Apache Lucene

CS410-Tech-Review

FALL-2021

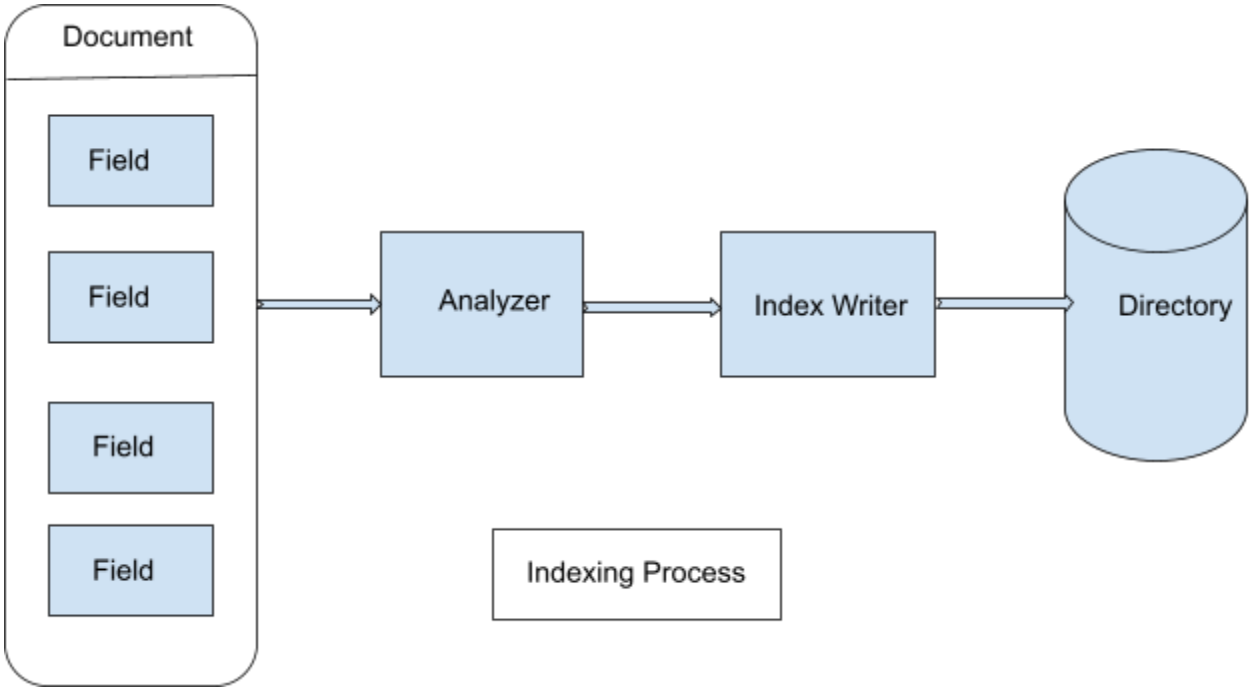
Uttam Roy

Apache Lucene is a powerful open-source, high-performance, scalable information retrieval (IR) JAVA library, written by Doug Cutting in 1999. Currently Lucene is maintained by Apache Software Foundation. Lucene has interfaces to use in other languages like C/C++, python etc.

We do web searches everyday multiple times. The database can search in a database, but can not search across multiple websites. A full-text search engine can query in multiple files over multiple web sites very quickly. With the increasing volume of data over the web we need efficient and fast search engines. Apache Lucene adds document search capability to software applications in an efficient and simple way.

Lucele uses following operations and has classes to use for these:

1) Build the document: build the documents from the raw content,
2) Analyze the document: Lucene analyze the document to find suitable parts for indexing
3) Indexing the document: this process creates the index of the document so that the document can be retrieved based on a few keys without using the entire document.
4) User Interface for Search: Lucene provides user interface so that use can enter text and search.
5) Build Query: Using the entered text, Lucene builds a query object to inquire built-in index databases to get details.
6) Search Query: Using the above query object, Lucene checks the index database and gets the relevant content documents.



Above diagram explains the indexing process. IndexWriter is the major component of the indexing process. Lucene makes use of Inverted Indices to make the indexing process fast.

Lucene has following basic terms:

1. document: a record which represents the source of data, the unit of search, the document is also returned as search results. Documents are collections of fields.
2. field: a field is a pair of items (name,value). When a field is inversely indexed, it is available for search.
3. index: a collection of documents, typically with the same schema
4. corpus: the entire set of documents in an index
5. inverted index: internal data structure that maps terms to documents by ID
6. term: value extracted from source document, a word from the text, smallest unit on which search is run, used for building the inverted index.
7. vocabulary: the full set of distinct terms in a corpus
8. uninverted index: array of all field values per field
9. doc values: alternative way of storing the uninverted index on-disk
10. Directories: Lucene index stores data in file system, to get better performance it can store in memory.
11. Analyzers: Analyzer uses tokenizer to make tokens. It also does stemming, filtering stop words, and normalizes by removing accents and other character markings.

Lucene has its own query language for searches. Lucene has a dynamic and easy to write query syntax. It allows users to use specific fields for search, boosting, the ability to use boolean queries, and other functionalities.

Lucene is a good tool for full-text search, but it has some limitations like performance, machine learning, time to value etc. For any changes it creates a new record without deleting the previous record, so over time, memory gets overloaded and gives slower performance. Machine learning in search engines helps to improve relevancy and results. Lucene relies on index matching and does not use ML, it costs more for business to get better results. Lucene is free to use, and implementing search applications is quick, but optimization to get faster speed and better results for big projects costs more. Alternatives to Lucene are Sajari, ElasticSearch, Solr, Algolia etc.

Reference:

Introduction to Apache Lucene, https://www.baeldung.com/lucene

Lucene in Action, Second Edition : Michael McCandless, Erik Hatcher, Otis Gospodnetic

Lucene 4 Cookbook : Edwood Ng, Vineeth Mohan

Basic Concepts, http://www.lucenetutorial.com/basic-concepts.html

Lucene Tutorial, https://www.tutorialspoint.com/lucene/index.htm

Apache Lucene, https://en.wikipedia.org/wiki/Apache_Lucene

Apache Lucene, https://lucene.apache.org/

Introduction to Lucene, Shubham Aggarwal, https://linuxhint.com/introduction-to-lucene/

Best Alternatives to Lucene,
https://www.sajari.com/alternative-to/best-lucene-alternatives

Lucene: The Good Parts, Andrew Montalenti, https://blog.parse.ly/lucene/