illumino—IIJ's Internal Data Analytics Platform

3.1 Introduction

As the use of IT in business processes continues to expand, many people no doubt feel that information volumes are growing more and more all the time. And commensurate with the volume of information, the number of servers, network devices, and applications managed also increases. This is why the amount of data to be managed is increasing at an accelerating pace beyond the amount we would normally want to deal with.

As the amount of data to be managed increases, devising ways of handling the data efficiently is key. There are many considerations when it comes to storing and managing data, including storage efficiency improvements, data integrity, ease of analysis, increases in analysis speed, and data security. In many projects up till now, only data saving and integrity, or a limited subset required for the task, were used. The data we analyze consists of not only structured data but also a lot of unstructured data such as text logs, which can take a bit of time and effort to analyze. Moreover, application architectures are moving away from large, unwieldy affairs and toward becoming more standardized and decentralized through microservices, and advanced analysis methods are now indispensable.

Being able to use a system capable of managing and making advanced use of data as a common platform would not only make it possible to focus on enhancing the value of services and systems, which is ultimately what we should be doing, but it would also help us to generate new value by using data in ways we were previously unable to. Enter our new data analytics platform, illumino.

This article describes illumino's features, the issues it has so far resolved, and how this was achieved.

3.2 Introducing illuminok

3.2.1 About the illumino Data Analytics Platform

■ Challenges

IIJ has many systems running, the number and scale of which grow every year. So far, the approach to managing the data generated by these systems was to implement the necessary and sufficient measures for doing so on a system-by-system basis. While some projects had evolved with the implementation of advanced data analytics, many services were left operating on a "necessary and sufficient" basis given cost and man-hour issues. Underutilized data here not only represents potential value for those systems but may also embody value for the company as a whole.

Solutions for utilizing data have evolved rapidly in recent years, making it easier than ever to efficiently manage large amounts of data and perform advanced analysis, but the problem with implementing solutions on a system-by-system basis is that it scatters the necessary cost outlays as well as the management and analytical expertise.

■ Solution

To address these issues, we built a common platform that makes it possible to implement data management and analysis easily and at a low cost—namely, the illumino data analytics platform. This solution provides, as a service, the data storage and analysis tools necessary for analyzing data along with system operations and implementation support from dedicated engineers.

This solution is available to anyone running a project within IIJ. Because it is provided as a service, it becomes available immediately after someone applies for it.

Let's review the benefits of this solution.

- Data storage
  Can store large volumes of data safely at low cost
- Analysis tools
  Advanced tools available
- System operations
  No need to handle this on the user project end
- Implementation support
  Dedicated engineers / data scientists are available to assist

With many of our internal systems now using illumino, the utilization and analysis of data at IIJ has been making forward progress.

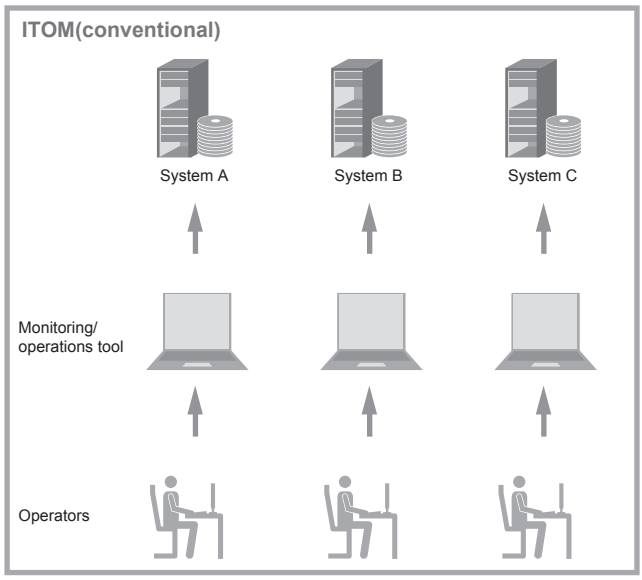
3.2.2 What is Data Analytics?

■ ITOA

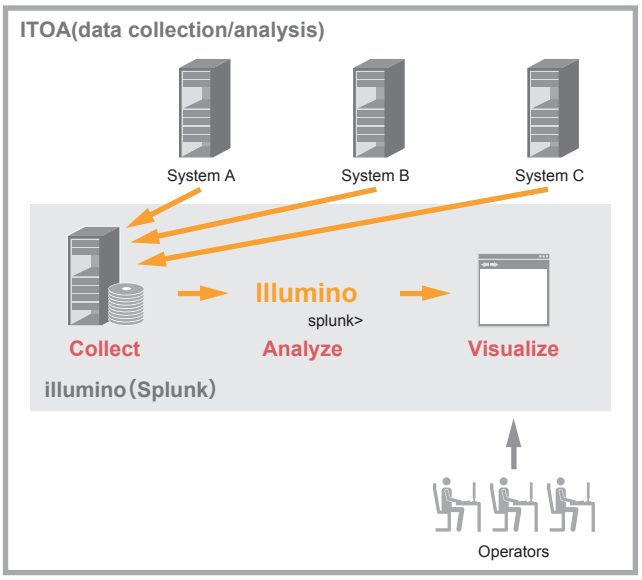
No doubt many people sense an increasing reliance on IT in their regular activities. With that increasing reliance, a high level of stability in system operations is being demanded. The systems, meanwhile, are increasing in number and the services based on them are becoming larger and more complex. Naturally, amid this ongoing evolution of systems and services, IT operations must also change. To provide the requisite highly stable level of operations, we need to transition from the conventional ITOM (IT Operation Management) paradigm to ITOA (IT Operations Analytics).

Say a particular service experiences a failure. It was often the case with conventional IT operations for the investigation team to examine data managed separately on each system, ask the relevant organizational units to cooperate in the investigation, and repeat the process until the problem was solved. Naturally, as the scale and complexity of services increases, it is often the case that the system on which users are aware of an error, or on which monitoring tools and the like have detected an error, is not the direct cause, and the number of associated systems (i.e., systems suspected of being the cause) also increases.

ITOA, as the name suggests, is a means of solving issues through the analysis of IT operations. Enabling the efficient collection of large amounts of data along with high-speed searching and advanced analytics logic makes it possible to automatically search data (logs) relevant to the error(s) detected when a fault occurs and swiftly track down the root cause. Trying out a range of data analysis also makes it possible to predict potential issues and new areas in need of attention.



Data is managed on a system-by-system basis



Data is collected and analyzed/visualized
Facilitates integrated operations

Figure 1: How ITOA Works

Overview of Splunk

Data collection, high-speed searching, and advanced analysis are key to making ITOA work. Of the many ITOA-related solutions out there, IIJ selected Splunk Enterprise from Splunk Inc.

Our reasons for selecting Splunk Enterprise include:

1. Had been used within IIJ before
   We had used it on some projects and thus already had a strong skillset
2. Enterprise grade track record
   Splunk has a track record of being used in large-scale, high-capacity projects worldwide, and we believed it would perform well in the use cases we envisioned for the IIJ common platform
3. Flexible system configuration
   We wanted to build it as an internal IaaS for reasons relating to connectivity with and security of the data input side of the system. But it also offers the flexibility to connect with SaaS offerings such as Splunk Cloud Platform and Splunk Observability.

Data collection

Data comes in all sorts of forms, and in almost all cases it needs to be processed a number of times to achieve the desired result. To achieve this with typical data-using applications, you had to structure the data beforehand to make it easy to work with.

In the process of utilizing and analyzing data, we commonly come across new issues and situations in which a new approach is needed. In such cases, pre-structured data can actually be more difficult to work with, or the structure may be unsuitable to begin with such that the data is not up to task, and at times we have to redo everything from the data structuring step.

Consider a situation in which we find ourselves needing to take our analysis back further in time. The original data will often be archived for storage efficiency reasons, and it will not always be in a state that is easy to work with. Even if it is in a workable state, the degree of processing difficulty rises in proportion to the amount of data.

Preprocessing and structuring the data can reduce the amount of data and increase search speeds, which can be quite beneficial for certain purposes. But what if you could eliminate the time and effort involved in preprocessing, reduce data storage costs, and increase search speeds?

A major advantage of Splunk is that data can be stored in an unstructured, unprocessed form and searched at high speed. Users input raw data without worrying about the data type. Many ways of getting data into Splunk are available—Splunk-specific forwarders are easy to set up, and it can also work with common forwarders like Fluentd.

Even before getting into data utilization requirements, simply putting data into illumino for storage and management already offers decent benefits. It's no exaggeration to say that the act of collecting data itself can create value, because bringing lots of data together and analyzing it as a whole enables the discovery of new value.

Harnessing high-speed search infrastructure

Data forwarded into illumino (Splunk) is manipulated in various ways to enable high-speed searching. With randomly stored data, search times increase in proportion to data volume, even if ample high-speed storage, CPU, and memory resources are available, and the server costs are also high.

The concept of time series is key here. A lot of data is timestamped. On illumino (Splunk), data is always sorted chronologically.

The statistics show that the searches we need to perform are often on relatively recent data. The most recent data is stored on high-speed SSD storage, and older data is stored on low-cost object storage. illumino (Splunk) returns results seamlessly even when searching data on object storage, so users need not worry about the storage lifecycle. Searches on data in object storage are slightly slower, but the illumino operations team monitors the object storage operating status and tunes the system to achieve optimal data allocation.

We build high-performance search servers with ample CPU and memory in parallel. Depending on the search specifics, a lot of CPU and memory is at times needed to provide the processing power for dealing with unstructured data, but we do maintain sufficient responsiveness for real-world use. Standalone systems (services) are costly, so there are limits to how much server performance is affordable, but common infrastructure allows the costs to be distributed and thus makes abundant infrastructure available. This allows us to use the strengths of Splunk and the cost advantages of common infrastructure to achieve high-speed searching while keeping a lid on the costs borne by each individual user.

Advanced use of Splunk's high-speed searching

For reasonable amounts of data and search details, we are able to achieve sufficient responsiveness using the powerful search capabilities of Splunk products and the illumino environment. But things are less than efficient in some cases when using thousands or tens of thousands of devices within the basic infrastructure to search through huge amounts of data, or when complicated search conditions need to be applied.

In cases like this, you can also structure some of the data imported into Splunk, either when it is imported or at a time of your choosing. The original, raw data is also preserved in these cases, so there is no loss of analytical flexibility. Although the structured data does consume additional storage, the increase in the amount of storage used can be kept relatively small because the system uses something akin to reference pointers to the original data.

Conventional methods often require privileged users—like data administrators—to preprocess data to structure it and so forth, but a big advantage with Splunk is that users can set this up themselves. Clearly, the ability of users to try this out themselves in conjunction with the requirements analysis step will make it possible to achieve the desired output sooner.

Some level of skill with Splunk is needed if users are to take advantage of these Splunk features. At IIJ, therefore,

support is provided by a dedicated team as a means of reducing learning costs on projects that use the system.

Visualization

Being able to search massive volumes of data at high speed is not the only thing. It is also important that users are presented with results that they can recognize as meaningful data. This is where data visualization comes in.

Extracted data is often presented in a table-like format. Splunk provides "visual effects" that let the user easily transform table data into various types of charts and the like.

In addition to simple line graphs and bar graphs, the user can easily work with dozens of other effects via a GUI. Visualized searches can be executed periodically and presented as a report, and the user can set up a dashboard to display searches on a single screen.

Analysis language

Data search and analysis is accomplished via Splunk's own simple but powerful language, SPL (Search Processing Language). SPL is similar to SQL as used in relational databases in that you specify a data source and filter results according to user-defined conditions, but what makes it powerful is the analysis features that follow that. SPL is a "one-liner" programming language whereby commands are chained using the "|" (pipe) character. Data extracted by a search can simply be passed into an analysis command using a pipe.

For example, say you want to count the number of occurrences of something in chronological order. You can simply add " | timechart count" after the search result. If you wanted to display the number of occurrences for each HTTP status chronologically in a web server log, you would use " | timechart count by status". This is intuitive and simple, so anyone can get up and running with it quickly.

3.3 Challenges and Solutions

The discussion so far has been about the illumino project and its internal system, Splunk. This section looks at challenges and solutions in actual illumino use cases.

3.3.1 Storage, Management, and Visualization of Large Amounts of Data

Challenges

IIJ has many systems running internally, and we use a common system for these systems' network and server infrastructure unless there is any special reason not to. This is the so-called "internal IaaS" approach. The systems are made up of thousands of servers and network devices, and until now, we had been collecting the various logs and metrics in a dedicated system.

This infrastructure has a relatively long history, and although we are able to see information from the logs and metrics using simple visualization tools, we had not been making any further use of or performing any additional analysis on the data. The large scale of the infrastructure means that the amount of data is also large, and reducing the costs involved in storage and management has been an issue we need to address.

Improving data storage and management

The first step was to start bringing this data into illumino. We had already been collecting data using tools like Fluentd, and since the data forwarded to illumino does not need to be structured, it can be sent directly via Fluentd, so all that had to be done was to add illumino as a forwarding destination. That is, we were able to improve cost performance just by adding a simple setting.

Data utilization

To look at the data on conventional systems, you had to view it graphed in a predetermined format on a dedicated system webpage or ask the relevant team to extract the data for you.

With illumino, authorized users can search the data themselves. They can produce graphs equivalent to those of a conventional system with a few lines of SPL, and since the original data is preserved, they can now set the search period, data category, and the like as they see fit.

A whole range of systems use this internal IaaS offering, and the importance of log and metric analysis varies from system to system. Some are more rigid to data changes while others are more tolerant. Among those that are more rigid, there are those for which CPU load is an important indicator, while disk I/O loads are important on others. Simple analysis and visualizations are provided in common report types and dashboards, and for some projects, custom reports and dashboards that focus on the indicators important for each system are created, providing analysis and visualizations suited to the operations in question.

There are also examples of logs from each system that uses the internal IaaS being imported and the analysis and visualizations being linked with the internal IaaS data to create visualizations of the relationships between each system and infrastructure usage. Visualizing system usage and infrastructure loads in an integrated manner like this helps to improve operational precision.

3.3.2 Data Sharing between Systems/Services

■ Challenges

Customers who use IIJ's services and teams that use internal systems often use multiple systems in combination. They also use combinations of components such as servers and networks, and use combinations such as a business application and a security service. Since individual teams run each of the systems, there was no linking of the data being managed. In the event of a system failure or when conducting surveys/analyses of usage, data had to be looked at across the systems being used, and coordinating the relevant teams took time and effort.

■ Data collection and sharing

As a shared system that can handle unstructured data, illumino is able to bring together data from a range of systems. It offers sufficient advantages just in terms of analyzing data from single systems, but the advantages are even greater when a number of systems and services are used in conjunction with each other.

Say you are using a business application and a security service in combination. Suppose an error appears on the business application screen. And suppose the user reports the details of the error: account name, time of error, etc. Based on that report, the system operator will

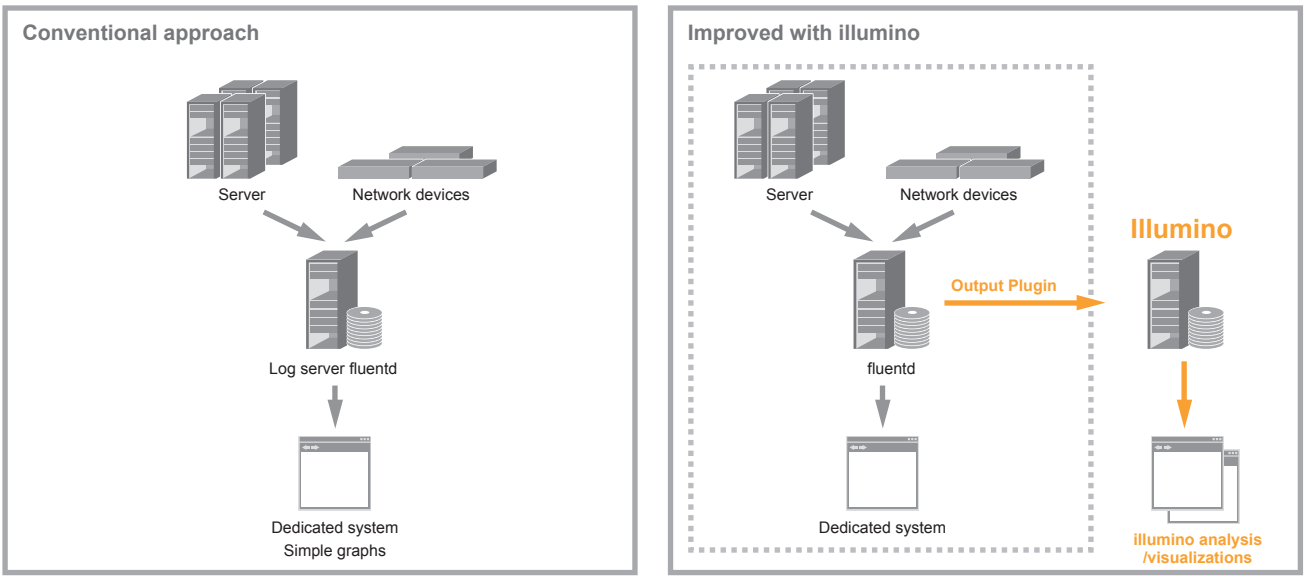


Figure 2: How Data is Utilized with illumino

start looking through logs and the like to identify the cause. If the business application turns out to be the cause, this is all relatively simple. If you find an error indicating the cause in the logs or whatnot, you can then proceed to the next response step.

What if it's the security service and not the business application that is to blame? Even if the business application log shows an error, this may only go so far as to suggest that the cause lies in the security service, and there may be cases in which no error appears at all. In such cases, the next step would be to investigate the security service side of things. The business processes are designed so that the relevant teams can coordinate smoothly with each other, but some degree of time and effort is still required if this inter-team coordination is not systematized.

Collecting all the data in illumino makes it possible to search and analyze the data in a way that takes into account relationships between datasets. In the above example, the security service can be investigated in conjunction with the investigation of the business application, where the overall investigation started.

Creating an operations tool that links multiple systems in a way that suits the business process design makes it possible to greatly reduce investigative time and effort.

Data security

Although the benefits of data collection and sharing are large, data security also requires careful consideration. Useful data often encompasses sensitive data. It is important to closely examine each system's data security situation and to manage what is forwarded to the shared system and the scope of what is disclosed and shared.

illumino achieves a high level of data security by setting the system up on dedicated internal infrastructure to protect the data and by configuring fine-grained data access permissions.

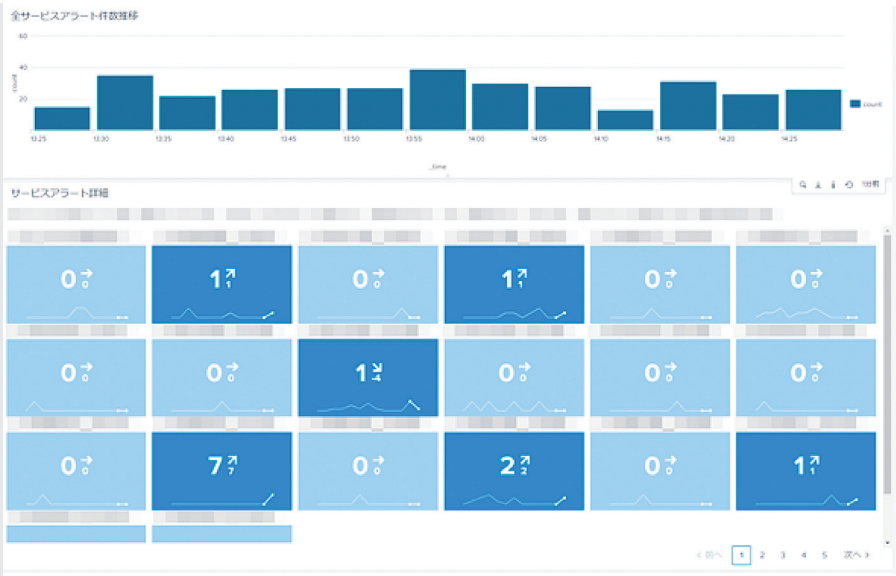


Image of integrated service monitoring dashboard
Multiple services can be monitored on via single screen

Figure 3: Image of Combined Service Monitoring Dashboard

The permissions need to be managed to a high level of quality, so there is a considerable load and costs involved, but we use role-based permissions management, which facilitates flexible, smart operations in accord with user needs.

3.3.3 Machine Learning

Challenges

One method of converting large amounts of data into useful data is machine learning. It has become relatively easy to acquire machine learning expertise in recent years, and there are real-world examples and solutions on offer, so the need for machine learning within IIJ is also increasing. We found ourselves unable to keep up with this rising need, however, as the task of setting up infrastructure for processing large amounts of data and the time and effort involved in data manipulation do present a high bar.

What is machine learning?

Machine learning is a method of extracting patterns from the original data for prediction and analysis. Analyzing large amounts of complex data using statistically verified methods makes it possible to identify patterns that were difficult to find with a rule-based approach in which operators set hypotheses and performed the design and implementation themselves. The main use cases for the patterns extracted are anomaly detection, prediction, and classification.

One great advantage of Splunk is that it works very well with machine learning by virtue of being a system capable of collecting large amounts of data for high-speed searching and analysis (Figure 4).

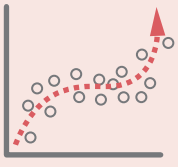
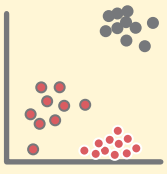
Anomaly detection	Predictive analytics	Clustering
		
• Deviation from past behavior—Historical AD • Deviation from peers—Cohesive AD • Unusual change in features • Training data available	• Features > predict values—Prediction • Features > predict category—Classification • Time > predict values—Forecasting • Preventive maintenance • Anomaly detection based on deviation from predictions • Training data available	• Identify peer groups—Clustering • Correlation between target variable and feature • Reduce feature granularity • Use rare groups to detect anomalies • Training data unavailable

Figure 4: Anomaly Detection, Prediction, and Clustering via Machine Learning

Machine learning basics

The machine learning logic itself is not unique to illumino (Splunk). The main machine learning logic provided by Splunk is actually implemented in Python and is open source.

In actuality, you could do machine learning without illumino so long as you have some basic Python skill and knowledge of open source machine learning software. But there are a range of benefits to be had from going through illumino (Splunk).

Advantages of using machine learning with illumino

To perform machine learning, you first need to prepare data to serve as a learning source (learning data). The accuracy of the learning data greatly affects detection quality.

Steps in creating source data suitable for machine learning include:

- Data extraction: Extracting appropriate learning data from a large volumes of original data
- Cleansing: Correcting and filling in noise and missing data
- Data conversion: Normalizing datasets to correspond in terms of data scale

On illumino (Splunk), this preprocessing is done using SPL. The user needs to decide how to extract, correct, fill in, and normalize the data, but once the decisions are made, the processing is easily expressed in SPL. The learning data thus created can be visualized using illumino's graphing features and the like, which is also useful in evaluating accuracy.

Once the learning data is ready, it is fed into the machine learning logic and turned into a model. This process of feeding the data into the machine learning logic is also expressed in SPL. In addition to expressing it in SPL, the user can also feed data into the logic and evaluate accuracy via a GUI. Accuracy can be evaluated visually via the GUI by, for instance, displaying the distribution of the learning data as a graph or using a slide bar to see what changes when different threshold parameters are used. There are data showing that this preprocessing accounts for about 80% of the machine learning process (Figure 5). Using illumino can greatly reduce preprocessing time and effort.

After creating the model, you evaluate it against the test data. The evaluation process can also be expressed in SPL, and the results can be used to issue monitoring alerts, easily visualized in graph and other formats, and

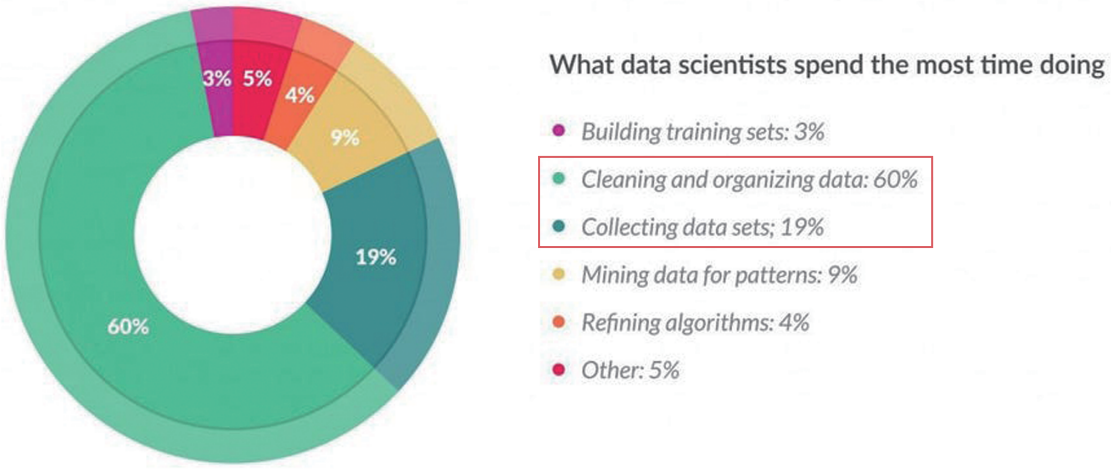


Figure 5: Time Spent on Machine Learning Tasks[*1]

*1 "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says", Forbes, March 23, 2016 (https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/).

so forth. The ability to complete all steps in the machine learning process on illumino has greatly lowered the bar to its adoption.

Example of machine learning in action

Here, we look at abnormal traffic detection as an example of how IIJ uses the system internally.

On one particular system, we monitor network traffic as one of the system health checks. Previously, the only monitoring configuration we did was to set fixed upper and lower bounds for traffic. As the system is aimed at business users, we know that traffic patterns differ greatly between day and night on weekdays and on holidays. We did at times observe trends in traffic that clearly differed from what was normal, even if not the sort of failure that would cause a complete interruption of service, so there were concerns about some equipment malfunctioning or experiencing abnormalities due to external factors, resulting in certain users being impacted. This was particularly difficult to detect based on fixed thresholds during nighttime and on holidays when traffic levels are relatively low.

To address such cases, we used machine learning to analyze past data and implemented threshold settings based on the probability density distribution (DensityFunction). By performing machine learning on past data on times when the system was operating normally, we can calculate, from a statistical perspective, what we term “rare upper and lower bounds”. By monitoring traffic data with these bounds as the threshold values and conducting a detailed investigation whenever the bounds were exceeded, we were able to identify phenomena that we were previously unaware of.

Fortunately, no failures have occurred since we implemented this machine learning-based monitoring, but if and when failures do occur, we will be able to respond swiftly and with precision.

3.4 Conclusion

We have discussed challenges around the utilization of data at IIJ, the solutions offered by the illumino project, and an example of the system in action and its effects.

The need for data utilization is bound to continue increasing ahead. At the illumino project, we are working to have illumino used on more systems and services and continuously striving to enhance our service offerings.

Takahisa Kudo
Analytics & Management System Development Section, Platform Development Department, Network Division, IIJ
Mr. Kudo joined IIJ in 2008. He is engaged in illumino project planning/operations and serves as illumino system administrator. He is also an internal evangelist for Splunk.