BA INSIGHT

Enterprise Integration with Elastic

Technical Overview of BA Insight's ConnectivityHub and Indexing Connectors for Elasticsearch

Introduction

Search is about rapidly obtaining relevant information and insights from large amounts of data. It is fundamental in enabling today's knowledge workers to get their jobs done. There are several key considerations when enabling internal search to meet your user's requirements:

- The *speed* at which relevant information is delivered
- The availability of data they need
- The relevance of results to the user
- The ability to analyze and summarize the data

There are multiple trends driving increased demand for internal search:

- Users demanding information
    - Over the last two decades, the internet has become more intelligent, social networking has boomed, and mobile computing is more prevalent than ever. All these factors have contributed to the user's " internet like search" expectations. Business users and consumers have grown accustomed to this "on-demand functionality" and the ability to quickly find/access information, even as the volume of data available to us has skyrocketed.
- Increasing complexity in enterprise IT
    - While the volume of different types of data within and outside the enterprise is increasing exponentially, the complexity of enterprise IT environments is also increasing dramatically.
- Increasing supply of unstructured content
    - The volume, velocity, variety, and value of unstructured content in today's digital world are all rapidly increasing. Enterprises are using an increasing number of specialized systems to improve productivity within specific functions and/or business processes. Information has become siloed within the enterprise and not available to all users. This results in decreased productivity, employee frustration, and duplication of effort. There is an opportunity for internal search to identify relevant and valuable information across an expanding variety of systems and make that information available to many more users.

The 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 they now expect to find this information through a Google-like natural language search. Results should be refinable, and users should have access to authoritative sources, all available

through a single, intuitive and modern user interface. One employee with access to information tailored to their needs can move mountains.

BA Insight provides a modular portfolio of products to address internal search needs. The solution integrates machine learning, cognitive computing, and enterprise systems to power a new generation of AI-driven internal search solutions. Internal, internet-like search is the next generation of technology.

This paper describes the capabilities and architecture of ConnectivityHub, which is the platform used to build our Ingestion Connectors, which provide high performance and secure crawling of content from many different enterprise systems. Our connectors are used to populate an Elasticsearch cluster by securely indexing both full text and metadata from source systems into Elasticsearch, thus enabling a single searchable result set across content from multiple repositories.

Ingestion Connectors

BA Insight's ingestion 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 Ingestion Connectors

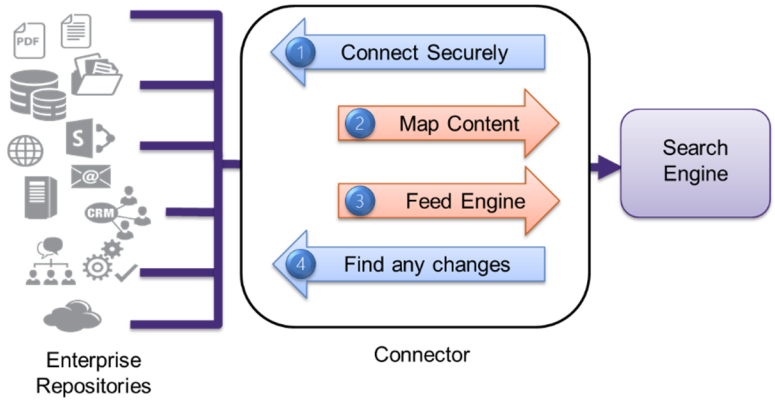
Capturing content is fundamental to search - if it's not indexed, you can't find it! Yet many organizations struggle to incorporate all needed content into search – it is much harder than it may seem. An understanding of the basics of ingestion connectors sets the groundwork for the rest of this paper.

What Do Connectors Do?

Ingestion 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 to have a connector developed specifically for that content source.

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 that has changed since the last crawl.



What Makes Connectors for Search Challenging?

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

- Unstructured content:
    - Ingestion 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 ingestion connectors.
- High throughput:
    - To make content complete, a copy of every desired item must be indexed. This requires very high throughput. Many installations have millions, if not hundreds of millions, of documents. Even 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, in practice, be required. This is out of the range of any approach that processes an item at a time, including BPM systems and per-user “Oauth” based approaches.
- Light touch:

  - The flip side of high throughput is the need to minimize impact on the source, which is usually a production, business-critical system. Ingestion 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 a lot more people than, for example, BI. As we will detail later, security for search can be particularly tricky. 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 for it to open the original item or document. However, many enterprise systems require a specific method to return that item. Ingestion 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.

ConnectivityHub

ConnectivityHub is the heart of BA Insight's ingestion connectors. It acts as a scalable hub for information integration and includes robust testing and administration features.

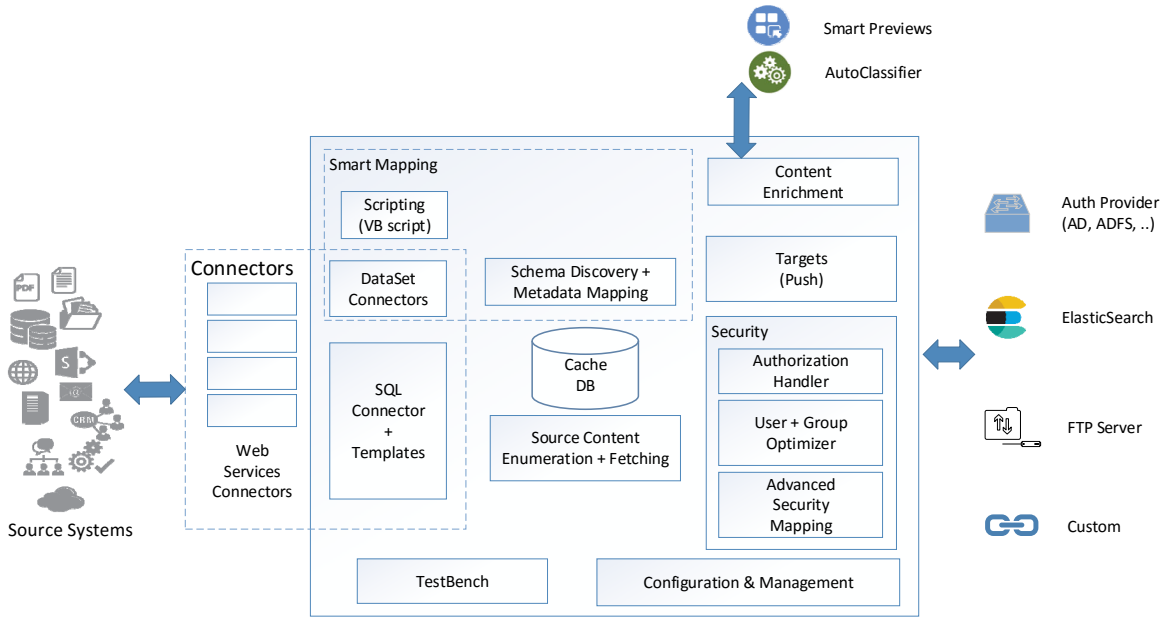


Figure 1 ConnectivityHub Architecture

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

- Connectors of various types (Web Services and SQL) 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 includes scripting and schema management. This also provides powerful capabilities when combined with dataset connectors, such as associated crawls to combine results from multiple sources into a single logical item.
- Content Enrichment allows smart processing of content pushed to the index via Classification, Machine Learning, Image or Video processing (OCR, text to speech, etc.) and Natural Language Processing.
- Targets provide a mechanism to direct content into a search engine, or even another location, such as an FTP site.

Secure Unified View

BA Insight's ConnectivityHub provides full security and operates at high throughput to minimize crawl times while maintaining a light touch on all source systems. It only requires read access to the source systems, and no client software is 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, which currently number more than 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 and to present 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, delivers a single unified result set with appropriate relevancy ranking and aggregations (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

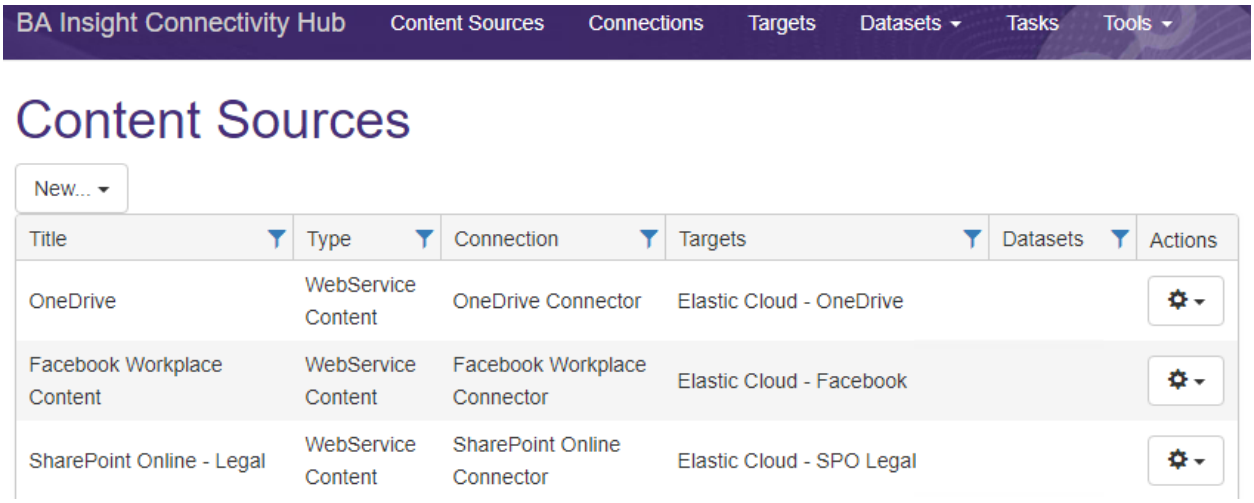
ConnectivityHub provides powerful security integration across the heterogeneous security schemes used by different source systems. It identifies and maps security schemas from any system to support the early binding security needed for responsive and accurate search results. AD-based systems benefit 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. As more source systems are 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

ConnectivityHub makes it easy to administer and configure connectors, metadata mapping, and content targeting for all connections. It provides facilities that simplifies 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 sources, connections, targets, datasets, tasks, and tools), administrators have contextual information and actions.



Connections and content sources are defined to declare which source system to index, and targets are defined to specify which search index to populate. Crawling is handled via scheduled jobs with access to historical data for straightforward administration.

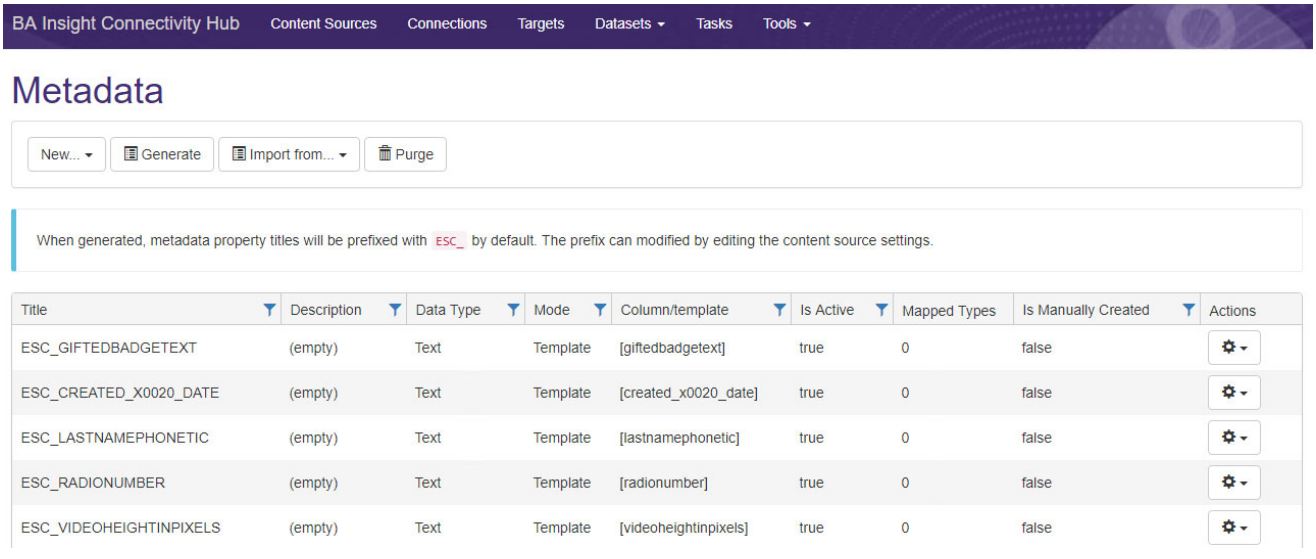
An integrated test bench makes it possible to test connectivity to confirm correct configuration of the connectors. Administrators can test the output of any defined content source, display all the properties returned for each item, and provide visibility into performance, security, and metadata contents. It is not necessary to crawl content to use the test bench, which means that it is often used to check a deployment without populating the search index or to troubleshoot indexing without replacing or updating items. An example of the output of the test bench is shown below.

Last Update: 12/9/2016 5:13:42 PM:Zone:Utc
Last Update From Enumerator: 12/9/2016 5:13:42 PM:Zone:Utc
Property: escbase_itemaclsddl Type: STRING Value: O:S-1-5-21-1169601201-886993797-1750357412-1103D:(A;;FA;;;S-1-5-21-1169601201-886993797-1750357412-1103)
Property: escbase_class Type: STRING Value: ARRAY (1): webservice
Property: escbase_id Type: STRING Value: 1
Property: escbase_extension Type: STRING Value: spw
Property: escbase_fileextension Type: STRING Value: doc
Property: escbase_crawlurl Type: STRING Value: spworks://WebService_x0020_Content/data.SPW?contentid=3&id=1&subid=&fid=d57f0103-072d-4a89-abb7-5eab6abe2977%3a%3a%2bS%2bT%2bS&fsubid=b4067782-488c-4ad9-a846-2aba26cd188f%3a&encoded=True&site=&csid=0&stype=test
Property: escbase_container Type: STRING Value: https://baidemoenv.sharepoint.com/Legal/0904-Policy-Electronic-Evidence-Discovery-Christiansen.doc
Property: 6 Type: STRING Value: https://baidemoenv.sharepoint.com/Legal/0904-Policy-Electronic-Evidence-Discovery-Christiansen.doc
Property: escbase_datastorename Type: STRING Value: BA Insight Team Site
Property: escbase_datastoreid Type: STRING Value: d57f0103-072d-4a89-abb7-5eab6abe2977::STS
Property: escbase_contentid Type: INTEGER Value: 3
Property: ows_SiteName Type: STRING Value: SharePoint Online - Legal
Property: SiteName (0b63e343-9ccc-11d0-bcdb-00805fccce04:4) Type: STRING Value: SharePoint Online - Legal
Property: escbase_public Type: BOOLEAN Value: False
Property: Url (49691c90-7e17-101a-a91c-08002b2ecda9:9) Type: STRING Value: https://baidemoenv.sharepoint.com/Legal/0904-Policy-Electronic-Evidence-Discovery-Christiansen.doc
Property: urn:schemas.microsoft.com:sharepoint:portal:isdocument Type: INTEGER Value: 1
Property: ESCBASE_MIMETYPE Type: STRING Value: application/msword
Property: Title (f29f85e0-4ff9-1068-ab91-08002b27b3d9:2) Type: STRING Value: ARRAY (1): CITL 5.1.1
Property: Author (f29f85e0-4ff9-1068-ab91-08002b27b3d9:4) Type: STRING Value: ARRAY (1): Mike Gregory
Property: ESCBASE_ITEMACLSDDL Type: STRING Value: O:S-1-5-21-1169601201-886993797-1750357412-1103D:(A;;FA;;;S-1-5-21-1169601201-886993797-1750357412-1103)
Property: LANGUAGE Type: STRING Value: en
Property: ENTITIES Type: STRING Value: ARRAY (10): COUNSEL,information,Policy,one,COMPANY NAME,action,individuals,Litigation,Workforce,Legal Hold
Property: DOCUMENTSUMMARY Type: STRING Value: This is one of a set of Information Asset Protection Policies adopted by <COMPANY NAME> ("Company") which are intended to provide reasonable assurance of the protection of valuable Company information and system assets, compliance with applicable laws concerning the confidentiality, integrity and availability of information; and accountability for systems use and information transactions. Unless emergency legal action is required (such as a temporary restraining order to prevent serious harm), <LEGAL COUNSEL> shall be

Smart Mapping

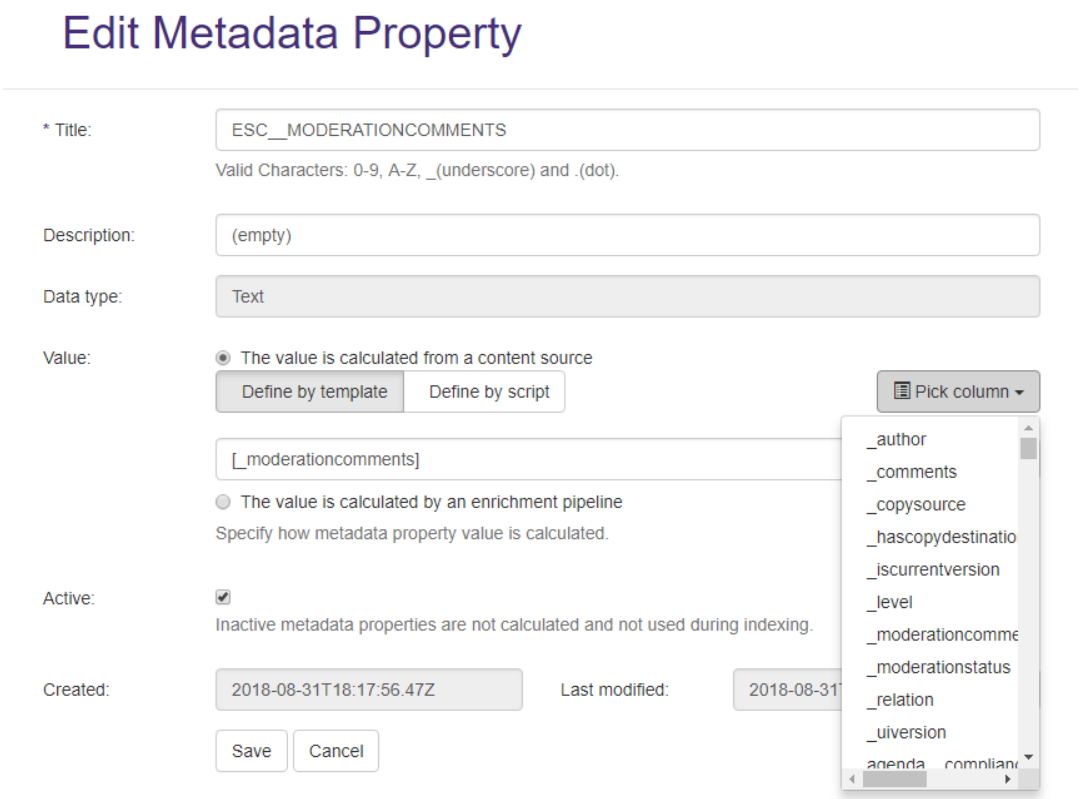
Mapping metadata schemas between source systems and a common search index is one of the most important tasks in setting up a search deployment. It is also one of the most laborious.

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.

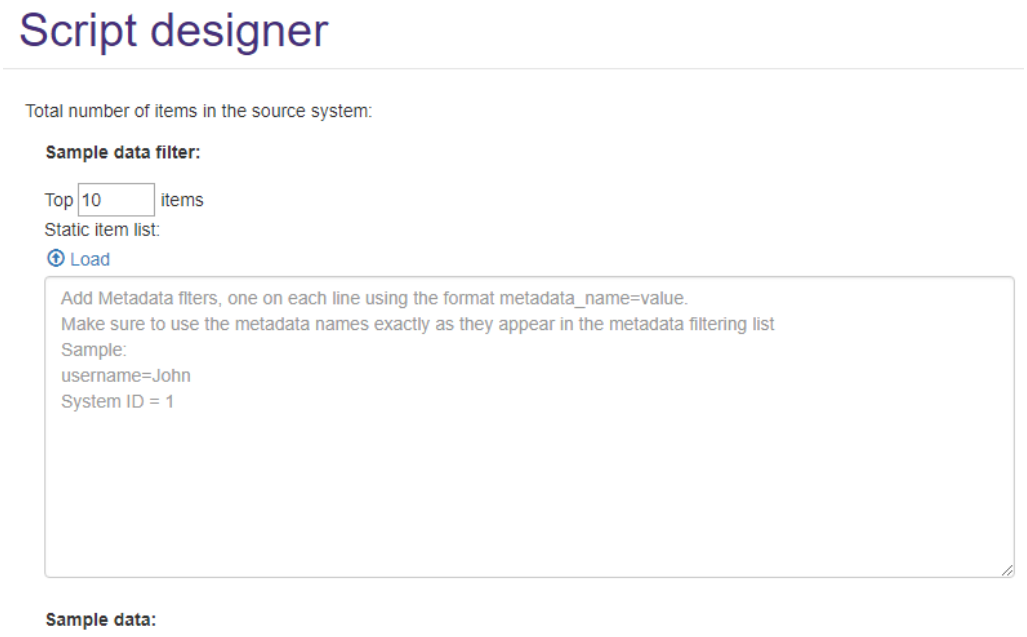


Connections, content sources, items, and metadata can be configured and extended to customize the content and tailor how search fields are populated. Custom scripting, using familiar VBScript syntax, can

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 test bench, 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 content systems, and it can be filtered to allow for fine granularity when matching the metadata to the content. This ultimately provides a richer and more accurate search result to the end user.

Dataset connectors essentially look up and “join” information across various 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 faceting and navigation. When exploring market information, the user also sees customers in that specific market segment.

Content Enrichment

For more complex data processing scenarios such as classification, multi-media processing, translation or concept extraction, ConnectivityHub can be seamlessly integrated with BA Insight’s AutoClassifier product to apply more advanced processing logic. Such logic includes machine learning-based models to classify content, OCR and object recognition, summarization and key concept extraction, sentiment analysis, etc. This information can be subsequently used to feed the search engine or decide how the content should be processed. For instance, content identified as sensitive or containing personally identifiable information can be filtered out at indexing time regardless of where the content originated.

No-search Targets

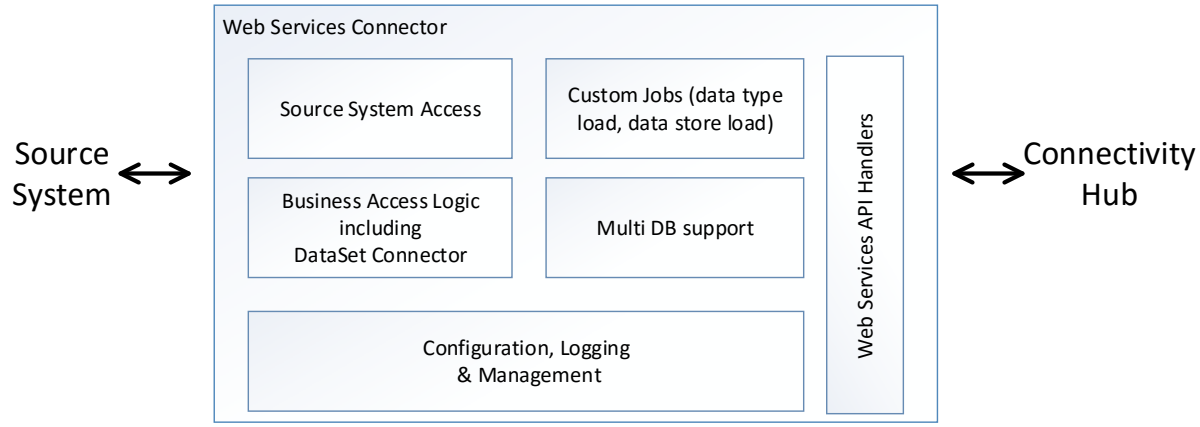
Besides the standard search index scenario, ConnectivityHub can also be used to push content to other destinations such as an FTP site. The mechanism can be, for instance, used to maintain two systems in sync or to notify a remote system when documents matching specific criteria exists, are created or updated.

Ingestion Connectors

Along with ConnectivityHub, BA Insight provides a wide range of Ingestion 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 can be tailored, 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 several functions and communicate to ConnectivityHub through a published Web Services-based API. The structure of these connectors is shown below.



Connectors are designed for high throughput and light touch when selecting and extracting content. They are agentless – i.e. no software needs to be installed on the source system and can communicate over a network to remote systems. They only require read access, so there is no risk of compromising source systems.

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

Creating New Connectors

BA Insight has extensive experience in creating and maintaining ingestion connectors, and a proven process for approaching new systems. ConnectivityHub provides facilities for testing, troubleshooting, and optimizing content extraction, which makes creating new connectors faster and simpler. Connectors

built on this framework inherit many powerful features such as Smart Mapping and content enrichment 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 professional services to create custom connectors and/or mentor developers who wish to create connectors.

Scalability and Performance

ConnectivityHub 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 minimal performance impact. It is multi-threaded and scales out so that higher throughput can be gained by adding more hardware resources. Throughput of hundreds of documents per second (DPS) can be achieved on relatively modest hardware.

Typically, the bottleneck in enterprise search deployments, when using BA Insight connectors, is not the actual connector – it is the source systems themselves. For this reason, there are several mechanisms built into ConnectivityHub 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. ConnectivityHub successfully eliminates this bottleneck by implementing a user group synch job offline, which performs such user group loading and mapping prior to index time.

Summary

Ingestion connectors are harder 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.

BA Insight's ConnectivityHub 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. ConnectivityHub operates seamlessly with Elasticsearch.

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, Elastic Cloud; and Elastic Cloud Enterprise; 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.