The Latest Developments in Jubatus, an Online Machine-learning Distributed Processing Framework

Takashi Hayashi, Masayoshi Umeda, Masato Sawada, Kaori Isagai, Akihiro Yamanaka, and Mitsuaki Tsunakawa

Abstract

Mobile terminals and other devices are now adopting multiple sensors, which generates voluminous data sets (big data). This necessitates rapid analysis of big data to understand the latest trends and current events. In this article, we introduce the latest developments in the open source community and commercial support activities for the distributed processing framework called *Jubatus*. Case studies are introduced to show that it offers deep analysis of big data in real time.

Keywords: real-time analysis, parallel and distributed architecture, online machine learning

1. Introduction

The popularity of the Internet and the rapid adoption of information and communication technology mean that large volumes of a wide variety of data sets are being generated. Examples include user data from various social networking services (SNSs) such as Twitter[*1] and Facebook[*2], log files from network equipment and servers, and data sent from vehicle-mounted sensors and home appliances.

We introduce here the online machine-learning distributed processing framework Jubatus and some examples of its application. It moves beyond simple statistical aggregation and keyword searches of big data and offers real-time and deep analysis of big data.

2. Jubatus

NTT's laboratories and Preferred Infrastructure, Inc. (PFI) commenced work on the Jubatus project in 2011 and released it as open source software (OSS) in October of that year [1–3]. Jubatus was developed with two goals in mind, and three features were added in order to achieve these goals, as shown in the green circles in **Fig. 1**.

The first goal is highly scalable processing performance for real-time analysis, which is achieved through the parallel and distributed architecture of Jubatus. The second goal, deep analysis, is realized by the adoption of machine learning techniques.

Real-time analysis enables quick response through sequential processing of continuously generated streams of data without temporary storage. The parallel and distributed architecture enables scale-out by adding servers in the same way as Hadoop and other large-scale data processing platforms. Given that it is often difficult to set rules to govern system processing in advance (since the rules are unknown, it is difficult to express constraints as rules, or the rules change over time), Jubatus adopts the machine learning approach to learn rules from data examples.

*1 Twitter is a registered trademark of Twitter, Inc. in the United States and other countries.

*2 Facebook is a registered trademark of Facebook, Inc. in the United States and other countries.

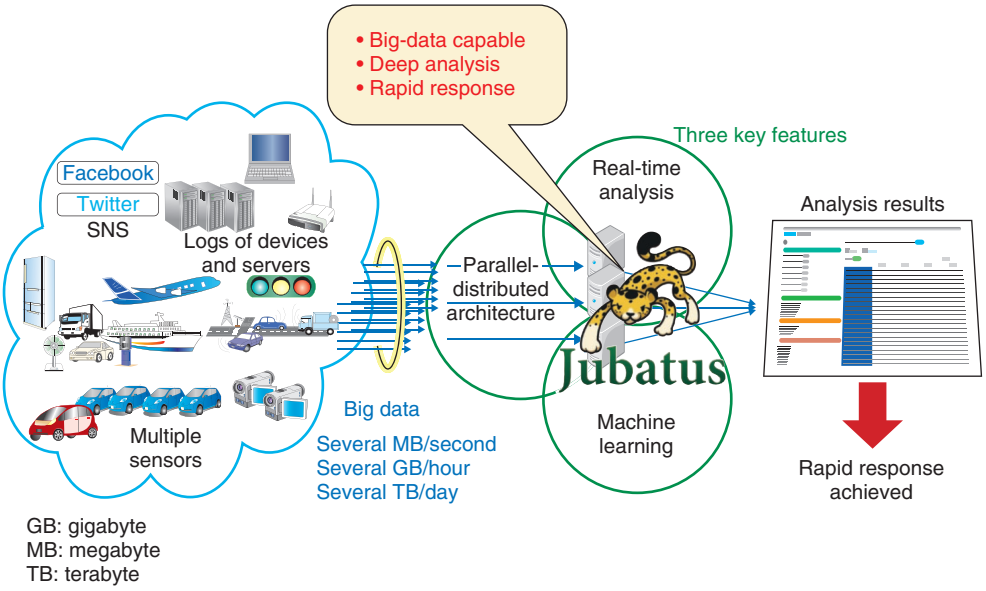


Fig. 1. Jubatus.

Jubatus has offered for the first time in the world the combination of real-time analysis, parallel and distributed processing, and machine learning [4].

3. Jubatus application examples

The three features of Jubatus are useful in detecting the onset of abnormalities or failures. For example, in the case of a factory, rapid response (real-time analysis) can be achieved by sequentially processing the data and logs output by a variety of sensors. The parallel and distributed architecture of Jubatus makes it easy to add extra servers whenever needed, so real-time response is possible even when the service area is extremely large and the number of sensors is enormous. Rather than adding dedicated systems to catch faults, the anomaly detection algorithm of Jubatus and its machine learning approach enable anomalies to be identified from the data of many sensors even if the detection pattern is unknown to the operator.

An example of the machine learning approach being applied to server log data is shown in **Fig. 2**. Principal components are extracted and plotted in three dimensions; data points that are unusual are shown in red. This representation yields regular and irregular patterns and enables unknown anomalies to be predicted.

To expand the application area of Jubatus, researchers at NTT's Machine Learning and Data Science Center are working in tandem with researchers involved in the *himico* initiative [5] on a proof of concept (PoC)[*3] project targeting areas such as network fault detection.

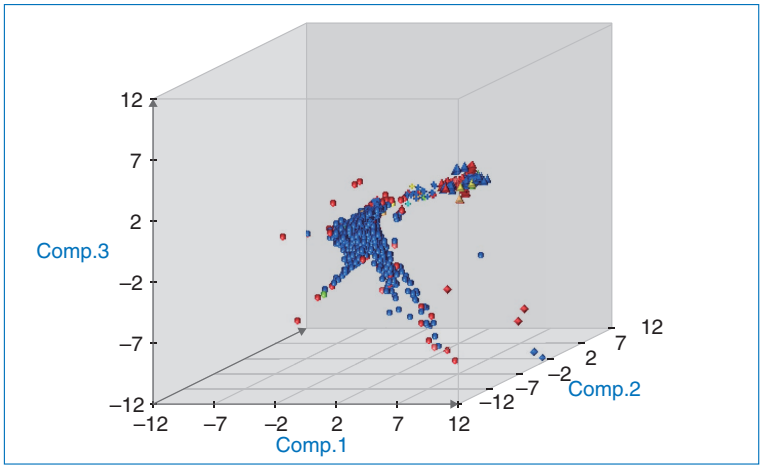
4. Latest development trends

Since its release as OSS in 2011, Jubatus has been repeatedly upgraded (nine times in 2014 alone), with emphasis placed on improving productivity. Consequently, it now offers an extremely efficient programming environment. In addition, data scientists and professionals in different fields are continually enhancing its various analysis algorithms. The algorithms currently available in Jubatus are listed in **Table 1**. As one example, the Bandit algorithm allows multiple solutions to a problem to be assessed in parallel, even while the service is in operation, and it determines which solution best suits the user's current situation. This algorithm is extremely useful for product recommendations and advertisement placements.

The analysis algorithms listed in the table take the form of plug-in modules, and user-written analysis modules can be loaded as plug-ins that can be freely selected and applied as needed.

JubaQL is the latest of our research results to be published. JubaQL offers an interactive interface with

*3 PoC: Minimal realization of a proposal sufficient to confirm its viability.



• Detecting unusual new patterns (shown in red) in log data; each data point has 41 dimensions.
• All data must be sampled to provide total inspection.
• This anomaly detection algorithm is based on unsupervised learning.

Fig. 2. Predicting failure from irregularities.

Table 1. Analysis algorithms implemented in Jubatus.

Analysis algorithm	Function	Use case example
Classification	Classifying data into categories	• Spam mail blocking • Categorizing tweets
Recommendation	Recommending items similar to given item	• Recommending items for EC sites • Linking search requests to ads
Regression	Estimating value from given data	• Predicting power consumption • Predicting share prices
Statistics	Aggregating frequency, standard deviation, maximum value, and minimum value statistics	• Sensor monitoring • Detecting data anomalies
Graph mining	Finding centroid and shortest route, etc., for given graph	• Social community analysis • Network structure analysis
Anomaly detection (Outlier detection)	Finding abnormal values (outliers) within a given data set	• Fraud detection • Fault detection
Clustering	Assigning given data points to a specified number of groups without using training data	• Segment analysis • Topic classification
Burst detection	Detecting bursts in traffic without using thresholds	• Detecting traffic congestion • Detecting SNS storms
Bandit	During service operation, tailoring service choices to suit the user's circumstances	• Personalized item recommendations • Linking users to ad sites

EC: electronic commerce

SQL (Structured Query Language)-like syntax to access the rich online machine learning functions available. Some of the benefits provided by JubaQL are shown in **Fig. 3**. Prior to its release, client programs that were needed for repeatedly evaluating different combinations of analysis processes had to be written in Ruby, Python, or another programming language. JubaQL offers highly efficient analysis processing and enhanced productivity.

5. OSS community activities and commercial support system

As mentioned above, Jubatus has been released as

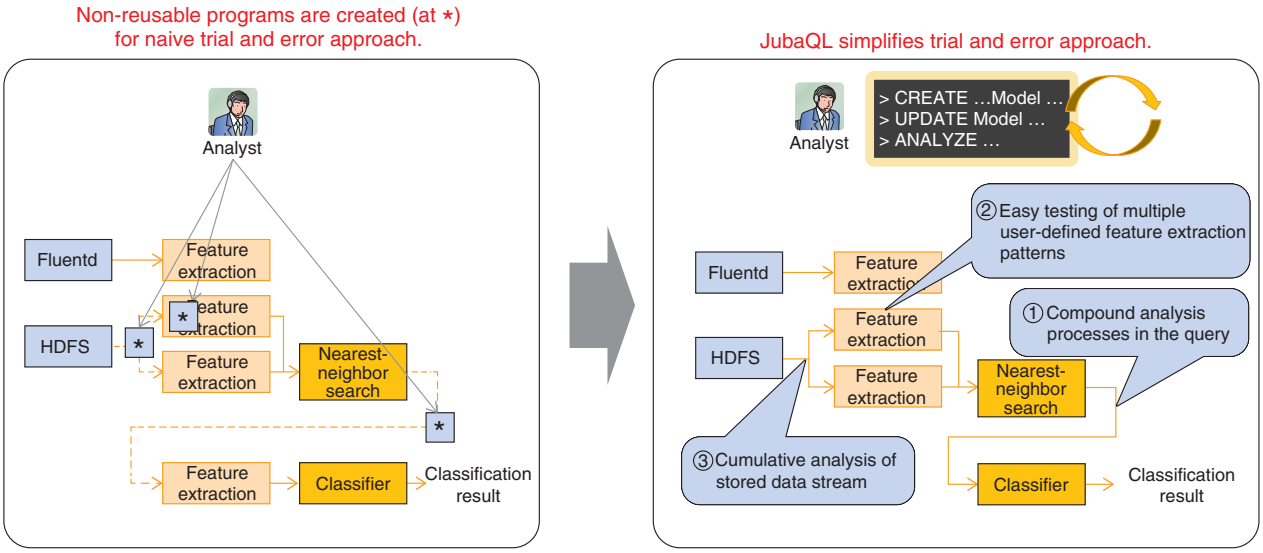


Fig. 3. Benefits of JubaQL.

an OSS project. Many of the distributed processing tools needed to undertake big data analysis such as Hadoop have been published as OSS projects and are attracting the interest of many developers and researchers. This has created an extensive ecosystem that has yielded new use cases and examples of data utilization. Our aim in releasing Jubatus as an OSS project is to connect with more users and thus discover new use cases and ideas for data utilization.

Our OSS project is extremely lively with over 2000 code commits, and we place emphasis on continually improving its usability. Lively community activities are also taking place, with some 590 participants attending two hackathons, four hands-on meetings, three casual talks, and one casual meetup event. The casual talks have been very interesting, as they introduced usage reports from organizations other than the original developers, NTT laboratories and PFI, which confirms that the use of the system is expanding.

Another significant indicator is the drive to introduce Jubatus to commercial products and thus expand its use by businesses. This will require careful utilization of commercial knowledge through selection of the most suitable analysis algorithm and performance tuning. To better support businesses, NTT Software Corporation established a commercial Jubatus support service in January 2014; it offers a wide variety of support functions to enhance the commercial application of Jubatus.

6. Case studies

6.1 Case 1: WatchBee

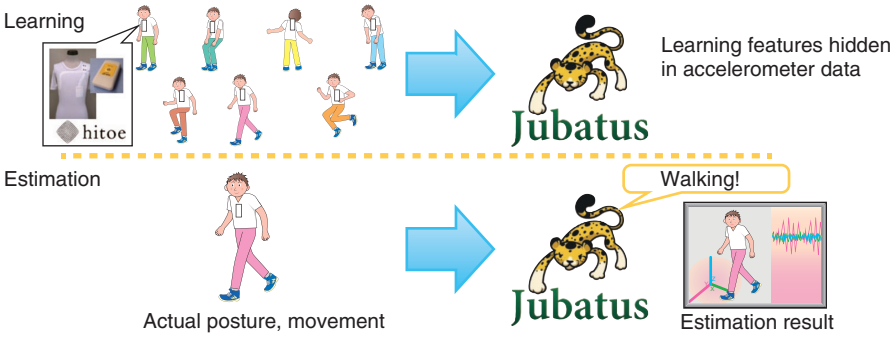
Jubatus is used in the summary display function of WatchBee [6], a reputation analysis service provided by NTT IT Corporation (NTT-IT). This function analyzes large-scale data such as SNS data, extracts the features of each kind of data, and groups similar data based on the correlation between features. Rapid data analysis is essential for this function because large quantities of new data are created every minute, and WatchBee must analyze that data every 30 minutes. The online processing of Jubatus has substantial advantages for this kind of deep and large-scale real-time analysis.

6.2 Case 2: hitoe

Jubatus is well-suited for analyzing sensor data. With conventional techniques, threshold values are necessary to classify human poses and motions when analyzing their three-axis accelerometer data[*4] acquired from wearable devices such as *hitoe* [7]. However, it is difficult to determine the proper threshold values. Furthermore, optimum values may differ from person to person because of their different physiques and manners of poses and gestures.

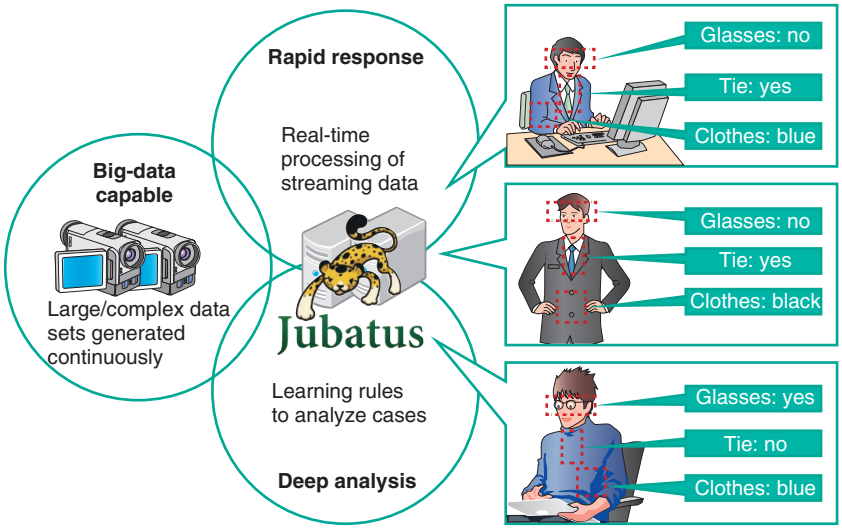
Jubatus can automatically find the criteria for

*4 Three-axis accelerometer data: To determine sensor orientation, data captured on the sensor's X, Y, and Z axes are processed to identify the direction of gravity.



• Learning and estimation functions of Jubatus (machine learning) identify poses and actions such as standing, sitting, and walking, by processing the output of the hitoe terminal.

Fig. 4. Example of Jubatus used with hitoe.



• Images tagged with features offer enhanced search functionality.
• Real-time detection of interesting abnormal events (crime prevention, disaster prevention) and customer identification

Fig. 5. Analyzing surveillance videos.

establishing thresholds by learning from the sensor data of each pose and motion (**Fig. 4**). Personalized criteria for every individual are generated by Jubatus since it is able to distinguish the differences in personal physiques and in the manner of poses and gestures. Furthermore, Jubatus can improve the accuracy of criteria by using its incremental learning function on misclassified data.

6.3 Case 3: Surveillance video analysis

Jubatus can also be used to analyze footage captured in surveillance videos (**Fig. 5**). Current security systems are inefficient, as the stored videos must be manually rewound to around the time of the incident and then manually reviewed. This can allow crime precursors to be overlooked. The classification power of Jubatus can be used to analyze the videos as they are captured and to extract notable features such as the color of clothing, the presence of a backpack, or actions such as using a smartphone or reading a book. The image data are tagged with the features, which greatly facilitates subsequent processing and search

operations.

We intend to utilize the latest video recognition technology such as *deep learning* [8] to further enhance the accuracy of attribute classification.

7. Future developments

Current activities are mainly directed towards strengthening the analysis algorithms of Jubatus and enhancing the usability of JubaQL. A long-term goal remains raising user awareness of Jubatus. Future activities will focus on introducing Jubatus to business applications, improving service quality (higher operability and reliability), and adding peripheral functions. Finally, we intend to greatly expand the application area of Jubatus by introducing it in new fields such as AI (artificial intelligence) and IoT (Internet of Things).

References

[1] Website of Jubatus, http://jubat.us/

[2] Twitter account of Jubatus, https://twitter.com/jubatusofficial

[3] Website of GitHub, Jubatus, https://github.com/jubatus/jubatus

[4] K. Horikawa, Y. Kitayama, S. Oda, H. Kumazaki, J. Han, H. Makino, M. Ishii, K. Aoya, M. Luo, and S. Uchikawa, "Jubatus in Action: Report on Realtime Big Data Analysis by Jubatus," NTT Technical Review, Vol. 10, No. 12, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201212fa5.html

[5] NTT press release issued on February 18, 2015 (in Japanese). http://www.ntt.co.jp/news2015/1502/150218a.html

[6] Website of NTT-IT, WatchBee (in Japanese). http://www.ntt-it.co.jp/product/watchbee/

[7] K. Takagahara, K. Ono, N. Oda, and T. Teshigawara, ""hitoe"—A Wearable Sensor Developed through Cross-industrial Collaboration," NTT Technical Review, Vol. 12, No. 9, 2014. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201409ra1.html

[8] Website of Deep Learning, http://deeplearning.net/

Takashi Hayashi

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.E. and M.E. in electrical engineering from Keio University, Kanagawa, in 1997 and 1999. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 1999, where he researched database management system (DBMS) technology. From 2003 to 2010, he developed video conference systems at NTT Bizlink. He is currently engaged in research and development (R&D) of a scalable distributed computing framework for big data. He is a member of the Information Processing Society of Japan.

Masayoshi Umeda

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.E. in electrical engineering from the University of Electro-Communications, Tokyo, in 1991. He joined NTT Information Telecommunication Networks Laboratory in 1991 and researched and developed DBMS technology. He is currently working on the development of Jubatus.

Masato Sawada

Senior Research Engineer, NTT Software Innovation Center.

He received an M.E. in information and computer sciences from Osaka University in 2000. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2000. He has been researching and developing a large-scale full text search engine. He is also currently involved in the R&D of data processing software using machine learning.

Kaori Isagai

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

She received a B.E. and M.S. in information engineering from Nagoya University, Aichi, in 1994 and 1996. She joined NTT Multimedia Network Laboratories in 1996 and engaged in R&D of large scale Internet protocol networks. From 2011 to 2015, she developed session control systems for NGN (Next Generation Network) at NTT EAST. Since moving to NTT Software Innovation Center in 2015, she has been incubating and developing the market for Jubatus.

Akihiro Yamanaka

Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received an M.S. in mathematics from Waseda University, Tokyo, in 2010. He joined NTT in 2010 and is involved in data analysis.

Mitsuaki Tsunakawa

Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. in mathematics from Tsukuba University, Ibaraki, in 1990 and joined NTT Communications and Information Processing Laboratories the same year. He has been researching DBMS technology and the integration of heterogeneous information sources. He is currently engaged in R&D of a scalable distributed computing framework for real-time analysis of big data. He is a member of the Institute of Electronics, Information and Communication Engineers.