Big Data Analysis Platform Supporting Telecom Carrier Operations

TAKEDA Kimiko, YAMASHITA Takayuki

Abstract

Telecom carriers are expected to provide services matching the tastes and likings of individual subscribers quickly and in a timely manner in order to increase the sales per subscriber. To achieve this, not only the data scientists but also the content providers themselves should be expected to quickly and intuitively analyze the data of their customers and services. They would then be able to create and improve their services by optimally utilizing their own knowledge and expertise (self-service business intelligence, or hereinafter called "self-service BI"). This paper introduces the series of techniques from the introduction up until the establishment of a self-service BI environment, which is to be added to the existing analysis environment used by the data scientist.

Keywords

impact analysis, metadata, self-service BI, data source, data mart, DM

1. Introduction

One of the major sources of income for the telecom carriers is the profit earned from data communications of mobile devices such as smart phones. To increase the income from communication fees, the carriers need to consider not only to improve the communications infrastructures (networks and base station placements) but also to challenge the provision of services with higher added values.

An example of such high-added-value services is the content distribution services deployed on smart phones, which have the goal of increasing the communication fee income by enhancing customer pleasure. In other words, what is needed is quick, timely provision of attractive services that match the taste and liking of each customer.

On the other hand, in order to provide services that are fitting for each customer, it is required to choose the necessary and accurate data from the large amount of available data, perform analysis that matches the purpose and comprehend the results.

Under this situation, contents providers that have the requisite knowledge and expertise regarding services are required to have self-service Business Intelligence (hereinafter abbreviated to "self-service BI") for conducting quick analyses from the required perspectives. Below, we introduce the self-service BI techniques that range from introduction to establishment.

2. The Present Status and Issues of Big Data Analyses of Telecom Carriers

2.1 Present Status

(1) Analysis operations

The big data analysis can roughly be divided into routine analysis and as-needed analysis operations, each of which can further be divided into the advanced and normal analysis operations. These analyses are usually entrusted from the contents providers to the data scientists who compile the analysis reports and provide them to content providers. This means that the content providers have to withstand a certain waiting time and the fixed format of the analysis reports makes it difficult to verify the data easily from the various perspectives that they may need (**Fig. 1**).

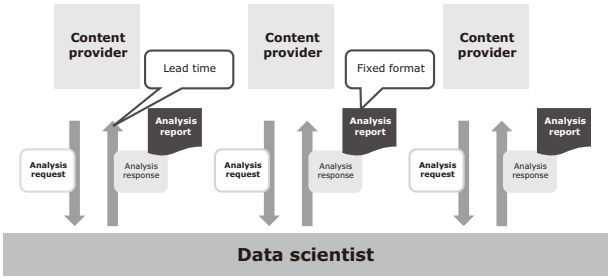

Fig. 1 Example of analysis operations.

(2) Routine analysis

This type of analysis collects the data in a certain time sector, such as daily, weekly or monthly according to a routine method and compiles a routine report. The main routine analyses are as enumerated below. Type 1) occupies the major part of the routine analysis cases.

1) Aggregated usages of each service (Normal analysis)
2) Aggregated entries/withdrawals of each service (Normal analysis)
3) Aggregated totals of Key Performance Indicators (KPI) and Key Global Indicators (KGI)
4) Extraction of customer segments with similar needs, obtained by cluster analysis of service continuation periods, service usage frequencies and utilized contents (Customer segmentation) (Advanced analysis)
5) Webflow line analysis based on the bounce rates, exit rate, CVR[*1] and cart abandonment rate, etc. (Advanced analysis)

(3) As-needed analysis

This type of analysis compiles a limited report on a specific need such as a campaign. Its operational range includes detailed reports on results obtained from routine analyses and targets the extraction of procedures such as those used in campaigns. The main as- needed analysis items are enumerated below, with type 7) occupying the major part of the as-needed analysis cases.

1) Issues are extracted from the entry, usage and withdrawal service areas and countermeasures are suggested. (Advanced analysis)
2) Visualization of the entry and withdrawal paths of good customers (customer journey) (Advanced analysis)
3) Prediction of the probability of entry into other services (Entry prediction model) (Advanced analysis)
4) Prediction of the probability of withdrawal from a service (Withdrawal prediction model) (Advanced analysis)
5) Extraction of recommendations of contents (Recommend model) (Advanced analysis)
6) Prediction of contact methods and timings for the fostering of good customers (Advanced analysis)
7) Extraction of the target customers of procedures such as campaigns (Customer segmentation) (Advanced analysis)

(4) Data source[*2]

The main data sources used in the routine and as-needed analyses are the customer attributions data, customer behavior data (viewing history, search history, download logs, etc.) and log data of various contents. Their major characteristics are that the data is sometimes of a very large scale from some millions to billions of items and that a large variety of data formats are handled[*3].

2.2 Present Issues and Improvement Measures

Fig. 2 shows the issues of the analysis operations, their causes and corrective measures by focusing on the content providers as those requesting the analysis operations shown in Fig. 1.

With regard to the traditional analysis operations, the content providers ask the data scientists to compile analysis reports and the scientists then develop and submit such reports. This method is inevitably accompanied by waiting

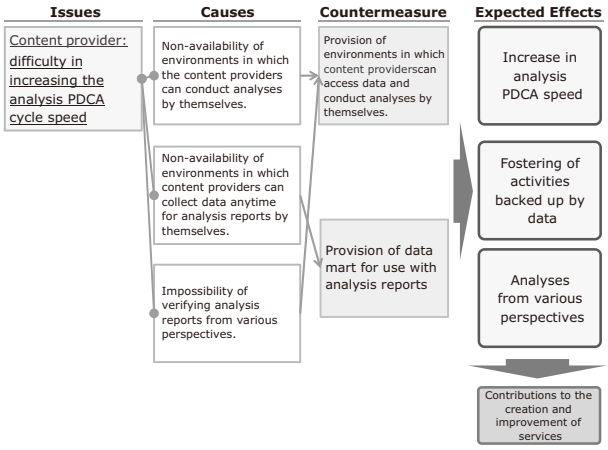

Fig. 2 Issues and their countermeasures.

[*1] CVR (Conversion Rate): Rate of results obtained from a website such as advertising site. The results include request of materials and purchases of services.
[*2] The data sources refer to databases, files, etc.
[*3] Usage of data is restricted in some cases due to the laws and regulations of different countries.

time for the report compilation, which sometimes makes it impossible to run the analysis PDCA cycle at a high speed. The specific causes of this issue are as follows:

1) Non-availability of environments in which content providers can conduct analyses by themselves.
2) Non-availability of environments in which content providers can collect data for analysis reports anytime by themselves.
3) Impossibility of verifying analysis reports from various perspectives.

The countermeasures for solving these issues include the introduction and establishment of the self-service BI with which the content provider can conduct analyses by themselves (countermeasures for 1) and 3)) as well as the Data Mart (DM) as the data source for the self-service BI (countermeasure for 2)).

3. Introduction and Establishment of DM and Self-service BI

3.1 Introduction of DM

The DM should be introduced without delay in the following flow so that the analysis PDCA can be advanced in a fast cycle.

(1) Design of DM

The DM can be classified into the master data and achievement data. The master data is the collection of basic information such as the customer and service names, while the achievement data is the collection of exchange-related data such as the usage counts and download counts, most of which are time-series data. Such a large amount of time-series data should be partitioned on a per-weekly, monthly or quarterly basis to improve the access performance. Since the data is accessed per service, data should be grouped per service in order to localize the access records.

(2) Selection of data sources

In order to generate a DM, it is required to obtain data from multiple data sources (master and achievement data).

This is an issue related to (3), "Implementation of DM creation processes". However, creating a DM by joining multiple data sources at once leads to a massive consumption of resources, such as of CPU and memory. To avoid such situations, it is necessary to incorporate a mechanism that limits the number of joined data sources to two or three and executes the process several times.

(3) Implementation of DM creation processes

The number of achievement data items is much larger than the master data. When such data is processed, the processing performance is affected greatly by the way the data sources are joined. It is therefore necessary to select a joining method that matches the environment.

The GUI tool utilizing metadata can be introduced to create the process and improve the maintenance efficiency. The tool can be used to collect metadata (definition and attribute information) from various data sources and visualize the locations and definitions of data as well as the flow of data and processing. This makes possible the instant identification (impact analysis) of the data and processing flow, thereby enabling quick investigation of the cause of producing improper operational data (**Fig. 3**). In addition, collection and distribution of the various data can be performed efficiently and flexibly by adding

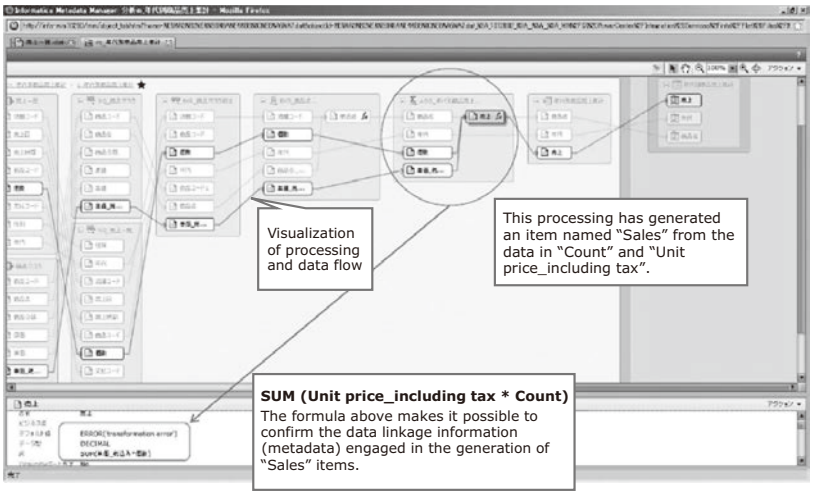

Fig. 3 Example of impact analysis.

or modifying the adopted data source via the GUI as required.

(4) Registration of processing in scheduler
The amount and number of DMs are variable depending on services. At the same time, extracted DMs may sometimes contain data that can cause system faults (such as the line feed code and comma). Based on such events, it is necessary to set scheduling by executing the parallel processing, synchronizing transactions and check points restarting with the batch processing so that the processing can be completed within the specified period.

(5) Masking of data
Data is masked as required during DM creation. However, there are cases in which data analysis in the specified schedule becomes difficult. This is because the masking such as hashing takes a longer processing time than expected. Consequently, the processing method should be implemented by taking the data masking time into consideration.

3.2 Introduction of self-service BI

The self-service BI is organized by the server and client. The server provides a disclosed analysis environment and data security and the client is responsible mainly for the closed analysis environment that is intended mainly for individuals. The assessment of the introduction of the self-service BI is performed from the following viewpoints.

(1) Selection of self-service BI tools
For the client, an analysis tool with intuitive operation that is easy for the end users to use is to be selected.
For the server, an optimum tool is selected by conducting studies from the viewpoints of (2) to (7) below.

(2) Availability
An environment in which the person in charge of services can analyze data anytime is required to advance the analysis PDCA cycle at a high speed. For this purpose, it is necessary to introduce servers that not only allow the HW but also the tools to deal with an available configuration.

(3) Performance
Slow response is a cause of disfavor for users, so an environment with full awareness of performance criteria should be prepared. As it is inefficient to analyze the big DM directly every time from the self-service BI client, a data set (extracted files) is sometimes created as a subset of the DM on the client side.

(4) Monitoring
This is associated with availability, but a mechanism for monitoring should be implemented to allow quick application of countermeasures in a case where an unexpected event occurs in the self-service BI environment.

(5) Authentication
Implementation of direct data source access from the self-service BI client is possible but, to ensure security related to data, it is recommended to implement data access control through the self-service BI server. In addition, it is also necessary to consider the linkage with the authentication platform at the corporate level and to define user operation.

(6) Authority
Similarly to the authentication, the data read/write and the publication of analysis results are controlled from the self-service BI server.

(7) Version up
Because of the short-lifecycle tendency of the BI tool, it is a product subjected to frequent version upgrading. Since a mixed presence of different versions could render data sharing, etc. difficult, a system for SW configuration management is implemented in order to achieve version upgrading smoothly.

3.3 Establishment

The self-service BI and DM should be established after introduction. The major tasks to be enforced in the establishment assessments are as follows. These tasks are expected to impart the self-service BI and to improve the skill of the content providers.

(1) Provision of templates and guidelines
Templates and guidelines for analysis are provided periodically for the content providers who are end users of telecom carriers. The templates should be created by gathering the opinions of end users and by selecting the most common factors.

(2) Provision of periodical training
Periodical training and hands-on experience for new users are to be held. To improve the data literacy as well as the knowledge of the self-service BI, opportunities for studying general knowledge items, including mathematics, statistics, database and SQL, should also be provided.

(3) Introduction of help desk
A help desk is installed near the service provider to enable quick, face-to-face responses to inquiries.

(4) Recommendation of BI tool qualification acquisition
Acquisition of BI tool qualifications offered by BI tool vendors is recommended in order to achieve an improved analysis capability and to enhance motivation.

(5) Visualization of data and processing

A system for visualizing data is to be introduced so that the analysis target data can be found efficiently and that data illegality can be detected as quickly as possible.

4. Conclusion

In the above, we introduced the flow from introduction to establishment of the self-service BI and DM as analysis infrastructures for supporting the operations of the telecom carriers by using big data. Big data analysis is currently in the stage of Business Intelligence, or is evolving "from data to information". In the future, it is expected that Business Analytics, investigative type data mining (self-service BI) and Artificial Intelligence will attract further attention, which is the stage evolving from "information" to "knowledge and expertise".

Authors' Profiles

TAKEDA Kimiko
Senior Expert
Common Carrier Solutions Division

YAMASHITA Takayuki
Assistant Manager
Cloud Platform Division

The details about this paper can be seen at the following.

Related URL:

Informatica
https://www.informatica.com/
Tableau Software
http://www.tableau.com

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese English

Vol.10 No.3 Special Issue on Telecom Carrier Solutions for New Value Creation

Vol.10 No.3
July 2016

Special Issue TOP

General Paper

NEC Information