January 2008

IBM Global Name Recognition Version 3.1

Distributed Search Approach

David Biesenbach
IBM Global Name Recognition

Global Name Recognition - Distributed Search

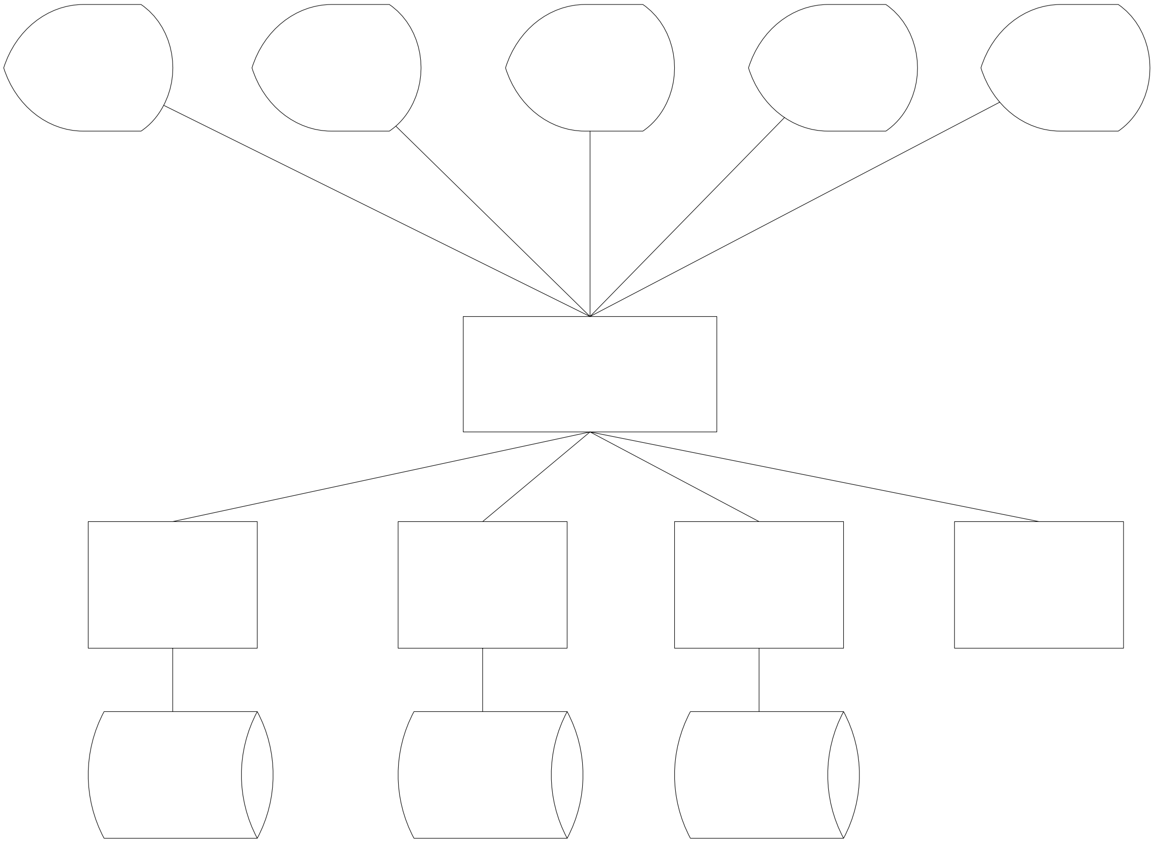
Introduction

Global Name Recognition - Distributed Search (GNR-DS) is an application that quickly and accurately searches for similar names in large lists of name data. Configured properly, a list of 200 million names can easily be searched in less than a second.

It is a server that accepts requests and provides responses via XML over tcp/ip socket connections. Unlike its predecessor, NHServer, it can be scaled to meet size and performance requirements. It is also backwards compatible with all NHServer client applications (e.g., NHSGUI), and operates on all GNR supported platforms.

Architecture

The drawing below shows a typical implementation of GNR-DS. It is the optimal configuration for a four CPU machine, or a two CPU machine where the CPUs are dual core. Processes can also be distributed across separate machines on the same network. The machines must use the same chip and operating system (e.g., AIX/PowerPC, Linux/Intel).



Clients can be any application that conforms to the GNR-DS XML message format. A custom client can be developed easily using GNR NameWorks which provides a Web Services interface with GNR-DS.

The communication manager accepts requests from clients, routes them to the searchers and combines responses before sending them back to the requesting client. It is the single point of access to GNR-DS for client applications.

Each searcher manages an equal portion of the entire name list. So, if you have a name list of 30 million names and three searchers, each searcher would manage 10 million name records. Every searcher should have a devoted CPU/core.

Note the right most searcher marked as “Searcher Adds”. While not required, GNR-DS works best if one Search is devoted to handling additions that occur after system start-up. Note that the “Add Searcher” does not require its own CPU.

GNR-DS can also be scaled small. That is, a simple configuration with the communications manager and one searcher which also supports adds will work fine on a single CPU machine.

Unique Name Mode

With a large number of names (tens to hundreds of millions) there will be many duplicates. Common names (e.g., Jose Gomez) can appear thousands of times. Not only is it wasteful to store and search the duplicates, returning hundreds of results with the same name will overwhelm users and quite possibly cause valuable matches to go unnoticed.

GNR-DS provides a unique name mode to overcome these problems where only de-duplicated names are loaded into searcher memory, and the original data is stored on disk. By default, only the unique names are returned, and the user is given the option to perform an additional query to see the original name data.

To support the de-duplication and splitting of name data, GNR-DS provides the name pre-processor utility (NPP). It will read a comma delimited text file and transliterate, regularize, parse, classify and de-duplicate the names and produce subsets of the data ready to be loaded into GNR-DS searchers. The actions taken are all configurable; for example, the generation of alternate parses can be disabled. NPP can also be used to support the processing of data for non-unique modes of operation.

Other Features

In addition to the ability to scale, GNR-DS provides several other valuable new features:

- Even without scaling it is at least twice as fast as previous versions of GNR.
- It provides a new query message that combines search parameters with the query name information. Previously, this required two messages.

- It provides a batch search utility (dsFile) which will read in a file of transactions and write the results to another file.