

# An Architecture for Scaling Federated Search

Presented by Andy Alsop, Vice President of Business Development  
 Deep Web Technologies, Inc.  
 ASIS&T 2009 Annual Conference



## Introduction

Federated search gives researchers access to a diversity of high-quality academic, technical, business and scientific databases. Much of the content these databases hold can only be found in the "Deep Web," a part of the Web that Google and other popular search engines can't reach. Federated search bypasses low-quality content sources and goes straight into the Deep Web, giving researchers important benefits:

- Searches that yield more relevant documents.
- More precise searches with better recall.
- More diverse searches that span sources in many scientific disciplines, leading to the cross-fertilization of ideas.

## Why Scale?

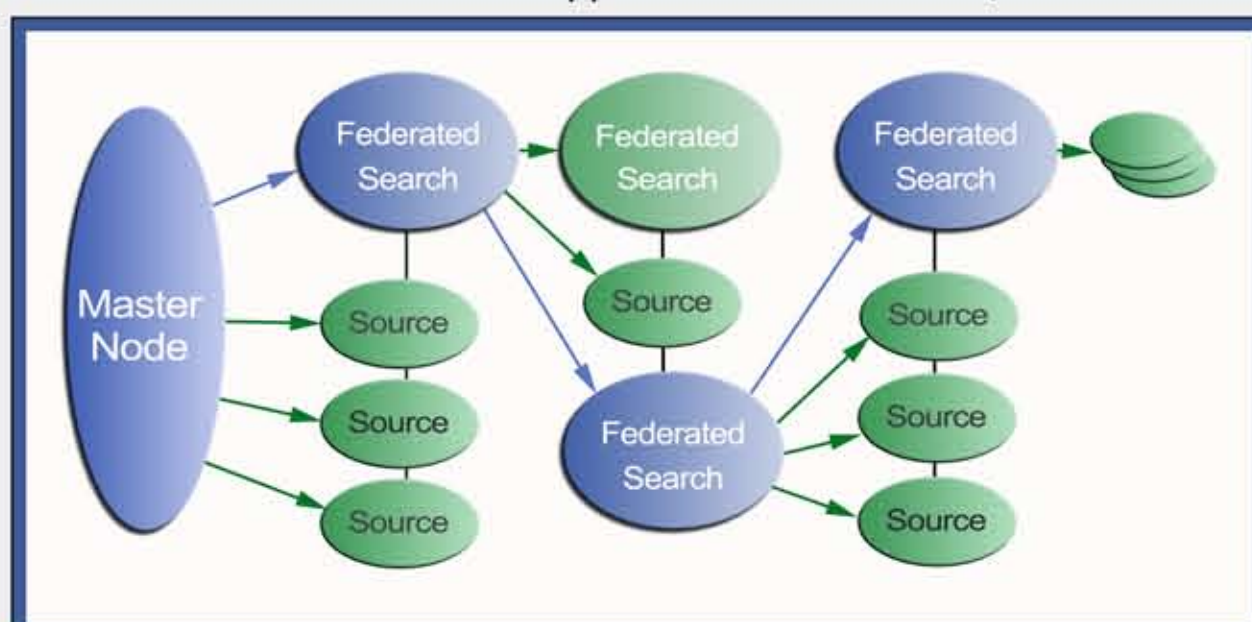
Federated search applications typically search around a dozen sources. A large traditional federated search application may search 40 or 50 sources. But building applications that search 1,000 or more sources simultaneously yields compelling new benefits:

- A much larger set of potentially relevant search results from which the application can choose.
- Accelerated cross-fertilization of ideas driven by large multi-disciplinary searches.
- One-stop research.

## Approaching Scalability

Our approach is to divide and conquer. We developed a way to hierarchically combine multiple federated search applications instead of just adding lots of sources. Our strategy uses these elements:

- A top-level federated search application is the single point where the user and the system interact.
- The top-level application searches other federated search applications and individual content sources.
- This combination of federated search applications goes on recursively: applications search more applications and sources, which in turn search other applications and sources, and so on.



## ScienceResearch.com

**"The world's science, all in one place."**

**ScienceResearch.com models scalable federated search**

**Searches 400 sources from one search box**

**Allows exploration of credible sources in a number of scientific disciplines**

**Just one of a number of state-of-the-art federated search applications built by Deep Web Technologies**

## The Architecture

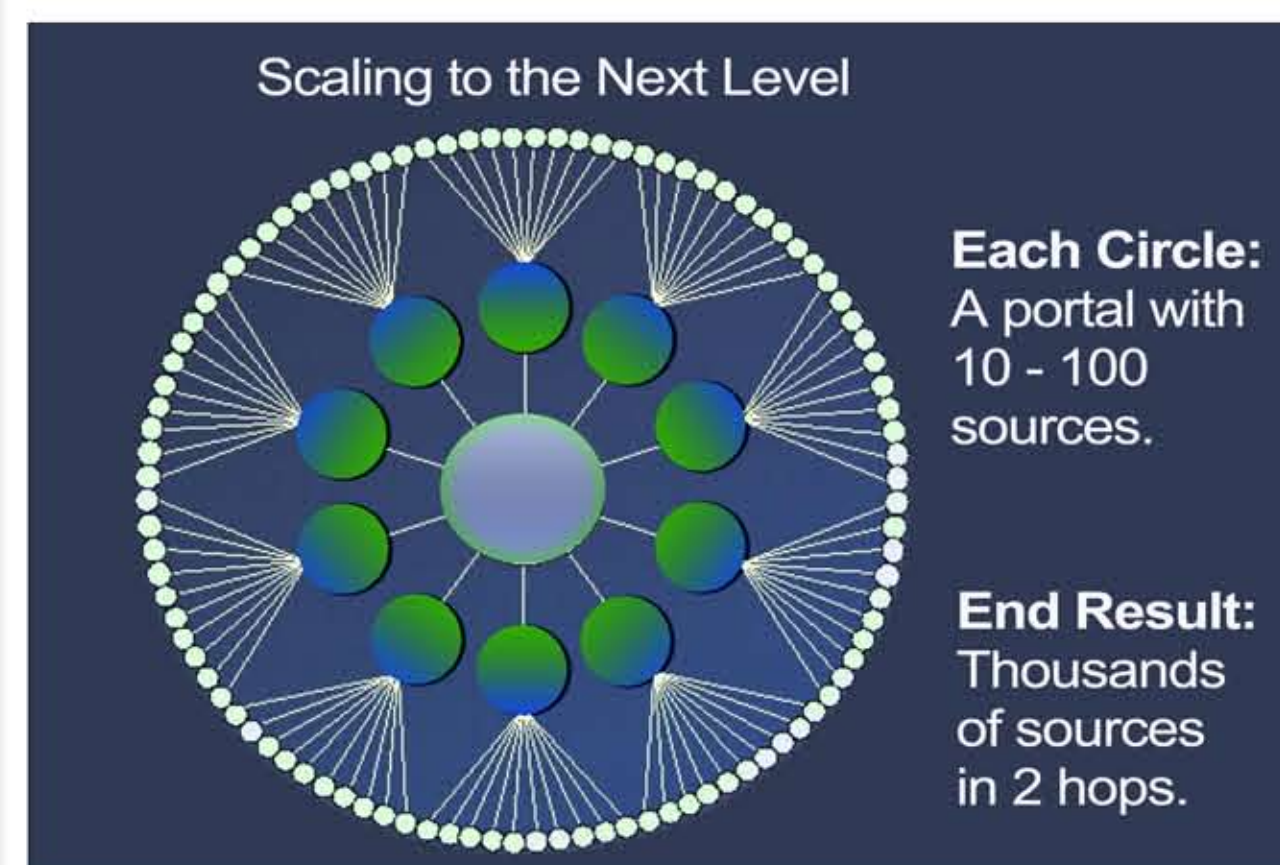
We created our hierarchical architecture by combining a number of elements into our divide and conquer approach:

- Distributed computing to spread the computation and network loads, i.e. to load balance. In particular, aggregation of search results from different sources and their relevance ranking lends itself to distributed computing.
- A mechanism for providing failover to redundant hardware components.
- A streamlined approach for creating, testing, monitoring, and updating thousands of sources.
- Automated source selection.
- Optional placement of distributed computing systems – compute nodes – geographically near the content providers to minimize network traffic.
- A mechanism to query and select a subset of sources from a federated search engine. This is required to eliminate duplicate sources across multiple engines.
- Development of interoperability standards for federating search engines from different vendors.

## The Reach

The number of sources that scalable federated search can access increases geometrically with the number of levels:

- One scalable federated search application, searching 10 applications which each in turn search 10 content sources, has a reach of 100 sources.
- Adding just one layer of 10 scalable federated search applications increases the reach to 1000 sources.
- Millions of search results can be analyzed by federated search engines, filtering the best ones to the user.



- The outer layers perform much of the processing so that the inner-most layer can provide the best results to users without being overwhelmed with itself having to rank and sort potentially millions of results.
- Linking together already existing federated search applications not only greatly magnifies the reach of scalable federated search but it does this at a very low incremental cost; the owners of the federated search applications are already monitoring and managing their connectors.

## Challenges

Scaling federated search is more complicated than merely adding a thousand new sources to an existing application. Beyond the management of vast computational, network and storage resources, other issues must be considered:

- Management of the creation, monitoring and repairing of many connectors within a single application.
- Careful selection of the right sources to minimize queries to sources that aren't likely to return relevant results.
- Seamless interoperability of federated searches by multiple applications.
- Identification and selection of a subset of sources to search – important when multiple applications search some of the same sources.
- Performance of asynchronous search in the hierarchical environment so that results can be streamed to higher-level search engines, allowing display of incremental results and improved handling of search engines that return no results.

## Next Steps

Deep Web Technologies continues its research and development efforts to advance the architecture and implementation of hierarchical and scalable federated search engines. To promote the hierarchical approach to other vendors and organizations we are pursuing a number of activities:

1. Building increasingly larger federations with more layers.
2. Working with other federated search vendors to develop and promote industry standards for creating hierarchical federations.
3. Encouraging content providers to make their search interfaces friendly to hierarchical federation.
4. Continuing research of automated source selection.
5. Continuing research of approaches to computer assisted source configuration and maintenance.

## Explore Further

Other Federated Search Engines using a scaled architecture:

[www.science.gov](http://www.science.gov)

Science.gov searches over 40 databases and 1,950 selected websites, offering 200 million pages of authoritative U.S. government science information, including research and development results.

[www.worldwidescience.org](http://www.worldwidescience.org)

a global science gateway—accelerating scientific discovery and progress through a multilateral partnership to enable federated searching of national and international scientific databases and portals.

