Next Generation Business Intelligence and Analytics

Quoc Duy Vo
Department of Computer Science, SUNY Korea, Incheon, South Korea
Department of Computer Science, Stony Brook University, New York, USA
rayvo@sunykorea.ac.kr

Jaya Thomas
Department of Computer Science, SUNY Korea, Incheon, South Korea
Department of Computer Science, Stony Brook University, New York, USA
jaya.thomas@sunykorea.ac.kr

Shinyoung Cho
Department of Computer Science, SUNY Korea, Incheon, South Korea
Department of Computer Science, Stony Brook University, New York, USA
sycho@sunykorea.ac.kr

Pradipta De
Department of Computer Sciences, Georgia Southern University, Georgia, USA
pde@georgiashouthern.edu

Bong Jun Choi
Department of Computer Science, SUNY Korea, Incheon, South Korea
Department of Computer Science, Stony Brook University, New York, USA
bjchoi@sunykorea.ac.kr

ABSTRACT
Business Intelligence and Analytics (BI&A) is the process of extracting and predicting business-critical insights from raw data. Traditional BI focused on data collection, extraction, and organization to enable efficient query processing to derive insights from historical data. With sources of data sources growing steadily, traditional BI&A are evolving to provide intelligence at different scales and perspectives: operational BI, situational BI, self-service BI. In this survey, we review the evolution of business intelligence systems from traditional settings. We focus on the changes in the back-end architecture that deals with the collection and organization of the data, as well as, the front-end applications, where analytics services and visualization are the core components. The survey provides a holistic view of Business Intelligence and Analytics for anyone interested to get a complete picture of the different pieces in the emerging next generation BI&A solutions.

CCS Concepts
•Applied computing → Enterprise computing → Business process management → Business intelligence

Keywords
Business Intelligence (BI); Traditional BI; Operational BI; Situational BI; Self-service BI; Integrative Data Analysis; Heterogeneous Data.

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICBIM'18, September 20–22, 2018, Barcelona, Spain.
© 2018 Association for Computing Machinery
ACM ISBN 978-1-4503-6545-1/18/09 …$15.00.
DOI: http://doi.org/10.1145/3278252.3278292

The Business Intelligence can be expressed as the automated process to collect raw data from heterogeneous sources and organize them in a systematic manner such that models and insights can be derived from the data to improve business processes. The best practice in enterprise BI architectures is split into back-end architecture, associated with the data collection and data organization, while the front-end is about the analytics and displaying the data to the user. The data source is often generated as different transactions are processed, and stored in the Online Transaction Processing server, also called Operational Data Sources. From the OLTP servers, data is extracted, and transformed, and stored in a data warehouse, which is a structured data repository. Different query optimization techniques can be applied to the data warehouse. The analytics query can run on the data warehouse. However, for complex queries, and often for faster response to multiple queries, several subsets of the data warehouse are created, called the data marts.

The data sources for business intelligence is evolving as messages over company intranets to even personal profiles of employees and customers from the web can augment the traditional data sources. The mobile devices and other sensor data also add to the data sources. Many of these data sources are not structured, like texts from messages posted on online social networks, or data from different sensors. This makes it challenging to design the data warehouse as a relational database while maintaining query efficiency. With more data, there is an opportunity for the analytics engines to discover more insight, but there are challenges to designing the necessary tools. The increase in data opens opportunities in expanding the scope of BI beyond just a mechanism to analyze trends from historical data. Situational BI can combine real-time data from sensors and other personal information in real-time to infer insights that are not traditionally available [1], [2]. Operational BI is concerned with providing real-time insights to the business operations, like the call center operatives who can benefit by getting instant feedback on their work [1], [3], [4]. Even the analytics is evolving, with the notion of self-service BI, where the user may compose the analytics rules based on meta-information about the data exposed to her [4]. These new approaches to BI, however, must be carefully

RIGHTSLINK

orchestrated such that the enterprise governance and compliance models are not violated.

In this survey, we capture the shifting trends in BI architecture. For the back-end, we show how different technology evolutions are transforming the architecture. For the front-end, where analytics engines play the pivotal role, we focus on different trends in Machine Learning that are enabling the evolution of BI from the traditional historical analysis tool. We also discuss how challenges in enforcing the enterprise governance models are being addressed in this landscape. The rest of this survey is organized as follows. Section II summarizes traditional BI and modern features of the next generation BI. Section III presents data architecture of recent BI systems, followed by introducing the next generation front-end architecture in Section IV. Finally, we conclude our work in Section V.

2. PRELIMINARIES

2.1 Traditional BI

Traditional BI systems use reporting mechanisms to access transaction data stored in the data warehouse to make business decisions. Since all information stored in the database layer is historical, analyzing such data can help us to detect patterns and predict business trends.

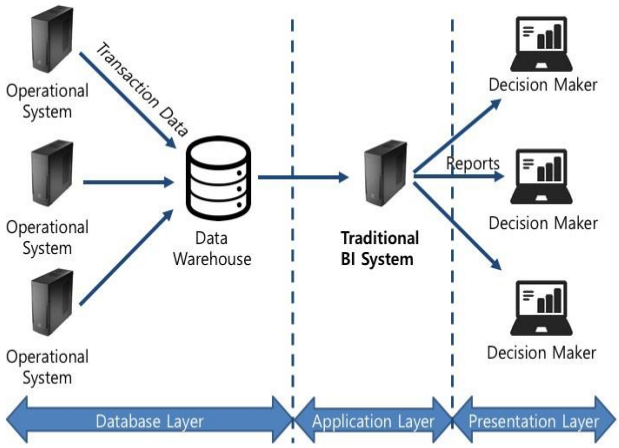

Figure 1. A Traditional BI

As shown in Figure 1, a traditional BI system consists of three separate layers: the presentation layer, the application layer, and the database layer. With a three-tier architecture, it is a big challenge to fulfill service level objectives like maximal response time and minimal throughput rates due to the difficulty in predicting execution times as the low-level layers cannot be correlated well to the high-level layers. Although the typical BI system can give us a forward view of the business, it is well-known that traditional BI systems are slow, rigid, time-consuming, and they place a burden on IT. To this end, many types of research have been investigated by adding features toward these problems and outline many directions for future BI. The following presents modern features of the next generation BI in the literature.

2.2 Modern Features of Next Generation BI

2.2.1 Operational (Real-time) BI:

The competitive pressure of today's businesses has increased the need for near real-time BI (also called operational BI). The goal of operational BI is to reduce the latency between when the transaction data is acquired and when that data is available for analysis so that they can take appropriate actions when an event occurs. Toward this goal, businesses wish to detect the patterns or temporal trends over the streaming operational data from events.

2.2.2 Situational (Situational Awareness) BI:

Situational BI is important for companies to become aware of events (e.g., positive or negative comments about their new products, natural disasters) occurring in the world that may affect their business [5]. However, such external information, which mostly comes from the corporate intranet, external vendor, or internet, are unstructured and need to be integrated with structured information from the local data warehouse to support decision marking in real-time.

2.2.3 Self-service BI (SSBI):

SSBI is an approach to data analytics that enables end users to create analytical queries and reports without the IT department's involvement. To free up IT staffers to focus on other tasks, the user interface in SSBI applications must be user-friendly, intuitive and easy to use, so that people who may be not tech-savvy can access and work with the corporate information. The user also should be allowed to access or extend not just IT-curated data sources, but also non-traditional ones.

3. DATA ARCHITECTURE

3.1 Background and Challenges

The traditional architecture of business applications consists of three separate layers, which are presentation, application, and database. Due to the correlation between low-level data management operations and high-level processes, the execution time is hard to be predicted. The workload management solutions are usually built on top of general-purpose database management systems, which require time delays when executing requests in parallel. This creates challenges for modern business applications to be able to work as operational or real-time business intelligence.

Today's enterprises use an extraction, transformation, and load (ETL) model to extract data, perform transformations, and load the transformed data into the data warehouse. This model relies on two types of processes which are vital to business operations: online transaction processing (OLTP) and online analytical processing (OLAP). OLTP is used to manage business processes, such as order processing, while OLAP is used to support strategic decision making, such as sales analytics.

Although the workloads from OLTP and OLAP are traditionally executed on the same database system, OLAP workloads perform mostly read-only operations on large amounts of business data updated by OLTP workloads. Therefore, when both workloads were performed on the same data in a single database, the transaction processing performance might be unpredictable due to resource contention. High synchronization overhead is required to handle the resource contention, which results in low overall resource utilization.

RIGHTSLINK

Table 1. A Categorization of recent BI systems in terms of modern features

System	HyPer [3]	MobiDB [10]	SIE-OBI [1]	SAP BusinessObjects [4]
Approaches	- Hardware-assisted replication mechanisms - Copy-on-write mechanism	- Queuing Approach	- Data extraction algorithm - Information Correlation	
Achievements	- Fast OLAP query response time. - High throughput rates for both OLTP and OLAP	- Low latency - High throughput rates for both OLTP and OLAP - Optimum space overhead	- Reduce latency - Reduce effort to build data	- Low IT costs and workloads - Real-time response - Flexible and scalable information structure
Operational BI	O	O	O	O
Situational BI	X	X	O	O
Self-Service BI	X	X	X	O
Year	2011	2011	2012	2009

In addition, commercial DBMS use special techniques, such as shadow copy [6], to handle mixed workloads with low-performance overhead. Different workloads are separated by performing them on different logical copies of the data. This may cause additional space overhead, which increases the infrastructure requirements and costs. Therefore, managing these mixed workloads (OLTP and OLAP) in the data management systems is a big challenge for current disk-based DBMSs [7].

3.2 Recent BI Systems

3.2.1 Traditional BI Systems:

In this section, we present existing traditional BI techniques which can perform OLTP transactions and OLAP queries on the same database without interfering with each other. Due to the contradiction of the dramatic explosion of the dynamic data volume, the integration of these mixed workloads on the same system requires extreme performance improvements.

(i) In-Memory Database (IMDB): In most today's BI systems, the mixed workload comprised of OLTP and OLAP on a single system can be handled by using in-memory (or main-memory) database (IMDB). This technique requires the system to store all data in main memory, which is considered as faster than disk-optimized databases since the internal optimization algorithms are simpler and use fewer CPU instructions. When querying the data, this approach provides faster and more predictable performance than disk by eliminating the seek time. However, IMDB systems can be said to lack of durability due to the loss of stored information when the device loses power or is reset. Many IMDB systems have proposed different mechanisms to support durabilities such as snapshot files, transaction logging, non-volatile DIMM, non-volatile random-access memory, and high availability [8].

(ii) Hybrids with on-disk database: Although main-memory is becoming large enough to handle most OLTP database, it may not always be the best option. Using the access patterns of the OLTP workloads, where some records are "hot" (frequently accessed), others are "cold" (infrequently or never accessed), recent systems tend to store the coldest records on a fast-secondary storage device, and hot records should reside in memory to guarantee good performance. For instance, authors [9] enable a main-memory database to migrate data to a larger and cheaper secondary storage. To reduce OS paging I/O and improve the main memory hit rates, the relational data structures are reorganized using the access statistics of the OLTP workloads. Recently, Siberia has been introduced as a framework for managing cold data in the Microsoft Hekaton IMDB engine [6]. Like [9], it does not require a database be stored entirely in main memory.

3.2.2 BI Systems with Modern Features:

In this section, we present different BI systems with modern features: operational BI, situational BI, and self-service BI. Table 1 shows a categorization of recent BI systems in terms of modern features.

While the H-Store system is limited to only OLTP transaction processing, a recent system, called HyPer, can handle mixed workloads from both OLTP and OLAP at extremely high throughput rates using a low-overhead mechanism for creating differential snapshots [3]. This system employs the lock-less approach which allows all OLTP transactions are executed sequentially or on private partitions.

Like H-Store, MobiDB is a special-purpose main- memory DBMS which can guarantee serializability and mixed workloads using the queuing approach [10]. Instead of processing the incoming transaction and periodic business queries right away, the MobiDB adds them to a queue and processes them later. These requests are first analyzed to estimate how long they would take to be performed adaptively execute the queued requests.

To alert business managers of situations that can potentially affect their business, M. Castellanos et al. propose a novel platform, called SIE-OBI, that integrates the required functionalities to exploit relevant fast stream information from the web [1]. The authors proposed novel algorithms which extract and correlate the information obtained from the web with historical data stored in the data warehouse to detect the situation patterns. Only relevant information is extracted from two or more disparate sources of unstructured data, typically one internal slow text stream and one external fast text stream. This platform is created to reduce the time and effort of building data flows that integrate structured and unstructured, slow and fast streams, and analyze them in near real-time.

RIGHTSLINK

BusinessObjects provides easy, self-service access to decision-ready information with SAP BI platform [4]. With a flexible architecture, BusinessObjects allow business managers to put decision-ready information within business users' reach, but with an increased responsiveness, low IT costs and workloads.

4. NEXT GENERATION FRONT-END ARCHITECTURE

4.1 BI: Analysis

The growing volume of data in many businesses makes cost-effective manual data analysis virtually impossible. The use of data mining techniques in business not only handles the volume and variety of data but also helps to take a proactive knowledge-driven decision and enhance business intelligence in general.

Data mining is a broad term that includes several the process as data modeling techniques, statistical analysis, and machine learning in search for consistent pattern or relationship and determining some predictive information in the data analyzed from large amounts of data [11]. Machine learning leverage to data mining algorithms to improve the predictive analysis. Machine learning techniques have become popular in business intelligence as they can handle growing volumes and varieties of available data and makes the computational processing that is cheaper and more powerful.

The two broad categories of machine learning algorithms are supervised and unsupervised [12]. Supervised learning involves the training data that contain sets of training examples, whereas unsupervised learning draws an inference from training data without labeled responses. These learning techniques are employed to carry out the predictive analytics task, build an analytic model at a lower level, search for predictable behaviors, business rules and look for answers on predicting performance and prescribing specific actions or recommendations. The techniques include regression modeling, clustering, neural networks, genetic algorithm, text mining, decision tree, and more.

4.2 Machine Learning Applications in Business Intelligence

In this section, we will discuss the few applications of machine learning techniques in time series forecasting and sale forecasting. The description of time series data includes high dimensionality, volume and continuously evolving. The analysis of time series data is a powerful analytics tool as it helps to address questions as rate of change in user behavior with time, co-variance between the product, marketing promotion strategies, current trend in product sale, profit monitoring, determine anomalies etc. in nearly all enterprises as sales, manufacturing, mobile companies, hospitals, etc.

4.2.1 Application Based on Support Vector Machine:

The financial prediction on the stock price is of great interest for the investors as well as the analytics. However, the prediction about the current time to buy or sell any stock is not an easy task as the price is influenced by several many factors. Support Vector Machine (SVM) is the most common technique used to carry out such predictions. Support Vector Machine (SVM) is a popular supervised machine learning algorithm, which can be used for classification or regression problems. SVM algorithms can capture complex relationships between data samples without carrying out difficult transformations. In [13], an application of SVM is described in financial time series forecasting using a single data source. In this work, the authors present a modified SVM algorithm (ASVM) that consider adaptive parameters to handle the structural changes in the financial data. The experiment was carried out on five real futures contracts collated from the Chicago Mercantile Market.

4.2.2 Application Based on Genetic Algorithm:

The prediction of future sale by an enterprise is termed as sales forecasting, which is a part of its critical management strategy. The machine learning techniques as Genetic algorithms (GA) are suitable candidates for this task since GAs are most useful in multiclass, high-dimensionality problems where heuristic knowledge is sparse or incomplete. Neural networks are considered efficient computing models for pattern classification, function approximation, and regression problems. The sales forecasting problem for printed circuit board (PCB) sales is addressed in [14], where the model is built by integrating GAs and Neural Network. The study was carried out for PCB electronic industries in Taiwan, where the feature of data included monthly sales amount, total production square measures, etc. The PCB sale is studied in [15] using integrated genetic fuzzy systems (GFS) and data clustering. The approach experimented on the PCB data source with parameters as pre-processed historical data, Consumer price index, Liquid crystal element demand and Total production value of PCB.

4.2.3 Limitations and Challenges:

The techniques discussed above considers single data source for forecasting, and the data source contains features such as Last, Change, Open, High, Low, Previous, Volume etc. However, today the stock markets databases are flurried from a wide range of complex data from diverse sources as market data, reference data, exchange, security description, fundamental data as enterprise financial, analyst report, filing etc., and even data from social media that may include blogs, web feeds etc. Moreover, the financial data is highly dynamic and volatile which raise a need for some integrative approach that combines data from the other sources and contributes towards the accuracy of the forecast. Similarly, in the case of sales forecasting, the prediction is very complicated owing to influence by internal and external environments.

4.3 Integrative Data Analysis

In today's enterprise working world, the data is heterogeneously generated at a much faster rate, characterized by huge data sets and varied data types, where the nature of data being structured, semi-structured and unstructured as flat files, cubes, videos, images, audio, weblogs, text, and e-mail etc. This poses a challenge to the existing traditional storage and analytic solutions which do not cope well with the heterogeneity of data.

The data from multiple sources can provide an insight to increase productivity, improve policy making, support performance measurement, and can help in strategic planning. These insights can help to achieve benefits including improved customer satisfaction, quality improvement, increased accessibility and analysis of information, timeliness and better information utilization. The need to handle the heterogeneous data and automated analytics algorithm can be resolved by doing the integrative predictive analysis. It supports the study of the integrated data using machine learning algorithm that continuously evolve the accuracy of predictive models and enable it to adjust.

RIGHTSLINK

4.3.1 Involvement of Machine Learning Techniques:

An integrative analysis task involves machine learning techniques for the integration of the available training data from different data sources to better analysis and generates a proactive response. A single data source may not contain all the required information about the data object. Thus, when we combine multiple sources of information for a data object it helps to add on some different or missing information that may lead to a better prediction accuracy. Integrating data from multiple sources and making decisions from these combined sources is becoming common to enhance the prediction performance for different applications in healthcare, image classification, the stock market, etc.

The two most common integrative analyses are Multiple Kernel Learning (MKL) [16] and the Bayesian Network (BN). Multiple kernel learning refers to a set of machine learning methods that use a predefined set of kernels, where kernel selection depends on the notions of similarity that may exist in the data source. MKL learn an optimal linear or non-linear combination of kernels as part of the algorithm that results in data integration. Multiple kernel learning algorithms have been developed for supervised, semi-supervised, as well as unsupervised learning. The Bayesian network consists of a directed acyclic graph (DAG) and a set of local distributions. Each node denotes a random variable, which can be an attribute, feature or hypothesis. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. Often the existing Bayesian network learning algorithms are modified to integrate multiple heterogeneous and high-dimensional omics data, where the Bayesian data integration varies a set of parameters relevant to the computational aspects of the task.

4.3.2 Application Based on Multiple Kernel Learning:

We discuss the application of MKL for the predictive analysis in case of the stock market. In stock price prediction it is observed that relying on the study of the time series historical data is not sufficient, rather by considering multiple sources of information such as news and trading volume one can significantly improve the stock market volatility prediction [17]. The experiments carried out on HKEx 2001 stock market datasets shows that the use of multiple kernel learning approach on the multiple data sources results in higher accuracy and lower degree of false prediction as compared to single source data.

The work in [18], shows that analyzing communication dynamics on the internet and using stock price movements may provide some new insights into relations between stock prices and communication patterns. They use MKL to combine information from time series data, stock price and stock volume with other data source: news and comments that included the frequency of News, the frequency of the comments, average Length, Standard Deviation of the length of comments, number of Early/Late response etc. The experiment was tested for the stock of Amazon, Microsoft, and Google, and it was found that MKL prediction model outperforms other baseline methods such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error(RMSE). These results motivate the use of integrative data analysis for other domains.

5. CONCLUSION

In this paper, we compared the next generation business intelligence (BI) with the traditional BI and showed the different features of the shifting trends in BI. To do so, we discussed the back-end and front-end architecture of BI. At the back-end level, we discussed the changing trend in system architecture. At the system level, there are three features in the next generation BI: operational, situational, and self-service BI. Operational BI focus on providing a near-real-time response to the data analysis request. To support the agile response, the architecture of BI systems is evolved to reduce the latency of response. Situational BI allows an enterprise to use real-time data from external sources to give the insight to make an accurate decision. The data from multiple sources, however, can occur low data quality and inaccurate data. Self-service BI, which allows users to participate in analysis process rope, makes the governance of data difficult. We discussed what kind of issues can be concerned in the situational BI. At the front-end level of the business intelligence architecture, we show the use of machine learning techniques for predictive analysis. In addition, we emphasize on the need for the use of the machine learning techniques for integrative data analysis that can handle data heterogeneity and may help to achieve a better predictive accuracy.

6. ACKNOWLEDGMENTS

This research was funded by the MSIT, Korea, under the "ICT Consilience Creative Program" (IITP-2017-R0346-16-1007) supervised by the IITP, and by the KEIT, Korea, under the "Global Advanced Technology Center" (10053204).

7. REFERENCES

[1] Castellanos, M., Gupta, C., Wang, S., Dayal, U., and Durazo, M. (2012). A platform for situational awareness in operational BI. Decision Support Systems, 52(4):869-883.

[2] Lo¨ser, A., Hueske, F., and Markl, V. (2009). Situational business intelligence. *Business Intelligence for the Real-Time Enterprise*. Springer. pp. 1-11.

[3] Kemper, A., and Neumann, T. (2011). Hyper: A hybrid OLTP&OLAP main memory database system based on virtual memory snapshots. *Proceedings of ICDE*. pp. 195-206. Hannover, Germany.

[4] SAP BusinessObjects. (2018). Retrieved June 21, 2018 from https://www.sapbi.com/.

[5] Lo¨ser, A., Hueske, F., and Markl, V. (2008). Situational business intelligence. *Business Intelligence for the Real-Time Enterprise*. Springer. pp. 1-11.

[6] Elmasri, R. and Navathe, S. B. (2014). *Fundamentals of database systems*. Pearson.

[7] Kuno, H., Dayal, U., Wiener, J. L., Wilkinson, K., Ganapathi, A., and Krompass, S. (2010). Managing dynamic mixed workloads for operational business intelligence. *Databases in Networked Information Systems*, Lecture Notes in Computer Science, 5999. Springer, Berlin, Heidelberg.

[8] Kallman, R., Kimura, H., Natkins, J., Pavlo, A., Rasin, A., Zdonik, S., Jones, E. P. C., Madden, S., Stonebraker, M., Zhang, Y., Hugg, J., and Abadi, D. J. (2008). H-store: A high-performance, distributed main memory transaction processing system. *Proceedings of VLDB Endow.*, 1(2), 1496-1499.

[9] Stoica, R., and Ailamaki, A. (2013). Enabling efficient OS paging for main-memory OLTP databases. *In Proceedings of DaMoN*. NY, USA.

[10] Seibold, M., Kemper, A., and Jacobs, D. (2011). Strict SLAs for operational business intelligence. *In Proceedings of the*

RIGHTSLINK

IEEE International Conference on Cloud Computing. Washington, DC, USA.

[11] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press.

[12] Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine Learning: An Artificial Intelligence Approach*. Springer Science & Business Media.

[13] Cao, L. J. and Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks*, 14(6), pp. 1506-241.

[14] Chang, P. C., Wang, Y. W., and Tsai C. Y. (2005). Evolving neural network for printed circuit board sales forecasting. *Expert Systems with Applications*, 29(1), pp. 83-92.

[15] Hadavandi, E., Shavandi, H., and Ghanbari, A. (2011). An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: Case study of printed circuit board. *Expert Systems with Applications*. 38, 9392-9399.

[16] Gonen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12, pp. 2211-2268.

[17] Wang, F., Liu, L., and Dou, C. (2012). Stock market volatility prediction: A service-oriented multi-kernel learning approach. *In Proceedings of the IEEE International Conference on Services Computing*. Honolulu, HI, USA.

[18] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., and Sakurai, A. (2011). Multiple kernel learning on time series data and social networks for stock price prediction. *In Proceedings of ICMLA*. Honolulu, HI, USA.

RIGHTSLINK