Use Case — High Tech
Intel Big Data Enablement — Online Media
Text Search and Recommendations

Intel and Cloudera Help Improve a Company's Text-Based Searches Resulting in Increased Revenue

Intel and Cloudera redesign a company's search engine, providing text-based recommendations, improved search results, faster performance, and more revenues.

Why Intel and Cloudera

Intel and Cloudera take the guesswork out of Hadoop. Using a unique collaborative approach, we delivery excellent performance, security, and quality distribution, built on open standards. Working with more vendors across the ecosystem, only a solution built on CDH can ensure freedom from lock-in, enabling you to build a robust big data solution to meet the needs of your business today and into the future.

- Uniquely aligned product roadmaps for software and hardware to drive innovation faster, providing more industry firsts than any other Hadoop alternative.
- Deep partnerships with virtually every provider in the data center, streamlining the process for building Big Data solutions.
- Proven track records of identifying the driving industry standards, so you don't run the risk of stranding yourself on an island.

An independent advertising agency that specializes in social media and brand-creation hits a performance wall with its legacy SQL solution. The Company updates its operations to a Cloudera enterprise data hub (EDH) and realizes significant performance improvements and revenue increases.

Built entirely upon the Cloudera distribution of Hadoop (CDH), the new portal ingests and processes large amounts of raw data. Cloudera Search, a fast and powerful content-based search engine that is bundled with Cloudera Enterprise (subscription), quickly provides the Company's customers with results and text-based recommendations. The entire Cloudera system delivers a fault-tolerant, redundant platform with industry-leading security features.

Results

The Intel/Cloudera solution yields the following improvements:

- Number of results served per minute increases from 20,000 to 35,000 (75% increase).
- Total clickthrough rate (CTR) per visitor increases 15%.
- Revenues more than double (103% increase).

Business drivers

Companies that rely on social media for their livelihoods need to be nimble and responsive to their customers' needs. In a data-driven environment where slowness means loss of revenue, companies utilizing outdated technologies are at a disadvantage.

Before Cloudera, the outdated and underpowered SQL that the Company's server was built on did not have the ability to keep up with the growing number of search queries, due to the Company's high demand. The Company's potential exceeded their capacity to store and process enough data to serve their customers. They needed a solution that would enable faster performance, as well as the ability to expand, with minimal changes to the application.

The Company's technology team asked:

- Are we losing business because our customers are not getting the information in a timely manner?
- Should we beef up our legacy system to handle larger workloads or invest in new technology designed specifically for high-volume Big Data loads?

Solution details

The Company contacted Intel to help solve its Big Data needs. Based on our recommendations, they decided to use a multinode Hadoop cluster running CDH.

Figure 1 shows how the various components interact with each other. The Company collects vast amounts of data from social media portals such as Twitter and Instagram. This data must be stored in a system that provides readily available information to customers whenever they need it. Customers also require fast search capabilities along with an option to tag based on their interests, which then provides them content in future searches.

Because data in a Cloudera system is stored in an HDFS cluster, the Company instantly gained the inherent advantages of a scalable, fault-tolerant file system and index storage. The massive quantity of social media data was no longer a barrier, as it had been under the Company's outdated system. The Company could simply add affordable nodes to increment capacity as their data storage and processing needs grew.

By deploying Cloudera Search—which is optimized for high traffic and advanced full-text search capabilities, and is integrated with Apache Flume* and Lily HBase Indexer* (a collaborative effort between NGDATA and Cloudera)—the Company achieved near real-time (NRT) indexing at scale. End-users noticed the performance improvements immediately.

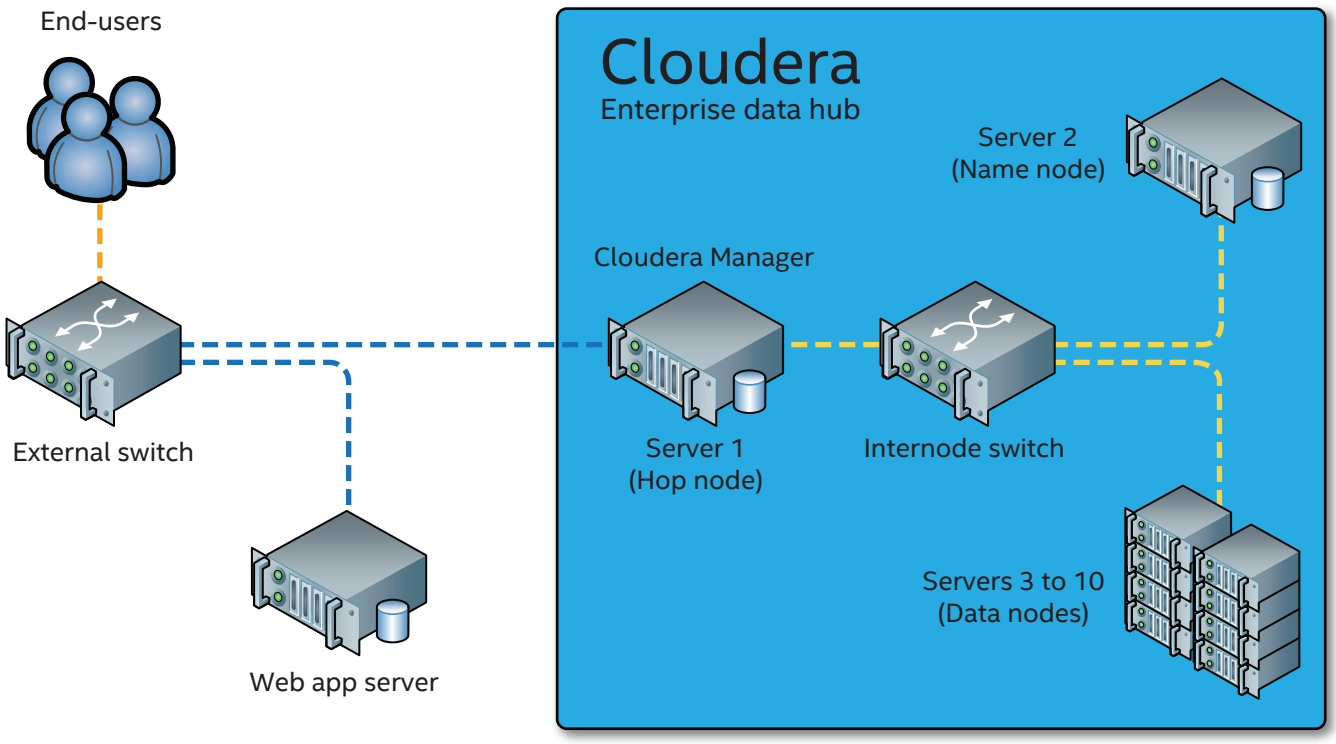
Based on Apache Solr*, Cloudera Search sifts through all server content data in the HDFS cluster. Solr*, which uses lucene classes to create an "inverted index", maintains a posting list that maps words, terms, and phrases along with the corresponding places where they occur.

A web application, connected directly to the Cloudera cluster, serves content to end-users via a user-friendly web interface.

Benefits of the new system include:

- **High-performance content searches.** Using Cloudera Search on top of HDFS has made real-time search results and recommendations available to customers. This gives the Company a substantial performance improvement when compared to their legacy proprietary SQL system.
- **Fault tolerance, redundancy, and linear scalability.** With the use of CDH, the Company is no longer dependent on a single instance of SQL server, but relies on the entire fault-tolerant, scalable cluster. In the event of increased data searches or workload, the Company can add to the

Figure 1 **Cloudera-powered search results and recommendations.** End-users submit queries to the web application server via worldwide web. The Cloudera enterprise data hub processes queries and rapidly returns search results and recommendations to end-users.



cluster, increasing processing power without changing the application.

- **Instant revenue gains.** Immediately upon implementation, the number of customer search queries returned per minute increased by 75%. Moreover, search results became more accurate, which resulted in increased clickthrough rates for graphic and text advertisements and a subsequent doubling of ad revenues.

Cloudera Enterprise

Intel built the Company's portal on the Cloudera distribution of Hadoop (CDH), which handled everything from ingesting data to providing recommendations. CDH provides an efficient and cost-effective platform for Big Data solutions and enables new solutions for the Company to provide personalized, relevant content to its clients.

Cloudera Search has brought full-text, interactive search and scalable, flexible indexing to the Company's EDH. Through its unique integrations with CDH, Cloudera Search gains the same fault tolerance, scale, visibility, security, and flexibility provided to other workloads. Its user-friendly, interactive search offers even nontechnical users a common, intuitive method for accessing and querying data across large, disparate data stores with mixed formats and structures—while simultaneously allowing access to the same data for advanced workloads.

Powered by Solr's standard APIs, Cloudera Search also allows users to express queries through natural language, logical operators, and even advanced regular expressions, with capabilities like "fuzzy" matching expressions that "find similar" results.

Cloudera Search provides various advantages, such as:

- Apache Solr-based for easy integration.
- Robust and scalable index storage in HDFS.
- Scalable batch indexing through MapReduce.
- Simple and reusable data extraction.
- Extended file format support.

Summary

With Cloudera, the Company not only benefits from better performance, enhanced security, role-based access authorization, fault tolerance, and batch/parallel processing, they now have a solution that accommodates their current and future needs. They turned to Intel to help them migrate from an outdated system to a Big Data platform that scales economically and will continue to grow as they do.

Let us help your business too.

Spotlight on Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop™. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Cloudera's open source Big Data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,600 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

For more information, visit **www.cloudera.com**.

cloudera®

Meeting your needs

We look forward to meeting with you to define your requirements and meet your objectives.

- **Accelerate time to value:** Achieve real-time cost savings, respond to market trends, and drive innovation.
- **Secure Big Data:** Deploy a sustainable Big Data program that doesn't put your organization, or you, at risk.
- **Maintain control:** Work with a partner who educates your team so you become self-sufficient.
- **Increase business potential**: Create and execute a plan that helps you adapt now, and in the future.

Hadoop sizing guide

		Cluster size		
		Small	**Medium**	**Large**
CPU		**Intel® Xeon® Processor E5 v3**		
Storage (TB)		<72 TB	72 to 570 TB	>570 TB
Node count	**Master**	2 to 3	4 to 7	≥8
	Slaves	<12	12 to 95	≥ 96
Memory (GB)	**Master**	64 GB	128 GB	≥256 GB
	Slaves	48 GB	96 GB	≥128 GB
Network		1 Gbps	10 Gbps	10 Gbps

Hardware configuration is highly dependent on workload. A high storage density cluster may be configured with a 4 TB JBOD hard disk, while a compute intensive cluster may be configured with a higher memory configuration.

Contact us

Contact your sales rep or e-mail us.

Intel.com/bigdata/services

332602-002