November 29, 2012

IBM Content Collector for SAP Applications – Sizing, Configuration, and High Availability –

White Paper

Version 1.0

Before reading this paper check for the latest version at
http://www.ibm.com/support/docview.wss?uid=swg27036773

Raiko Nitzsche
Software Engineer

Martin Russold
Quality Assurance Lead

IBM ECM Development
IBM Germany Research and Development Lab

Abstract

IBM Content Collector for SAP provides advanced archiving capabilities that help SAP users run their businesses more efficiently. It provides access to business information of all types, supports the transfer of older information to lower-cost disks and tapes, and improves system performance.

This white paper gives information about advanced topics as sizing and high availability setup of IBM Content Collector for SAP Applications.

Contents

1 Introduction

Before reading this white paper, get familiar with the product IBM Content Collector for SAP Applications. Use the following web page as a starting point: http://www.ibm.com/support/docview.wss?uid=swg27036331

This web page gives you access to the product documentation and all other important information that is available for IBM Content Collector for SAP Applications. We strongly encourage you to use the provided information to get a good understanding of the product.

The goal of this white paper is to give you some guidance regarding sizing, configuration, high availability of Content Collector for SAP, and the usage of a load balancer in the context of Content Collector for SAP. As the white paper will be updated from time to time, check the following web page for the latest version of this white paper:
http://www.ibm.com/support/docview.wss?uid=swg27036773

The white paper is organized as follows:

Chapter 2 describes the environment where Content Collector for SAP runs. In addition, it covers all possible physical setups and explains the common SAP archiving scenarios. Read this chapter before you continue with any other chapter.

Chapter 3 introduces the standard configurations for different performance requirements. To choose the right configuration for you it is necessary to determine the performance requirements first. A list of items guides you through this activity.

Chapter 4 describes how to configure a Collector Server instance based on the hardware configuration that you selected in Chapter 3.

Chapter 5 shows how to set up high availability and how to use a load balancer.

2 Content Collector for SAP environment

This chapter provides a short overview of the environment where Content Collector for SAP runs so that you understand the role of Content Collector for SAP in an SAP environment. The paper concentrates on sizing information and the high-availability setup of Content Collector for SAP. However, it is necessary to understand the complete SAP archiving environment before you can size it and set up high availability.

SAP provides an interface to external storage systems for archiving documents or data. This interface is called *SAP ArchiveLink*. Each storage system that wants to connect to an SAP system must support SAP ArchiveLink. The external storage system can be a simple file-based archiving tool or a sophisticated enterprise content management system. Because most storage systems have their own API - mainly because they provide more functionality than SAP ArchiveLink - a translator is needed between SAP ArchiveLink and the API of the storage system. Content Collector for SAP has been designed to translate the protocols of SAP ArchiveLink and several IBM storage and content management systems, as shown in figure 1.

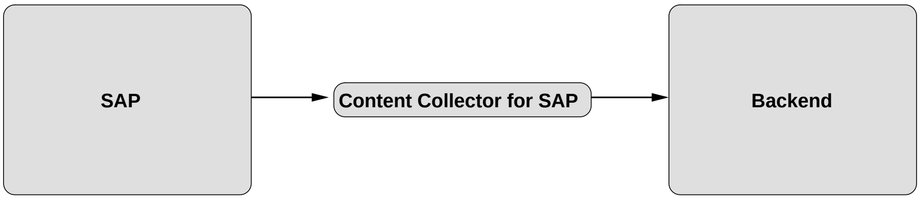

Figure 1: Content Collector for SAP integration in SAP archiving environment

In such an environment, the overall performance is determined by SAP and the storage or content management system. Therefore, it is not sufficient to size and tune Content Collector for SAP only. Of course, you must ensure that Content Collector for SAP does not become the bottleneck. It must be able to handle the data throughput between SAP and the storage or content management system. But, although it is always good to have some reserves, an oversized environment is expensive and does not improve the performance.

2.1 Physical setups

You can deploy Content Collector for SAP in three different ways:

- On the SAP servers (figure 2)
- On a separate physical server (figure 3)
- On the enterprise content management system (figure 4)

Technically, there is no preference for any of the three deployment methods. Nevertheless, only the methods that are shown in figure 3 and figure 4 are normally used. By installing Content Collector for SAP on its own server, you have full control over the sytem: from choosing the preferred operating system to fine-tuning the server for maximum document throughput without affecting the storage or enterprise content management system. On the other hand, by installing Content Collector for SAP on the enterprise content management system, you do not need an additional server that requires maintenance. In addition, the network connection between Content Collector for SAP and the storage or enterprise content management system is only virtual and hence faster and with lower latency.

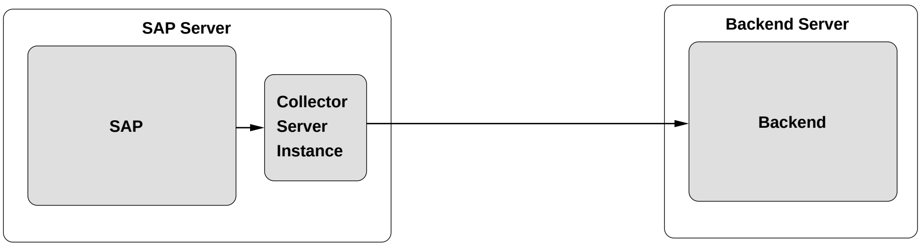

Figure 2: Collector Server on the SAP server

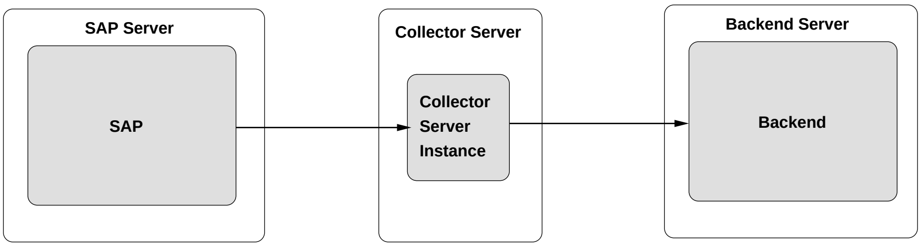

Figure 3: Collector Server on a separate server

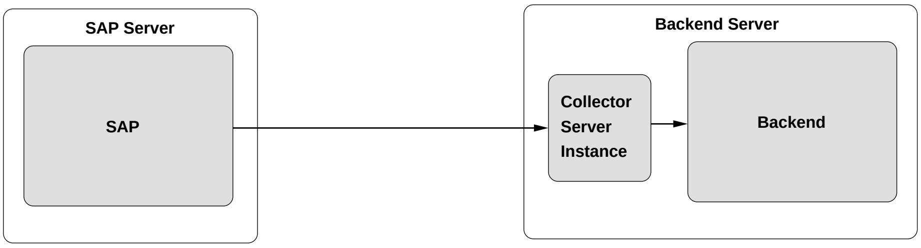

Figure 4: Collector Server on the enterprise content management system

Which installation scenario to choose depends on the enterprise content management system used. If only a single system is used, you can install Collector Server as a protocol translator directly on the server. In a more complex enterprise content management system environment, for example, where multiple enterprise content management systems use the same Collector Server, the installation on a separate server is preferable.

2.2 SAP archiving scenarios

In a typical SAP archiving environment, the following types of documents and data are archived.

- Outgoing documents
- Print lists
- Incoming documents
- Data

Outgoing documents are documents that are generated by the SAP applications, such as invoices or reports. SAP archives such documents in the external storage system by using ArchiveLink. The document type is normally PDF and the document size typically ranges from 10 to 100 KB, although, in some business areas, sizes of several megabytes are common.

The amount of outgoing documents mainly depends on the size and the use of the SAP system itself, for example, what kind of tasks in the company are executed inside SAP and how many users work with the system. The number of generated documents per day and at peak times is a key input for the sizing of Content Collector for SAP and mainly determines the required throughput of the system.

Print lists are documents that are generated by SAP and are formatted for printing. Print lists usually have a size of up to 500 KB. The frequency for generating such print lists is normally much lower than for the outgoing documents. When you size your system, add the print lists to the total number of documents generated per day.

Incoming documents are documents that are generated outside of the SAP system and are available in electronic form, for example, by scanning documents. Those documents are typically archived before they are linked to an SAP business object. The document type is typically TIFF, but newer documents are also available as PDFs. Because they are mostly scanned documents, their size is rather large depending on the number of pages in a document.

Incoming documents are already stored in the archive system and only a short identifier is transferred to the SAP system. Thus, this scenario does not require large system resources - at least compared to the outgoing documents scenario. The actual document only needs to be transferred from the archive system to SAP when the user accesses it with the attachment link in the business object. This is a manual interaction compared to the automatic document generation and the archiving of the outgoing documents. Thus, the required bandwidth and performance depends on the number of concurrent users in the SAP system that are actively looking at archived documents.

If outgoing and incoming documents are used with the same Collector Server instance, you can simplify the sizing by adding all documents that are handled per day and in peak hours. From the perspective of the ArchiveLink servers, all those documents need to be transferred from one side to the other.

Data is extracted by SAP from its internal databases to a special data file. SAP then archives the file in the archive system. This is an administrative task that normally happens rarely and at scheduled times. The generated data files typically reach sizes of several hundred megabytes, but the time window allowed to transfer the files to and back from the archiving system is larger than for the other documents, since this is not an adhoc interactive usage scenario. Normally, data archiving is automatically covered by the sizing process. But for the capacity sizing of the content management system, this data must be taken into consideration.

3 Hardware configurations

This chapter describes three standard hardware configurations for different performance requirements.

Each configuration is based on the throughput of documents of the following sizes:

- Small documents with an average size of 20 KB
- Medium documents with an average size of 500 KB
- Large documents with an average size of 5 MB

Depending on the gathered input data for the sizing, these standard configurations can be used as a starting point for further adjustments to the actual usage scenario.

3.1 Gathering input for system sizing

Before starting with the sizing of the Content Collector for SAP system, the following information needs to be collected.

- The dominating document size, that is the average size of the such documents that are archived or retrieved the most
- The total number of documents archived and retrieved to/from the Collector Server per day (docs/day) and at peak hours (docs/hour)
- How many users are expected?
- What operating system is planed for use?
- How many parallel requests are expected?
- What is the expected workload characteristics?

3.2 Small environment

Document throughput

Document size	Document throughput docs/h
20 KB	180.000
500 KB	45.000
5 MB	4.500

Hardware requirements

- Dual-core CPU
- 2 GB RAM

- One hard disk (min. 7200RPM and min. 100 IOPS) or
  10 MB RAM file system (2 x size of largest archive file - REO or ALF)
- 1 Gbit network card

When running Content Collector for SAP on a separate server that could be an **IBM System x3250** or **IBM Power 710**, for example. Otherwise, the described hardware needs to be available on the server.

3.3 Medium environment

Document throughput

Document size	Document throughput docs/h
20 KB	750.000
500 KB	200.000
5 MB	20.000

Hardware requirements

- 4 cores
- 4 GB RAM
- 3 hard disks (10K-15K RPM and min. 150 IOPS) as RAID 0 or
  1 SSD with more than 1000 IOPS or
  50 MB RAM file system (2 x size of largest archive file - REO or ALF)
- 1 Gbit network card

When running Content Collector for SAP on a separate server that could be an **IBM System x3250** or **IBM Power 710**, for example. Otherwise, the described hardware needs to be available on the server.

3.4 Large environment

Document throughput

Document size	Document throughput docs/h
20 KB	1.500.000
500 KB	400.000
5 MB	40.000

Hardware requirements

- 8 cores
- 8 GB RAM

- 6 hard disks (15K RPM and min. 150 IOPS) as RAID 0 or
  1 SSD with more than 2000 IOPS
  100 MB RAM file system (2 x size of largest archive file - REO or ALF)
- 2 Gbit network cards (one for incoming connections from SAP and one for outgoing connections to storage or content system)

When running Content Collector for SAP on a separate server that could be an **IBM System x3550** or **IBM Power 710**, for example. Otherwise, the described hardware needs to be available on the server.

4 Collector Server configuration

4.1 Collector Server instance

Content Collector for SAP supports the concept of instances. An instance is the smallest work unit of Content Collector for SAP. The number of instances is not limited by Content Collector for SAP. You can configure an arbitrary number of instances. However, as server resources are limited, only a certain number of instances are feasible.

Depending on your usage scenario, it might be necessary or desirable to run multiple instances. One Collector Server instance can only connect to one SAP system. However, this connection is only necessary for certain usage scenarios (for example, incoming documents). In this chapter we assume that only one Collector Server instance is used.

A Collector Server instance consists of the following components: dispatcher, engine, and agent (Figure 5).

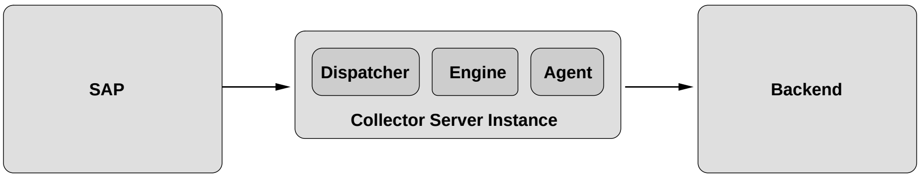

Figure 5: SAP environment with a single Content Collector instance

The dispatcher communicates with the SAP system and the agent communicates with a storage system or an enterprise content management system. The engine interprets and translates the messages that it receives through a dispatcher or an agent, and then forwards the translated messages to an agent or a dispatcher. The engine starts the dispatchers and agents during its startup phase.

The number and the type of dispatchers and agents that are initialized during the startup phase are configured in a server configuration profile, which is by default named `archint.ini`. Each Collector Server instance has its own server configuration profile.

Different types of dispatchers and agents exists.

A Collector Server instance can have one or more HTTP dispatchers, RFC dispatchers and client dispatchers. In addition, it can have one or more TSM agents, OD agents, CM agents, and P8 agents (Figure 6).

It is also possible to set the number of dispatchers or agents to zero. In this case, the dispatcher or agent is not started by the engine.

The type of dispatchers and agents depends on the implemented SAP archiving scenario. The following sections refer to dispatchers and agents in general.

The server configuration profile specifies the port number where the configured dispatchers accepts incoming requests from the SAP system.

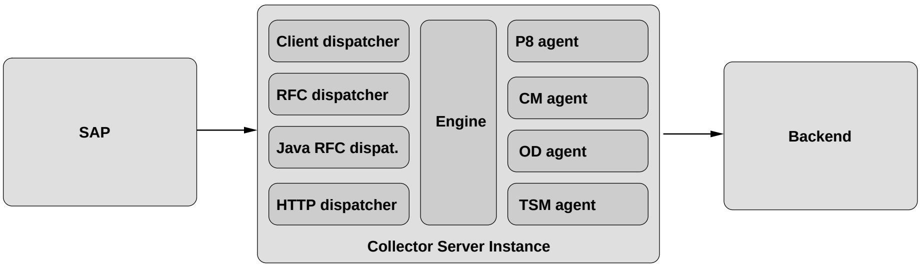

Figure 6: Overview of the Content Collector for SAP integration in an SAP environment

Each server configuration profile contains one or more *archive sections*. The parameters of such an archive section identify a particular storage system or enterprise content management system. The name of the archive section corresponds to the name of a content repository defined in SAP.

For the purpose of this white paper, it is sufficient to remember that the server configuration profile defines the following:

1. The port where the Collector Server instance runs
2. The dispatchers and agents that are started by the engine during its startup phase

All other settings in a server configuration profile have a minor impact on the sizing or the high-availability setup of Content Collector for SAP.

You must configure a Collector Server instance that corresponds to the standard hardware configuration that you identified in Chapter 3. Therefore, the following sections will present a small, medium, and large instance configuration.

4.2 Small configuration

A small configuration consists of two dispatchers and two agents. Use this configuration if you have a small environment (see Chapter 3.2). The following shows a sample minimum configuration for a Content Manager backend system:

```
WEBDPS          2
DISPATCHERS     0
ARCHWINS        0

ADSMAGENTS      0
ODAGENTS        0
CMAGENTS        2
P8AGENTS        0
```

4.3 Medium configuration

A medium configuration consists of four dispatchers and four agents. Use this configuration if you have a medium environment (see Chapter 3.3). The following shows a sample standard configuration for a P8 backend system:

```
WEBDPS          4
DISPATCHERS     0
ARCHWINS        0

ADSMAGENTS      0
ODAGENTS        0
CMAGENTS        0
P8AGENTS        4
```

4.4 Large configuration

A large configuration consists of 16 dispatcher dispatchers and 16 agents. Use this configuration if you have a large environment (see Chapter 3.4). The following shows a sample large configuration for a TSM backend system:

```
WEBDPS         16
DISPATCHERS     0
ARCHWINS        0

ADSMAGENTS     16
ODAGENTS        0
CMAGENTS        0
P8AGENTS        0
```

You can configure an arbitrary number of dispatchers and agents for an instance. Remember, however, that the other systems involved (for example, SAP system, enterprise content management system) then also have to be able to process this amount of data.

A large configuration is usually sufficient to fulfill most requirements. Nevertheless, if you have identified that your requirements exceed a large configuration, please feel free to ask your Content Collector for SAP contact for help.

5 High-availability scenarios

Before setting up Content Collector for SAP for high availability, you must be familiar with the Content Collector for SAP environment (Chapter 2). You also have to understand the concept of a Collector Server instance (Chapter 4). It is assumed that the SAP Content Server HTTP Interface is used for the communication between the SAP system and Content Collector for SAP. This interface is based on HTTP and therefore stateless. This means that there is no relationship between, or dependency on, any requests.

There are many possibilities and products that can be used to set up high availability. Therefore the following sections focus on the specifics of Content Collector for SAP.

5.1 High-availability capabilities and limitations

Content Collector for SAP does not offer built-in high availability. The product has only few dependencies and can, therefore, work self-contained. You can integrate the product into any high-availability solution, which is transparent to the applications running on each server.

Chapter 2.2 introduced the following SAP archiving scenarios: outgoing documents, print lists, incoming documents, and data. Depending on the scenarios used, different type of high-availability solutions (for example a load balancer) can be used.

The SAP archiving scenarios can be divided into the following groups:

1. Scenarios driven by SAP system
2. Scenarios driven by Content Collector for SAP

Scenarios driven by SAP system

Outgoing documents, print lists, and data archiving belong to the scenarios that are driven by the SAP system. Content Collector for SAP does not ensure the integrity of the data provided by SAP. The SAP system is always the master system that controls the data flow and maintains data integrity. Due to the protocol that is used by SAP, the SAP system knows whether a document or data is correctly stored in the archive system. If an error occurs during the archiving process, SAP receives an error code or a timeout. SAP then considers the action failed and the data to be still inside of SAP. It will then retry to store the document or data later.

The following paragraphs describe, from a high-level point of view, which steps are necessary to set up high availability for a Collector Server instance.

Step 1: Create two Collector Server instances, each on a different servers. Both instances must use the same server configuration profile (archint.ini).

Step 2: Determine whether a Collector Server instance is operational or not. Content Collector for SAP offers a monitoring tool called *archcheck* for this purpose. With this tool, you can query the operational status of a Collector Server instance. It checks whether requests can be routed through the dispatcher, the engine, the agent and the storage system or enterprise content management system. Depending on the return code you can identify if a Collector Server instances is operational. A detailed description of the *archcheck* tool can be found in the product documentation.[1]

Step 3: Set up a virtual IP address. The setup of a virtual IP address depends on the high-availability solution. In addition, change the configuration of the SAP system so that SAP sends its request to the configured virtual IP address. The high-availability solution used forwards the request from the virtual IP address to one of the Collector Server instances.

Thus, the scenarios that fall into this category can be set up for high availability through a HTTP/HTTPS load balancer, for example. Depending on your needs, you can set up the system as active-passive or active-active.

Scenarios driven by Content Collector for SAP

Archiving incoming documents is a scenario that is driven by Content Collector for SAP. In such a scenario, the communication path is reversed and the high-availability model changes. Because the communication is not based on HTTP/HTTPS, high-availability solutions based on this protocol cannot be used. Low level (for instance a TCP load balancer) high-availability solutions, however, can be used.

[1] The monitoring tool *archcheck* is only available from IBM Content Collector for SAP Applications Version 3.0. If you use IBM Content Collector for SAP Applications Version 2.2, the administration tool *archadmin* can be used instead. However, this tool has limited capabilities. Compared to *archcheck* it can only verify whether the engine is operational.

6 For more information

Use the following web pages to get the latest version of this white paper and to get more information about IBM Content Collector for SAP Applications:

Sizing, configuration, and high availability for Content Collector for SAP
http://www.ibm.com/support/docview.wss?uid=swg27036773

Content Collector for SAP Applications Version 3.0 publication library
http://www.ibm.com/support/docview.wss?uid=swg27036331

Content Collector for SAP Applications Version 2.2 publication library
http://www.ibm.com/support/docview.wss?uid=swg27022108

IBM Deutschland Research & Development GmbH
ATTN: Dept. 5248
Schoenaicher Strasse 220
71032 Boeblingen
Germany

Produced in Germany
November 2012

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. Any statements regarding IBMs future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED AS IS WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements (e.g. IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided.