Hitachi Open Middleware for Big Data Processing

Jun Yoshida
Nobuo Kawamura
Kazunori Tamura
Kazuhiko Watanabe

OVERVIEW: The quantity of data being handled by corporations is growing rapidly and the ability to utilize this "big data" effectively will be one of the keys to future corporate development. Examples include the use of real-time data such using sensor data to detect abnormalities in plant and machinery, batch-processing-style trend analysis such as the use of sensor data collected over a long period to conduct failure analysis of plant and machinery, and improving the speed of daily batch processing such as calculating sales and order data totals or nightly batch processing of database updates. In response to these needs, Hitachi supplies platforms such as stream data processing middleware and parallel and distributed processing middleware to support the processing of big data. With data volumes expected to continue expanding rapidly, Hitachi is also working with The University of Tokyo, etc. on the research and development project of an ultra-high-speed database.

INTRODUCTION

WE have entered an era of "information explosion," in which the quantity of data handled by corporations is growing rapidly in response to factors such as advances in sensor technology and the widespread adoption of broadband and portable devices. For example, making effective use of large quantities of access log or sensor data, in ways that can create new businesses is one of the keys to future corporate development. In existing systems, the time required to process batch jobs is becoming longer as the quantity of data increases, and this is starting to impinge on the time available for other services. Given this situation, new business value can be unlocked by processing a batch job quickly that previously took several days to execute.

Hitachi is working on research and development of these open middleware technologies to facilitate the efficient processing of big data.

This article gives an overview of big data processing and Hitachi's open middleware, which is evolving to meet its requirements.

OPEN MIDDLEWARE FOR BIG DATA PROCESSING

Challenges to be Overcome by Big Data Processing

The technologies required for big data processing can be broadly divided into the following two categories, which represent challenges to be overcome:

(1) Real-time processing

Technologies for processing an endless flow of big data such as geospatial information services or the use of sensor data to detect abnormalities in plant and machinery

(2) Faster batch and tabulation processing

Technologies for improving processing speed to avoid lengthening execution times as data volumes rise when performing daily batch processing, such as calculating sales and order totals or nightly batch processing of database updates.

To overcome these two challenges, progress is being made on software technologies that take advantage of developments in hardware. Stream data processing middleware is one example of a technology for achieving real-time processing (1). Taking advantage of the improved performance and decreasing cost of memory, this technology can achieve high-speed near-real-time performance by processing big data in memory. For faster batch and tabulation processing (2), meanwhile, a parallel and distributed processing middleware technology that has attracted attention in recent years is the open source Hadoop[*1] software. Hadoop takes advantage of the improved performance and lower cost of IA (Intel[*2] architecture) servers to speed up batch processing by operating a large array of these servers in parallel.

Prospects and Challenges for Hadoop

Hadoop allows high-speed batch processing to be performed with ease, without requiring operators to

*1 Hadoop is a trademark of the Apache Software Foundation.
*2 Intel is a trademark of Intel Corporation in the U.S. and/or other countries.

be concerned about such things as the complexities of parallel and distributed processing or how data is allocated. Hadoop is an open source program with good future prospects, and applications for it are being sought on corporate systems around the world.

In one typical example, it is used to generate product recommendations from web access logs at a consumer web site to help motivate customer purchases. Beyond web access logs, other potential applications include the use of sensor data collected over a long period to conduct failure analyses of plant and machinery, or statistical analyses of geospatial information.

However, while Hadoop can simplify the process of implementing high-speed batch processing, its applications are limited. For example, it cannot run existing batch programs written in languages such as COBOL (Common Business Oriented Language) and programs need to be rewritten to suit the Hadoop processing model. Other problems include a lack of flexibility in areas such as how data allocation is performed, and an inability to give a reliable indication of how long batch processing will take.

Supply of Open Middleware and Associated Services

Hitachi supplies a range of open middleware products for big data processing (see Fig. 1).

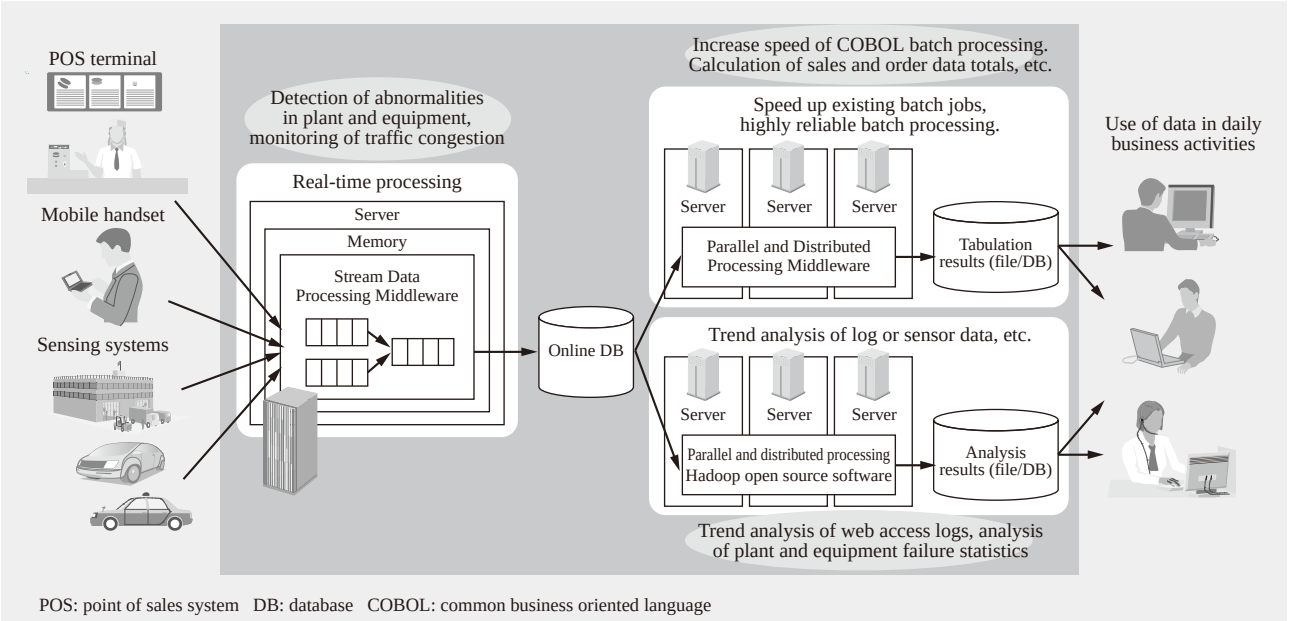
For real-time processing, it offers the stream data processing middleware.

For faster batch and tabulation processing, Hitachi offers a support service for the open source Hadoop software. However, Hadoop does not suit all applications and the parallel and distributed processing middleware is available to cover tasks for which Hadoop is unsuited. Features of the parallel and distributed processing middleware include the ease with which existing batch processing can be migrated, flexibility in how data is allocated, and reliable batch processing completion times.

STREAM DATA PROCESSING MIDDLEWARE

Features

The stream data processing middleware is a middleware package for the real-time, in-memory processing of continuous flows of big data as it is generated. Tabulating and analyzing large quantities of data at high speed allows the timely detection of "abnormal situations." The middleware uses scripts written in CQL (continuous query language), an extension of the SQL (structured query language) commonly used with databases, to provide an easy way to specify tabulation and analysis scenario definitions. This means that users familiar with SQL will have no trouble producing scenario definitions.



POS: point of sales system DB: database COBOL: common business oriented language

Fig. 1—Concepts Used in Big Data Processing Technologies for Business Systems.
Potential applications include using the stream data processing middleware for real-time processing such as detection of abnormalities in plant and equipment, using the parallel and distributed processing middleware for highly reliable batch processing such as calculating sales and order data totals, and using the open source Hadoop software for tasks such as trend analysis of web access logs.

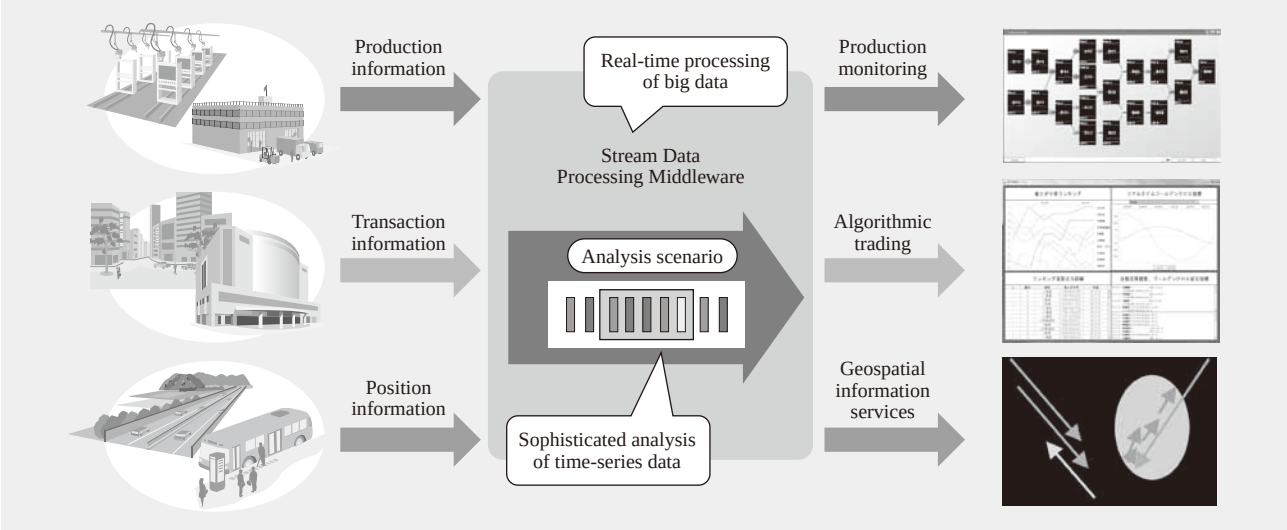


Fig. 2—Overview of Stream Data Processing Middleware.
The continuous flow of big data is processed in realtime as it is generated. This allows applications such as geospatial information services that perform real-time analysis of position information from objects. Monitoring conditions are easily defined as scenarios using CQL (continuous query language).

Example Applications

It is anticipated that the stream data processing middleware will be used for real-time processing in a wide range of different fields. In addition to the detection of abnormalities in plant and equipment and other maintenance services, potential examples include implementing corporate compliance requirements such as preventing unauthorized web access, algorithmic trading in which buy and sell orders are generated automatically based on an analysis of factors such as stock prices and turnover, and recommendation services that use position data from GPS (global positioning system) terminals (see Fig. 2).

The stream data processing middleware has already been chosen by a Japanese exchange to provide a service for publishing market indices, where it provides a high-speed distribution service with world-leading levels of performance. The system is able to track fluctuations in the prices of the stocks that make up an index, and then update and distribute it at a pitch in the order of milliseconds (compared to seconds previously).

Examples of applications that take advantage of the advanced capabilities of the stream data processing middleware for analyzing time-series data are also increasing. One such is proactive preventive maintenance of IT systems, which are becoming larger and more complex due to advances such as virtualization and cloud computing. This involves analyzing the large quantities of log data collected by these systems to look for correlations and other trends so that potential faults can be identified and prevented before they occur.

PARALLEL AND DISTRIBUTED PROCESSING MIDDLEWARE

Features

As existing batch jobs at corporations have become black boxes, there are risks associated with making any changes. The parallel and distributed processing middleware is a middleware package that can reuse existing batch jobs and distribute them across multiple servers for faster parallel execution. Because a number of servers are used, if a fault occurs on one of them, its processing can be shifted to another server. Minimizing the scope of faults in this way significantly reduces recovery time (see Fig. 3).

In addition to preventing any impact on other jobs due to “overruns” (tabulation jobs run as nightly batch processing taking longer than scheduled), the middleware can also ensure that scheduled execution times are achieved even when data quantities increase in the future due to business growth.

Example Applications

In addition to preventing overruns, the ways in which the parallel and distributed processing middleware is used include being able to create new tasks by shortening execution time compared to the past. One example is the tabulation of data from POS (point of sale) systems, which are normally run as daily batch processing to calculate sales totals. Making it possible to perform tabulation and analysis at one-hour intervals instead of overnight speeds up decision making, such as product procurement and placement.

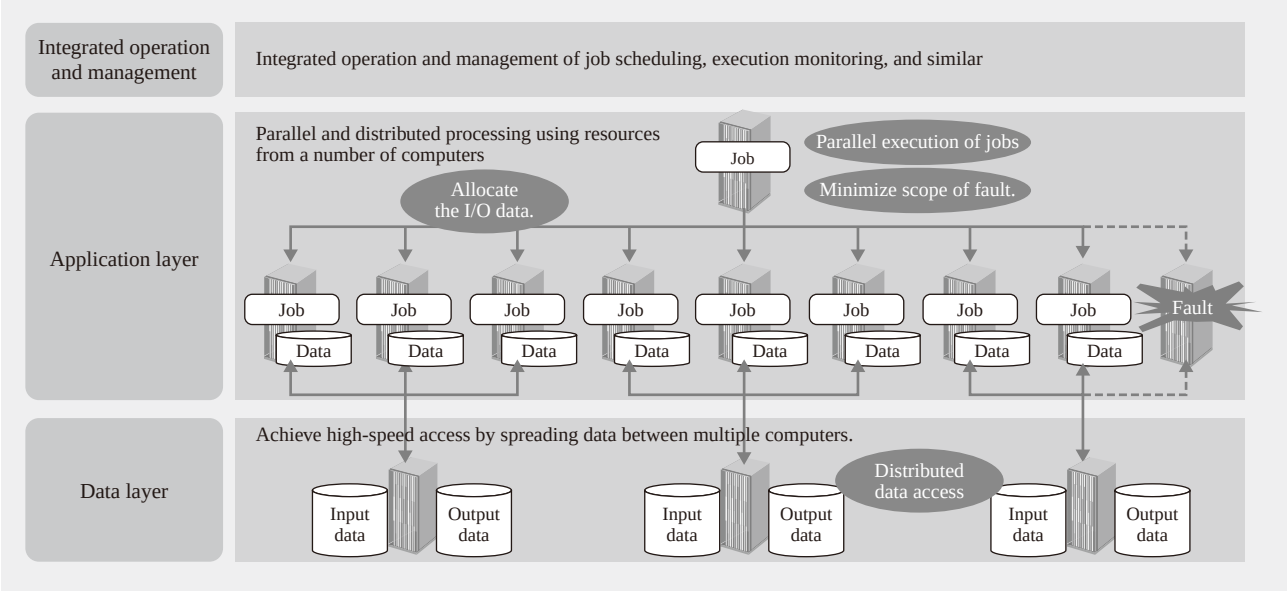


Fig. 3—Overview of Parallel and Distributed Processing Middleware.
The parallel and distributed processing middleware uses parallel and distributed processing to speed up batch processing. Features include the ease with which existing batch processing can be migrated, flexibility in how data is allocated, and reliable batch processing completion times.

Improving the speed and reliability of all aspects of batch jobs delivers major advantages in mission-critical fields, such as databases for systems that handle big data, situations in which it is desirable to complete tabulation processing within a designated time, and settlement and bank transfer applications in the finance industry that require strict exclusive control.

The parallel and distributed processing middleware is also recognized as an ideal solution for speeding up batch jobs that use existing COBOL programs, many of which remain in use in core business systems. It has been demonstrated that the level of program modification required to when migrating COBOL programs to the parallel and distributed processing middleware environment is only 1%.

FUTURE-ORIENTED RESEARCH AND DEVELOPMENT: ULTRA-HIGH-SPEED DATABASE

Large petabyte-class databases will be required if data quantities continue to increase rapidly in the future. Unfortunately, existing commercial databases take a long time to process such large amounts of data and are getting toward the point where they can no longer keep up. Accordingly, Hitachi, Ltd. is working with The University of Tokyo, etc. on the research and development project entitled “Development of the Ultra-high-speed Database Engine for the Era of Super-large Databases and Demonstration & Evaluation of Strategic Social Services with This Database Engine at Their Heart,” which is supported by the Funding Program for World-leading Innovative R&D on Science and Technology (FIRST).

This project is developing an ultra-high-speed database engine that operates on a new principle called “the out-of-order execution principle” that was devised by The University of Tokyo. The aim is to enhance industrial competitiveness and help provide safety and peace of mind, with potential applications including the development of products that are tailored to specific needs based on an understanding of the customer’s lifestyle and life stage, or to assist with quality assurance and more efficient inventory management through the use of traceability in production and distribution.

CONCLUSIONS

This article has given an overview of big data processing and Hitachi’s open middleware, which is evolving to meet its requirements.

In addition to supplying maintenance support and products designed to use technologies, including those for stream data processing and parallel and distributed processing, Hitachi is a one-stop supplier of a range of different solutions. One of these is the Big Data Distributed Processing Assessment Service.

To get the best out of big data in business, it is necessary to deepen one’s understanding of

each of these technologies, and to pay attention to effectiveness evaluations when making decisions on which technology to use. Accordingly, Hitachi also offers a consulting service on how to analyze and utilize big data, and on the use of the PaaS (platform as a service) products of Hitachi Cloud Computing Solutions to provide systems on which the stream data processing middleware, the parallel and distributed processing middleware, or Hadoop are already configured in order to offer a testing support service that can assist customers in testing on an actual system in preparation for installation.

Hitachi also intends to continue collaborating with The University of Tokyo on the development of an ultra-high-speed database engine, and to continue developing open middleware with the aim of processing big data efficiently and applying it in business.

REFERENCES

(1) Distributed Processing of Big Data, http://www.hitachi.co.jp/Prod/comp/soft1/big_data/ in Japanese.
(2) A. Arasu et al., "STREAM: The Stanford Stream Data Manager," IEEE Data Engineering Bulletin **26**, No. 1 (Mar. 2003).
(3) Welcome to Apache Hadoop!, http://hadoop.apache.org/
(4) "Development of the Fastest Database Engine for the Era of Very Large Database and Experiment and Evaluation of Strategic Social Services Enabled by the Database Engine," http://www.tkl.iis.u-tokyo.ac.jp/FIRST/index.html in Japanese.
(5) M. Kitsuregawa et al., "Vision and Preliminary Experiments for Out-of-Order Database Engine (OoODE)," DBSJ Journal **8**, No. 1 (Jun. 2009) in Japanese.

ABOUT THE AUTHORS

Jun Yoshida
Joined Hitachi, Ltd. in 1998, and now works at the Business and Technology Emerging Business Laboratory, Software Division, Information & Telecommunication Systems Company. He is currently engaged in market development of Big Data business.

Nobuo Kawamura
Joined Hitachi, Ltd. in 1981, and now works at the Innovative R&D Project Center, Software Division, Information & Telecommunication Systems Company. He is currently engaged in the development of the ultra-high-speed database engine for the era of super-large databases through The University of Tokyo FIRST Program. Mr. Kawamura is a member of the Information Processing Society of Japan (IPSJ).

Kazunori Tamura
Joined Hitachi, Ltd. in 1991, and now works at the Network & Transaction Department, Software Division, Information & Telecommunication Systems Company. He is currently engaged in market development of the Stream Data Platform.

Kazuhiko Watanabe
Joined Hitachi, Ltd. in 1988, and now works at the Operating System Department, Software Division, Information & Telecommunication Systems Company. He is currently engaged in the parallel batch job project.