Apache Solr: Enterprise search platform

Alok Kumar

Introduction

Over the past decade, web presence - both external facing, for customer and partners; and internal for employees, management and team, has become a crucial factor in almost every organization from start-ups to trillion dollar companies. Consequently, the amount and type of content available has grown manyfold, to an extent that organizing and making use of the online content has become a significant challenge. Even if one restricts the content to text (and not binary data like images), the solution for searching and retrieving meaningful content from this massive volume of data has created its own domain of text search and retrieval theories, open source and proprietary tools and platform specializing in the same. Apache Lucene has emerged as the de-facto base library for many of such tools and platforms, the most oldest one being Apache Solr which provides a production grade enterprise platform implementation of Lucene. However, there are a lot of relatively newer players in the search market and we try to assess if Solr is still relevant among the various tools and platforms.

What makes a search platform “Enterprise-grade"

Any software claiming to be enterprise grade should provide certain features, as detailed below. Solr features as they pertain to search are also described along with the respective feature:

1. Scalable: Any enterprise scale platform should be able to handle large scale of traffic and scale relatively linearly with the increased demands. Solr can support massive amounts of traffic and is being used by many large sized companies like Netflix, eBay, Disney, Reddit and many others. The processing or the computing power could come from the underlying docker/Kubernetes cluster thus imparting the Sole platform scalability and on-demand computing addition as needed.
2. Administration and Monitoring: Administrative interfaces are built in the platform , making it easy to control the platform. Solr publishes a lot of metrics via JMX which provides insight into the platform and instances.
3. Fault tolerance and Load Balancing: Automatic recovery from crashes is a very important feature of any enterprise platform. Solr uses Apache Zookeeper for cluster coordination and

configuration to provide fault tolerance and high availability (Zookeeper is a proven service for maintaining configuration, naming, distributed synchronization - and is used in many distributed platforms like Hadoop).

Solr also uses Zookeeper's database to figure out which servers need to handle any search request, making it flexible and truly distributed; queries could be sent to any server.

4. Security: Solr can be secured using one of the many pluggable standard approaches like SSL, Authentication and Role based authorization. Organizations can also create their own authorization plugins using Solr API.
5. Standard interfaces and extensibility: Solr interaction is based on standard HTTP REST-like API, and the message format could be one of the many open standards like JSON, XML, CSV etc. It also publishes many well-defined extension points to make connecting index and query time plugins easy. The whole project being an Apache project, also allows enterprises to change or modify any piece of code to suit their requirements.
6. Advanced search features: Solr provides many other features making it one of the most popular search platforms currently available:
   a) Full-text search capability
   b) Near real-time indexing
   c) Geospatial search
   d) Rich document parsing
   e) Query suggestions, auto-complete, spell check etc.

The good, the bad and the ugly

"Faster, Cheaper, Better - pick two" - is an old saying among engineers. While originally intended for non-software products, it probably applies to software as well, the idea being that no product can excel in everything, due to the very nature of the capabilities. While Solr presents a compelling proposition when looking for an enterprise search platform, those features come at a cost which may deter its adoption among some organizations. Some of the criticisms that Solr has faced are:

1. Too complicated to configure
2. Limited metrics
3. Steep learning curve
4. Limited analytics
5. Few supporting products in the ecosystem
6. Slower refreshes due to single cache architecture

Many newer entrants like Elasticsearch has tried to shine in the areas traditionally cited as the weaknesses in Solr (Elasticsearch is based on the same Lucene core). Elasticsearch, for example, is considered easy to learn and cuts down implementation time significantly for organizations adopting a new search platform. There are other open source search platforms like Sphinx and Terrier. Some of the products, like Splunk, while not directly in the category of search platforms, also compete with Solr when it comes to usage and functionality.

The truth, however is rarely binary. While it is true that many new search platforms have emerged with new features and approaches, Solr use and adoption still makes it one of the top three players in the market. It is partly due to its features and maturity, and partly due to the support of the open source community that keeps adding new technologies and features to it, while improving its core functionality. The newer Solr versions, for example, provides similar JSON search APIs as Elasticsearch overcoming some of its limitations.

Conclusion

Search platforms, like other enterprise tools, are evaluated not only based on their core capability, but also on other features which provide much needed robustness, scalability, operational and monitoring features, security and ease-of-use. Solr is one of the oldest products, used in many large size organizations. It is true that there are many more options now - both open source and proprietary, Solr community and developer support still make it relevant, and would continue to do so in the near foreseeable future. The author believes that the inherently different architectures between different search platforms make choosing one a nuanced decision, one that has to be tailored keeping in mind the organization's requirements, maturity and size. Given great variations in those factors, different search platforms, Solr among the prominent ones would find sizable adoption among enterprises.

References

Apache Solr [Online]: https://solr.apache.org/

Search Engine Rankings [Online]: https://db-engines.com/en/ranking/search+engine

Yigal, Asaf, 2020: Solr vs Elasticsearch [Online]: https://logz.io/blog/solr-vs-elasticsearch/

Klimenko, Ana: Sphinx vs Elasticsearch vs Solr [Online]: https://greenice.net/elasticsearch-vs-solr-vs-sphinx-best-open-source-search-platform-comparison/