BA INSIGHT

Bringing Content to SharePoint On-premise and SharePoint Online Search through Secure Connectivity to Enterprise Systems

Technical Overview of BA Insight's Connector Framework and Indexing Connectors for SharePoint Servers and SharePoint Online

Introduction **2**

Indexing Connectors 2

The Fundamentals of Indexing Connectors **3**

What Do Connectors Do? 3

What Makes Connectors for Search Challenging? 3

The Connector Framework **4**

Secure Unified View 5

Unparalleled Security 6

Administration and Configuration 6

Smart Mapping 7

Feeding Search Indices 10

No-search Targets 10

Indexing Connectors **11**

Supported Source Systems 12

Creating New Connectors 13

Scalability and Performance **13**

Advanced Scenarios **13**

Summary **14**

About BA Insight **14**

Introduction

BA Insight makes search intelligent by connecting machine learning, cognitive computing, and enterprise systems to power a new generation of intranets and cognitive search solutions.

This whitepaper describes the capabilities and architecture of the Connector Framework and Indexing Connectors, which provide high performance, secure crawling of content from many different enterprise systems. These securely index both full text and metadata from source systems into the search engine, enabling a single searchable result set across content from multiple repositories. They also support a variety of important scenarios beyond traditional search.

The entire idea of multiple search locations, repositories, vaults, etc., for employees is antithetical to the way the modern worker operates. It's a world of information that they need to access. Users need to quickly find relevant information and expect to find this information through a Google-like natural language search, available filtering to refine information, and access to authoritative sources; all available through a single, modern user interface. One employee with access to information tailored to his or her needs can move mountains. Internal, internet-like search is the next generation of technology.

This paper intends to provide an overview of BA Insight's Connector framework and indexing connectors, which are used to populate a single SharePoint on-premise or SharePoint Online index. This index provides users with search results from the key authoritative sources they expect.

Indexing Connectors

BA Insight's indexing connectors provide secure connectivity and information integration with many of the most common data repositories available. They handle the persistent security mapping, data connectivity, context mapping, and enrichment that businesses need to leverage their information to make crucial business decisions.

The Fundamentals of Indexing Connectors

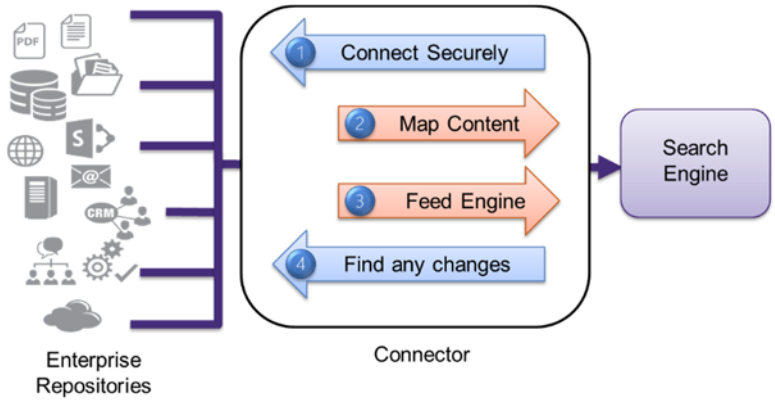
Capturing content is fundamental to search - if it's not crawled and indexed, you can't find it! Yet many organizations struggle to incorporate external content in search – it is much harder than it may seem. An understanding of the basics of indexing connectors sets the groundwork for the rest of this paper.

What Do Connectors Do?

Indexing connectors extract content from source systems and transmit it to a search engine for indexing. Each enterprise repository typically has a specific way to extract content (access method or API), a particular layout of content (schema), and specific security capabilities. Therefore, each type of system may need a connector developed specifically for it.

A connector establishes a secure connection to the source system and maps the content, including metadata and attachments, from the source system schema to the search engine schema. It then extracts content and feeds it to the search engine in a process called crawling. There are two main types of crawls:

- Full crawls, which extract all desired content
- Incremental crawls, which extract only content which has changed since the last crawl



What Makes Connectors for Search Challenging?

Many types of systems have connectors. For example, most database systems have adapters to a variety of systems and ETL (extract, transform, and load) facilities. Some business process management (BPM) and workflow systems have connectivity to enterprise systems. SharePoint has a Business Connectivity Services (BCS) facility that allows different systems' content to be surfaced as an external list. However, none of these facilities will suffice for enterprise search.

Several key requirements for connectors used for enterprise search make them more difficult than they may seem:

Unstructured content: indexing connectors must work with unstructured content as well as structured data. Large documents, attachments, and complex systems with customer-configurable schema are typical. This is not what ETL systems are designed for, but it is essential for indexing connectors.

High throughput: in order to make content complete, a copy of every desired item must be indexed. This requires very high throughput. Many installations have millions of documents or even hundreds of millions. Just one million documents indexed at one document/second would take over 11 days for a full crawl. Throughput of hundreds of documents per second can be required in practice. This is out of the range of any approach that processes an item at a time, including BPM systems and external lists.

Light touch: the flip side of high throughput is the need to minimize impact on the source system, which is usually a business-critical system in production. Indexing connectors must ensure that they will not impact the performance of these systems.

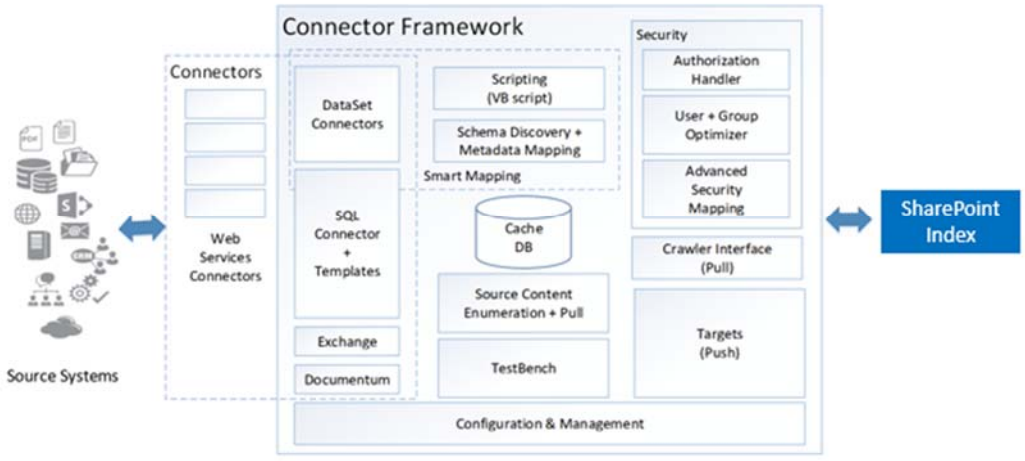
Security: it is essential that users see only content that they are entitled to see, especially because search is often used by many more people than, for example, BI. But security for search is particularly tricky, as we will detail later. ETL systems often disregard security as they move content into a data warehouse.

Click-through: with enterprise search, the source system retains the master information and the search index has only a representation (pointers). Users expect to click on a search result and be working on the original item or document. However, many enterprise systems require a specific method to bring up that item. Indexing connectors must include the click-through method for each source system, which may also vary by the type of content referenced.

To compound this, applications of enterprise search tend to be heterogeneous. They have multiple sources, with different types of information, different schema, and different security models – all in one combined result set. It's no wonder that so many people underestimate connectors, only to find themselves in difficulty.

The Connector Framework

The heart of BA Insight's indexing connectors is the Connector Framework. It acts as a scalable hub for information integration and also includes robust testing and administration features.



A set of components and capabilities plug into this hub, including:

- **Connectors** of various types (Web Services, SQL, and dataset connectors) which integrate securely with complex business systems without installing software on production systems.
- **Security Integration** across the heterogeneous security schemes used by different source systems. 'Early binding' security makes it possible to deliver secure, high-performance search solutions.
- **Smart Mapping** of content and metadata including scripting and schema management. This also provides powerful capabilities when combined with dataset connectors, such as associated crawls.
- **Targets** provide a mechanism to direct content into different places, in addition to search indexing. For example, content can be copied to a SharePoint list, along with its metadata and security settings.

Secure Unified View

BA Insight's Connector Framework provides full security and operates at high throughput to minimize crawl times – while maintaining a light touch on all source systems. It requires only read access and no client software needs to be installed on any source system server. It is scalable and incorporates redundancy for reliability as well as scale-out in content size and indexing throughput.

A library of pre-built content connectors, currently over 70, is available for a broad range of sources including both structured and unstructured content. Full support for attachments provides access to all the content in a source system. Flexible configuration allows you to index only the back-end system content you desire, presenting it to end users in the manner they demand.

The result is seamless and simultaneous access to all content. A single consolidated search index, referencing content from many repositories, is surfaced as a single unified result set with appropriate

relevancy ranking and faceted navigation. Common, consistent metadata can be created across all sources (using BA Insight's AutoClassifier) to provide great findability and navigation. This approach maximizes the value of an organization's existing ERP, CRM, ECM, and messaging systems by securely unlocking and surfacing this information in a unified view.

Unparalleled Security

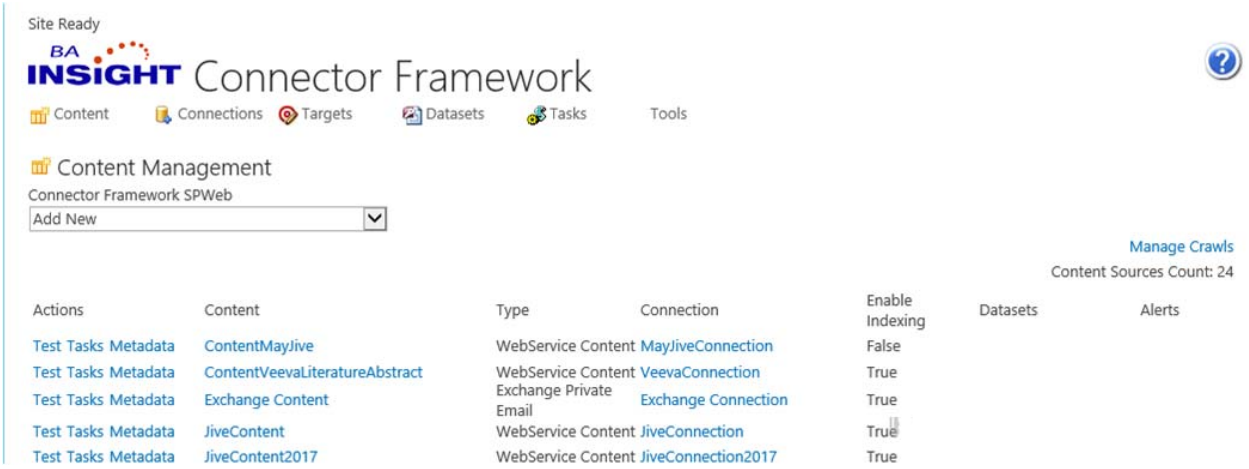
The Connector Framework provides powerful security integration across the heterogeneous security schemes used by different source systems. It identifies and maps security schemas from any system to support the early binding security needed for responsive and accurate search results. AD-based systems benefit from automatic AD group binding; non-AD systems benefit from advanced security mapping that goes beyond the claims-based security of native search platforms. This means you can handle the toughest and most sophisticated security challenges across heterogeneous systems and ensure rigorous adherence to all permission and access protocols.

Advanced Security, including role-based and attribute-based security, handles the complex security scenarios that arise with sophisticated source systems such as OpenText Documentum or dynamic authentication providers such as CA SiteMinder. The more source systems included in a search application, the more complex the security tends to be. For example, deployments with connectors to multiple different cloud systems pose daunting security issues even if each system is relatively straightforward by itself. BA Insight's capabilities for advanced security are specifically designed for heterogeneous, complex search security scenarios.

Administration and Configuration

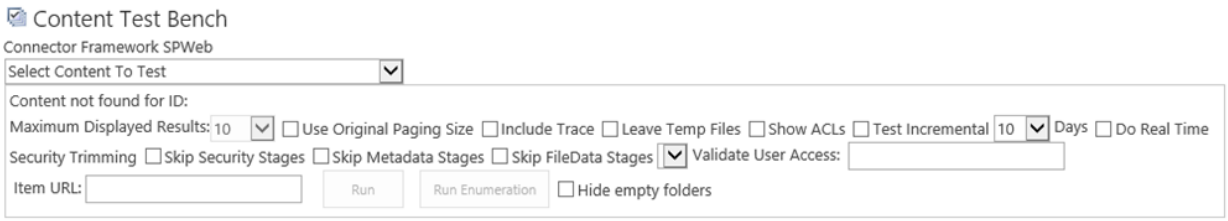
The Connector Framework makes it easy to customize and administer connectors, metadata mapping, and content targeting for all connections. It provides facilities that simplify configuration, operation, and troubleshooting of the overall system - reducing administrative effort and speeding problem resolution.

The figure below shows an administrative screen focused on content connections. From each of the tabs along the top (content, connections, targets, datasets, tasks, and tools), administrators have contextual information and actions.



Once deployed, crawling is monitored and managed transparently through SharePoint's familiar crawl management tools. Scheduled jobs for common tasks such as security sync, backups, and mailbox management enable straightforward administration.

An integrated Test Bench makes it possible to rapidly deploy, configure, and test connectivity. Administrators can test the output of any defined Content Source, display all of the properties returned for each Item, and provide visibility into performance, security, and metadata contents. It is not necessary to crawl content to use the Test Bench, so it is often used to check a deployment without populating the search index or to troubleshoot indexing without replacing or updating items.

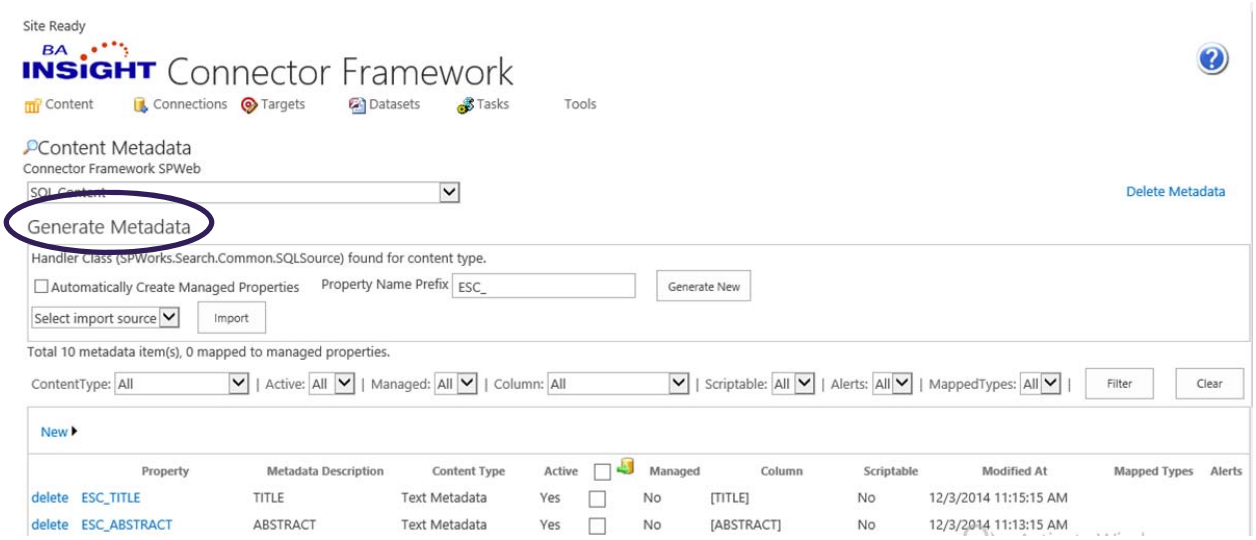


Smart Mapping

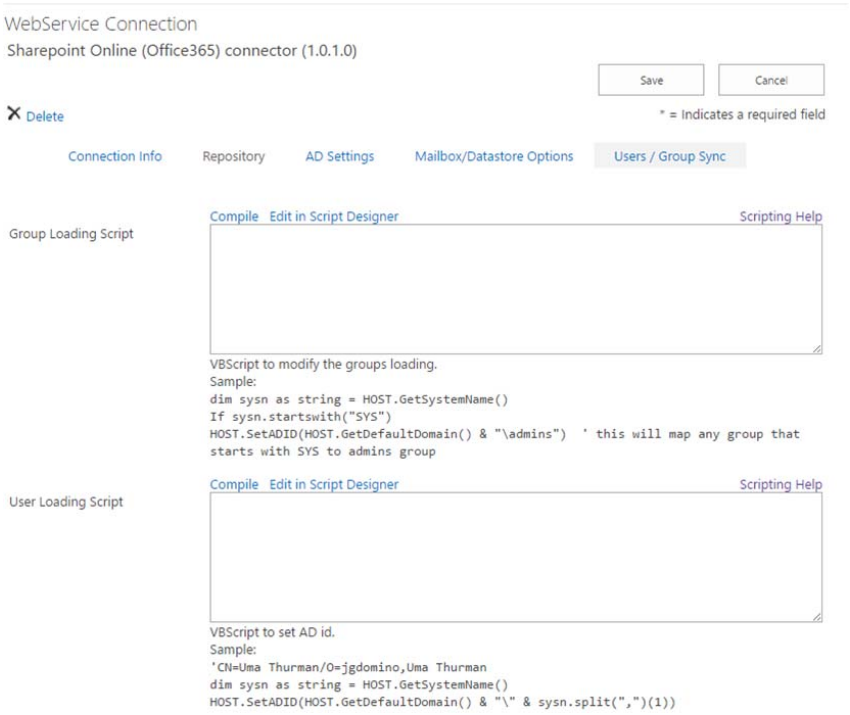
Mapping metadata schemas between source systems and a common search index is one of the most important tasks in setting up a search deployment. It is also one of the most laborious. Although SharePoint search sets up a default mapping between crawled properties and managed properties, it only

does so for content in stored in SharePoint. For all other content sources, administrators must do all of the mapping manually.

Smart Mapping makes this process much simpler by auto-generating property names and the metadata mapping based on the source system schema. It also tracks any manual modifications or overrides and respects these whenever the mapping is refreshed.



Connections, content sources, items, and metadata can be configured and extended to customize the content and tailor how search fields are populated. Custom scripting, using familiar VBScript syntax, can be applied to connections, security, crawling, and metadata to handle even the most demanding applications.



A Script Designer allows you to tweak and test scripts right from the testbench, and a library of sample scripts gets you going quickly.

Dataset Connectors are another element of Smart Mapping. These provide a way to enrich indexed content with metadata from an associated content source. Metadata can be retrieved from multiple metadata content systems, and it can be filtered to allow for fine granularity when matching the metadata to the data content. This ultimately provides a richer and more accurate search result to the end user.

Dataset connectors essentially look up and join information across systems during the crawling process.

For example, a dataset connector can combine customer data across an ERP system and a CRM system. Imagine crawling customer billing records from the ERP and retrieving the market segment designation and sales territory of the customer from the CRM system. The user performing a search can then see this information associated with each record and use it for refinement and navigation. When exploring market information, the user also sees customers in that specific market segment.

Feeding Search Indices

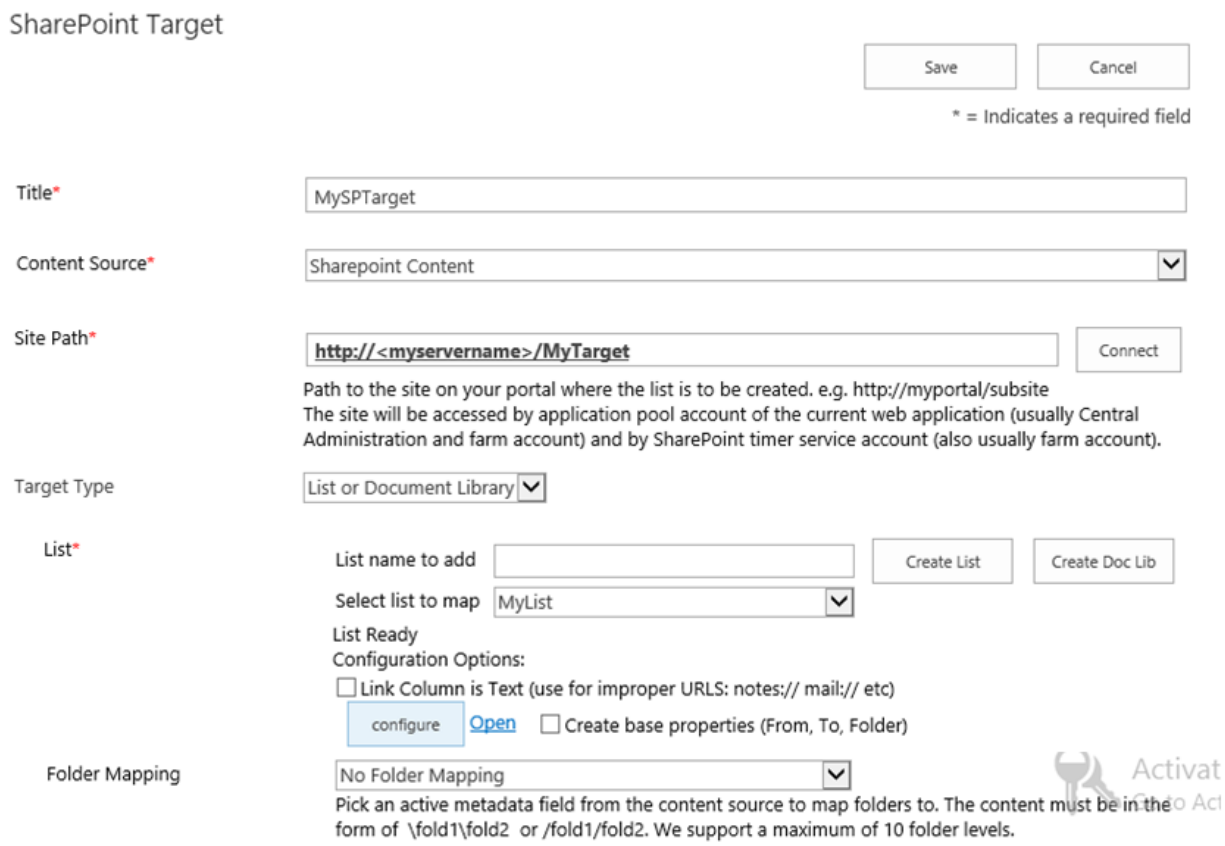
When deployed within SharePoint, The Connector Framework provides the following method to feed the search engine:

Crawler Interface – lets the SharePoint crawler poll the connector framework and pull data from it. In this model, crawls are scheduled and managed within SharePoint and look the same to administrators as 'built-in' crawls. Metadata mapping may still be managed through the Connector Framework, but the search engine can re-map the results into index fields with its own mechanisms.

No-search Targets

Targets provide a mechanism to direct content into different places, in addition to search indexing. For example, content can be copied to a SharePoint list, along with its metadata and security settings. Targets support a wide variety of advanced scenarios and are a powerful tool to the solution developer.

With Targets, a synchronization schedule is used to move new content to specific locations. This is analogous to the full crawl and incremental crawls used by search. Pre-built targets include SharePoint Lists and Libraries, the SharePoint User Profile Store, and several specialty targets. Custom Targets, fileshare Targets, and database Targets can also be created on demand.

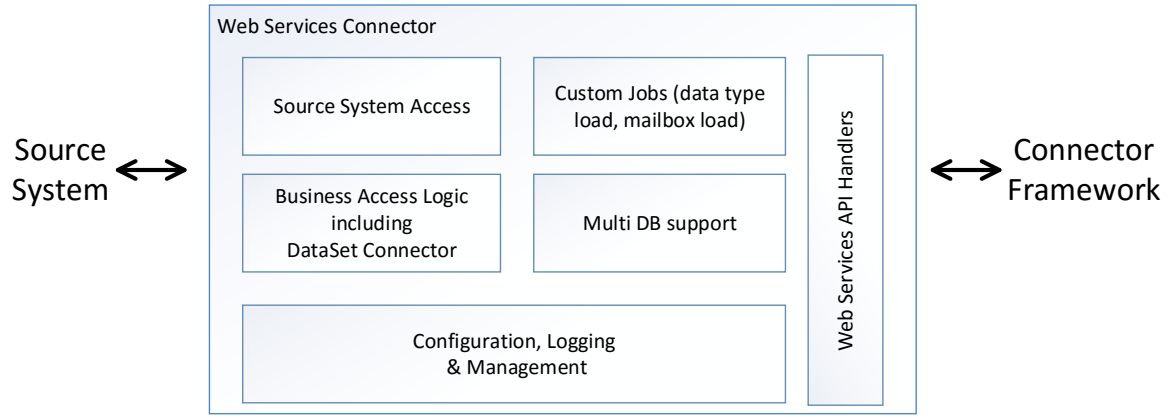


Indexing Connectors

Along with the Connector Framework, BA Insight provides a wide range of Indexing Connectors. Each connector is developed and maintained for a particular source system. There are two types of connectors:

SQL-based Connectors: for source systems that expose content via an underlying database. These connectors use a common framework with template-based administrative screens. The SQL calls are available for tailoring, either for performance optimization or to support advanced scenarios.

Web Service Connectors: for source systems that publish APIs for content access. Web services connectors include a number of functions and communicate to the Connector Framework through a published Web Services-based API. The structure of these connectors is shown below.



All connectors share a high throughput, light touch approach to selecting and extracting content. They are all agentless – i.e. they do not require any software to be installed on the source system and can communicate over a network to remote systems. They need only read access, so there is no risk of compromising source systems.

Many of BA Insight's Indexing Connectors also can act as Dataset Connectors. For example, a SQL system may have an associated file system for raw storage, or a file-based system may have an associated database holding metadata. In these cases, both the file and the metadata are indexed as a single item using an associated crawl.

Supported Source Systems

Connectors are available to over 70 systems of a variety of types:

- Aderant
- Amazon Aurora
- Amazon RDS
- Amazon S3
- Alfresco
- Azure SQL Database
- Box
- Confluence
- CuadraSTAR
- Deltek
- Elite / 3E
- EMC eRoom
- File Share
- Google Drive
- Google Cloud SQL
- HP Consolidated Archive (EAS, aka Zantaz)
- HPE Records Manager/HP TRIM
- IBM Connections
- IBM Content Manager
- IBM DB2
- IBM FileNet P8
- IBM Lotus Notes
- IBM WebSphere
- iManage Work
- Jive
- LegalKEY
- LexisNexis InterAction
- Lotus Notes Databases
- MediaPlatform PrimeTime
- Microsoft Dynamics CRM
- Microsoft Exchange Server
- Microsoft Exchange Online
- Microsoft Exchange Public Folders
- Microsoft SQL Server
- MySQL
- NetDocuments
- Neudesic The Firm Directory
- Objective
- OneDrive for Business
- OpenText Documentum
- OpenText LiveLink/RM
- OpenText eDOCS DM
- Oracle Database
- Oracle WebCenter
- Oracle WebCenter Content (UCM/Stellent)
- PLC/Practical Law
- PostgreSQL
- ProLaw
- Salesforce.com
- SAP ERP
- SAP HANA
- ServiceNow
- SharePoint Online
- SharePoint 2016
- SharePoint 2013
- SharePoint 2010
- SharePoint 2007
- Sitecore
- Any SQL-based CRM system
- Veeva Vault
- Veritas Enterprise Vault (Symantec eVault)
- West km
- Xerox DocuShare
- Yammer

Creating New Connectors

BA Insight has extensive experience in creating and maintaining indexing connectors, and a proven process for approaching new systems. The Connector Framework provides facilities for testing, troubleshooting, and optimizing content extraction, which makes creating new connectors faster and simpler. Connectors built on this framework also inherit many powerful features such as Targets and Smart Mapping,and present a consistent and effective interface to administrators.

There are two main facilities for creating new connectors. A Universal SQL connector toolkit supports secure indexing for any SQL-based source system with the flexibility to tailor the way database content is composed and transformed into indexed items. Developers can also use the Web Services API to integrate crawling into their system or to create new connectors themselves. BA Insight also provides services to create custom connectors and/or mentor developers who wish to create connectors.

Scalability and Performance

The Connector Framework allows users to process search data from different source content systems and add metadata information at high throughput. Essentially, the retrieved content is passed through with barely any performance impact. It is multi-threaded and scales out so that higher throughput can be gained with more hardware resources. A throughput of hundreds of documents per second (DPS) can be achieved on relatively modest hardware.

Typically, the bottleneck in enterprise search deployments with BA Insight connectors is not the connectors – it is the source systems themselves. For this reason, there are several mechanisms built into the Connector Framework to optimize the performance of access to source systems.

For incremental crawls, a powerful facility is used to find new items efficiently. This source system enumeration works even in the absence of efficient change logs on the source system; it speeds incremental crawls dramatically. Another significant source of stress and potential performance issues to the source system is the lookup and translation of security accounts for each indexed item. The Connector Framework successfully eliminates this bottleneck by implementing a user group synch job offline, which performs such user group loading and mapping prior to index time.

Advanced Scenarios

The Connector Framework, Indexing Connectors, and DataSet Connectors are flexible and extensible. They can be configured in a wide variety of ways to meet challenging requirements and handle advanced scenarios. In this paper we can barely scratch the surface, but a few examples should help in understanding the possibilities.

By combining DataSet Connectors with a fuzzy lookup, metadata can be normalized – resulting in cleaner content in the search index and better search results. For example, imagine a project management system hosting information about company projects. Even if the name of the project is misspelled or incomplete, a DataSet Connector can retrieve matching projects so each document indexed from any source includes a managed property containing the official name of the project.

Using this facility, along with a target for the User Profile Store, supports a consolidated user profile based on content from a variety of different sources. This improves people search, social networking, and other applications driven from the user profile.

Targets provide a simple content selection and relocation facility, as well as a content migration capability. For example, combining a SharePoint Library target with workflows, and/or the SharePoint Content Organizer, supports very flexible content selection, migration, and replication across systems and domains. Synchronizing and moving content to a specific system can be done using a custom target for that system.

Scripting can be used to call remote facilities. Examples include translating information into other languages, calling other systems to provide related information, or finding the top news stories related to an item and populating links on the indexed item so that users can click through to the story.

Summary

Indexing Connectors are more complex than they may seem. They require a balance of high performance and light touch, along with rigorous security and easy administration across a wide range of sophisticated source systems. BA Insight has a proven architecture, a wide range of supported connectors, and extensive experience to ensure secure successful search deployments.

The BA Insight Connector Framework provides a robust, flexible hub for secure, high throughput content integration. Powerful security integration across heterogeneous and complex security schemes is built-in. Smart Mapping provides automatic mapping of metadata properties; Dataset Connectors that support lookup and content normalization; and flexible content processing. The Connector Framework operates seamlessly with SharePoint Search.

About BA Insight

As an innovator in AI-driven search, BA Insight's best of breed approach helps companies make search intelligent by providing technology that connects machine learning, cognitive computing, and enterprise systems, powering a new generation of intranets and cognitive search solutions. Our customers have the freedom to leverage the best search engines and cognitive computing capabilities available, providing users with an internet-like search experience while saving them precious time looking

for needed information. We support multiple search platforms including Azure Search; Elasticsearch and Elastic Cloud; and SharePoint search (online, on-prem, and hybrid).

Our modular software product portfolio features SmartHub, delivering a personalized, internet-like user experience; connectors, providing secure connectivity to a wide variety of systems; classification, increasing findability using auto-tagging, text analytics, and metadata generation; and analytics, providing valuable data to make intelligent decisions about your intranet.

Hundreds of organizations and over 3.5 million users benefit from BA Insight's software on a daily basis to provide compelling intranets that people love to use. This includes respected organizations such as the Australian Government Department of Defence, CA Technologies, Chevron, DLA Piper, Keurig Green Mountain, Mars, Pepsi, Pfizer, and Travers Smith. BA Insight is a Microsoft Gold Certified Partner, a member of the Microsoft Enterprise Cloud Alliance, and an Elastic Partner.

Visit www.BAinsight.com for more information and follow us at @BAinsight.

