(Press Release)

October 26, 2011

Nippon Telegraph & Telephone Corp.
Preferred Infrastructure Corp.

Leading Development of Scalable Distributed Computing Framework for Real-Time Analysis of Big Data

-- Released as Open-Source on October 27 --

Nippon Telegraph & Telephone Corporation (NTT, of Chiyoda Ward, Tokyo, CEO: Satoshi Miura), and Preferred Infrastructure Corporation (PFI, of Bunkyo Ward, Tokyo, CEO: Toru Nishikawa) have developed an infrastructure technology called “Jubatus” (1st Edition)[*2], which is capable of high-speed, real-time analysis of large-scale data, referred to as “Big Data” [*1]. Conventional batch processing methods periodically process data in batches and put newly arrived data on hold until the next batch execution. These methods are inadequate for Big Data applications such as real-time trend analysis, for which the timeliness of data is a critical requirement. By providing the capability of analyzing the latest data in real time, Jubatus can help create value-added services in a wide range of areas such as fraud detection, forecasting of market, economic and stock prices, natural disaster prediction, parts and materials procurement estimation for manufacturing, health-risk assessment, and predictive techniques in life and natural sciences. This development is a result of open innovation between NTT Information Sharing Platform Laboratories and PFI Corporation. It will be released as open source on October 27 on the Jubatus OSS Web site, http://jubat.us/ , as a public domain software contributing to the utilization of Big Data.

Technical Background

In recent years, the term “information explosion” is used in various fields to describe the rapid increase in the amount of published data available. It is important for companies to generate business intelligence by utilize this “Big Data” proactively and effectively.
Currently, Big Data is analyzed by temporarily storing it in a cloud environment composed of a server farm, and periodically processing it in high-speed batches. Hadoop[*3], is one such system that is gaining recognition and popularity.
However, the world is changing extremely quickly, and there is a growing need for technology able to perform sophisticated, real-time analysis of large volumes of data, arriving in time sequence, without storing it. Applications such as SNS analysis or detection of abnormal traffic or unauthorized access use it in order to implement high-speed decision making based on sophisticated analysis and forecasting.

Technical Overview

Jubatus is a large-scale, distributed, real-time analysis framework with the objective of continuous, high-speed, deep analysis of high-volume data (Figure 1). Jubatus achieves continuous, high-speed processing of high-volume data by dividing the large-volume of data among multiple servers and processing it sequentially and in parallel. Deep analysis requires use of sophisticated statistical processing and machine learning, and implementation in a distributed environment requires a framework allowing multiple servers to share intermediate results. Such sharing requires frequent communication between servers, and this can become a bottleneck to overall performance if a suitable communication method is not devised.
Accordingly, Accordingly, Jubatus not only ensures the real-timeliness and accuracy of data analysis, but also increases robustness by exchanging the intermediate results among multiple servers in a loose manner and thereby reducing the communication overhead between servers.

Technical Features

The main features of Jubatus are as follows (Figure 2).

(1) MIX Processing system

This processing system has the following three functions.
<1> MIX Computation: Arranges the aggregate computation logic, depending on the data analysis logic.

<2> MIX Protocol control: Determines how data is aggregated and redistributed when checking intermediate analysis results among the servers.

<3> Membership management: Performs tasks such as recovery from server faults and adding more servers in order to ensure continuous data processing, before data overflow can occur.
Even with simultaneous parallel analysis, having all servers wait for each other to compare intermediate results at each iteration will clearly result in a bottleneck. We were able to ensure that each server can run autonomously without slowing down by having servers exchange and mix intermediate results with other servers at suitable time intervals, rather than at every iteration. The balance in achieving both real-time nature and scalability is adjusted within the range allowable by the application, so that the precision and strictness (overall consistency) of the aggregate results can be relaxed (Figure 3).

(2) Pluggable architecture

Analysis engines, analysis modules, and data storage methods (local, distributed) can be combined and rearranged flexibly (plugged-in, out) due to the definition of shared interfaces.

(3) Workflow definition

It is possible to define and control execution of paths and parallel execution between process components easily and flexibly, from data input, to applied analysis, analysis engine and others.

At this time, we have implemented and evaluated a multi-value classifier for online machine learning as the first instance of analysis engines for Jubatus.

Future Developments

In order to further advance R&D and contribute to the development of information processing technology for Big Data, NTT and PFI Corp. are working to promote the spread of real-time large-scale data analysis infrastructure and related business by expanding the Jubatus community and businesses built on it. We are considering an “SNS analysis application” service in particular. This application will perform sophisticated analysis, such as categorization, fuzzy search, real-time filtering, and relevancy ranking, of the large volumes of real-time SNS data generated every day, so that it can be used for marketing and other applications. Figure 4 illustrates the concept of SNS analysis applications using Jubatus.
Other applications include:

“Sensor data analysis”
“POS data analysis”
“Log data analysis”
“Financial data analysis”
“Behavioral analysis”

References

See the Preferred Infrastructure Web site: http://www.preferred.jp
PFI Corp. is a venture company with excellent research and development staff in the fields of natural language processing and machine learning. The high-performance analysis engine in Jubatus has been produced from PFI technical expertise.

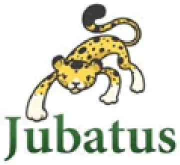
Terminology

*1 Big Data

Refers to data sets that are very large and have complex structure, so they are difficult to manage and process using conventional technologies. Although not clearly defined, these data sets are usually several-hundred-terabyte or peta-byte-class in size, do not have fixed form and are real-time in nature. They can include, for example, data from RFID tags or other types of sensors, or text from blogs or other new communications tools.

*2 Jubatus

A name selected from “Acinonyx jubatus,” the scientific name for the Cheetah. We hope to form a “Jubatus community” through release of the system as open source.



Jubatus Logo

*3 Hadoop

An open-source clone of Google’s infrastructure system (MapReduce, BigTable, GFS, etc.). It is the representative example of a batch-processing large-scale distributed processing infrastructure. The Hadoop community is quite mature.

Attachment·Reference

- Figure 1: Jubatus positioning
- Figure 2: Architecture overview
- Figure 3: Jubatus mechanisms
- Figure 4: SNS analysis application concept

Inquiries

NTT Information Sharing Platform Laboratories
Planning Dept., Promotions Office

TEL: 0422-59-3663
E-mail: islg-koho@lab.ntt.co.jp

Preferred Infrastructure Corp.
Jubatus Group

TEL: 03-6662-8675
E-mail: info@preferred.jp

Information is current as of the date of issue of the individual press release.
Please be advised that information may be outdated after that point.

NTT Press Releases Index

NTT Press Releases

- Latest Press Releases
- Back Number
- Japanese is here

Search Among
NTT Press Releases

January 1997 -
November 2021
Search

▲ Page Top