Chris Toombs

CS410 Fall 2020

Technology Review

Apache Lucene: Open-Source Search Engine Library

In the rapidly advancing space of search, companies and users are constantly looking for more advanced methods for indexing and retrieval of search topics and documents, both in the world of search engines as well as application development. Apache Lucene is an open-source search engine software library developed specifically to meet these needs (Apache Lucene). While Lucene is purely the Java implementation of this library, several additional products have been created to extend Lucene's core functionality such as Apache Solr and Elasticsearch (Apache Lucene). This technology review will provide the reader with a brief overview of the capabilities of Apache Lucene and will continue into a brief overview of common extensions of Lucene: Apache Solr and Elasticsearch. Some industry use cases will be provided to give the reader a general sense of how this product is used in a real-world setting.

Created by Doug Cutting in 1999, and named after his wife's middle name, Lucene joined the Apache project in 2005 (Apache Lucene). Lucene is a library which enables text-based search of documents given a user query. The main functionality of Lucene provides functionality for an in-memory index, parsing of queries, and ranking of documents based off a given query. The in-memory index of Lucene is full text that can be sourced from a variety of sources, such as databases, file stores (E3, S3, Linux, etc.) or websites. Lucene is a highly performant library as it will search the in-memory index versus searching text in documents directly. Indexing in Lucene is performed by using an IndexWriter (to add documents to the index) and using an IndexSearcher (to retrieve documents from the index). Documents in Lucene are defined as a unit of text and index. Each Document is comprised of one or more fields, which are essentially just a key: value pair. For example, as Field can be named "UserID" with the value of the Key "UserID" being "1234". The indexing portion of Lucene involves around adding Documents into the index (Lucene Tutorial).

Lucene allows for fast indexing, with capabilities of over 150GB worth of indexing hour on modern hardware. Despite this throughput, only a small RAM cache is needed with Lucene utilizing a 1MB Heap. Incremental and Batch indexing is similar in performance and a typical index only takes up around 20-30% of the size of the text indexed (Lucene Features). Lucene provides ranked searching, multiple query types (including wildcard), sorting by field, faceted search and error-proof search suggestions - it is highly scalable and performant (Lucene Features).

In a world moving to distributed systems, it was important to extend the capabilities of Lucene. This extension allowed for the rise of Apache Solr: a search platform based off Apache Lucene (Apache Solr). Apache Solr allows (and extends) the many features of Lucene but in a distributed fashion. A large amount of RESTful JSON APIs have been developed for Solr to allow for integration with several

programming languages (Apache Solr). Solr is widely used in industry to enable enterprise search and analytics because of its ability to distribute load of the index across two or more Solr Servers. This sharding (partitioning) of the index will allow users of Solr to experience fast query retrieval despite the size of the underlying index. In a simple sense, Solr is more suited to enterprise applications where big data tools have already been implemented (such as Apache Hadoop and Apache Spark) on large amounts of static text.

Like Apache Solr, ElasticSearch is another framework which also extends the Lucence framework by providing distributed performance on text-based search. It is currently the most popular implementation of indexed based search, although it is a slightly heavier install than Apache Solr. ElasticSearch requires 1GB of HEAP memory while Solr only requires 512MB. In terms of install, ElasticSearch is ~300MB install while Solr is ~190MB ("Solr vs ElasticSearch"). ElasticSearch is completely JSON based as well and still provides a large number of RESTful APIs to users for integration within apps. Within a use case perspective, ElasticSearch is more suited to a web-based application where information is brought in and out in a JSON format ("Solr vs ElasticSearch"). In recent times, ElasticSearch developers have put more time into making the tool more resilient, and many teams are now beginning to use ElasticSearch as a primary data store for applications. ElasticSearch has a large amount of documentation on the internet and is complimented by a large number of books and guides on the internet. Currently, Elastic Co. (maker of ElasticSearch) comments that they currently have over 3000 corporate adopters of the Elastic Stack (Elastic Co). Some large companies who utilize ElasticSearch within their companies are Uber, Walmart and HEB (Elastic Co).

Given the quick adoption of the Apache Lucene product, along with some of its frameworks (such as Solr or Elastic search) we can expect to see the continued improvement of in-memory distributed text search capabilities. With much of the world moving their applications to a cloud / distributed setup, these tools can expect to see continued growth in the future. ElasticSearch will continue to rise in popularity given the amount of community support and documentation available to users. With the large amounts of streaming and user generated data (think social media) in the world today, it will be interesting to see how these in-memory index solutions improve in terms of a search perspective. Small improvements in performance can potentially save the world exponential amounts of time and energy given the volumes of data the world is dealing with. In conclusion, Lucene (along with it's extensions of Apache Solr and ElasticSearch) will continued to see increased adoption amongst companies and users who are in need to fast text retrieval and analytic capabilities.

Works Cited

Apache Lucene (2020, Nov 03). Retrieved November 03, 2020, from https://en.wikipedia.org/wiki/Apache_Lucene

Lucene Tutorial (2020, Nov 03). Retrieved November 03, 2020, from https://www.lucenetutorial.com/basic-concepts.html

Lucene Features (2020, Nov 03). Retrieved November 03, 2020, from https://lucene.apache.org/core/features.html

Apache Solr (2020, Nov 03). Retrieved November 03, 2020, from https://lucene.apache.org/solr/

Solr vs. Elasticsearch: Who's The Leading Open Source Search Engine? (2020, August 07). Retrieved November 03, 2020, from https://logz.io/blog/solr-vs-elasticsearch/

Elastic Co. (2020, Nov 03). Retrieved November 03, 2020, from https://www.elastic.co/customers/