CASE STUDY

Intel® Xeon® Processor E5 Family
Big Data, Business Intelligence
Entertainment and Media

Provide Recommendations to Online Users with Big Data. Analysis of the Infrastructure That Uses Apache Spark*

Develop data analysis infrastructure with implemented recommendation feature/ behavior analysis from Apache Spark* on a server based on Intel® Xeon® Processor E5 Family.

DMM.com Labo

DMM.com Labo Co., Ltd.
Headquarters: Ebisu Garden Place Tower 14F, 4-20-3 Ebisu Shibuya-ku, Tokyo
Founded: April 3, 2000
Capital: 30 million yen
Business description: System development and operation, providing network infrastructure, web marketing
http://labo.dmm.com/

Issues

- **Customized support for recommendations and behavior analysis:** Recommendations and behavior analysis were handled individually for each division, leading to uneven accuracy.
- **Operational framework for recommendations:** Some divisions handed over data to outside companies for analysis, but customized fine tuning was difficult with outsourcing.

Solutions

- **Recommendation feature and behavior analysis feature from Apache Spark*:** Develop infrastructure that implements recommendation feature and behavior analysis feature from Apache Spark* on a server with built-in Intel® Xeon® Processor E5 Family.
- **Cloudera Enterprise:** Introduce platform for developing integrated system infrastructure that makes use of Apache Spark* and Apache Kafka*.

Benefits

- **Release of commercial service with implemented recommendation feature:** Based on the purchase history from multiple services, it is now possible to recommend similar and related products with high accuracy.
- **Achieving real-time analysis:** By importing large amounts of log data in real time, it is possible to obtain values in real time.

Successfully insource recommendation feature and make behavior analysis real-time

The DMM.com group is involved in a variety of businesses including video streaming of movies, anime, etc. e-commerce, rentals, online games, FX trading, 3D printers, English conversation classes, robots, solar panels, etc. One part of this group, the corporation DMM.com Labo (henceforth, DMM.com Labo), is involved in system development, operation, and web marketing for the whole group. In order to respond to customer needs which have diversified due to expansion into more business fields, the company has set itself to work on creating a recommendation infrastructure that uses big data to analyze user characteristics and introduce users to recommended products. Building an analysis infrastructure that utilizes the high-speed analysis engine Apache Spark* (henceforth, Spark*), using Cloudera's "Cloudera Enterprise" as processing infrastructure for big data, a recommendation feature and real-time behavior analysis used by multiple services of the group was implemented. The Intel® Xeon® Processor E5 Family was adopted for use in hardware that comprises Spark's* infrastructure, supporting high-speed analysis processing that makes use of big data.

The Intel® Xeon® Processor E5 Family geared towards multi-core, multi-thread computing and contributing to parallel distributed processing by Spark*

"While aiming to implement recommendations that use Apache Spark*, we were able to increase I/O performance through hardware cache that uses SSD."

– DMM.com Labo Co., Ltd.
System Headquarters
CTO Office Chief
Norio Ojima

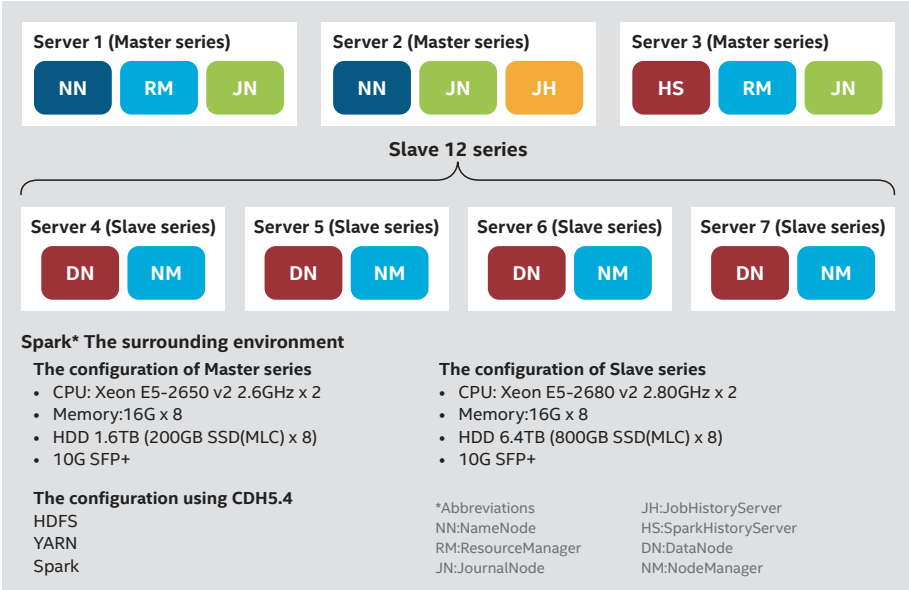
Improving predictive accuracy with customer and market analysis

The DMM.com group reached sales of 98.1 trillion yen in 2014, and reached 17 million members in October 2015. The group began with the distinct image of an internet company, but in recent years it has been developing a variety of businesses such as operating the start-up base "DMM.make Akiba" with equipment needed for hardware development, and the robot carrier business "DMM.make Robots". In this situation, increasing prediction accuracy through customer and market analysis in order to provide customized products and services. Norio Ojima from the R&D Team Chief Service and Application Engineering explains: "The mechanisms of IT and cloud computing are essential for the business growth of DMM.com. Considering future prospects, I believe we must tackle big data analysis as soon as possible and forge a system that will be at the core of our business."

Before examining the use of big data, DMM.com Labo offered "searches" of accumulated data, product "recommendations" tailored to customer attributes, and "behavior analysis" of customers. Matters were left to individual operational divisions, and although the use of these three tools had already taken shape in some divisions, there were other divisions which had barely begun. There were also issues with the accuracy and speed of searches and analysis, so the development of company-wide integrated infrastructure became necessary. Yuichi Tanaka from the Office of the CTO System Development Main Division explains.

"Some divisions handed over data to outside companies for analysis, but customized fine tuning could not be done with outsourcing. For example, even when we wanted to increase accuracy by adding important factors from users' browsing status to recommendations based on purchase record, it was necessary to communicate closely with the outsourcing contractor, and it took up to half a year until completion in some cases. Also, although some divisions developed the recommendation feature internally, it was not on the assumption of company-wide use, so it cannot be used across divisions. We therefore decided to develop company-wide

Spark* analysis infrastructure



Spark* The surrounding environment

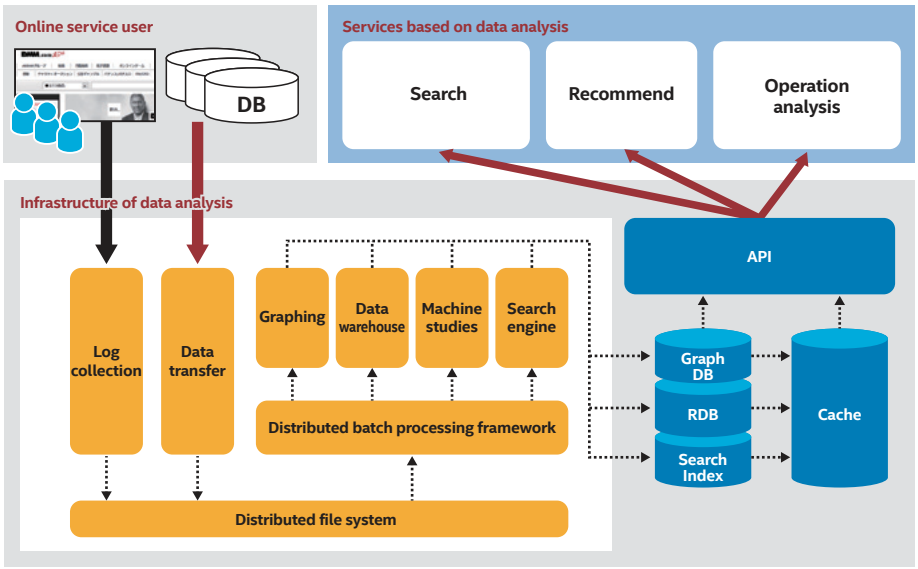
The configuration of Master series
- CPU: Xeon E5-2650 v2 2.6GHz x 2
- Memory:16G x 8
- HDD 1.6TB (200GB SSD(MLC) x 8)
- 10G SFP+

The configuration of Slave series
- CPU: Xeon E5-2680 v2 2.80GHz x 2
- Memory:16G x 8
- HDD 6.4TB (800GB SSD(MLC) x 8)
- 10G SFP+

The configuration using CDH5.4
HDFS
YARN
Spark

*Abbreviations
NN:NameNode
RM:ResourceManager
JN:JournalNode
JH:JobHistoryServer
HS:SparkHistoryServer
DN:DataNode
NM:NodeManager

Data analysis infrastructure (overview)



infrastructure for behavior analysis, searches, and recommendations."

Apache Spark* selected due to high processing speed and a complete library

While considering analysis infrastructure for integrating all services of the DMM.com group, DMM.com Labo, which thought the adoption of open source that could be manipulated by the company was necessary, tested a recommendation system based on Apache Hadoop* (henceforth, Hadoop*) that used Apache Mahout*, a library for machine learning algorithms. However, when it was actually analyzed, it became clear that the processing speed for large amounts of data was slow, and it took half a day to a full day until completion. At this point we turned our attention to Spark*, which allows for a higher analysis speed. Compared to other approaches, the in-memory processing of Spark* was vastly faster, and by running processing on memory using the libraries provided such as MLlib and GraphX, real-time processing could be ensured. We were considering Spark's* run infrastructure, but decided to adopt Cloudera's "Cloudera Enterprise" for platform development. Mr. Ojima looked back on the reasons for selecting Spark*. "Back in May 2015 when we were examining our choices, Cloudera was actively promoting the Hadoop*-based Spark*, and since not only Spark*, but Apache Kafka* (henceforth, Kafka*) for importing large-scale log data in real time, libraries necessary for recommendations, and features necessary for log searches were all included, we felt that it matched our needs well."

Adopting the Intel® Processor based on clock rate and thread count guarantee I/O performance with disk cache

To build big data analysis infrastructure using Spark*, the key factors for server hardware were memory capacity, clock frequency of processor, and thread count.

Mr. Tanaka said that large capacity memory is necessary for handling Spark* with its in-memory processing, and thread count was also important for Spark*, which has parallel processing at its foundation.

And the server adopted to meet those requirements is the Intel® Xeon® Processor E5 Family. Engineer Yasuhiro Azebu from the Server Engineer Infrastructure Division explains. "There was no server other than the Intel® Xeon® Processor that fulfills the requirements. We principally adopt IA (Intel® Architecture) for our other services as well, and taking into consideration the balance between performance and cost, we adopted the Intel® Xeon® Processor E5 Family for the recommendation infrastructure using Spark* as well."

The analysis infrastructure using Spark* started off at the beginning of development at the end of 2014 with only 15 servers, but 10 more were added after a half year of operation so that 25 servers are now in operation. Servers with a built-in Intel® Xeon® Processor E5 Family were also adopted for the entire ecosystem around Spark*, which means that a total of about 35 IA servers are currently in operation.

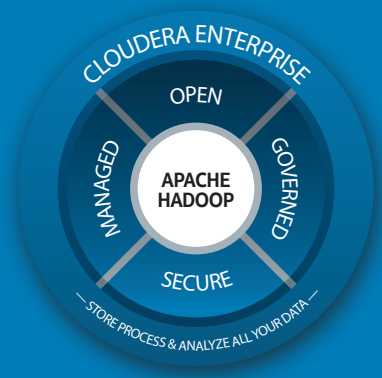
Among the 25 IA servers that comprise the infrastructure of Spark*, when it comes to the master system where applications operate,

Cloudera Enterprise

Hadoop*, extensive security and integration/control features for enterprises

Cloudera Enterprise is a platform comprised of Apache Hadoop* and the software components needed for Hadoop*. It is comprised of "CDH" which is widely used for global enterprises, the automatic cluster management feature "Cloudera Manager" for effectively handling Hadoop*, and the Hadoop* technology support service "Cloudera Support" run by Hadoop* developers and expert teams.

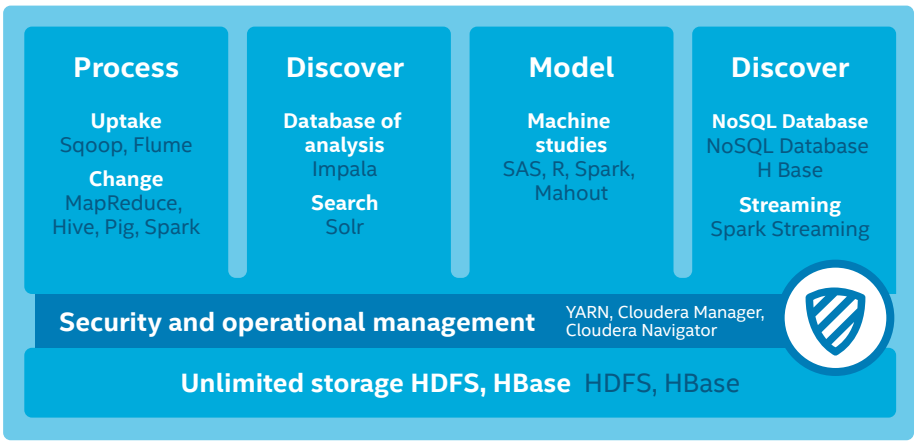
The core of the system CDH includes the distributed file system "HDFS", "Apache Spark*" which implements real-time processing, the large-scale distributed data store "Apache HBase*", "Apache Kafka*" which imports large-scale data in real time, the open-source interactive SQL "Cloudera Impala*", the open-source interactive search engine "Cloudera Search*", etc.



More details can be found here:
http://www.cloudera.co.jp/

Cloudera Enterprise

Cloudera Enterprise is an enterprise-oriented data management platform with integrated 100% open-source Hadoop* distribution, system management, data management, and comprehensive support.



thread count is attached more importance. A 2-socket server with two built-in 8-core Intel® Xeon® Processors E5-2650 v2 (2.60 GHz) was adopted, and Intel® Hyper-threading technology that supports two threads in one core brings the total number of threads operating to 32. With the database slave system, there are many processes that monopolize threads, so more importance was attached to the clock frequency than the thread count, and the 2-socket server with two built-in Intel® Xeon® Processors E5-2680 v2 (2.80 GHz) was adopted.

As for the feared I/O performance issues with Spark* and Hadoop*, the bottleneck was eliminated by building in 128 GB memory. Furthermore, in place of HDD, SSD with its high-speed disk access was adopted. Eight 800 GB SSD were built into each server of the slave system to achieve a total of 6.4 TB. Eight 200 GB SSD were also built into the master system, achieving a total of 1.6 TB.

Mr. Azebu explains: "Because the analysis infrastructure must support our company's core, there could be no comprising on performance, and importance was attached to fulfilling the speed requirements. The architecture was built such that hardware would not become a bottleneck, and a balance was struck between price, disk aggregation rate, and data aggregation rate."

Full service with recommendation feature and behavior analysis started in September 2015

Adopting Cloudera Enterprise, the analysis infrastructure with Spark's* recommendation feature implemented on IA servers started operation in December 2014. While accumulating log data, product data provided by each business division was read in, and in September 2015 the service with recommendation feature and behavior analysis was released, marking the Begining of full-scale operation.

The originally recommended platform and features and now being specified in similar and related products to users with a great deal of accuracy based on the purchase records from several services provided by the DMM.com group. Also with regards to how the data collected in real time is used for recommendations, we saw the advantage of being able to control it at our own discretion.

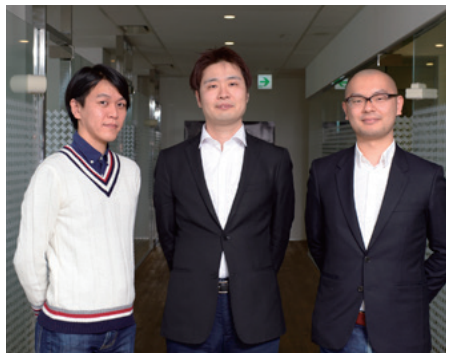
Behavior analysis is currently being used for the purpose of business intelligence, with users' log data being analyzed on an internal control screen, used for making all sorts of decisions on matters such as the purchase price of a user. Mr. Tanaka explains that the system is structured such that with behavior analysis, marketers can write SQL and issue queries on their own.

The difficulty of scaling up to more effectively use Spark*

With Spark* and Kafka*, DMM.com Labo built a company-wide integrated system infrastructure, and implemented recommendation and behavior analysis features. A future-oriented perspective is already in place. It is still in the research and development stage, but we have started an initiative to recommend products, taking into consideration market trends in real time, from product information and social streaming data being browsed by the user, using the Spark* streaming feature. Mr. Ojima told us that with monthly PVs exceeding 2.5 billion, and more than 1.7 million members, the key factor for rapidly giving recommendations is how the system can keep pace with the expansion of services.

Moving forward, the issue on the hardware side is how to scale up while balancing costs. Mr. Azebu explains: "5.6 TB of disk capacity on the database system's servers is just not enough in the world of big data. The issue now is expanding laterally and increasing the number of servers while keeping costs down in order to more effectively use Spark*." He also mentions disk speed as an issue to consider, and has high hopes for the faster, next generation connection standard "NVMe*".

The analysis infrastructure using Spark* that was first released in Asia by DMM.com Labo will continue to evolve while incorporating many features.



System Development Main Division
Office of the CTO Yuichi Tanaka (left)
Infrastructure Division
Server Engineer Yasuhiro Azebu (center)
Service and Application Engineering
R&D Team Chief Norio Ojima (right)

For more details on the Intel® Xeon® Processor E5 Family, please visit:
http://www.intel.co.jp/xeonE5/

Printed in USA JPN/1511/PDF/SE/DCG/ME 333413-001US