



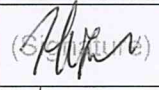


2017 Best Practices Awards

Leading Innovations in Analytics, Business Intelligence, and Data Warehousing

Deadline: April 27, 2017

** Indicates a required field.*

*Name of Nominated Company <i>(exactly as you want it to appear in print)</i>	ING Life Insurance Korea (nominated company)
*To which category are you applying? <i>You must submit a separate award entry for each category to which you are applying. Limit is two categories per organization. If you enter in two categories, be sure each award entry is tailored to its category.</i>	Check one below: <input checked="" type="checkbox"/> Advanced Analytics and Data Science <input type="checkbox"/> Emerging Technologies and Methods <input type="checkbox"/> Enterprise Data Warehouses <input type="checkbox"/> Data Management Strategies <input type="checkbox"/> Big Data <input type="checkbox"/> BI, Visual Analytics, and Data Discovery <input type="checkbox"/> BI and Analytics on a Limited Budget
*Lead Business Sponsor or Driver at Nominee's Firm Name, Title, Phone, E-mail, and Role	Chan-Soo Kim, Managing Director at 2e consulting, +82 10 2041 4799, cskim@2e.co.kr , Data Science Team Manager
*Signature and Date	 2017-04-26
*Lead I.T. Contact at Nominee's Firm (must be different person than above) Name, Title, Phone, E-mail, and Role	Dong-Heon Kim, Deputy IT Manager at ING life Insurance Korea, +82 10 7182 8183, Dongheon.kim@inglife.co.kr , Data Architect
*Signature and Date	 2017-04-26
Contact at Solution Sponsor's Company (If Applicable) Name, Title, Company, Phone, E-mail	Jong-Yeol Hyun, Senior Consultant at 2e consulting, +82 10 4255 8202, jacobgreen1984@2e.co.kr , Eun-Bi Shin, Senior Consultant at 2e consulting, +82 10 3127 7502, eunbis@2e.co.kr Min-Kyu Park, Senior Consultant at 2e consulting, +82 10 7765 4136, martin@2e.co.kr
*Signature and Date	 2017-04-26
May TDWI personnel contact you about speaking at a TDWI event?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

**Note: Your award entry form is not considered complete until you print and email back this page to TDWI at bwoodbridge@tdwi.org. The information contained in your entry form is used solely for the purpose of selecting winners for the Best Practices program, but is otherwise considered confidential by TDWI staff and judges. If you are*

selected as the Best Practices winner, your signature authorizes TDWI to promote your organization in TDWI public relations and marketing efforts.

A. BACKGROUND - Respond to all questions below.

Company Description. Describe your (the Nominee's) company in one paragraph (50 words maximum).	ING Life Insurance Company has continued growth in Korea since establishment in 1987 and made an increasing contribution to the overall advancement of Korea's insurance market.
Project Description. Describe the business purpose of your project in one short paragraph (300 words maximum).	The project was performed to construct the analysis environment, and to develop the Big Data reference model that encompasses the overall field of the insurance business in Korea.

B. SHORT QUESTIONS - Respond to as many of these questions as you can.

Depending on the solution you are describing in this award entry form, it is possible that some of the following technical questions may not apply to you. If your employer prohibits the disclosure of financial information, you may need to skip those questions. Even so, note that judges favor award applications that are as complete as possible.

Other Contests. Has this project been submitted to other contests? If so, which ones and when?	No
Rollout Date. What month and year did the system being nominated officially go into production?	On 31 December 2016 the big-data architecture and hybrid recommender system were developed and has been operated until now.
Active Users. How many business users use the system at least once a week?	10
Types of Users. What percentage of the users fall into the following categories?	40% Casual Users (View reports several times a week) 60% Power Users (Explore data regularly) ____% Customers/Suppliers ____% Other (Please specify): _____ 100%
Source Systems. What number of distinct source system applications does your solution draw from?	Enter an Integer: 3
Load/Update Intervals. What percentage of data is loaded in the following intervals?	____% Quarterly 20 % Monthly ____% Weekly 80 % Daily ____% Less than daily. Please specify the interval and update mechanism: _____
Data Volume. How much data is managed by your solution? Express this in whatever terms you commonly use, like records, tables, files, gigabytes, terabytes, etc.	<ul style="list-style-type: none"> ➤ Customers: more than 5 million ➤ Management accounts: more than 10 million ➤ Transaction data per month: more than 1 million ➤ The number of related tables: approximately 10 thousand tables

<p>Architecture and Method. Briefly list any methods or architectures that your solution employs. (Be brief; you'll describe these in detail later in your essay.)</p>	<ul style="list-style-type: none"> ➤ Software Architecture: HDFS, Spark, R ➤ Machine Learning algorithms: SOM(Self Organizing Map)-based two-step clustering, Item-based & User-based Collaborative Filtering
<p>What is the 2016 maintenance budget of your system? (Please put a check the correct range at right):</p>	<p><input checked="" type="checkbox"/> Less than \$100,000</p> <p><input type="checkbox"/> \$100,000 to \$500,000</p> <p><input type="checkbox"/> \$500,000 to \$1 million</p> <p><input type="checkbox"/> \$1 million to \$2.5 million</p> <p><input type="checkbox"/> \$2.5 million to \$5 million</p> <p><input type="checkbox"/> \$5 million to \$10 million</p> <p><input type="checkbox"/> \$10 million +</p>
<p>Team. How many full-time equivalent staff are on the current BI team, including external consultants and contractors? What percentage is external to the company?</p>	<p>10 Number of FTEs on Current Team</p> <p>0 % of External Consultants/Contractors in above</p>
<p>Roles. How many FTE staff fill the following roles? (Include external consultants in your FTE count.)</p>	<p><input type="checkbox"/> Business sponsors/drivers</p> <p><input type="checkbox"/> <u>1</u> Project managers</p> <p><input type="checkbox"/> BI architects/developers</p> <p><input type="checkbox"/> ETL architects/developers</p> <p><input type="checkbox"/> <u>3</u> Data architects/data modelers</p> <p><input type="checkbox"/> Subject matter expert</p> <p><input type="checkbox"/> <u>2</u> Business requirements analyst</p> <p><input type="checkbox"/> <u>4</u> Data analysts (e.g. people who analyze data)</p> <p><input type="checkbox"/> DBAs</p> <p><input type="checkbox"/> DW Administrators</p> <p><input type="checkbox"/> Data scientists/statisticians</p> <p><input type="checkbox"/> Trainers</p> <p><input type="checkbox"/> Consultants/Contractors</p> <p><input type="checkbox"/> Other notable roles with more than one FTE:</p>
<p>Initial Roll Out. Please indicate how long it took to roll out the initial system, the total cost to roll out that system, and the years until payback.</p>	<p>Time (from approval to initial roll out): 5 months</p> <p>Start Date: 2016. 8</p> <p>End Date: 2016. 12</p> <p>Cost (including HW, SW, Services, Labor): US \$2.5 billion</p> <p>Years Until Payback (or estimate): -</p>
<p>Stewardship/Governance. Describe the steering committee(s) or person(s) that set direction for the system. One paragraph total.</p>	<p>2e consulting was established in 1996 in Korea. It has been equally competing with global consulting firms and has been giving various professional consulting services on financial, public, logistic sectors. 6 members from 2e consulting had worked together to complete the ING life Big-data project. The project team consists of 1 project manager, 1 business analyst, 1 data architect, and 2 data analysts.</p>

C. BEST PRACTICES ESSAY

The open source based analysis environment was constructed by 2e consulting for the first time in the Korean financial industry, and the Big Data reference model that encompasses the overall field of the insurance business was developed. It is a very successful case of minimizing trial-and-error when domestic insurance companies introduce Big Data, and on providing support of analysis to be applied on the work.

Insurance data with high potential value and applicability

In the Korean insurance and financial field, most of the works are process through the online environment, and accumulated data sizes are big and the types are various. Depending on the data form, there are voice data (call consultation recording), text data (call counselor memo, claims memo, VOC memo, etc.), and structured data (customer information, FC information, contractual information, product information and homepage log information, etc.). In addition, as the new market combined with IT and finance has appeared, the financial industry is facing radical environmental change. There are continuous attempts to discover the value inherent in the data occurred internally/externally to find the opportunity within the environmental change of the financial industry. However, there are insufficient success cases and application effects of analyzing Big Data clearly reflected with the characteristics of the insurance company, and also almost no experiences on constructing the Big Data analysis environment, so each insurance company is suffering from continuous trial-and-error on the Big Data analysis business. Therefore, development and sharing are required on the standardization cases regarding algorithm, analysis environment, and governance that each insurance company can refer to in common.

Analysis Environment

In the financial industry, processing and analyzing data are very massive and enormous compared to other industries. The number of transaction data accumulated each second from the millions of customer accounts is approximately more than several billions for each table. In this environment, even requesting simple query generated great cost, so many companies generally adopted the scale-up method of highly-advanced server for various analyses. However, in this project, the approach was made in the scale-out method for the first time in the Korean financial industry to construct the open source software based analysis system environment. In place of purchasing one highly-advanced server with very expensive cost, we have constructed the multi-node cluster server which is composed of 7 nodes and has equal or higher performance.

Software Architecture

To compose the architecture for process and analyze large amount of data, the updated version of Open Source Software was actively introduced. For the basic framework, centralized architecture was realized through YARN based on HDP (Hortonworks Data Platform), and for the overall monitoring and operation, the design was made to enable management with Ambari and HUE through the web interface. Also, Spark was used to enable high-speed processing of the files stored in HDFS, and R and Sparkling Water were mounted additionally to supplement the insufficient statistical algorithm and mechanical learning.

Building Data Marts

The basic RDB data for product recommendation and actual result forecasting was all mounted on the HDFS through the arrangement processing, and customers' consultation memo and voice recording were duplicated after passing through the STT&TA process in the separate server. After the ETL process, the data was processed based on the Spark processing speed about 10~100 times faster than the existing RDB to construct the mart for analysis on the HDFS.

Customer segmentation before building recommender system

Collaborative filtering utilizes the product satisfaction index on the product possessed by the customer to recommend the product possessed by the customer similar to the target customer among the products not possessed, or recommend the similar product to the product possessed by the target customer. The collaborative filtering is intuitive and easy to understand, and the recommendation performance is also outstanding to be loved by many companies. However, the recommendation analysis is enabled only through the product preference index, so we can't say that the product was recommended after fully understanding the customer.

To enable customized product recommendation, the customer must be first understood, because to recommend a proper product, the customer's status must be understood. For example, recommending a 10% discount coupon to a restaurant on a customer in financial difficulty is a recommendation without understanding the customer's circumstances. To understand the customer most efficiently, listening to the customer's situation through 1:1 consultation with the customer is the best, but great time and effort are required to handle each and every one of the customers. Also, it isn't easy to fix the consultation time with the customer.

How can the customer be understood quickly and easily? We were able to find the clue on this answer from the cluster analysis. It is impossible to identify the characteristics of all customers, but it isn't difficult to cluster the similar customers into n groups to identify the characteristics of the relevant cluster. In addition, it was considered that the most important factor on recommending the product to the customer was the economic power of the customer. Therefore, the cluster analysis was performed for the purpose of identifying the economic power of the customer.

Modify the product level for recommender system

ING Life Insurance was releasing new insurance product regularly, so even though the names of the new & existing insurance products were different, there were cases of having the same contents (coverage details). In addition, due to the characteristics of the

insurance industry, most of the customers have 1~2 insurance products, so there was an issue of sparse matrix on recommendation analysis.

To solve this issue, we increased the product recommendation unit from the insurance product to the coverage unit to use it on the recommendation analysis model. By modifying the recommendation unit to coverage, the different product names but with same coverage details were classified as the same product. Also, one insurance product has several coverages, so there was an effect of reducing the sparse matrix issue that occurred on recommendation analysis.

Design user-item matrix for up and cross-selling

The conventional recommendation analysis is performed in the method of recommending the product that the customer may prefer among the products that are not possessed by the customer. This method is possible for cross-selling, but up-selling is not possible. In the conventional recommendation method, when the product purchasing amount is less than the average amount even though the customer is possessing the product, the up-selling recommendation method of recommending as much as the average amount additionally was added to realize the recommender system that enables both up-selling and cross-selling.

The analysis procedure of the recommender system realized in this project is show as follows:

1. Total customers are clustered into n groups (Self Organizing Map based two-step clustering)
2. Select one of the n groups
3. Calculate the average coverage amount for each coverage within the cluster:
 $\text{Sum (Customer coverage amount) / No. of customer}$
4. Calculate customer coverage preference score:
 $\text{Coverage preference score per customer} = \text{Customer coverage amount} - \text{Average coverage amount per coverage}$
5. Use the customer coverage preference score to produce the binary user-item matrix:
 $\text{Coverage preference index per customer} = 1$ if the customer coverage preference score is equal or higher than 0, and 0 if the customer coverage preference score is smaller than 0
6. **Perform recommendation analysis (Below)**
7. Select other cluster to repeat 3~6 times

6. Perform recommendation analysis (Details)

Among the customers, when the sum of the coverage preference index is 3 points or higher, perform recommendation through A method

A.1. Produce various candidate recommendation model: Item-based and user-based Collaborative Filtering with different hyper-parameters

A.2. Select the best model through model verification

A.3. Perform coverage recommendation with the best model

A.3.1. (up-selling) When the target customer is possessing the recommended coverage: The customer is possessing the amount less than the average coverage amount, so enable additional recommendation as much as the average coverage amount

A.3.2. (cross-selling) When the target customer is not possessing the recommended coverage: Recommend the relevant coverage product in the standard of average coverage amount

Among the customers, when the sum of the coverage preference index is less than 3 points, perform recommendation through B method.

B.1. Recommend 3 most popular coverages within the cluster that the target customer is included in

B.1.1. (up-selling) When the target customer is possessing the recommended coverage: The customer is possessing the amount less than the average coverage amount, so enable additional recommendation as much as the average coverage amount

B.1.2. (cross-selling) When the target customer is not possessing the recommended coverage: Recommend the relevant coverage product in the standard of average coverage amount

(1) Business Impact. What is the business value of the BI/DW/DM or analytics initiative?

The reference model that internalizes the Big Data analysis in the standard of insurance company's value chain is developed, operated and managed. This provided the effect of preparing the infrastructure of decision-making support system based on analysis, and mounting the preemptive business response system. Through the analysis model, the information appropriate to the customer's need was provided, customer customized products were developed, and loss ratio was minimized based on the risk forecasting model to be directly involved in the overall insurance industry. Also, constructing the data analysis environment by applying the open source technology has the cost reduction effect, and the execution performance is highly improved.

The product recommendation analysis result can be utilized on searching the customers subjected for visit in advance, and strengthening the reminder of needs on the existing contracted customer by using the customer cluster. In addition, management is possible for each cross-sell/up-sell/insurance re-planning target of each individual customer, and the cluster information can be utilized not only on the recommender system, but also on establishing the campaign strategy. Information related to new product purchasing customer recommendation, insufficient coverage information utilization per customer, and customer loyalty level can be extracted to be used in establishing the marketing strategy.

(2) Maturity. To what degree has the solution's 'vision' been implemented? Has the solution been operating long enough to corroborate business impact and growth?

Currently, the constructed analysis environment and the developed analysis model are operated stably. For the next step, the developed analysis model is to be improved continuously, and the environment that can perform the analysis through the new technology is to be realized.

(3) Relevance. Does the BI/DW/DM or analytics solution exemplify best practices that other companies can adopt?

Presentation was given on the construction of the analysis environment in the insurance company and on the analysis model to other domestic insurance companies. There are continuous inquiries, and benchmarking is enabled not only on the analysis environment infrastructure and details on the analysis model, but also on the data governance, composition of the organization, and budget, etc.

(4) Innovation. Does the BI/DW/DM or analytics solution use an innovative design or approach?

The innovation points for each field realized in this project are shown as follows.

1. Construction of Big Data customized analysis environment:

■ Various open sources converged to construct optimized Big Data analysis environment

2. Hybrid recommendation analysis combined with cluster analysis and recommendation analysis:

■ Realization of customized product recommendation for each similar customer through the cluster analysis

■ Use the recommendation unit as the coverage unit to solve the sparse matrix issue that occurs on insurance product recommendation

■ Realization of the recommender system that is possible for up selling and cross selling

D. SUMMARY / ABSTRACT

Based on your essay, please summarize your overall project in NO MORE than 300 words.

In the standard of the insurance industry value chain, the project was performed on internalizing the Big Data analysis on all fields of insurance work, and to construct the analysis environment. The internal data of the insurance and financial fields are large in size and have various types to have high potential value and applicability. However, the insurance industry lacks the cases of introducing Big Data analysis and application effect sharing, so to solve these issues, the Big Data analysis reference model was developed to be shared.

For the first time in the financial industry, the approach was made in the scale-out method to construct the open source software based analysis system environment. Total of 7 nodes were composed, and the multi-node cluster was constructed with the cost of purchasing a highly-advanced server, but with the equal or higher performance. Based on HDP (Hortonworks Data Platform), centralized architecture was realized through YARN, and the overall monitoring and operation was designed to be managed with Ambari and HUE through the web interface. Also, Spark was used to enable high-speed processing of the files stored in HDFS, and R and Sparkling Water were mounted additionally to supplement the insufficient statistical algorithm and mechanical learning. The basic RDB data for analysis were all mounted on the HDFS through the arrangement processing, and customers' consultation memo and voice recording were duplicated by passing through the STT&TA process in the separate server. The analytic mart was constructed based on the Spark processing speed that is approximately 10~100 times faster than the existing RDB.

For the customer product recommendation, the hybrid recommender engine system applied with cluster analysis (Self Organizing Map based two-step clustering) and collaborative filtering was realized. Before the recommendation analysis, the customers were clustered based on their economic power, and the customer type was identified in various perspectives through this. After the customer cluster, the collaborative filtering algorithm was applied for each cluster to perform the recommendation analysis. Here, there was an issue of sparse matrix due to the characteristics of the insurance industry of the customer not purchasing various insurance products. To solve this issue, the product recommendation unit was increased from the insurance product to the coverage unit to be used in the recommendation analysis model. As a result, the issue of sparse matrix was not only solved, but better recommendation analysis result was obtained. Also when the customer possesses a product, but the product purchase amount is less than the average amount, the up-selling recommendation method of additionally recommending as much as the average amount was added to the conventional recommendation approach (cross-selling). Thus the recommender system of enabling both up-selling and cross-selling was realized.

Currently, the Big Data analysis reference model and the analysis environment developed in the standard of insurance industry value chain within ING are being operated stably. This provided the foundation of analysis based decision-making support system, and brought the effect of enabling the preemptive business response system. Through the analysis model, the information appropriate to the customers' needs was provided, customer customized products were developed, and loss ratio was minimized based on the risk forecasting model, etc. to be directly involved in the overall insurance industry. In addition, there was cost reduction effect by constructing and operating the data analysis environment applied with the open source technology, and the execution performance was highly improved.