Strategic Management for Real-Time Business Intelligence

Konstantinos Zoumpatianos, Themis Palpanas, and John Mylopoulos

Information Engineering and Computer Science Department (DISI),
University of Trento, Italy
`{zoumpatianos,themis,jm}@disi.unitn.eu`

Abstract. Even though much research has been devoted on real-time data warehousing, most of it ignores business concerns that underlie all uses of such data. The complete Business Intelligence (BI) problem begins with modeling and analysis of business objectives and specifications, followed by a systematic derivation of real-time BI queries on warehouse data. In this position paper, we motivate the need for the development of a complete Real Time BI stack able to continuously evaluate and reason about strategic objectives. We argue that an integrated system, able to receive formal specifications of the organization's strategic objectives and to transform them into a set of queries that are continuously evaluated against the warehouse, offers significant benefits. In this context, we propose the development of a set of real-time query answering mechanisms able to identify warehouse segments with temporal patterns of special interest, as well as novel techniques for mining warehouse regions that represent expected, or unexpected threats and opportunities. With such a vision in mind, we propose an architecture for such a framework, and discuss relevant challenges and research directions.

1 Introduction

Strategic Management (SM) is concerned with the continuous evaluation and control of a business and the environment within which it operates; it assesses internal and external factors that can influence organizational goals, and makes changes to goals and/or strategies to ensure success. SM has been practiced since the '50s, thanks to seminal contributions by Alfred Chandler, Peter Drucker and others who emphasized the importance of well-defined objectives and strategies that together determine and guide organizational activities [19]. Specific analysis techniques have been developed to support SM processes, including the strengths-weaknesses-opportunities-threats (SWOT) analysis technique widely used in practice [14].

SWOT analysis focuses on internal strengths and weaknesses, as well as external opportunities and threats that may facilitate/hinder the fulfillment of organizational objectives. Once such SWOT situations are identified, they need to be continuously monitored in real-time to assess the degree to which they occur and keep track of their evolution over time (monitoring). In addition, operational data need to be continuously scanned in search of emerging SWOT situations or unexpected data (outliers).

Being able to identify such threats and opportunities requires systems able to process data as they evolve, as well as, algorithms able to discover trends and deviations hidden among multiple layers of aggregated information. The problem of aggregating data over multiple dimensions has been studied within the domain of data warehousing, where OLAP technology [6] provides to the analysts the ability to explore and query these aggregates, in search of interesting market segments. Traditionally data warehouses considered time as a common dimension inside a data cube. Although, as Chaudhuri and Dayal noticed, a dimension of special significance [3]. Moreover, the notion of time has also been studied from the perspective of temporal databases, where it became known as temporal data warehousing [2, 16]. In such systems two distinct temporal information types are identified: the validity time and the transaction time. The validity time refers to the interval within which a record "holds", and the transaction time refers to temporal information (i.e., a point in time). For a detailed review the reader can look at [8]. Even though temporal data warehouses solved the problem of dimension and record evolution, they did not provide any means for monitoring trends and deviations in streams of data, and they also failed to capture the evolution dynamics of specific market-segments.

It becomes clear at this point that treating time as a common dimension can be rather limiting. This is especially clear in applications where predicting future incidents or the demonstration of complex behaviors over time, as well as their causes, is important. Moreover when this kind of temporal analytics have to be provided in a real-time and concise manner all over the data-set, data-warehouses have to treat the dimension of time, wherever it is found (validity time or transaction time), as a first class citizen and not solely rely on external tools for temporal data mining.

Various systems have been proposed, in order to solve this problem, that adopt time-centric data representation techniques. These include the CHAOS system, which tackles the problem taking into account the business rules and their continuous evaluation against the data warehouse [9]. Others works have concentrated on creating OLAP architectures able to handle streaming data, capture their evolution dynamics over the time [5, 10] and identify interesting segments of the data cube [15].

While all of them tackled the problem from various perspectives, to the best of our knowledge, there have been no comprehensive implementations that combine trend identification, as well as deviation detection query answering mechanisms, with the high-level information of the strategic objectives of a company. Additionally, there have been no tools that allow this kind of analytics to be done in a systematic way, allowing the automatic (and ad-hoc) generation and evaluation of such analysis queries.

The main thesis of this position paper is that taking into account SM models leads to truly effective Business Intelligence (BI), enabling the pain-free analysis of real-time SWOT situations, a task that is critical to modern organizations. The intricacies and special characteristics of this synergy have not been carefully studied before.

In this study, we describe the development of a system that extends the traditional functionality of data warehouses in two complementary ways. First, we propose the tight integration of SM models in BI solutions. Such an integration ensures that all the important, strategic questions are evaluated, thus better aligning the needs of the business analyst with the data analytics performed on the warehouse. Second, we describe two additional classes of queries, namely, trend and deviation detection queries, which are necessary for monitoring the progress of the Strategic Objectives of an organization and perform SWOT analysis. Note that these types of queries have to be natively supported by the data warehouse (current systems do not offer this functionality), in order to allow for their efficient execution in the context of real-time BI. In addition, we discuss how these queries can be semi-automatically generated from the Strategic Objectives, reducing the burden for the analysts, and streamlining the process.

2 BI Models for Strategic Management

Consider a fictitious multinational electronic sales enterprise (hereafter ElectronicE) selling a range of products in Europe, with a strategic objective to increase its market share by 5% over the next week, as a result of a large scale search engine and social networking ad campaign (objective increase-market-share, or IncreaseMS). Because of the nature of electronic sales and electronic campaigns, the enterprise needs to be able to monitor the impact of their advertising in real-time. During strategic planning, this objective can be refined into sub-objectives through AND/OR decompositions of IncreaseMS, such as the development of new product lines (NPLines), the introduction of targeted advertizing (TAdvert), or the creation of new eCommerce channels (NChannels) for reaching ElectronicE's customer base. Out of this planning exercise, one or more strategies are adopted for achieving IncreaseMS. Each such strategy is a plan consisting of concrete actions that can be carried out by ElectronicE or its partners. SWOT analysis adds to this planning exercise a risk component. What are the factors that could affect positively/negatively the achievement of IncreaseMS? An economic downturn (EconomicD) in a region would drive sales down. Lack of knowhow (LackKH) within ElectronicE would slow down progress on the development of new eCommerce channels, while a wrong choice of target market segment (WrongMS) can affect the impact of targeted advertizing. On the other hand, a successful design for a new product (GoodD) would definitely help with the fulfillment of IncreaseMS. The elements of this scenario can be captured in a modeling language such as the Business Intelligence Model (BIM) [12]. One can think of BIM as a language in the same family as the Entity-Relationship Model, but with the concepts of 'entity' and 'relationship' replaced by concepts such as 'goal', 'situation', 'indicator', 'process' and more. Figure 1 shows a simple BIM model for our scenario. The figure captures the concepts mentioned above, and relates them through relationships such as AND-refinement (the goal IncreaseMS is refined into three sub-goals), influence (with a + or - label to indicate whether a situation influences positively/negatively a

goal) and association (each indicator is associated with at least one goal and/or situation). Indicators (short for Key Performance Indicators, or KPIs) measure the degree of fulfillment of a goal, or the degree of occurrence of a situation, and possibly other things (e.g., the quality of a product or the degree of completion of a process). For example, the increase in sales over a period of time (SalesC) measures how well we are doing with respect to the objective IncreaseMS, while a negative change to Gross Domestic Product (GDP) measures the degree to which we have an economic downturn situation in a geographic region. In a similar vein, lack of knowhow is measured by delays in developing new eCommerce channels (indicator time-to-completion, or TimeTC). Indicator values are often normalized so that they fall in the range [-1.0, 1.0]. Moreover, thresholds are provided to mark whether a goal is fulfilled, partially fulfilled, denied, etc. [1].
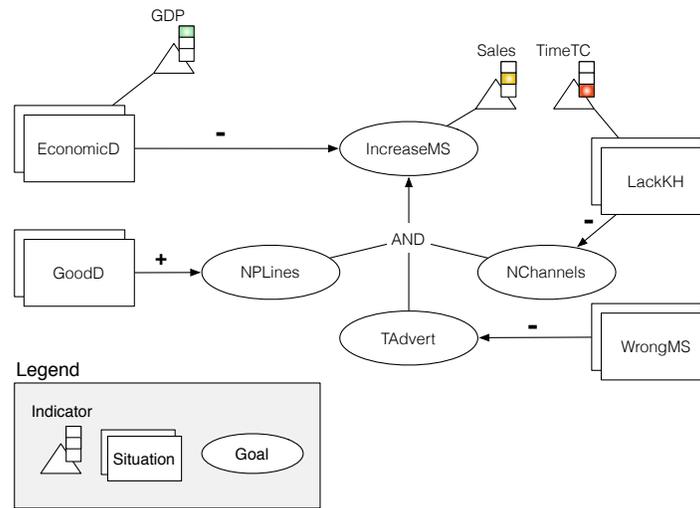

Fig. 1. A simplified strategic model

3 Analysis for SM

There are two facets to the analysis we want to support. First, there are dimensions to the problem at hand (fulfilling IncreasedMS) that identify with the dimensions of the underlying data cube, in our case (i) time (over what periods did/didn't we do well?), (ii) region (Trentino/Italy/Europe) and product (specific product/product line/all products). This means that for any question we want to answer, such as "How are we doing with IncreaseMS?" there will have to be series of queries that answer the query at different levels of granularity.

Second, the analysis to be conducted can have three different modalities:

1. **Monitoring and Impact** – this modality applies when we are trying to keep track of the status of an objective or situation, i.e., we are trying to

answer "how are we doing?" questions: are we on track with IncreaseMS? has there been any change to the status of situation EconomicD? what is the impact of a locally applied strategy?

2. **Outliers** – here we are looking for surprises, both positive and negative. Such a modality is very important for organizations, since it ensures that interesting information that is not part of the strategic model arrives to the analyst. For example, sales have jumped in Lombardia threefold relative to the rest of Italy.

3. **Explanation and Troubleshooting** – in this modality, we are analyzing the reasons for some observed behaviors. For example, if we are doing badly or very well with respect to IncreaseMS (Monitoring and Impact modality), we need to pinpoint where (what regions? ... time periods? ... for what products?) and for what reasons this happens. The reasons would be the root causes for observed performance, such as weaknesses and threats that correlate with observed performance. Additionally, the identification of outliers needs to be followed by diagnosis (why?) and possible followup action, remedial or perfective (replicate in other regions a successful strategy).

It is important to emphasize the importance of the outliers modality of analysis. Peter Drucker and others have pointed out since the 70s that we live in an Age of Discontinuity where extrapolating from the past is hopelessly ineffective [7]. Accordingly, strategic analysis should not fall into the linearity trap and always look out for surprises, because they are now the norm.

4 SM Enabled Business Intelligence

Monitoring a large, high-dimensional data cube for patterns of interest, as well as for unexpected patterns, is not a trivial task. It is hard to analyze such data cubes at the most detailed level, because of their sheer size, and at the same time summarization techniques, which can reduce the amount of data to operate on, may hide various interesting patterns (e.g., anomalies) that could be identified as threats or opportunities. In systems where we are interested in supporting continuous trend queries (e.g., checking for the fulfillment of the Strategic Objectives of a company), it is of great importance to be able to capture the fine-grained temporal dynamics of each market segment within the data cube, so that we are able to perform queries considering the temporal aspects of our data.

Moreover, current data warehouses rely on pre-specified hierarchies for each dimension in the database, for example, the City, Region, Country hierarchy. Nevertheless, within the current global market such hierarchies may not always make sense to consider and, additionally, their full extent may not be known beforehand. This means that new interesting groupings could be discovered that would complement the analysis based on the existing dimension hierarchies.

Building a system able to cope with a high workload of continuous monitoring queries corresponding to SWOT analysis, leads to the following set of challenges.

1. Real-time updating of the data warehouse does not necessarily mean instantaneous updates [26]. Instead, the update process should follow the business requirements. This means that we need to be able to assess how often we need to update which data in regards to our goals and their current status. This is something that we can judge from our Strategic Model and the evolution dynamics of the data currently available in the data warehouse.
2. The system should include mechanisms for generating a set of focused exploratory queries based on the Strategic Model, as well as, algorithms for answering them efficiently. Such queries should allow for the **Modeling and Impact** Strategic Management analysis modality. Their systematic evaluation should be decided based on the Strategic Model and with efficiency in mind (i.e., decide *which* queries to execute *when*). Further on we should be able to support the **Explanation and Troubleshooting** analysis modality by providing mechanisms that are able to explain why the status of our goals and indicators is positive or negative with regards to our expectations. This means that we need to be able to provide insight for the correct exploration of the data cube behind the data that affect the specific parts of the Strategic Model.
3. Monitoring the Strategic Objectives of an organization is vital, but it is tricky since there may exist Threats and Opportunities that do not directly affect the Strategic Model. For this reason, multidimensional outlier detection techniques have to be employed, such that we identify all the interesting parts in the data: these techniques should use as little input information as possible, and (as pointed out earlier) go beyond the dimension hierarchies explicitly defined in the data warehouse. In order to fully support both the **Outliers** and the **Explanation and Troubleshooting** analysis modalities, explanation techniques have to be deployed in order to explain such deviations in the data.

Our solution proposal, depicted in Figure 2, involves a Strategic Model layer that the user defines based on the business schema, a Query Generator layer that creates continuous queries in regards to the model, as well as a Query Engine that is able to answer trend and outlier detection queries on top of a specialized and efficient Data Warehouse structure. The architecture also includes a Dashboard for the visualization of results (which can also be generated from the Strategic Model [20]). We elaborate on these components in the following paragraphs.

The Strategic Model layer allows the user to specify the Strategic Objectives of the organization, the indicators that affect them, as well as their interconnections and decompositions. Additionally, the ability to perform ad-hoc queries to the system is given to the analyst, so that queries can also be formed manually. The Query Generator layer is responsible for automatically generating continuous queries based on the Strategic Model. These queries are evaluated against the data warehouse, in order to provide feedback with respect to the Strategic Goals in real time. Recent research [24] has shown that it is possible to generate data warehouse queries from the Strategic Objectives of an organization for monitoring its Strategic Goals. Such techniques have to be extended and
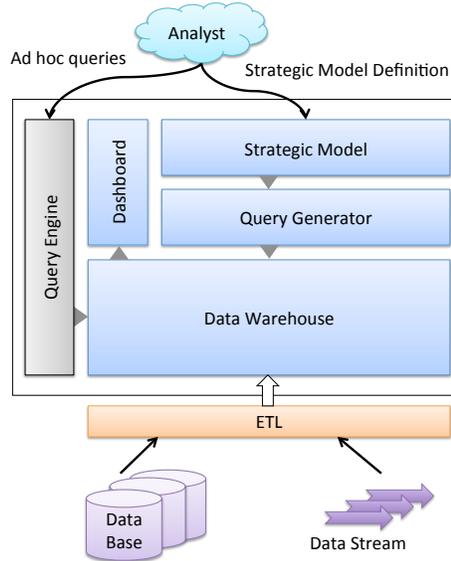
Fig. 2. System architecture

incorporated in this layer, in order to interpret high level formalizations into low level operations, such as trend and outlier detection queries, on top of the data warehouse.

The Data Warehouse layer is responsible for storing the data received as soon as they are available (requirement 1), ideally in a streaming fashion. This means that the warehouse has to be designed in such a way that it can keep expanding as data arrive, with the minimum storage and processing cost overhead. Moreover since we are interested in performing fine-grained temporal queries on top of our data, it has to be able to inherently support trend and outlier detection queries using a temporal representation of the data available in each cell. At this point we argue in favor of the use of time series based representations for each market segment, in order to efficiently capture and reason about temporal dynamics.

Finally, the Query Engine has to be able to answer the queries generated by the Query Generator, as well as the ones that where directly submitted by the user. This component incorporates specialized time-series based algorithms for identifying trends, aggregating parts of the cube, as well as identifying outliers within the data cubes. An example trend query is the following: "Will the current sales trend that we observe up to now, within a time window W, in the market segment S help as achieve the goal of increasing our market share by 5% (IncreaseMS)?". Additionally, an outlier detection query could be of the form "Are there any interesting sub-markets within Europe?". The results of such a query could be unexpected patterns, such as sub-markets in Europe where the

company is doing surprisingly worse than others, or expected patterns such as a sub-market that is responding to the company's strategy.

It is also important that the Query Engine is able to compute and monitor trends, and identify outliers with respect to *ad hoc* market segments, not relying on pre-specified hierarchies and groups, as they need manual setup and may not always be representative of the data. An example could be the Countries, Cities hierarchy, where we cannot always expect that all sales within a country or a city follow the same patterns. As a result, we should be prepared to find outliers in groupings not explicitly specified by the hierarchy, for example it may be the case that interesting trends appear in regions within a country.

5 Previous Work

Data warehouses support decision making by providing On-Line Analytical Processing tools (OLAP) [6] for the interactive analysis of multidimensional data [11]. Such tools allow a person (analyst) to quickly acquire important information drilling in and out of the most interesting aggregates of a database, a task that is fairly complex considering the large number and sizes of dimensions [23]. The common case for updates to a data warehouse is the Extract-Transform-Load (ETL) processing [25], i.e., data are extracted from the sources and loaded to the data warehouse during specified time intervals. As applications are pushing for higher levels of freshness, data warehouses are updated as frequently as possible, giving rise to Active Data Warehousing [13, 22, 21].

While traditional data cube technology is good for static aggregates, it fails to explain trends in the multi-dimensional space [5] (requirement 2). For this reason, linear regression analysis in time series based data cubes has been presented in [10, 5]. The basic idea was to do linear regression analysis in the maximum level of detail an analyst is willing to inspect, in order to identify the trend of the data in each cell and use these low level trends for calculating the ones in the upper layers. Furthermore, the efficient generation of logistic regression data cubes was studied in [27], where the authors introduced an asymptotically lossless compression representation (ALCR) technique, as well as an aggregation scheme on top of it, in order to eliminate the need of accessing the raw data for the calculation of each distinct cell in the data cube.

With regards to the problem of reducing the size of a time series based OLAP data cube, a tilted time-frame scheme aiming at reducing the storage costs, as well as a popular-path based partial materialization of the multi-dimensional data cube for reducing the processing time are presented in [5]. Additionally in [9] an extended Haar wavelet decomposition technique has been proposed for reducing the amount of data, as well as, for providing a reduced hierarchical based synopsis method for the time domain.

There has been very little work on identifying the interesting parts of a time series based data cube. In [5] the idea of exceptional regression lines has been presented, i.e. regression lines with a slope greater than or equal to an exception threshold. Moreover, in [15] the problem of identifying the top-k most anomalous

parts of a regression-lines-based cube has been studied. The main observation, on which their solution was based, was that subjects which are parts of a hierarchy, should behave similarly to the parent market segments. Even though this method will work for the hierarchies existing in the warehouse, we envision an approach able to identify a wider range of anomalies (e.g., anomalies not directly correlated to the existing hierarchies).

We note that combining the use of temporal data representations as first class citizens, techniques for outlying market segments identification, and the Strategic Objectives modeling, gives rise to a complex problem that none of the current works addresses. CHAOS [9] is the system that is the most relevant to our proposal. This system is able to create a multi-dimensional cube for streaming data, apply event processing above it, and visualize the interesting parts of the cube. It is able to answer queries of the form: *"Report the set of biggest changes within a market segment M and a time window W"*. Using the results of such queries it is able to evaluate complex business rules and infer events of special significance. Nevertheless, the business rules that CHAOS considers are defined relative to the data warehouse schema, and not relative to the business schema, creating a disconnect to the Strategic Objectives of the organization. Moreover, CHAOS does not describe any techniques for trend and outlier analytics, which are indispensable for performing SWOT analysis and answering SM queries.

Additionally a set of related works that could be used on top of our proposed system are the ones related to Sentinels [18] and the OODA concept [17]. The OODA concept describes the loop of Observation (are data normal?), Orientation (what is wrong?), Decision (user analysis) and Action (course of action) on top of a set of KPIs. Sentinels are causal relationships between KPIs mined in the data warehouse which are used to trigger early warnings, thus helping the analyst perform OODA cycles efficiently. Such sentinels could be mined and integrated in our Strategic Model as situations that can affect our goals. Related to that is the work on Bellwether analysis [4] which aims at identifying efficient predictor queries for estimating target future values for other ones. Such bellwether queries can also be additionally used to identify the future values of our Indicators.

6 Conclusions

In this work, we motivated the need for a system that is able to continuously monitor a data warehouse based on queries generated from the Strategic Model of an organization. This system should be able to identify trends in regards to these pre-specified objectives, and also to monitor the warehouse for expected or unexpected threats and opportunities in the data as well as their causes. Furthermore, we presented the related work and its gaps, we discussed the challenges of building such a system and presented an architecture of a system that treats the trends and outlier detection techniques, in all the layers of the system, as first class citizens, which is crucial for the kind of analysis we propose.

Acknowledgements

References

[1] D. Barone, L. Jiang, D. Amyot, and J. Mylopoulos. Composite indicators for business intelligence. 2011.

[2] P. Chamoni. Temporal structures in data warehousing. *DEXA*, 1999.

[3] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1), March 1997.

[4] B. Chen, R. Ramakrishnan, J.W. Shavlik, and P. Tamma. Bellwether analysis: Searching for cost-effective query-defined predictors in large databases. *ACM TKDD*, 3(1), March 2009.

[5] Y. Chen, G. Dong, J. Han, B.W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. *VLDB*, 02, 2002.

[6] E.F. Codd, S.B. Codd, and C.T. Salley. *Providing OLAP (on-line Analytical Processing) to User-analysts: An IT Mandate*, volume 32. Codd & Date, Inc., 1993.

[7] P.F. Drucker. *The age of discontinuity: Guidelines to our changing society.* Harper and Row, New York, New York, USA, 1968.

[8] M. Golfarelli. A survey on temporal data warehousing. *International Journal of Data Warehousing*, 5, 2009.

[9] C. Gupta, S. Wang, I. Ari, and M Hao. Chaos: A data stream analysis architecture for enterprise applications. *CEC*, 2009.

[10] J. Han, Y. Chen, G. Dong, J. Pei, B.W. Wah, J. Wang, and Y.D. Cai. Stream Cube: An Architecture for Multi-Dimensional Analysis of Data Streams. *Distributed and Parallel Databases*, 18(2), 2005.

[11] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis. *Fundamentals of Data Warehouses.* Springer-Verlag, 2003.

[12] L. Jiang, D. Barone, and D. Amyot. Strategic models for business intelligence. *Conceptual Modeling ER 2011*, 2011.

[13] A. Karakasidis, P. Vassiliadis, and E. Pitoura. ETL queues for active data warehousing. *IQIS '05*, 2005.

[14] R. Lamb. *Competitive strategic management.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[15] X. Li and J. Han. Mining approximate top-k subspace anomalies in multi-dimensional time-series data. *VLDB*, 2007.

[16] A.O. Mendelzon and A.A. Vaisman. Temporal Queries in OLAP. *Proceedings of the 26th International Conference on Very Large Databases*, 2000.

[17] M. Middelfart. Improving business intelligence speed and quality through the ooda concept. DOLAP '07, New York, NY, USA, 2007. ACM.

[18] M. Middelfart and T.B. Pedersen. Implementing sentinels in the targit bi suite. In *ICDE*, 2011.

[19] R. Nag and D.C. Hambrick. What is strategic management, really? Inductive derivation of a consensus definition of the field. *Strategic Management*, 955, 2007.

[20] T. Palpanas, P. Chowdhary, G. Mihaila, and F. Pinel. Integrated model-driven dashboard development. *ISF*, 9(2-3), July 2007.

[21] T. Palpanas, R. Sidle, R. Cochrane, and H. Pirahesh. Incremental maintenance for non-distributive aggregate functions. *VLDB*, 2002.

[22] N. Polyzotis, S. Skiadopoulos, P. Vassiliadis, A. Simitsis, and N.E. Frantzell. Supporting Streaming Updates in an Active Data Warehouse. *ICDE*, 2007.

[23] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-Driven Exploration of OLAP Data Cubes. *Advances in Database Technology*, 1377, 1998.

[24] V.E.S. Souza, I. Garrigós, and J. Trujillo. Monitoring Strategic Goals in Data Warehouses with Awareness Requirements. In *ACM Symposium on Applied Computing*, 2012.

[25] P. Vassiliadis, A. Simitsis, and P. Georgantas. A generic and customizable framework for the design of ETL scenarios. *Information Systems*, 2005.

[26] H.J. Watson, B.H. Wixom, J.A. Hoffer, R. Anderson-Lehman, and A.M. Reynolds. Real-time business intelligence: Best practices at continental airlines. *Information Systems Management*, 23(1), 2006.

[27] R. Xi, N. Lin, and Y. Chen. Compression and aggregation for logistic regression analysis in data cubes. *TKDE*, 21(4), 2009.