# Intel and Cloudera Help Design a Content Recommendation Engine for Chinese Content

**Using Intel algorithms customized for the written Chinese language, a regional media publisher increases readership and advertisement revenues.**

## Why Intel and Cloudera?

Intel and Cloudera take the guesswork out of Apache Hadoop*. Using a unique collaborative approach, we deliver excellent performance, security, and quality distribution, built on open standards. Because we work with hundreds of vendors across the ecosystem, a solution built on Cloudera Enterprise 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 many industry firsts with Hadoop.
- 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.

A Chinese language media company serving the international sinophone market decides to modernize its infrastructure to keep pace with the shift in news distribution. As readers have changed their consumption habits, the Company wants to adopt search engines, social media, mobile apps, and video content designed specifically for Chinese logograms—to cater to the needs of its clientele.

Intel develops a Chinese language content recommendation engine that automatically creates a personalized news portal for the Company's readers. This portal rests entirely on Cloudera, which handles everything from ingesting data to providing recommendations.

## Results

- News recommendation engine supports written Chinese and successfully analyzes complex Chinese data from news feeds.
- Cloudera allows real-time news recommendations based on each reader's historical behavior.
- The number of unique daily visitors grows by 10% and page views grow by 60%.
- The Company sees a 20% increase in the number of subscriptions.
- Increases in the unique visitors, page views, and clickthrough rates drive daily revenues from graphical and text-based advertisements to 120%.
- The number of subscriptions increases 20%, yielding a gain of 56% in monthly revenue.
- Editors review 50% more articles per day, and annual editorial costs drop 33%.

## Business drivers

The Company wanted to grow readership by providing fresh content—news, videos, and interactive games—based on each reader's personal interests—without having editors spend a substantial amount of time manually updating the news platforms with interesting topics.

They realized they needed a recommendation engine that would analyze internal historical news and video data, and target personalized recommendations to users individually. The recommendation engine would create article recommendations for each known user, presenting relevant articles that would encourage them to read more articles and spend more time on the Company's website. The solution also needed to support the logographic architecture of the Chinese language for its analytics.

With the launch of its daily news website and a mobile app a few years earlier, the Company had already become an early adopter of the digital news platform in the industry. Today, the Company's daily newspaper is the leading digital news platform in terms of page views and unique visitors from Hong Kong and Taiwan.

Nevertheless, growing and retaining audience in the face of competition for mindshare is an ongoing challenge. To keep readers interested, news articles on the top page of each news platform must be refreshed frequently, based on editorial picks, number of views, and timeliness.

The Company had been relying on manual editorial recommendations, meaning a staff of editors would spend a considerable amount of time choosing which content should be provided to readers. Despite the editorial staff's efforts, all readers were presented with the same articles on the news platforms, regardless of their interests. As a result, it was difficult to keep a reader's attention, even with the use of social media. Given the characteristics of the Company's content and reader base, traditional data analytics tools were not suitable for creating the customized recommendations they needed to keep readers engaged on the news platforms.

The Company contacted Intel because of its experience in working with large enterprises on diverse algorithms, such as content filtering. Intel's data scientists and Hadoop* engineers collaborated with the Company to come up with possible solutions to attract more page views and unique readers, and increase the time readers spent on the news platforms.

## Solution details

With the rise of digital technology, Chinese language publishers may experience the same scalability issues other growing companies encounter, but they also face unique challenges due to the logographic nature of written Chinese. Since the Company's articles are based on Chinese language content, the recommendation engine needed to be capable of handling these characteristics of the Chinese language.

Compared to most European languages, Chinese uses different methods to convey time (verb tense), count (singular/plural), and even tonality (pitch) in speech to differentiate homonymous words. In written form, Chinese is radically different. For one, it does not use a phonetic alphabet, but rather a logographic system where each symbol represents a word. To complicate matters, certain logogram combinations may form new words with different meanings, much like compound words in English (such as "greenhouse"). But Chinese does not separate word-grams with spaces as English does, so the algorithm has to check for possible combinations contextually. Grouping logograms into words is a natural language processing (NLP) technique similar to phrasal grouping in languages like English.

Unfortunately, many NLP algorithms do not work with the same efficiency in Chinese, so we had to derive language contexts—using an n-gram model—before linguistic processing.

For example, 長實牽頭發 may be broken down into the following 4-, 3-, and 2-gram possibilities:

| 4-gram | 3-gram | 2-gram |
|---|---|---|
| 長實牽頭 | 長實牽 | 長實 |
| 實牽頭發 | 實牽頭 | 實牽 |
| 牽頭發展 | 牽頭發 | 牽頭 |
| 頭發展的 | 頭發展 | 頭發 |
|  | 發展的 | 發展 |
|  |  | 展的 |

One objective of the recommendation engine (Figure 1) was to enhance the user experience by creating personalized article recommendations for each known user.

By presenting a user with relevant articles, the Company hoped the user would read more articles and thus spend more time on the site.

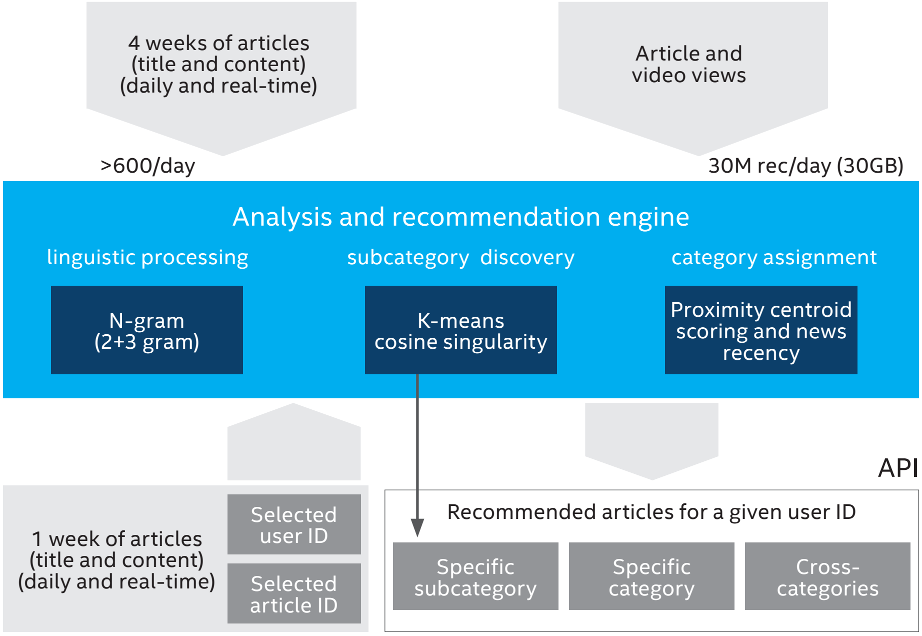

**Figure 1. Analysis recommendation engine.** We use Jaccard, Pearson, and cosine similarities to measure the similarity of articles. We also use two different metrics at various stages of the analysis to evaluate cluster fit: the Davies-Bouldin index and the silhouette coefficient. Data is stored in HBase in 12 different tables, representing the types of result sets that can be queried for: article, category, and generic.

In this solution, the Hadoop cluster ingests viewership data and article data. Articles are classified and then each user's behavior is analyzed to create a personalized model. Once the model is created, newly arriving articles can be classified and recommended to a user based on his or her profile. Recommendations are stored in HBase and exposed to web applications through a RESTful (representational state transfer) API.

There are three distinct components of this architecture:

- Learn the user's preferences.
- Provide recommendations based on learned preferences.
- Analyze the recommendations to see if they make sense.

The Company's website hosts two distinct types of news: daily and real-time. The news content and purposes are very different, so it made sense to examine each news type independently. That is why we created two completely separate models, one for daily news and one for real-time news.

## Cloudera Enterprise

The Company selected Cloudera for its efficient and cost-effective Hadoop platform. Big data solutions based on Hadoop allow publishers to provide personalized, relevant content to readers.

Intel helped the Company develop its article recommendation engine using Cloudera and Intel® Xeon® processor E5 v3 family-based servers. Data scientists from Intel helped architect, design, and build the analytics engine to meet the Company's unique requirements.

The Company can now process the large number of view histories as a basis for the news recommendation engine, which analyzes readers' views on articles to understand their behavior, deriving likes/affinities to subsets of news topics. Data is obtained from two sources:

- Daily and real-time articles from the content management system (CMS).
- The number of views that are tracked for specific articles.

Hadoop provides data scientists with options such as Mahout (a scalable machine-learning library) as well as streaming, which allows creating and running MapReduce jobs using any executable script, such as Python.

The Company's data scientists are building predictive analytical models through data mining, and producing business intelligence solutions that run on Cloudera.

## Summary

The Company continues to work with Intel to improve its article recommendation engine as well as explore other opportunities beyond providing traditional news content. With a better understanding of its readers' preferences, the Company can provide content that caters to their interests. By getting more relevant content, readers spend more time on the Company's digital platforms, which in turn gives the Company more time and added opportunities to build one-on-one relationships with its readers.

Through this increased engagement and understanding of its readers' preferences, the Company can further enhance the user experience with more responsive content and innovative products and services that evolve along with readers' changing behaviors and preferences.

Thanks to the power of Cloudera and the expertise of Intel's Big Data scientists, the Company now has a sustainable and scalable enterprise that generates business value from its massive data growth. They understand their readers' behavior better now and can attract and retain more loyal users with relevant, personalized content.

Let us help your business too.

## Spotlight on Cloudera

Cloudera is revolutionizing enterprise data management by offering a 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 1,900 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.

## 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** | | <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** | **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 email us at Hadoop-services@intel.com

Intel Big Data Professional Services:
Intel.com/bigdata/services

The results cited in this document are based on research and testing conducted by Intel and Intel's customer and are for informational purposes only.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Cost reduction scenarios described are intended as examples of how a given Intelbased product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725 or by visiting Intel's website at http://www.intel.com/design/literature.htm.