

Connect ▶ Find ▶ Act

Longitude v4.0 for Microsoft Enterprise Search

Technical White Paper

November 2010

Martin Muldoon

Director Product Marketing

martin@ba-insight.net

Abstract

BA-Insight's Longitude v4.0 for Microsoft Enterprise Search optimizes Microsoft Search Infrastructure. Optimizations include an advanced search interface which provides users with a more productive search experience, automatic relevance tuning, auto classification, analytics and reporting, and connectors that enable the secure indexing and surfacing of business data stored in ERP, ECM, CRM, and custom applications.

This white paper covers **Technical Architecture, Security, Capacity Planning, Deployment, Customization, and Management** topics related to the Longitude product offering. The information is intended for System Administrators or Technical Managers who are already familiar with Microsoft Search deployment and administration.

BA-Insight's Longitude Optimizes Microsoft Search Server 2008/2010, SharePoint 2003/2007/2010, FAST Search for SharePoint 2010, and FAST ESP. This document is applicable to all of the aforementioned search platforms which will be referred to as Microsoft Search throughout the course of this document.

The information contained in this document represents the current view of BA-Insight on the issues discussed as of the date of publication. Because BA-Insight must respond to changing market conditions, it should not be interpreted to be a commitment on the part of BA-Insight, and BA-Insight cannot guarantee the accuracy of any information presented after the date of publication. This document is for informational purposes only. BA-INSIGHT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, AS TO THE INFORMATION IN THIS DOCUMENT. BA-Insight may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from BA-Insight, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

Introduction	4
Architecture Overview – Longitude Search	6
Process Flows – Longitude Search	7
Indexing Content	8
Unstructured Data Querying Process	11
User selects item in search result to assess relevance	12
Supported File Formats	13
Supported Content Sources	13
Supported Languages	13
Supported Content Types	13
Availability	14
Reliability	15
Scalability	15
Integration	15
Security	16
SharePoint Upgrade Patch	16

Introduction

BA-Insight's Longitude software offering consists of two complimentary products (depicted in Figure 1 below). **Longitude Connectors** extend Microsoft Search indexing capabilities to enable indexing of content that Search Server, SharePoint, or FAST can't index "out of the box". **Longitude Search** is an advanced user interface based on Microsoft SilverLight that provides the user with several advanced capabilities. The enhanced user interface collects data as users are searching. This information is then used to automatically tune the ranking algorithm, auto-classify data, and provides the administrator with search analytics and reporting capabilities that can be used to further optimize the search experience within and organization.

Longitude Search is typically installed on a dedicated server, or virtual server, to minimize setup, configuration and ongoing maintenance of the solution. Longitude Connectors are based on Microsoft's Business Connectivity Services technology and are installed on the index servers directly.

Longitude is not a search engine. Longitude fully leverages the index, index engine, relevance ranking algorithms and security model of the respective search system deployed. Throughout this document, you will read the words "Enhanced Document", "Enhanced Processing", or "Enhanced Indexing". They refer to the processing that occurs on the Longitude server to provide end-users with advanced search capabilities.

This technical overview addresses the most commonly asked questions regarding the technical architecture of Longitude and consists of the following topic areas:

- **Architecture Overview** – Logical and physical architecture.
- **Process Flows** – How Longitude interacts with the search engine during key processes.
- **Configuration and Customization** – Corpus definition, SharePoint integration settings, and security settings. Customization of pages and web parts
- **Security** – How Longitude leverages Active Directory

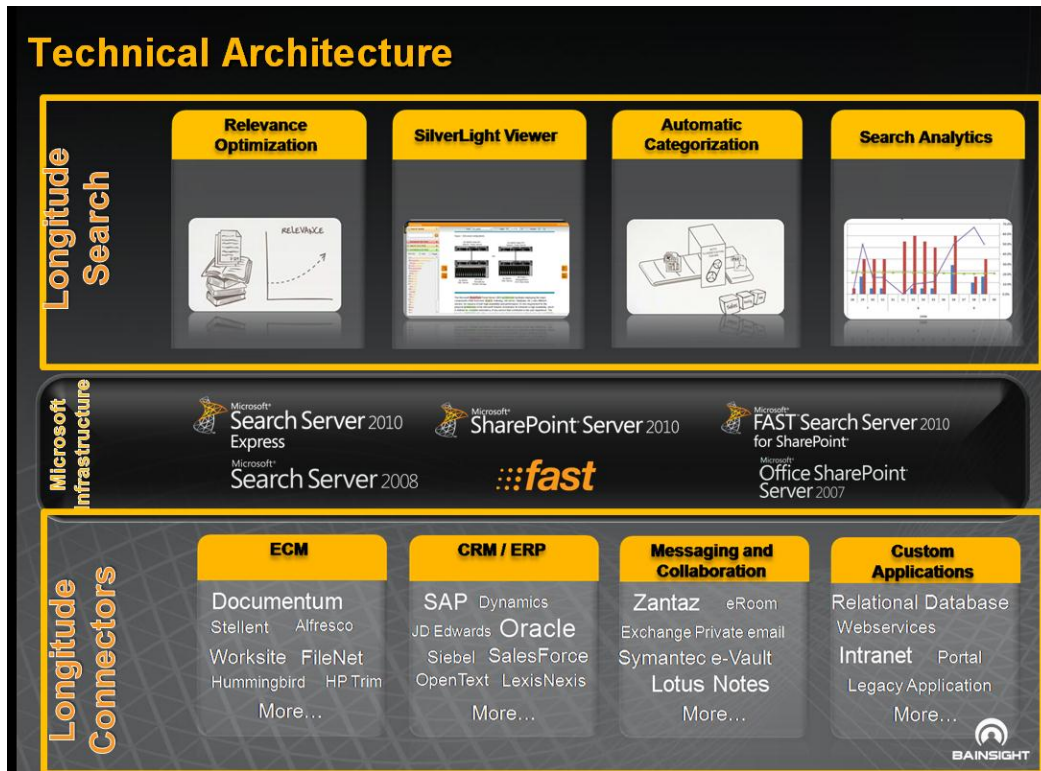


Figure 1

Architecture Overview – Longitude Search

Overall Architecture

Longitude v4.0 builds on the solid foundation provided by the .NET framework 3.5, Windows Server 2005 / 2008, and SQL Server 2005 / 2008 – All 32 and 64 bit.

Longitude Search is an extension of Microsoft Search technology. Longitude Search is compatible with all versions of Search Server, SharePoint Search, and Microsoft FAST.

SharePoint Services, and Microsoft SharePoint Foundation cannot be extended by Longitude as these systems not extensible. BA-Inight's recommendation is to deply Microsoft Search Server to index these content sources. Longitude can then extend these systms to provide users with state of the art search.

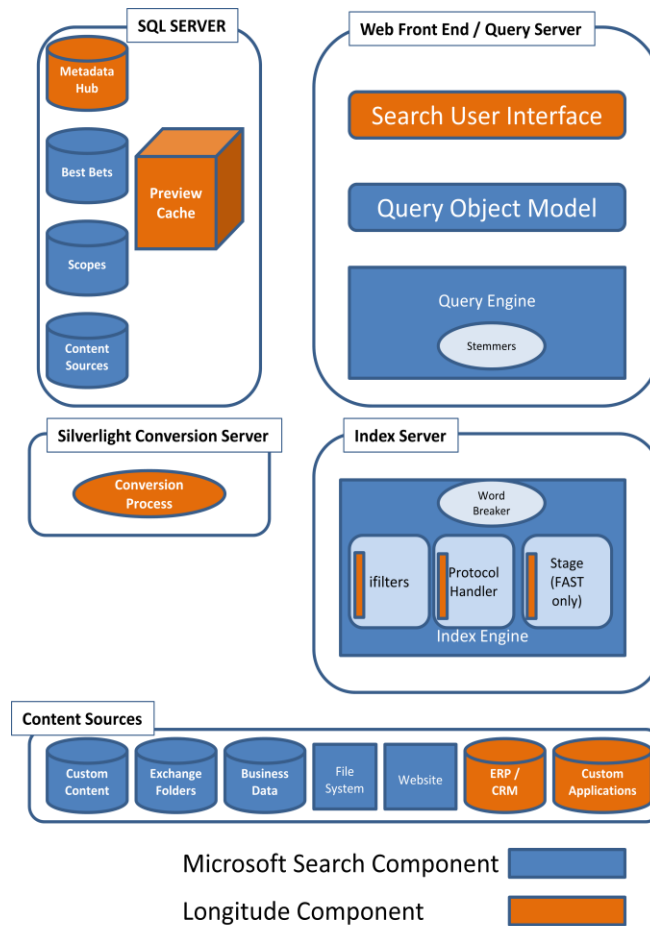


Figure 2

As depicted in Figure 2 above, Longitude Search consists of four primary components. The interface components consist of a collection of webparts that are loaded as a SharePoint Solution on the web front end(s).

A core feature / functionality provided by Longitude is a preview of the relevant section of a document when a user clicks on a link in a search result. For performance reasons, the previews are generated at index time. Creating the preview is a CPU intensive process, therefore BA-Insight recommends to any organization indexing more than 50,000 documents to run this process on a separate piece of hardware, or virtual environment.

The third component in a Longitude Search installation is an additional SQL database instance. The purpose of the database is to store the cache of previews that are generated at index time. The size of the cache is typically 40% of the content being indexed. For example 100MB of files would require enough storage for an 40MB of previews. Longitude also stores metadata in this database, but overall size of this metadata is negligible relative to the footprint of the previews.

The fourth component in Longitude is a custom iFilter that is loaded onto the index server. The purpose of the custom iFilter is to monitor the index process. Longitude must provide previews for all newly indexed documents and remove previews for documents that have been deleted. The iFilter keeps the index and the cache of previews synchronized.

Process Flows – Longitude Search

This section of the document details the interaction between Longitude and Microsoft Search during key processes.

Indexing Content

Longitude fully leverages the Microsoft Indexing Engine. It honors existing security, content sources being crawled, and indexing schedules.

The Microsoft indexing engine crawls the selected content sources and passes a copy of the document back to SharePoint for indexing. Longitude leverages and extends the crawling process. A copy of the document is temporarily passed to the SilverLight Conversion Server where it is converted to a SilverLight format. The temporary copy of the source document is then discarded. The SilverLight rendition is compressed, and stored in an SQL Server database. As Longitude doesn't index content itself, the impact of the interaction between Longitude and SharePoint is negligible relative to the overall indexing time.

Certain documents may not be converted to SilverLight because of their internal structure or password protection. Certain document types, such as AutoCAD or MS Visio, are not currently

Longitude leverages SharePoint indexing engine and is compatible with any content sources that SharePoint Search can crawl and index. In other words, it is not limited to SharePoint content.

supported by the SilverLight conversion process. In such a scenario, the documents do appear in the search result if relevant, they simply don't exhibit an enhanced result behavior.

Figure 3 depicts the interaction between Longitude and the Search engine during the indexing process.

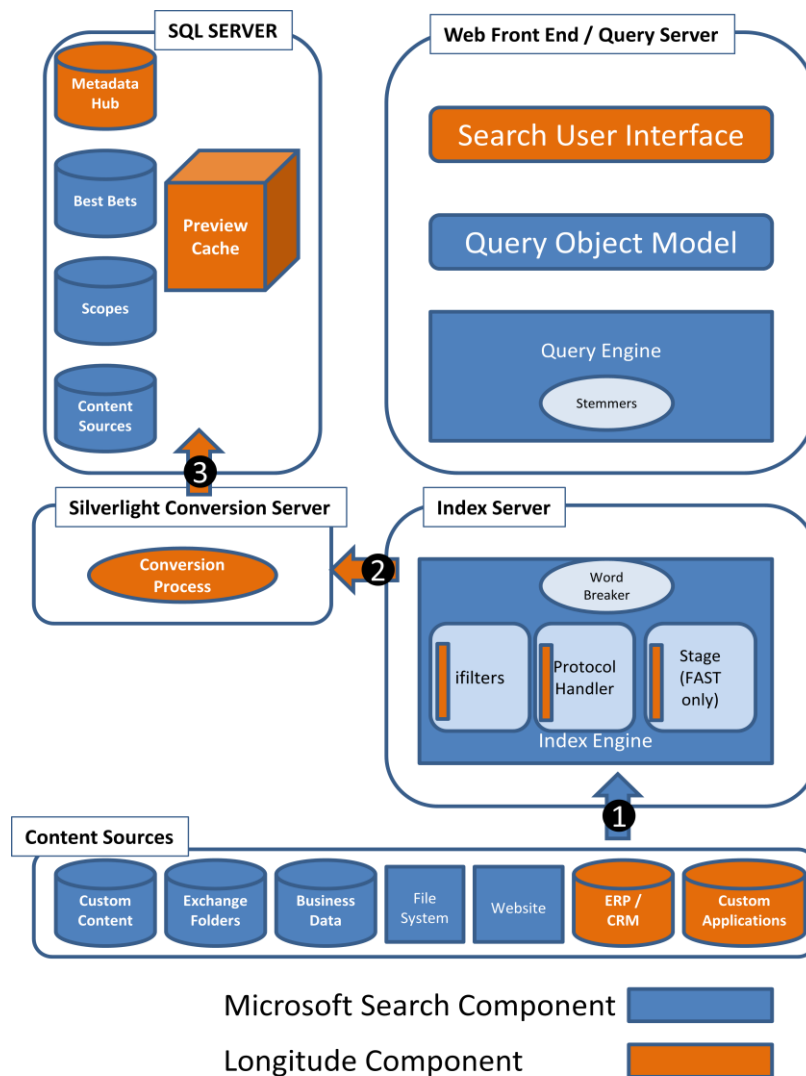


Figure 3

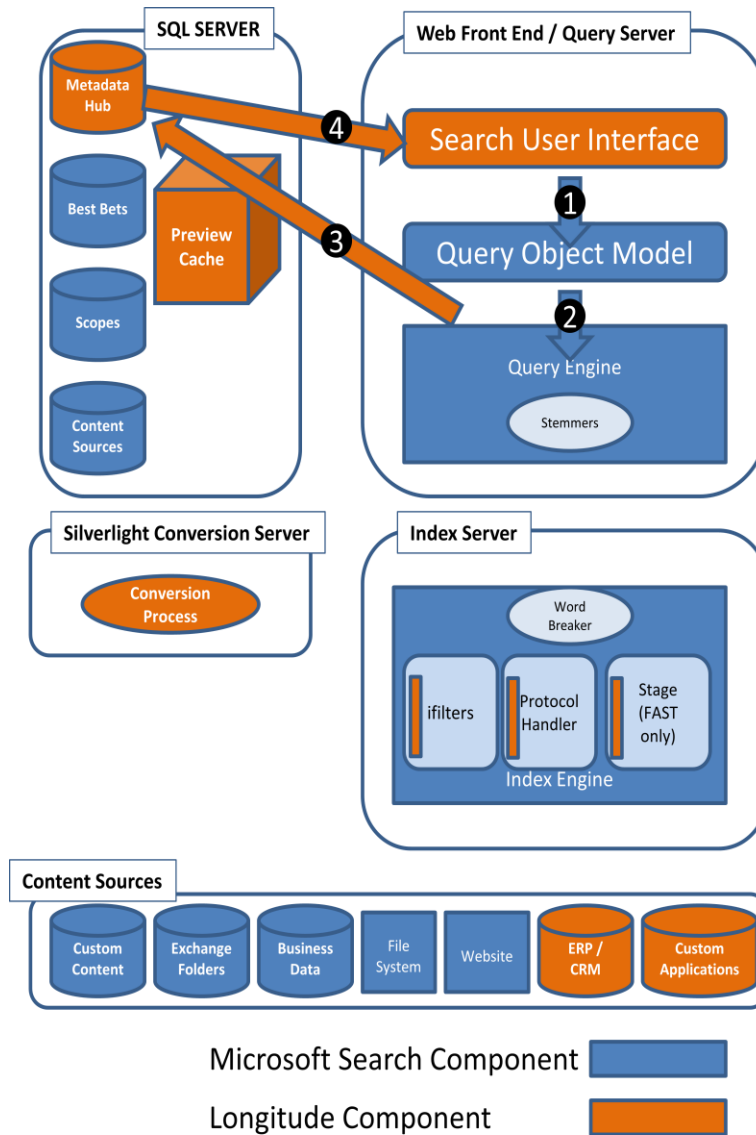
Process Flow

As depicted in figure 3 above, Key interaction between the search engine and Longitude is as follows:

1. SharePoint Indexes content sources according to the index schedule.
2. Longitude’s custom iFilter captures and passes documents to the SilverLight Conversion Server during the indexing process. SharePoint indexing process is not impacted.
3. Documents are converted into SilverLight and stored in MS SQL Server Longitude database. The collection of SilverLight documents is referred to as the SilverLight Cache. Longitude also extracts metadata from the documents during the conversion process and stores it into MS SQL Server Longitude database. The collection of metadata is referred to as the metadata hub.

Unstructured Data Querying Process

Figure 5 below illustrates the interaction between Longitude and the search system when a user executes a query.

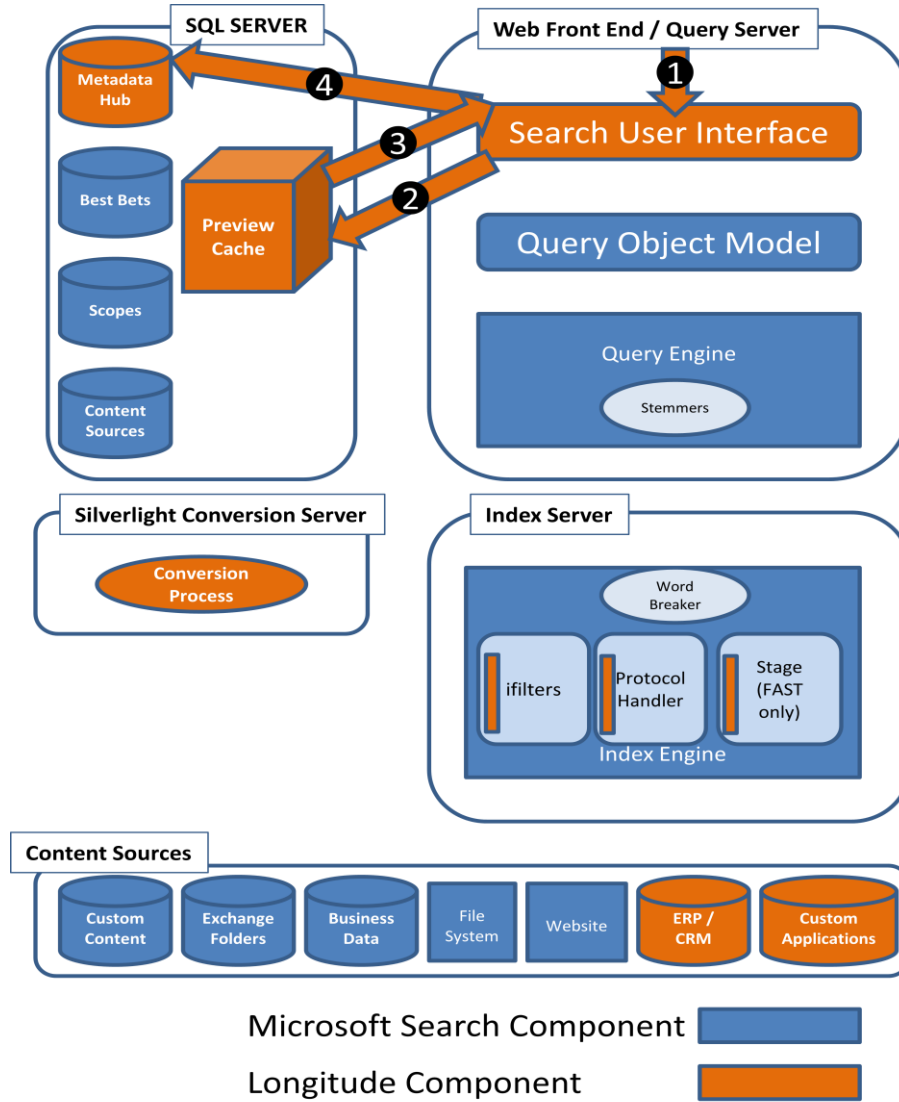


Process Flow

1. User executes a query against the search engine.
2. Query engine receives request and stems keywords. Index is queried for stemmed keywords based on security credentials of the individual executing the query. The index returns the list of relevant documents on page N that the user has (at minimum) read-access to.
3. The metadata hub is queried for the extended metadata of each document

- The enhanced search result is returned to the user

User selects item in search result to assess relevance



Process Flow

- User selects an Search result item.
- The SilverLight Cache is queried for the relevant SilverLight preview
- The SilverLight preview is returned to the query engine
- The user’s activities are logged in the metadata hub. Longitude records what user’s preview and what actions they take on documents.

Supported File Formats

Longitude support all major file formats. Any format that is not supported is still searchable and will appear in the search result, but will not have preview, hit-highlighting, and other functionality provided by Longitude.

Longitude functionality applies to content stored both in and out of SharePoint. A document in SharePoint or a file server are treated no differently in a search result.

Word (2000, 2003, 2007, 2010)	Zip
Excel (2000, 2003, 2007, 2010)	HTML
PowerPoint (2000, 2003, 2007, 2010)	Images (Gif, Png, Jpg) coming in v4.1
Tiff, PDF (automatic OCR-ing coming in v4.1)	RTF
Email coming in v4.1	Additional file types on request.
Text	

Table 1

Supported Content Sources

Longitude leverages the Microsoft Search crawling and indexing process. It follows the same schedule for indexing, and retrieves documents from any of the content sources crawled by SharePoint, i.e. file systems, SharePoint document libraries, MS Exchange Public Folders. Longitude Connectors (Separate Product) extends the reach of Microsoft Search to all of the major ECM, ERP, CRM, Messaging, and custom application content repositories.

Supported Languages

Longitude currently supports all languages Latin based languages. Additional languages can be readily added, but will be done so on per request basis.

Availability

Availability is a function of the redundancy built into the Search Server / MOSS / SharePoint 2010 / FAST farm. If the web front ends, search front ends, and SQL database servers are redundant and/or failover is enabled, then the Longitude enhanced search result is fault tolerant. The SilverLight cache is stored in a SQL server cluster/mirror and the business logic and web parts are In the event of a Longitude failure, Microsoft Search defaults back to a standard search result. At no time is search unavailable.

Reliability

Reliability is directly linked to the reliability built into the MOSS farm, whether it's the redundancy of web front and SQL database servers, or redundancy of hard drives in the database server. Many companies have a disaster recovery plan where at minimum the data is mirrored to a backup environment, but often the entire MOSS farm is full operational in a backup data center.

Scalability

Scalability is directly linked to the scalability of the Search farm. In particular, the number of concurrent and total users supported is a function of the web front ends and SQL database server's horsepower. See the Capacity Planning section for more detail on properly sizing the hardware for the Longitude Server and Additional hard drive spaced required to store the Longitude database.

Integration

Longitude has been designed from the ground up to take full advantage of your existing Microsoft Search infrastructure. It leverages the Microsoft Search indexing process and ranking algorithm, to retrieve and process the targeted documents during indexing.

The Longitude search interface has been implemented as a set of custom web parts, complying with the Microsoft Framework recommendations. Extending the search page is relatively easy by connecting your own web part to Longitude's. Specifically, you can customize the search result layout using the XSLT template, add/remove web parts, modify the page layout. But, even though it is technically feasible, we do not recommend that you extend the Longitude web parts to add special functionality. BA-Insight has seen companies extending Longitude's functionality in ways that break the upgrade path. BA-Insight's recommendation to submit your enhancement requests to BA-Insight support. Provided the request is can be leveraged by other companies, BA-Insight will typically implement a feature recommendation free of charge, or at very little cost to the interested organization.

Security

Only SharePoint Shared Service Provider can communicate with the Longitude Server and Longitude database. The security component trusts communication from Longitude signed assemblies only. Those customers who choose to reuse the Longitude web parts, or attempt to query directly the Longitude Server and database, will be directed by the security component to run the query against SharePoint Index to ensure that proper access rights.

The integrity and protection of the data residing on the Longitude SQL database is guaranteed through numerous restrictive gatekeepers. The Longitude server is a self-contained environment only interacting with your network through two interfaces, a network share and a secured web service. Both are configured to allow only a limited set of accounts to interact with the Longitude server. Calls are only authorized from a pre-defined user account and a signed application ensuring the validity of the requests. Moreover, any query to the Longitude database that has not been initiated by Longitude's signed assemblies dismissed. Longitude database calls have to be done using a very specific account to be answered, such as the Longitude account and or the Search Center application pool account.

SharePoint Upgrade Path

Longitude extends rather than replaces the Microsoft Search engine through interface points defined by Microsoft. Upgrading to MOSS, or SharePoint / FAST 2010 is warranted to customers. BA-Insight product development is tightly integrated with the development cycle of the Microsoft Redmond SharePoint product team, with ties to all key Group Program Managers in the Search and BI arena. As a result, no new features in SharePoint collide with Longitude's added functionality, which gives the customer peace of mind in terms of protecting its investment in Longitude.



United States

445 Hamilton Avenue, SUITE 1102
White Plains, New York 10601
United States of America
Tel | 914.220.8395

Europe

Vester Voldgade 106,
DK-1552 København V
Denmark
Tel | +45 21 14 48 17 6500

Web | www.BA-Insight.net
Email | info@ba-insight.net