International Journal of Advanced Science and Technology
Vol.59, (2013), pp.113-122
http://dx.doi.org/10.14257/ijast.2013.59.09

# Information Retrieval Architecture for Heterogeneous Big Data on Situation Awareness

Jongwook Woo

*Computer Information Systems Department*
*California State University Los Angeles*
*jwoo5@calstatela.edu*

***Abstract***

*The paper proposes the model of the data integrations systems with which the buyer/user searches and aggregates organizational entity from multiple heterogeneous data sources in order to have a contract with the proper entity for the new project. The requirement for the systems is to retrieve and integrate heterogeneous data which is non- or semi-structured. And, the systems need to be alerted when the data is changed. The proposed model supports services to retrieve and integrate an entity(s) and it also alerts the update of the entities to the user. Besides, the systems store and compute Big Data using Hadoop platform.*

***Keywords:*** *information integration, heterogeneous data sources, situation awareness, information retrieval, Big Data*

## 1. Introduction

It is not easy to find out the proper IT solution providers, especially for government projects as people has manually researched and aggregated the profiles of the possible business contractors in order to address the possible IT solutions. The profile needs to include the relevant business licenses, organizations' structures, and data related to Air Force SEP, ISO 9000, CMMI (Capability Maturity Model Integration), and Agile. However, the legacy approach involves manual search, manual aggregation, and the data must be constantly refreshed, which is expensive, labor intensive and time consuming processes. As web and its n-tier architecture has been popular in the world, we could build a web application that can continuously search for information sources, aggregate and rank the results in a persistent database, and alert the users the change of the interesting entities at SA (Situation Awareness) situation to provide an up-to-date profile of organizational entities. The web solution will reduce the labor intense and avoid government over-commitment and abandonment of processes during an emergency, quick-response situation.

The paper illustrates the possible systems design and approach to implement this web application that provides built-in functions to retrieve, aggregate, manage data and profile qualifications for IT solutions.

In order to measure the qualification of an entity to have a contract, we can use the Air Force System Engineering Process (SEP), ISO 9000, CMMI, and Agile. However, it is labor intensive to search and aggregate data from heterogeneous data sources to asses an organizational entity. Furthermore, the heterogeneous data sources can be both structured (Database, XML and JSON files) and non-structured (raw data files and web). Besides, data from the data sources is large scale data that is more than tera- or peta-bytes. It also becomes worse when the user monitors the status of the interesting entities continuously over time, especially at the moment of situation awareness. Therefore, we need a web application that can continu-

ISSN: 2005-4238 IJAST

ously search information sources, and aggregate and rank the results in a persistent database to provide an up-to-date profile of organizational entities.

The paper is composed of the following sections. Section II illustrates related works. Section III presents background such as system engineering, information integration and retrieval. Section IV shows the proposed model. And, finally, section V is the conclusion

## 2. Related Work

Naveen *et al.*, [31] implement Ebox system that provides integrated access to multiple heterogeneous data sources relevant to providing situational awareness during emergency response situations. The user of the system can receive the useful information from the multiple sources at the emergency situation.

Woo *et al.*, [16-18] design an information integration model on N-tier architecture with a global XML schema for a specific domain, which is a format that each heterogeneous data source generates XML data to be migrated to a global data source. Woo *et al.*, [25, 26] also illustrates and presents a possible e-Business architecture integrated with an enterprise search engine.

Woo *et al.*, proposes Market Basket Analysis algorithms using Hadoop and HBase. The algorithms are built in MapReduce that runs on Hadoop parallel systems with multiple nodes [11-14].

## 3. Background

Situation Awareness (SA) is defined as "the ability to identify, process, and comprehend the critical elements of information about what is happening to the team with regards to the mission" [30]. In this paper, SA is defined as the situation that the user is alerted when the profiles of the interesting organizational entities are changed. The background of the processes can be described in the following. First, system engineering approaches - CMMI, SEP, and ISO 9000 - are presented. Second, data migration and integration concepts are explained. Finally, data retrieval technologies for structured and non-structured data are illustrated.

### 3.1. System Engineering Approaches

There is a standard to measure the qualification of the enterprises that can provide IT solutions. There several system engineering approaches such as CMMI, SEP, and ISO 9000, which are used to profile the qualifications of Small Business and Large Business Contractors for implementing IT solutions.

System Engineering Process (SEP) is a process that applies system engineering techniques - technical management and problem-solving processes - to develop all kinds of systems. SEP has a meta model that is composed of four processes: Agreement, Project, Technical and Evaluation. Agreement process is to establish an agreement with the customer, in order to build a new system. Project process is to plan the project and be modified during technical process. Technical process designs and develops the project. Finally, evaluation process is to repeatedly check out whether the requirements are met, valid, and consistent [5, 6]. DoD system engineering is composed of 8 Technical Management Processes and 8 Management Processes. Technical Management Processes are equivalent to the Systems Analysis and Control portion to support and control the application of the Technical Processes during system development in order to meet program or project objectives [6].

 

International Standard Organization (ISO) 9000 addresses Quality Management families that consist of standards and guidelines relating to quality management systems and related supporting standards. The families fulfill the customer's quality requirements and applicable regulatory requirements while meeting customer satisfaction and achieving continual improvement of its performance in pursuit of these objectives. ISO 9001 is one of the families, which illustrates the requirements of quality management system. Proper quality management improves business, create a more efficient, effective operation, enhance marketing, and increase customer satisfaction. ISO 9001 sometimes makes people concern because of the amount of money, time and paperwork required for its registration [7, 8].

Capability Maturity Model Integration (CMMI) is a process improvement approach that provides organizations with the essential elements for effective process improvement in software engineering and organizational development. Thus, it guides process improvement across a project, a division, or an entire organization. An enterprise can be appraised not certified in CMMI so that it can be awarded a maturity level rating (1-5) or a capability level maturity profile. By the appraisal, the external customers and suppliers can be informed how well the organization’s processes to CMMI best practices. Besides, customers can use the appraisal for the contractual requirements [1-3]. We can refer to the web site of Software Engineering Institute at Carnegie Mellon University, which lists the appraisal results of many organizations [4].

Agile provides rapid results and more frequently visible business benefits with light and adaptive iterations so that it has been adopted in software engineering project before the web [32]. But, Agile has become popular lately to work with iterations and sprint teams. A section of a project needs to be delivered successfully to the next sprint team. And, if not, the section should be iteratively retested and updated before shipped.

Lately, people adopt both CMMI and Agile together as CMMI focuses on what the project does and Agile focuses on how to develop the project, which can complement each other. CMMI enables an agile team to give the team a system engineering practices for a large project. And, Agile provides a method to develop the project, which is missing at CMMI [32].

### 3.2. Data Migration and Integration

In order to integrate data from multiple heterogeneous data sources, we need two processes: data migration and data integration. Data migration is to retrieve and collect data from the data sources and store them into the third data source in a format specified. There are many issues to resolve for it. First, we must know the data sources. If data sources are at the same organization, we should connect data sources with data source connection information by a security authentication and authorization. If data sources are at the different organizations, it will become more complicated. Thus, people contract and share data as feed data such as in RDF and RSS format, which are either XML or Json data. Otherwise, we need to search the web sites of related data, collect the relevant data, and transform them to the desired format [17-19].

After data is collected, we need to integrate data into a DB, that is, global integrated DB, by converting them to the proper format of the DB. Data integration processes are: first, to check if each data exists in the DB and then update data; second, to remove or combine duplicated data from the heterogeneous data [17, 18]. If the data is a large scale data that is more than tera- or peta-bytes, Hadoop big data systems can be used to store and compute Big Data. When the data is updated or deleted, SA situation occurs.

It can be simply implemented with email and Short Message Service (SMS) that sends text message to phone.

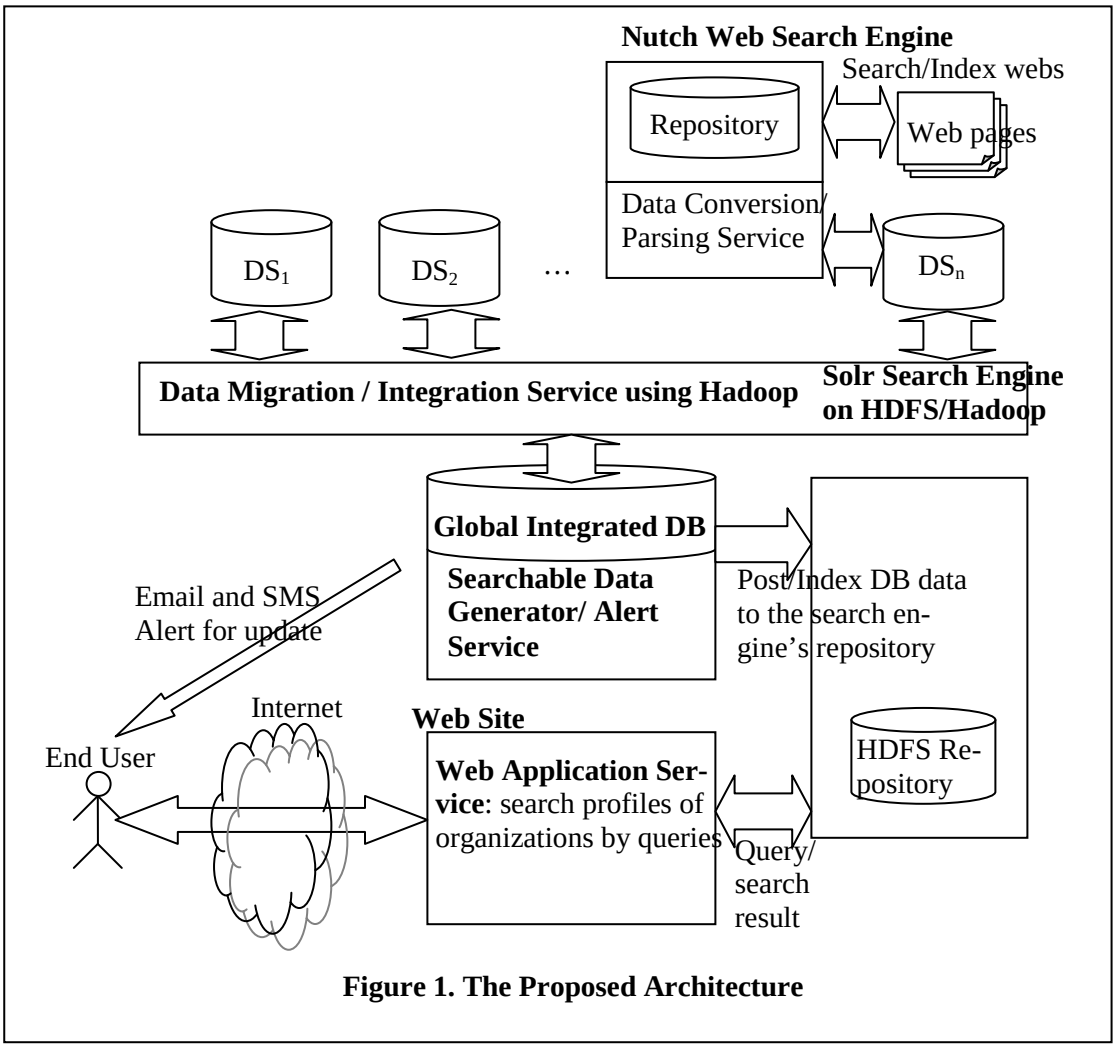


**Figure 1. The Proposed Architecture**

### 3.3. Search Engines

We can use enterprise search engine that searches local DBs, for example, Solr and Fast. After searching the data sources, the engine indexes data specified and stores indexed data into the repository of the engine. Another is web search engine that searches web sites.

Lucene is one of open source apache software projects supported by the Apache Software Foundation. It is the text search engine library in Java [19]. Solr is an open source enterprise search engine on the Lucene APIs. Solr runs on servlet engine and has many features such as hit highlighting, faceted search, caching, replication, and a web administration interface with XML/HTTP and JSON APIs [20]. FAST [23] is an enterprise search engine and a product of Microsoft. It supports geographical search function as well as all features of Solr. LocalLucene is to support geographical search capabilities to Lucene. LocalSolr provides the geographical search capabilities by using Local-Lucene APIs [21, 22].

 

As data becomes large-scale data lately, that is, Big Data, Solr search engine is integrated with Hadoop platform that can store and compute Big Data on HDFS and Hadoop core functions. Thus, we can even implement the architecture using Hadoop platform for Big Data [33].

## 4. Proposed Model

### 4.1. Objectives

The paper proposes the model that searches and aggregates organizational entity from multiple heterogeneous data sources. Besides, when the profiles of interesting entities are updated, the user should be alerted. The model needs several processes: 1) selecting the data sources that have the profiles of entities, 2) migrating and integrating data from the data sources, 3) searching for the proper entities over the integrated data source, and 4) the user can keep the interesting entities and be alerted by email and SMS when the entities are updated. We assume that we have the information of the data and data sources that we can access. Thus, the selecting process 1) is based on the algorithm and use cases that follow the way to research and aggregate data manually to assess an organizational entity.

### 4.2. Implementation

Figure 1 illustrates the proposed architecture. It is mainly composed of web search engine, data migration/integration service, enterprise search engine, and web application service. We implement the model in open source tools and APIs because the tools are free and proven in the market. Apache Solr search engines on Hadoop platform can handle and integrate Big Data.

**a) Implementing Data Migration/Integration Function**

Web Search Engine is needed to collect non-structured data, for example, html documents, from web sites. The search engine periodically crawls through, searches for, and indexes data of web pages at the web sites given. We suggest to use Apache Nutch [24] engine for web searching by specifying the web sites given by buyers of data. And, in order to retrieve and store the data at the local data source from the repository of the search engine, we need to implement and provide Data Conversion and Parsing Service. After the periodic web data search and indexing, the service retrieves data of the repository by queries. And, the service parses and converts collected data to the desired data format for data integration and stored into local data source $DS_n$. We combine Apache Solr [19-22, 32] search engine and Nutch so that we can use Solr's search APIs in order to collect, parse, and covert data, then finally store converted data into local data source $DS_n$. Besides, Solr runs on Hadoop in order to handle Big Data from the heterogeneous data sources.

Data Migration and Integration service retrieves heterogeneous data periodically from multiple data sources. And, the service parses and converts the data to the specified format by using the dictionary and mapping service. Then, the data in the format needs to be merged into the Global Integrated DB.

For example, data of DS1 has data format {local ID1, column 11, column 12, …., column 1l}. And, data of DS2 has data format {local ID2, column 21, column 22, …., column 2m}. And, global data has data format {global ID, list of local IDs, column 1, column 2, …., column n}. The columns of data DS 1 can be mapped to the columns of the global data as follows: [column 11 -> column 1, column 13 -> column 2, …, column 1l -> column n].

Figure 3.2 illustrates how to integrate profiles' data of an organization into $DS_3$ from $DS_1$ and $DS_2$. IDs of Tables 1 and 3 are mapped to Local IDs of Table 2. Columns name, city, and state of Tables 1 and 3 are mapped to the correspondent columns of Table 2 respectively. CMMI Maturity level of Table 1 is mapped to Maturity column of Table 2. And, ISO 9001 certified column information of Table 3 is mapped to Certified column of Table 2.

**b) Implementing Search Function**

Once we have integrated data at the global integrated DB as shown in Figure 2, we need to index the columns for relevant search, which is related to machine learning algorithm. We use Solr [20] enterprise search engine to search for the proper organizational entities based on dynamic machine learning search algorithm. The developer (or administrator) of *Solr* needs to define columns to index and specify the relevancy weight of the columns.

**Table 1. ABC Inc's Profile with CMMI Appraisal Result from $DS_1$**

| id | name | city | state | zip | Team Leader | Maturity Level |
|---|---|---|---|---|---|---|
| *1111* | *ABC Inc* | *Santa Monica* | *CA* | *90040* | *Tom Cruise* | *5* |

**Table 2. ABC Inc's Profile at Global Integrated Table from $DS_3$**

| id | Local IDs | name | phone | city | state | Maturity | Certi-fied |
|---|---|---|---|---|---|---|---|
| *00012123* | *{1111: DS1, 123:DS3}* | *ABC Inc* | *310-340-xxxx* | *Santa Monica* | *CA* | *5* | *Y* |

**Table 3. ABC Inc's Profile with ISO 9001 Certificate from $DS_2$**

| id | name | phone | city | state | zip | ISO 9001 certified |
|---|---|---|---|---|---|---|
| *123* | *ABCInc* | *310-340-xxxx* | *Santa Monica* | *CA* | *90040* | *Y* |

**Figure 2. Data Integration Example**

For example, if Table 3 has the data {00012123, {1111: DS1, 123:DS3}, "ABC Inc", "Category: IT", "310-340-xxxx", "Santa Monica", CA, '5', 'Y'}, we can specify indexable columns for category "Category: IT", CMMI Maturity level '5', and ISO 9001 certified 'Y'. And, we could give weight 50% to the category, 75% to CMMI Maturity Level, and 75% to ISO 9001 certified columns respectively. Then, when the search engine display the search results, the highly relevant search results will be displayed first. Once we have more information how to retrieve and sort data from the buyer, we can define the relevancy as required. Besides, we can also implement dynamic relevancy – machine learning algorithm - that dynamically update the relevancy by giving the different weights to the columns from the information from the buyer.

As shown in Figure 1, Our Searchable Data Generator Service periodically retrieves data from the global DB and generates XML or Json files. And, it posts XML or Json files to the Solr search engine and the search engine indexes the XML (or Json) elements as specified above. Then, the indexed data are stored at the repository of the search engine. XML (or Json) file looks as shown in Figure 3. Besides, Data Alert Service is for SA, which keeps the user's

 

contact information and interesting data. When the profile data is updated, the service sends an alert to the user.

Our web site has search functions that are built with Solr search APIs. Thus, when the user searches for organizational entities by queries, the APIs request for search results. And, the search engine looks for the entities' information at the repository then it sends the response back to the user with the relevant search results that are displayed on the user's web browser. For example, Figure 4 shows the example GUI view for search. The user chooses ">3" for CMMI maturity level that is greater than 3, "Yes" for ISO 9001 certified, and "IT" for category in order to search for entities that are good fit to the queries.

**c) Use Cases**

```
<docs>
  <doc>
    <field name = "id"> 00012123 </field>
    <field name = "name"> ABC Inc </field>
    <field name = "zip"> 90030 </field>
    <field name = "category"> IT </field>
    <field name = "maturity"> 5 </field>
    <field name = "certified"> y </field>
  </doc>
  <doc>
…
</docs>
```

**Figure 3. XML Documents Example for Solr**

In order to build the systems shown in Figure 1, we need several use cases that direct what to implement. The use cases with the services are illustrated as follows:

**Goal:** A user searches for and assesses the profiles of entities at the web site to find the proper contractors for IT solutions, which are collected from the multiple heterogeneous data sources.

**Assumption:**

(1) The buyers give us the steps and processes to search and retrieve the profiles of the organizational entities.

(2) Data sources include web sites that need to be searched for by web search engine.

**Use Cases for Web Search Engine:**

(1) Web Search Engine periodically crawls through, searches for, and collects the entities' web sites given by buyers.

(2) Data Conversion/Parsing Service converts the collected data to the proper format and stores them to the data source.

**Use Cases for Data Integration/Migration:**

(1) Data Integration/Migration Service periodically retrieve data from the multiple data sources – including data source at the web search engine.

(2) The service combines by following the mapping information and stores the data to the global integrated DB.

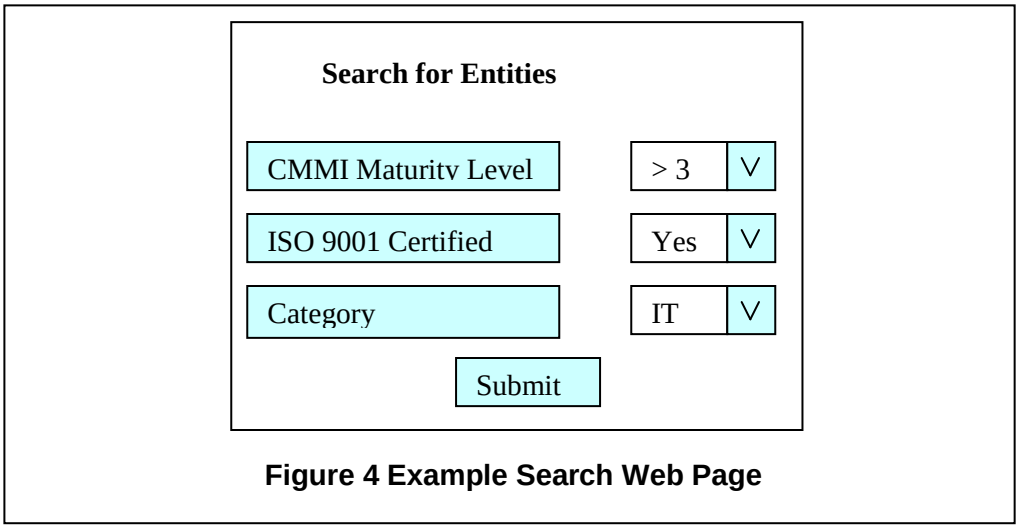


**Figure 4 Example Search Web Page**

**Use Cases for Data Indexing:**

(1) Searchable Data Generator Service periodically generates XML files from the global integrated DB.

(2) The service posts the XML files to the enterprise search engine.

**Use Cases for Situation Awareness:**

(1) Data Alert Service keeps the user's contact information for the user's interesting data.

(2) Data Alert Service sends an alert to the user when the user's interesting data is updated.

**Use Cases for User:**

(1) The user opens the web site and logs in

(2) The user chooses the queries from the drop down menus given and clicks on submit button.

a. Or/and the user types in the queries at the text field and clicks on submit button.

(3) The site moves to new page that displays the search results that show the profiles of the entities by relevant order.

## 5. Conclusion

The paper presents the model that integrates heterogeneous data sources and retrieves information from the integrated database in order to find the proper IT businesses that

 

satisfy the system engineering standards. Besides, the profile data of the businesses are updated periodically so that the model has the function to update the data on SA. It is to assist buyers to easily search for the proper vendors who are qualified for the buyers' projects.

The model is built in Nutch web search engine and Solr on Hadoop platform to collect web and local Big Data using J2EE. And, the data is collected in XML (or Json) and integrated into local DB using Hadoop platform. The model is implemented on N-tier web architecture so that the buyers can conveniently find the proper vendors.

## References

[1] "Capability Maturity Model Integration", http://en.wikipedia.org/wiki/Capability_Maturity_Model_ Integration.
[2] "Capability Maturity Model", http://en.wikipedia.org/wiki/Capability_Maturity_Model.
[3] "Standard CMMI Appraisal Method for Process Improvement (SCAMPISM) A, Version 1.2: Method Definition Document", CMU/SEI-2006-HB-002, Software Engineering Institute, Carnegie Mellon University. **(2006)**.
[4] "Published Appraisal Results", http://sas.sei.cmu.edu/pars/pars.aspx, Software Engineering Institute, Carnegie Mellon University.
[5] "Systems engineering process", http://en.wikipedia.org/wiki/Systems_engineering_process.
[6] "Systems engineering process", https://acc.dau.mil/CommunityBrowser.aspx?id=250180, ACQuipedia, Defense Acquisition University.
[7] "ISO 9000 and 14000", http://www.iso.org/iso/iso_catalogue/management_standards/iso_9000_iso_14000.htm, International Organization for Standardization.
[8] "ISO 9000", http://en.wikipedia.org/wiki/ISO9000.
[9] "Six Degrees: The Science of a Connected Age", Watts, Duncan, New York: W.W. Norton & Company, **(2003)**.
[10] "Linked: How Everything is Connected to Everything Else and What it Means for Business, Science, and Everyday Life", Barabasi, Albert-Laszlo, New York: Plume, **(2003)**.
[11] "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing", J. Woo and Y. Xu, The 2011 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas **(2011)** July 18-21.
[12] "Market Basket Analysis Algorithm with NoSQL DB HBase and Hadoop", J. Woo, S. Basopia, Y. Xu, S. Ho Kim, The Third International Conference on Emerging Databases (EDB 2011), Songdo Park Hotel, Incheon, Korea, **(2011)** August 25-27.
[13] "Apriori-Map/Reduce Algorithm", Jongwook Woo, The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2012), Las Vegas **(2012)** July 16-19.
[14] Apache Hadoop Project, http://hadoop.apache.org/.
[15] "The Comparison of J2EE and .NET for Enterprise Information Systems", J. Woo, Journal of Scalable Computing: Practices and Experience (SCPE), vol. 8, no. 1, **(2007)** June, pp. 131-140.
[16] "Information Integration Model in e-Business", J. Woo, D.-Y. Kim, W. Cho, M. Jang, The 2009 international Conference on e-Learning, e-Business, Enterprise Information Systems, e-Government, and Outsourcing, Las Vegas, **(2009)** July 13-16.
[17] "Integrated Information Systems Architecture in e-Business", Jongwook Woo, Dong-Yon Kim, Wonhong Cho, MinSeok Jang, The 2007 international Conference on e-Learning, e-Business, Enterprise Information Systems, e-Government, and Outsourcing, Las Vegas, **(2007)** June 26-29.
[18] "Enterprise information integration: successes, challenges and controversies", A. Y. Halevy, N. Ashishy, D. Bittonz, M. Careyx, D. Draper, J. Pollock, K. Arnon Rosenthal and V. Sikkayy, Proceedings of the 2005 ACM SIGMOD international conference on Management of data.
[19] "Lucene", http://lucene.apache.org/, Apache Software Foundation.
[20] "Solr", http://lucene.apache.org/solr/, Apache Software Foundation.
[21] "Local Lucene", http://www.gissearch.com/locallucene".
[22] "Local Solr", http://www.gissearch.com/localsolr.
[23] "FAST Esp", http://www.microsoft.com/enterprisesearch/en/us/Fast.aspx , Microsoft Systems.
[24] "Nutch", http://nutch.apache.org/ , Apache Nutch Project.
[25] "e-Business Architecture with Search Engine", Jongwook Woo, The 2008 US-Korea Conference on Science, Technology, and Entrepreneurship (UKC-2007), Information Technology (IT), San Diego, **(2008)** August 14-17.

[26] "e-Business Architecture with Enterprise Search Engine", Jongwook Woo, the 9th KOCSEA (Korean Computer Scientists and Engineers Association in America) Technical Symposium, VA, **(2008)** October 25-26.
[27] "State of the Data Integration Market 2008-2009", An Oracle White Paper, **(2008)** November.
[28] "New Research Report Quantifies Economic Benefits of Informatica Data Integration Platform and ICCs", Press Release, http://www.streetinsider.com/Earnings/New+Research+Report+Quantifies+Economic+ Benefits+of+Informatica+Data+Integration+Platform+and+ICCs/4654711.html, **(2009)** May 15.
[29] "Making the Most out of Mashups", http://www.ciostrategycenter.com/kpix/Res/res_strategies/ making_most_out_of_mashups/index.html, Courtney Macavinta.
[30] "Situation Awareness", http://www.uscg.mil/auxiliary/training/tct/chap5.pdf, US Coast Guard.
[31] "The Software EBox: Information Integration for Situational Awareness", N. Ashish, J. Lickfett, S. Mehrotra and N. Venkatasubramanian, IEEE Intelligence and Security Informatics (ISI), Dallas, **(2009)** June.
[32] "CMMI or Agile: Why Not Embrace Both!", H. Glazer, J. Dalton, D. Anderson, M. Konrad, S. Shrum, Technical Note CMU/SEI-2008-TN-003 CMU Software Engineering Institute, **(2008)** November.
[33] "Cloudera Search Installation Guide", https://www.cloudera.com/content/cloudera-content/cloudera-docs/Search/latest/PDF/Cloudera-Search-Installation-Guide.pdf, Cloudera Inc.

## Author

**Jongwook Woo** is currently an Associate Professor at Computer Information Systems Department of California State University, Los Angeles. He received the BS and the MS degree, both in Electronic Engineering from Yonsei University in 1989 and 1991, respectively. He obtained his second MS degree in Computer Science and received the PhD degree in Computer Engineering, both from University of Southern California in 1998 and 2001, respectively. His research interests are Information Retrieval/Integration/Sharing on Big Data, Map/Reduce algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier Architecture application in e-Business, smartphone, social networking and bioinformatics applications. He has published more than 40 peer reviewed conference and journal papers. He also has consulted many entertainment companies in Hollywood.

 