Title: Using open source technology to extract biochemical data from big data repositories: the NCBI PubChem experience.

Authors: Lewis Y. Geer, Lianyi Han, Asta Gindulyte, Bo Yu, Paul A. Thiessen, Evan E. Bolton, Yanli Wang, Stephen H. Bryant.

Affiliation: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894.

Presenting author's email: lewis.geer@nih.gov

Project URL: http://pubchem.ncbi.nlm.nih.gov/search

Code: http://lucene.apache.org/solr/

Abstract: A rapid increase in the amount of publicly available chemical and biological information has revolutionized biomedicine but has posed challenges for traditional query engines and methods due to size and complexity. This issue is particularly acute in databases that serve as repositories for large scale projects, such as NCBI's PubChem, which contains over 200 million bioassay readouts and nearly 50 million small molecule structures. PubChem is an NIH funded public database containing information on the biological activities of small molecules and used by tens of thousands of unique users per day. Nontraditional query methods, such as NoSQL, promise to address scalability and access to tractable subsets of data while opening opportunities for discovery and analysis across disparate data sources. In particular, we have used and contributed to the Apache Solr search engine in order to create the PubChem Search service: http://pubchem.ncbi.nlm.nih.gov/search, which has a variety of search types, such as text and chemical similarity, that can be executed across collections of small molecules, bioassay readouts, protein sequences, and patents. This search service is able to execute subsecond searches while joining records across collections with many millions of records. For example, searching the chemical compound collection by structure retrieves chemically similar molecules from a collection of 50 million chemicals as the user draws the chemical structure, giving immediate feedback. This chemical search can be done in any of the collections joined to the chemical compound collection, such as searching patents or bioassay readouts by chemical structure. This powerful cross database searching allows for specification of multiple queries to winnow down large scale data to information of interest. In this presentation, we describe the architecture of this search engine along with novel features and interfaces. The overall SaaS architecture consists of information retrieval algorithms and high speed joins that are coordinated by a query optimizer exposed via a RESTful interface accessed by a thin javascript client running in a browser. The architecture is intended to be generalizable to most types of standard biomedical data. Features developed during the creation of this service, such as the high speed joins, were contributed back to the Apache Solr project.