Paper 11780-2016

PROC IMSTAT Boosts Knowledge Discovery in Big Databases (KDBD) in a Pharmaceutical Company

Yoshitake Kitanishi, Ryo Kiguchi, Akio Tsuji, and Hideaki Watanabe, Shionogi & Co., Ltd.

ABSTRACT

In recent years, Big Data has been in the limelight as a solution for business issues. Big-Data-mining implementations have begun in a variety of industries. The variety of data types and the velocity of increasing data have been astonishing, whether represented as structured data (stored in a relational database) or unstructured data (e.g., text data, GPS data, image data, etc.). In the pharmaceutical industry, Big Data means real-world data, e.g., electronic health records, genomics data, medical imaging data, social network data, etc. Handling these types of Big Data often requires a special environmental infrastructure for statistical computing. This document covers two case studies: 1) the implementation of SAS® In-Memory Statistics for Hadoop™ (IMSTAT) as a large-scale parallel computation environment; conversion from business issue to data science issue in pharma, and 2) data handling and machine learning for vertical and horizontal big data by using PROC IMSTAT; the importance of analysis result integration; and caution points of big data mining.

INTRODUCTION

SAS® may be the most used programming language in pharmaceutical companies, rather than other languages (e.g. C, Java, etc.). A possible reason for this is that statistics are important for inferring a population from clinical or non-clinical trials. Thus, many SAS programmers work for pharmaceutical companies. Meanwhile, 'Big Data' has become a familiar phrase nowadays. To investigate various hypotheses, pharmaceutical companies attempt to analyze combinations of datasets from a wide variety of Big Data. With this background, launching new pharmaceutical products becomes more challenging. Each pharmaceutical company must discover as many drug species as possible and enhance the probability of succeeding in clinical trials. Big Data is one of the measures to advance this issue.

However, because Bio-statisticians and SAS programmers in pharmaceutical companies have been conducting analyses by using simple calculation environments (personal computers, single servers, etc.), they are unfamiliar with parallel computing environments. We built a Hadoop distributed file system (HDFS) to handle Big Data in-house, along with a machine learning environment such as Mahout system. However, when we tried analysis implementation, we discovered that handling Hadoop requires considerable skill. It was difficult to increase the number of users; in fact, they did not increase. At that time, we renewed our awareness that many SAS users work in our office.

In this specific pharmaceutical company environment, IMSTAT is a very useful system from the viewpoint of effectively using the Hadoop as IT asset. After the IMSTAT system construction, the 'data science cycle' began in our activities. In situations where a wide variety of data exists, e.g., open databases, commercial databases, and in-house databases, many ideas for solving various research questions began to emerge. The analysis flow proceeds as follows:

1. Data selection, ETL (Extract, Transform, Load);
2. Comprehension of the data feature, determination of the analytical approach, implementation of the analysis;
3. Visualization of the output; interpretation, explanation, and submission of the proposal.
4. Review and retrial from a heightened perspective.

Figure 1 shows the analysis flow.

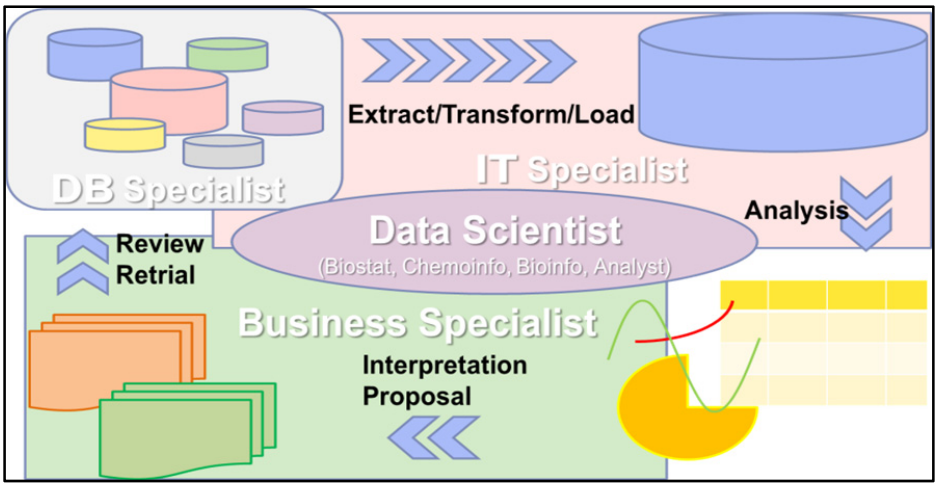

Figure 1. Data Science Cycle

In this document, we describe the points below.

- R&D Big Data positioning in pharmaceutical companies
- IMSTAT and relevant tools, Visual Analytics, Data Loader, in our system architecture
- Variety of Big Data
- Analysis and visualization approach for Big Data

R&D OF BIG-DATA POSITIONING IN PHARMACEUTICAL INDUSTRY

Real world data (RWD) is typical Big Data in the pharmaceutical industry. The data comes from medical actions by doctors, e.g., electronic health records (EHR), claim data, disease registries, etc. RWD is classified as un-controlled data in pharmaceutical R&D. Meanwhile, clinical and non-clinical trial data are usually collected under controlled situations. The analysis approaches to these two types of data are basically different. Un-controlled data is for hypothesis formulation: we usually find associations/rules using machine learning. Controlled data is for testing a hypothesis, usually by using inferential testing.

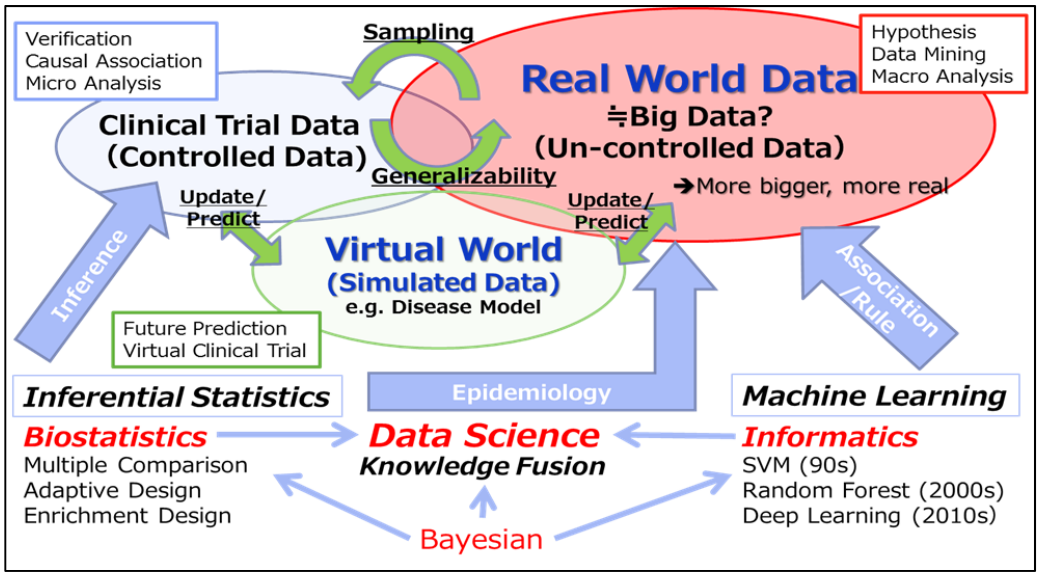

Figure 2. Data fusion between controlled data and uncontrolled data

In addition, real-world data is bigger than clinical trial data. Therefore, the IT approach to these two types of data is also different. Technologies to handle Big Data have been developed rapidly. Figure 2 shows a brief concept of the data fusion between controlled data and uncontrolled data.

IMSTAT AND RELEVANT TOOLS FOR OUR SYSTEM ARCHITECTURE

Powerful computing environments are necessary to help data scientists handle and analyze Big Data. However, despite such a demand, it is difficult for data scientists to build a parallel computing environment. The help of IT experts or computer scientists is needed. Thus, we teamed up with them for this project, and connected computers to construct a parallel computing environment. Programming language mastery became the next challenge. We were able to lower the programming hurdles for Hadoop by using IMSTAT. In addition, we used SAS® Data Loader for Hadoop as a data import system from various data formats/databases to HDFS, and SAS® Visual Analytics (VA) as a visualization system. These cooperative systems cover a series of flows from data input to result output. Figure 3 shows our system architecture and explains the background of the system.

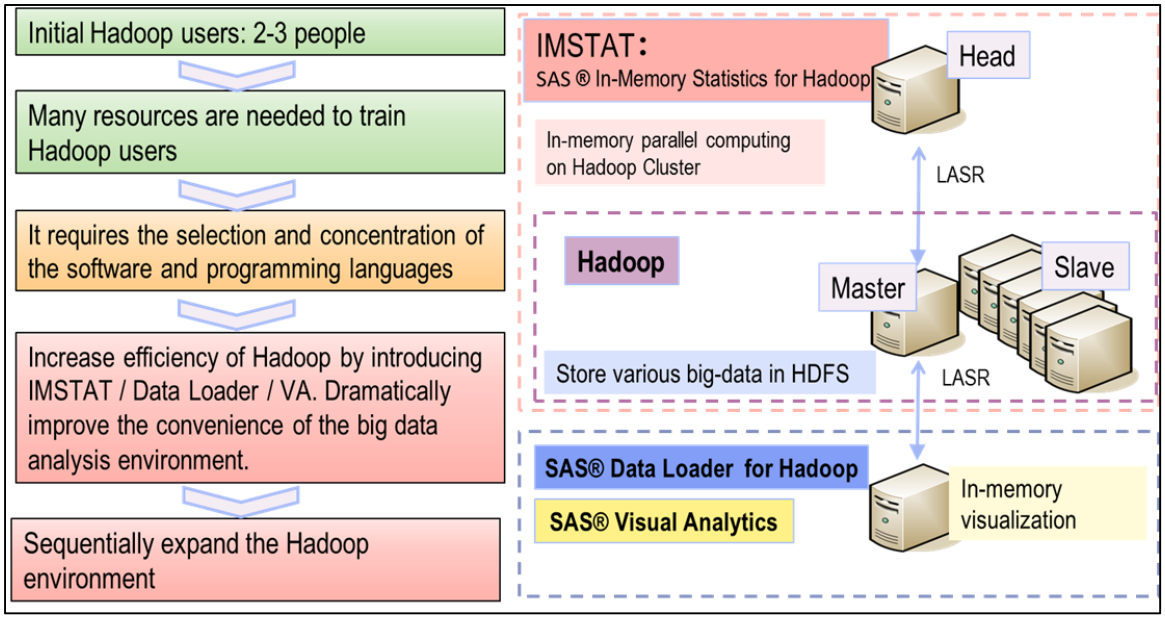

Figure 3. Background and system collaboration diagram, including IMSTAT / DataLoader / VA

VARIETY OF BIG DATA

Data generation and collection continues to accelerate. Big Data has a wide variety. However, it can be roughly classified into two types: vertical and horizontal. Transaction data is an example of vertical data. Point of Sales (POS) data is also classified as vertical. In the pharmaceutical industry, the Adverse Drug Event Report Database and Real World Data (Electronic Health Records, claim data, disease registries, etc.) are representative examples. The point of analyzing vertical Big Data is to efficiently find associations or rules.

On the other hand, genomic data and clinical trial data are examples of horizontal Big Data. This data type includes roughly three analysis approaches: multiple comparisons, dimension reduction, and variable selection. However, horizontal Big Data has the famous ‘Curse of dimensionality’ known as high dimensional issues (e.g. increasing the noise, shortage of observations to estimate well, etc.). Whether a method is robust to this issue may be the key factor when selecting an analytical method. Table 1 summarizes this information.

Table 1. Vertical and Horizontal Big Data

Big Data Type	Analysis approaches	Analysis Types (Examples)
Vertical (Rule Detection)	Sorting Algorithm Contingency Counting Algorithm	Contingency Table Analysis *Association Rule Mining* Signal Detection Time-to-Event Analysis etc.
Horizontal (Feature Selection)	Multiple Comparison	Bonferroni False Discovery Rate (FDR) etc.
	Dimension Reduction	Principal Components Analysis Canonical Correlation Analysis etc.
	Variable Selection	*Random Woods* *Lasso (Least absolute shrinkage and selection operator)* *Elastic Net* etc.

ANALYSIS AND VISUALIZATION APPROACHES FOR BIG DATA

Sample code and examples that we found useful in our big data analysis are described in this section. Association Rule Mining, which is often used in POS data analyses, is an analysis method that can also be applied to the pharmaceutical industry, e.g., adverse drug event prediction, diagnosis prediction, etc. Association Rule Mining is a very powerful tool for finding rule combinations.

Association Rule Mining [IMSTAT]

```
proc imstat data=LASRLIB.Dataset;
arm                                    /*Association Rule Mining*/
item=COL1                              /*Item Colum*/
tran=ISR                               /*User ID Colum*/
        / minItems=1 maxItems=5        /*Item Counts*/
 itemsTbl                              /*Temporary table*/
          support(LOWER=15)            /*Support*/
          rules(confidence(LOWER=0.9) /*Confidence*/
          numrhs(upper=5 lower=1)      /*N of items in the right side*/
 numlhs(upper=5 lower=1))              /*N of items in the left side*/
rulesTbl                               /*Temporary table*/
;
run;
```

A recommendation system is an application of the Association rule. Based on the enormous amount of information on people with similar background factors, the system might then be able to predict what will happen to an individual. The Recommend System predicts an outcome based on three types of data: background, rating, and event.

Recommend System [IMSTAT]

```
proc recommend  recom = LASRLIB.Datset;
 add LASRLIB.recomend /                      /*Recommendation project name*/
   item = item
  user = ID
  rating = count;
 addtable LASRLIB.rating  /type = rating               /*Rating table*/
 vars=(ID code eventname  count);
 addtable LASRLIB.event    /type = item;               /*Event table*/
 addtable LASRLIB.demo      /type = user;              /*Demo table*/
run;
method knn / label = "knn" k = 20 positive  /*Algorithm for similarity*/
similarity = pc seed = 1234;
run;
  predict / method = knn  label="knn" Num = 5          /*Prediction*/
             users = ("000001"~"0XXXX");
run;
```

Random Woods is a powerful variable-selection tool for vertical Big Data. Variable selection is based on importance. The degree of importance is calculated based on the statistics that increase or decrease when a new variable is added to the tree.

Random Woods [IMSTAT]

```
proc imstat DATA=LASRLIB.Dataset;
RANDOMWOODS OUTCOME/                              /*Response variable */
      INPUT=(COL1 COL2 COL3 COL4 COL5 COL6)    /*explanatory variable*/
      NOMINAL=(COL3 COL5 COL6)                  /*Category variable*/
      M=4                        /*Sampling number of explanatory variable*/
      LEAFSIZE=5                               /*Size of Leaf*/
      MAXBRANCH=2                              /*Maximum number of branch*/
      MAXLEVEL=10                              /*Depth of tree*/
      BOOTSTRAP=0.8                            /*Default:1-exp(-1)*/
NTREE=3000;                                    /*Number of tree*/
run;
quit;
```

Lasso and Elastic Net are powerful variable-selection tools used in genome analysis. The Lasso Approach has attracted attention in sparse modeling.

Lasso (Least absolute shrinkage and selection operator) [SAS/STAT®]

```
proc glmselect data=work.Data plots=all ;
         model OUTCOME=COL1-COL10
           / selection=lasso(steps=1000 choose=AIC) ;
run ;
```

Elastic Net [SAS/STAT®]

```
proc glmselect data=work.Data plots(stepaxis=normb)=coefficients ;
         model OUTCOME=COL1-COL10
           / selection=elasticnet(steps=1000 L2=0.1  choose=AIC) ;
run ;
```

The Sankey diagram is a graph representation method that allows the intuitive visualization of a complex event. For example, a Sankey diagram mounted on Visual Analytics is useful when expressing disease and prescription transitions.

Sankey diagram [SAS® Visual Analytics]

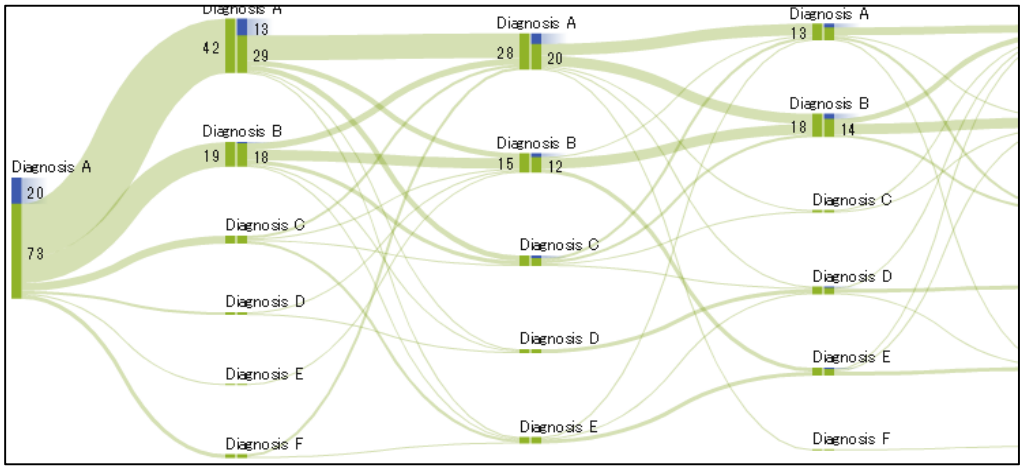

Figure 4. Visualization of a diagnosis transition using the Sankey diagram.

CONCLUSION

The keywords for tackling big data analysis are ‘Team building,’ ‘IT environment construction,’ ‘Analysis method selection,’ ‘Communication’ and 'Easy-to-understand result output'. However, the last 2 keywords are often overlooked, but important. Therefore, teamwork should be considered as the first priority. Analysis execution speed is important to accelerate the thought process. A computing environment that can handle huge amounts of data without stress is a great advantage. Once a data scientist captures the characteristics of the data, it is important to apply the appropriate analysis method for Big Data. Understanding the Data Science Cycle from a bird's-eye perspective and implementing the cycle is extremely important. We hope this paper will be useful for Big Data Analysis in pharmaceutical companies.

REFERENCES

Bellman, R.E. 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

SAS Institute Inc. 2015. SAS® LASR™ Analytic Server 2.7:Reference Guide. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2015. SAS® Visual Analytics 7.3: User's Guide. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2015. SAS® Data Loader 2.3 for Hadoop: User's Guide. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2015. SAS/STAT® 14.1 User’s Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yoshitake Kitanishi

Shionogi & Co., Ltd.

yoshitake.kitanishi@shionogi.co.jp

Ryo Kiguchi

Shionogi & Co., Ltd.

ryo.kiguchi@shionogi.co.jp

Akio Tsuji

Shionogi & Co., Ltd.

akio.tsuji@shionogi.co.jp

Hideaki Watanabe

Shionogi & Co., Ltd.

hideaki.watanabe@shionogi.co.jp