Teja Pitla (tpitla2)
CS410
UIUC
Department of Computer Science
November 2022

Applications of Apache Lucene and Elasticsearch

Apache Lucene is a powerful open-source search engine software library. It provides tokenization, analysis, indexing and search capabilities. The search engine is highly performant and is scalable. Large amounts of documents can be indexed rapidly on modest hardware, both in batch fashion and incrementally. It is possible to also horizontally scale Lucene based applications across multiple nodes to support larger datasets. It is also available in multiple programming languages and can be directly used or extended from any custom application. Although it is beyond the scope of this document, Lucene can also be used to implement a recommender system. Apache Lucene forms the basis for many other projects including Elasticsearch. Elasticsearch is an enterprise search server that internally uses Lucene and provides a broad range of use cases. These tools can be utilized, extended, and combined in a number of interesting ways and a discussion of these applications is useful for those interested in text search applications.

Apache Lucene is a high-performance cross platform text search engine library. It can be used in any application that relies on document indexing and text search. Lucene allows for a document to be defined as a collection of fields, where a field is either a binary, numeric or text type. It is possible to search only certain fields of a document and it is also possible to build and search multiple indexes. Internally, Lucene represents these indexes as an inverted index. The standard Lucene distribution comes with analyzers that can provide tokenization and handling of stop words. The search results are returned ranked based on relevance to the query terms. Lucene provides a pluggable API for using different search models including Vector Space Model or Okapi BM25. Although it is straightforward to build a custom search-based application using Lucene by simply importing the library and calling the Lucene API, the more common use case is to use an off the shelf solution such as Elasticsearch.

Apache Elasticsearch is a distributed analytics engine that is based on Lucene. It supports a broad range of use cases that are related to search and analytics. Whereas Lucene is a lower-level library that by default requires an application on top of it to make use of more primitive search capabilities, Elasticsearch comes out of the box with support for common enterprise search uses. Out of the box Elasticsearch can support clustering to support much larger quantities of data in real time. One very popular application is the monitoring of logs. In an enterprise or big data environment with many software systems in operation, there will often be massive amounts of logging data that is continuously generated. It can be difficult to monitor, interpret and search this data but a system like Elasticsearch offers a scalable way to analyze this data. These analytics capabilities naturally extend to provide infrastructure monitoring and security/threat detection. One other particularly interesting capability that further extends the scalability of this search system is the use of HDFS or Hadoop Distributed Filesystem. HDFS is a distributed file system that is horizontally scalable and provides high performance and significant fault tolerance. One could maintain a large data lake of text documents on an HDFS cluster and then make use of the Elasticsearch-Hadoop connector to interface an Elasticsearch cluster and an HDFS cluster to provide truly massive scale real time search and analytics capabilities. Additionally, Elasticsearch can also interface with object storage over S3 protocol or use other public cloud provided data stores on Azure, Google, etc. The combination of these datastores, Elasticsearch and a scalable ETL solution to aggregate data provides a surprisingly simple architecture for large scale text search/analytics platform. All these tools including Apache Lucene, Elasticsearch and Hadoop are open source and available for use for free.

Apache Lucene and Elasticsearch are open-source tools that provide a highly accessible solution for state-of-the-art text search capabilities. They are extensible and cross platform tools that contain a broad range of capabilities that can be applied to numerous use cases. Apache Elasticsearch is particularly useful and popular in enterprise applications. Elasticsearch is also an appropriate tool in the big data space due to its ability to handle massive scale text data by scaling horizontally with its clustering support. Additionally Elasticsearch supports interoperability with many other big data technologies such as Hadoop/HDFS, S3, etc. The combination of these tools offers a scalable, fault tolerant and elegant solution for large scale text search, data reporting, dashboarding, threat detection, infrastructure monitoring and many other forms of analytics.

References:

https://lucene.apache.org/core/

https://svn-us.apache.org/repos/asf/lucene/java/site/docs/index.pdf

https://www.elastic.co/what-is/elasticsearch-business-analytics

https://dl.acm.org/doi/abs/10.1145/2652524.2652556

https://www.elastic.co/elasticsearch/

https://www.elastic.co/what-is/elasticsearch-hadoop

http://ceur-ws.org/Vol-1823/paper8.pdf