Enterprise Content Management for SAP

Stefan Momma
Manfred Oevers
Andrew Stalnecker
Jörg Stolzenberg
Introduction

IBM® Enterprise Content Management (ECM) portfolio software enables the largest organizations in the world to make better decisions, faster. By gaining control of structured and unstructured information, organizations can access, collaborate, and influence business decisions in new ways, making content a first-class source of insight.

With industry-specific IBM ECM solutions, organizations can capture, manage, and share content throughout the information lifecycle, helping to ensure compliance, reduce costs, and maximize productivity.

The IBM ECM portfolio includes a wide array of capabilities that integrate with existing systems to help organizations maximize the value of information, including IBM DataCap document imaging and capture, social content management, advanced case management, IBM Information Lifecycle Governance solutions, and IBM Content Analytics with Enterprise Search.

This IBM Redpaper™ publication describes the business goals that can be achieved by extending the SAP archiving infrastructure with IBM Enterprise Content Management portfolio solutions. It explains how the IBM ECM solutions support and extend the existing SAP data and document archive functions. It provides examples for enhancing the implementation of core end-to-end business process solutions by seamlessly integrating SAP and IBM ECM components.

This paper describes the following topics:
► “Enterprise content management business goals” on page 4
► “ECM for SAP use cases and solution architecture” on page 10
► “Business process enhancements through ECM for SAP solutions” on page 21
► “Data governance: Managing growth and compliance” on page 35
► “References” on page 43
Enterprise content management business goals

This section explains the drivers that motivate an organization-wide approach to content management. This approach enables a unified view and management of all business content.

First, this section describes the importance of a single source of truth for all of the information spread across an organization. Challenges are associated with such a requirement, such as managing the volume of information and fulfilling legal requirements for retention and disposal of information.

This section also describes the types of information in an organization:

- Information in databases, typically referred to as structured information
- The vast variety of all kinds of documents that exist throughout a company, both in paper and electronic form, known as unstructured information

This section shows how such a holistic approach naturally extends into the information contained in an SAP system, in the form of the raw SAP database content, and in the collection of documents associated with this content as attachments, reports, and other documents in textual or image form.

Finally, this section outlines how an integrated and federated solution based on IBM Enterprise Content Management portfolio software supports the goals of various stakeholders in an organization, such as line-of-business decision makers, service-providing information technology (IT) departments, and legal counsels. The solution, based on IBM Enterprise Content Management portfolio software, provides a common approach, a common collection of governing rules, and, as a result, a common view on all relevant business content.

Information lifecycle management: More than just archiving¹

Any decision-making process in a modern organization has to rely on the availability, timeliness, and accuracy of current business information.

This requirement is associated to several challenges:

- The amount of information that has to be collected, maintained, surveyed, evaluated, and managed grows at a massive rate.
- The vast variety of information resources, information duplication, and uncontrolled versioning makes an efficient discovery process extremely complex.
- An increasing number of rules and regulations influence how information has to be retained or disposed.

Managing the overall growth of data

People who are not directly involved in supporting an IT infrastructure are often surprised to hear that during the last five years, in a typical organization, there has been an average 50% year-over-year growth in the volume of data. However, that statistic is not surprising to IT managers, who struggle to find a way to support rapidly growing volumes of data with an often flat or only slightly increasing budget.

¹ Unless stated otherwise, when the term archiving is used in the context of this paper, it refers to the operation that stores content in a repository outside of the SAP system, where it is still active, and immediately accessible. Only in the case of data archiving, which is described later in this paper, does archiving refer to taking inactive content out of the database for performance maintenance reasons.
Also no surprise to IT managers is that much of the data under management is debris that is outdated and duplicated many times across multiple systems, with no real value to the business. Such over-retention results in direct and indirect costs on several levels:

- Overspending on a more complex IT environment:
  - More IT resources required to support larger systems
  - More storage and higher demands to maintain predefined system performance levels
- Higher costs for e-discovery processing. With larger amounts of information, much of it not valuable, processing any requests for information takes longer and results in higher review fees.
- More legal risk inherent with a larger information set.

The proliferation of collaboration products and social media platforms has added both to the diversity and to the volume of data that needs to be growth-managed, and emphasizes the need for an organization-wide solution to manage enterprise content.

**System performance**

Constantly growing amounts of data not only increase the cost of storage required to keep the data, it also has a detrimental effect on system performance of business-critical applications (apps). High volume accumulation of transactional data in a high-performance business application usually results in a deterioration of application performance, jeopardizing service level agreements (SLAs) for guaranteed response times.

**Storage costs use up the IT budget of an organization**

If storage of data is not actively maintained, and permanent growth of data without controlled disposal of obsolete data is permitted, an increasing percentage (not just amount) of the IT budget is spent on storage alone. The perception prevails that storage is getting cheaper, but the percentage of a typical IT budget spent on storage grows. Identifying and actively managing information debris can result in storage savings of up to 30 to 50 percent.

The IBM ECM product portfolio offers a family of content collection and archiving products designed to help curtail over-retention by attacking the problem at the source, and minimizing the amount of unnecessary information (the debris) that IT and other stakeholders must manage.

**Data governance and discovery**

Compliance and corporate information oversight demands require proper governance and defensible disposition of practically all types of content, including archived structured data.

**Information discovery is essential**

Access to all business-relevant information within the organization needs to be seamless and easy. The search, classification, and categorizing capabilities of IBM ECM portfolio products provide targeted results from the information sets managed by the content repositories.

A common client infrastructure based on IBM Content Navigator as the integration platform acts as a state-of-the-art front end to all discovery services. It is built on open standards using Hypertext Markup Language version 5 (HTML5) and Dojo components. This approach enables delivery of information to the correct client at the correct time.
An enterprise-wide approach is needed
The implementation of an organization-wide content management strategy with federated, deduplicated content requires that all players in the company’s decision process network are involved. They must work together on defining a common strategy and a common understanding of the rules that govern the management and organization of all of the company’s information throughout its entire lifecycle.

Information “islands” or silos must be abandoned. The uncontrolled evolution of diverging, and often contradictory, rules for governing content must become an integrated lifecycle model agreed upon by all stakeholders.

The information lifecycle and its governance
This section reflects on the lifecycle of information as it occurs in a typical organization, and provides observations on the constraints and challenges that come with it.

Inception of information
For most organizations, the delivery of a product or service depends on the exchange of documents that are part of the record for all transactions. How efficiently organizations manage documents can have a huge effect on the quality of the experience for their customers, patients, students, or constituents.

For all documents that have not been created electronically from the beginning, moving away from the handling of paper as soon as possible in the lifecycle is the first step in this direction. The need for a sophisticated capture solution to digitize the document content is a key requirement.

Smarter document capture uses technologies that convert documents to searchable images, automate data entry, identify documents, check data quality, and format data for adequate use by business systems. By automating labor-intensive, error-prone manual processes, IBM Datacap can accelerate document processing capacity, improve the quality of the processing results with significantly less manual intervention, and reduce cost.

More importantly, IBM Datacap can remove many of the obstacles that degrade service quality to help organizations create deeper engagements with their customers.

Content creation, revision, and approval workflows
In addition to digitized paper documents, electronically created documents encompass an increasingly large portion of the unstructured content that has to be managed. The creation process, and the revision with proper version control and collaboration between different actors that play a role in the creation of the document, can be modeled in the information lifecycle under the control of IBM ECM software.

In addition to their state-of-the-art storage organization, search, and retrieval capabilities, IBM content repositories provide powerful workflow engines that can be used to model business workflows, such as approval processes, based on an electronic document flow through the organization. Security and audit functionalities ensure that all business decisions follow a well-defined protocol in an auditable way, compliant with the governing rules.

These workflow capabilities can, and in many cases have to, be extended and intertwined with the business workflows modeled in the SAP infrastructure of the organization. Experience shows that organizations have business requirements for all of these processes inside and outside of their SAP systems. A critical task is to provide the correct information to the correct client at the correct time. Implementing fully integrated workflows is crucial to satisfy this requirement.
IBM Defensible Disposal

The IBM Defensible Disposal solution applies to the entire organization, which must include content that is also related to SAP systems. As mentioned before, treating SAP structured and unstructured content in a siloed solution counteracts the proposed integrated information lifecycle model, compromises the completeness and accuracy of discovery operations, and raises the cost of disposal under a common retention rule governance.

As indicated earlier, not all of the information that is created and stored in a content repository retains its business value indefinitely. Business transactions have a natural lifecycle. The information associated with the business transactions should be retained according to the retention rules, and subjected to disposal processes to keep data growth under strict control.

It is mandatory, under the legal rules defining compliance and auditability of the business processes, that the disposal process follows a common strategy across all divisions and departments that are involved in the handling of business records. Simply speaking, the decision about when certain content can be disposed of cannot default to the IT department, which might make such a decision indiscriminately, based on the age of documents.

The IT department may be unaware of the fact that certain types of content fall under different and often complex rules for their retention and disposal. An example of such complexity is the rules governing human resources (HR) documents about applicants that were not hired. In certain legislative regions, these documents must be disposed of within short time frames, for example, within a month after the final decision. Failing to do so can result in substantial fines.

Construction plans, design documents, and quality records for complex infrastructures, alternatively, have retention periods that are counted in decades rather than months. Retention periods of 30 or more years are not unusual in such contexts. Note that such complex rules are not only specific to document types or business applications, but can also be different across countries with different laws, which can make the required setup in a multinational organization even more complex.

Legal holds overlay the general rules for retention of different types of information based on type, owning organization, and so on. The legal holds apply whenever business operations documents are subject to litigation, and are put under a hold order issued by a court of law to prevent their disposal before the end of the investigation.

IBM Information Lifecycle Governance (ILG) solutions help change the keep everything approach by delivering capabilities that the various information stakeholders, including legal departments, IT, records management, privacy officers, and business users, can use to address information governance issues in their domain. These IBM solutions help to drive down the costs and risks associated with over-retention by using tools that enable the following outcomes:

- IT staff is enabled to understand and manage, by system and employee, the content collection, archiving, and retention criteria, and the procedures established by the organization. Furthermore, the staff can implement an archiving program that reduces cost by reliably retaining what is important and required, and deleting what is not.
- Attorneys and paralegals are enabled to automate legal hold processes, and coordinate evidence collections across the organization to respond to requests for information more quickly and cost effectively.
- Records managers are enabled to develop and manage global retention policies, and to coordinate compliance and disposition across multiple systems and jurisdictions.
Privacy officers are enabled to assess and communicate privacy duty by data subject and data location, including overlapping obligations.

IT staff is enabled to determine which systems appear to have the highest cost and risk profiles, and to enable them to address management of systems by information value.

These capabilities generate a more complete picture of the information inside an organization, and enable information management decisions based on fact and certainty. With this confidence comes the ability to implement a defensible disposal program that can have a real effect on the amount of data under management and the associated cost and risk.

The IBM Value-Based Archiving solution and, in particular, the IBM Content Collector family of products support defensible disposal efforts with a larger set of capabilities that attack the data growth problem at the source system. This approach immediately reduces the amount of information debris under management inside an organization, and the cost and risk associated with over-retention.

**Information lifecycle governance applied to SAP systems**

The previous sections of this paper explain the business requirements that justify the need for an organization-wide content management strategy that helps to meet the challenges imposed by data growth, discovery needs, and rules for legal compliance when handling business structured and unstructured information.

As mentioned, the information contained in an SAP system, which in the majority of cases is not just one source but a key source of business operations information, must not be treated as an isolated silo. It must be integrated into the overall strategy.

**Structured information in SAP systems**

Managing and controlling the size of SAP applications provides sustainability of system performance while monitoring infrastructure cost.

Application size management is primarily performed by moving infrequently used or obsolete business transaction data, which is no longer involved in active business transactions, to less costly storage while maintaining transparent access to such data. Examples include data needed only for summary reporting or infrequent auditing purposes.

The SAP system provides the tools to archive data that is declared *business complete*, and stores these in a proprietary aggregation format in archive files in a secure ECM repository. The SAP system also provides transparent access to the archived data directly from the ECM repository. The effects of this maintenance operation are two-fold:

- System performance is kept under control by limiting table sizes within manageable boundaries.
- Storage cost on high-performance storage that it is needed for immediate transactional processing of table data can be reduced, because infrequently used data is moved to less-costly storage.

The management of SAP data too often exists as an IT island, disconnected from the rest of the information management of the organization, retention, and information governance standards. However, as SAP implementations grow, and as new SAP enterprise resource planning (ERP), customer relationship management (CRM), and other component modules interact with more of the overall organization, an SAP archiving solution needs to be more comprehensive and interconnected.
This approach helps ensure proper information governance for SAP system data and content, both within and outside of the SAP system. In fact, the need for scheduled, defensible disposition of SAP data and content often drives the need for archiving itself. This motivation is on a par these days with infrastructure cost reduction and overall business efficiency drivers. For more information, see “Data governance: Managing growth and compliance” on page 35.

**Unstructured information in SAP systems**

Unstructured information is associated with most SAP business processes:

- Invoices in an accounts payable process
- Purchase orders and billing documents in a supply chain or financial application
- Generated reports summarizing business transactions in just about any application

This information frequently needs to be accessed from outside the SAP application:

- Customer service department employees need to see all of the information about the previous transactions of a customer. They want to provide the customer with a digital copy of their bill to reduce mailing cost.
- As part of a litigation, the legal department, which has no direct access to the SAP system, has to provide all of the contract documents from a given time period, matching certain textual search criteria, to fulfill a court order.

The critical business value in these examples is the ability to provide *all* business information, both structured and unstructured content, to the correct person at the correct time. Providing access to unstructured content directly within the business process enables the person to make the correct decision quickly. Details about some solution use cases are in “ECM for SAP use cases and solution architecture” on page 10.

In summary, both the structured and the unstructured information in an SAP system are an integral part of the overall information landscape of an organization, and therefore must be managed in the same infrastructure of ECM, following the same strategies, and the same governance rules as non-SAP-related content.

**IBM proposes a base structure of an integrated ECM solution**

The previous sections of this paper describe how an organization-wide content management strategy, including archiving and defensible disposal of information, extends the value of an organization’s SAP landscape. Several examples demonstrate how to enhance the implementation of core end-to-end business process solutions by seamlessly integrating SAP system components with IBM ECM components.

The following list summarizes these value points:

- The IBM ECM/ILG solution extends the SAP business solution by providing the correct information at the correct time within all business processes across an organization.
- A strictly value-based archival and disposal model enables an organization to maintain system performance and reduce infrastructure cost.
- An organization-wide retention and hold model agreed upon by all stakeholders provides a framework for maintaining full compliance with complex governing rules under a common defensible disposal strategy.
- Cross-system, end-to-end solutions that implement complete business processes can be achieved with the seamless integration of SAP business modules and IBM ECM products, such as IBM Datacap and IBM Content Navigator as a common ECM front end.
The IBM solution is integrated and certified by SAP. Moreover, IBM is rated as a market leader in this space by the leading industry analysts. “ECM for SAP use cases and solution architecture” on page 10 focuses on the SAP archiving process and its integration with IBM ECM portfolio components. It describes the core use cases and shows how organizations can use IBM Content Collector for SAP Applications to implement them.

The core use case can be extended with state-of-the-art document capture functionality implemented through IBM Datacap. Additionally, IBM Content Navigator is introduced as the document-centric integration platform of ECM.

ECM for SAP use cases and solution architecture

IBM ECM portfolio software can enhance the value of an SAP business infrastructure with document and data archiving support. IBM Content Collector for SAP Applications plays a central role as the gateway product that operates between the SAP system and the ECM environment.

“Business process enhancements through ECM for SAP solutions” on page 21 describes the integration with business applications, such as an accounts payable solution.

Before describing the use cases for SAP archiving (in “SAP archiving use cases” on page 11), the following high-level use cases must be introduced:

- Archiving data and documents originating from SAP systems.
- Making documents already stored in an ECM system available to SAP users by linking them to SAP Business Objects.
- Enabling SAP users to access documents stored in the ECM repository outside the SAP graphical user interface (GUI). Such documents can be anything, from scanned images, faxes, forms, and email to documents originating in electronic form. The documents stored in the ECM repository by an SAP system can also be made available to other non-SAP applications and capabilities of the ECM platform, such as classification, records management, and e-discovery.

SAP archiving standards

Starting with SAP Release 2.2, SAP clearly recognized the need to enhance the SAP software through a robust archiving facility, both for offloading the operational SAP database and processing business documents that are in an external archiving system. With SAP Release 3.0, all key applications and modules offered archiving functionality that was included as part of the standard SAP package. The SAP Business Framework provides the interfaces necessary to integrate these external functions with SAP systems.

SAP ArchiveLink

The primary interface for integrating storage and content management systems into an SAP system is called SAP ArchiveLink. It was introduced with Release 2.2 and enhanced in subsequent releases.

Additionally, the SAP Hypertext Transfer Protocol (HTTP) Content Server interface was defined in SAP Release 4.5 as a subset of SAP ArchiveLink that focuses on content rather than storage management. HTTP Content Server interface is a general, cross-application interface that connects the enabled SAP business applications to a content server, and enables them to process documents in logical content repositories.
This content server can be a database, a file server, an SAP system, or an external archive. The following list describes the supported SAP ArchiveLink 4.5 functions:

- HTTP Content Server interface
- Business Application Programming Interface (BAPI) for bar-code-based archiving scenarios and creation of SAP work items
- Object Linking and Embedding (OLE) functionality for storing inbound documents or PC files, and starting external viewing applications on Microsoft front ends

SAP ArchiveLink has not been changed since SAP Release 4.6. Note, however, that the current SAP ArchiveLink specification has made the requirement to support OLE optional, so vendors might drop support for OLE in future versions of their SAP ArchiveLink software.

Although SAP ArchiveLink is focused on the management of content and storage, it is not suited by itself to address compliance use cases, such as decommissioning existing systems, managing data retention rules, and collecting and preserving data for legal cases through legal holds.

**SAP Information Lifecycle Management**

SAP introduced SAP Information Lifecycle Management (ILM) to manage the lifecycle of productive or archived data and documents. With ILM, data and documents are stored in an ILM Web Distributed Authoring and Versioning (WebDAV) server, and the following retention management capabilities are provided:

- Placing retention policies on documents and data to determine how long they need to be kept in the archive to comply with governmental rules and regulations
- Placing legal holds on data and documents that are needed in the context of a legal case
- Enabling controlled disposal of documents that have reached the end of their retention period and are no longer needed as part of legal holds

For more details about ILM, see “Data governance: Managing growth and compliance” on page 35.

**SAP archiving use cases**

This section considers various use cases for SAP archiving and several important dimensions:

- Data versus document archiving
- Early versus late and simultaneous archiving
- Use of bar code technology
- Access to documents from outside of an SAP system

**Data archiving**

Business records that are no longer needed on a day-to-day basis can be packaged in an SAP-defined format called Application Development Kit (ADK) file, and archived using the SAP ArchiveLink or HTTP Content Server protocols. By doing so, organizations can keep the SAP database at a manageable size.

This archived data is still accessible by an SAP system in a transparent manner, while at the same time reducing storage costs, increasing productivity, and improving system performance. For more detail, see “Data governance: Managing growth and compliance” on page 35.
Document archiving

For legal and internal policy reasons, companies must keep documents pertaining to their business operations for a certain period of time. Filing the documents in paper form has several disadvantages because of their physical nature. For example, they must be duplicated when more than one user needs them. Tracking their physical location reliably can be cumbersome and error-prone.

Processing documents electronically has the following benefits:

- Organizations can use cost-efficient storage media.
- All authorized users can access the documents without being delayed by conventional archive inquiries.
- Several users can access the same documents at the same time.
- Disaster recovery (DR) procedures can be fully automated. Workflow processes, such as an approval procedure, can be defined consistently organization-wide, and can be fully automated.

The following types of documents can be identified:

- Incoming documents. These documents include, for example, supplier invoices that reach the company by way of mail or telefax, and are typically stored as digitized images.
- Outgoing documents. Documents created electronically, printed, and sent to their respective recipients. Archiving the electronic form of these documents enables for fast retrievals for customer inquiries or audits.
- Reports and print lists. A specific type of outgoing documents, generated by an SAP application, that are usually printed. By using an archive a physical printout becomes obsolete. The electronic journal, as opposed to its paper equivalent, is easily searchable, and can contain hyperlinks to other documents to enable convenient cross-reference. For example, a specific entry in a document journal might refer to a scanned original document. Furthermore, the archived lists can be used as input to other applications, such as data mining systems, for advanced analytics purposes.
- Desktop files. Documents that are created by applications, such as office applications or common agent services (CAS) applications. They can be archived and later accessed by an SAP system using Desktop Office Integration (DOI), based on SAP Business Connector to SAP Central Instance (BC-CI) integration, or using the document management system (DMS).

Early archiving

In early archiving, an incoming document is first captured into a repository, for example, IBM FileNet® Content Manager. It is then made available for processing by SAP applications before the associated SAP business object, for example, a sales order, is created. The incoming document drives the creation of the SAP business object. This approach eliminates paper processing from the beginning of the process, and separates the scanning process from the creation of the business object.

When the document is captured and assigned to an SAP document type, an SAP user is notified that a new document is ready to be processed. The user then creates the business object associated with the document type. The notification process uses SAP Business Workflow, and a link between the business object and the document in the repository is established.
Simultaneous archiving
In simultaneous archiving, all document entry and SAP object processing steps are carried out by the same SAP user. Overall, the process is the same as in early archiving except that the SAP work item, which is created at link time, is assigned to the current user.

Late archiving
In late archiving, the creation of the SAP business object comes first, and linking to the corresponding incoming document happens later in the process. In practical terms, this process is like a traditional paper-based process. The SAP business object already exists when the incoming document is captured into a repository, and a link between the business object and the incoming document is established.

Use of bar code technology
Document archiving can include bar code linking, whereby a bar code that is applied to the scanned document is captured by an SAP ArchiveLink-compliant repository and uploaded to the open bar code table of the SAP system for linking. Bar code scenarios are particularly suited when existing paper-based processes should not be disturbed by the introduction of scanning and archiving technology. Documents are still distributed in paper form to the point where the business object is created and the linking occurs.

Similar scenarios where the electronic processing and the physical handling of paper remain uncoupled for some time include the handling of receipts in expense processing, and the processing of contracts that require physical signatures at completion time. In both cases, the use of bar codes helps to bring paper and electronic records together.

Access to documents from outside an SAP system
The archiving use cases described so far are within the context of the SAP ArchiveLink protocol and specification. There is one use case, however, that goes beyond what is offered by SAP ArchiveLink.

All SAP ArchiveLink operations are based on the concept that the SAP database represents a master index catalog of all documents, including those whose content resides in the external archive. SAP ArchiveLink therefore passes no business information to the external archiving system except a Universally Unique Identifier (UUID), which the SAP software uses to identify a document throughout its lifecycle.

For many SAP users however, a requirement exists to access archived documents, including those from non-SAP contexts, independently of the SAP system (without using the SAP GUI). To support federated searching of the external archive by business information, such as customer number or fiscal year, the corresponding document attributes must be transferred from the SAP system to the external archive to be searchable there. This process is commonly referred to as index transfer.

This transfer of business information from SAP Business Objects referring to the corresponding documents in the external archive, to the external archive as document attributes, is usually achieved by proprietary function modules. These modules must be installed on the SAP system and configured to collect and transfer the required document attributes from the SAP system to the archive.
Implementing use cases with IBM ECM products

The previous sections describe the use cases that are relevant for SAP archiving, some of which even go beyond what is defined by the SAP ArchiveLink standard. The following IBM ECM portfolio products can be used to implement these use cases:

- IBM Content Collector for SAP Applications
- IBM Datacap
- IBM Content Navigator

Architecture of IBM Content Collector for SAP Applications

The architecture of IBM Content Collector for SAP Applications enhances the SAP business infrastructure with data and document archiving. With the components in this architecture, you can archive and view data and documents, and print lists. You can also link archived documents with SAP business objects.

Figure 1 shows the high-level architecture of IBM Content Collector for SAP Applications. Content Collector for SAP acts as a gateway that interfaces between the SAP system and the ECM repositories.

IBM Content Collector for SAP Applications supports the archiving of data and documents using both versions of SAP ArchiveLink into four back-end systems:

- IBM FileNet Content Manager
- IBM Content Manager Enterprise Edition
- IBM Content Manager OnDemand
- IBM Tivoli® Storage Manager
IBM FileNet Image Services can be used as a back-end for IBM Content Collector for SAP Applications, accessed transparently through the Content Federation Services agent for FileNet Content Manager.

The centerpiece of the architecture is the engine of the IBM Content Collector server, which distributes the incoming requests from an SAP system to the back-end systems and returns the responses back to the requesting SAP system through the use of dispatchers and agents.

The user interface for IBM Content Collector for SAP Applications is based on IBM Content Navigator, the common user interface of all IBM ECM repositories.

The following paragraphs provide more detail about the components of IBM Content Collector for SAP Applications.

**Server**

The server of IBM Content Collector for SAP Applications is the central component that handles all operations. It contains the engine and the components that connect to SAP and to all back-end systems. This set of components, operating in its own port range, is called an instance.

The engine processes all inbound and outbound communication in a bidirectional way. Communications include archival and retrieval requests from SAP and their translation into search, retrieval, and archival requests to the attached ECM repository also.

The connections to SAP on the left side of Figure 1 on page 14 are implemented as dispatchers, which can be started in a configurable number in order to address different workloads. The SAP ArchiveLink based dispatchers are RFC and HTTP based, depending on whether the newer 4.5 version or the previous 3.1 version of the protocol is used. The SAP ILM dispatcher is also HTTP based, but uses the WebDAV protocol.

The following archiving protocols that use SAP Business Connector are supported:

- SAP ArchiveLink (BC-AL) versions 3.1 and 4.5
- SAP HTTP Content Server (BC-HCS), which is a subset of the full SAP ArchiveLink specification, comprising the pure server-to-server communication without any front-end scenarios.
- SAP Information Lifecycle Management (BC-ILM), which is an SAP extension of the WebDAV protocol. WebDAV is the protocol for the Extensible Markup Language (XML)-Data Archiving Service, referred to as SAP DAS in Figure 1 on page 14.

The connections to the repositories translate the repository-agnostic requests from the engine into repository-specific requests by using the respective API. For each back end, a separate agent exists. The number of agents is also configurable in order to adapt to different workloads.

The IBM Content Collector Server is also scalable horizontally by starting additional instances of the entire collector server, where each instance can operate independently from the other by assigning individual port ranges. This mode of operation is also used when multiple SAP systems use the archival services that are provided by IBM Content Collector for SAP Applications.
Components related to the client interface

IBM Content Collector for SAP Applications adopted IBM Content Navigator as the basis for all of its client functionality as described in this section.

For more information about IBM Content Navigator, its integration and customization capabilities, and its extensibility, see *Customizing and Extending IBM Content Navigator*, SG24-8055.

As shown in Figure 1 on page 14, IBM Content Collector for SAP Applications provides a single client interface, which offers functionality to perform the following actions:

- Configure an IBM Content Collector instance.
- Create and administer profiles for archiving, document linking and index transfer.
- Administer, schedule, execute, and monitor tasks based on those profiles.

It also integrates with the advanced document viewing capabilities of IBM Content Navigator, based on IBM Daeja™ ViewONE technology.

The plug-in architecture of IBM Content Navigator allows for a seamless integration with IBM Content Navigator components of other ECM products such as IBM Case Manager or IBM Datacap.

IBM Content Collector for SAP Applications uses the IBM Content Navigator plug-in architecture in order to provide the functionality mentioned previously. See Figure 2.

![Image of IBM Content Collector for SAP Applications configuration using IBM Content Navigator](image)

**Figure 2** IBM Content Collector for SAP Applications configuration using IBM Content Navigator

**Configuration feature**

In the instance configuration feature, administrators create and maintain all IBM Content Collector for SAP Applications instances. One configuration feature (in one Content Navigator instance) can maintain instances on multiple hosts, covering multiple SAP systems (in general one per instance), and multiple back-end repositories.
One instance configuration collects all the necessary information about the instance itself and the characteristics of the participating systems on the SAP and the back-end repositories side such as these examples:

- User credentials for the SAP system and the back-end repository
- Communication details:
  - Communication protocols and ports
  - Security options and certificate handling
- Logical archive configuration:
  - Document classes that correspond to SAP document types
  - Folder structure of the logical archive
  - Document metadata mapping
  - Viewing options based on document type and MIME type information
  - Repository-specific configuration items

The complexity of the logical archive configuration depends on the choice of the back-end repository and, therefore, the functional capabilities that are supported by that repository.

The configuration feature uses active connections to the SAP system and to the back-end repository to provide as much information about the running systems as possible. With this information, you can perform consistency checks across the entire configuration as early as possible, and the information significantly reduces the number of configuration errors that might occur in a purely manual operation.

**Document viewing options**

Depending on the SAP business application that is used, the types and formats of attached documents can vary greatly. With that, the need for a highly configurable set of options for viewing these documents is necessary. Traditionally, the SAP system includes many options for viewing documents from within the SAP GUI, but external viewers can also be added and integrated.

The availability of IBM Content Navigator as the common user interface for the content repositories adds even more flexibility to this use case. IBM Content Navigator includes the IBM Daeja ViewONE viewer, which is highly configurable for many viewing scenarios. The logical archive configuration section in the configuration feature provides the options to activate these viewing methods.

**Administration feature**

In the IBM Content Collector for SAP Applications administration feature, for each SAP archiving use case (described in “SAP archiving use cases” on page 11) profiles are created that describe the operation performed and the content parameters of that operation, including the following information and details:

- Information about the source of the documents that are processed, that is, whether they are from an external source, such as a capture solution destination, or whether they already reside in a content repository.

- Information about specific document linking methods, to distinguish between these:
  - SAP barcode processing
  - The creation of SAP work items that trigger business specific SAP workflows

- Details about the mapping of attributes that synchronize document properties with corresponding business object metadata in the SAP system
Under this general setup, three types of profiles are configurable:

- Archiving profiles
- Document linking profiles (which also cover the special case of manual linking)
- Index transfer profiles

**Archiving profile**

An archiving profile is used to describe the necessary transactions to transport a set of documents from an external source into the repository, and simultaneously create the association of these documents with corresponding SAP business objects. (In that sense, every “archiving” operation is implicitly combined with a “linking” operation.)

The archiving profile must provide the following set of information:

- The origin of the documents
  These can be from an external capture solution, or from some other type of application that deposits documents into a file system location that the archiving process monitors for new entries, which are then archived.
- The target document class on the back-end repository
  This controls the set of metadata that will be assigned to the document.
- The linking method and the corresponding document type on the SAP system side, if the linking method is Create Work Item.

Two basic linking operations are available, as specified by the SAP ArchiveLink protocol:

- Creating an SAP workflow work item
  Based on content that is placed onto a work queue in FileNet Content Manager or into a work-basket in IBM Content Manager Enterprise Edition repositories, SAP workflow work items that are based on standard work tasks can be created and be configured to start the appropriate SAP transaction according to each SAP document type.
- Creating SAP external bar code entries
  Based on barcode information that is present as document metadata, a link to the associated business object is established if the values of the open external and the internal barcode tables match.

**Document linking profile**

A linking profile contains all the necessary information for performing the operation of associating documents that are already present in the document repository with SAP business objects in the SAP system.

Similar to the archiving profile, the linking profile also distinguishes between the two main methods of creating the SAP-based information:

- Create a work item, as described for the archiving profile (see “Archiving profile”).
  In addition to the triggering of a specific workflow on the SAP system side, additional metadata (taken from the document's metadata) can be specified. These are then added as metadata to the SAP business object during the document linking action.
- Create external SAP bar codes (as described in “Archiving profile”).
**Index transfer profile**

In some business scenarios, you might need to access documents that are related to SAP Business Objects outside of the SAP system, but within the business context. Documents that are created in an SAP system, and archived in IBM Content Manager Enterprise Edition, IBM Content Manager OnDemand, or FileNet Content Manager by using IBM Content Collector for SAP Applications, do not contain any searchable business data-related attributes by default. To support such a scenario, IBM Content Collector for SAP Applications provides a function to transfer business data from SAP to the ECM repository as document properties. This capability is called index transfer.

Through the use of folders, a structured document hierarchy that, as an example, exists in a contract management solution in SAP, can be mirrored in the content repository, providing seamless access to documents from within and from outside of the SAP system.

In an index transfer profile, you specify the required parameters to synchronize selected metadata between business objects in the SAP system and their associated documents in the content repository (for repositories that do support enrichment with metadata).

An index transfer profile includes the following information:

- The SAP business objects tables involved (for example, a BKPF table in an FI application). These can also be user-defined tables.

- The document types that should receive the metadata.

- Mapping information that associates SAP business object metadata with configured properties of the document types in the repository. The corresponding data types must be compatible and commensurate.

**Operation feature**

The operation feature of the IBM Content Collector for SAP Applications plug-in is the place where the profiles that are created in the administration feature are put to use in day-to-day operations.

Based on these profiles, tasks are created that describe the operational aspects:

- Task schedules
  
  Each task can be configured to run only once or repeatedly.

  Recurring tasks are assigned a task schedule, describing on which day of the week at which hour and minute the task is scheduled to run. With this function, you can plan task resource usage according to anticipated system load and resource availability.

- Task status
  
  Planned tasks can be monitored for their individual progress, and for their overall status at task termination, by indicating the number of processed items and potential error status.

  Recurring tasks can be suspended and resumed.

- Task auditing
  
  For all tasks that are administered through the operations feature, the task manager component of IBM Content Navigator provides auditing facilities to document the task activities.
Client API
External applications, for example document capture solutions such as IBM Datacap, communicate with the Collector Server by using a public client API. External applications can integrate with IBM Content Collector for SAP Applications through the use of this client API. The client API supports the archival of documents and the subsequent action of either creating SAP Work Items or sending bar codes to link documents to SAP Business Objects.

IBM Datacap is integrated in this way, and integrations with IBM Business Partner applications were also implemented. For more details, see “IBM Datacap” on page 20.

Figure 3 on page 21 shows the IBM Datacap integration schematically. A document is first scanned and processed by IBM Datacap, which extracts certain information, for example, an invoice number, from the scanned document. This information is written to a properties file, and the custom IBM Datacap action calls the client API and passes the relevant information to it. Metadata, such as the invoice number, can be stored in the archive or the SAP system along with the scanned document.

Scalability and high availability
IBM Content Collector for SAP Applications architecture is designed for scalability and high availability (HA).

The IBM white paper IBM Content Collector for SAP Applications — Sizing, Configuration, and High Availability explains how to size an IBM Content Collector for SAP Applications solution, and how to scale it and configure it for HA. It is available at the following web page: http://www.ibm.com/support/docview.wss?uid=swg27036773

The SAP-driven scenarios for outgoing documents can be made highly available by using standard techniques, such as a load-balancer for the HTTP traffic.

IBM Datacap
The ability to efficiently convert paper records into digital form by scanning and making them available to SAP systems as an incoming document is a key ingredient in any initiative for reducing manual paper processing. See Figure 3 on page 21.

IBM Datacap quickly and easily captures, manages, and integrates enterprise content while extracting critical information. IBM Datacap offerings include easy-to-use customization with high-volume document capture. The IBM Datacap product family consists of many components. In the context of this paper, consider the following components:

- IBM Datacap is based on a service-oriented architecture (SOA) capture and automation solution that includes both web and thick clients. It is a document ingestion product.
- IBM Datacap Studio is used to configure the IBM Datacap solution by defining and assembling the document hierarchies, recognition zones and fields, rules, and actions.

Like IBM Content Collector for SAP Applications, the user interface of IBM Datacap is based on IBM Content Navigator.

For more information about IBM Datacap, see Implementing Document Imaging and Capture Solutions with IBM Datacap, SG24-7969-01.
Business process enhancements through ECM for SAP solutions

Even in a “paperless world,” the majority of today’s business processes are document-centric. A tight and smart combination of the strength of SAP in ERP systems with the sophisticated capabilities of the IBM ECM infrastructure enables business users to benefit from an integrated cross-system workflow management. An important aspect to understand is that those solutions and enhanced features go significantly beyond the SAP ArchiveLink specification, which requires only a simple imaging and archiving system.

Objectives of a document-oriented workflow management

The operation of a document-oriented workflow management system usually pursues one or many of the following objectives:

- Enable end-to-end processing
- Increase the degree of automation
- Decrease error rates
- Reduce handling expenses
- Relieve users from routine tasks
- Adhere to compliance rules
Examples include HR file management, invoice processing, contract management, processing of order sheets, and delivery receipts.

**SAP-centric versus ECM-centric process management**

Basically, two fundamental approaches exist for an integrated workflow management:

- **ECM-centric.** A separate business application that communicates through interfaces with the SAP system. This application is typically an integral part of the ECM system, either a workflow component or a case management application.
- **SAP-centric.** The business application is deployed directly on the SAP system.

The ECM-centric process management configuration is usually covered by functions of the ECM systems. Their archiving and imaging capabilities are needed to satisfy the SAP ArchiveLink use case, but they are typically full-featured ECM systems. Configuration and maintenance are in general easier than they are for the complex SAP ERP system. However, the main disadvantage is that SAP is still the leading system from a business perspective, because all of the business data is kept in the ERP system.

More than that, all related data is stored there. For instance, the processing of invoices requires access to purchase order or even vendor contract data. An accounts payable solution running inside the ECM system has to synchronize all the basic and related data with the ERP system, which process can be error-prone and difficult to manage.

Business applications directly deployed on the SAP system ensure an integrated workflow management by default. Typically, they are installed as an ABAP module within its own name space. Therefore, they behave like any other standard SAP module. The following list describes the benefits of this approach:

- The entire workflow management happens within a single system. The user does not need to learn different UIs.
- There is direct access to data of the underlying SAP system. Therefore, no configuration or maintenance of external interfaces is necessary.
- Any relevant business data from any other SAP module required for processing an incident is directly accessible.
- Even non-SAP GUI users can easily be included through web UIs.
- Not a standard capability, but SAP Online Service System (OSS) integration can be highly advantageous.

The SAP integrated business application should be integrated into the SAP OSS. By doing so, the support process for the business application is the same as for the SAP system itself. Ideally, a support case is directly routed from SAP level 2 support to the application provider.

Further considerations in this section are based on the SAP-centric process management approach because, based on practical experience, this is the preferred approach for SAP-centric organizations.

However, a fundamental requirement for both approaches is that the ECM system is well integrated with the SAP system. This consideration goes far beyond support for the SAP ArchiveLink interface. It includes synchronization of documents, their metadata, and embedding them into the business context between both systems.
Components of an ECM for SAP Solution

Solutions for an integrated workflow management typically consist of the following components (see Figure 4):

1. Capturing the documents.
2. Processing document content and integrating it into the SAP business process.
3. Filing documents and archiving them into the repository.

These components can be delivered either from multiple vendors or from one source. A single vendor is no guarantee of a high-quality integration, because solutions from a single vendor are often compiled through acquisitions. The quality of the integration, which is an important criteria for a seamless end-to-end process, is sometimes revealed during its usage in the field. The same concept applies to solutions where multiple vendors work together and one of them appears as the main contractor.

Fortunately, ECM systems adhering to the established SAP ArchiveLink standard enable a flexible and independent combination of capturing and repository system.
Capturing solution components

This section has more detail about the individual components of the capturing solution.

Capturing

The capturing process extracts the business-relevant data from documents coming from various sources and converts them into an electronic format (see the example in Figure 5).

![Invoice Image]

*Figure 5  Examples of invoice data to be recognized and captured*

In Figure 5, the areas enclosed in red frames represent the data that must be extracted:

- The vendor name and their postal address
- The date and serial number of the invoice
- Identification of the purchase order corresponding to the invoice
- Position data
- Total sum of the invoice

The documents can be submitted either in hardcopy form (for example, invoices, delivery notes, applications, and so on) or in an electronic format (for example, through electronic data interchange (EDI)). For EDI, the format conversion can be limited to a pre-processing to make it workable for further processing in the SAP system. This conversion can even be omitted if both business partners have already aligned their exchange formats. The format is standardized, for example, in the automotive industry.

Even in times of electronic documents, the capturing of business data from paper documents remains an indispensable function for the foreseeable future. For invoices and delivery notes in particular, the flawless capturing of the header and position data is a basic requirement to achieve a high throughput without manual interaction.
The recognition of the semantic content (in which position on the various forms can that specific invoice information be found) is a special requirement in a mass capture scenario. For instance, which content represents the order number, the invoice number, address data, invoice date, and so on. Of course, the position data of multi-page invoices and delivery notes has to be recognized accurately as well.

In contrast, OCR recognition (capturing data and transforming it into electronic character sequences) is the easier task.

The following technologies are the basic approaches to the challenging task of recognizing semantic content:

- Free-form recognition
- Form-based recognition

Sometimes the technologies are applied in combination, such as in IBM Datacap. These technologies have a demanding theoretical background, for which a detailed description goes beyond the scope of this book.

Free-form recognition identifies the semantic meaning from the content itself. Form-based recognition concludes the semantic meaning from the position information, for example, a previously performed training mode tells the software at which location on the page to find the invoice number. The assumption here is that the vendor uses a certain standard layout for their invoices, which is typically a safe assumption.

The need for a training phase is not a challenge that is unique to form-based recognition. Free-form recognition requires a certain training also. The payoff for the effort spent for a precise adjusting of the recognition to a particular form is a high automation rate without any manual interaction in production. In the retail industry in particular, with its numerous vendors and a huge number of bills and receipts, a high automation rate can cause a significant increase in efficiency.

The details of mass and batch processing, such as holding or re-sorting of batches, monitoring, and so on, are not described in this paper.

An important aspect of an integrated process management, however, is the connection with the SAP system. The transfer of the scanned documents occurs together with the extracted data in XML format, as a batch or as single documents. It can be achieved either through a shared directory in a universal and flexible way, or through a direct network connection, for example, using web services.
Depending on how the invoice approval is organized, a first comparison of an invoice with the corresponding SAP purchase order data can be a reasonable approach, as shown in Figure 6. This approach does not substitute the final invoice processing in SAP Financials (FI). However, it can help sort out obviously incorrect documents at an early stage, in addition to correcting minor deviations, such as splitting a single order item into two invoice items.

![Comparison of invoice items (left side) with order items (right side) during recognition process](image)

**Processing**

The question can arise: Why introduce an additional module from a third-party vendor for accounts payable, for instance, when SAP already provides predefined workflows for invoice processing in it, as standard, in the SAP Financials (FI) module? Furthermore, predefined tasks are already part of the SAP ArchiveLink customization for the late archiving and early archiving scenarios, for example, TS3001128 and TS3001117 (see “SAP archiving use cases” on page 11).

The simple answer is: The main benefit is adding the ability to extend functionality and to simplify configuration and operation. These benefits are illustrated by the examples in the following list:

- Better support for mass processing

  At first glance, the standard SAP ArchiveLink scenarios provide the capability for mass processing of inbound documents, for example, late archiving with bar code. However, by default, the SAP system is not prepared for the flexible processing of extracted data coming from a capture system.
Posting FI records with externally captured invoice data is feasible through the SAP transaction MIRO (Enter Incoming Invoice). However, a flexibly configurable job control and queue management to buffer, control, and monitor large amounts of input data as shown in Figure 7, is available through only the added third-party module.

The upper section of Figure 7 shows all elements of the input queue (this example has only two) and the lower section can display the corresponding purchase order items for the selected invoice. Other options, such as displaying the invoice line items, contract data, and so on, are configurable too.

This centralized queuing and monitoring is a key capability in larger deployments. As an example, the capability to serve the invoice processing for 53 SAP systems simultaneously was the main criteria for vendor selection at a large German automotive supplier.

Figure 7  Centralized queue management and monitoring for all incoming input: Capture application example
Simpler process configuration when compared to native SAP Workflow

The processing of invoices requires a simple configuration of workflows in conjunction with the option to modify them easily and on short notice. For example, you can reroute the invoice processing from one operator to another, or include an additional approver, as shown in Figure 8.

Figure 8  Simple and flexible configuration by adding another operator to an approval process

Seamless integration with office systems

Examples of this integration include dragging an email with an attachment into a personnel file to file it, as shown in Figure 9, or compiling response letters directly from the accounts payable solution automatically.

Figure 9  Filing an email into digital personnel file by dragging it
Easy access to data and functions of other SAP modules directly from within the workflow application

The resolution of any discrepancies during invoice processing often requires queries from other modules. For example, access to the vendor contract is needed to validate discount conditions on the invoice. Figure 10 and Figure 11 on page 30 illustrate this example where a vendor contract can be verified directly from within the workflow.

Figure 10   Direct access to the vendor contract from the invoice receipt list (part 1 of 2)
Figure 11 shows the accessed contract details.

Figure 11 Direct access to the vendor contract from the invoice receipt list (part 2 of 2)

- Bidirectional coupling between the capture system and the repository

  A holistic and integrated process also requires a tight coupling between the capture and the archive system. Making the purchase order and vendor data available during the capture phase provides an easy way to rule out any misrouted documents in an early phase of the process, before they get into the invoice receipt book.

Repository and filing system requirements

The back-end repository and filing system must be able to fulfill a set of requirements that extend beyond the bare ability to safely store digitized content in an efficient manner. This section outlines these requirements.

**Basic requirements**

The main function of the repository system is to safely store the digitized documents, and to provide fast retrieval later when accessed from an SAP transaction. The technical connection occurs mainly through the SAP ArchiveLink interface. To some extent, the simplicity of this interface in its pure form limits the ECM vendors, because only low-level archiving functionality is provided.

In a pure SAP ArchiveLink solution, the only connection between the SAP system and the archiving system is the document identifier. The SAP ArchiveLink interface does not provide a way to transport any additional, potentially valuable, information, such as user information, business context, and so on. The need for extra features that add value to the solution available in today's ECM systems becomes apparent as soon as the documents filed from the SAP system are used outside of the standard use cases.
Beyond the initial low functional requirements for an SAP ArchiveLink content repository, be sure to look more closely at the additional capabilities that such a repository must provide to be suited for the purpose:

- Scalability to process large volumes of documents.
- Stability of the vendor, because the purchase and operation of an archiving system are inherently designed for long-term usage.
- Flexibility connecting storage subsystems, or the capability to directly connect to a hierarchical storage management (HSM) system. IBM Content Collector for SAP Applications offers these capabilities.
- Pre-processors and user exits to enable additional operations, such as customized document format conversion. For example, you may want to convert a document from Tagged Image File Format (TIFF) to Portable Document Format (PDF).

**Requirements for integrated processing**

As mentioned previously, an integrated ECM SAP solution requires a direct link between the SAP system and the ECM system, which goes far beyond the capabilities defined in the SAP ArchiveLink specification. This sophisticated link basically means that the SAP business context to which the SAP-linked documents in the ECM system belong is at least partially synchronized between SAP and the ECM system.

The following examples show what this criteria means in practical terms. Consider an invoice with an attached document as a starting point (see Figure 12).

![Figure 12 FI invoice with an attached document](image)

In the heterogeneous SAP/non-SAP enterprise environment that is the basis of our examples throughout this book, business users work on the ECM system, where they operate on documents from the SAP system and on documents of non-SAP origin. These workers want to see not only the bare image linked to the FI invoice, but at least some of the SAP business metadata assigned as properties to the particular document.
If the workers cannot see this information, there is no chance to properly identify and retrieve the document on the ECM side, because the only default attribute is the document title (see the lower right side of Figure 14 on page 33).

An example of such a business context as it can be presented in an ECM environment is shown on the left of Figure 13. The UI is based on IBM Content Navigator. For a certain vendor, the associated invoices with their SAP document name are shown. The folder name at the lowest level reflects the SAP invoice number, and the folder contains the attached documents. The invoice folder plus a folder containing the vendor’s contracts represent an excerpt of the associated business context.

Both folders are structured as subfolders of a certain vendor. This example can easily be extended with further business information, depending on the information requirements in this context.

An important point to recognize is that the business information in the SAP system (the source of the information in the ECM system), and the representation of that information in the ECM system, must be kept in sync. From a business perspective, SAP is the leading system. Therefore, all of the information accessible in the ECM system is derived from the SAP system. As information changes (some of it never, some of it quite frequently), it must be reflected on the ECM side in near real time.

![Figure 13](image.png)  
*Figure 13  Access to SAP business data and context from within the ECM system (example)*
In some cases, the ECM users might need to directly access the original SAP FI business object. For this purpose, the invoice folder in our example contains a link to the FI document (see center section of Figure 14). When clicked, the SAP web GUI opens and the user is taken directly to the invoice record transaction without having to navigate the SAP system manually (see Figure 14).

Figure 14  Access to the original SAP FI record from within the ECM system

Figure 15 shows the same information in Microsoft Office.

Figure 15  Same representation of the SAP business context in Microsoft Office
ECM SAP solution architecture

This section describes, as an example, the high-level architecture of the invoice processing frequently used when working with IBM and SAP Business Partners who specialize in providing SAP integrated solutions.

Figure 16 shows the architecture and the process flow through the components. The general use case of the integrations enabled by this architecture is to automate the capturing, processing, linking, and archiving of incoming documents to SAP processes. The example described in this section and depicted in Figure 16 is based on an invoice processing.

Figure 16  Basic architecture and example process flow

The following list describes the architecture components shown in Figure 16:

- IBM Datacap
- SAP ERP system (SAP ERP Central Component (ECC) 6 or later)
- ECM-centric business application, deployed on the SAP system
- IBM Content Collector for SAP Applications
- IBM repository:
  - IBM FileNet Content Manager
  - IBM Content Manager Enterprise Edition
  - IBM Content Manager OnDemand
  - IBM Tivoli Storage Manager

The following activities are included in the flow shown in Figure 16.

**Processing in IBM Datacap:**

1. Scan the invoice.
2. Capture header and invoice line item data.
3. Access corresponding order data from the SAP business application and verify that the invoice line items match the order data.
4. Send the scanned invoice image and the verified invoice data to the business application in the SAP system. Create a work item in the SAP system for further processing.
**Processing inside the SAP business application:**

5. Start and run the approval process. Archive the invoice image using SAP ArchiveLink and IBM Content Collector for SAP Applications, and link it to the work item. Post the invoice, create the invoice record, and finalize the processing.

6. The invoice document can now be retrieved as an attachment to the invoice record.

Figure 17 outlines the protocols and technical interfaces between the components.

---

**Data governance: Managing growth and compliance**

The previous sections in this paper extensively describe the document-centric approach to ECM in an SAP context. This section focuses on the use of IBM Content Collector for SAP Applications in combination with IBM ECM repositories for SAP data archiving, following the standard SAP ArchiveLink and the SAP ILM data archiving models.

**Business drivers**

“Enterprise content management business goals” on page 4 provides information about how structured data (the content of the database underlying the SAP system) contributes to a large extent to the data growth that any SAP system has to maintain and control. At the same time, this type of data also falls under the regulatory set of rules regarding its retention and disposal.
According to their usage patterns and their lifecycle characteristics, the following variations of structure data must be distinguished:

- Data that occurs in high volume, typically from short-lived transactions, with corresponding short-term business relevance.
- Data that falls under regulatory control due to compliance legislation, such as the Health Insurance Portability and Accountability Act (HIPAA), Sarbanes-Oxley, or other laws, typically with long-term retention requirements.
- Data that falls under the previous category of long-term retention requirements, but that resides in SAP systems that have reached, as a whole, the end of their system lifecycle.

The following sections provide more detail on each case.

**High volume, short-lived transactional data**

A large portion of the data stored in an SAP system originates from transactions of diverse time duration. Consider how long this data must be available instantly, as opposed to only occasionally, only in summary form, or whether it is no longer needed at all. These considerations all determine the archiving strategy.

Purchase data provides an example with a typical lifecycle, starting with the initiation of the purchase, followed by the process of an order, the execution of the order, delivery of the goods or services ordered, the billing process, and finally the payment.

After the final payment, individual order data typically only has to be available for auditing or summary reporting purposes. The usefulness of the data and the need for immediate access to it diminishes rapidly. As an element of a closed business process, the data is no longer of immediate importance. This type of data is ubiquitous, but in particular, and with increasing frequency, this is the case for low-cost purchases of goods and services through the Internet.

For example, the purchases of music, ebooks, and video downloads, which happen instantaneously, with delays only in the order of magnitude of minutes between the order, the delivery, the billing, and the closing of the payment process. After this point, the individual transaction is considered completely closed, it is no longer of value in the live system, and there are no further actions necessary on it other than for purposes that are mostly statistical in nature.

Also part of the nature of such transactions is that they are extremely high in frequency. Therefore, they contribute massively to the data growth in the corresponding database tables that represent the transactions.

At the same time, the performance of the database in such a business scenario is of utmost importance. Therefore, a mandatory step is that preventive maintenance of the database is performed on a regular basis to preserve excellent performance for access to the database tables, and to keep transaction execution time to a minimum.

These high-frequency transactions contribute significantly to database growth, with all the associated consequences of additional cost for storage, administration, and other volume-related cost factors. Their usefulness in the database is clearly time-bound.

An operational method to address this issue is required. Obsoleted transaction data must be moved to sections of the environment that do not need to deliver high performance, and can, therefore, be provided at a lower cost. This approach does not jeopardize the still-required, albeit limited, access to the data.
Business data with regulation for long-term retention

On the other side of the lifecycle spectrum are the situations where data has long rather than short lifecycles in an organization. Industries, such as pharmaceutical companies or healthcare providers, have to comply with regulations that require their data and documents to be accessible in their original form for 20 years or longer. Although the active part of the lifecycle of such business data can be rather short, legal regulations increasingly require that such data remains accessible for auditing, litigation, and other documentation purposes.

Figure 18 shows the lifecycle for business data records (and the associated attachments), with the relative importance of the data displayed as a function over the life term.

Figure 18 shows the following data:

- Starting with creation time, and during the period where changes to the business data are still happening, importance is high, and remains so during the phase where immutability is reached, and finally the transaction can be marked as business complete.
- After the business complete point, the importance of the data drops significantly, until the decision is made to archive and offload the data by moving it out of the live database.
- The individual business data is still accessible, and might encounter short periods of increased relevance.
- During business audits, larger collections of archived business data are retrieved. During these periods, the data is of higher importance to support passing the audit objectives.
- The highest revival of importance that archived business data experiences is during the imposition of legal holds, typically as part of a litigation. The legal hold prevents unconditionally the disposal of data for the duration of the hold.
- After the hold is released, the rules of retention become reactivated, and the business data eventually reaches the end of life, at which point it can (and in many cases must) be finally disposed of.
To fulfill these requirements without clogging the live SAP system, while at the same time maintaining accessibility of the data, SAP protocols for archiving and retention can be enhanced significantly through the addition of IBM middleware and secure storage solutions that ensure the immutability of the archived data.

**SAP system consolidation and decommissioning**

Another important core scenario for the archiving of inactive data is found in the context of *system decommissioning*.

Typically, SAP systems are not monolithic, long-term systems in an organization-wide SAP landscape. Instead frequently the case is that production systems are regrouped, reused, consolidated, or reorganized on a regular basis. With each of these operations, the question of data preservation as described in “Business data with regulation for long-term retention” on page 37, must be addressed. It is not always necessary, as described previously, to preserve all data from existing live SAP systems in its live state.

Regulations for retention and controlled destruction of business data apply to decommissioned systems just as they do for data from live systems, and therefore the triage of data (described in “Leading practices for establishing data archiving” on page 41) applies equally to the decommissioning case.

All stakeholders in the business units of an organization must be closely involved in the process of deciding which business data must be preserved, over what periods of time, at what level of granularity, and when the data can or has to be securely disposed of.

Different models can be applied, depending on the amount of data that must be preserved from the system to be decommissioned, the business relevance of the data over time, and the need for immediate or delayed accessibility. The following list describes some of the options:

- Complete preservation of a system in virtualized form (made possible by current virtualization techniques).
- Transfer of the data into a consolidated new system.
- Export of data subsets into formats that can be accessed and interpreted without the need to restore into a live SAP system. The Data Retention Tool (DART) format is an example that auditors often accept as evidence, because there are viewers for the format available. The viewers provide all of the audit-relevant information, without the need to keep the data in a live SAP system.

**SAP infrastructure for data archiving**

SAP provides an infrastructure with data archiving capabilities that supports the business needs outlined in “Business drivers” on page 35. It fulfills the business needs by providing archiving capabilities without system interruption, and it ensures continued read access to the archived data in a transparent manner. Additionally, compression technology built into the archiving process provides additional business value by significantly reducing the cost that is directly caused by the volume of storage.

SAP data archiving comes in the following base varieties:

- Standard data archiving using the SAP ArchiveLink protocol that is also used for document archiving. This archiving method treats the data archive as a constantly growing data store, with no explicit means to control or enforce disposal. Read access to the archived data is also achieved in a transparent manner using SAP ArchiveLink.
- Archiving in the realm of SAP ILM, with the added benefit of an explicit method to control retention periods for defined collections of data. It also provides the ability to set legal
holds that override standard retention and disposal schedules for those cases where data has been put under the control of a legal dispute, and must be preserved until the dispute is closed and the hold can be released.

ILM is based on a WebDAV model, relying on established standards for XML-based archiving, rather than on the proprietary ADK archiving format. From a business perspective, the decision to use ILM-based data archiving has to take into consideration that SAP licenses ILM separately, at additional cost, and that it requires the presence of SAP Business Warehouse.

SAP promotes this type of archiving also in the context of system decommissioning, where data from existing systems that are no longer actively used still has to be read-accessible for compliance reasons. It should be noted that external vendors offer comparable functionality with respect to retention and legal holds capabilities, without the need to implement a full SAP ILM solution.

The following sections provide more information to help you choose between the two models based on operational and cost considerations and how they fit into the overall concept of enterprise content management using IBM ECM repositories.

Data archiving and the choice of IBM ECM content repositories

Structured data (pure database content) has limited value outside of the SAP system context. Its business value and its contribution to business decisions is often significant, but it is linked strictly to direct SAP operations that are executed and evaluated from within the SAP system. Operations such as content search or other types of non-SAP analytics typically do not apply to this data, unless it is extracted and transformed through other methods outside the scope of this book.

This state of affairs directly influences the choice of repository that is best suited to act as the ECM back-end for this type of archiving. The repositories can be selected on the basis of their performance characteristics and their cost effectiveness. The selection does not have to be based on advanced content use capabilities, such as content search, metadata analysis, or content aggregation.

Data archiving based on SAP ArchiveLink can be implemented with all four IBM ECM content repositories. The decision about which one to choose can be made based on the following conditions:

- Availability of existing repositories in the organization
- Archiving needs for unstructured data
- Cost considerations
- Availability of administration skills for a particular repository type

All four repository types support a tiered approach to data storage based on access performance needs, and ensure a high level of data protection.

In the most basic of possible setups, IBM Tivoli Storage Manager is often the repository of choice due to its high performance and low cost characteristics, combined with its ability to interact transparently with secure storage. For the data archiving model that is based on SAP ILM, there is only one IBM content repository that is supported when archiving through IBM Content Collector for SAP Applications: IBM Tivoli Storage Manager.
SAP ArchiveLink-based data archiving

Data archiving based on SAP ArchiveLink makes use of the existing infrastructure for document archiving, and employs a customized proprietary file format for compressed storage of bundles of archived data. Data that is archived from the live SAP system is combined, bundled, and compressed into a proprietary file format: ADK or Reconciliation Object (REO). The two terms are used in a synonymous manner. The customization of the data archiving operation must take into account the following conditions:

- The availability of SAP ArchiveLink support in the business module for which the data needs to be archived. Most SAP business modules support ArchiveLink or the Content Server interface in one form or another, and are therefore suited for this operation.
- The combination of database content into archivable objects using detailed table analysis. Be sure to account for the interdependencies of SAP business objects, and how they can be combined into archivable objects that constitute a closed business transaction that can be removed from the live system.
- The applicability of retention regulations for the identified data. Even though SAP ArchiveLink has no explicit retention handling capabilities, the rules governing retention and disposal can also be applied to the stored archive files. Organizing and aggregating the data into individual files that have common properties regarding their retention policies simplifies this process.
- The scheduling of archive operations, either in manual or in automatic mode. As part of this decision point, disposal rules on an archive file basis can also be taken into account.

Depending on the business module, and its particular implementation of the archiving of its business objects, several modes of access to archived data are available.

Depending on the SAP application, archived data can either be viewed transparently from within the business transaction, or it can be interacted with separately, for example, in a document relationship browser. Typically this access to archived data does not imply that the data is restored from the archive back into the live system (even though this option also exists, but is only used in exceptional cases). The archive data is retrieved into the application interface for viewing purposes only.

Archiving objects: Logical units of business process elements

Data archiving enables you to move data (in compressed and standardized format) to either the local file system directly connected to the SAP system, or, through SAP ArchiveLink, to all connected content repositories.

Storing the archived data outside of the active SAP system has several advantages:

- Keep the active system lean.
- Provide advanced protection against data tampering if, for example, certified storage systems are used.
- Improved compliance capabilities.
- Choosing an archive location physically separated from the location of the active system adds physical security and a higher level of protection for the archived data.

Associated with the archiving objects are archiving rules that determine when, and at which intervals, archiving objects will be moved from the live system to the archive. These rules can be based on temporal constraints, such as the posting date for a booking, and can be augmented with additional metadata, such as customer names or product group.
Leading practices for establishing data archiving

The need for data archiving is often ignored in the initial setup of SAP systems, and only becomes apparent when the database encounters the first wave of performance effects caused by database size. A standard practice should be to establish a data archiving process early on, rather than introducing the archiving process as a last resort before a total system performance breakdown.

This is an important point, because the setup of a well-structured data archiving system requires careful planning across multiple tiers of the organization. There must be an active interaction between the organization responsible for the operation of the SAP system (and therefore for its performance) and the SAP system business users and user groups. Also, stakeholders who are not directly related to the SAP system, but are responsible for setting governance rules and other compliance standards, must be included in the design process.

During this interaction between the interested parties, important system facts must be established:

- Expected data growth
- Detailed structure of authorizations
- Legal rules for data protection and retention times, and standard procedures to establish the scope of legal holds
- Business process interdependencies

Figure 19 shows a simplified decision tree that should be applied to the data in the live system, where a triage-like sequence of decisions must be made based on data properties such as continued usage, the ability to summarize, and ability to remove without disruption of the live system.

A simplified decision tree that should be applied to the data in the live system is shown in Figure 19.
Data archiving using SAP ILM

Compared to SAP ArchiveLink-based data archiving, the adoption of SAP ILM has significantly less traction in the market. This fact is mostly attributed to the complexity of the setup, and to the extra cost for the required infrastructure of such a solution.

The following list includes the main use cases for data archiving using SAP ILM:

- Decommissioning of existing systems, where data from these systems still needs to be available for compliance reasons
- Business scenarios with strict retention rules or high potential of legal holds on structured data, where the retention rules can be tied to organizational structures
- Highly structured rules for identifiable subunits of the organization that map naturally into the hierarchical organization of the WebDAV-based data archive

Preparation

The implementation of ILM-based data archiving requires detailed preparation steps on the SAP side regarding the modeling of the affected data. In broader terms, this pertains to establishing a data model and retention plan alongside the business structure of the organization, reflecting the structure of rules that govern the following business aspects:

- The data that has to be retained
- The retention periods in all their variations
- Prescribed destruction times, for example, for certain types of HR data
- Definition of structural models that describe the extent of potential legal holds on structured data, based on organizational and business process modelling

The hierarchical model derived from all of these considerations defines collections of data with common properties (a property index). These properties can be used to identify groups of resources that are then assigned common retention properties.

The model provides a hierarchical, unambiguous representation of archived data, with standardized access methods using unique Uniform Resource Identifiers (URIs).

SAP ILM-enabled archive back-end

As with SAP ArchiveLink, in the ILM scenario the SAP system is also independent of the service provider for the archiving protocol. IBM provides an ILM-enabled repository that is accessed through IBM Content Collector for SAP Applications.

The intrinsic hierarchical organization of the archived data is mapped transparently into a hierarchical WebDAV archive implemented with IBM Tivoli Storage Manager. As an intermediate, IBM Content Collector for SAP Applications is used to translate the ILM WebDAV model into the IBM Tivoli Storage Manager storage model.

After this modelling is established on the SAP side, the model must be made known to the ILM-enabled archive (in this case the IBM Tivoli Storage Manager archive through the IBM Content Collector for SAP Applications server). It can then be used to carry out all necessary operations of archiving data, setting retention schedules at the required level of detail, applying legal holds for active litigations, and eventually disposing of the data in an auditable manner.
Comparison of SAP ArchiveLink and ILM-based data archiving

The decision about which archiving model to choose in a given situation depends on the business needs:

- ADK-based archiving. The main benefit is its support of database maintenance in a simple and transparent manner. It is the method of choice where none of the other requirements exist.

- ILM-based archiving. This method is the choice if complex retention and legal hold rules must be implemented that are not available, at least not internal to the protocol, for ADK-based archiving. The standardized XML-based storage format enables for more transparent access to the archived data, if such access is required from outside of the SAP system. This capability is further supported through the use of the property indexes that make the archived data transparently searchable using metadata queries.

Adding the value of IBM middleware and storage solutions for SAP data archiving purposes

In summary, both data archiving protocols of SAP support the implementation of external archive providers using standardized interfaces. IBM ECM products provide the necessary flexibility regarding the choice of repository. IBM storage solutions ensure the required security conditions to achieve full compliance with the imposed regulations.

All IBM ECM repositories provide storage hierarchy solutions that will store the data in the most cost-effective way based on usage patterns and accessibility requirements. Full integration into the IBM ECM product portfolio enables the archiving solution to extend beyond the base requirements of the SAP archiving standards into complete end-to-end solutions.

References

These websites are also relevant as further information sources:

- IBM Enterprise Content Management portfolio

- IBM Information Lifecycle Governance solutions

- IBM Value-Based Archiving solutions

- IBM Datacap

- IBM Content Collector for SAP Applications

- IBM Content Collector for SAP Applications Knowledge Center
  [http://www.ibm.com/support/knowledgecenter/SSRW2R_4.0.0](http://www.ibm.com/support/knowledgecenter/SSRW2R_4.0.0)

- IBM Content Collector for SAP Applications: Sizing, Configuration, and High Availability
Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Boeblingen Center.

Stefan Momma is a Software Architect in the Enterprise Content Management division of IBM SWG at the IBM Research and Development Lab in Boeblingen, Germany. He is responsible for the product architecture of IBM ECM for SAP solutions. Before his current assignment, Stefan focused on full-text search in IBM DB2® and IBM Enterprise Content Management portfolio products. He holds a Diploma degree in Computer Science and Computational Linguistics from the University of Stuttgart.

Manfred Oevers is a Development Manager with the IBM ECM division of IBM Software Group, in IBM Germany. Manfred has 14 years of experience in the IT industry. He has worked as a high-performance computing specialist, emerging technology services, life science development, IBM WebSphere® software development, ECM Lab services, and release manager. He holds a degree in Physics from the University of Bielefeld, and a Ph.D. in Theoretical Particle Physics from the same university.

Andrew Stalnecker is a Technical Sales Consultant with the IBM Sales & Distribution division in North America. For the past 14 years, he has helped clients identify requirements, design, and architect solutions based on the IBM ECM portfolio. He is known as the “go-to” person for SAP ECM integration.

Jörg Stolzenberg is the Product Manager for IBM Content Collector and for ECM SAP integrations. After spending 12 years in the ECM industry, mostly in technical positions, Jörg joined IBM FileNet in 2003 as a Technical Alliance Manager for SAP, based in Walldorf, Germany. From 2005 on, he was also in charge of the Product Management of FileNet’s SAP connectors. Since the acquisition of FileNet by IBM, Jörg is the Product Manager for IBM ECM for SAP product portfolio. From 2009 to 2012 he was also in charge of IBM Content Classification. Since 2012, Jörg’s responsibilities also include IBM Content Collector. Jörg holds a Doctoral degree in Mechanical Engineering.

The project that produced this publication was managed by Marcela Adan, IBM Redbooks® Project Leader with ITSO, Global Content Services.

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at: ibm.com/redbooks/residencies.html
Stay connected to IBM Redbooks

- Find us on Facebook:
  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:
  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:
  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
- Stay current on recent Redbooks publications with RSS Feeds:
  http://www.redbooks.ibm.com/rss.html
Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user’s responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.
This document REDP-5268-00 was created or updated on September 15, 2015.

Send us your comments in one of the following ways:

- Use the online Contact us review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an email to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

- Daeja™
- DB2®
- FileNet®
- IBM®
- Redbooks®
- Redbooks (logo) ®
- Tivoli®
- WebSphere®

The following terms are trademarks of other companies:

- Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.