Module 8: Search and Indexing

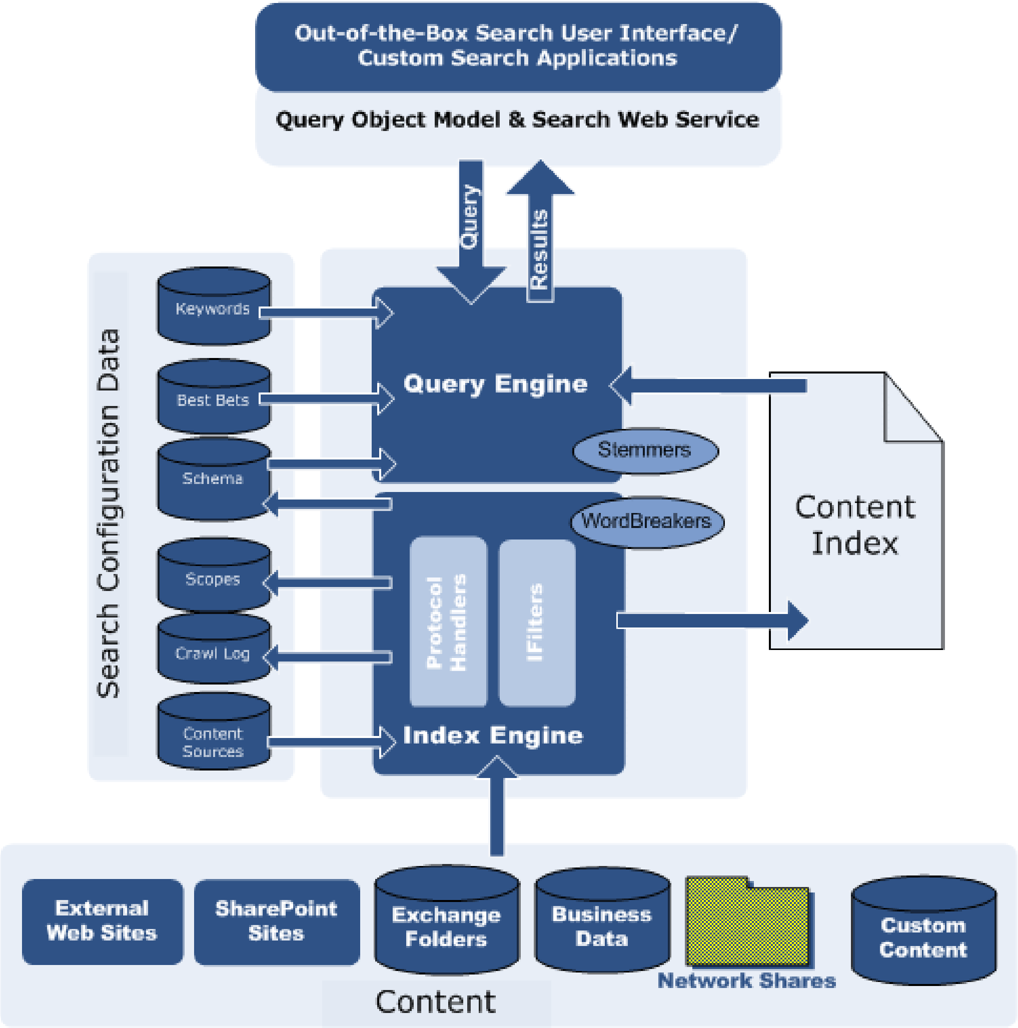
Overview

- Search Architecture
- Configuring Crawl Processes
- Advanced Crawl Administration
- Configuring Query Processes
- Implementing People Search
- Administering Farm-Level Settings

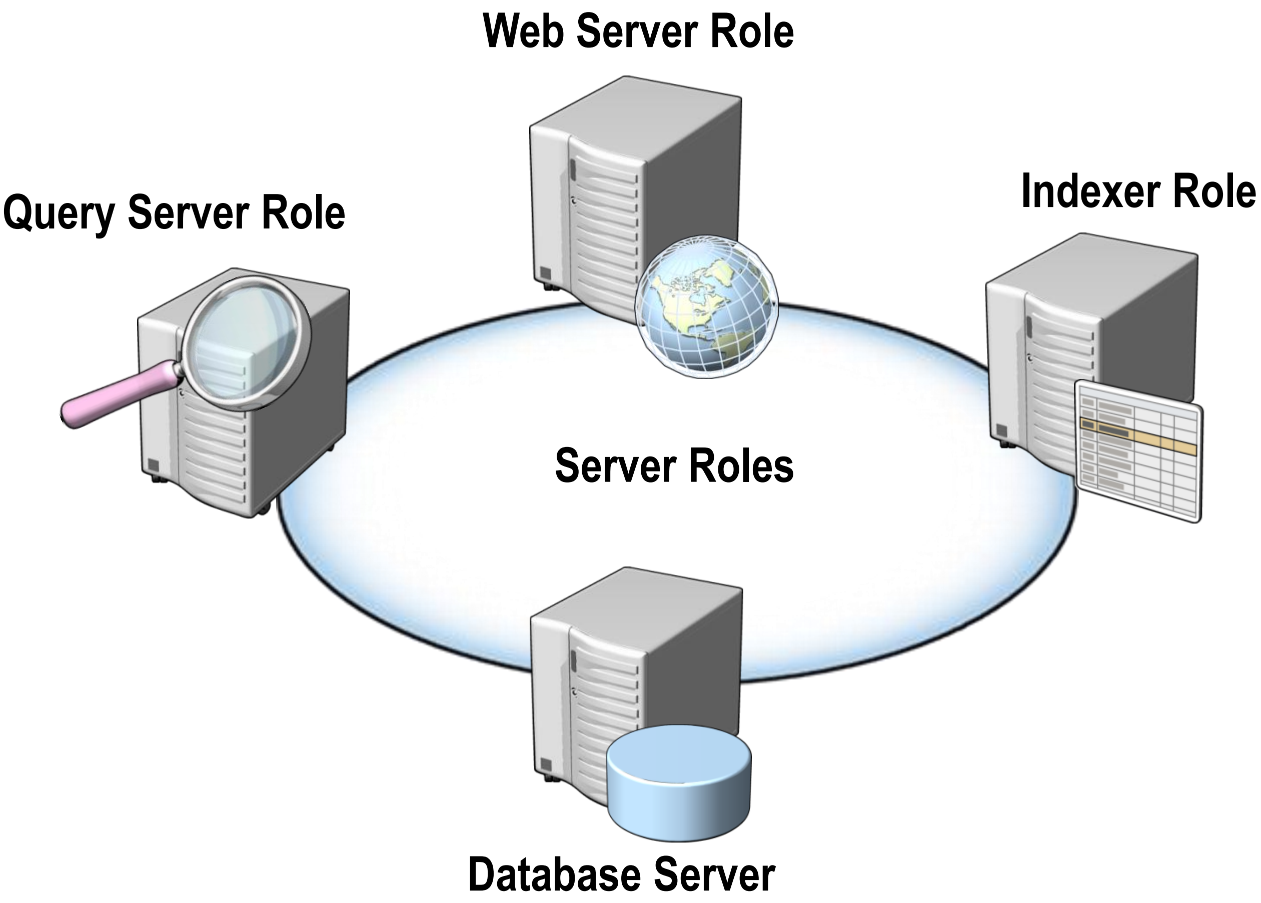
Lesson 1: Search Architecture

- Indexing and Search Architecture
- Server Roles
- Indexing Processes
- Protocol Handlers
- iFilters
- Word Breakers and Stemmers
- 32-Bit and 64-Bit Architectures
- Index Propagation
- Query Processes

Indexing and Search Architecture



Server Roles



Indexing Processes

1. The indexer retrieves the start addresses of content sources
2. The indexer invokes a protocol handler to connect to and traverse the content source
3. The protocol handler identifies content nodes, such as files and Web pages
4. The protocol handler retrieves system-level metadata and access control lists
5. The protocol handler invokes the iFilter associated with the content node type
6. The iFilter retrieves content and metadata from the content node
7. Content and metadata are parsed by the word breaker and are added to the full-text index
8. Metadata and access control lists are added to the search database

Protocol Handlers

Protocol Handlers

Connect to and traverse content sources over a given protocol. Identify content, invoke iFilters, retrieve system-level metadata, and return content and metadata streams to the index engine.

- **Protocol Handler Characteristics**
  - Web Protocol Handler
  - SharePoint Protocol Handler
  - File Protocol Handler
  - Exchange Public Folder Protocol Handler
  - Business Data Catalog Protocol Handler
  - Lotus Notes Protocol Handler

iFilters

iFilters

Open content nodes in their native format. Filter out embedded formatting and retrieves content and properties.

- **iFilters included in Microsoft Office SharePoint Server 2007**
- **Additional iFilters**

Word Breakers and Stemmers

- **Word Breakers in the Indexing Process**

  Identify breaking characters, such as white spaces and punctuation and then identify words to be indexed

  Language-specific word breakers and compound words

- **Word Breakers at Query Time**
- **Stemmers**

  Inflectional forms: Nouns and Verbs

- **Stemmers in the Indexing Process**

  Morphological Analysis

- **Stemmers at Query Time**

  Morphological Generation

- **Enabling Language-Specific Stemmers**

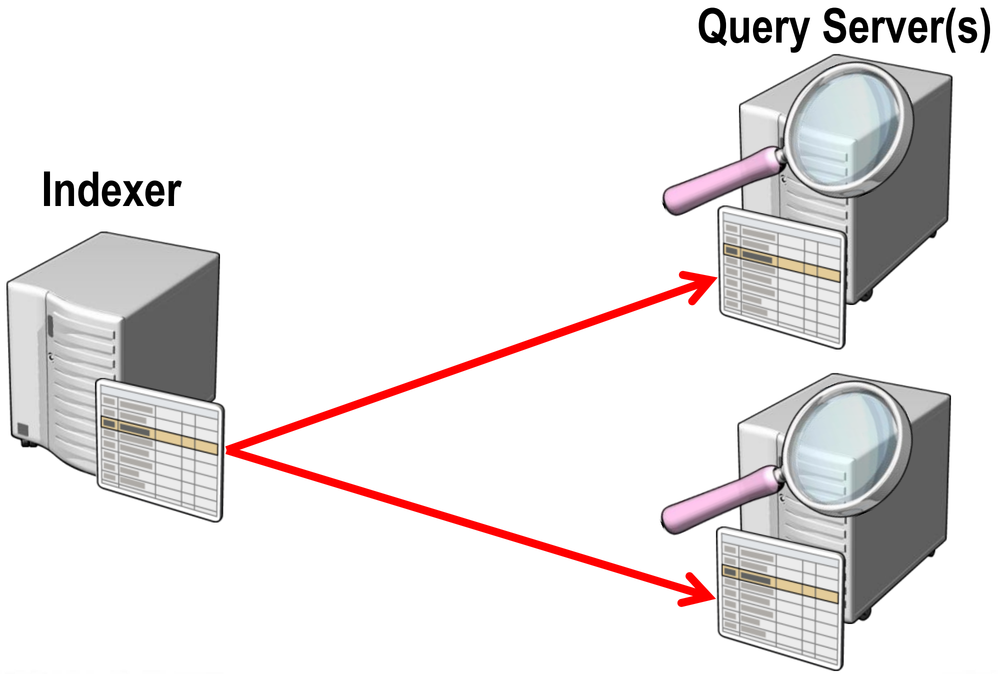
32-Bit and 64-Bit Architectures

- **Query Servers and Web Servers**
- **Index Servers**
  - Availability of Protocol Handlers
  - Availability of iFilters

Index Management and Propagation

- **Master and Shadow Indexes**
- **Continuous Propagation from Index Server to Query Servers**

  Between 3 and 30 seconds for an indexed document to be searchable



Query Processes

1. **Query terms are collected by a Web server**
2. **Query terms are supplemented with contextual information**

   Who is the user?

   Where is the user?
3. **The Web server initiates the query by:**

   Contacting a query server to run the query on the full-text index

   - Stemmers and thesaurus expansion are used (if activated)

   Contacting the search database for managed properties and access control lists
4. **The Web server security-trims the results and returns them to the caller**

Lesson 2: Configuring Crawl Processes

- Creating Content Sources and Crawl Schedules
- Creating Crawl Rules
- Full and Incremental Crawls
- Optimizing Crawl Schedules

Creating Content Sources and Crawl Schedules

- **Content Source:**

  A specification of a protocol handler with at least one start address

  Example: Web Content Source with a start address of http://moss.litwareinc.com

- **Office SharePoint Server 2007 Capabilities:**

  Up to 500 content sources per Shared Service Provider

  Up to 500 start addresses per content source

- **Crawl Schedules Associated with Content Sources**

  Allows for segmentation of the corpus into manageable sections

Creating Crawl Rules

- **Adapt the Behavior of the Typical Crawl Process**
  - Addresses can be pattern matched for special treatment
  - Supports exclusion rules and inclusion rules
  - Supports altering the authentication mechanism
  - Supports crawling SharePoint sites as HTTP pages
- **Multiple Rule Order of Precedence**
  - Rules applied in a configurable order

Full and Incremental Crawls

- **Full Crawls**
  - Re-crawl existing indexed documents and new documents
  - Update crawl behaviors based on configuration changes in Office SharePoint Server 2007
- **Incremental Crawls**
  - Only crawl new or modified content
  - Dependence on Protocol Handler Characteristics
  - WSS 2.0
  - WSS 3.0
- **When Are Full Crawls Required?**
- **WSS 3.0 Change Log Management and Crawls**

Optimizing Crawl Schedules

- **Full Crawls vs. Incremental Crawls**
  - Full crawls will be required periodically
- **Corpus Size**
- **Content Volatility**
- **Document Formats and Locations**
- **Index Freshness**
- **Segmenting Corpus for Crawl Optimization**

Lesson 3: Advanced Crawl Administration

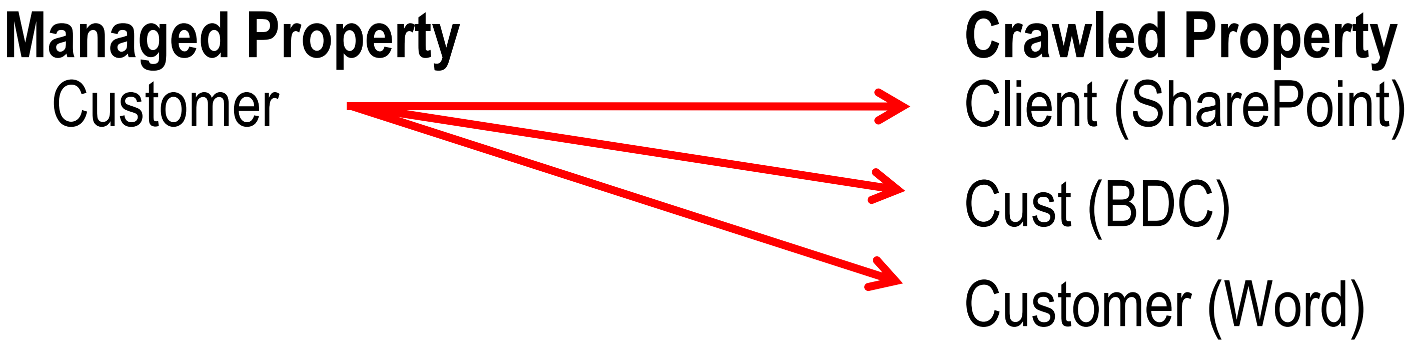
- Managing File Types and iFilters
- Implementing Managed Properties
- Implementing Server Name Mappings
- Configuring Content Access Accounts

Managing File Types and iFilters

- Protocol Handlers Load iFilters Based on Configuration Settings
- File Types and iFilter Mappings Managed at the Shared Service Provider Level
- Best Practice Is to Also Modify the DOCICON.XML file to Display Appropriate Icons in Search Results

  C:\Program Files\Common Files\Microsoft Shared\Web Server Extensions\12\TEMPLATE\XML\

Implementing Managed Properties



Managed Property

Customer

Crawled Property

Client (SharePoint)

Cust (BDC)

Customer (Word)

- **Managed Properties Used In:**
  - Scope definitions
  - Advanced Search Web part
  - Keyword query syntax
    - author:luis
  - Results display
  - Custom SQL queries

Implementing Server Name Mappings

- **Override How Search Results are Displayed**
- **Examples**

Hide file path:
file://moss/my_share mapped to
http://moss.litwareinc.com

Manipulated Forms-Based Authenticated URLs:
http://extranet mapped to
http://extranet.litwareinc.com

Configuring Content Access Accounts

- **Default Content Access Accounts**
- **Overridden Content Access Accounts**
- **Content Access Accounts and Versioning in Microsoft Office SharePoint Server 2007**

  Full Reader account recommended for most scenarios

Lesson 4: Configuring Query Processes

- Implementing Scopes
- Configuring Advanced Search Properties

Implementing Scopes

- **Scopes Are a Logical View on an Index**
  - Compiled and Efficient
- **Scopes Are Based On:**
  - Web address
  - Managed property
  - Content source
- **Scopes Are Defined by One or More Rules**
  - Include
  - Require
  - Exclude
- **Can Be Used Throughout the Search Experience**

Configuring Search Web Parts

- **Advanced Search Web Part**
  - Search Term Options
  - Managed Property Options
    - •Language filters
    - •Result types
    - •Property pickers
- **Result Web Parts**
  - Many properties, including stemming, views, duplicate collapsing

Lesson 5: Implementing People Search

- Indexing User Profiles
- Indexing Social Networks
- People Scope

Indexing User Profiles

- **What is People Search?**
- **People Search Based on Indexing User Profile Properties**
- **User Profiles can be Imported**
  - From Active Directory
  - From LDAP Directories
  - From Business Data Catalog Applications

Indexing Social Networks

- **What Is Social Distance?**
- **Why Is it a Useful Concept?**
- **How Is it Implemented?**

  My Site Colleague Tracker

  Outlook add-in for suggested colleagues

People Scopes and Results

- **People Search in Search Center**
  - Dedicated tab
  - Dedicated Web Parts
- **People Scope**
  - Based on a Managed Property Query
  - contentclass=SPSPeople

Lesson 6: Administering Farm-Level Settings

- Monitoring Enterprise Search Solutions
- Single Server Deployments
- Scaled-Out Database Server Deployments
- Scaled-Out Web Server Deployments
- Scaled-Out Query Server Deployments
- Consolidated Web and Query Servers
- Enterprise Search Indexing Performance
- Demonstration: Administering Farm-Level Settings

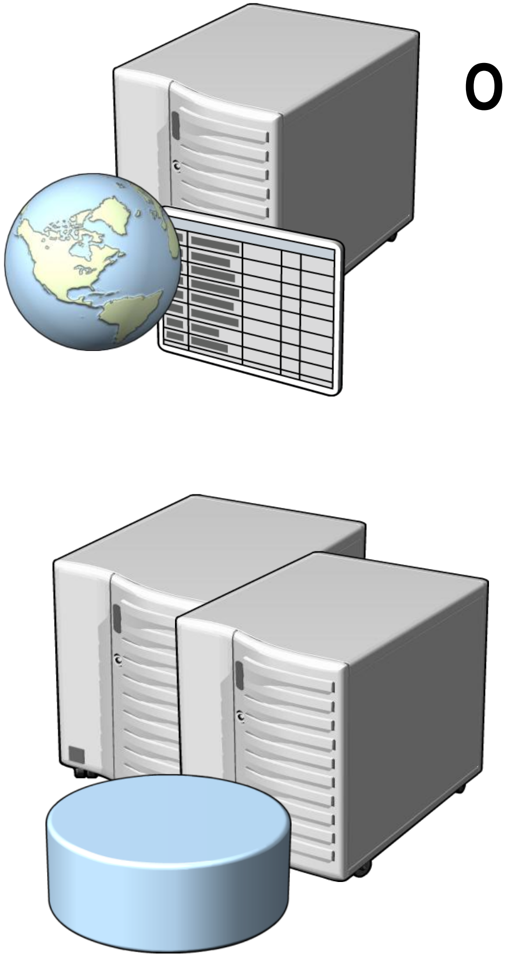
Monitoring Enterprise Search Solutions

- **Standard Counters**
  - Memory
  - Disk
  - Processor
  - Network
- **Search Counters**
  - Query Server
  - Index Server

Single Server Deployments

- Scaling Up Single Server Deployments
- Typical Usage

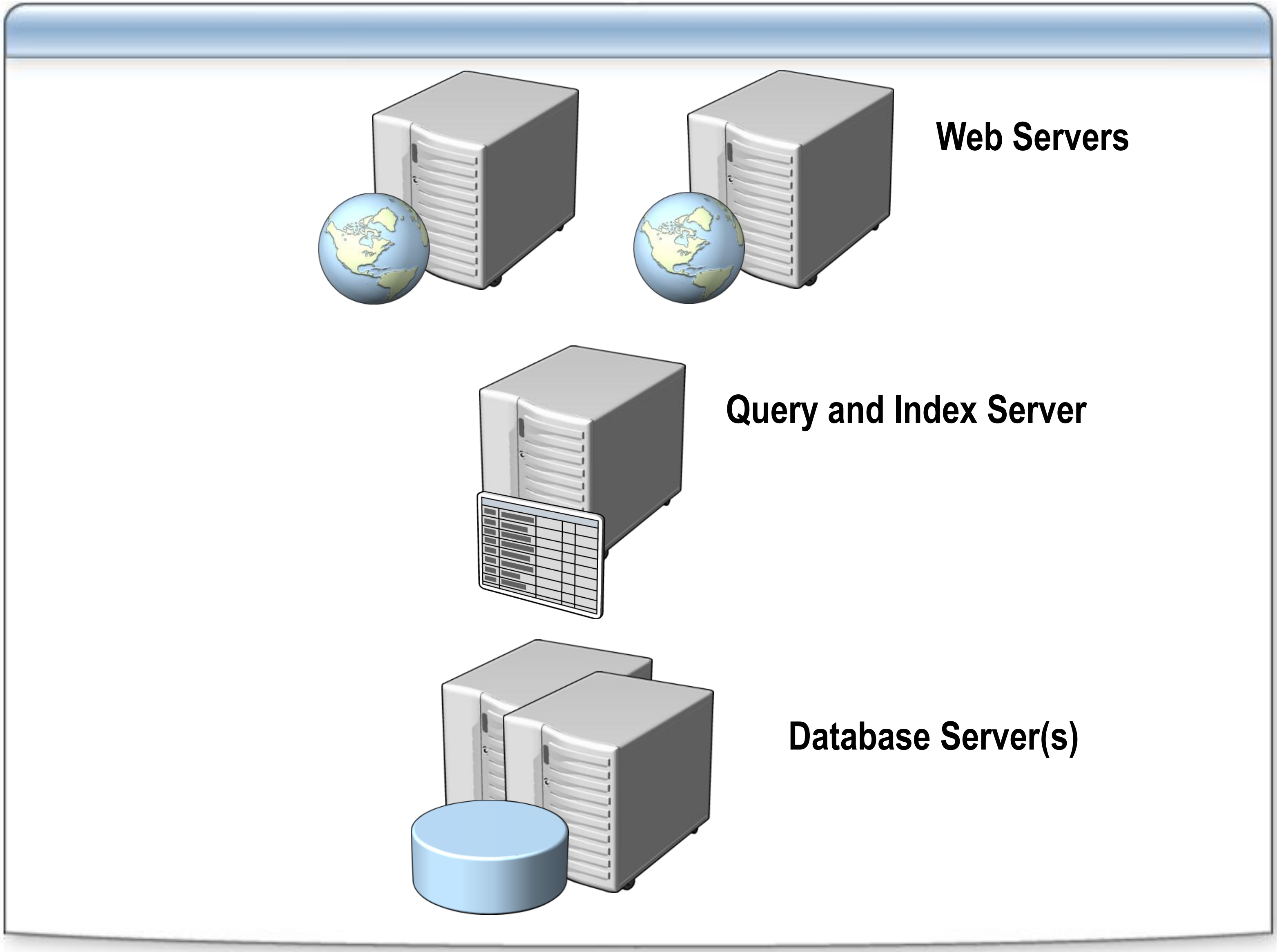
Scaled-Out Database Server Deployments



Office SharePoint Server 2007

Database Server(s)

Scaled-Out Web Server Deployments

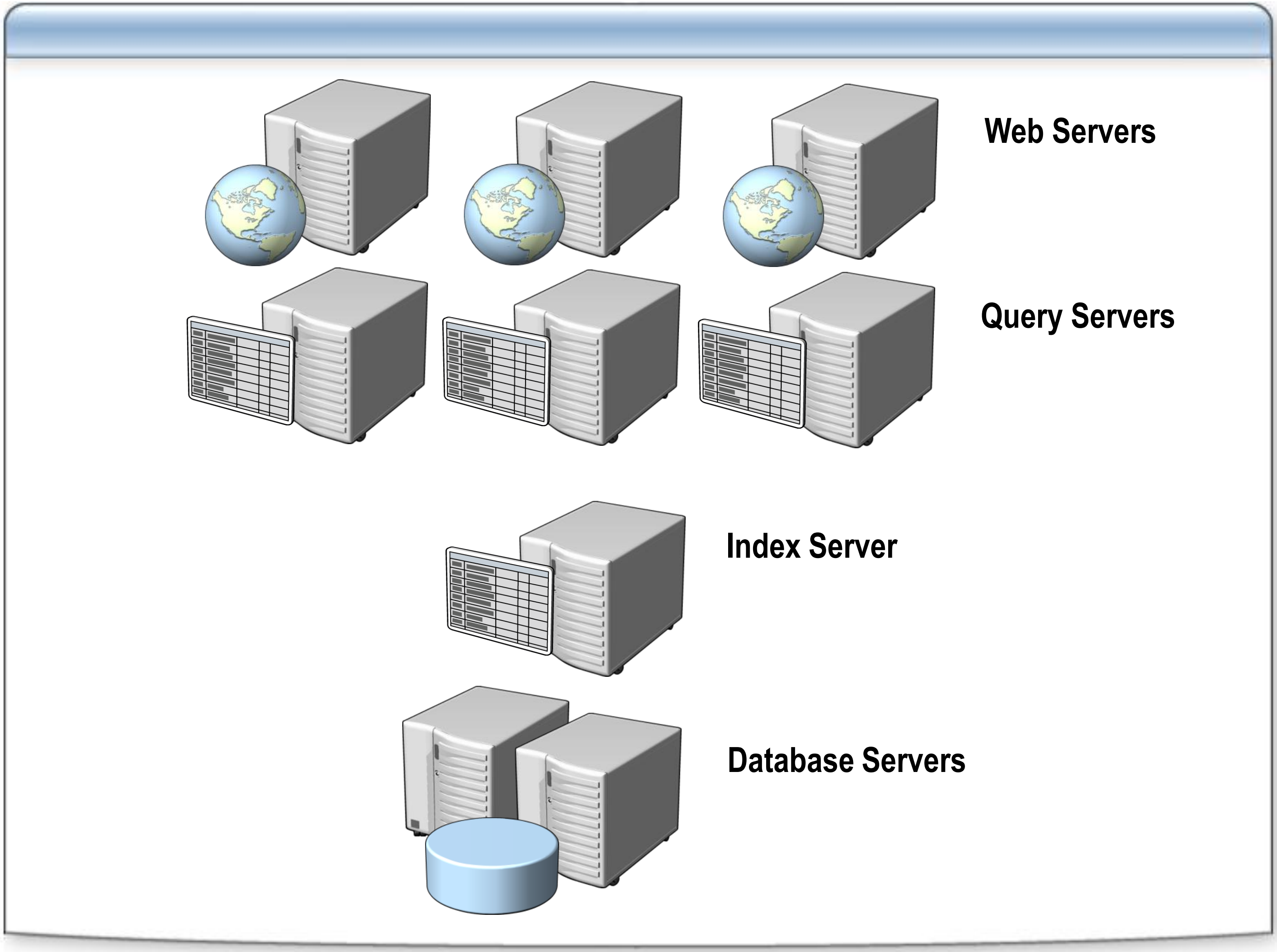


Web Servers

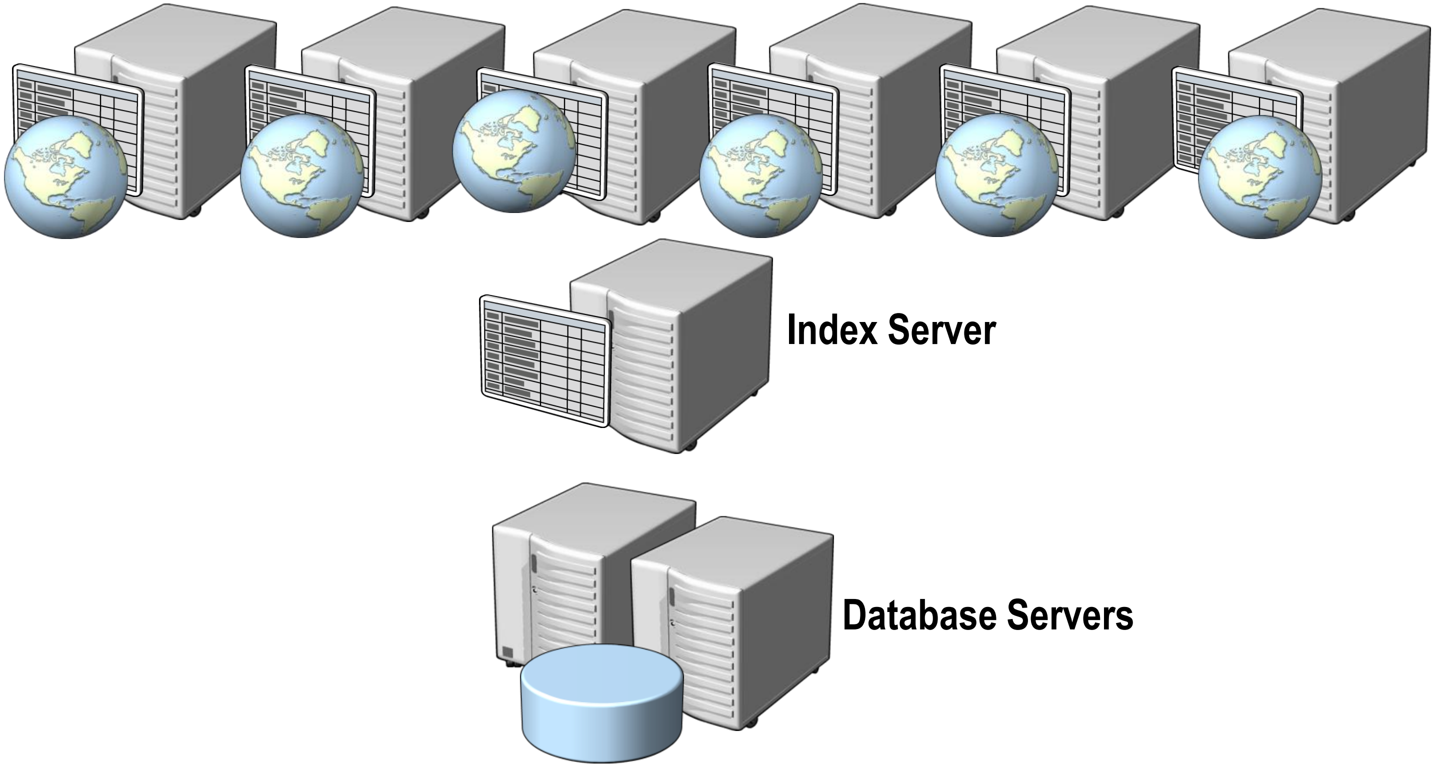
Query and Index Server

Database Server(s)

Scaled-Out Query Server Deployments



Collapsed Web and Query Servers



Web/Query Servers

Index Server

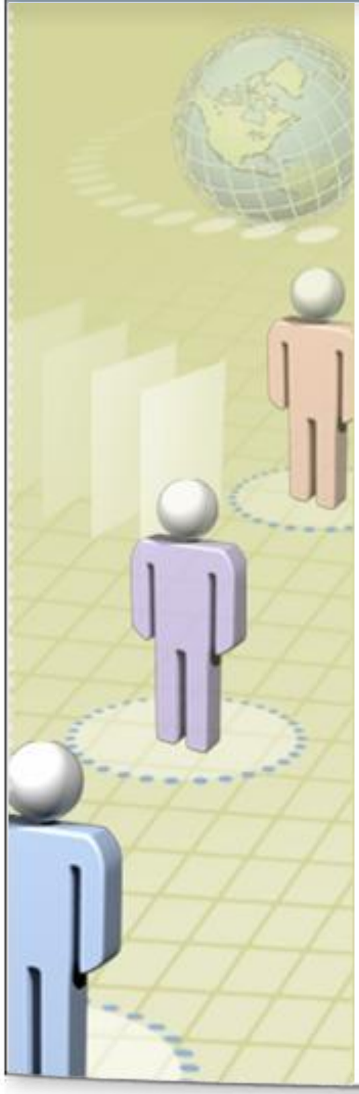
Database Servers

Compare and Contrast with Fully-Scaled Query Servers

Enterprise Search Indexing Performance

- **Crawler impact rules**
    - Parallel document indexing
    - Degree of parallelism
- **Database Load Options**
- **Dedicated Web Server Option**

Demonstration: Administering Farm-Level Settings



- Configuring Server Roles
- Tuning Indexing Performance
- Specifying Crawler Impact Rules