DATASTAX

Implementing Search in Web, Mobile, and IOT Applications

An Overview of DataStax Enterprise Search

Table of Contents

Introduction

Nearly every web, mobile and Internet-of-Things (IoT) application has search functionality that helps its users locate the information they want. Depending on the industry and need, the search requirements for an application may be very simple or astonishingly complex.

While the need for search capabilities is ubiquitous in modern applications, it is surprising to find that most enterprises struggle with building high-performance, robust and cost effective search components into their business systems. Even though most database vendors have search functionality as part their platform, and various specialized search software exists in both open source and proprietary form, enterprises still wrestle with scaling their systems so they can process increasing data volumes and users, and keeping the search capabilities of their web and mobile applications always online.

This paper describes the general search requirements that most web, mobile and IoT applications have, and the common ways enterprises have tried to deploy search systems in the past. It then describes how NoSQL database systems are fast becoming the standard database platform for these types of applications and why DataStax Enterprise, with its integrated enterprise search capabilities, can make developing powerful search components for an application fast, easy, and cost effective.

Why Search?

Whether it's a retail web application that smartly guides its user through their buying decision process or a mobile entertainment app that seems to know what each particular user wants, search technology is behind the scenes, personalizing every experience. Search functionality has become critical in every web and mobile application for helping users navigate directly to the products, services, and media they are interested in.

General Search Requirements

While every application is different, there are certain core search requirements that most every modern web and mobile application has. When it comes to specific search features, the most routine must-have's include the following:

- **Directed navigation** – the ability to guide a user through a series of choices to find what they want, most oftentimes accomplished through faceted search and other search features like wildcarding, groupings, and more.
- **Breadcrumbs** – an important part of navigation for retail sites, they help inform a user where they are within a site and assist with upward navigation.
- **Query assistance** – involves a multitude of aids that assist the user in entering a search request and delivering search results that both directly and indirectly match the user's request. Assistance can be in the form of auto-completion, spell checking, synonym and acronym extensions, and providing "more like this" based results.
- **Relevancy control** – helps handle the ranking of search results and is typically adjusted quite a bit to ensure the proper placement of products and services an enterprise wishes to promote.
- **Spotlights** – part of relevancy control, it allows businesses to directly affect search queries to ensure products and services are consistently ranked high in search results.
- **Personalization signals** – assists in smartly personalizing search results using data such as current user location (e.g. enabled with geospatial search) and historical search patterns.
- **Rich document handling** – allows for the inspection of document contents that may be in formats such as Adobe PDF, Microsoft Word, etc.
- **Content integration** – provides a blending of core search results with other integrated and related data (e.g. product reviews for products listed in search output).
- **Analytics** – supplies the ability to understand events such as failed searches, conversion metrics and more.

From an architectural perspective, a search subsystem needs the following to support global web and mobile applications:

- **Continuous uptime capabilities** – users must be able to use an application's search component without encountering any outages or downtime of search functionality.
- **Multi-data center and cloud support** – web and mobile applications are "follow you everywhere" in nature, and as such, it is important that a search subsystem be able to span multiple data centers and cloud zones around the globe so that search operations are fast regardless of a user's location.
- **Scalable performance** – the search system must be able to grow in an online fashion to accommodate increasing data volumes and user connections, while delivering consistent performance. Further, new data should be quickly indexed and made available for search as fast as possible.
- **Standards based interfaces** – the search API's should include usual and customary interfaces such as HTTP, XML, and others.

Traditional Deployment Strategies for Search

IT organizations learned long ago that, even though RDBMS's provided search options within their engines, search traffic caused resource contention for data and compute resources in ways that impacted the performance of transactional (OLTP) work. Because of this, in the same way that IT groups separated OLTP and analytics workloads, they began to also break out search into a different system.

These "mixed workload" situations typically result in "sharded" data management environments that have separate data platforms for OLTP, analytics and search functionality, and an application that is specially coded to access each distinct vendor's data platforms. In addition, data is constantly extracted-transformed-and-loaded (ETL'd) between the three data platforms as data common to OLTP, analytic, and search requests must be present on all systems.

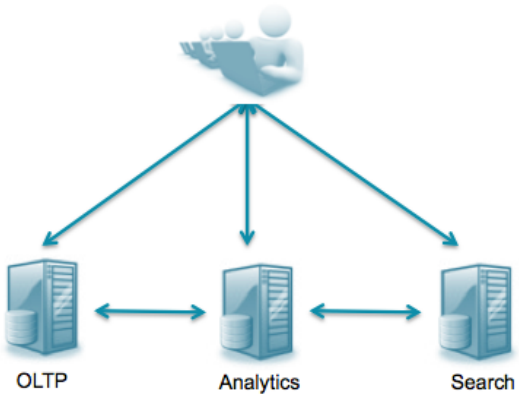

Figure 1 – Typical Sharded Application

Even with NoSQL databases beginning to usurp the place of RDBMS's where web, mobile, and IoT applications are concerned, the data platform sharding approach is still being utilized with only the vendor names changing (e.g. OLTP might be handled by Apache Cassandra™, analytics by Apache Hadoop™, and search by Apache Solr™).

Most agree that, regardless of the data platforms used, sharded systems are difficult to manage and maintain and can deliver higher than expected total cost of ownership even when "free" open source software is used. Businesses needing to solve their mixed workload problem want an easier approach that also offers a way to quickly build robust search functionality into their applications.

What is DataStax Enterprise?

As previously alluded to, modern web, mobile, and IoT applications have evolved past centralized systems that made use of relational databases (RDBMS's) as their data management foundation. These modern applications require a database platform that is able to meet the scale, performance, and data distribution needs of radically-connected systems.

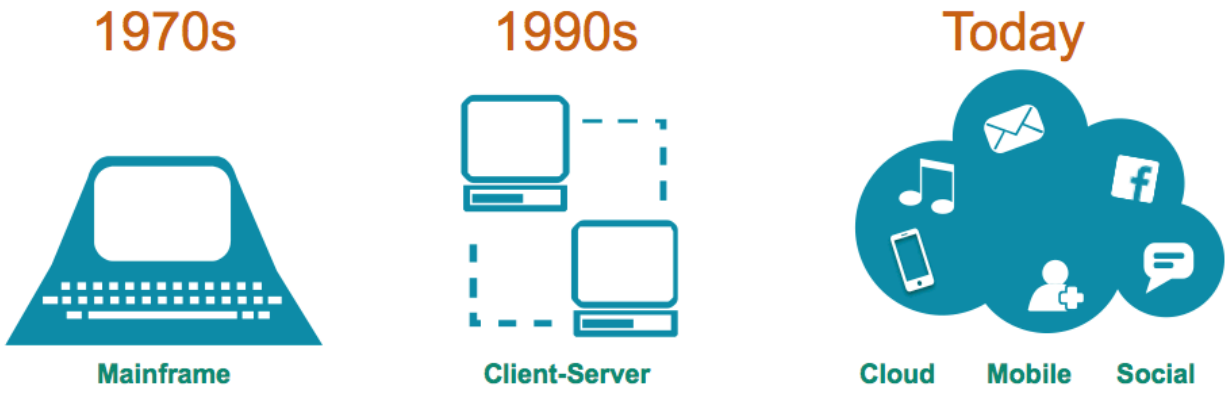

Figure 2 – The evolution of data-driven applications.

To meet these requirements, DataStax Enterprise (DSE) delivers Apache Cassandra in database platform that meets the performance and availability demands of IoT, web, and mobile applications. It gives enterprises a secure, fast, always-on database that remains operationally simple when scaled in a single datacenter or across multiple datacenters and clouds.

Figure 3 – Components that make up DataStax Enterprise.

DataStax Enterprise provides everything needed to deploy NoSQL and Cassandra in production environments. A certified version of Cassandra that is tested and optimized for production applications is included, along with advanced security features to protect sensitive data, management services that automatically perform key maintenance and tuning functions, advanced visual management and administration capabilities, and around-the-clock expert support.

DSE also solves web, mobile, and IoT application's "mixed workload" problem by smartly integrating analytics and search functionality into the platform along with built-in workload isolation and replication abilities that keep OLTP, analytics, and search workloads separate from one another. This functionality of DSE eliminates the need for multiple data management providers and application sharding.

What is DSE Search?

DataStax Enterprise supplies built-in enterprise search functionality (DSE Search) on Cassandra OLTP data in a way that is tailor-made for the search requirements of modern Web, mobile and IoT applications. Built on Apache Solr, DSE Search provides full Solr compatibility along with additional enterprise search capabilities that enable enterprises to quickly build robust search components into their systems.

Including enterprise search capabilities in a DSE cluster is very easy. DSE Search can be enabled in a new or existing cluster by provisioning nodes that are devoted to search operations.

For web or mobile applications that include mixed workloads involving OLTP, analytics, and search, an administrator provisions however many nodes are needed to service each workload, with those nodes operating as a distinct group of compute resources within a single cluster. Data is then automatically replicated between the OLTP, analytic, and search node so there is no need to manually ETL data between different systems and platforms.

DSE Search delivers all of the features and functionality of Solr, with additional enterprise search, uptime, and performance benefits included as well. The following sections highlight the key benefits of DSE Search that make it well suited for the search requirements of today's web, mobile, and IoT applications.

Enterprise Search Support

DSE Search meets the general search requirements of web and mobile applications previously listed. Use cases supported by DSE search include general web, full-text, faceted (categorization), hit prioritization and highlighting, log mining, rich document (PDF, MS Word, etc.) analysis, geospatial, and social media match ups.

Distributed Enterprise Search

DSE Search is built for distributed web and mobile applications that need to easily search data contained in large data stores. The divide-and-conquer architecture allows for consistently fast response times across large data volumes that may be distributed across multiple machines and locations.

Multi-Data Center and Cloud Search

DSE Search's distributed capabilities include the capacity to run across multiple data centers and cloud availability zones in Active-Active-Active-nActive manner, which allows search operations on OLTP data to be easily carried out in different geographical regions. The key benefit is that search results can be sent back to users in those locations in the fastest possible time.

DSE Search differs from other search solutions like Elasticsearch in that it allows for full write and read operations in multi-Active manner, whereas products like Elasticsearch are limited by their master-slave architectures and can only support Active-Passive configurations where a master and read slave machines are used for multi-DC operations.

Always-On Search

DSE Search is perfect for applications that need search functionality that is always available and never goes down. DSE's always-on architecture, built on Cassandra, ensures 100% uptime for search operations. DSE Search allows a user to create multiple copies of search data across multiple nodes, data centers, and clouds so even if certain machines or data centers go down, data is always available for search tasks.

Online Elasticity

Additional capacity (i.e. more search nodes) can be added online so search workloads can easily scale to meet incoming data and customer demands.

Live Search

DSE Search contains a unique “Live Indexing” feature that allows new data entered into the database to be immediately available for search. Whereas typical search systems may have gaps involved between when new data enters the system and when it is ready for search operations, DSE’s Live Indexing feature indexes fresh data making it quickly available for search.

With Live Indexing, enabled indexing throughput doubles, and indexing throughput remains linear with CPU cores on each node.

Secure Search

DSE Search offers native support for user authentication, user authorization, data encryption and firewall configuration. DSE Search does not require costly external plug-ins or external security configurations. All DSE Search settings and enablements are sourced from a single configuration space.

Security features enabled in DSE Search include LDAP, Active Directory, and Kerberos authentication, client-to-node and node-to-node encryption, and data auditing, which provides administrators with a view into events happening in their DSE cluster.

Fault-Tolerant Search

DSE Search includes options to automatically retry search queries that fail due to a node going down, with other replicas containing the same data being transparently accessed. Another option provides the ability for partial results to be returned when the use case allows for it.

Workload Isolation and Management

DSE Search fully supports workload isolation and management, ensuring that search workloads do not compete with OLTP or analytic workloads for data or compute resources. Cassandra’s powerful replication abilities automatically copies and moves data among nodes so there is no need to extract data from transactional databases and load them into another search system. Everything is contained within one database cluster.

Automatic Query Routing

To help ensure fast search response times, DSE Search automatically routes a search request to the best performing replica in a cluster that holds the data needed to satisfy the request. The system takes into account a number of factors including each search node’s uptime, current workload, and network distance to the user, and sends the request to the node that is able to handle the request in the most optimized manner.

Solr Compatible

Helping power DSE Search is a production-certified version of Apache Solr. DSE Search inherits all the power and capabilities of Solr and builds on top of it to create even more powerful enterprise search functionality. Anyone familiar with Solr can immediately begin to develop with DSE Search using the same Solr API’s.

Using Solr and Lucene as its foundation, DSE Search merges the ability to perform complex transactional queries with the solubility and high availability of Cassandra. In addition, DSE Search’s Solr compatibility layer allows legacy search workloads to be seamlessly transferred to DataStax Enterprise with no modification to existing client code or behavior.

Integration with Cassandra Query Language

Search/Solr syntax is integrated with the Cassandra Query Language (CQL), which enhances CQL in a way that allows it to operate as a powerful search language. Solr syntax (e.g. a wildcard search) may be passed directly through a CQL WHERE clause so that data can be searched for via CQL in addition to the native Solr API's.

The CQL language and transport is cluster aware meaning that it is aware of changes to the schema or cluster topology in real-time. If a node gets added to the cluster it will be automatically be added to the connection pool of all clients. If a node is removed for maintenance or fails it will be removed from the connection pool of all the clients minimizing high latency request timeouts or failures. Because the CQL protocol is cluster aware it is able to avoid added complexity, fragility, and the cost of load balancers.

Integration with Spark Analytics

DSE Search integrates with Apache Spark™ SparkSQL in the same way it does with CQL. Solr search syntax may be passed through a SparkSQL WHERE clause on DSE analytic nodes running Spark, which greatly expands Spark's analytic query ability and combines both search and analytic functions in one statement.

Visual Management and Monitoring

DSE Search functionality and operations can easily be visually provisioned, managed, and monitored with DataStax OpsCenter.

Common Use Cases

In addition to typical search application usage, there are a number of common use cases that benefit from DSE Search.

Document and Message Search

A common use case for DSE is either as a document or message store. In the instance of a document store, the documents may be all documents, or correspondence pertaining to a customer account. These records could be chat messages, correspondence, financial statements, reports, transactions, or records.

This model typically has two user personas: the auditor and the end user. The auditor is concerned with all documents across all users that match a certain criteria. The end user is concerned with which of their documents match a particular key phrase, or were sent by a particular sender.

In this model both the metadata for each record, and the full text body of each record is indexed. When an auditor performs a search, all records across all users can be obtained. Conversely, when a user performs a search, the result set is filtered to only records associated with the user account. Additionally, DSE Search provides additional functionality to pass locality information so that when user facing, high volume, low latency queries are performed, the fan out processing normally done by other search software is avoided and only a single node needs to participate in the query processing.

Master Data Management (MDM)

Master Data Management is a paradigm where a central repository contains all information about an item. An item could be simple like a t-shirt that has 30 attributes, or an item could be complex like a TV and have over 600 attributes.

With MDM, it is typical that every field for an item be both indexed and searchable. Data sources are typically pulled from multiple sources such as suppliers, shippers, and the vendor's own internal processes. Typical applications include product catalog for vendors, retailers or manufacturers.

The catalog could service either internal process or external customers. In both instances the focus is on a large volume of low latency queries.

Real-Time Search Analytics

Real-time analytics typically operates over an event stream of many small machine generated events. This could be log data from servers or marketing and revenue data from an online retailer.

With real-time search analytics, the emphasis is to quickly generate a report of events, users, etc., which satisfies some conditions. Examples could include “find all servers where this exception occurred”, “count all users which have logged on in the previous 30 days”, and so on. The emphasis is on counting or identifying records which match search criteria.

Hybrid Search/Batch Analytics

Real-time search analytics can be limited when the process can’t perform aggregations or do deeper more complicated calculations without help. To perform these types of calculations, DataStax Enterprise integrates Apache Spark, which is a perfect fit for such use cases.

As mentioned earlier in this document, Spark is a batch analytics framework with advanced functionality such as graph abstractions and machine learning, and DSE’s unique integration with analytics and search allows customers to initiate a batch job that uses a search query as it’s source. This greatly reducing the number of records that must be processed, and thus reduces response times. Using this methodology it’s possible to reduce complex batch-reporting times to seconds or possibly sub-second.

DSE Search Customer Examples

The following DataStax customer examples illustrate how DSE Search is being deployed in enterprise environments.

Clear Capital

Clear Capital is the premium provider of data and solutions for residential and commercial real estate asset valuation and collateral risk assessment for large financial services companies. Clear Capital facilitates the ordering, tracking and delivery of valuation reports by leveraging massive data sets, human-based review and automated review tools. Clear Capital currently stores valuation data for over 90% of all properties in the U.S.

With the largest source of valuation data for residential and commercial properties in the U.S., Clear Capital recognized a market opportunity to develop a software application that can digest their massive database of properties to deliver highly accurate and recent valuation data quickly to financial institutions and banks. They understood relational database systems were not built to support high volumes of small transactions, linear scalability, real-time performance, and continuous availability – all key requirements for their cutting-edge valuation review management software as a service platform.

Clear Capital’s deciding factor to go with DataStax Enterprise was DSE Search and powerful analytics abilities. A major component of Clear Capital’s appraisal system is their geospatial search and analysis capability. This level of analysis allows Clear Capital to gain valuable market insights based on aggregate statistics, which feeds into the accuracy of the valuations delivered. “DataStax Enterprise powers our geospatial search and delivers real time insights to our customers who rely on the most recent data to support their market-level decisions,” said David Prinzing, solutions architect at Clear Capital. “This was a significant reason why we chose DataStax Enterprise to power our system.”

Penn Mutual

Penn Mutual is a life insurance and annuities company that has operated since 1847. Across almost 170 years of business, Penn Mutual has been dedicated to help people do more in life by creating solutions that deliver the complete value of life insurance across all life’s stages.

In 2010, Penn Mutual's Information Management and Technology Division, the IT arm of the business, started a project called "Core Services" aiming to merge all data domains spread throughout the company into a single source by marrying their service oriented architecture and master data management capabilities into a comprehensive system. Penn Mutual started out with a traditional RDBMS approach for the persistence layer of their Core Service, but soon realized that it could not meet their requirements for application performance or scalability with the existing RDBMS footprint without a large cost commitment.

Penn Mutual chose DataStax Enterprise for their MDM system, with a prime motivator being DSE Search. DSE Search allows Penn Mutual to offer traditional data access services and ad-hoc query to create more data discovery type applications, with the end result being improved ability to find information and pull reports.

Internet Identity

Internet Identity (IID) is a cyber security company that provides the platform to easily exchange cyber threat intelligence between enterprises and governments. Fortune 500 companies and large government agencies leverage IID to detect and mitigate threats.

With high volumes of multi-structured data pulled from a wide array of sources, IID struggled to keep up with the ongoing flow of malicious threats due to their reliance on legacy relational database technology. They made a move to NoSQL and chose DataStax Enterprise in part because of DSE Search.

DSE Search allows IID to easily search and index data, while DSE's integrated analytics allow them to quickly identify security threats and deal with abnormal behavior. IID's CTO remarked, "The fact that DataStax integrated three core elements with an operations console on top of it that allows us to monitor and measure was enormous. And I have all of this along with the scalability and availability of Apache Cassandra, advanced security capabilities and 24x7 support – all for one-fifth the price that I would pay for a relational database."

Conclusions

DataStax Enterprise with DSE Search provides everything modern Internet Enterprises need to build scalable and high-performance search capabilities into their Web, mobile, and IoT applications. For more resources and downloads of DataStax Enterprise, visit www.datastax.com today.

Appendix A – Search Feature Comparison

This section provides a general feature comparison between DSE Search, Apache Solr, and Elasticsearch.

	DSE Search 4.7.0	Solr 4.10.2	ES 1.6.0
INDEXING			
Schema Creation (Schema-less)	Yes	Yes	Yes
CJK Support	Yes	Yes	Yes
Partial Document Updates (Atomic)	Yes	Yes	Yes
Live Indexing	Yes	No	No
SEARCH/QUERY			
Query Syntax	CQL + JSON (Solr format)	key/value pair based using / and () to delineate and nest queries.	JSON

Distributed Group By/Collapse by field	Yes	Yes	Yes
Full Text Search	Yes	Yes	Yes
Geospatial queries	Yes (Solr)	Yes	Yes
Field Types and analyzers	Yes	Yes	Yes
Query Rescore	Yes	Yes	Yes
Auto-Mapping	Yes	Yes	Yes
Query time join	Yes	Yes	Yes
Deep paging	Yes	Yes	Yes
Per-segment filters	Yes	No	Yes
Multi-threaded queries	Yes	No	No
Auto-retry	Yes	No	No
Distributed Search			
Distributed Queries	Yes	Yes	Yes
Node Discovery	Cassandra	Requires ZooKeeper	Zen Discovery
Coordination	Cassandra	Requires ZooKeeper	Self-contained
Automatic Shard Rebalancing	Cassandra	No	Yes
OPERATIONS			
Continuous Availability	Yes	No	No
Integration of analytics and search workloads	Yes	No	No
Data Resiliency	Yes	No	No
Shard Splitting	Cassandra Ring	No need to reindex when adding more shards	Requires reindexing
Linear scaling	Yes	Yes	Limited
Durability	Yes	NA	NA
Partition Tolerance	Yes	Need ZooKeeper	Split-Brain
Consistency	Eventual (tunable)	Sync	Sync / Async
Admin Interface	Yes	Yes	Yes
Fine grained memory statistics	Yes	No	No
Multi-data center write/read anywhere	Yes	No	No
SECURITY			
Authentication System Support (LDAP)	Yes	No	No
Encrypted Communications	Yes	No	No
Audit Logging	Yes	No	No
Kerberos	Yes	No	No
API			
Format	XML, CSV, JSON	XML, CSV, JSON	JSON

HTTP REST API	Yes (Through Solr API)	Yes	Yes
Binary API	Yes	Yes	Yes
JMX Support	Yes	Yes	No
Drivers	Comprehensive	Solr4J	Comprehensive
Output	Solr JSON output	Solr JSON output	JSON in, JSON out
OTHER			
Analytics	Yes	Through 3rd Party	Through 3rd Party
Integration with Spark analytics	Yes	No	Partial
Manageability	OpsCenter	Admin UI	Marvel

About DataStax

DataStax delivers Apache Cassandra in a database platform purpose built for the performance and availability demands of Web, Mobile, and IOT applications, giving enterprises a secure always-on database that remains operationally simple when scaled in a single datacenter or across multiple datacenters and clouds.

DataStax has more than 500 customers in 38 countries including leaders such as Netflix, Rackspace, Pearson Education, and Constant Contact, and spans verticals including web, financial services, telecommunications, logistics, and government. Based in San Mateo, Calif., DataStax is backed by industry-leading investors including Lightspeed Venture Partners, Meritech Capital, and Crosslink Capital.