Report on Apache Big Data North America 2016 and Spark Summit 2016

Takeshi Yamamuro and Tsuyoshi Ozawa

Abstract

Two key conferences were held this year on open source projects concerning big data processing systems: Apache Big Data North America 2016 and Spark Summit 2016. This article reports on some of the topics discussed at the conferences.

Keywords: distributed processing, Apache Hadoop, Apache Spark

1. Apache Big Data North America

Apache Big Data North America is one of the largest conferences related to open source projects on big data processing and is supported by the Apache Software Foundation. The conference features interesting presentations given by users and developers on various big data processing systems using Apache open source software (OSS) products such as Hadoop [1], Spark [2], Kafka [3], and Cassandra [4]. This is a key conference for OSS developers and typically has higher numbers of developers than other events. Lively discussions were held at the conference that continued even during the coffee breaks.

1.1 Conference summary

Apache Big Data North America 2016 [5] was held in Vancouver, Canada from May 9 to 12. Many people attended the conference from hardware vendors to content providers, including representatives from Intel Corporation, Netflix, Inc., eBay Inc., Yahoo Japan Corporation, and Recruit Holdings Co., Ltd. These companies are active users of OSS products.

1.2 Business use cases

The notable keywords appearing in the titles of the presentations at the conference were Spark, Hadoop, Kafka, and Cassandra. Of the total presentations, 55 of them, or more than 40%, were related to these keywords.

In a keynote speech entitled, "How Netflix Leverages Big Data," Brian Sullivan, Director of Streaming Analytics at Netflix, talked about how open source products are being used in Netflix's services. Netflix is a worldwide video streaming service used by 81 million subscribers. A huge amount of video content—more than 125 million hours—is uploaded and distributed per day. Surprisingly, Netflix's communication traffic occupies more than one-third of all the Internet traffic in North America. Netflix creates and improves its services by using big data. By analyzing over 3 petabytes of data each day, Netflix can create, deliver, and provide excellent content. They utilize OSS developed in community supported by the Apache Software Foundation such as Hadoop, Spark, Kafka, and Cassandra.

2. Spark Summit

The Spark Summit conference was first held in 2013. Since 2015, it has been held three times a year on the east and west coasts of North America and in the EU. Developers present the latest features of Spark, an open source cluster computing framework, and users share Spark use cases and know-how on using it.

Spark was developed by the AMPLab team at the University of California, Berkeley. Developers at

Databricks are now taking the lead in this community. In addition to the underlying mechanisms necessary for distributed processing, Spark includes easy-to-use libraries for SQL queries, machine learning, streaming, and graph data processing. A report on a 2015 questionnaire survey [6] on Spark indicated that the Spark community continues to grow year by year and that Spark is one of the most popular OSS products worldwide. The development community is highly active, and version 2.0 was officially released at the end of July 2016.

2.1 Conference summary

The 7th Spark Summit 2016 [7] was held from June 6 to 8 in San Francisco. It was supported by many companies including IBM Corporation, Microsoft Corporation, Intel, and EMC Corporation. The conference was attended by more than 2500 people from over 720 companies.

2.2 Development community trends

At the conference, the main developers talked about the upcoming features in version 2.0, which is characterized by continuousness and structuralization. The new features provide effective processing of stream data such as sensor log data. Spark version 1.6 has a feature called DataFrame/Dataset for processing static data with schema. Version 2.0 has a new feature for stream data called Structured Streaming. This Structured Streaming feature is intended for users who need to process log data in near-real time. Another session at the conference focused on memory and query management. All indications are that Spark has great potential to grow because of the increasing number of developers and users.

2.3 Business use cases

Spark includes numerous libraries necessary for analytical processing in various use cases. In particular, many use cases concerned Spark Streaming. This function was implemented in version 0.7. Spark Streaming is similar to Structured Streaming, which is currently under development, as it deals with dynamic data. Spark Streaming has certain limitations, however, so it is expected to be replaced over the long term by the sophisticated functions of Structured Streaming.

Presentations were also given at the conference on the analysis of user logs generated daily for Microsoft's search engine Bing and on the behavior analysis of the popular Airbnb, a rental service for lodgings and guest houses. The common point of these two cases is that they use Kafka, a distributed message processing infrastructure. Kafka is used to store generated data, and Spark fetches the data from Kafka. Because Kafka and Spark both use the programming language Scala, they mesh well. Consequently, Kafka is becoming the de-facto standard to input into Spark Streaming.

In the case of the Weather Company, a weather forecasting company purchased by IBM in January 2016, petabytes of data are generated every day, and Spark is used for data processing. Thus, in recent years, the number of large-scale examples has gradually been increasing, so it is expected that the scale of applications will continue to grow even more.

3. Activities of NTT Software Innovation Center

NTT Software Innovation Center performs research and development on big data processing infrastructures, which is the common theme between the two conferences reported in this article. Therefore, we would like to introduce one of our efforts related to these conferences.

We gave a presentation at Apache Big Data North America 2016 on Hadoop's function to manage computer resource allocation, which we developed with other OSS developers [8]. At the conference, we discussed various issues with OSS developers. For example, we discussed how to use various devices on the Hadoop framework with researchers from the University of Toronto. Devices such as a GPGPU[*1] and FPGA[*2] are the most important accelerators for processing big data. Hadoop's current resource manager does not control these accelerators because it is not clear how they should be allocated to various tasks.

However, two main bottlenecks can occur with the accelerators. One is a CPU[*3] intensive bottleneck that can occur with, for example, scientific computing and machine learning. The other is a network I/O[*4] intensive bottleneck, which can occur when extracting a specific pattern from a large amount of data. A separate solution must be found for these bottlenecks. Unfortunately, the problem is that both bottlenecks can occur in one use case. This means we must solve the problems under these constraints, which involves a trade-off.

[*1] GPGPU: general-purpose graphics processing unit

[*2] FPGA: field-programmable gate array

[*3] CPU: central processing unit

[*4] I/O: input/output

It is difficult to inject accelerator management into the Hadoop framework in terms of interchangeability. However, as the use cases of Hadoop are extended, it will be necessary to solve the problem of the utilization of accelerators in Hadoop.

References

[1] Apache Hadoop, http://hadoop.apache.org/

[2] Apache Spark, http://spark.apache.org/

[3] Apache Kafka, http://kafka.apache.org/

[4] Apache Cassandra, http://cassandra.apache.org/

[5] Apache Big Data North America, http://events.linuxfoundation.org/events/apache-big-data-north-america

[6] Spark Survey 2015 Results, https://databricks.com/blog/2015/09/24/spark-survey-2015-results-are-now-available.html

[7] Spark Summit 2016, https://spark-summit.org/2016/

[8] Apache Big Data 2016, https://apachebigdata2016.sched.org/event/6M0N/yarn-a-resource-manager-for-analytic-platform-tsuyoshi-ozawa-ntt

Takeshi Yamamuro

Senior Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.

He received an M.E. from Faculty of Science and Technology of Sophia University, Tokyo, in 2008. He joined NTT in 2008 and has been working on database management systems. His research interests include compression and hardware-aware algorithms (e.g., SIMD, NUMA, and GPU). He received the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2015. He is a member of IPSJ and the Database Society of Japan (DBSJ).

Tsuyoshi Ozawa

Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. in information and system engineering from Chuo University, Tokyo, in 2008 and an M.E. in computer science from the University of Tsukuba, Ibaraki, in 2010. He joined NTT in 2010 and is working on distributed processing frameworks such as Hadoop. His research interests include distributed computing and distributed databases. He received the Computer Science Research Award for Young Scientists from IPSJ in 2013 and the 9th Japan OSS Incentive Award from the Japan OSS Promotion Forum in 2014. He has served as a committer since 2014 and as a member of the project management committee since 2015 in the Apache Hadoop community. He is a member of the Association for Computing Machinery (ACM), IPSJ, and DBSJ.