DOI: 10.2298/CSIS120922030L

Design and Implementation of E-Discovery as a Service based on Cloud Computing

Taerim Lee[1], Hun Kim[1], Kyung-Hyune Rhee[1], and Sang Uk Shin[1]

[1] Pukyong National University,
Busan, Republic of Korea
{taeri, mybreathing, khrhee, shinsu}@pknu.ac.kr

Abstract. Recently, as IT Compliance becomes more diverse, companies have to take a great amount of effort to comply with it and prepare countermeasures. Especially, E-Discovery is also one of the most notable compliances for IT and law. In order to minimize the time and cost for E-Discovery, many service systems and solutions using the state-of-the-art technology have been competitively developed. Among them, Cloud Computing is one of the most exclusive skills as a computing infrastructure for E-Discovery Service. Unfortunately, these products actually do not cover all kinds of E-Discovery works and have many drawbacks as well as considerable limitations. This paper, therefore, proposes a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS(E-Discovery as a Service) to make the best usage of its advantages and overcome the limitations of the existing E-Discovery solutions. EDaaS enables E-Discovery participants to smoothly collaborate by removing constraints on working places and minimizing the number of direct contact with target systems. What those who want to use the EDaaS need is only a network device for using the Internet. Moreover, EDaaS can help to reduce the waste of time and human resources because no specific software to install on every target system is needed and the relatively exact time of completion can be obtained from it according to the amount of data for the manpower control. As a result of it, EDaaS can solve the litigant's cost problem.

Keywords: E-Discovery, EDRM, Cloud Computing, SaaS.

1. Introduction

Due to the wide distribution of digital devices such as computers, smart phones and rapid advances in various IT technologies, Internet has become a part of our daily life and automated information processing system has been used more and more in our work. As a result of it, electronic documents have been rapidly getting used. This situation has had an impact on the judicial systems and brought big changes on them. In litigation, particularly on civil

litigation in the US Federal Courts, the parties are required, if requested, to produce documents which are potentially relevant to the issues and facts of the matter. This is a part of the process called “Discovery”. When it involves with the electronic documents, or more formally, “Electronically Stored Information (ESI)”, it is called as E-Discovery. Especially nowadays, the growing number of legal cases for civil or criminal trials where critical evidences are stored in digital storages has been submitted as the digital forms of information with a high preference. Moreover, business owners and professional executives are growing more interested in E-Discovery since the number of lawsuits is rapidly increasing among business corporations due to conflicts of interest. And also, many global firms specially aimed at U.S. are reconstructing their business processes and deploying the professional E-Discovery service solution to cope with fast-growing IT compliances effectively apart from ERP (Enterprise Resource Planning) solutions because E-Discovery is also one of the most notable compliances and a specialized field for IT [13]. As IT Compliance becomes more diverse, companies have to take a great amount of effort to comply with it and prepare countermeasures.

The major objective of E-Discovery works is to win a suit. To achieve this goal, the litigants have to secure crucial evidences closely related to litigation issues and apply them to prove their legitimacy. In the E-Discovery procedures, the Potentially Relevant Documents are said to be responsive. The actual E-Discovery works are performed by both jurists and IT experts who are collaborating with each other. When the litigation is filed, an attorney or a legal team hired by the litigant analyzes the contents of the petition and identifies major issues of the litigation at first. Then, they produce a keyword list about evidences which must be secured on the basis of the litigation issues and deliver it to IT experts. By using the generated keywords as well as the specialized tools, IT expert or a special team searches related data as potential evidence and visualizes them for review. After that, attorneys review and analyze again the extracted data from various points of view such as suitability, sensitivity and confidentiality. Finally, all evidences are produced by passing through the procedures mentioned above for a presentation in the trial [1]. Although this procedure sounds easy, it is very complicated works and there are many cases which this procedure is not going well because of several unexpected variables such as system error, data loss, and etc.

When people do an E-Discovery, there are two important factors that have to be obligatorily considered besides winning a suit. One is time and the other is cost. Recently, the volume of ESI that must be reviewed for relevance continues to grow and continues to present a challenge to the parties. So, the cost of E-Discovery can easily be in the millions of dollars. According to some commentators, these costs threaten to skew the justice system can easily exceed the amount at risk. Discovery is a major source of costs in litigation, sometimes accounting for as much as 25% of the total cost. Overwhelmingly, the biggest single cost in E-Discovery is for attorney review time - the time spent considering whether each document is responsive

(relevant) or not. Traditionally, each document or email was reviewed by an attorney. As the volume of ESI continues to grow, it is becoming increasingly untenable to pursue that strategy [7]. In addition, according to FRCP(Federal Rules of Civil Procedure), litigants must submit all evidences within 120 days from the day of lawsuit filed [10]. 120 days seem to be enough time to make evidences but the reality is different. Because that period contains a lot of tasks, such as a checking the litigation issues, a discussion about whole e-Discovery schedule or evidence submission format. If litigants cannot prepare suitable evidence within the fixed period by a law, the case is definitely lost. So, attorneys and their clients are looking for ways to minimize the cost and time of E-Discovery.

To comply with their request, many E-Discovery vendors have competitively developed and released their own service system or software applying the state-of-the-art technologies and Cloud Computing is one of the most exclusive skills as a computing infrastructure for E-Discovery service. Unfortunately, this business is still at a preliminary stage. So, a present level is a simple and partial combination between the existing E-Discovery technologies and Cloud Computing factors for performance enhancement. On the other hand, there are some solutions which implement all E-Discovery functions based on Cloud Computing through a complete platform conversion. However, these products actually do not cover all E-Discovery works and have many drawbacks as well as considerable limitations [3].

In this paper, therefore, we design a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS(E-Discovery as a Service) in order to make the best use of its advantages and overcome the limitations of the existing E-Discovery solutions. The goal of EDaaS is to put all required functions during a whole E-Discovery procedure on the cloud service. This means EDaaS enables E-Discovery participants to smoothly collaborate by removing constraints on working places and minimizing the number of direct contact with target systems. What those who want to use the EDaaS need is only a network device for using the Internet. Moreover, EDaaS can help to reduce the waste of time and human resources because no specific software to install on every target system is needed and the relatively exact time of completion can be obtained from it according to the amount of data for manpower control. As a result of it, EDaaS can solve the litigant's cost problem. Compared to the previous version of this paper appeared in MIST 2012 [2], we improve the EDaaS architecture to expand its functionalities and additionally propose the framework to clarify a configuration of EDaaS. Also, we suggest the way of performance improvement and implement the prototype version of EDaaS.

This paper is organized as follows. Section 2 introduces the background and related work of this study. Section 3 explains how to design and how to use the EDaaS. Section 4 describes three implementation methods for differentiated functions of EDaaS and shows the result of implementation as the prototype. Section 5 then analyzes the practicality of EDaaS to confirm its advantages and limitations. At last, Section 6 presents our conclusion and future work.

2. Background and Related Work

2.1. E-Discovery and EDRM(Electronic Discovery Reference Model)

Electronic discovery (or E-Discovery), first introduced by Federal Rules of Civil Procedure amendments on December 1 2006, refers to Discovery in civil litigation which deals with information in electronic format referred to as ESI (Electronically Stored Information) [10]. This is the result that reflects the modern trend that Discovery's main target is ESI. According to these rules, each company has the responsibility to produce their own evidence for winning the suit, and the use of digital forensic tool is essentially necessary.

EDRM is specified legal requirements of E-Discovery mentioned in U.S. FRCP, and EDRM describes the details about tasks of E-Discovery works. This provides guidelines associated to E-Discovery procedure for standardization and describes functional specification of each phase. This guideline can be recognized as a universal standard because it has been developed in consultation with more than 60 leading E-Discovery-related organizations since 2006. Thus, most of the tools and techniques for E-Discovery are designed on the basis of this model [11]. Fig. 1 shows EDRM diagram which represents a conceptual view of the E-discovery process.

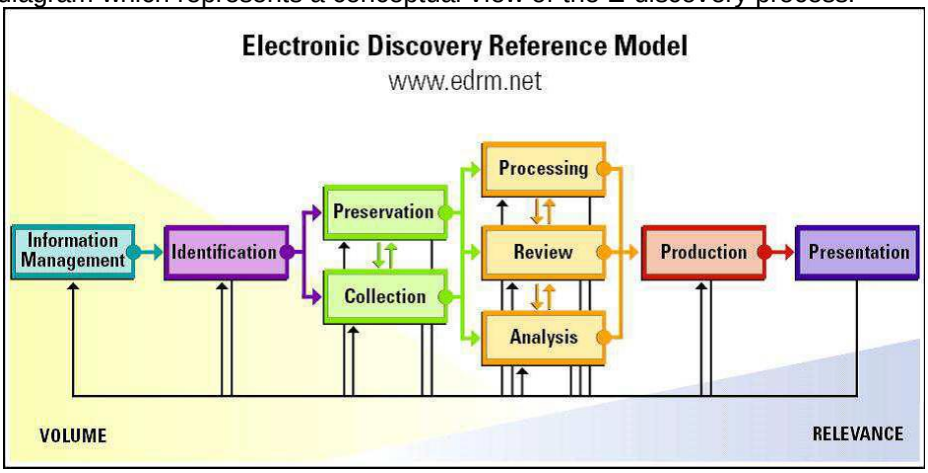

Fig. 1. Electronic Discovery Reference Model

2.2. Major Functions of Existing E-Discovery Service and Solutions

Table 1 shows the phases of e-Discovery and summary from specifications of each phase proposed by EDRM [11].

Most functions of existing E-Discovery service and solutions are focused on the following list of five phases(Collection, Processing, Review, Analysis and Production) because all these phases have a high level of dependence on tool's performance for efficiency improvement of E-Discovery works [3]. The primary technologies for implementing these tools are as follows:

- Document Indexing and Query Processing for an effective search operation
- Classification for removing of duplicated, patent or confidential documents
- Data Format Converting for using of integrated platform, prearranged evidence production and various format compliances
- Data Visualization for a cooperation of review and analysis operation
- Labeling and Tagging for a document selection based on the relevance with litigation issues

Table 1. The phases and the summaries from specifications of each phase proposed by EDRM

Phases	Summary of Specifications
Information Management	Phase to manage their own ESI according to organization's information management policy
Identification	Phase to determine scope of e-Discovery target and identify a real ESI for collecting and preserving
Preservation	Phase to protect ESI from a malicious attack or an intentional destruction
Collection	Phase to collect ESI from various types of storages
Processing	Phase to remove overlapping ESI or unrelated data with lawsuit from collected ESIs and convert the ESI to fit the format for an effective review
Review	Phase to sort sensitive ESI according to privilege, confidentiality, privacy
Analysis	Phase to analyze the collected ESI based on Litigation-related information (Litigation issue, Persons, Keyword, Important documents)
Production	Phase to product ESI with a format negotiated in advance
Presentation	Phase to submit ESI an effective way for being crucial evidence

2.3. The Impact of IT Compliance on the E-Discovery

Generally, GRC (Governance, Risk management, and Compliance) is the umbrella term covering an organization's approach across these three areas. Being closely related concerns, governance, risk and compliance activities

are increasingly being integrated and aligned to some extent in order to avoid conflicts, wasteful overlaps and gaps. While differently interpreted in various organizations, GRC typically encompasses activities such as corporate governance, Enterprise Risk Management (ERM) and corporate compliance with applicable laws and regulations [9]. Among them, Compliance means the conforming to the stated requirements. At an organizational level, it is achieved through management processes which identify the applicable requirements (defined in laws, regulations, contracts, strategies and policies as examples), assess the state of compliance, assess the risks and potential costs of non-compliance against the projected expenses to achieve compliance, and hence prioritize, fund and initiate any corrective actions deemed necessary. Widespread interest in GRC was sparked by the US Sarbanes-Oxley Act and the need for US listed companies to design and implement suitable governance controls for SOX compliance, but the focus of GRC has been shifted towards adding business value through improving operational decision making and strategic planning. It therefore has relevance beyond the SOX world [12]. Especially after the appearance of SOX, many countries and organizations make their own compliance in recent years, such as HIPAA, GLBA, or SB1386. These factors have resulted in the multiple companies demanding on a new type of supporting tool in order to satisfy various requirements of compliance. As a result of that, a large number of E-Discovery technologies related to Digital Forensics have been actively developed and several types of E-Discovery solution have been already released to the market.

2.4. Cloud Computing

Cloud Computing is the most prospective technology for the future of E-Discovery service. A definition of Cloud Computing by NIST(National Institute of Standards and Technology) [5] is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources(e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cloud Computing includes various types of services such as: Infrastructure as a Service(IaaS), where a customer makes use of a service provider's computing, storage or networking infrastructure; Platform as a Service(PaaS), where a customer leverages the provider's resources to run custom applications; and finally Software as a Service(SaaS), where customers use software that is run on the providers infrastructure.

Cloud computing has the five essential characteristics; rapid elasticity, measured service, on-demand self-service, ubiquitous network access, resource pooling. Cloud Computing structure consists of applications, servers, distributed file systems, distributed databases, caches, and cloud storage, mass data analysis, cluster management, server virtualization, etc. The user connects to the cloud service by using the web browser or the

dedicated client, and uses the provided application. Fig. 2 shows a simple SaaS structure of cloud computing system.

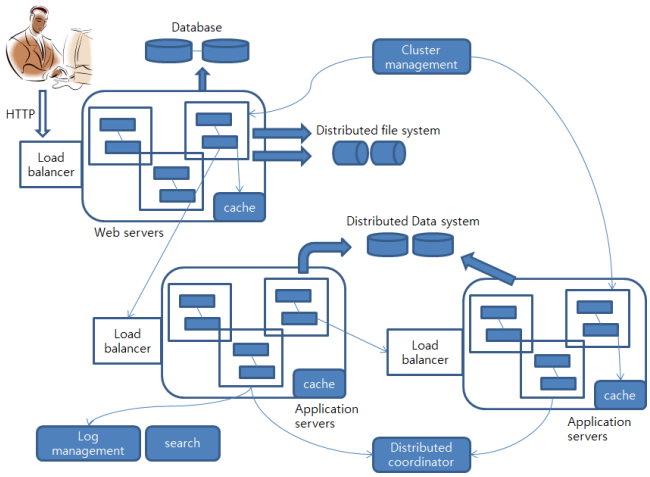

Fig. 2. A Simple SaaS Structure of Cloud Computing System

2.5. E-Discovery Market and Trend of Solution Development

Fig. 3 first introduced in GARTNER 2012 Report shows the famous vendors' position or role in E-Discovery market [3].

Fig. 3. Magic Quadrant for E-Discovery Software

This report was compiled based on the investigation of functionality and characteristics of various E-Discovery software and introduces about each vendor's strengths and cautions. The market covered by this Magic Quadrant contains vendors of e-discovery software solutions for the Identification, Preservation, Collection, Processing, Review, Analysis and Production of ESI in support of the common-law discovery process for litigation, regardless of delivery method. Among them, a vendor who belongs to the group of leaders and visionaries similarly has a clear intention to develop E-Discovery software based on Cloud Computing in a form of SaaS although there are some differences between vendors.

In general, the convergence is made by a partial phased combination and this kind of E-Discovery service consists of two software parts; one is an installation type which was developed at first to deal with many tasks from Collection to Processing phase and the other is cloud server which was implemented Review and Analysis platform. Using the first type software, E-Discovery specialists or hands-on workers can select potentially relevant documents from target system, and convert some documentary format for suitable to the integrated Review and Analysis platform and transfer them to cloud server. After that, various E-Discovery participants, especially company's legal team or attorneys from the external law firm, can review and analyze a relevance of documents as evidence at the same time with no limitations of place. This is an attempt to reduce wasted cost for Review and Analysis phase by improving work efficiency because this phase requires a lot of collaboration among various participants.

AccessData and Guidance Software are representative vendors who make this kind of product. The reason why they are all belong to the leaders group and choose the way of partial convergence is that they already have a powerful software with similar to the first type and they want to keep using and selling that. However in the real litigation cases, cooperation is required through the entire procedure of E-Discovery as well as Review and Analysis. Accordingly, it is necessary to combine additional phases from Identification to Production or to implement all functions on the complete Cloud Computing platform. At this point, vendors such as Xerox Litigation Service, Integreon are continually trying to develop solutions which implement a considerable portion of E-Discovery procedure by using Cloud Computing technologies. Unfortunately, they have not produced a noticeable outcome yet, so they are classified as the Niche Players Group.

Therefore, differentiation factors of our research as follows; the goal of our research is to suggest a new type of E-Discovery service by using Cloud Computing technology. As far as we know, there are no studies related to this goal. Thus, we will compare with famous commercial solution. Considering the trend of E-Discovery solution development, all vendors above mentioned are on the same page, but our attempts and methods to develop a solution are totally different. Simply put, our design and framework is to implement all functions which were required during a whole E-Discovery procedure on the Cloud Computing platform and our methods to conduct them have a distinctive differences from the methods of existing vendors. It means our

development result is the complete convergence of E-Discovery and Cloud Computing beyond the present level of convergence.

3. Design of EDaaS(E-Discovery as a Service)

3.1. Convergence of E-Discovery Solutions and Cloud Computing

In recent years, the quantity of a company's data which may become an object of E-Discovery potentially is growing larger day after day and E-Discovery participants are becoming more diverse. Especially, E-Discovery participants may include company's legal team, general employees, staffs, managers in each department, external law firm, or outsourcing company specialized in E-Discovery, etc. They are people who were closely related with litigation, E-Discovery works or litigant parties. So, nothing is more important than smooth cooperation among participants for the success of E-Discovery works. To reflect this circumstance, the recent trend of technical development for E-Discovery is the convergence of existing services or solutions with Cloud Computing. But even if a lot of famous vendors have been released a new convergence type of solution competitively, serious challenges still remain. Top priority challenge is the complete convergence of E-Discovery with Cloud Computing.

Before attempting to combine E-Discovery Solution with Cloud Computing, most of tools for E-Discovery were developed in a general form called installation type software. It means these kinds of tools must be installed at target system for use. So, E-Discovery participants need extra time for software installation beyond the total time required for E-Discovery works. In order to reduce waste time like this, pre-installing of an E-Discovery tool on every in-house system is time and cost consuming and obviously inefficient. Moreover, installation-oriented software can usually give no guarantee of steady operation pace because its operating efficiency definitely depends on the performance of system where it was installed. With all its faults, vendors don't make an effort to change a principle of their development method because they already have powerful software and they want to sell it consistently. However, it is time for a change. Therefore, we design a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS in order to make the best use of its advantages and overcome the limitations of the existing E-Discovery solutions.

3.2. EDaaS Architecture

The goal of EDaaS is to provide for all functions required during a whole E-Discovery procedure on the cloud service. That is, EDaaS is composed in the

manner of SaaS. To do this, each function will be implemented in the form of application, and each application will interoperate with separated cloud storages based on its purpose and E-Discovery work schedule. Fig. 4 shows the overview of EDaaS architecture.

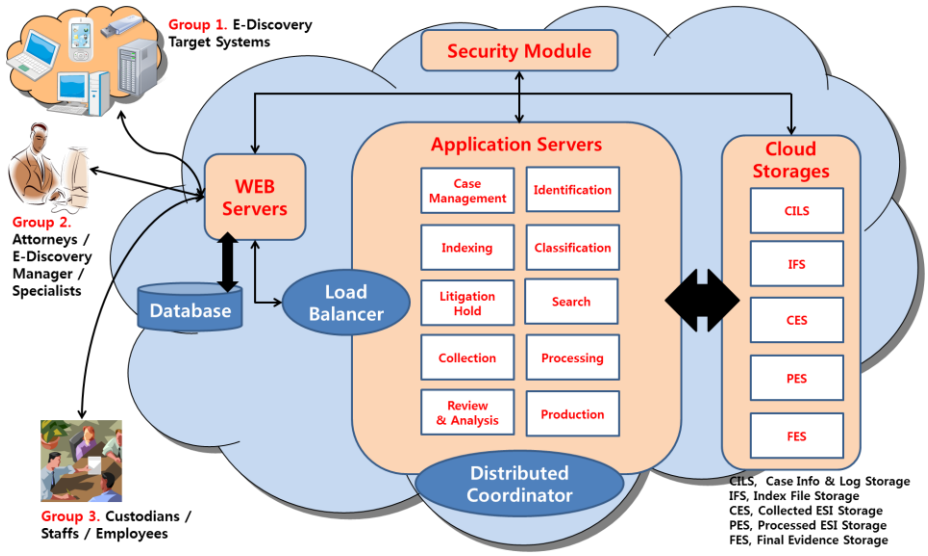

Fig. 4. The Overview of EDaaS Architecture

This architecture is a PIM(Platform-Independent Model). Generally, business applications from various problem domains usually comprise complex functionalities. If such functionalities would not be embedded into the PIM of a software system being designed, a programmer has to create latter a program code of such functionalities, or at least has to amend a generated program code, "by hand" [4].

Users of EDaaS can be divided into three groups. The first group 1 includes E-Discovery target systems which were identified that potentially relevant documents were stored and these systems will be connected for indexing and collection. The second group 2 includes those who have a responsibility to do an E-Discovery works because they were hired as specialists by a litigant such as attorneys in law firm, staffs in outsourcing company specialized in E-Discovery. Of course, if a litigant is a company and the company has a legal or E-Discovery team, these people also belong to the second group. The last group 3 includes those who are related to the litigation issues and have a duty to interview for Identification.

EDaaS consists of 4 parts for the E-Discovery service operation(WEB Servers, Application Servers, Cloud Storages, Security Module) and 2 parts for the system resource management(Load Balancer, Distributed Coordinator). Blocks depicted in Application Servers section of Fig. 4 are

service applications of EDaaS. The name and purpose of each application is shown at the next Table 2.

Table 2. The name and purpose of each application for EDaaS

Name	Target Users	Interoperated Storages	Purpose
Case Management	Group 2	CILS	Saving and managing the all information about case and E-Discovery works (litigation issue, participants, the progress of work, the people concerned, E-Discovery target systems, etc.)
Identification	Group 3	CILS	Providing a specific protocol and reply forms for interview to identify E-Discovery target systems
Indexing	Group 1 and 2	CILS and IFS	Creating index files of each target system for classification and search
Classification	Group 2	IFS	Classifying documents according to contents and updating index files by using the result
Litigation Hold	Group 2	N/A	Ordering target system to prevent users from modifying or deleting important data as potential evidence
Search	Group 2	CILS and IFS	Search for potentially relevant documents related with litigation issue and saving the search result (the path of document)
Collection	Group 1 and 2	CILS and CES	Making a copy of the relevant documents and creating hash values for file integrity
Processing	Group 2	CES and PES	Converting a document file format suitable for integrated Review and Analysis platform
Review and Analysis	Group 2	PES and FES	Providing an integrated platform, visualizing the contents of document, tagging relevant documents as evidence and moving them to FES
Production	Group 2	FES	Convert a document file to the negotiated evidence format and making a final report

In addition to applications, there are essential parts for EDaaS and Fig. 5 shows the entire framework of EDaaS.

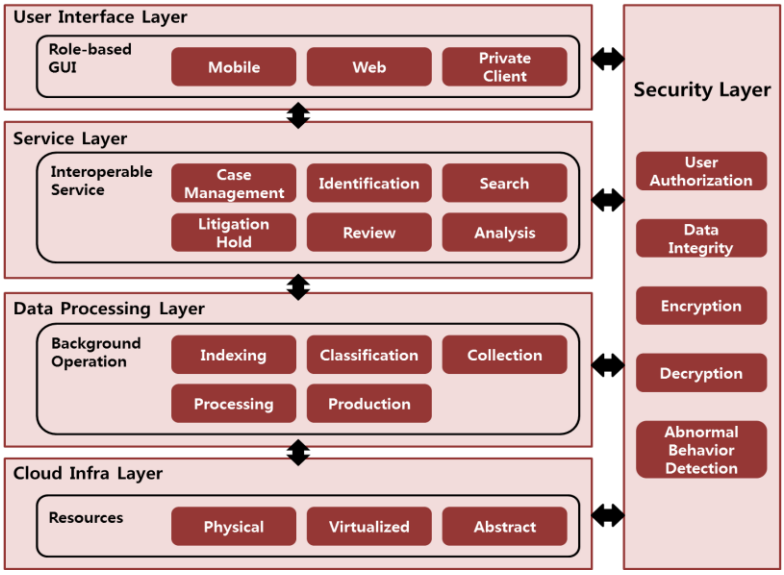

Fig. 5. The Framework of EDaaS

The framework consists of 5 Layers. To the exclusion of Service and Data Processing Layer which was composed of applications described in Table 2, there are 3 more parts; User Interface Layer, Security Layer and Cloud Infra Layer. Each Layer's role is as the following:

- User Interface Layer: Role-based GUI identifies a client's device type, such as Mobile, Desktop or a special device which was made for using a EDaaS only and provides an appropriate GUI for each device.
- Security Layer: It provides a series of functions based on cryptographic technique for user authorization, data integrity, etc. Particularly, this part can be implemented by using a special hardware as well as software. Also, it monitors a state of each layer from user's abnormal behavior.
- Cloud Infra Layer: This layer is for physical hardware of EDaaS. It's a basis part of networks, storages, virtual servers. EDaaS can provide an actual service based on these devices.

3.3. Use Scenario

In order to use the functions of EDaaS, all participants and target systems of E-Discovery have to connect the WEB Servers by using a browser. According to the WEB Server's request, Load Balancer assigns an available Application Server and then WEB Server sends a user's request to the Application Server. After that, Application Server executes a specific application corresponding to the user's request. Fig. 6 shows the mutual

relation between Applications and User Groups with E-Discovery Procedure of EDRM as the center.

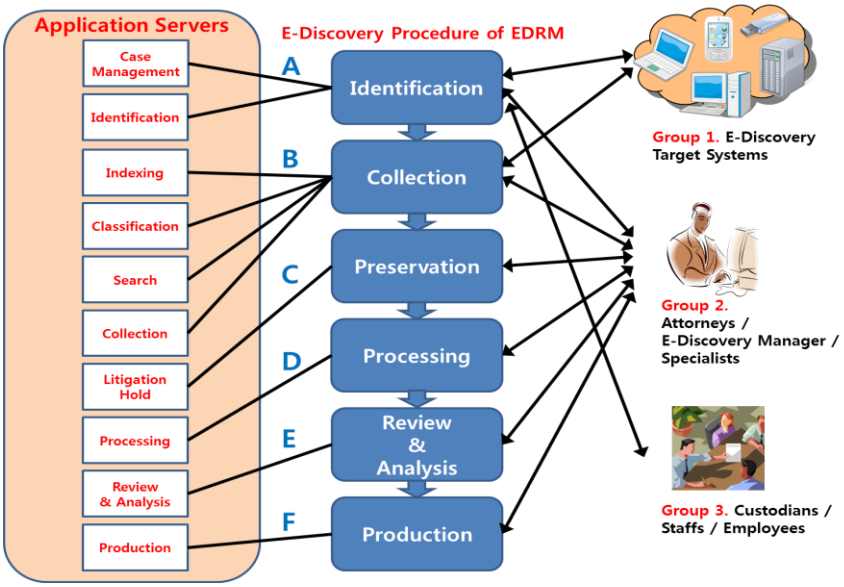

Fig. 6. The mutual relation between Applications and User Groups with E-Discovery Procedure of EDRM

Relation A to F means a bundle of EDaaS Applications to do an essential works for each process of EDRM. These relations reflect the realities of E-Discovery work flow. Full details are as follows:

- A : Once the litigation is occurred, the chief of E-Discovery team creates a database in CLIS and E-Discovery participants record all the information about the litigation and E-Discovery works by using the Case Management application. People those are involved in the litigation have to connect and give an interview personally according to the procedures of Identification. This can make the participants identify E-Discovery target systems.
- B : Identified systems are indexed by the Indexing application. Using an index, participants can search the potentially relevant documents for the future review of suitability as evidence, and the information produced by a Classification application can be used during this process. Because this application enables to remove duplicated documents and identify sensitive documents which are not supposed to make public such as patent or business secret. Classification result can be saved by updating index files with no extra storage. If target documents for review are decided, Collection application can be used to make a copy of each original document and save them to the CES.

- C : By using an Litigation Hold Application, E-Discovery manager have to protect original ESIs from potential threats, such as an intentional Digital Forgery and an accidental loss of data, etc.
- D : Copied files are converted their format suitable for integrated Review and Analysis platform and then they are saved to the PES by using the Processing application.
- E : Attorneys can review and analyze the processed documents and sort out them for the final submission of evidence.
- F : Before the submission, selected documents have to be converted to the negotiated evidence format by using the Production application.

In order to increase work efficiency, various participants can progress this whole process at the same time, regardless of sequence. Also, if the participants know that there are unintended mistakes, errors or failings by the evaluation of each Application's result, they can go back anytime to the troubled part for reworking.

4. Implementation of EDaaS Prototype

Despite the large number of methodologies, standards, and tools, development of large-scale information systems remains a challenging task. The percentage of unsuccessful development projects in terms of exceeding time and/or budget is constantly between 50% and 70%, from the early 80's to the late 90's. Thirty percent of all projects never reach deployment. Prototype-based methods intended to correct these shortcomings and to bring a software project closer to its users [6]. So, we first developed the prototype version of EDaaS which has basic functions with our proposed methods. The development environment for EDaaS is as follows:

- Operating System: Windows 7 Professional K Service Pack 1, IIS 7
- Integrated Development Environment: Microsoft Visual Studio 2010, .NET Framework 4.0
- Database: Microsoft SQL
- Open source library: Apache Lucene 3.1.0 (Indexing/Search), Apache Mahout 0.5 (Classification)

The implementation of Load Balancer and Distributed Coordinator for large-scale service was excluded from the EDaaS architecture because our focus is to develop the basic functions above all. So, we just implemented core parts for 3 components of EDaaS(Web Servers, Application Servers, Cloud Storages) as one in server computer and prepared a local network which was set for Network File System(NFS) to test. Each PC as a target system of EDaaS on this network was assigned static IP address.

4.1. Implementation Methods for Core Functions of EDaaS

In order to differentiate EDaaS from the existing E-Discovery service and solutions, we suggest the following three implementation methods:

- Remote Indexing: The most straightforward method to create index files at the cloud server-side is storing all of original documents in the cloud storage. Considering the amount of company's data is rapidly increasing, this method is very inefficient from the perspective of storage efficiency and making backup every day is also inefficient because people cannot expect when the E-Discovery work will be needed. Remote Indexing is an alternative to solve these problems. At the beginning, Indexing application of EDaaS creates a new user account which is equivalent to the administrator on target system. This function can be implemented in the form of web browser's plug-in. When this plug-in is installed with user's agreement once, it can start to create a new account by modifying the Windows Registry. Using this account, the application makes a reconnection with target system, and start creating index files by using OS dependent functions such as Network File Sharing or File System. Naturally, developers have to prepare additional methods to deal with communication errors for the stability of indexing operation.
- Classification: Making a dictionary of terms which were made up documents and vectorizing is required prior to create index files. The function for the automated document classification based on its contents can be implemented by using the information produced through these kinds of operations. To do this, developers can use the machine learning algorithms as the case may be. If the E-Discovery participants can decide categories of documents and prepare appropriate learning samples in advance, supervised-learning algorithms like Support Vector Machine will be useful. Were it otherwise, unsupervised-learning like K-means will be more useful [8]. In addition, using a distributed processing system like Hadoop [14] enables to reduce the entire operation time.
- Collection: The function for collection can be implemented in a similar way to Remote Indexing. Using an account created for Remote Indexing, all files in target system can be shared over the networks. The work necessary for collection is only copying files what user want. Above this, hash algorithms can be used to verify the originality and integrity of files. To do this, the application has to get hash values of files before making a copy and compare those values after copy operation.

4.2. Website for EDaaS

This website provides various interfaces. To use the EDaaS, all users have to register and log-in at first page. Through this site, administrator of EDaaS can manage all the information of users and create groups to authorize each user based on his or her grade. According to this grade, available functions of

EDaaS are decided and each user can identify these functions at second page after log-in. For example, second web page for Group 1 and 2 includes menus to request applications for Case Management, Indexing, Search, Review/Analysis and the other page for Group 3 to start an interview process for Identification. Fig. 7 shows the status of webpage when administrator logs to the EDaaS for the management of user's information and rights.

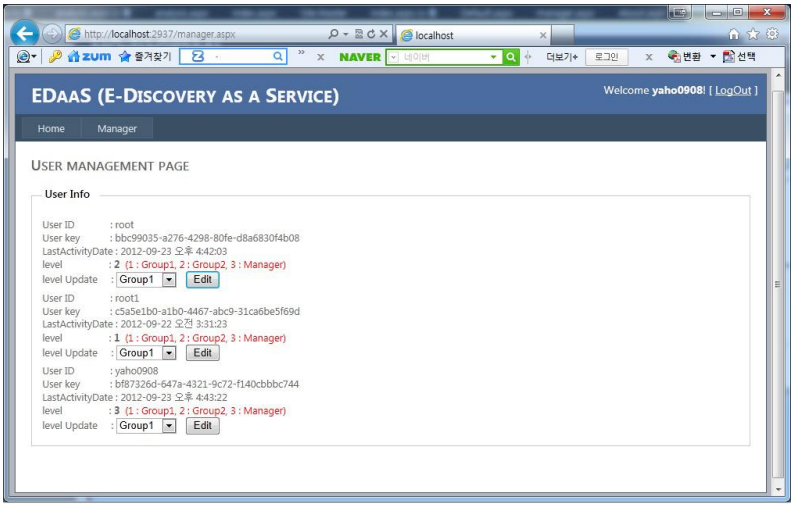

Fig. 7. EDaaS webpage for the management of user's information and rights

4.3. Basic Functions for Application Server of EDaaS

- Indexing and Search : These functions were implemented by using the Apache Lucene Library based on Java. In this prototype, we restricted the document format of E-Discovery target to .TXT text file and applied a simple Boolean search method. Indexing application was made with the C# Thread to run in the background.
- Classification : This function was implemented by using the Apache Mahout Library based on Java and Hadoop Map-Reduce. The biggest reason why we use the Mahout is the interoperability with Lucene. Generally, extra methods for vectorization of each document are required prior to perform a classification. However, Lucene index file is what Mahout only needs for vectorization. Also, it provides various algorithms for document classification, but we choose a K-means clustering method first because it enables to classify documents automatically without training set.[1]

[1] A training set is a set of data used in various areas of information science to discover potentially predictive relationships. Training sets are used in artificial

– Collection : EDaaS prototype runs on network which was set for NFS. So, EDaaS can collect potentially relevant documents by making copy of them.

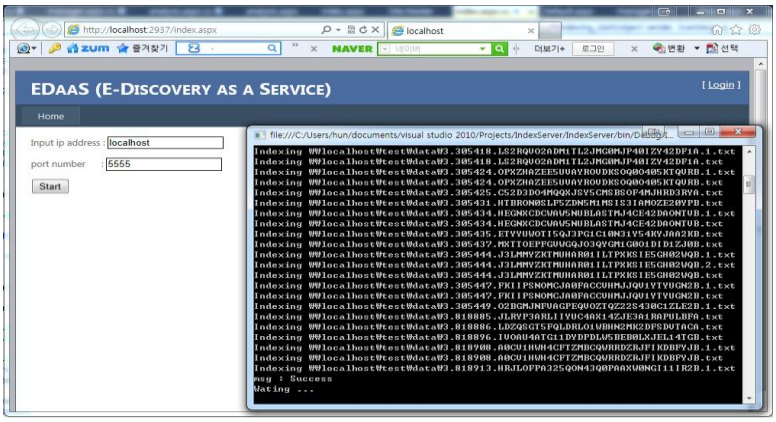

Fig. 8. The Capture of Remote Indexing Operation

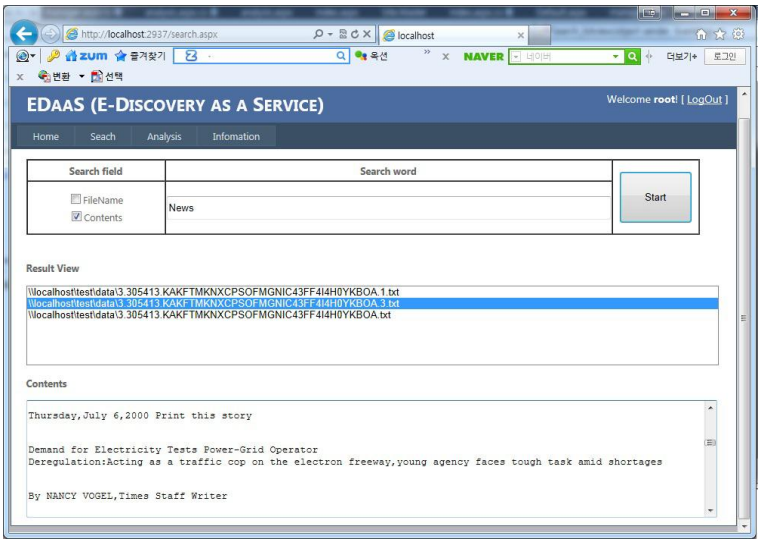

Fig. 9. The Capture of Search and Review Operation

Fig. 8 is the capture of Remote Indexing operation and the console in the right side is to check the logs. Also, Fig. 9 is the capture of Search and Review operation. If the user clicks the one of search result in the middle of page, he can identify the contents of each document.

---

intelligence, machine learning, genetic programming, intelligent systems, and statistics.

5. Analysis about Practicality of EDaaS

Until a recent date, Information Retrieval to find evidence used to be considered the most important function of E-Discovery solution, so evaluation methods for the performance of solution mostly focus on this kind of function. Considering the object of E-Discovery solution, that is quite natural, but it does not fit for informing advantages of EDaaS because it was designed from the another viewpoint. In addition, as far as we know, there are no studies related to this purpose of EDaaS. Therefore, we explain its advantages through the comparison with a typical existing solution.

5.1. Advantages of EDaaS

E-Discovery participants can use EDaaS anytime and anywhere if they have a device for using the Internet. This means no specific software to install on every target system is needed. Especially, the more E-Discovery target systems, the better EDaaS is; it can reduce the waste of time and human resources for the software installation. Moreover, it is difficult to get an estimated time of completion in the case of using the installation type software because its operating efficiency definitely depends on the performance of system where it was installed. If the litigant has to hire persons to the number of target systems for the rapid progression of E-Discovery work, it will cost a huge amount of money. On the other hand, EDaaS can give a relatively exact time of completion according to the amount of data. This information is very useful for the placement of human resources. For this reason, EDaaS can solve the litigant's cost problem. With these advantages, EDaaS enables for participants to collaborate smoothly by removing constraints on working place and minimizing the number of direct contact with target systems. Table 3 shows a comparison with AccessData Summation to explain advantages of EDaaS based on Cloud Computing. Founded in 1987, AccessData Group is a privately held company, with a workforce of over 450, that has addressed the E-Discovery market since 2008 and it has been most famous vendor recently. Also, Summation is the integrated solution which was redesigned to run on the powerful and proven AccessData technology core in 2010 [3].

5.2. Limitations of EDaaS

There are two considerations for practical use of EDaaS. The first is the performance of indexing. The biggest influence is the read/write time for the physical storages on the local system indexing, but remote indexing of EDaaS is additionally influenced by the communication time. So, it is necessary to verify whether or not this tradeoff is tolerable through the experiment. The second is the OS function of Network File System for

Remote Indexing and Collection. Windows OS uses 4 static ports(137, 138, 139, 445) for the sharing service of file and printer. The problem is most ISPs and companies prevent using these ports for security reasons. Furthermore, private local network continues to increase, using the function of NFS as it is with systems on the external network is becoming more difficult. It means additional actions like port forwarding are required to implement Remote Indexing of EDaaS.

Table 3. A comparison with AccessData Summation

Phases	EDaaS	AccessData Summation
Software Installation	N/A	All the target systems of E-Discovery
Extra burden on Installation	N/A	Time, cost, and human resources
Concurrent Users	No limitation	Only one user per system
Working Place	No limitation	Only the place where the system installed it is
Performance	Stable and Predictable (except for network)	Unstable and Unpredictable (It depends on the performance of each system installed it)

5.3. The Future Development Direction for Improving EDaaS

For the performance enhancement of Remote Indexing function, we will bring a Hadoop Map-Reduce technique and implement that function in the form of distributed processing. It is capable of solving the potential Big Data problem. Also in order to prepare when the Remote Indexing is not available because of the network configurations such as the restriction of service port, IP sharing router or VPN, we will develop the additional software in installation type. The ultimate goal of this software is to enable the sharing of file system through the specific port. After expansion of Remote Indexing is complete, experiment for performance evaluation will be done by comparing with local indexing method. Above these works, we will implement the rest of EDaaS architecture and update EDaaS prototype by adding useful techniques for search and review to make it suitable for real E-Discovery business.

6. Conclusion

In this paper, we designed a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS in order to make the best use of cloud computing advantages and overcome the limitations of existing E-

Discovery service or solutions. And then, we explained the structure and framework of EDaaS and suggested a series of a use scenario. Also, we introduced the prototype of EDaaS which was implemented by using three implementation methods; Remote Indexing, Classification and Collection. Finally, we analyzed the practicality of EDaaS and talked about the considerations for the way of improvement.

From now on, complete realization of EDaaS and upgrading its functions based on the study for the improvement of performance will be our future work. It will be performed in accompaniment with the suggestion of a better method for Remote Indexing and the expansion of target OS in order to overcome the limitations of EDaaS prototype.

Acknowledgement. This research was partly supported by Basic Science Research Program(No. 20110006097) and by Next-Generation Information Computing Development Program(No.2011-0029927) through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(Corresponding author: shinsu@pknu.ac.kr).

References

1. Cohen A.I., Kalbaugh G.E.: ESI Handbook : Sources, Technology and Process. Aspen Publishers, Mckinney, USA (2010)
2. Lee T., Kim H., Rhee K.H., Shin S.U.: A Study on Design and Implementation of E-Discovery Service based on Cloud Computing, Journal of Internet Services and Information Security, Vol.2, No.3/4 (MIST 2012 Volume 2), 65-76. (2012)
3. Logan D., Childs S.: Magic quadrant for E-Discovery software. AccessData Company, USA (2012). [Online]. Available: http://accessdata.com/gartner-2012 (current September 2012)
4. Luković I., Popović A., Mostić J., Ristić S.: A Tool for Modeling Form Type Check Constraints and Complex Functionalities of Business Applications. Computer Science and Information Systems, Vol. 7, Issue 2, 359-385. (2010)
5. Mell P., Grance T.: The NIST definition of cloud computing. National Institute Standards and Technology, USA (2011). [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (current September 2012)
6. Milosavljević G., Perišić B.: A Method and a Tool for Rapid Prototyping of Large-Scale Business Information Systems. Computer Science and Information Systems, Vol. 1, Issue 2, 57-82. (2004)
7. Roitblat H.L., Kershaw A., Oot P.: Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review. Journal of the American Society for Information Science and Technology, Vol. 61, Issue 1, 70-80. (2010)
8. Sebastiani F.: Machine Learning in Automated Text Categorization. Journal of ACM Computing Surveys, Vol. 34, No. 1, 1-47. (2002)
9. Tarantino A.: Compliance Handbook(Technology, Finance, Environmental, and International Guidance and Best Practices). Wiley, USA (2007)
10. The Committee on the Judiciary House of Representatives.: Federal Rules of Civil Procedures, USA (2010). [Online]. Available:

http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/2010%20Rules/Civil%20Procedure.pdf (current September 2012)
11. The Electronic Discovery Reference Model.: EDRM Framework Guides, USA (2009) [Online]. Available: http://www.edrm.net/resources/guides/edrm-framework-guides (current September 2012)
12. The Free Encyclopedia Wikipedia.: Governance, Risk Management, and Compliance (2011). [Online]. Available: http://en.wikipedia.org/wiki/Governance,_risk_management,_and_compliance#Integrated_governance.2C_risk_and_compliance (current September 2012)
13. Volonino L., Redpath I.J.: e-Discovery For Dummies. Wiley, USA (2009)
14. White T.: Hadoop: The Definitive Guide 1st Edition. O'Reilly, USA (2009)

Taerim Lee received his Bachelor and Master of Engineering degrees from Pukyong National University, Busan Korea in 2008 and 2010, respectively. He is currently doing a Ph.D. program in Department of Information Security, Graduate School, Pukyong National University. His research interests include digital forensics, e-Discovery, cloud computing, and machine learning.

Hun Kim received his B.S. degree in Major of Computer and Multimedia Engineering from Pukyong National University, Busan, Korea in 2012. He is currently pursuing his master's degree in Department of Information Security, Graduate School, Pukyong National University. His research interests include digital forensics, e-Discovery, Cloud System and security.

Kyung-Hyune Rhee received his M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea in 1985 and 1992, respectively. He worked as a senior researcher in Electronic and Telecommunications Research Institute (ETRI), Daejon, Korea from 1985 to 1993. He also worked as a visiting scholar in the University of Adelaide in Australia, the University of Tokyo in Japan, the University of California at Irvine in USA, and Kyushu University in Japan. He has served as a Chairman of Division of Information and Communication Technology, Colombo Plan Staff College for Technician Education in Manila, the Philippines. He is currently a professor in the Department of IT Convergence and Application Engineering, Pukyong National University, Busan, Korea. His research interests center on multimedia security and analysis, key management protocols and mobile ad-hoc and VANET communication security.

Sang Uk Shin received his M.S. and Ph.D. degrees from Pukyong National University, Busan, Korea in 1997 and 2000, respectively. He worked as a senior researcher in Electronics and Telecommunications Research Institute, Daejeon Korea from 2000 to 2003. He is currently an associate professor in Department of IT Convergence and Application Engineering, Pukyong National University. His research interests include digital forensics, e-Discovery, cryptographic protocol, mobile/wireless network security and multimedia content security.

Received: September 22, 2012; Accepted: March 08, 2013