Global Expansion of Apache Hadoop/ Apache Spark Activities at NTT DATA

Ravindra Sandaruwan Ranaweera and Akira Ajisaka

Abstract

Apache Hadoop and Apache Spark are open source software projects that have been attracting worldwide attention as platforms for big data applications. NTT DATA has been developing systems using Hadoop/Spark since Hadoop/Spark's early days and providing one-stop technical support for the entire NTT DATA Group. Furthermore, in addition to participating in community-based development to reflect the feedback gained from technical support service into Hadoop/Spark itself, NTT DATA has been presenting the know-how accumulated from these activities at numerous events to raise its global presence.

Keywords: big data, parallel distributed processing, global

1. Hadoop/Spark

An increasing number of companies are attempting to create new businesses or expand existing businesses based on the results of analyzing huge amounts of various types of data known as big data. Apache Hadoop was introduced in 2006 and enables massive amounts of diverse data to be stored, processed, and analyzed in a realistic period of time at a reasonable cost. Apache Spark, meanwhile, was developed by postgraduate students at the University of California, Berkeley to efficiently perform repetitive and complex processing of big data, which Hadoop is not good at doing.

At present, Spark is becoming increasingly popular thanks to a number of key features. For example, it supports programming languages such as Python and R widely used in the data analysis industry, incorporates a machine-learning library, and performs various types of optimization.

2. Hadoop/Spark activities at NTT DATA

NTT DATA began using Hadoop in 2008 before Hadoop began attracting attention worldwide. At that time, Hadoop was still immature, and there were no functions satisfying the strict requirements pertaining to availability and operability demanded by enterprise customers. The NTT DATA team working on Hadoop, however, felt that Hadoop could bring significant benefits to enterprise customers. With this in mind, the team conducted tests on applying NTT DATA's know-how in system integration in order to meet such strict availability/operability requirements, and published that know-how in the form of a test report released in 2010 [1].

This report helped to make NTT DATA's work on Hadoop well known, and since then, NTT DATA has gone on to construct and provide systems using Hadoop/Spark in all sorts of industries including telecommunications, real estate, public administration, finance, media, and manufacturing. NTT DATA's involvement is not limited to just a portion of system development, but rather, it provides a wide range of services covering the entire system development process including planning, design, development, and support to provide its customers with systems that can lead to new business opportunities (**Fig. 1**). Providing versatile services in this way enables NTT DATA to satisfy the genuine needs of its customers.

Furthermore, to enhance services and raise the level of customer satisfaction, NTT DATA proactively

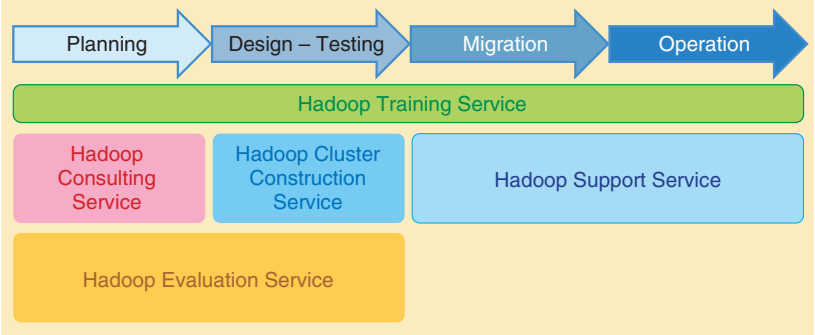

Fig. 1. NTT DATA Hadoop/Spark service menu.

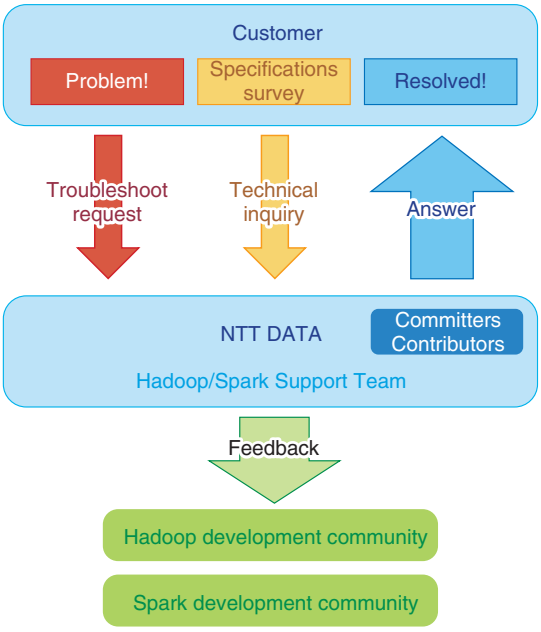

Fig. 2. Knowledge gained from Hadoop/Spark system development is fed back to the development community.

feeds back the knowledge it has gained from Hadoop/Spark system development to the development community (**Fig. 2**).

For example, if some type of problem should occur in system development due to a software bug, NTT DATA would interact with the community to fix that bug. In this way, there is no need for NTT DATA to install and manage its own software patches. For the customer, meanwhile, updating to a new version provides a fundamental solution to the problem while preventing its software from becoming detached from the software updates provided by the community.

NTT DATA has also developed a variety of Hadoop/Spark functions and merged them with the community. For example, in Spark, to simplify performance tuning and debugging, NTT DATA guided the development of the Timeline Viewer tool for visualizing which process has run or is running on which server in chronological order. A screenshot of the Timeline Viewer function whose development was headed by NTT DATA is shown in **Fig. 3**. This type of activity was carried out through discussions and collaborations with developers from around the world.

The Hadoop/Spark development community recognized the value of this work and elected several of these NTT DATA developers to the position of

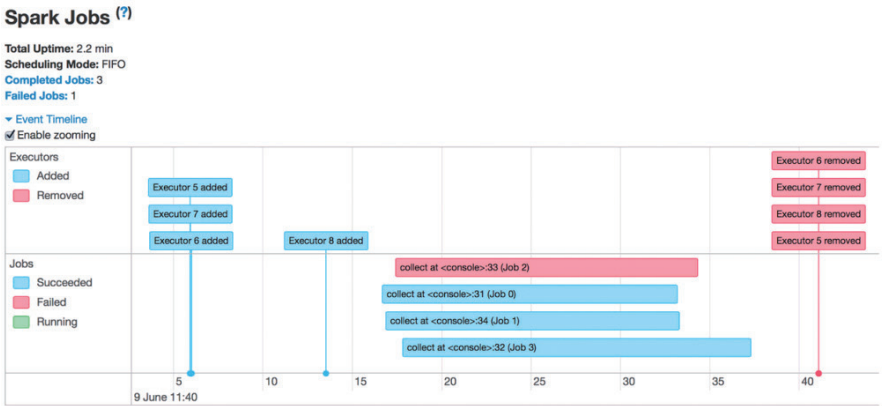

Fig. 3. Screenshot of Spark Timeline Viewer developed under NTT DATA leadership.

Photo 1. Hadoop/Spark Enterprise Solutions Seminar held by NTT DATA.

committer (a developer having the right to modify Hadoop/Spark source code), thereby making these developers the first Hadoop/Spark committers from a Japanese company. They continued with their Hadoop/Spark development activities even after becoming committers, resulting in their being promoted to members of the Project Management Committee that manages Hadoop projects.

In Japan, there are many companies that, while possessing large volumes of data, are not making use of that data. These companies who would like to analyze the large amounts of data in an efficient manner to generate new business value have come to NTT DATA asking for support. In response to these needs, NTT DATA holds the annual Hadoop/Spark Enterprise Solutions Seminar [2] to explain how enterprises can use Hadoop/Spark to create new value (**Photo 1**). This seminar introduces advanced use cases driven mostly by NTT DATA in a variety of industries within Japan as well as Hadoop/Spark use cases driven by NTT DATA Group companies outside Japan. It presents the benefits that Hadoop/Spark can bring to enterprises in an easy-to-understand manner and attracts many individuals from NTT DATA enterprise customers. In addition to the above, NTT DATA regularly holds a variety of seminars and meetups within Japan to introduce know-how of a more technical nature and to promote its technical expertise in Hadoop/Spark.

3. NTT DATA's Hadoop/Spark related global activities

NTT DATA does not share its knowledge and know-how only at seminars and events within Japan. Since it pursues system development using advanced technologies, it has been involved in a few world-first use cases applying those technologies. NTT DATA

Photo 2. Presentation by the NTT DATA Hadoop/Spark team in the USA.

actively participates in seminars outside Japan to introduce these cutting-edge use cases and to promote the company's technical expertise. To date, it has given presentations at a variety of conferences including ApacheCon, Apache: Big Data, Dataworks Summit (formerly Hadoop Summit), Strata Data Conference (formerly Strata + Hadoop World), Spark Summit, Kafka Summit, and Global Big Data Conference, and it plans to continue doing so [2]. Scenes from NTT DATA's Hadoop/Spark team giving presentations in the United States in 2017 are shown in **Photo 2**. No other Japanese company regularly presents such advanced results at global Hadoop/Spark-related events, and in this capacity, NTT DATA acts as a representative of Japan.

Although Hadoop/Spark-related technologies are gaining in popularity, the pool of Hadoop/Spark technical personnel around the world is not growing. The reasons given for this include differences in past technologies and in ways of thinking, and the need for knowledge covering a wide range of technologies involving hardware, operating systems, databases, networks, and distributed processing. It is an unfortunate fact that there are relatively few Hadoop/Spark developers within the NTT DATA Group despite being a company whose employee numbers have grown to 110,000 worldwide through strategic acquisition of firms around the world. Under these conditions, non-Japanese NTT DATA Group companies often request support from the Japan-side Hadoop/Spark team for system development involving big data analysis.

For example, requests are being received from the United States, Europe, and Asia for help in creating and reviewing proposals, designing system architectures, developing applications, and performance tuning. End users (customers) have also recognized the technical expertise of the Japan-side Hadoop/Spark team that responds to such requests. In particular, Hadoop/Spark technical personnel in Japan have received praise from global customers for their ability to support an entire system, in contrast to the trend outside Japan to focus only on one specific area.

4. Future development

At NTT DATA, we wish to expand the use of Hadoop/Spark even further to help our customers create new value. Furthermore, in addition to Hadoop/Spark, we plan to incorporate the latest research results in open source middleware for distributed processing to improve its performance and scalability. We are presently holding discussions with researchers from the NTT laboratories on the latest research trends in distributed processing, and we hope to continue these discussions at an even deeper level going forward.

References

[1] NTT DATA, "Project Report on 2009 Industry-Academia Collaboration Project on Software Engineering Practice," Mar. 2010 (in Japanese). http://www.meti.go.jp/policy/mono_info_service/joho/downloadfiles/2010software_research/clou_dist_software.pdf

[2] Website of NTT DATA Hadoop/Spark Enterprise Solutions Seminar (in Japanese), http://oss.nttdata.com/hadoop/event.html

Trademark notes

Ravindra Sandaruwan Ranaweera
Software Engineer, OSS Professional Services, System Engineering Headquarters, Platform Engineering Department, NTT DATA Corporation.

He received his B.E. and M.E. in network optimization from the University of Electro-Communications, Tokyo, in 2012 and 2014. He joined NTT DATA in 2014 and has since been working as a platform engineer/consultant on distributed systems using Apache Hadoop/Spark.

Akira Ajisaka
Software Engineer, OSS Professional Services, System Engineering Headquarters, Platform Engineering Department, NTT DATA Corporation.

He received his B.E. in engineering and his M.E. in applied mathematics and physics from Kyoto University in 2009 and 2011. He joined NTT DATA in 2011 and has since been working on distributed systems using Apache Hadoop. He is an Apache Hadoop committer and Project Management Committee member.