Connector Framework

Making Microsoft SharePoint® and FAST® Search Universal and Actionable

"IDC has been surveying information workers worldwide since 2001 to find out how much time they spend on a variety of information tasks. Our studies show that searching and analyzing information ranked as one of top three on the list, making these tasks relatively straightforward candidates for better automation. BA Insight 's technology is designed to remove the inefficiency out of these tasks."

Sue Feldman, Vice President Search and Discovery Technologies, IDC

BA Insight
January, 2014

Contents

Introduction

The Connector Framework is a SharePoint Search connector framework (a.k.a. Protocol Handler framework) that provides the means to connect and index data from an external line of business systems into Microsoft Office SharePoint, and Microsoft Search Server. This ability increases the scope of enterprise search to include those external systems.

Connectors securely integrate over 30 enterprise systems into Microsoft SharePoint and FAST Search, providing knowledge workers with a single point of access to all information, people, and expertise across the enterprise.

With rapid and cost-effective out-of-the-box deployment, comprehensive security-mapping, and full Active Directory integration, Connectors maximize the Return on Investment (ROI) of an organization's ERP, CRM, ECM, and messaging systems by securely unlocking and surfacing this information in a unified view.

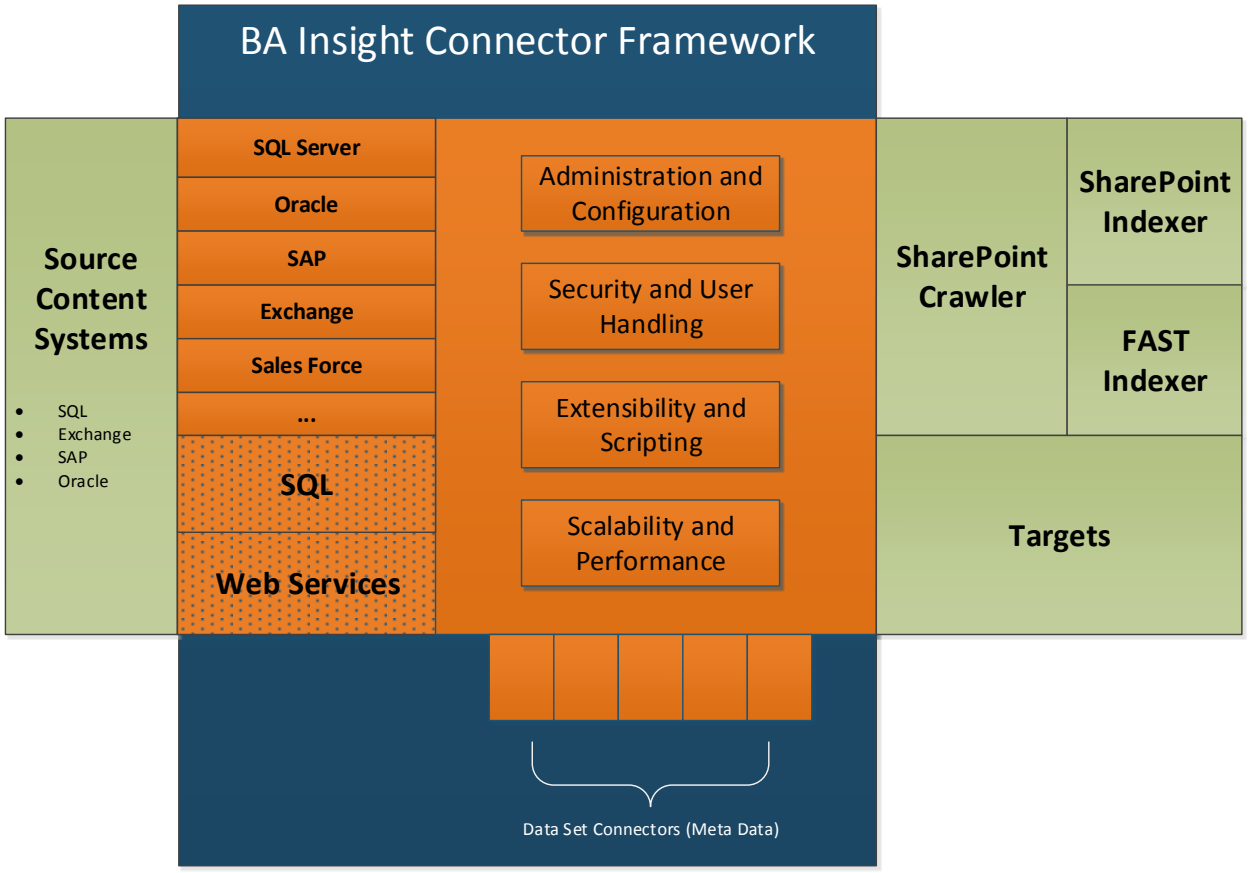
The Connector Framework is built based on the following principles:

- Robust
  Empower FAST or SharePoint Search to surface any structured or unstructured data from virtually any content repository.
- Flexible
  Index only the back-end system content you desire, and present it to end-users in the manner they choose.
- Secure
  Swiftly map and strictly honor any security model
- Scalable
  Optimized for multi-threaded, fully-distributed performance

This document provides a high level overview of the Architecture and Implementation of the Connector Framework. It highlights unique features that are part of the Connector Framework which sets it apart from any other Enterprise Search Solution for Microsoft Office SharePoint.

Connector Framework

Connector Framework Architecture



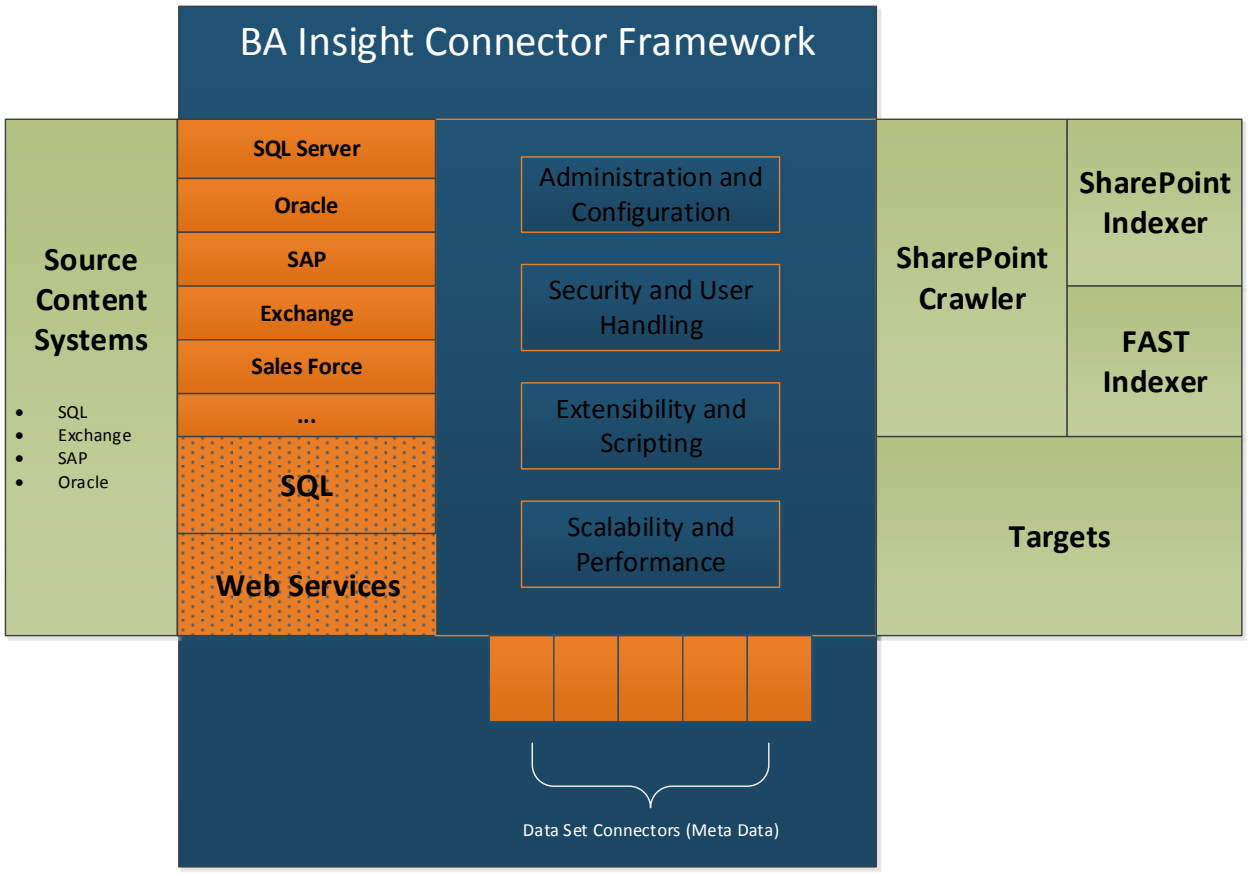
Technical Architecture

The Connector Framework is built using Microsoft .NET with no reliance on custom Windows Application DLL or Windows Services. The main components are:

- **Administration and Configuration**
  - Admin Pages – Manages the configuration settings of the connection string, as well as the creation and crawling schedule of content source.
  - Content Definition lists – Manages the definition of content (a.k.a. items) crawled and indexed.
  - List Events – Handles various background processing tasks such as temp cache cleanup.

- **Security and User Handling**
  - User Group Synch – Synchronizes the list of system users with the list of Active Directory Users on its own schedule, offline from the main content crawling and indexing, in order to not slow down indexing.
  - Security Trimmer – Implements MS security trimming interface.
- **Connectors**
  - Content Connectors – Connect virtually any source content system through codeless configuration capabilities. Content sources are either connected through standard SQL and/or web services connectors or through custom developed built-in connectors.
  - Dataset Connectors – Connect to additional content sources to enrich search results by adding relevant metadata.
- **Extensibility and Scripting**
  - SQL and Web Service integration allows to connect virtually any source content system
  - Script complex content and metadata gathering through the administration interfaces.
- **Scalability and Performance**
  - Multi-Threading capabilities allowing users to scale the performance as needed based on the environment.
  - Pass-Through of content with minimized performance impact.
- **SharePoint Connection**
  - Protocol Handler – Implements MS indexing interface. Connect to SharePoint or FAST indexer.
  - Targets – The Connector Framework supports target systems which get the content pushed to them instead of pulling their content data from the Connector Framework. Targets include FAST, SharePoint, and custom targets. Targets can be implemented to push content into a file or database for example.

Connectors



Content Connectors

The Connector Framework takes advantage to both ADO.NET and Web Services to connect to virtually any databases (Oracle, MS SQL Server, IBM DB2, MySQL, OLE DB) and packaged or custom applications with a Web Services interface.

Configuration templates that map the Connector Framework to the following Enterprise Content Management and Enterprise Resource Planning applications which are available are:

- Aderant
- Microsoft Dynamics
- Oracle E-Business Suite
- Peoplesoft / Siebel
- Salesforce.com / Force.com
- SAP Business Suite
- LexisNexis Interaction CRM
- IBM FileNet / Content Manager / DB2
- IBM Lotus Notes Application Databases
- Autonomy Worksite
- OpenText Hummingbird / LiveLink
- Oracle CMS / Stellent

- OpenText LegalKEY
- Oracle RDBMS / SQL Server / DB2
- Real Practice
- Web Services
- West KM
- EMC Documentum
- Alfresco
- Xerox DocuShare
- HP Trim Context (TowerSoft)
- EMC eRoom
- Microsoft Exchange Private Mailboxes
- Symantec KVS E-Vault
- IBM Lotus Notes Mailboxes

The Connector Framework leverages standard SQL and web services connector templates to rapidly integrate with any content source that provides the ability to feed content through any or both of those technologies. Many content connectors combine SQL and web service capabilities to provide content to the Connector Framework.

Dataset Connectors – Metadata

Dataset connectors enable dynamic enrichment of retrieved content information with specific metadata from other content systems that contain related information. This provides a richer search result to the end user.

Dataset connectors can either be:

- SQL based
- Web Service based
- SharePoint based
- Force.com (SalesForce) based

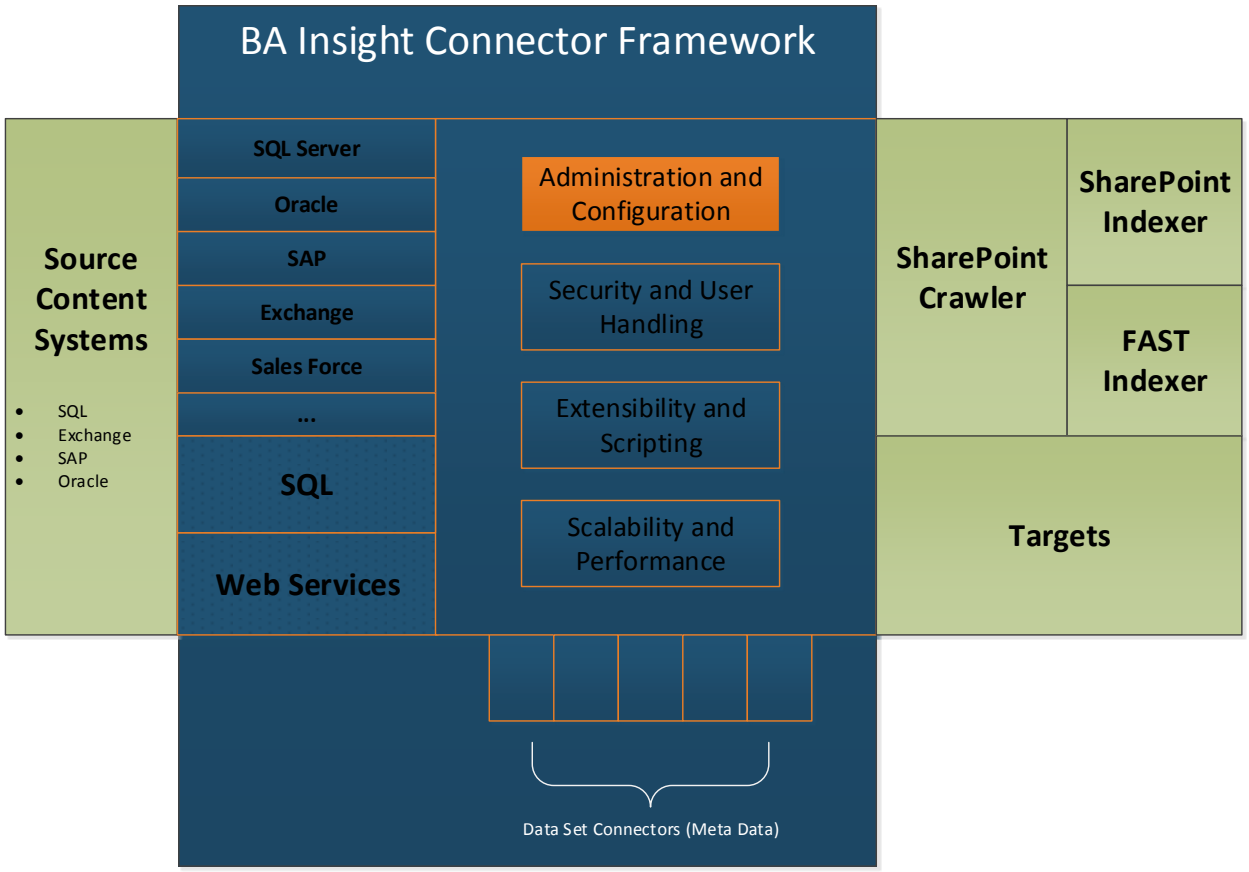
The built-in scripting interface and configuration system allows to tailor the metadata collection to the needs of any installation and infrastructure.

Metadata can be retrieved from several metadata content systems which allows for example to provide/enrich search results on a customer data record from an ERP system to be enriched with information about that same customer from a CRM system. Such information could for example be to retrieve the market designation from the CRM system while the content is gathered from the ERP system. This would then allow the user performing a search to be able to see customers in a specific market if he or she searches for a market.

The retrieved metadata can be additionally filtered to allow for finest granularity when matching the metadata to the data content.

Text Metadata Sample View:

Administration and Configuration



The initial installation of the Connector Framework is guided by the actual setup wizard, which is fully automated and takes minutes to run. The actual configuration of the connector mapping template, if not already provided by BA-Insight as a pre-configured map, takes anywhere between 1 to 5 days, based on hundreds of successful customer deployments.

There are two types of Connectors:

1. SQL Database Connectors
2. Web Services Connectors

The following two sections describe the specific configuration settings required for each type of connector.

SQL Database Connector Configuration

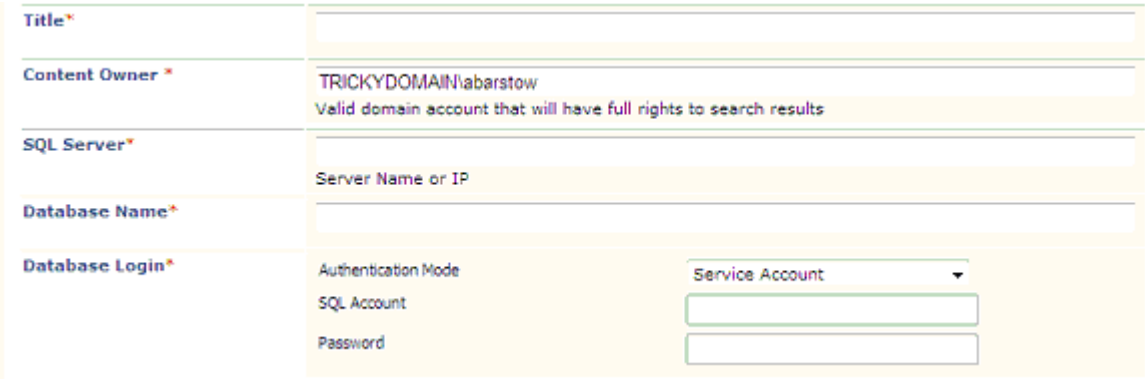
The SQL Database Connector administrative interface exposes three key setting groups of the various system components:

- **Connection Settings** – information needed to connect to the relational database and Active Directory (AD) for security accounts.

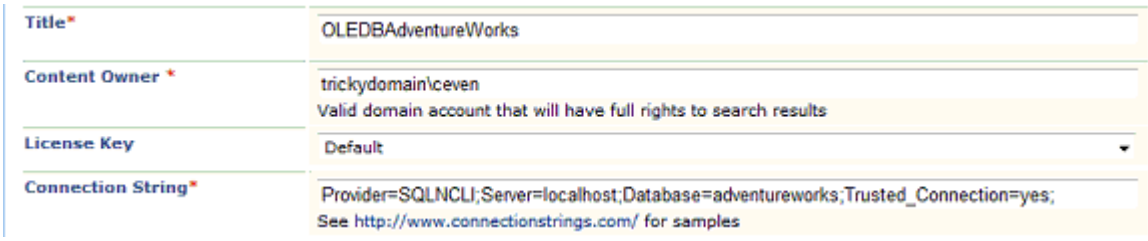
- **User Group Synch Settings** – information needed to schedule the load of the security information from the relational database and map it to AD accounts.
- **Content Settings** – information needed to define the metadata and unstructured content to be crawled, including an optional modified date field for incremental crawls and item security credentials by user or group.

Connection Settings

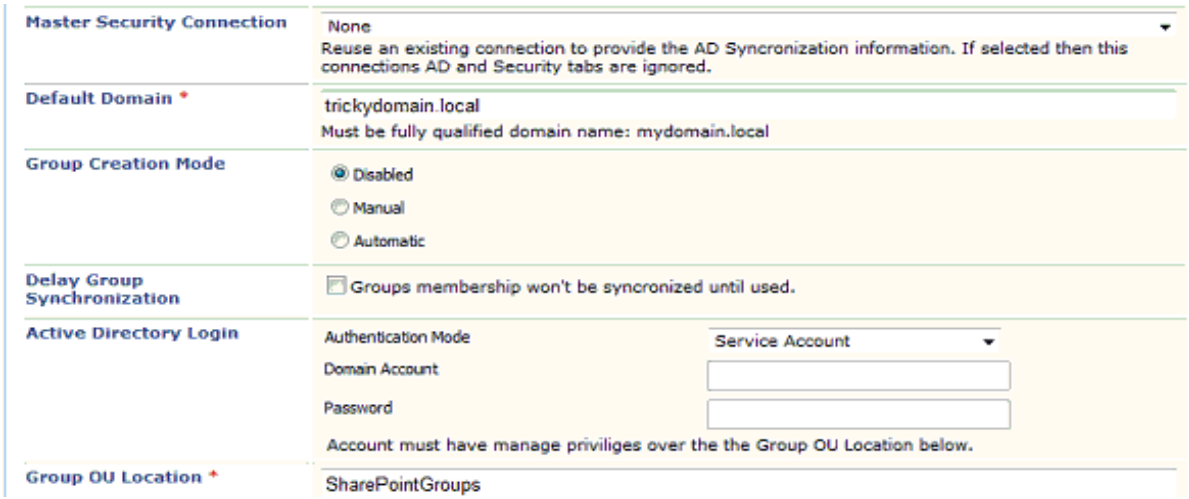
The first step in installing and configuring the SQL database connector is to configure the connection settings. As the screenshot below depicts, the requested information includes the server name and security account authorized to connect to the database.



Alternatively, an OLE DB connection can be defined with the appropriate connection string information.



Active Directory (AD) connection settings must also be provided, as shown in the screenshot below.



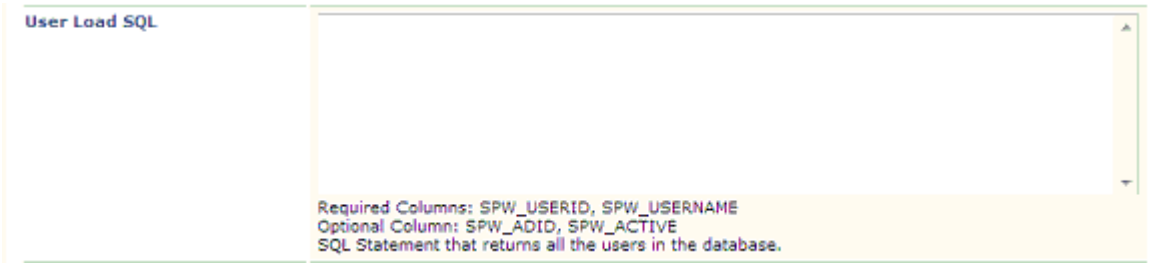
User Group Synch Tool

The second step in installing and configuring the SQL database connector is the configuration of the user group mapping from the SQL database system users to the AD accounts.

The user group synch tool performs the following tasks:

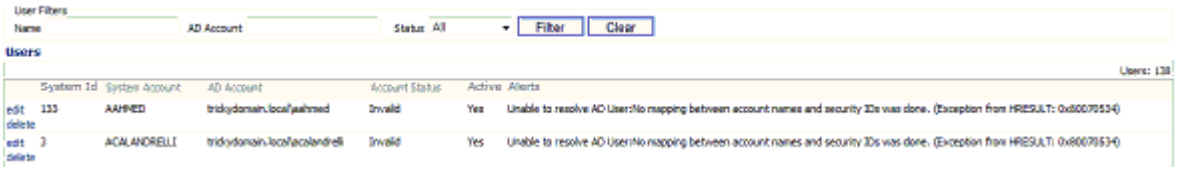
- Sets AD connection settings provided in the section above to create a connection to Active Directory.
- Loads the database system user and group accounts, if any, based on Administrator defined SQL scripts, as depicted below.

User Load SQL Sample View:



- Automatically maps the database system users to AD user accounts based on name matching rules when run. The Administrator has the ability to override the automatic mapping by editing or deleting certain entries.
- Automatically creates AD groups under a distinct organizational unit (to be created by the AD team) for which it has read/write access. It keeps the AD groups synchronized with the database system groups/roles based on a job schedule defined by the Administrator.

The screenshot below depicts the mapping table between a system user account and AD account.

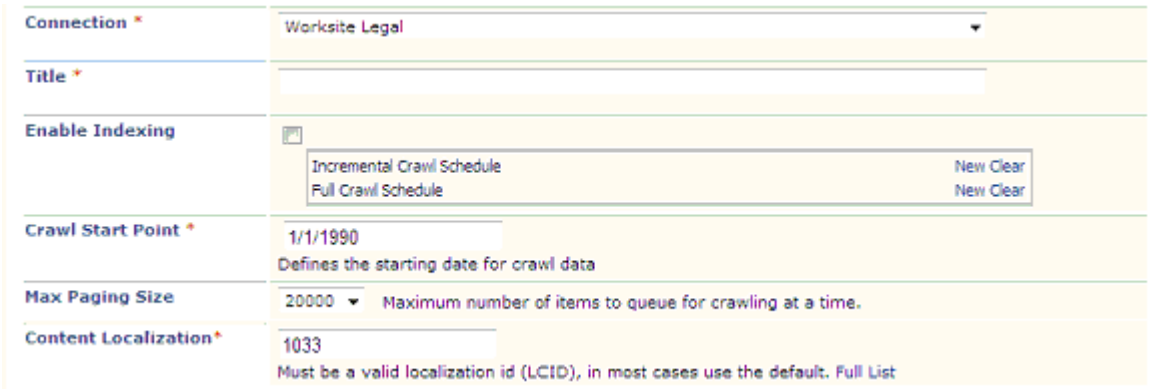


Content Settings

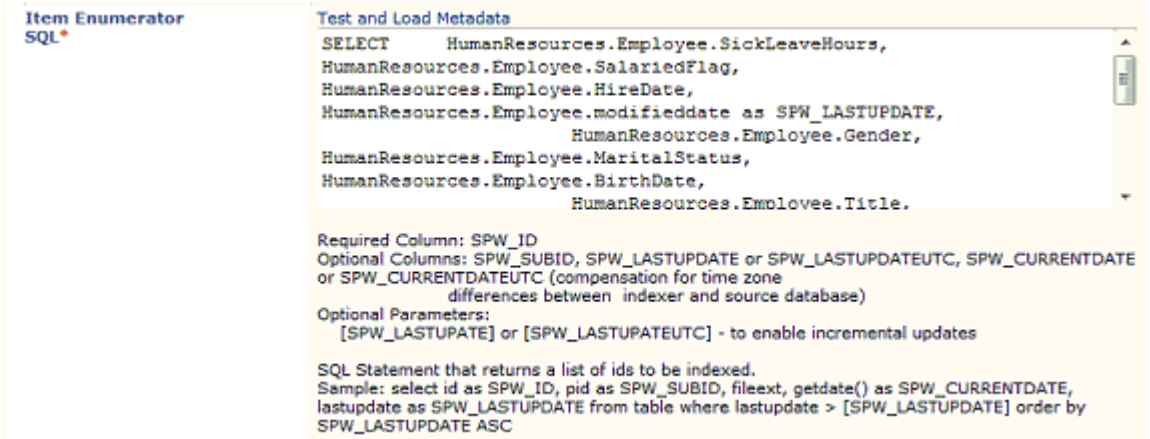
The third and final step in installing and configuring the SQL database connector is the configuration of the content to be crawled.

Content Configuration

Content Source - the Content Source crawl schedule is specified, and automatically published to SharePoint.

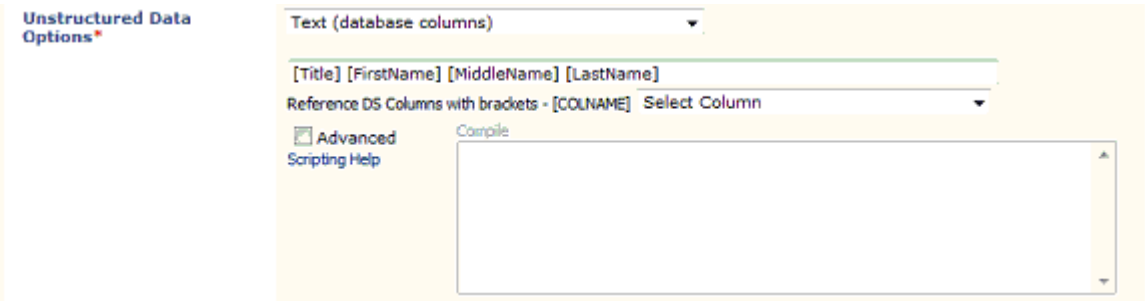


Item Enumerator - The Item Enumerator specifies the record ID and modified date, if any, in order to perform efficient incremental crawls where only the changed items since the last crawl are being indexed.

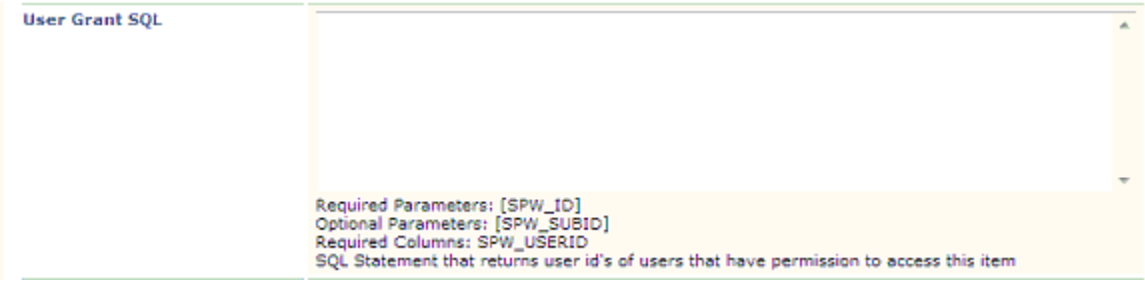


Item Definition - The Item Definition specifies the record detail, including the metadata and related unstructured data, if any, stored in the database or an external file system. The metadata must include

certain default properties such as the URL and Title. Note that the Managed Properties in SharePoint are automatically generated from the Item Definition properties.



Item Security - The Item Security specifies the record security credentials based on the database system's internal user and group ID's, which will be mapped automatically to an AD account and group during indexing.



Note that Real-time security trimming can be enabled as well by simply selecting a checkbox, if there is a requirement to check security credential changes in between crawls. Since the real-time security trimming is done real-time at query time, be aware that it adds overhead to the query response time.

Web Services Connector Configuration

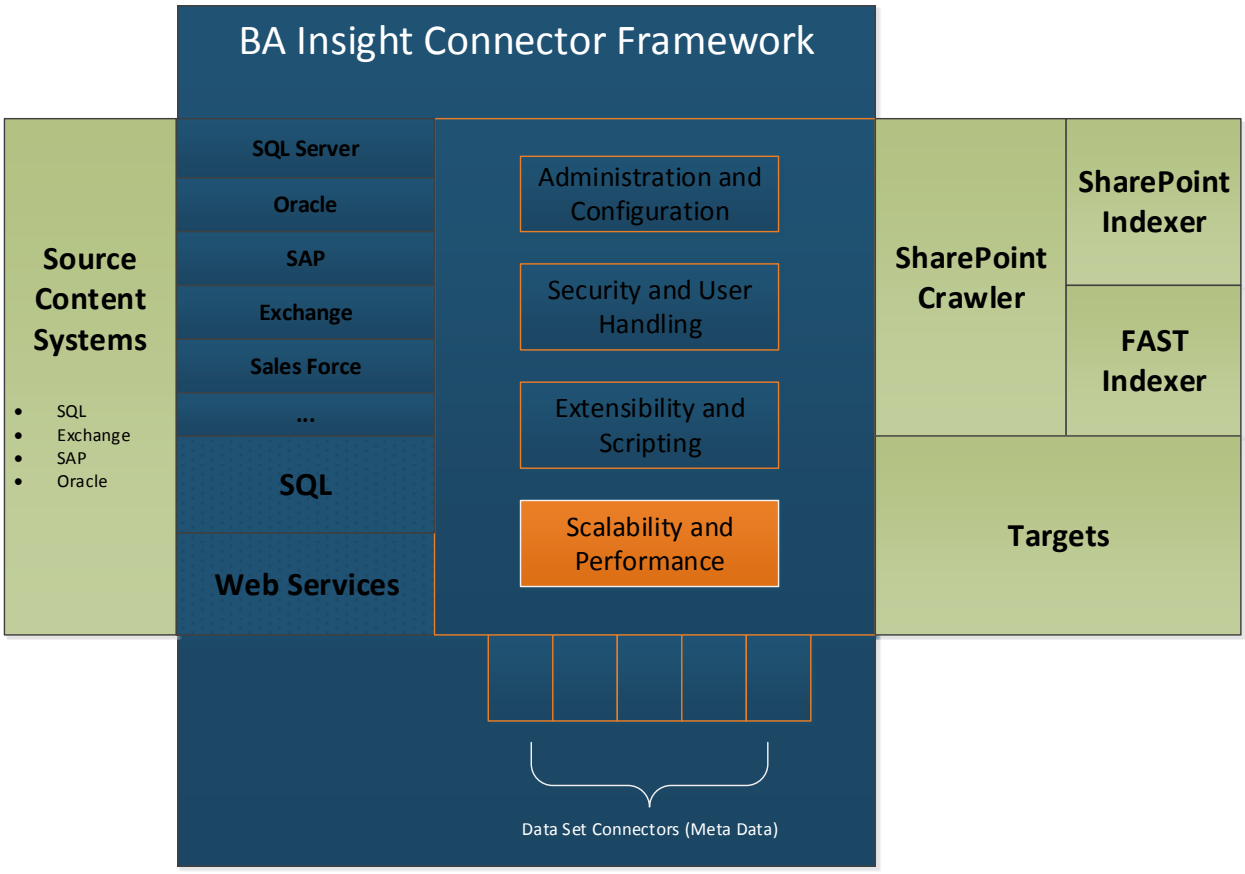
The Web Services Connector requires the implementation in C#.NET of our custom web service interface below in order to take advantage of the Connector Framework.

Most ECM, CRM, and ERP systems, such as Alfresco, Lotus Notes, Salesforce.com, SAP, etc., have defined a Web Service interface that can be reused and wrapped around to map into the required Web Service interface. The methods to be implemented are listed below:

Methods to Implement	Description	Required
DoCrawl()	Returns the list of id's for the items to index.	*

DoItemData()	For each ID this method returns all the metadata and any attached files to be indexed.	*
DoDescribe()	Provides the ability for your web service to tell us which features you implemented and what information you want the end user to provide you with in our admin system. If you require a username, password, or maybe a custom filter passed through to your web service, then that option would be set here.	*
DoGetAvailableDataStores()	Allows your web service to provide a pick list of data sources to the admin.	
DoGetDatastoreTypes()	Allows your web service to provide users with a pick list of the data types to crawl: Client, Project, etc.	
DoGetGroups()	For AD mapping support; returns a list of Groups in data source.	
DoGetGroupsGroups	For AD mapping support; returns a group hierarchy	
DoGetGroupsUsers()	For AD mapping support; returns a member list for group.	
DoGetUsers()	For AD mapping support; returns a list of Users.	
DoRealTimeSecurityCheck()	Allows additional security check.	

Scalability and Performance



Scalability

The Connector Framework has been tested at the Microsoft Technology Center on an eight quad-core CPU indexing server. It can reach indexing speeds ranging from 20 to 500 records per second, based on the size of the record data and the number of file attachments. Incremental crawls reach a speed of 5,000 records per second, as the Connector Framework effectively crawls the subset of records modified. SharePoint out-of-the-box will instead crawl every record from the source system to detect if it has changed, potentially amounting to millions of queries versus just a few queries with our Framework.

Real-world experience has shown that a poorly executed incremental crawl can bring a source system down. In fact, BA-Insight has redeveloped certain out-of-the-box connectors, such as Lotus Notes, as it found that the out-of-the-box version, while free, simply didn't work.

Another significant source of stress and potential performance issues to the source system is the lookup and translation of security accounts into AD accounts for each indexed record. BA-Insight connector architecture successfully eliminates this bottleneck by implementing a user group synch job offline, which performs such user group loading and mapping prior to index time.

Performance

The Connector Architecture allows users to process search data from different source content systems and add metadata information with minimal performance impact. Essentially, the retrieved content is passed through with barely any performance impact.

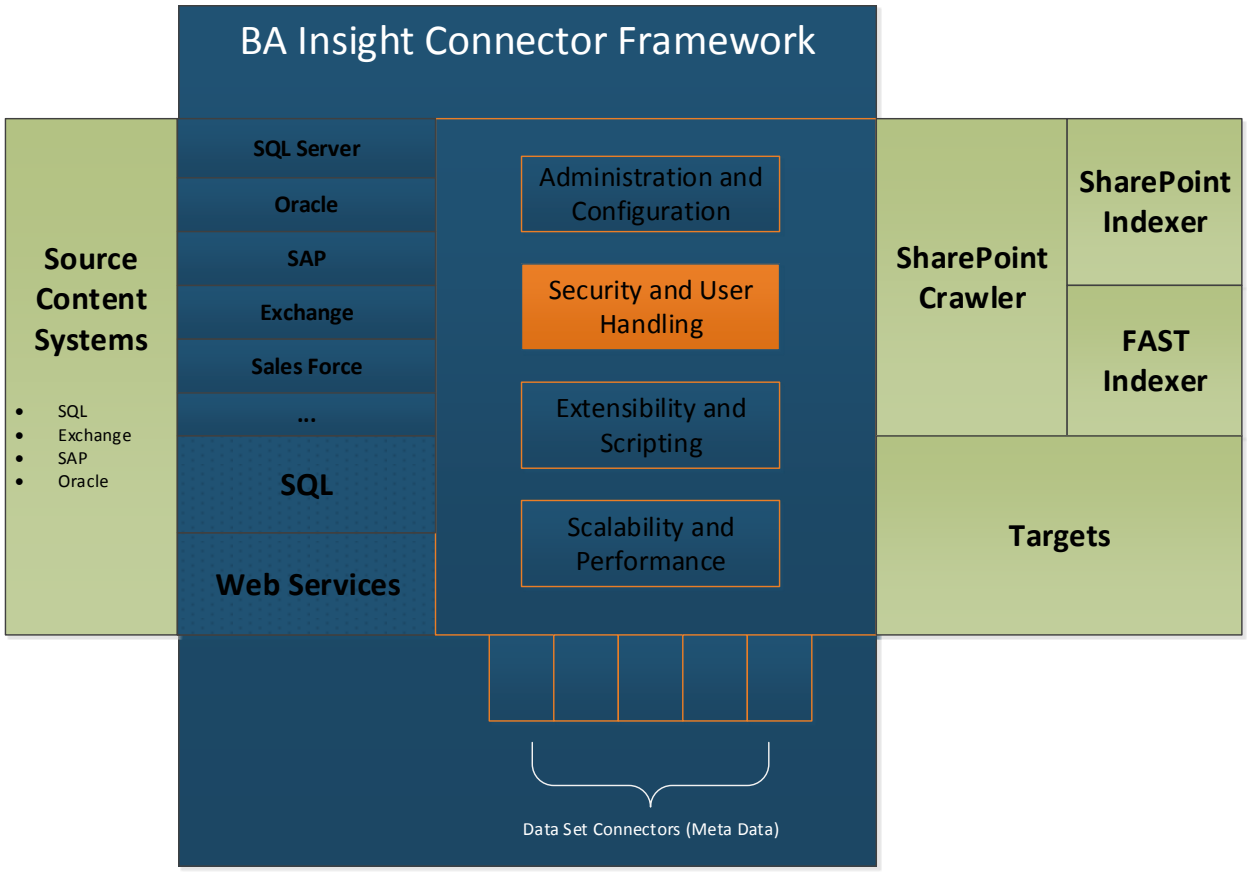
The Connector Framework is implemented to retrieve data multi-threaded. Due to this architecture, the Connector Framework's performance is not the bottleneck seen in many enterprise search setups.

Based on the experience of many implementations the performance question comes down to:
How fast is responsible?

The Connector Framework allows users to retrieve data from many different content source systems which each have their own architectural and performance limitations. Issues have arisen in the past where the connector to the source system was actually trying to retrieve too much information too fast. This data retrieval in turn caused the source content system to perform slower and at unacceptable speeds.

This performance issue can also be seen on the different targets and in SharePoint. If the IT infrastructure for a specific installation cannot support the full speed of the Connector Framework, then the number of threads which perform the data processing will have to be adjusted in order to scale the performance to the existing systems.

Security and User Handling

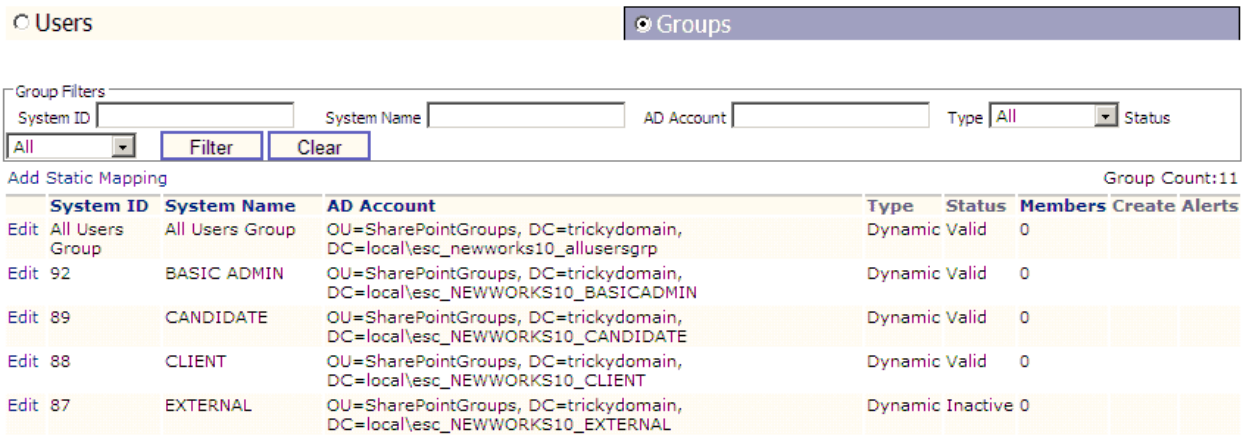


Microsoft Office SharePoint Server Search and Microsoft Search Server rely on two mechanisms for securing search results so that users only see the items they are allowed to see.

AD Synchronization

The first mechanism is based on traditional Windows/Active Directory (AD) Access Control List (ACL) security. When items are indexed, they are associated with a fixed access control list much like every file in your file system. At query time the current users AD credentials are applied to this ACL to determine if access is granted. This security model is highly scalable and typically used, but it has some limitations. For example, it cannot be used to easily secure non-AD based systems, or if the SharePoint authentication method is not AD-based. In these situations, there is no way for your credentials to be applied.

Security Mapping Sample View



The Connector Framework provides a synchronization process that can be scheduled automatically to create Active Directory Global security groups and manages the user membership of those groups to ensure that they match exactly the membership of the source systems. This allows for group membership changes to be reflected in search results between crawls as frequently as the sync process is scheduled. The actual job is single-threaded and does not impact the performance of the installed AD systems.

In order to mitigate access concerns from network administrators, a new Organization Unit (OU folder) is created in AD to manage these groups in. By restricting the AD write access to only this particular OU, and limiting the write privileges to only create and manage groups, no other system can be impacted.

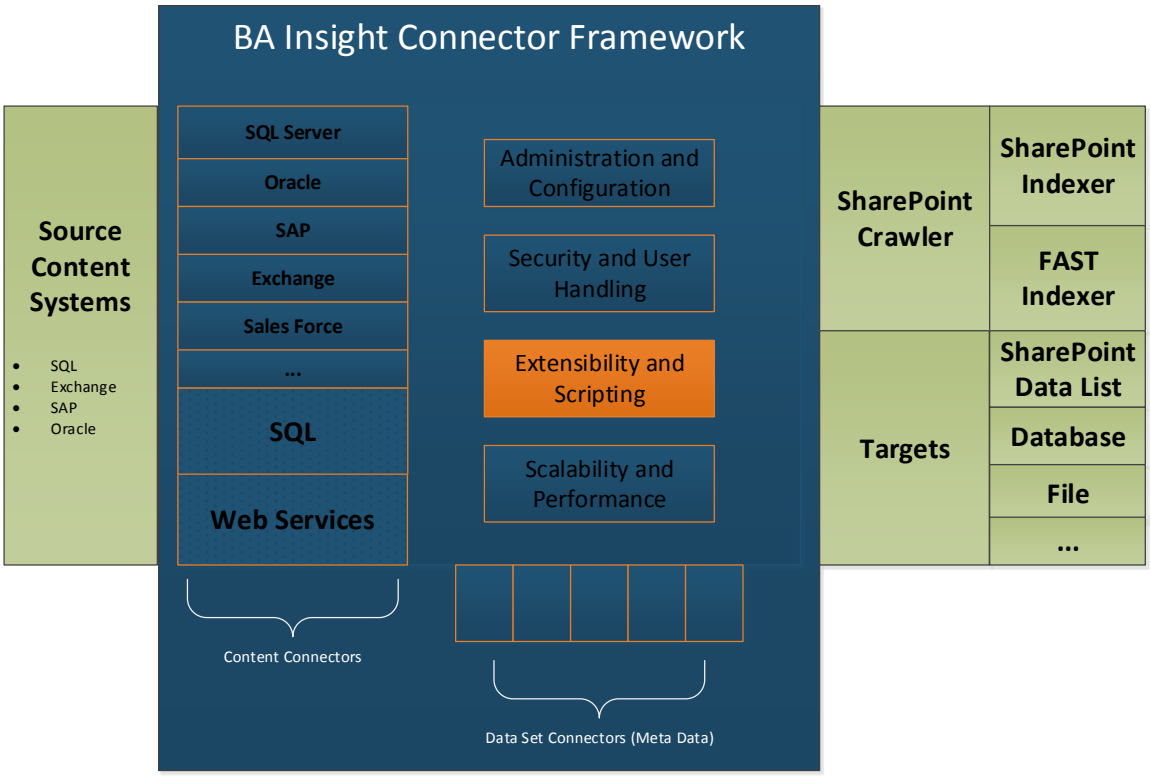
Real-Time Security Trimming

The second mechanism, which can be used in combination with the first, is called Real Time Security Trimming. This method relies on custom callouts to the source system to verify that the results about to be presented to the user are allowed. The benefit of this model is that it is not AD-dependent and can be used to secure most systems. The downside is that it can only be applied to the first few hundred results and if the user does not have permissions to those results, no results are displayed for the user.

The connector framework, including the Universal SQL and Web Services ones, provides a straightforward way to define custom real-time trimmers.

The Connector Framework supports both security models, and for most of the connectors these models are applied in combination to ensure both scalability and real-time security trimming. In order to support AD-based ACL security against systems that have their own security model (e.g., Hummingbird, WorkSite, Documentum, and other document management systems), users and security groups are mapped into a corresponding Active Directory entry that is created automatically by the Connector Framework. For systems that already have some form of AD synchronization, existing groups can be reused.

Extensibility and Scripting



Custom Content Connectors

Custom content connectors allow to integrate any kind of content source to be indexed by the Connector Framework. There are three types of custom content connectors that are available to integrate content source systems:

Custom Database Connectors

Custom Database Connectors allow users to connect and extract data from any database. SQL Server is natively supported, as are any databases that support OLEDB.

Template Based Connectors

Template Based Connectors are essentially pre-configured Custom Database Connectors which are pre-loaded with various configuration entries from a previously defined template. They are database connectors that have all the SQL and scripting already provided based on the content source system. Only the connection to the content source system has to be specified.

Custom Web Service Connectors

The Web Services Connector is an extensible API based connector that allows customers to develop custom connectors to any API based system. The Connector Framework provides a standardized Web

Services interface (WSDL) which can then be integrated to enable the indexing of data from that web service.

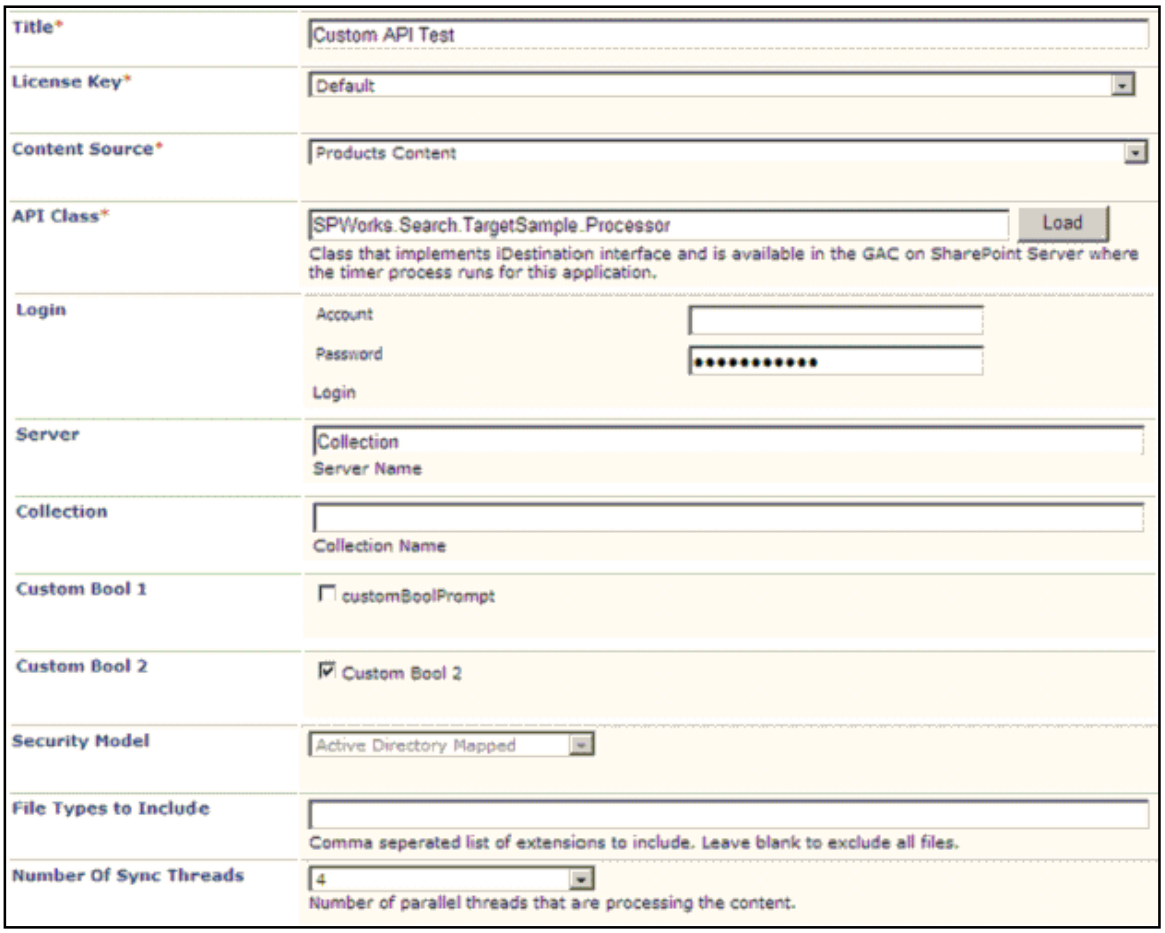
Targets

Targets are alternate destinations for the content that is extracted from the content source systems. Targets have the content pushed to them from the Connector Framework at configurable intervals.

The Connector Framework supports synchronizing content from any content source system to any target system. The supported out-of-the-box targets include:

- FAST
- SharePoint List
- SharePoint Document Library

The Connector Framework offers the capability to create custom targets as needed. Custom Targets can be included by implementing a specific Custom API provided through the Connector Framework. The target can then be configured within the Connector Framework.



Scripting

The Connector Framework provides the possibility to integrate custom VB.Net based scripts to perform various customizations to the retrieved data. There are three different main uses for scripts:

Text and Metadata

Scripts provide the ability to control the text that is sent to the search index for the Author, Title, URL and all other string metadata.

Real-Time Security Mapping

The real-time security scripting allows users to write custom security logic in order to provide advanced means of controlling access to an item.

User and Group Loading

This scripting functionality allows users to override the loading of users and groups. For example, a static user out of active directory can be set instead of mapping to the current user.

Summary

This whitepaper has provided an overview of the Connector Framework and how it can be implemented into your overall Enterprise Search based solution. The robust architecture along with the ability to scale across large amounts of data can maximize your user experience and add true value to your search endeavors.

About BA Insight

BA Insight provides software that enables organizations to rapidly implement powerful search-driven applications at a fraction of the cost, time, and risk of other alternatives. With our Knowledge Integration Platform, our customers deliver a remarkable user experience, classification, and connectivity to a wide variety of systems. It can function as a comprehensive solution or be implemented in a phased approach to meet growing organizational needs.

We serve visionary organizations such as ADP, Australia Department of Defence, Bayer, Chevron, ConocoPhillips, Deloitte, Ford Motor Company, Green Mountain Coffee, Pfizer, Rio Tinto, The Procter & Gamble Company, U.S. Army, and the U.S. Department of Veterans Affairs. Visit www.BAinsight.com for more information.

BA Insight's Connectors securely integrate over thirty enterprise systems into Microsoft SharePoint and FAST Search, providing knowledge workers with a single engine to locate relevant information, people, and expertise across the enterprise, wherever they reside. With rapid and cost-effective out-of-the-box deployment, comprehensive security-mapping and full Active Directory integration, Connectors provide organizations the ability to deliver a unified search experience and information access across all enterprise systems, including CRM systems such as Salesforce.com® and Microsoft Dynamics® CRM; ERP systems including SAP® and Oracle®; ECM systems including WorkSite®, Documentum®, FileNet®, Livelink® and Hummingbird®; and Email and Archiving Systems including Lotus Notes®, Microsoft Exchange®, EVault®; and more.

BA Insight's Search optimizes the enterprise search experience by making search results both insightful and actionable. With 's patented Search technology, every search result – regardless of its type, format, or location – can be instantly previewed, internally searched without the need to download the file, and analyzed for relevancy leveraging a number of intuitive data visualization tools. Throughout the search process, knowledge workers can identify and act upon relevant search results quickly and efficiently, with features including parametric search, real-time OCR and hit highlighting, and swift assembly of relevant results into multiple file-types for immediate action.