Teikoku Databank, Ltd.

Shortening the time to process billions of textual data items from several days to 30 minutes

TEIKOKU DATABANK

Overview

The need

Teikoku Databank, Ltd. owns data on millions of companies. To stay competitive, it wanted to analyze this proprietary information in combination with "big data" gathered from the Internet.

The solution

The company deployed IBM® InfoSphere® BigInsights™ software in a server cluster. Based on Apache Hadoop, InfoSphere BigInsights software enables distributed processing of large data sets.

The benefit

Now Teikoku Databank can process billions of items of textual data in 30 minutes. It is analyzing 4.75-fold more data for customers, which has improved its competitive advantage.

Teikoku Databank's history goes back more than 100 years to the establishment of Teikoku Koshinsha in 1900. Based on the corporate philosophy of "supporting economic activities and contributing to the development of society as a reliable information partner", they are developing their business in areas such as corporate credit research, credit risk management services, database services, marketing services and e-commerce support services.

Corporate credit research, which is the main business, entails "investigating the real facts of the asset status, sales results, reputations, etc. of another party when there is a commercial transaction among companies"—for example, an investigation that clarifies various types of business information. At an early stage, they introduced computers into this corporate credit research to improve business efficiency and customer service.

Teikoku Databank was the first in the industry to introduce computers, back in 1968. The primary aim was to streamline business operations, such as for financial processing, but they were also trying to find ways to use the accumulated corporate data in different forms. As a result, in 1972 they developed the "COSMOS1" corporate financial database. In 1974 they also began offering "COSMOS2," which put corporate profiles into the database.

"We introduced IBM InfoSphere BigInsights into the server cluster for data analysis and performed distributed processing of big data. It has become possible to process billions of items of textual data in 30 minutes."

—Mr. Satoshi Kitajima, MBA Statistician, SPECIA Team, Business Analytics Division, Market and Business Intelligence Department, Teikoku Databank, Ltd.

Mr. Satoshi Kitajima, MBA Statistician, SPECIA Team, Business Analytics Division, Market and Business Intelligence Department, Teikoku Databank, Ltd.

They presented the "COSMOSNET" online service in 1988 and the "ATTACK" integrated marketing service in 1990. In 1999 they made their first real entry into the e-commerce (EC) support business and developed "COSMOSNET/EC." In these ways, they are always developing leading-edge services and presenting them to customers.

Using big data to improve quantity and quality of information

Teikoku Databank has been providing their customers with reliable corporate information based on their credit research for more than 100 years, and owns a huge amount of corporate data such as corporate credit report files of 1.6 million companies, "COSMOS1" financial statements of 4.4 million terms worth of information gathered from 680,000 companies, a "COSMOS2" corporate profile database of 1.42 million companies, and other corporate data for 4.1 million companies.[1] Recently, however, information published on the Internet has been starting to have a significant effect on company business, so responding to this situation has become an urgent task.

Mr. Satoshi Kitajima, an MBA Statistician in the SPECIA Team of the Business Analytics Division of the Market and Business Intelligence Department, explains: "We have detailed knowledge of corporate information that has been investigated and combined in order to answer questions such as 'That company is what kind of company?' and 'How much can they be trusted?' We haven't kept information similar to what is found on the Internet, about the kinds of products owned or how products are evaluated in the market. However, customers have been asking us to present corporate information that includes such Internet information."

Mr. Kitajima adds: "Recently, in order to understand the business situation, attention has been focused on the 'big data' found on the Internet. We knew that, in the future, big data would be extremely important as information for analyzing business situations. However, with just our existing technology it was difficult to respond to the customers' needs, so it became important for us to introduce new technology in order to quickly extract from big data the business data that met the customers' needs."

Solution components

Software
- IBM® InfoSphere® BigInsights™

Also, in terms of the amount of data, although they have corporate credit reports for 1.6 million companies, that was sufficient only for presenting business information, and they had to improve the quality of the data as the amount of data grew for each company into the future.

Mr. Kitajima says: "By using big data, it is possible to use many keywords, such as product names, and it is possible to improve the accuracy of the analysis. As large amounts of individual information are accumulated, the 'individual' characteristics that exist within the data that were hard to understand in the past, can become clear. Our goal was to present 'only-one' corporate information by using 'collective intelligence' that had not previously existed in the world."

As the first step toward solving this problem, Teikoku Databank decided to combine its existing corporate information with the big data on the Internet and develop a new extracting service named "List of Companies in Specific Industry" to be used in fields such as marketing. For the technological foundation they used IBM InfoSphere BigInsights software, which is software for analyzing big data that makes use of distributed processing functions based on Apache Hadoop.

Evaluation of IBM support and results, and use of InfoSphere BigInsights

The project to create a service for using big data began in November 2011. Through the use of a system architecture that introduced InfoSphere BigInsights software, verification within the company began in May 2012. The company data extracting service began in November 2012. Now the company data extracting service comprises a server cluster for crawling Internet data, a data analysis server cluster that uses InfoSphere BigInsights software and a server cluster for product manufacturing.

Mr. Kitajima says: "The 'List of Companies in Specific Industry' service presents a pull-down menu including things such as 'Data Center Management Businesses,' 'Caregiver Services' and 'Healthy Foods.' Previously, keywords were extracted from corporate credit reports that

"We investigated the products of multiple vendors, but our evaluation of IBM's technical support was the big reason for choosing IBM InfoSphere BigInsights."

—Mr. Kengo Sawayama, System Planning Section, Corporate Planning Department, Teikoku Databank, Ltd.

Mr. Kengo Sawayama, System Planning Section, Corporate Planning Department, Teikoku Databank, Ltd.

were based on our unique 'Visit and Confirm Onsite Investigation' method, but by adding to this the information obtained by crawling the Internet, we have increased the amount of data used for creating the presentations by a factor of 4.75. Thus, the number of data items that can be presented has also been greatly expanded. Also, by combining this with our company's existing database, it is possible to perform compound searches that would not be possible with just Internet information such as sales figures, number of employees, age of the president, designated bank transactions and building floor space. Internet information contains a lot of information that is not directly related to the business, or is simply a rumor, or is unverified, but by using a system of machine learning, it was possible to extract more detailed and more accurate corporate information to match the meaning of the search keywords."

With regard to the implementation of the "List of Companies in Specific Industry" service, Mr. Kengo Sawayama of the System Planning Section of the Corporate Planning Department says, "Until now, development concentrated on business systems that used internal corporate data. Since a corporate data search service uses a lot of company-external data, distributed processing based on Hadoop is effective. However, since we had insufficient experience or know-how concerning Hadoop, we decided to use BigInsights."

With regard to the reason to choose InfoSphere BigInsights software, Mr. Sawayama adds, "We investigated the products of multiple vendors, but our evaluation of IBM's technical support was the big reason for choosing IBM InfoSphere BigInsights. An additional reason for the decision was the fact that many companies both within Japan and abroad have already used it."

"By implementing a service that uses the highly reliable corporate information … that we had previously cultivated, as well as the big data that is rapidly processed with BigInsights, we achieved strong differentiation from our competitors."

—Mr. Satoshi Kitajima

With regard to the circumstances that led to the success of this project, Mr. Kitajima says: "In order to quickly respond to customers' needs, we decided not to outsource the development but rather do it internally. For development, it is important to have team cooperation among the user departments and system departments, and we use agile development so that the desired data can be obtained by the people who want it. First, we set up a system where the knowledgeable people create programs in the Ruby language, the content is understood by the team and then anyone can respond. We are not a group of specialists with a lot of experience, so we are following this approach while working on personnel development."

Processing billions of textual data items in 30 minutes

For the system structure, initially they tried crawling the Internet and processing the information using a single server without parallelization, but it was slow and unusable, not completing even after several days. A test calculation indicated that it would take several more weeks, with some of the processing stopping in the middle. When InfoSphere BigInsights software was introduced into the same server, the processing speed was greatly increased.

Mr. Kitajima says, "In the initial stages, when we introduced BigInsights into a single server and operated in Hadoop quasi-dispersion mode, the same processing was completed in several days. After we had verified the effectiveness of distributed processing based on Hadoop, we introduced IBM InfoSphere BigInsights into the server cluster for data analysis and performed distributed processing of big data. Now it has become possible to process billions of items of textual data in 30 minutes."

"We certainly want to aggressively develop methods of using more corporate data. ... We also expect BigInsights to have a big role to play in the future use of big data."

—Mr. Satoshi Kitajima

With regard to the effect from the viewpoint of the system architecture, Mr. Sawayama says: "At first, some things about BigInsights were unfamiliar, but there was support from IBM, and we soon were able to use it. Thanks to IBM, it was possible to build the system in a short time."

Mr. Sawayama adds: "The GUI tools for system management provided by BigInsights were also useful. Usually, to verify the HDFS directory structure on Hadoop it is necessary to enter and display commands, but in the case of three layers even entering the commands is hard work. With BigInsights, in the File Manager Format GUI tool, it is possible to display the directory as a tree structure, so it is very convenient."

And with regard to the plan that IBM presented, Mr. Sawayama says: "In the plans of other companies, the only way they tried to improve performance was by increasing the number of servers. On the other hand, IBM's plan was based on results that made use of the characteristics of Hadoop, where CPU performance, such as the number of cores, is more important than the number of servers. With this plan we could tell there was a difference in technical strength."

With regard to the business effect after implementing the corporate information search service that introduced InfoSphere BigInsights software, Mr. Kitajima says, "By implementing a service that uses both the highly reliable corporate information of about 1.6 million companies that we had previously cultivated, as well as the big data that is rapidly processed with BigInsights, we achieved strong differentiation from our competitors."

Continuing to aggressively take advantage of all kinds of data

As for Teikoku Databank's future plans, Mr. Kitajima says: "We certainly want to aggressively develop methods of using more corporate data, upon which our business is based. For example, there is the analysis of social network data, customer sales data and Web access log data. Companies want to use the big data that they own, but we know that there are many companies that don't know where to start. In such cases, by incorporating the corporate data that we have, for example, they could cover the range from simple statistics to complicated analyses. We also expect BigInsights to have a big role to play in the future use of big data."

For more information

To learn more about IBM InfoSphere BigInsights, please contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/software/data/infosphere/biginsights

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
March 2013

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] Figures as of November 2012

Please Recycle

IMC14806-USEN-00