Database Extractor

Version 3.0

Extracting data from a database or other content repository that it can be processed by a search engine or an XML based application may sound simple at first. When one starts getting into the details of how to deal with vast amounts of data quickly, documents stored in fields with meta-data in related fields, relationships spanning several tables, deletions and changes in the database, and many other hurdles, the job starts to sound like something that a company is likely to pay a Database Administrator a great deal to accomplish and even more every time something changes. Hence 30 Digits has developed the Database Extractor.

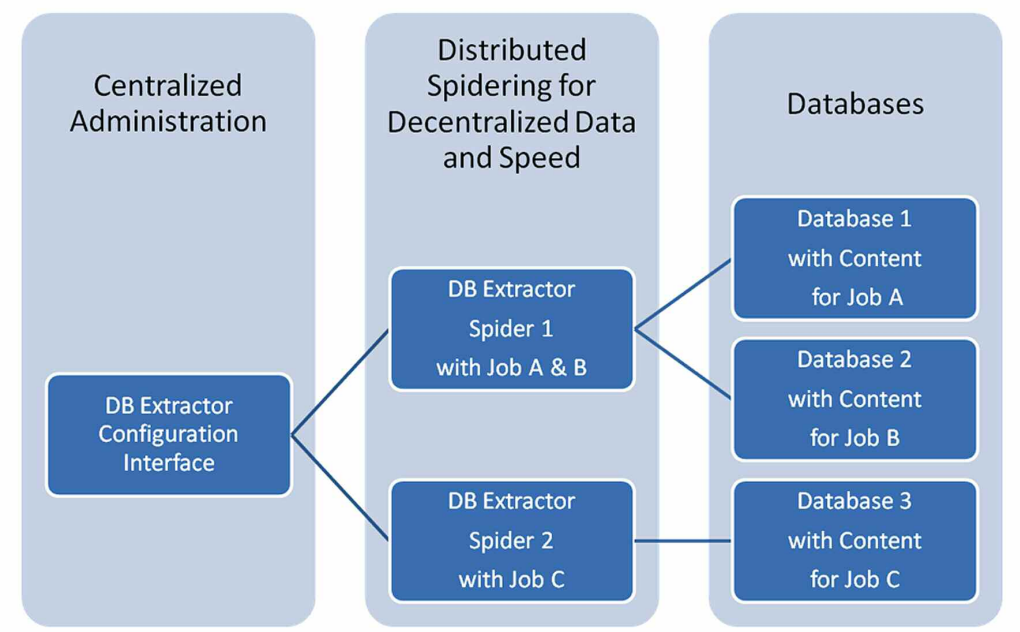
The Database Extractor is designed to crawl any database type with little to no knowledge of the tables, fields, and architecture. It has a simple user interface that is easy to use for non-technical users that need to extract the data but do not have the time to spend learning a database system and may not have the assistance of a database administrator. It also retrieves binary files like Microsoft documents, PDFs and images. The context of the data in its relationship to other data in related fields and tables can be maintained as the Database Extractor easily handles 1 to 1, 1 to many, and many to many relationships. Once all the data is extracted from the database, it is then fed to the desired search engine or other application in XML or another defined text format.

Supported Databases

The Database extractor uses JDBC drivers to connect to the various database management systems. Thus, any database that already has a JDBC driver (which is nearly all) can be crawled by the Database Extractor. Here is a list of some of the most common: Oracle, MS-SQL, MySQL, Postgres, ODBC, MS Access, Ingres, Informix, IBM AS/400 and DB2. See the following page supplied by Sun for a more comprehensive list see http://developers.sun.com/product/jdbc/drivers.

Architecture

The architecture has been elegantly designed that it has a single point of administration while allowing multiple instances of the spider component to be run on several different machines. This allows scaling for large implementations even across multiple geographic locations while maintaining accuracy and ease of administration.



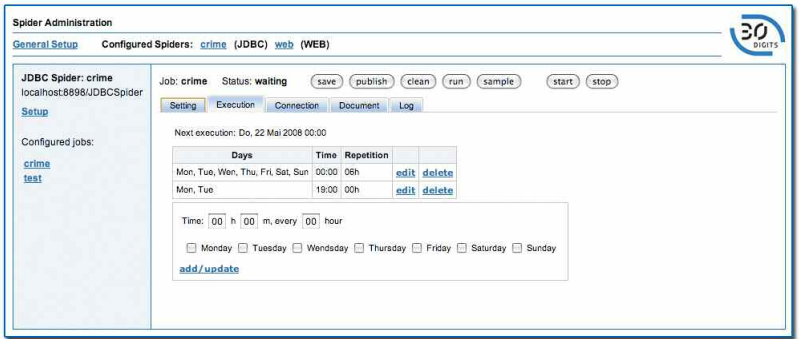
Setup and Administration

An often overlooked aspect of such a system is the effort involved in setting it up and administering it. The development team at 30 Digits have seen how complex and manual most of the solutions on the market are and have thus taken great effort to create a method for extracting content that is easy and intuitive while remaining flexible and powerful. Some of the main ways this has been achieved are by providing the following:

- Graphical User Interface (GUI) available via web browser for all of the configuration and administration options instead of textual configuration files which are often nearly impossible to decipher and lack documentation of how to setup and no central way to administrate them.
- The system uses AJAX extensively to load lists in the background to help fill out information.
- Simple field property modification like changing the title name of a field or even more complex operations like switching field variables from names to numbers for search engine optimization.

- Advanced features are available when an administrator wants to go deeper but not necessary for beginners or basic setup.

- Regular polling times for checking new data can be set very granularly and flexibly (i.e. every Monday, Wednesday, and Friday at 0, 6, 12, and 18 hours).

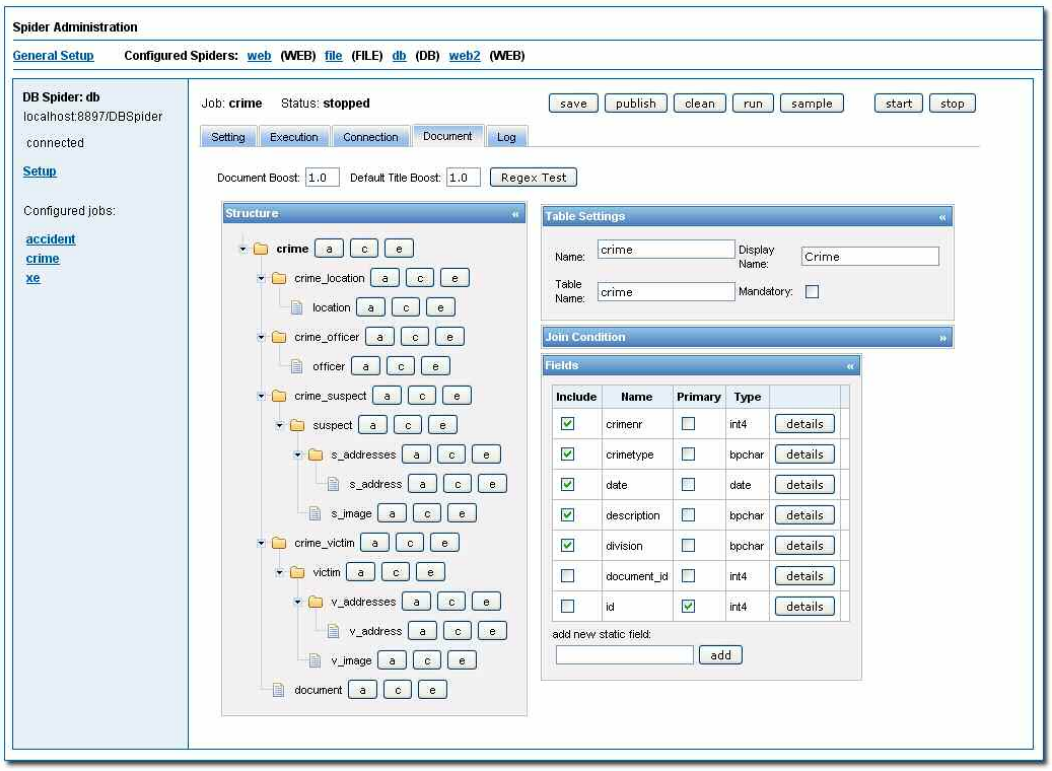


- When information is changed or deleted, the system can be set to either replace or delete the information in standard mode or keep time stamped versions of the document in the versioning mode.

- Load of spider on the sources can also be configured for faster spidering or less intensity on the server

- The Database Spider allows the building of structured text documents (typically XML) through an easy to use interface by loading table information dynamically. It replaces the need for a DBA writing highly complex SQL queries and DB Views.

- Sampling of spider output before jobs are run. This can save tremendous amounts of time as different configurations can be tried before committing jobs that could have long run times. It also allows the capturing of errors at the beginning instead of waiting to analyze a log file after hours of spidering.



Blobs and Clobs (Binary Data & large text files)

Databases are often used to store large amounts of text in a field of variable size (CLOBs) and binary files (BLOBs) like MS Word documents and PDFs. With this information is often valuable meta-data. The Database Extractor has the ability to extract both CLOB and BLOB fields and associate their meta-date with them even across tables. The Extractor is even able to extract meta-data from the binary files themselves. It handles a large assortment of file types. Here is a list of some of the major ones: Microsoft Word, Excel, PowerPoint, PDF, HTML, TXT, RTF, MSG, XML, and ZIP. New file types are constantly being added to the import process to assure all kinds of files can be processed.

Contact

For more information or to schedule a demo, contact us at one of the following:

Tel: +49 89 45 23 89 66

Fax: +49 89 45 23 89 70

contact@30digits.com

30 Digits GmbH

Agnes-Pockels-Bogen 1

80992 München

Germany