From Content Management to Enterprise Content Management

Cataldo Mega, Frank Wagner, Bernhard Mitschang

IBM Deutschland Entwicklung GmbH
Institut für Parallele und Verteilte Systeme, Universität Stuttgart
cataldo_mega@de.ibm.com
fkwagner@informatik.uni-stuttgart.de
mitsch@informatik.uni-stuttgart.de

Abstract: In this paper we will provide a step-by-step description on what it means to evolve the architecture of a traditional content management system into an Enterprise Content Management (ECM) system, explain the differences of both systems and motivate why this transformation is necessary. By analyzing business scenarios in the realm of different content management domains, we will explain why today's content management systems struggle when it comes to satisfy the need for performance, scalability and business resilience. Using the system design of IBM DB2 Content Manager as reference point we will outline and discuss some of the new key technical challenges found when promoting on-demand ECM services and look at their affordability. By detailing a few representative use cases we will perform a problem analysis, and an attempt will be made to present an enhanced ECM system design that makes use of a component ‚virtualization' approach in order to allow for a dynamic services infrastructure to be setup and which capitalizes on proven peer-peer and grid technology.

1 Introduction

Today, almost every company and particularly every enterprise faces the need to facilitate an "Any Time, Anywhere Access" to its key business data. The Information Lifecycle implied by the underlying business models typically requires the use of Enterprise Content Management systems (ECM systems) that allow both, business to business, and business to consumer applications access to content assets around the clock, 7 day a week, 365 days a year to an unknown number of web users. This is what is called 'information on demand' within e-business.

The main purpose for ECM systems is to manage and leverage any type of information, across multiple applications and business processes. Their core functionality is facilitating access to all the relevant collateral information that does accompany complex business transactions, from the respective content repositories. As an example, in today's regulated market all the emails flowing in and out of a Stock Broker Company must be monitored and archived due to compliance reasons. Because of this, in a large company up to tens of millions of emails and respective attachments must be captured, parsed, grouped and archived during a business day. That is, a substantial amount of files must be managed and processed during a day, independent of type, size and location.

As another example, given the affordability of new storage technology and the increase in compute and network power, in almost all business domains the amount of unstructured rich media data "consumed" by business processes is becoming larger every day and exceeds the accompanying operational (meta)data by orders of magnitudes.
One consequence is that the operational environment not only has to deal with the management of several relational database management systems but also with a growing number of rich-media content repositories.

In any case, the derived and imposed challenges for the supporting IT infrastructures are not only the management of this sheer amount of data, but also the need to satisfy the online transaction demand of an unknown, but huge number of end users and applications. It is the latter statement, an *unknown* but huge number of transactions, that represents the biggest technical challenge to the ECM IT infrastructure, because it implies that either: a) one has to oversize the ECM operational environment, thus becoming a financial burden or b) during peak load times a possible outage has to be considered. In both cases, the key non-functional requirements that ECM solutions have are:

- Performance and Scalability
- High Availability (HA) including continuous operations
- Backup and Recovery
- And last but not least affordability.

This paper is structured as follows. In Chapter 2 we will introduce what is understood as domain-specific content management systems and especially detail on the lifecycle management issue. Based on the scenarios set up in Chapter 3, we detail on the advancement of content management systems to enterprise content management systems. System characteristics and basic architecture are introduced in Chapter 4. Realization concepts towards an on-demand system strategy are discussed in Chapter 5. Here, basically two virtualization techniques are discussed. Finally, Chapter 6 summarizes the discussion and concludes the paper.

2 Domain-specific Content Management Systems

In the following we outline the 4 most relevant business domains in which a content management infrastructure is essential and which have influenced CM systems design development with their requirements, and still do.

Operational Content Management denotes the first generation of content management systems deployed more then 2 decades ago primarily by the finance domain, with bank and insurance companies as the early adopters. These systems were used as content repositories for operational content like: bank statements, invoices and recurring print output like monthly reports complementing transactional data. This is the domain of Document Image Management Systems (DIMS). They have a limited document lifecycle and implement a 'write once - read almost never' archiving philosophy. The archive retention period might exceed 10 years and their size reaches meanwhile the petabyte range.

The media market segment is the domain of the **rich-media archive systems**. This type of content management systems were developed almost 1 ½ decades ago and were pushing the technological edge.

Only with the advance in storage, computing and compression technology, rich-media archives became affordable to the point at which the media industry started digitizing their assets and adopting Digital Media Management Systems (DMMS) for their production.

With all data digitized, the media business operations as well became subject to a paradigm shift, away from physical assets management. Thus in turn allowing more re-use for the content and more efficient ways in which to produce it!

In the digital realm, physical data and card catalogs were consolidated by means of content management systems. New IT technology allowed better and more efficient operations and produced new business opportunities. The cost factor physical archive evolved into an investment, the content management system.

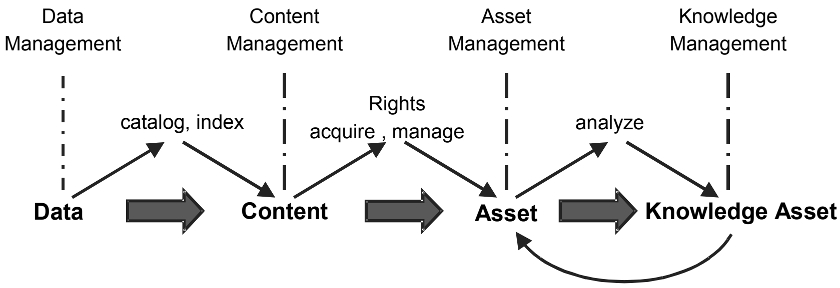

Figure 1 Evolving the information management infrastructure

Pushing further the envelope, with digital rights management (DRM) and knowledge management technology on the horizon, content along with the proper rights became assets and could be traded. And finally with the aid of search and mining algorithms knowledge could be extracted out from the content itself thus allowing automatic indexing and classification of content. All this requires an infrastructure that would support this kind of business operations. In Figure 1 we have sketched this evolution of the information management infrastructure.

The segment of **Web Content Management systems** (WCM) is in the need of a framework for creating, managing, integrating, web-enabling, and delivering unstructured digital content across the enterprise and beyond - to employees, customers and trading partners. Here we differentiate between web content authoring and web content runtime environment. The document lifecycle is more elaborated and implies complex workflows. Security plays an important role and performance is paramount for online-business. The size of content objects is usually small, below 1 MB but their number can grow large. It’s not unusual to find big corporation with hundreds of thousand or even millions of web pages.

The fourth domain is the emerging one. It is the domain of **collaborative and workgroup documents**. People across business environments are increasingly using unstructured data. In the course of any business process, they are leveraging web content, email, spreadsheets, calls, faxes, etc. to help conduct business more efficiently and effectively. The speed of LANs, the amount of computing power now at the desktop and a drop in the cost of storage and processing have aligned so that the use of unstructured data has become not only practical, but a critical element of conducting e-business.

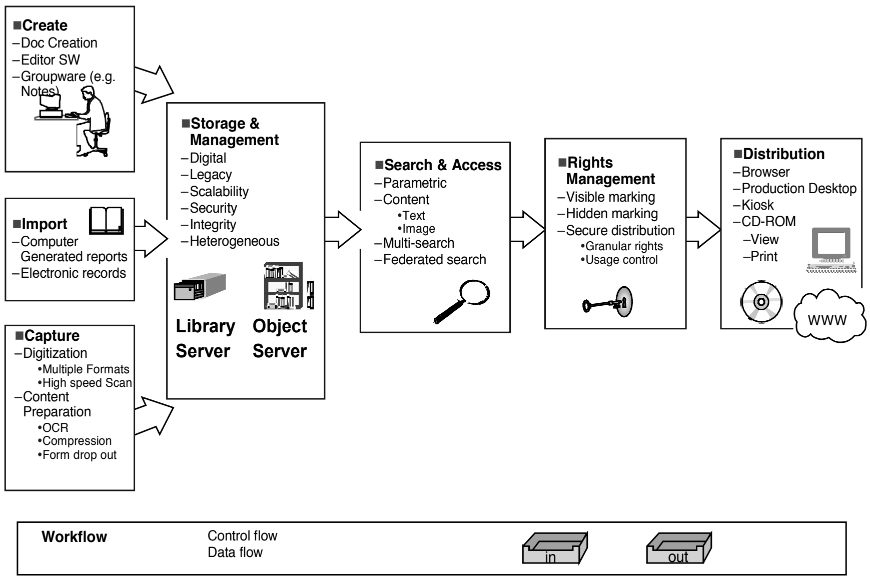

Figure 2 Outline of typical document lifecycle

This kind of content management platforms must provide a unified view of multiple repositories so that users can access all relevant content, independent of which application created it or where it is stored. Business processes in this domain use compound documents subject to a complex document lifecycle and workflows.

While outwardly dissimilar, all of these forms of content management systems serve similar needs and provide in essence the same kind of services.

As outlined in Figure 2 when synthesized, the content lifecycle can be expressed by the following set of services: create, store, catalog, manage, search, browse, access and distribute complemented by document centric workflow services. The difference is with the domain specific semantics these services offer and the context in which they are consumed. For example, content creation might relate to scanned documents, a picture or an article. Searching is different in databases, in plain text documents, in audio or videos. Distribution might be sending a fax, copying a file or playing an advertising video. Because of the fact that these disparate systems have so much in common, and because business content is undergoing exponential growth, companies are struggling to manage and leverage it effectively.

In the world of e-business, compartmentalization of content into separate silos is no longer a viable option. All unstructured data types share similar needs for mass storage, search and access, workflow, personalization, integration with business applications, rights management, version control, and rapid delivery over the Internet.

Instead of separate point solutions for each type of content, enterprise content management can – and should – be implemented on top of a set of unified ECM components as an extension of the e-business infrastructure.

3 Content Management Scenarios

Some of the above mentioned requirements can be found in the scenarios described below.

3.1 Scenario1: Corporate Message Monitoring and Retention

Every day more organizations need to monitor, parse, archive and manage e-mail, instant messages, and attachments to support regulatory compliance, litigation, and corporate policy and to improve IT performance and productivity. Formal electronic recordkeeping is the accountability of "corporate information assets recorded in electronic form."

This involves the following key business requirements and can be expressed as:

a. Protection against litigation due to the unavailability of electronic information to the "right people at the right time".
b. Control of large repositories of corporate records to determine when and what "information should be deleted."

c. Compliance to industry "regulations that must be adhered to in order to do business."

In the concrete case of today's regulated market, a Stock Broker Company for example must monitor and process all incoming and outgoing instant messages and emails. Government laws and regulations enforce archiving of the electronic mail correspondence associated to brokering transactions. For the underlying ECM infrastructure this translates into the fact that for a single company, in a single day up to 25 millions of emails and respective attachments must be monitored, parsed, interpreted and, if found relevant, archived for auditing purposes.

A sample activity diagram detailing the logical flow of events is shown in Figure 3.

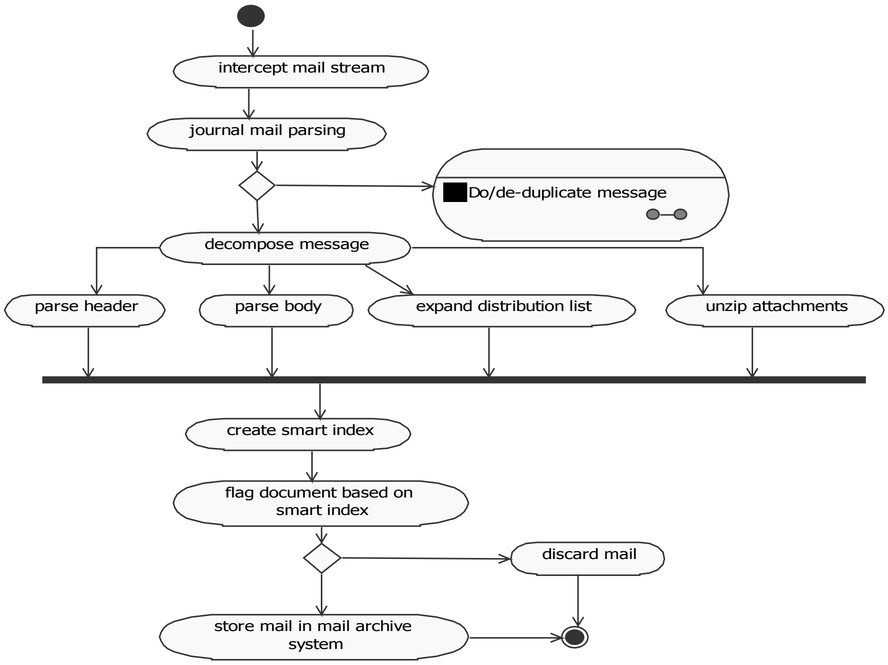

Figure 3 Sample mail monitoring and archiving activity diagram

As simple as this task might seem, it poses a few challenges to the distributed application development and deployment as well as to the flexibility an IT infrastructure must provide. For example, providing consistent response times to customers, despite workloads with significant deviations between average and peak utilization is paramount. Thus, they require flexible resource allocation in accordance with workload demands and priorities.

The key problem at hand is to support the sharing and coordinated use of the many diverse resources needed in dynamic, distributed environments like the one described above such to allow the construction of reliable, scalable, and secure distributed systems.

Given these requirements and the current rush driven by regulatory pressures, one feasible way to solve this problem is to decompose and distribute the mail archive and retention services through a network of content repositories.

A free form component diagram showing the logical decomposition of mail processing services is shown in Figure 4.

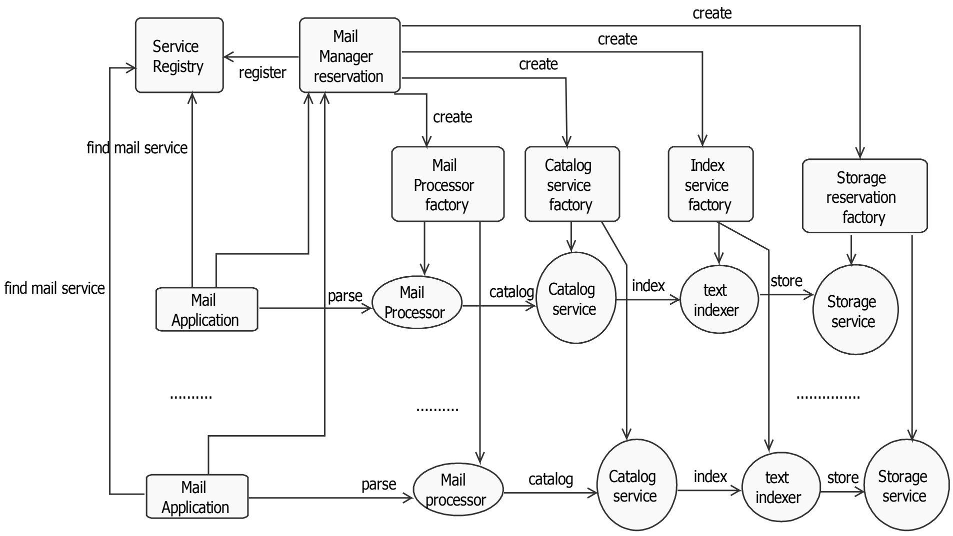

Figure 4 Distributed mail archive service infrastructure

3.2 Scenario2

Similarly, we see that in other domains due to the affordability and increase in size and power of new storage and server technology, the amount of unstructured rich media data "created and consumed" by business applications is exponentially increasing, exceeding the amount of accompanying metadata by orders of magnitudes.

In both scenarios, one can imagine that the corresponding volumes of transaction load and data throughput generated is quite consistent, such to represent a non-trivial challenge for ECM systems to cope with. As said before, the challenge lies not only with the lifecycle management of this sheer amount of data, but also with the need to satisfy the on-line transaction demand of an unknown, but huge number of end users and applications. In parallel with these developments, enterprises are engaging ever more aggressively in e-business and are realizing that much smarter resource provisioning technology is necessary in order for the IT infrastructure to be able to handle the associated unpredictability and rapid growth.

4 Enterprise Content Management as a consolidated Content Management Infrastructure

Historically ECM embraces three separate technologies with an emerging fourth one, collaborative workplaces, as already mentioned in section 2:

- Web content management
- Document management
- Digital media asset management.
- Workplace and collaboration management.

To be truly useful, a content management solution must address requirements for mass storage, search and access, personalization, integration with business applications, access and version control, and rapid delivery over the Internet. This commonality suggests that rather than comprise separate point solutions for each type of enterprise content, an ECM framework can — and should — be implemented on top of a set of unified ECM components, a natural extension of the e-business infrastructure, as shown in Figure 5.

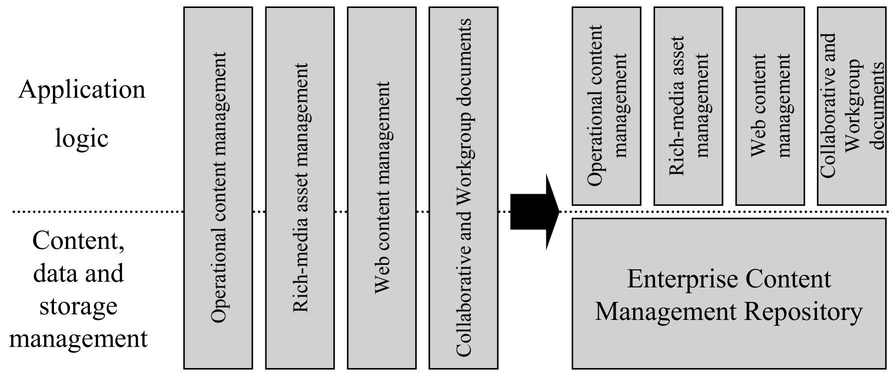

Figure 5 CM to ECM system transition

Then, applications across the enterprise can leverage common platforms and peripherals, lowering the total cost of ownership — not only in hardware and software, but also in system administration, training and custom development.

With an efficient ECM framework as depicted in Figure 6 applications can cost-effectively take advantage of the full range of enterprise content.

Legacy content can be easily published to Web sites and to portals designed to provide secure, personalized delivery to customers, partners and employees across and beyond the enterprise. The ECM infrastructure also allows emerging technologies such as digital rights management, XML-based Web services and Grid services to be implemented consistently and cost-effectively across all forms of enterprise content.

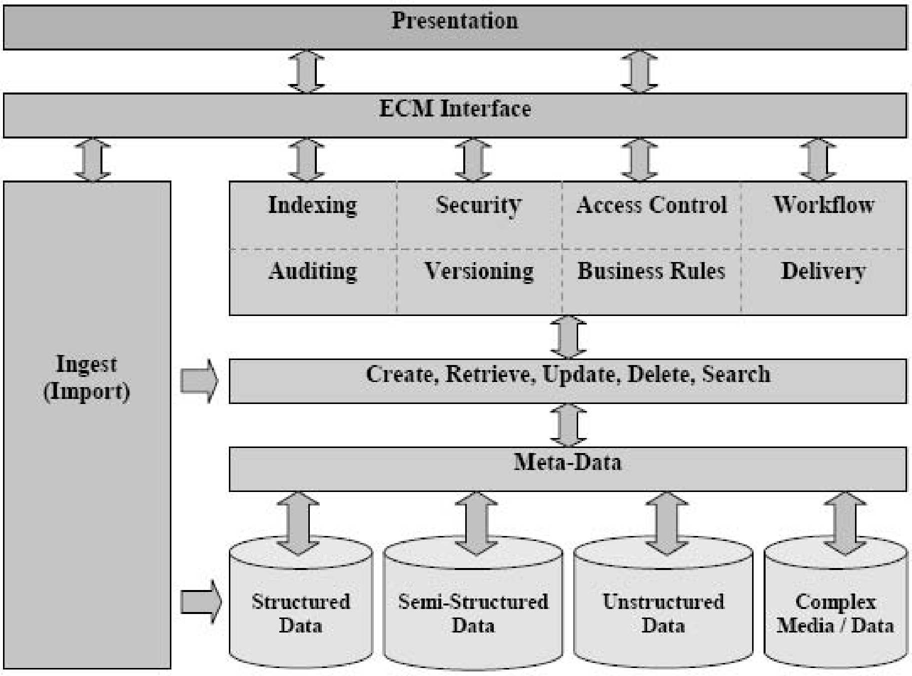

Figure 6 Generic architecture of an ECM system

Characteristics of an ECM infrastructure:

What turns a set of content management components into the basis for an ECM infrastructure?

Customers report that they are seeking five principal characteristics in an ECM infrastructure solution:

- Information Integration.
  The most important criterion is the ability of any business application to access, through a common software interface, all forms of enterprise content. It shouldn't matter whether this content consists of business documents, rich media or dynamic Web content, nor should its repository be a hindrance. In fact, customers want to be able to perform federated searches, in which a single query can retrieve content from a heterogeneous set of repositories distributed over the network. Moreover, customers want to use federated indexing data to personalize that content for portals and e-commerce applications.

- Information Lifecycle
  Another fundamental requirement of modern ECM systems is the capability to support the full content life cycle — from creation through review and approval, to a period of retention, ultimately followed by the destruction of the content.
  This requirement is best explained through the fact that businesses today might not be buried in paper, but they are inundated with all sorts of digital content, from many different sources and in different formats. Employees struggle with managing the content, finding information, and overcoming technological challenges. Therefore the employee workplace increases employee productivity by providing a single user interface for employees to create and manipulate content. Employees should easily find content because the employee workplace eliminates the need to search manually across several applications. Workflow can be used to automate business processes and accelerate task completion. In essence, the employee online workplace badly-needs to be integrated with corporate information, employee directory function, and collaborative capabilities with access to administrative and business applications. Document sharing, document management, and workplace content management are key capabilities which need to be provided.

- Repository scalability and robustness.
  Content repositories must be able to scale cost-effectively from small departmental solutions to enterprise wide applications used by ten thousands of employees, and to customer-facing e-business Web sites drawing millions of daily hits. Scalability also means the ability to manage huge content volumes—particularly important for repositories housing documents and rich media, which may occupy hundreds of gigabytes but must still support rapid response to queries. And it means the ability to cache and distribute content to the "edge of the network" to optimize delivery speed, yet maintain tight central management control.

- Business resiliency.
  To be more successful, businesses need to cope with pressures from an increasingly globalized economy. This implies a transformation to a fully integrated and resilient business across people, processes and information, including suppliers and distributors, customers and employees.
  The ECM infrastructure must encompass an end-to-end information technology environment which allows a business to align with its business goals and strategy for handling all failures as well as planned outages. The key objective is that the availability and resiliency needs and actions are integrated into the everyday ongoing operations of enterprise content management systems and therefore lead to predictable, successful outcomes in case of failure or planned outage.

- Openness.
  An ideal ECM infrastructure must not only be based on open standards, but must also have the flexibility to adapt to any application— including those of competing vendors—through a set of published application program interfaces (APIs). In addition, the ECM infrastructure should support leading infrastructure components from other vendors, including server platforms, database management systems, content repositories and packaged applications from competitors. Not surprisingly, this is the most difficult part for most ECM technology suppliers.

5 Realization of an on demand Enterprise Content Management infrastructure

The content an enterprise has to deal with is one of its most valuable resources. In some branches like insurances and financial institutions the work substantially depends on the availability of the content. Due to the globalization of the enterprises and the rise of the Internet, the content must be available all the time, 24x7x356.

In order to achieve the requirements derived from such business scenarios, one approach is to virtualize the content management system and to automatically provide required system resources to the different components on a as-needed basis. The fact that most often not all components have their peak workload at the same time allows the introduction of an allocation scheme to share resources and to re-assign them to the components that currently need them. Exploiting this synergy on the one hand produces an overall scalable system, which on the other hand also becomes more affordable. Dynamic resource allocation eliminates the need to configure excess resources for the peak hours and promotes a more cost effective resource sharing between the system components.

5.1 Enhancing the traditional CM System Architecture

Researches in Grid computing have shown that some of the latest generation Grid technology is capable of providing partial answers to some of the many challenges ECM systems are faced with today. In some cases Grid technology already provides solutions to part of the implied problems. As an example, Data Grid technology of the 3rd generation in use by several web based applications, show that it is possible to service hundred of thousands of on-line users which move Tera Bytes of data over the Internet in a reliable, secure fashion and with adequate speed.

For ECM systems to take full advantage of this technology, what is missing is an enhanced service integration layer. That is a middleware application framework that not only facilitates access to the business content assets but also promotes the efficient integration of business processes and ultimately business applications. We envision well behaved applications linked to a Business Integration (BI) hub through an adapter mechanism that supports communication to the content manager repository via a standard Business Integration Web services adapter architecture.

Within such a framework, performance and scalability is usually addressed via an intelligent workload management system, using schemes of system auto-monitoring and dynamic load balancing. In tandem with high availability the requirements can be accomplished by:

- Automatic and dynamic system (re-)configuration
- Automatic and dynamic routing of application requests

- Growing and shrinking the environment as needed for a more efficient resource usage, which helps satisfy the affordability.

To address this new kind of business needs legacy Content Management (CM) systems are evolving into Enterprise Content Management (ECM) systems. Their structure is being architected and designed around and for the above mentioned business needs, and still some of the existing technical problems need to be further analyzed, and designed, before a satisfying solution can be found.

In the following we discuss some approaches to virtualize an ECM system.

5.1 Step-by-step Virtualization of the ECM architecture

Virtualization is a trend in application and infrastructure technologies [XHLL 04]. What that means is the virtualization of services provided by applications and infrastructures. This is the decoupling of service consumers from service providers and can be achieved by the introduction of an indirection layer, the virtualized application service layer.

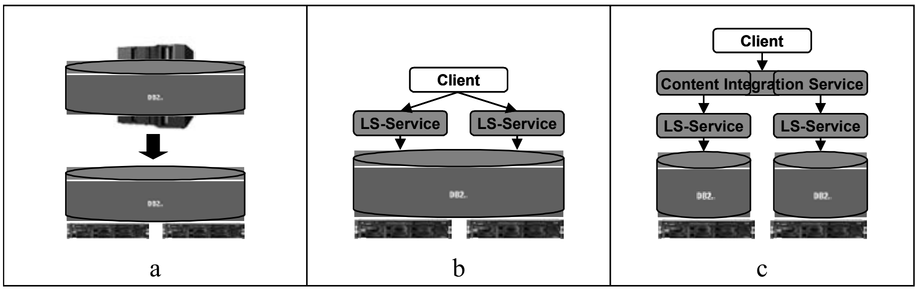

Figure 7 Steps towards a virtualized content management infrastructure

In our case virtualizing the infrastructure will allow dynamic resource configuration such that performance and availability goals regarding all system components can be fulfilled according to a Service Level Agreement (SLA).

To virtualize the infrastructure, we want to use products from the Tivoli family. For example the Tivoli Provisioning Manager for the configuration of the hardware and software, and Tivoli System Automation to automatically maintain the system based on policies.

To enable the dynamic provisioning of an application, a virtualization layer must be implemented between the clients and the applications. A client can't directly connect to the server, as there can be multiple alternative servers with changing addresses over time. So a client first looks up the available instances in a registry. Then one of those instances is chosen based on some metrics like e.g. current load. All parts of the content management system must be considered:

- The catalog, which is the central part of current work
- The content repositories
- The search components (text index, image search).

As outlined in Figure 7. our current project is evaluating two different ways to virtualize one of the key CM system components, the CM system catalog.

Figure 7a represents our first variant: Here we define the same single physical catalog on a distributed database which in turn is installed on multiple physical nodes. In this approach the virtualization layer is provided by the database it self.

Figure 7b shows the second approach where we implement a Grid service on top of the content management system. This approach allows us to test the services and to develop approaches for load-balancing and availability.

Figure 7c represents our third variant: Here we use a CM global logical catalog defined on a cluster of single physical catalogs, each of which defined in a single node data base, in turn installed on a single physical node. In this approach the CM Catalog virtualization layer will be based on a standard Grid services layer the IBM Grid Toolbox [IGT] as a provisional platform and is part of the development effort of this project.

5.2 Virtualization using distributed database technology

The Enterprise Server Edition of IBM DB2 provides a feature called "Database Partitioning Feature" (formerly called Enterprise Extended Edition), that allows to partition a database within one server or over multiple nodes [BFM+03]. This enables a scalable database system (in our case the CM Catalog) as additional resources can be added as needed. The distribution of the database is completely transparent to the client applications (the ECM clients). A client application sees the same database (Catalog) regardless of the node it uses to connect to the database.

When creating the database, the database administrator first defines subsets of the nodes called database partition groups. These groups are then used when creating table spaces to specify which nodes a table space spans.

In Figure 8 there are two single-node table spaces 1 and 2, and one table space 3 that spans both nodes. There are three ways to use these table spaces as illustrated in Figure 8:

- A table must be placed entirely into a single table space
- Typically multiple tables are placed in a single table space

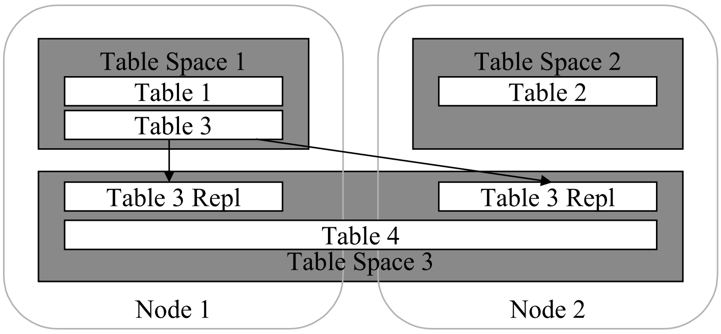

Figure 8 Sample database with single-node, replicated and partitioned tables

- By creating table spaces for every node, it is possible to explicitly place tables on one node or the other (Table 1 and 2).

- Using replicated tables, a variant of materialized query tables, it is possible to replicate a source table (Table 3) into a table space that is defined over multiple nodes. The source table is replicated to every node participating in that table space. In this way the table is locally available for queries at all the nodes. This mechanism is good for smaller tables with low update probability. To increase the utilization of the replicate tables, one should create indexes on those tables. Otherwise it is often cheaper to execute a part of a query on the node with the source table and use the indexes there instead of using the replica without an index.

- A third mechanism allows the horizontal partitioning of a table over a set of nodes (Table 4). For the partitioning key, a user definable subset of the columns of a table, the database calculates a hash value. Based on this value the database decides on which node to put that row. This is a simple mechanism to spread the tuples of a table evenly over a set of nodes that is especially useful for large tables.

Additional nodes can be added to the database and superfluous nodes can be removed from the database on demand. If the databases are explicitly placed on the nodes, the administrator has to adapt the distribution of the tables. Replicated and horizontally partitioned tables can be reorganized using the available “reorg” command.

To exploit IBM DB2 partitioned database (DPF) features we needed to modify the IBM DB2 Content Manager system catalog data-model. We derived a partitioning scheme based on the following principles:

- Tables with few updates and that are often used in queries are replicated to all nodes. As far as possible we defined the indexes that were defined on the source table also on the replicated tables.

- Tables with metadata about the content are partitioned over all nodes. Because there are multiple unique and primary keys defined on those tables, we tested several partitioning keys.

- Tables with no good partitioning key or tables that are updated too often to be replicate were left on one node.

The partitioning of the CM catalog schema poses different challenges. One of which is the large schema for database partitioning with approximately 120 tables interconnected with approximately 150 foreign keys. Other challenges are the inherent OLTP like ECM application workload model and IBM DB2 DPF partitioning constraints. Therefore we experimented with different configurations and adaptations:

- When partitioning a table over multiple nodes, IBM DB2 enforces that every primary key and unique key is a superset of the partitioning key. This allows the database to locally verify the global uniqueness of a key on one node. One consequence is that unique user defined attributes are no longer enforceable. This restriction also has consequences on the foreign keys, as their targets always have to be unique.

- There are implementation schemes in IBM DB2 Content Manager and IBM DB2 which are not well suited for this kind of partitioning and therefore lead to a high communication overhead for some SQL statements. This results in increased execution time which in some cases caused deadlocks for some of the operations. Therefore modification of the application code was necessary.

- The IBM DB2 Content Manager Catalog logic is implemented as a set of stored procedures. The stored procedures are executed on the coordinator node, the node the client applications connect to. This node can be different for every client, but IBM DB2 DPF uses only one execution plan for the given stored procedure independent of the node they are executed.

- All replicated tables must have a primary key. If this is not the case one has to define a primary key. However, this has an impact on updates of the replicated tables.

- It is not possible to define unique indexes on a replicated table.

The catalog of the IBM DB2 Content Manager library server, is implemented as a set of stored procedures in a IBM DB2 database, therefore using a distributed database for virtualizing the catalog is one alternative. In principle only the database schema must be adapted to reflect a good distribution of the database over the available nodes. In practice this goal cannot be achieved so easily because of the diverse nature of the inherent workload models executed by the ECM applications. First an ECM system usually hosts different ECM solutions each of which having its own specific OLTP workload and second their implied schemas might prevent a global partitioning approach that satisfies all needs.

5.3 Virtualization using Grid technology

The second approach we will take into account for virtualization is to implement Grid services based on the Open Grid Services Architecture [OGSA, FKT 04]. As a service-oriented architecture, OGSA supports the separation between users and applications through abstract service definitions. The users program against the service without knowledge about the internals of the service, and the application is flexible in how it provides the service.

The Open Grid Services Infrastructure (OGSI) is the basic part of OGSA. OGSI is based on Web services and primarily enhances them to support long-lived, persistent services that are often needed in distributed applications. Therefore it provides mechanisms for creating, naming, managing lifetime, monitoring, and grouping of services and for information exchange. One open source implementation of OGSI is the Globus Toolkit version 3 [Globus].

Based on OGSA, the Data Access and Integration Services Working Group of the Global Grid Forum [DAIS-WG] develops standards to access databases. The OGSA-DAI project [OGSA-DAI], which is part of the UK e-Science project, developed an implementation that supports the access to existing relational and XML databases.

We will define services that extend the OGSA-DAI services and that allow the access to all parts of an enterprise content management system. Thereby we will focus on the essential CRUD-operations to create, retrieve (including search based on metadata and full-text), update and delete the content. We will develop a prototypical implementation based on the IBM DB2 Content Manager. There already exists preliminary work from the eDiaMoND project [eDiaMoND]. But as they use IBM DB2 and IBM DB2 Information Integrator for their metadata management and not the catalog server with its stored procedures, their implementation primarily focuses on the content repository aspects.

First we can test those services by deploying one service on every node of our distributed database described above. This configuration allows us to use the Grid services for virtualization and load balancing, and to use the distributed database to do the information integration. In the long run we want to exploit technologies from peer-to-peer technologies for information integration. Some interesting projects in that field are edutella [NWQ+02] and Peer Data Management [HIST 03].

6 Summary and Conclusions

Based on an analysis of business scenarios in the realm of important content management domains, we provided an argumentation for and a step-by-step description of the transformation of a traditional content management system to an Enterprise Content Management system. Performance, scalability and business resilience have been identified as the major forces that drive this evolution.

Focusing on IBM DB2 Content Manager as reference point we outlined a first strategy to achieve what is called on-demand ECM services. Our approach is built on the ‚virtualization' of the ECM service and component architecture. We described and discussed virtualization at the database level using existing database technology and virtualization based on a dynamic services infrastructure which capitalizes on proven peer-peer and grid technology.

Further work will evaluate our approach by means of prototype measurements and analysis.

Literature

[BFM+03] Baragoin, C.; Fournil, A.; Miqui, A. H.; Latimer, G.; Schuster, S.; Zuzarte, C.: Up and Running with DB2 UDB ESE. Redbook SG24-6917-00. 2003.

[DAIS-WG] Data Access and Integration Services Working Group of the Global Grid Forum. http://www.gridforum.org/6_DATA/dais.htm

[eDiaMoND] http://www.ediamond.ox.ac.uk

[FK 04] Foster, I; Kesselman, C. (eds.): The Grid 2 – Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publishers, 2004.

[FKT 04] Foster, I.; Kesselman, C.; Tuecke, S.: The Open Grid Services Architecture. Chapter 17 in [FK 04].

[Globus] http://www.globus.org

[HIST 03] Halevy, A.Y.; Ives, Z. G.; Mork, P.; Tatarinov, I.: Schema Mediation in Peer Data Management Systems, Proceedings of the 19th International Conference on Data Engineering, 2003.

[IBM CM] IBM Content Management, http://www-306.ibm.com/software/data/cm/, 2004

[IGT] IBM Grid Toolbox, http://www-1.ibm.com/grid/solutions/grid_toolbox.shtml

[Schwab] IBM case study on Charles Schwab, http://www-306.ibm.com/software/success/cssdb.nsf/cs/BEMY-5UZP3F

[NWQ+02] Nejdl, W.; Wolf, B.; Qu, C.; Decker, S.; Sintek, M.; Naeve, A.; Nilsson, M.; Palmér, M.; Risch, T.: EDUTELLA: A P2P Networking Infrastructure Based on RDF, Proceedings of the Eleventh International World Wide Web Conference (WWW2002), 2002.

[OGSA] Open Grid Services Architecture, http://www.globus.org/ogsa/

[OGSA-DAI] Open Grid Services Architecture: Data Access and Integration http://www.ogsadai.org.uk/

[XHLL04] Xu, M.; Hu, Z.; Long, W.; Liu, W.: Service Virtualization: Infrastructure and Applications. Chapter 14 in [FK 04]