



Longitude v3.0 for MOSS Technical White Paper

Published: January 2007

Abstract

Longitude v3.0 for Microsoft Office SharePoint Server extends SharePoint Search, Business Data Catalog, and Excel Services to provide innovative features that empower users to find and understand information 10x faster than traditional search and business intelligence technologies.

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 MOSS deployment and administration.

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.

© 2007 BA-Insight Inc.. All Rights Reserved.

Contents

| | |
|---|----------|
| Introduction | 1 |
| Architecture Overview – Unstructured data..... | 2 |
| Overall Architecture | 2 |
| Unstructured Data Indexing Process | 3 |
| Unstructured Data Querying Process | 4 |
| Supported File Formats..... | 4 |
| Supported Content Sources..... | 5 |
| Supported Languages..... | 5 |
| Availability | 5 |
| Reliability..... | 5 |
| Scalability | 5 |
| Integration | 5 |
| Security | 6 |
| SharePoint Upgrade Path | 6 |
| Capacity Planning and Deployment – Unstructured Data | 7 |
| Estimating Capacity Requirements..... | 7 |
| Longitude Farm and MOSS Farm Topology | 7 |

Introduction

The Longitude software product is installed on a dedicated server, (or Appliance), to minimize configuration and maintenance efforts, as well as support the intensive processing of building the structured and unstructured data cache. Longitude plugs seamlessly into your existing MOSS infrastructure to deliver advanced search and BI for the masses capabilities.

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 and BI for the masses 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.

Capacity Planning and Deployment – Capacity planning and deployment scenarios.

Configuration and Customization – Corpus definition, SharePoint integration settings, and security settings. Customization of pages and web parts.

Performance Monitoring – Appliance performance and health monitoring.

Technical Architecture

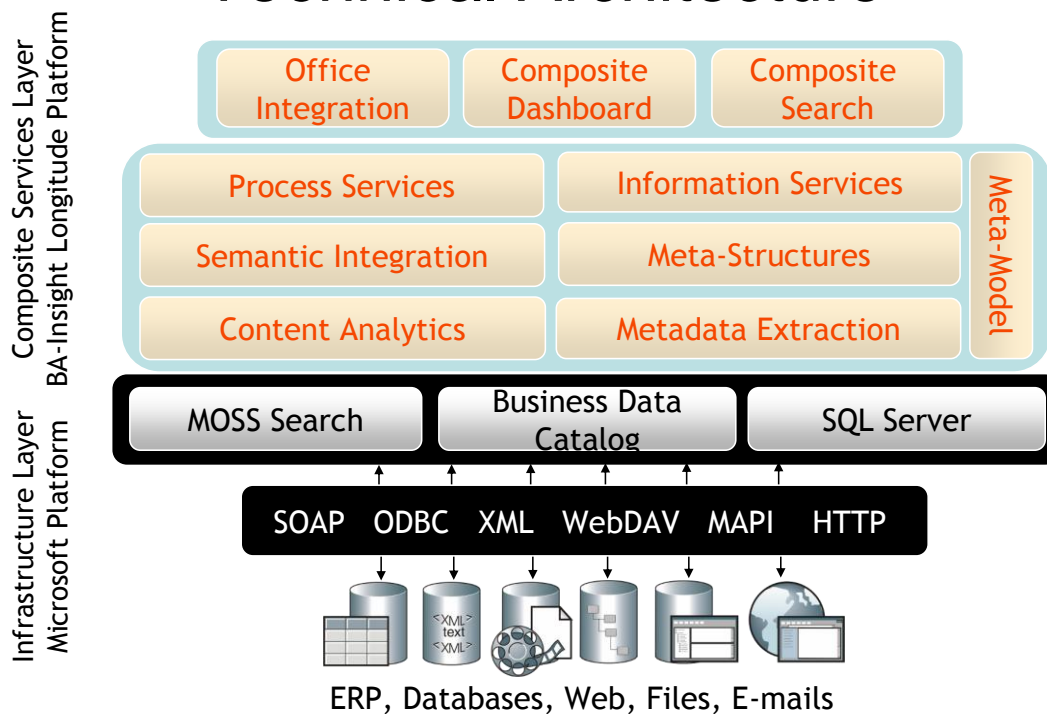


Figure 1.0

High-level technical architecture diagram

Architecture Overview – Unstructured data

Overall Architecture

Longitude v3.0 for MOSS builds on the solid foundation provided by the .NET framework v2.0, Windows Server 2003 SP1, and SQL Server 2005.

Longitude can be delivered as an Appliance or a software only package. Unlike the Google Appliance, Longitude is not a “black box”. It runs on Windows 2003 Server, complies to Domain network and security policies, supports Anti-Virus software, and be remotely and centrally managed by Administrators.

Longitude plugs into your existing MOSS infrastructure, as depicted by the diagram of a typical high-availability SharePoint server farm below.



Longitude does not replace the SharePoint Search engine! Longitude extends SharePoint Search guaranteeing future enhancements from Microsoft are fully leveraged from product updates

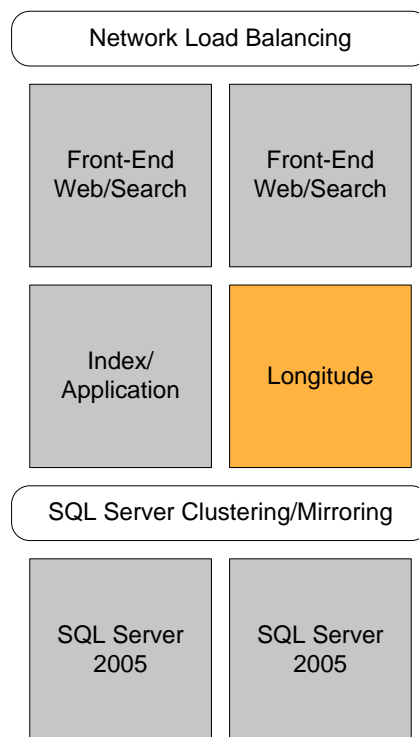


Figure 1.1

High-level physical architecture for baseline high-availability server farm

The core Longitude platform was built entirely on the Microsoft .NET framework version 2.0. Text analytics are based on the past five years of research work published by the ACM’s Special Interest Group in Information Retrieval (<http://www.acm.org/>).

The Longitude platform is powered by best of breed SVG conversion utility with the highest rendering quality achievable.

Unstructured Data Indexing Process

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

The SharePoint 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 Appliance where it is converted to an SVG format. The temporary copy of the source document is then discarded. The SVG rendition is compressed, and stored in the 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 SVG because of their internal structure or password protection. Certain document types, such as AutoCAD or MS Visio, are not currently supported by the SVG 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 (most relevant page preview, hit highlighting, etc.).



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.

As depicted in figure 1.2, the steps are:

Step 1a – SharePoint Indexing Server crawls the content sources, both portal and non portal content, based on the schedule defined by the SharePoint Administrator.

Step 1b – A custom IFilter defined for the supported file types passes a copy of the document to the Longitude server for further processing. In particular, It is converted to SVG and stored in the database.

Step 1c – A unique identifier has been added to the document properties for later retrieval (See next section about the querying process).

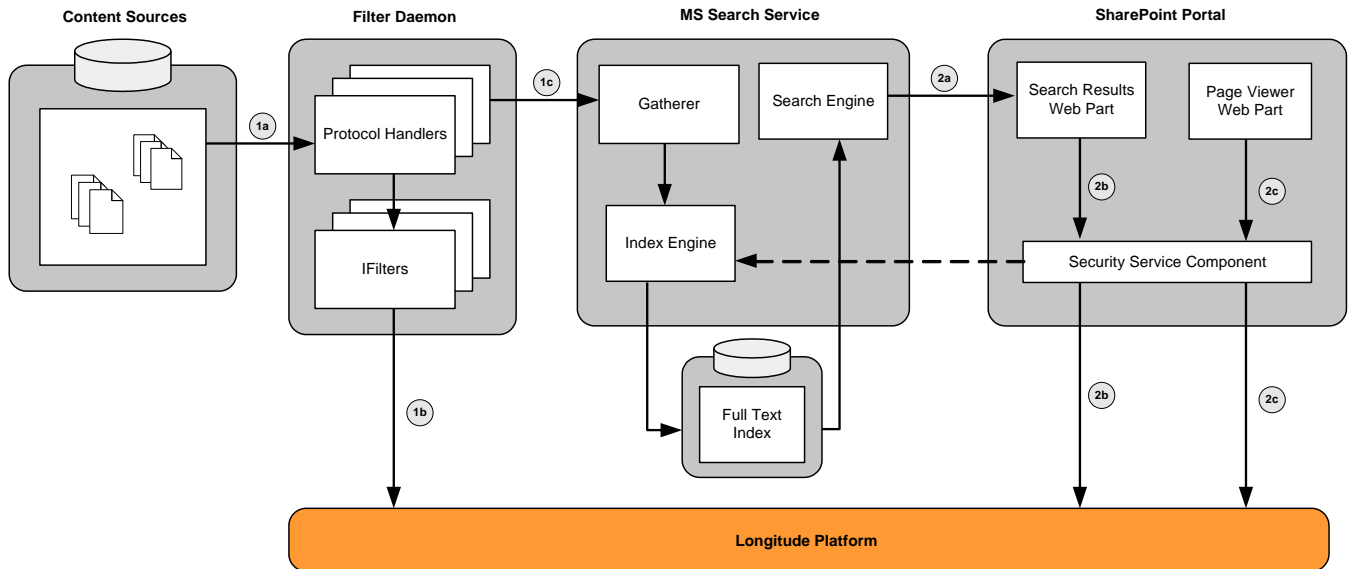



Figure 1.2

Longitude platform process interaction diagram

Unstructured Data Querying Process


As depicted in figure 1.2, the steps are:

 Stress Tests shows that the search result is returned in less than one second for up to 50 concurrent searches per second, on a dual-CPU Longitude Server, Medium MOSS farm (2 web front end, 1 SQL Server, 1 Index Server), and 2MM indexed items.

Step 2a – The Search Result Web Part queries SharePoint Index for list of documents that the user has access to. SharePoint Security defines what content the user has access to.

Step 2b – The Search Result Web Part then queries the Longitude server, via a dedicated and secure connection, for the enhanced search result, including most relevant page preview, top pages, document popularity, key concepts, taxonomy navigation, etc.

Step 2c – The Page Viewer Web Parts retrieves, via the dedicated and secure connection to the Longitude Server, the page being previewed by the end-user.

 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.

Supported File Formats

The Longitude product currently supports Microsoft Word, PowerPoint, Excel, Adobe PDF, WordPerfect, Text files, and Tiff out-of-the-box. Longitude architecture is extensible and can easily support additional file format if the need arises. Please refer to Configuration and Customization for more information. Formats that are not supported are still searchable through SharePoint. They will not exhibit “Enhanced” features.

Supported Content Sources

The Longitude Appliance leverages the current SharePoint 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, Lotus Notes, Documentum (via Third-Party Protocol Handlers), etc.

Supported Languages

Longitude currently supports all languages supported by MOSS, i.e. Documents in those languages will be fully supported in the retrieval, preview, and key concepts. Localization of specific actions in the Longitude Web Parts may be modified directly in the localization configuration file.

Availability

Availability is first and foremost a function of the redundancy built into the MOSS farm. If the web front ends, search front ends, and SQL database servers are redundant and/or have failover, then the Longitude enhanced search results are always on. The cache is stored in the SQL server cluster/mirror and the business logic and web parts are run on the web front ends. Medium to Large MOSS farm deployments typically deliver such high availability.

In the event of a loss of network connectivity between the Longitude Server and the SharePoint environment, or a hardware failure of either the Index server or the Longitude Server, only newer are not indexed (if index server failed) or enhanced (if Longitude server failed).



Note that Longitude servers can be setup as a cluster as well for added throughput. It has the benefit of providing failover in case one Longitude server goes down.

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 MOSS farm. In particular, the number of concurrent and total users supported is a function of the web front ends and SQL database servers 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 SharePoint technical infrastructure. It leverages the SharePoint 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. We have found

that it breaks the upgrade path of Longitude releases and require extra development/support work in the part of the customer to keep up with the Longitude releases. It is our recommendation to submit your enhancement requests instead to BA-Insight support.



Note that SharePoint 2003 search web parts used to be extensible, and Microsoft decided to seal all MOSS search web parts.

Security

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 SharePoint Search engine through interface points defined by Microsoft. Upgrading to SharePoint 2007 or SharePoint 14 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.

Capacity Planning and Deployment – Unstructured Data

Estimating Capacity Requirements

This section helps you plan your capacity requirements so that you can choose the appropriate Longitude configuration for your SharePoint deployment.

The key drivers are:

The size of the index (in terms of # of indexed items) of the MOSS Shared Service Provider Search service targeted by Longitude.

The refresh rate of the document corpus to be enhanced, i.e. the number of new or modified documents within a 24 hour period. Note that a 5% to 10% of the corpus size refresh rate is the recommendation provided by Microsoft when sizing up the indexing server hardware and farm topology. SharePoint Indexing logs can provide a realistic estimate of refresh rates and size of the index.

Longitude Farm and MOSS Farm Topology

One or more Longitude servers can be added to increase enhanced processing throughput of the Longitude Farm. Below is an example of a medium to large MOSS Farm and network communications between the MOSS Farm and Longitude Farm. Load balancing between Longitude Servers defined on each server for processing, so that if one server were to go down, the other servers would know how to share the load until the failed server is restored.

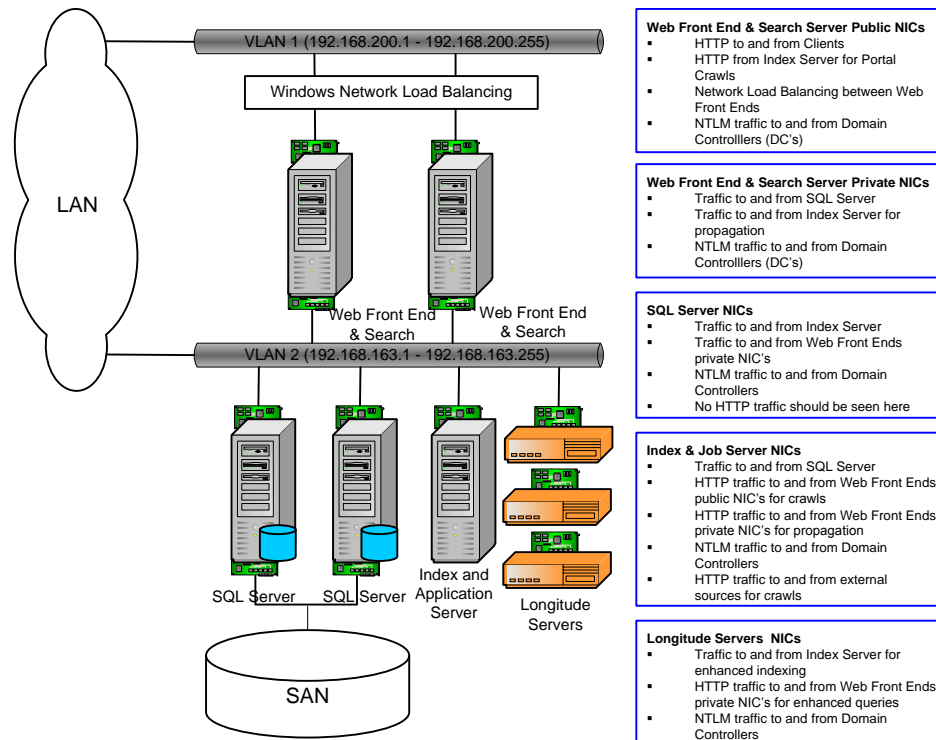


Figure 2.0

Network setup for high-availability (medium) server farm and Clustered Longitude Appliances