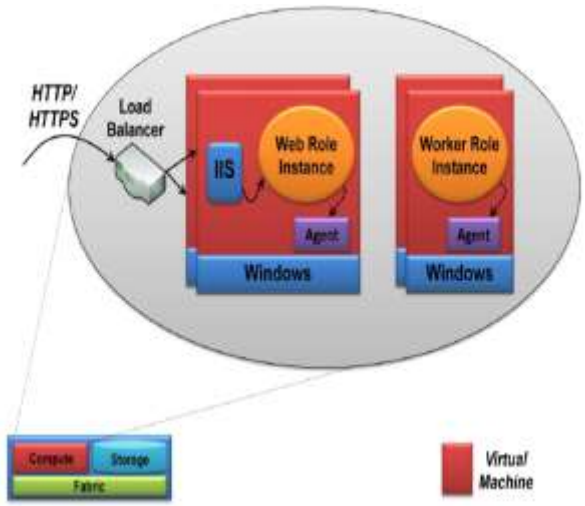


Figure 3

Based on these features and our service design contraints, we propose the Document Delivery Service to be implemented via the following Azure design pattern - *Scalable Web Application with Background Processing*. This cloud pattern (Figure 4) provides most of the features needed by our Document Delivery Service and also takes care of the initial design contraints. In the following few paragraphs we describe the details of such an implementation.

The front end of the web proxy is hosted using some number of web role instances (each instance can be though as a VM that provides a fixed computation and storage capability). More than one instance is needed in order to support scalability in terms of the number of users trying to access the service. Each user is supposed to log in using a username/password combination; this is to restrict the access to only participating individuals who register for this service. The access control data can be stored as a table in storage (the actual format for this storage in Azure is a named value entity and not relational but that difference doesn't matter for our purposes) which authenticates a registered user and then enables him to use the web proxy for accessing documents. A logged in user can now issue web requests to get the required document, these requests are handled by our web proxy and put in a queue.

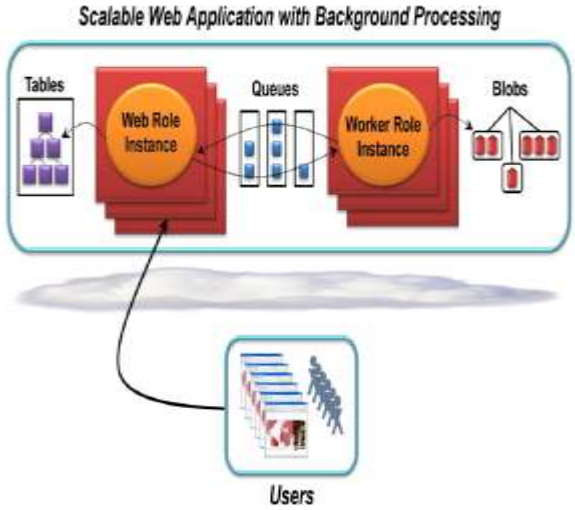


Figure 4

A worker role instance reads the queue of pending web requests for documents, retrieves the document and then the response is sent back to the web role to display to the user. Having a separate worker instance for the document retrieval serves two purposes, firstly it allows scalability with respect to the processing of the web requests and responses by the proxy and secondly it allows us to control the rate at which we send web requests to the document repositories. It is important so that we can maintain a fair usage of the resources on the document repository side as well. ACM and IEEE digital libraries impose a maximum concurrent usage restriction even on the institutional subscribers within the limits of fair usage. We propose to implement those limits by controlling the number of simultaneous web requests made by our proxy at any given time using the queue and worker role.

Once a web response is received typically in form of the required document (say a .pdf file) we cache this response and store it in from of the blob structure in Azure. So if next

time a request is made to the same document we do not have to forward the web request and we can just return the document from the cache. No attempt has been made to optimize this caching mechanism and for our purposes it just serves the case when the user hits refresh or somehow requests the same document again. We clear this cache periodically and at any given time only a fixed storage is used in form of blobs
The following shows the behavior exhibited by the web proxy based on the requests it gets –

a. If the request is for a white listed website like IEEE/ACM/Google etc. the proxy just forwards the request and gets the web response using queue and working role instance
b. For all other requests we just deny the request and show an error response to the user

This behavior ensures fair use of our system and makes sure that users don't come to use it just to access restricted sites in their network. We think this should not be a big inconvenience because we envision the usage of our Document Delivery Service in some sort of a controlled environment like that of a college computer center. Figure 5 depicts the typical proposed work flow of a requestor requesting a document through the Cloud Based Document Delivery Service. The boxes show the various implementation artifacts (Azure components) while the arrows show actions between the corresponding boxes. These actions are detailed below.

a. Authentication:

Each requestor is authenticated using the registered username and password provided during registration. Only after authentication the requestor is able to make requests through the web proxy to access documents.

b. Web Request:

A Web Request is made when an authenticated user uses the functionality embedded in the web proxy to request for documents from the digital libraries. This is done using the worker role which queues all the requests in an Azure queue.

c. Ensure Fair Use:

The process of document retrieval takes place through a worker role which will pick the requests from the queue and ensure that we do not unnecessarily bombard the digital libraries with sudden traffic. The queue based approach along with worker processes will help us ensure fair use and no negative impact on the digital libraries infrastructure.

Along with this a white list is maintained which will contain all the allowable domains, preferably only the links of the digital libraries which can be accessed using the cloud-based website.

d. Document Cache:

The document once retrieved will be cached in the Azure's blob storage. As of now we think of the cache as an instant backup in case of service or connection breakdown before the completion of delivery via web

e. Web Response:

This is the final delivery of the document to the requestor. The delivery will take place to the same web role, after all the above checks, from which the document was requested.

Our service is able to do document retrieval from IEEE/ACM digital libraries because we register the IP Address of the worker role instance for institutional membership with them. We ensure that we are fair to their system by enforcing checks in the number of requests that we make at any given time. This we hope will not be too limiting for the potential users of the service since they at present have no access to such information, besides most of the time we do not expect such high concurrent usage. We believe our implementation is able to address most of the issues in document delivery while keeping all the features of the proposed service. For the detailed component level diagram of our implementation and interaction between various layers, please refer Appendix.

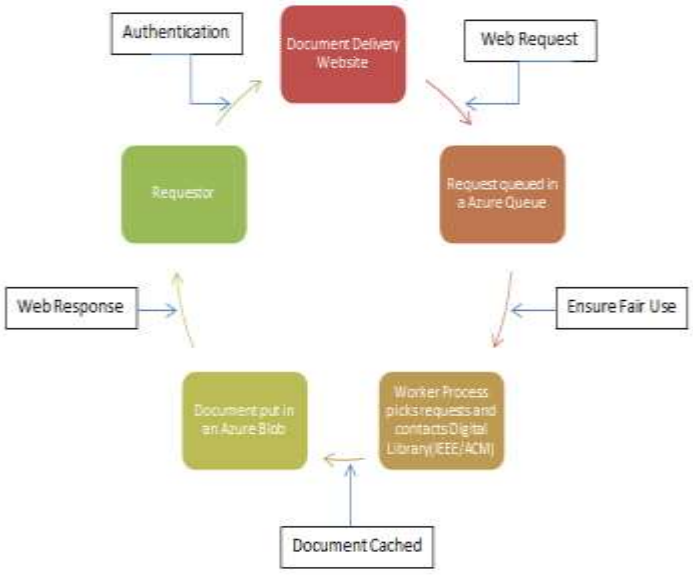


Figure 5

In the next section we will discuss some of the economic aspects of such an implementation and show how we are providing an economical solution to the emerging markets by leveraging the cloud technologies.

V. RESULTS AND DISCUSSION

In the following few paragraphs we try to discuss the economic benefit of such a service by taking typical membership costs of ACM and IEEE digital library as well as the cost of using Azure services. The calculations and results we reach may not be exact but we believe they are representative of the cost benefits one can get from such a service and make a good case for our proposed implementation.

Table I.

Sample of Consortia Subscription Savings

	# of Institutions	List Price
Tier 1	5	$2,000
Tier 2	3	$13,030
Tier 3	4	$14,208
Tier 4	2	$14,876
Tier 5	4	$15,165
Gov't	3	$15,165

The current membership cost [13] of IEEE Computer Society digital library is: - $17,595.The current membership cost [14] of ACM digital library ranges from $2000 to $15,165 depending on the size of the consortium. As is clear from Table I., the cost of membership is very high and way beyond the means of a non-funded college in an emerging market. Let us assume for the service we propose, to grant access to 10 colleges grouped together to access the digital libraries through a Document Delivery Service website.

As shown in the snippet (Figure 6) taken from Standard rates for Windows Azure the cost of compute instance for the website on the cloud [12] is fixed and would additionally depend on the number of instances we choose to run. For the purpose of this example (i.e. 10 colleges) one small compute instance will suffice which will just need around $80 per month. The service that we will be using to make our service framework functional would fall under 2 categories:-

a. Windows Azure Content Delivery Network(CDN)
b. Windows Azure Storage

Windows Azure Content Delivery Network costs $0.01/10k transactions. Windows Azure storage also has the same cost $0.01/10k transactions. For every request that has been made to fetch a document, there will be two steps:-

Standard Rates:

Windows Azure

- Compute
  - Small instance (default): $0.12 per hour
  - Medium instance: $0.24 per hour
  - Large instance: $0.48 per hour
  - Extra large instance: $0.96 per hour
- Storage
  - $0.15 per GB stored per month
  - $0.01 per 10,000 storage transactions
- Content Delivery Network (CDN)
  - $0.15 per GB for data transfers from European and North American locations*
  - $0.20 per GB for data transfers from other locations*
  - $0.01 per 10,000 transactions*

Figure 6

a. *Authentication of the Requestor*: This is done by confirming if the credentials of the user are present in the Windows Azure Table Structure using the registered username and password. This will make us liable to pay for the storage cost pertaining to Windows Azure.

   Per transaction Cost = $(0.01/10000)

b. *Document Delivery*: After the requestor is authorized, the request will be forwarded to the digital library; the document will be fetched and delivered. Depending on the location of the requestor, the Content Delivery Network costs will be added.

   Per transaction Cost = $(.01/10000)

Assuming 10 million transactions per month the cost is just $20 based on above calculations. Even after adding the $80 for compute instance we get a per month cost of $100 for running our service in Azure.

This gives a per college cost of just $10 per month or $120 per annum. Compared to the cost of subscribing to IEEE Xplore which is $17k per annum this cost is negligible. Hence each institution can now effectively access the IEEE digital library by just paying $1.7k per annum (the cost of using the cloud services being negligible).

As seen, the total cost to host the Document Delivery Service on the cloud is minimalistic and the cost of membership also comes down by a considerable margin with the Cloud Based Document Delivery Service. We believe that this service effectively enables the educational institutions to

get access to digital libraries which they would not have otherwise, due to economic constraints.

VI. Conclusion and Future work

We have described a framework for a Document Delivery Service which works on the available digital libraries with the main aim of making them available to more number of people. This service targets students in higher education institutions of developing regions which are not financially replete to partake individual membership of current digital libraries.

Currently, there are around 5 million students in Indian colleges. Apart from the elite 'A' class colleges, the others are low on funds and grants. They do not have enough resources to offer world class literature in different areas of specialization. They are also not able to acquire membership of the existing IEEE and ACM digital libraries which are very expensive. By bringing together 5-10 colleges and asking them to access the digital libraries through the Document Delivery Service we will be able to overcome most of the hurdles. The membership and the usage cost of the service will be split between the grouped institutions lowering the total cost to make it affordable for institutions.

Indeed, it seems that the idea of grouping together colleges for common subscription is gaining acceptance since ACM digital library has recently started providing membership to consortiums of institutions and the pricing is configured depending on the country's economic standing [14]. This is similar to the spirit of the service framework we propose. ACM's pricing strategy along with the Cloud Based Document Delivery Service will further increase the willingness of institutions to enroll in institution memberships for their students.

An area of future work is where we can provide the content of digital libraries as an Azure 'Data Set', which students can easily access without having to go via a website. Currently many such datasets are available from the Azure Marketplace such as the Mars Landing Dataset from NASA. The Document Delivery Service can be extended to a library so as to provide easy online access to the entire library's collection of journals and papers. Further, an ambitious implementation would be where we can also try to extend this set up for all the books in a library. This will require the creation of a digital library and to tackle obstacles such as IP and copyrights conflicts [2].

An important future work in this regards is to be able use this framework with research findings for rural areas in emerging markets [6] and be able to design a solution for accessing important political, social and policy documents. In this case as well, a Cloud Based Document Delivery Service (for rural areas in emerging markets) should be able to provide easy access to information on new laws, government writs, circulars, amendments and documents of importance to them.

References

[1] JNCASR, Document Delivery Service, http://www.jncasr.ac.in/library/workon/document_deliv.php

[2] Suchitra Patnaik, "Document delivery system in digital environment", in Proceedings of DRTC Workshop on Information Management, Jan 6-9, 1999

[3] Sumner Lemon, "Microsoft, HP Stretch Education Budgets in Emerging Markets", in PC World, March 9, 2010

[4] Shikoh Gitau, Gary Marsden, and Jonathan Donner, " Challenges facing mobile-only internet users in the developing world", in Proceedings of the 28th international conference on human factors in computing systems (CHI 2010), Association for Computing Machinery, Inc., 16 April 2010

[5] Ivatury, Gautam and Mark Pickens, "Mobile Phone Banking and Low-Income Customers: Evidence from South Africa.", in Proceedings of CGAP, 2006.

[6] Medhi, I., Ratan, A. and Toyama, K., "Mobile-Banking Adoption and Usage by Low-Literate, Low-Income Users in the Developing World", in Proc. Human Computer Interaction International, San Diego, USA, 2009

[7] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi, "Intermediated Technology Use in Developing Communities", in Proc. CHI 2010: HCI and the Developing World April 10–15, 2010, Atlanta, GA, USA

[8] Thomas Erl, Arman Kurtagic and Herbjörn Wilhelmsen, "Designing Services for Windows Azure", in MSDN Magazine, Jan 2010

[9] David Chappel, "Introducing the Azure Platform", in Azure Documentation, http://www.microsoft.com/windowsazure

[10] "Architectural Strategies for Cloud Computing", in Oracle White Paper in Enterprise Architecture, Aug 2009

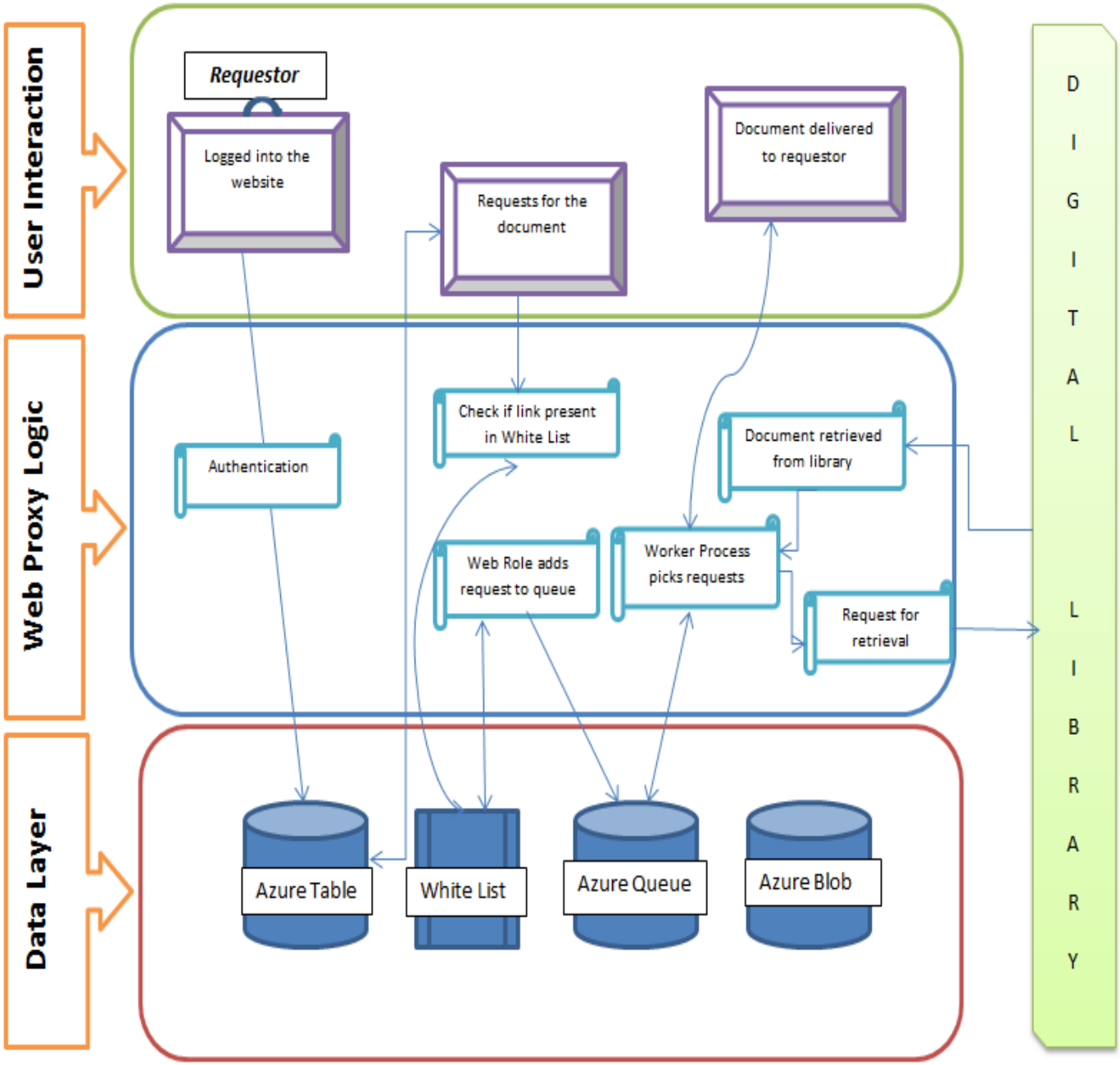
[11] Statistical information of Colleges & Universities in India, http://www.indiastudycenter.com/Univ/College-Statistics.asp

[12] Windows Azure Pricing Overview, http://www.microsoft.com/windowsazure/pricing/

[13] IEEE Computer Society Digital Library (CSDL) Pricing Details, http://www.ieee.org/publications_standards/publications/subscriptions

[14] ACM Digital Library Pricing Details, http://libraries.acm.org/academic/pricing

APPENDIX



Component Level Diagram of Document Delivery Service