Ultrahigh-Speed Data Analysis Platform "InfoFrame DWH Appliance"

KOCHU Daisuke, MIKUNI Yukinori

Abstract

With the aim of achieving timely extraction of valuable information, the need for *ad hoc* analysis of big data is increasing rapidly. This paper describes the reason that the InfoFrame DWH (Data WareHouse) Appliance is attracting attention as a high-speed analysis platform and introduces actual case studies of its use in marketing by *ad hoc* analyses of website access logs.

Keywords

big data, DWH Appliance, information utilization, ad hoc analysis, web access log

1. Introduction

Extraction of valuable information from big data requires not only the traditional analytic approach but also an *ad hoc* analytic approach that can appraise data from different angles. In this paper, we describe how the ultrahigh-speed data analysis platform "InfoFrame DWH Appliance" implements the requisite high speed and simplicity for the *ad hoc* analysis of big data. We will also introduce actual case studies of *ad hoc* analyses of the access logs of websites by introducing InfoFrame DWH Appliance.

2. Analytic Approach to Big Data

Recent progress in the fields of networks and devices and the activation of user community services such as SNS have significantly increased the volume of generated data. Collection, storage and analysis of big data that has previously not been manageable are now becoming possible. However, commercial utilization is still at an early stage in which methodology is still developing.

Extraction of valuable information from big data requires an *ad hoc* analytic approach that regards data from different angles rather than the traditional analytic approach as represented by sales and customer analyses.

For the sales and customer analyses methods traditionally adopted by many enterprises, the types of data to be collected and the indices for analysis were already arranged systematically so that there are few issues as to how to advance analysis. On the other hand, since the methodology for extracting valuable information to support planning and decision making as an enterprise strategy and for creating new services using big data is not yet established, the analysis should be advanced carefully. In other words, the analytic approach to big data is mainly via " *ad hoc* analysis," which is advanced through trial and error of analysis patterns that are not assumed in advance.

3. Big Data and *ad hoc* Analysis

Big data has the three properties of "large amount, wide diversity and high frequency" and consequently there are three requirements for *ad hoc* analysis.

(1) High speed
High-speed processing of data of large amounts of some tens of terabytes (TB) to a few petabytes (PB) is necessary.

(2) Simplicity
The ability to input, combine and analyze a large diversity of data is necessary.

(3) Promptness
The ability to analyze frequently generated data in a timely manner is essential.

Of these three requirements, the one that necessitates special consideration is "simplicity." With regard to big data analysis being advanced carefully at the present time due to the absence of an established methodology, it is difficult to assume the data types to be combined for analysis in advance, so it is necessary to analyze all data types as the analysis targets. It is because many data types have to be handled in the analysis of big data that "simplicity" is critical for collecting,

combining and analyzing various kinds of data.

4. High–Speed Analysis Platform for *ad hoc* Analysis

NEC offers InfoFrame DWH Appliance as the platform for high-speed, simple *ad hoc* analysis of big data. It is an optimized data warehouse appliance product that combines NEC’s high-performance and high-reliability server/storage systems with Netezza architecture and software capable of advanced data search/analysis (**Fig. 1**).

In the following subsections, we introduce details of how the InfoFrame DWH Appliance implements the high speed and simplicity required for *ad hoc* analysis of big data.

4.1 High Speed of InfoFrame DWH Appliance

In the processing of big data at a scale of some tens of TBs to a few PBs, what is important is “to prevent the data as far as possible from being migrated.” Unnecessary migration of big data leads to I/O bottlenecks and excessive processing in the CPU or memory, which will eventually exert serious effects on the system performance.

InfoFrame DWH Appliance filters data (elimination of unnecessary data at the same time as data is read from a disk, etc.) at as early a stage as possible in data streaming by using a universal device called the FPGA (Field Programmable Gate Array). The data processing in such a way in the proximity of a disk containing data makes it possible to avoid unnecessary migration of big data (**Fig. 2**).

For high-speed processing of big data, the InfoFrame DWH Appliance adopts the MPP (Massively Parallel Processing) architecture. In this architecture, symmetrical multi-processing hosts are connected to the grids of the MPP nodes in order to utilize the MPP grids for dealing with heavy loads such as during data search and analysis. The InfoFrame DWH Appliance ZA100 model is capable of processing a query with 92-processor parallel processing. The parallel execution of big data using the MPP technology allows the InfoFrame DWH Appliance to manifest a high performance in big data handling (**Fig. 3**).

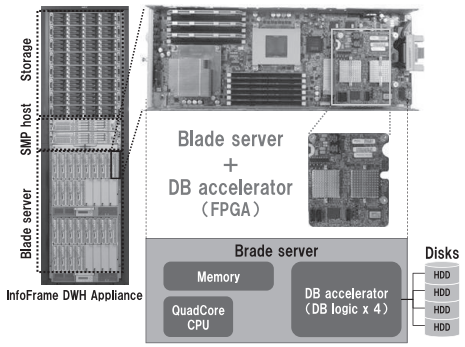

Fig. 1 Configuration of InfoFrame DWH Appliance.

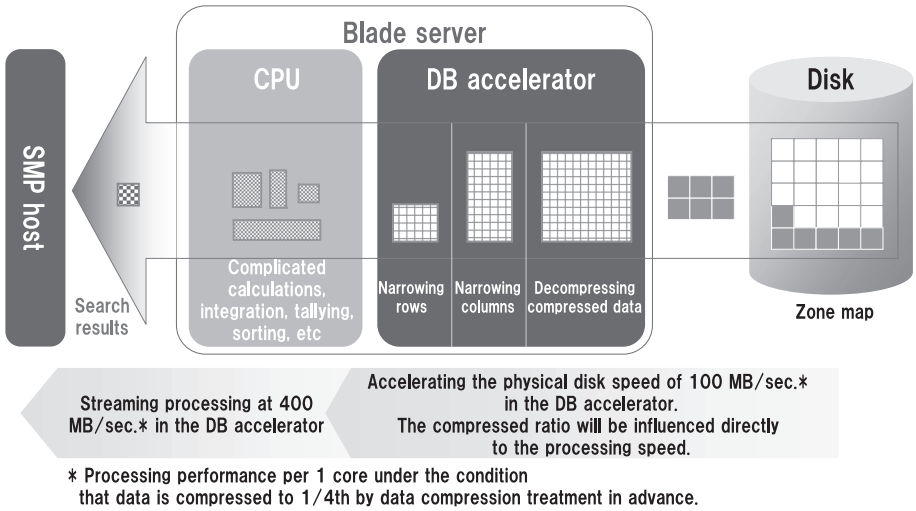

Fig. 2 Data streaming processing by FPGA.

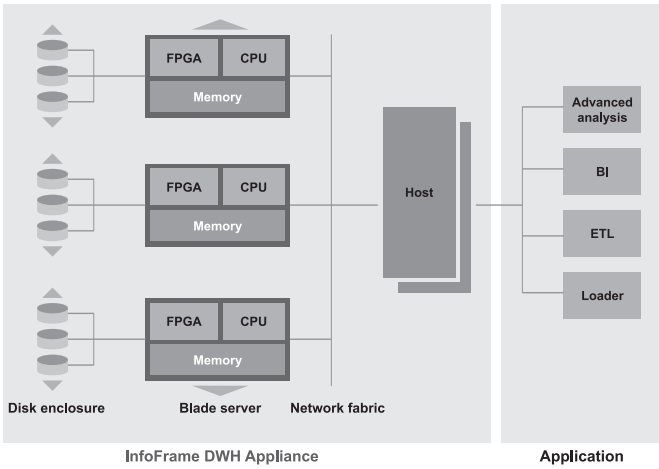

Fig. 3 MPP architecture of InfoFrame DWH Appliance.

4.2 Simplicity of InfoFrame DWH Appliance

InfoFrame DWH Appliance is designed with emphasis on "simplicity" aiming at a "simple appliance configuration." The data loaded in InfoFrame DWH Appliance is distributed automatically before being stored, without intervention of the database administrator. A query input to InfoFrame DWH Appliance is processed automatically at high speed in the MPP grids. This is possible almost without the advance preparation required in ordinary DB technology, such as database tuning (indexing, partitioning, etc.).

The "simplicity" of InfoFrame DWH Appliance manifests significant effects in the *ad hoc* analysis of big data. In the big data analysis, which sometimes involves combination of hundreds of data types, it is unrealistic to perform database tuning each time that a new data type is added. Hence the "simplicity" for which no tuning is required, even when a new data type is added. The data can then be used immediately after it is loaded in analyses. This procedure is critical for the *ad hoc* analysis of big data.

5. Actual Case Studies of Big Data Utilization Using a High–Speed Analysis Platform

Attempts at *ad hoc* analyses of big data have been performed successfully by enterprises in the network/communication industries aiming at achieving greater success for their businesses. In this section, we introduce actual case studies of customers who have introduced our InfoFrame DWH Appliance in order to enable prompt and smooth analyses of large amounts of stored web access logs and have utilized the results in marketing and for website improvement.

5.1 Usage of Web Access Logs

The web access log is a typical target of big data utilization that is currently at the development stage and it is actually already in use in various business domains. For example, some content planners/developers read the browsing situations (behaviors and actions in websites) of users from the access logs and analyze them in order to improve their websites and develop new services. Some of them also analyze the browsing histories of registered users and use the resulting data in recommendations. On the other hand, sales engineers analyze website access trends to use the results in formulating proposals for promotions and to encourage customer acquisitions.

5.2 *ad hoc* Analyses of Web Access Logs

According to the opinions of our customers who actually execute web access log analyses, the analysis of web access logs is tending to become *ad hoc* rather than typically traditional. For example, when performing an analysis by limiting the target users to those who have utilized specific services or who have browsed specific promotion pages, it is necessary to define the boundaries before each analysis. In order to identify the behaviors and trends of users in detail, it is necessary to combine and analyze several types of data in addition to the web access log data. As seen here, it is because the contents and data of analysis differ depending on the promotional theme or the purpose of planning that the analysis becomes *ad hoc* .

5.3 Web Access Log Analysis Using InfoFrame DWH Appliance

InfoFrame DWH Appliance also manifests high performance in the *ad hoc* analysis of a large volume of web access logs. At NEC, we provide a POC (Proof of Concept) program for customers who are considering introducing InfoFrame DWH Appliance. This provision is intended to let the customers confirm the "high speed" and simplicity" of InfoFrame DWH Appliance when using their own actual data.

Here is an example in which we have taken the web access

log of a customer and applied a POC program to it.

(1) With the previous system, the extraction/aggregation processing of about 1.6 million items of log data including the shop data, member data and contents took about 3-1/2 hours. This compares with the performance of the InfoFrame DWH Appliance that completed the same amount of processing in only 11 seconds.

(2) We verified the fuzzy search of a web access log with 350 million items (20 GB) using the InfoFrame DWH Appliance, which completed it in 3 seconds.

Our verifications of a few patterns of web access log analyses demonstrated a performance 576 times greater on average compared to those of the previous system, or even as great as about 1,000 times at maximum. Through the POC program, we were able to have the customers highly evaluate the InfoFrame DWH Appliance for its lack of need for special tuning and its minimal operational and management requirements. These results confirm that the introduction of the InfoFrame DWH Appliance makes it possible to provide the results of timely web access log analyses for the persons in charge of content planning/development and sales.

6. Conclusion

As we mentioned in the introduction of actual case studies of web access log analyses, one of the major characteristics of the analysis of big data is that it should adopt an *ad hoc* analytical approach. In the analysis of big data which should be advanced cautiously due to the absence of established methodology, the “high speed” permitting several trials and errors and the “simplicity” of inputting and combining a large diversity of data are important requirements. We have more than 50 corporate customers that utilize InfoFrame DWH Appliance (Netezza) as a solution that is capable of the *ad hoc* processing of big data by meeting both of the above requirements. In the future, too, we will continue our endeavors so that this solution is utilized as an analysis platform for use in extracting information from big data. This policy will serve the planning and decision making of enterprise strategy and help create new services using big data.

*Netezza is a registered trademark or trademark of International Business Machines Corporation.

Authors' Profiles

KOCHU Daisuke
Assistant Manager
IT Platform Solutions Division
IT Software Operations Unit

MIKUNI Yukinori
Group Manager
IT Platform Solutions Division
IT Software Operations Unit

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.2 Big Data

Vol.7 No.2
September, 2012

Special Issue TOP

Remarks for Special Issue on Big Data
NEC IT Infrastructure Transforms Big Data into New Value

◇ Papers for Special Issue

Big data processing platforms
Ultrahigh-Speed Data Analysis Platform “InfoFrame DWH Appliance”
UNIVERGE PF Series: Controlling Communication Flow with SDN Technology
InfoFrame Table Access Method for Real-Time Processing of Big Data
InfoFrame DataBooster for High-speed Processing of Big Data
“InfoFrame Relational Store,” a New Scale-Out Database for Big Data
Express5800/Scalable HA Server Achieving High Reliability and Scalability
OSS Hadoop Use in Big Data Processing

Big data processing infrastructure
Large-Capacity, High-Reliability Grid Storage: iStorage HS Series (HYDRAstor)

Data analysis platforms
“Information Assessment System” Supporting the Organization and Utilization of Data Stored on File Servers
Extremely-Large-Scale Biometric Authentication System - Its Practical Implementation
MasterScope: Features and Experimental Applications of System Invariant Analysis Technology

Information collection platforms
M2M and Big Data to Realize the Smart City
Development of Ultrahigh-Sensitivity Vibration Sensor Technology for Minute Vibration Detection, Its Applications

Advanced technologies to support big data processing
Key-Value Store “MD-HBase” Enables Multi-Dimensional Range Queries
Example-based Super Resolution to Achieve Fine Magnification of Low-Resolution Images
Text Analysis Technology for Big Data Utilization
The Most Advanced Data Mining of the Big Data Era
Scalable Processing of Geo-tagged Data in the Cloud
Blockmon: Flexible and High-Performance Big Data Stream Analytics Platform and its Use Cases

◇ General Papers

“A Community Development Support System” Using Digital Terrestrial TV