Searching and Indexing Big Data

By Dinesh Amatya

Lucene

- Lucene is a high performance, scalable Information Retrieval (IR) library.
- lets you add indexing and searching capabilities to your application
- can index and make searchable any data that can be converted to a textual format
- mature, free, open-source project implemented in Java

Lucene

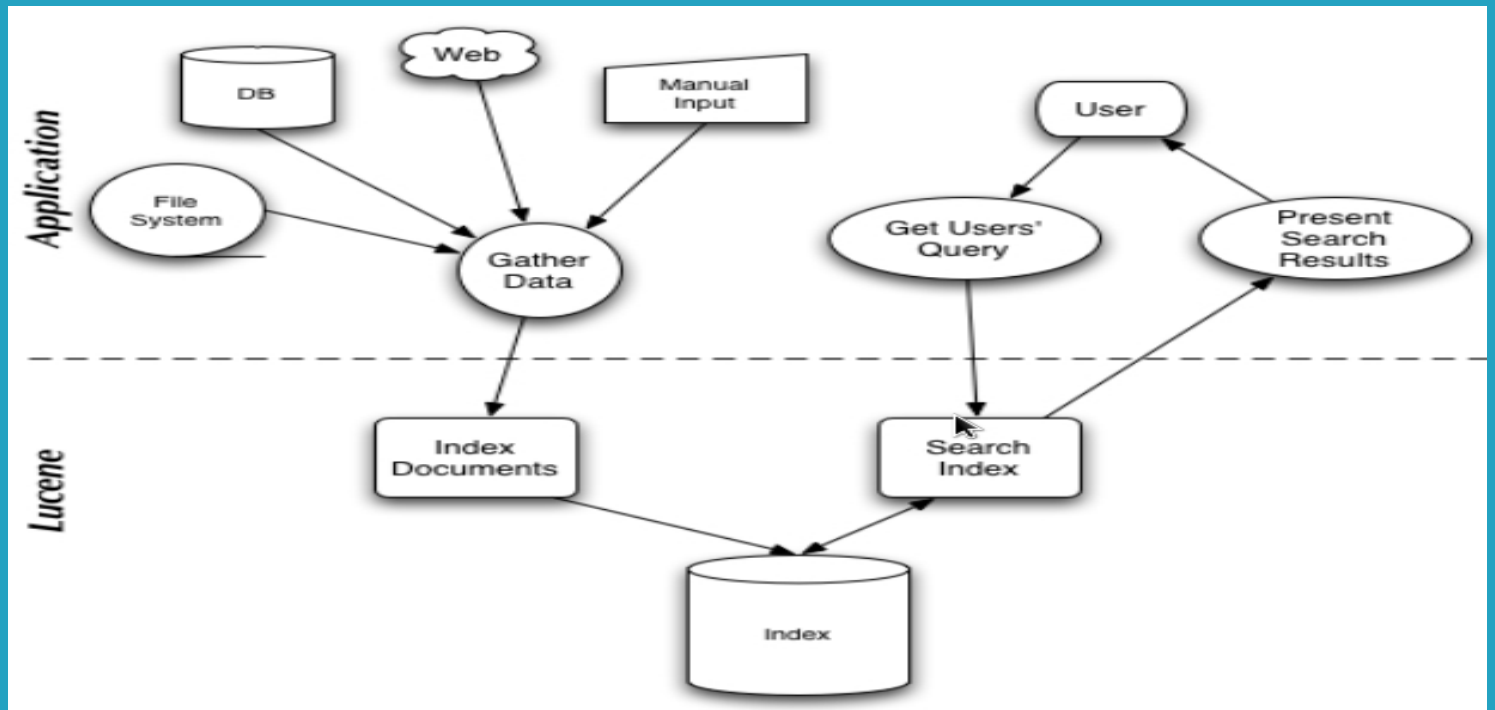
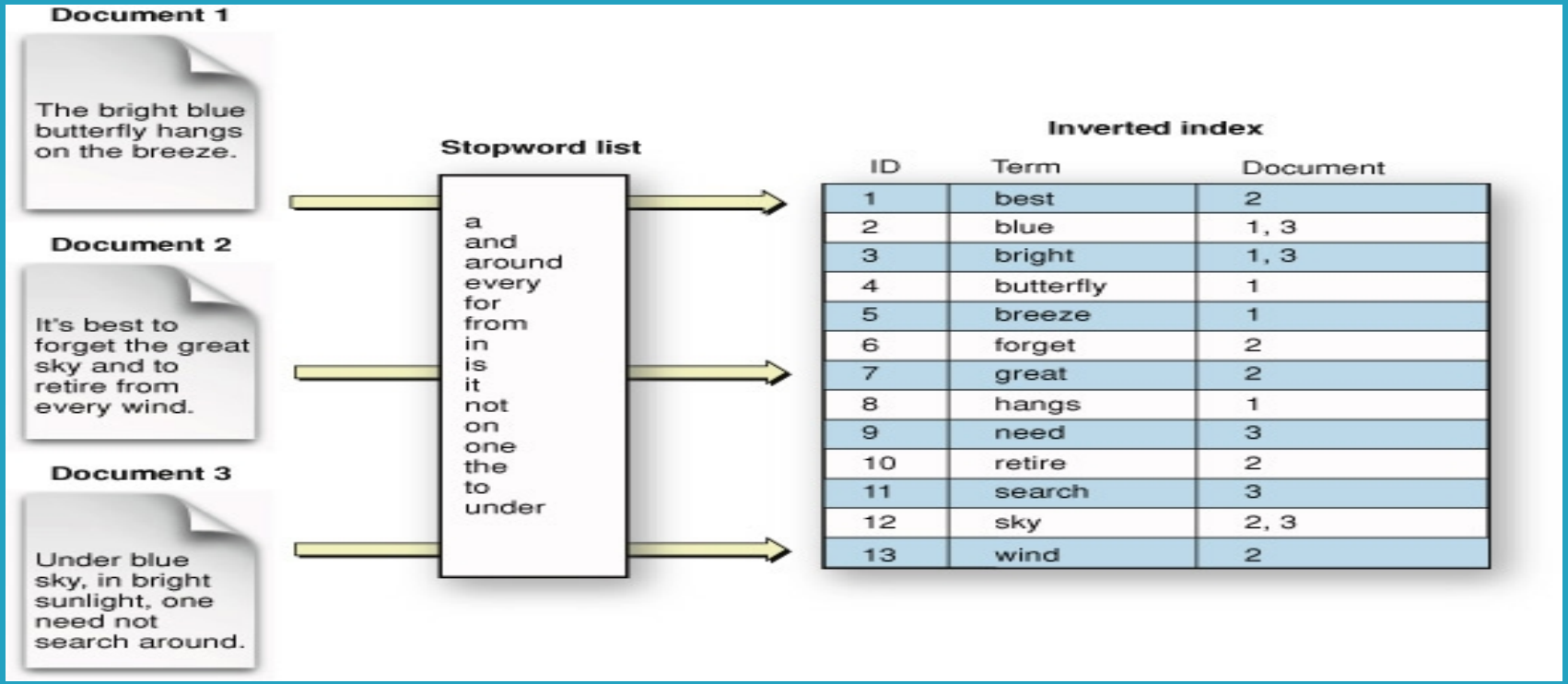


Figure 1.5 A typical application integration with Lucene

Basic Concepts : Indexing

- To search large amounts of text quickly, one must first index that text and convert it into a format that will let one search it rapidly, eliminating the slow sequential scanning process. This conversion process is called indexing, and its output is called an index.

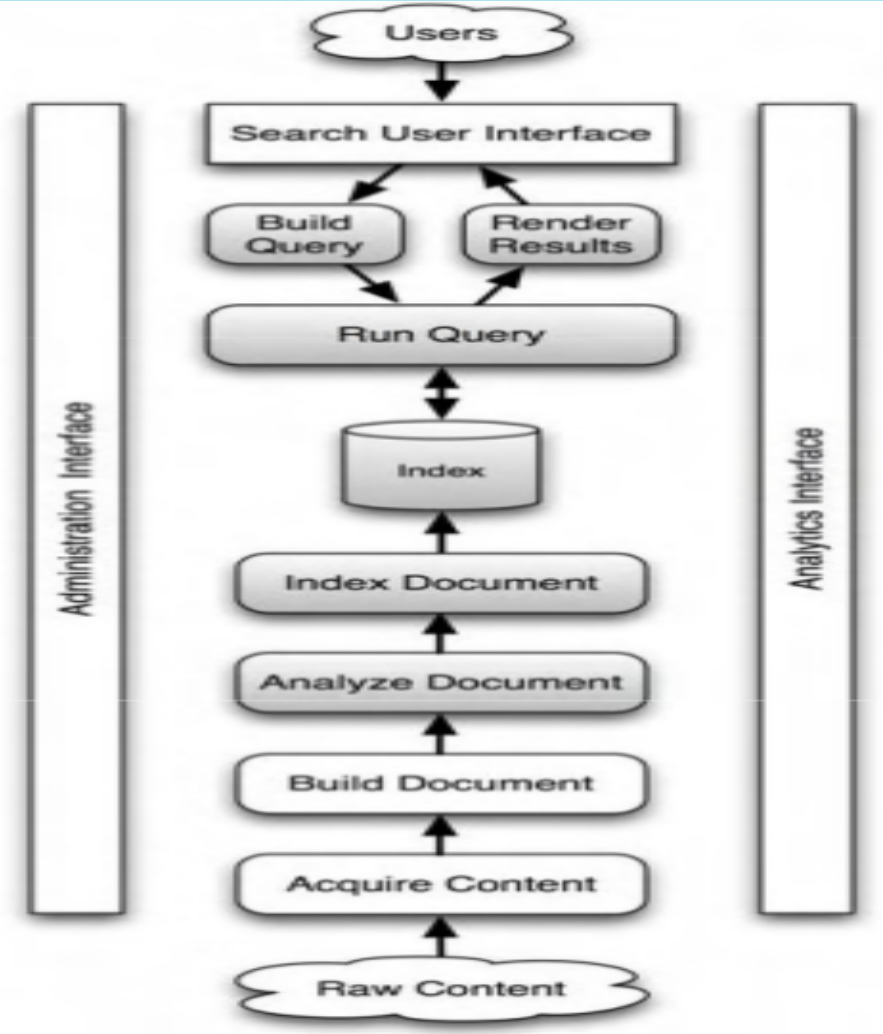
Basic Concept : Inverted Index



Basic Concept: Searching

- Searching is the process of looking up words in an index to find documents where they appear
- Quality of search described by
  - Recall
  - Precision
- Searches index instead of text

Typical Components of Search Application



Core Indexing Classes

IndexWriter
Document
Analyzer
Field
Directory

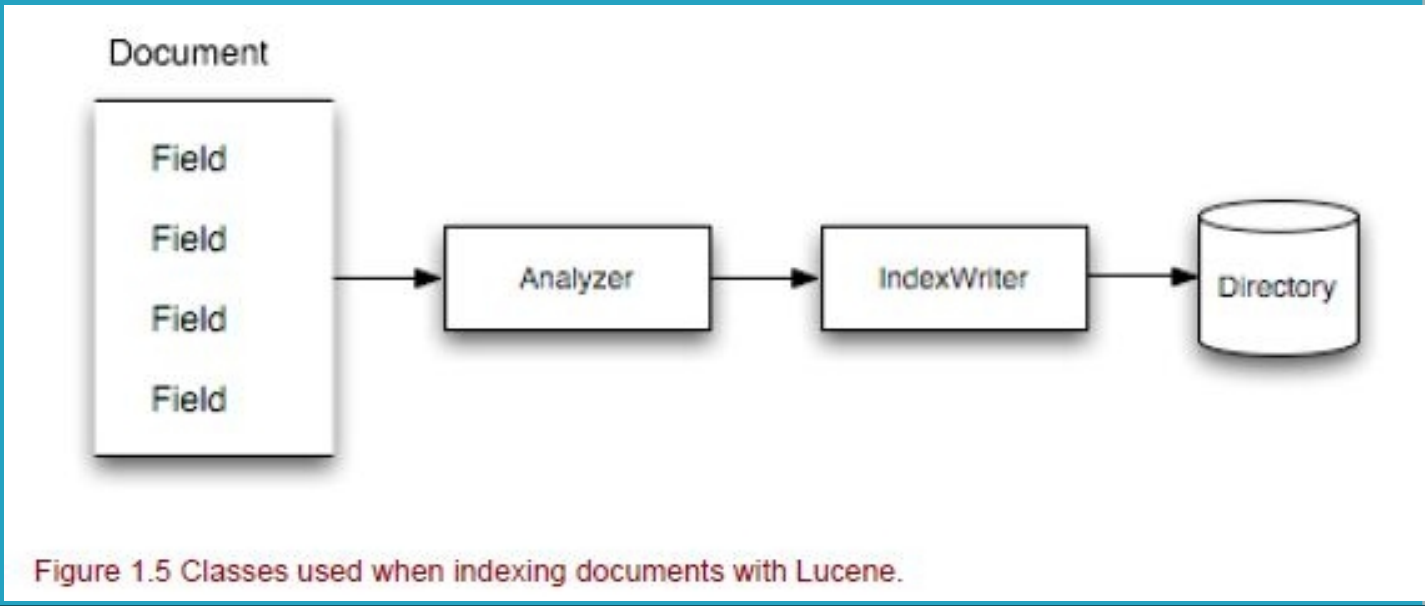


Figure 1.5 Classes used when indexing documents with Lucene.

Primary Analyzers available in Lucene

WhitespaceAnalyzer

SimpleAnalyzer

StopAnalyzer

KeywordAnalyzer

StanderdAnalyzer

Core Searching Classes

IndexSearcher

Term

Query

TermQuery

TopDocs

References

- https://en.wikipedia.org/wiki/Full_text_search
- Lucene in Action
- http://www.javabeat.net/using-the-built-in-analyzers-in-lucene/