TECHNIQUES FOR DATA WAREHOUSE MODELING IN CONTEXT OF AGILE BUSINESS INTELLIGENCE

Radek Němec
Faculty of Economics, Technical University of Ostrava, radek.nemec@vsb.cz

ABSTRACT:
The aim of this paper is to evaluate benefits of data warehouse modeling tools and techniques, in the context of typical features of an agile approach to building Business Intelligence system. The paper specifies 4 mostly discussed data warehouse modeling techniques and presents assessment of 2 aspects of agile BI that were identified as key in case of data modeling. The paper identifies several drawbacks and advantages of each technique in context of each aspect of agile BI and numerically evaluates their characteristics.

KEYWORDS:
Agile Business Intelligence, BI, data warehouse, data modeling techniques

1. INTRODUCTION

Current advances in information technology usage in business and evolution of businesses themselves place heavy burden on underlying implementation processes of business crucial information systems. Development actions need to be carried out very quickly since business requirements usually tend to change in a matter of weeks or even days, as illustrated by a Forrester Research survey among 226 U.S. companies (Forrester, 2010b). Current state of Business Intelligence (BI) system development shows that there are still many implementations that fail at fulfilling key principles of successful BI system deployment. Among factors that mostly influence the success of BI implementation is a strong BI strategy with clearly set vision and goal of BI system and healthy user expectation of BI tools functions (Biere, 2011). Despite these project and usage requirements there is also a technology layer which is also an important factor in BI system success. Means of storing data for the BI system usually in a data warehouse need to be modeled in a way that facilitates changes without severe intervention in the underlying architecture of the database. In last few years there were serious discussions about the agility of BI. Agile BI is a lately discussed way of planning, developing, managing and even using the BI system, while considering rapid nature of business requirements' changes and the need to deliver BI functionalities as fast as possible and according to customers' needs.

1.1. Methodology

The goal of this paper is to assess data modeling techniques referenced as commonly used to or viable for modeling information architecture of a data warehouse. The respective data modeling techniques will first be specified (results of a short research on data modeling techniques, commonly discussed in connection with data warehousing) and then studied in context of agile BI's methodological aspects.

As a part of techniques' assessment methodology there will be a numerical evaluation according to aforementioned aspects and their characteristics. The numerical evaluation will be represented by positive and negative points depending on what mechanisms suffice respective needs of agile BI's methodology.

2. AGILE BUSINESS INTELLIGENCE

Business Intelligence is commonly referred to as an umbrella term for a complex of technologies, databases and tools for computerized support of management and decision making process in a company (Turban, et al., 2007). The BI also encompasses processes that lead to design of a BI solution and its post-implementation management.

In the field of the BI solutions' design there are significant drawbacks reported through years of attempts on implementation of BI systems. Many BI projects fall short in requirement gathering and analysis phase which tends to take a long time to carry out (Howson, 2008). Business requirements also tend to change over time which even more complicates the traditional process of developing a BI solution. Requirements analysis is one of crucial sources for the design of data warehouse data model (Ballard, et al., 2006) and in the context of this paper it will be assumed as a key part of the overall BI design process.

The aforementioned problem in the requirements gathering and analysis phase (and subsequent design of data warehouse data model) is one of key areas which is the agile BI concept mostly concerned with. According to Forrester (Forrester, 2010a) agile BI methodology takes advantage of agile development principles as well as best practices while trying to answer several typical issues connected with BI development by emphasizing:

- iterative (agile) data warehouse development with use of prototyping techniques,
- responsiveness and flexibility of designed system,
- minimization of IT liaisons' involvement and encouragement of direct interaction with business users instead of creating extensive documentation.

There are 5 fundamental functionalities that an agile BI architecture; platform and application should provide (Forrester, 2010a):

1. **Integrated metadata** – integration of metadata of all BI systems' components to facilitate easy and fast traceability of problems and bugs all the way to the source.
2. **End-to-end integrated BI component management** – using either centrally managed metadata to auto-generate ETL and DDL scripts or architectural stack with data, data model, metadata and BI in one packaged solution.
3. **DBMSes built from the ground up for BI (reporting and querying)** – DBMS built and optimized mostly for BI purposes requires less data modeling, are highly optimized for exploration queries (not insert/update/delete) and as a result can react faster to new business requirements with less significant modeling, architecting, and development efforts.
4. **Data-driven, not schema-driven, BI** – traditional schema-based BI is complicated when it comes to handle very complex or unstructured data and changes in data sources. New generation data-driven BI tools directly reflect changes in sources and facilitate creation of reports according to current state of source systems (so called self-service BI).
5. **Ability to handle complex data structures** – truly agile BI applications have to seamlessly and innately integrate any type of data and content and provide capabilities to do faceted type search versus traditional OLAP to analyze complex data structures with thousands of dimensions and complex hierarchies.

Forrester (Forrester, 2010a) also points out the fact that the traditional strategic planning of BI is not deprecated at all. Rather they emphasize combined approach of strategic planning of big picture of the BI system on one side and rapid deployment of prototypes on the other side as well as proper assessment usability of agile BI methodology in respective use cases.

3. REVIEW OF DATA WAREHOUSE MODELING TECHNIQUES

After searching for resources that discuss data modeling techniques in the field of data warehousing (using Google search engine and SpringerLink journal search engine), 4 approaches were found. Following section presents short description of each technique.

3.1. Normalized Data Warehouse (E-R modeling) (3NF)

E-R modeling is a classical database modeling technique widely used since 80s for creation of traditional OLTP systems. E-R modeling uses normalization techniques to facilitate effective data storage and minimal information redundancy. Although 3rd normal form database is commonly optimized for effective storage of data, not complex ad-hoc querying (Patel, et al., 2012 p. 6), William Inmon advocates the E-R modeling as a better option for data warehousing purposes since current state of technology allows to maintain excellent querying results while having data model in a nearly same form as in source systems (Inmon, 2005). Inmon’s approach is commonly recognized as the Corporate Information Factory (CIF).

3.2. Dimensional Data Modeling (DM)

Dimensional modeling technique is a widely used technique for modeling multidimensional data warehouses that facilitates effective storage for querying large amounts of data to get the desired information. According to Kimball and Ross (Kimball and Ross, 2002) the dimensional modeling is based on a concept of **dimensional tables** that hold descriptive and lookup data for facts (metrics) and is usually denormalized (1st normal form) to gain superior query performance. **Fact tables** are highly normalized and hold information (at varying granularity level) on performance of a business process that is covered by the dimensional model and dimensional and fact tables form a star or snowflake schema. Star (snowflake) schemas can be combined into a galaxy schema where several dimensions are usually shared by multiple fact tables (shared dimensions). Kimball’s approach is commonly known as Enterprise Bus Matrix.

3.3. Data Vault Modeling (DV)

Data vault modeling is a novel technique for modeling a data warehouse that offers tools and mechanisms that leverage the best of E-R modeling and star schema modeling and is designed to handle dynamic changes to relationships between information (Linstedt, 2002). Data Vault modeling is meant to facilitate flexibility and scalability of a data warehouses’ data model and employs date/time stamping of data tuples to maintain historical evidence on evolution of data. The Data Vault model is composed of **Hub entities** (entities composed of unique business keys – a variation of a fact table), **Link entities** (physical representation of a many-to-many 3NF relationship and represents the relationship or transaction between two or more business components – business keys) and **Satellite Entities** (contain Hub key descriptive information – satellite is a subject to change over time therefore the structure must be capable of storing new or altered data at the granular level – a variation of a dimensional table). Each entity is normalized in 3rd normal form.

3.4. Anchor Data Modeling (AM)

„Anchor Modeling is an agile information modeling technique that offers non-destructive extensibility mechanisms enabling robust and flexible management of changes. A key benefit of Anchor Modeling is that changes in a data warehouse environment only require extensions, not modifications.“ (Regardt, et al., 2009) According to authors, anchor modeling enables robust and flexible management of changes that occur in source systems and thus mirroring them with ease in the BI solution. Anchor modeling is a highly decomposed relational database schema that takes advantage of a 6th normal form (an extension to a 5th normal form by addition of temporal validity of information). Anchor data model is composed of **anchors** (relations with only 1 column that represents a set of entities), **knots** (relations that represent a fixed set of entities that do not change over time; knot is used to manage fixed properties that are shared by many instances of an anchor), **attributes** (attributes are used to represent properties of anchors and can be defined as changing in time or static) and **ties** (represents associations between two or more entities - anchors).

4. DATA WAREHOUSE MODELING TECHNIQUES – AN ANALYSIS IN CONTEXT OF AGILE BI’S ASPECTS

According to defined functionalities of an agile BI application, architecture or technology, there are several aspects that can be directly influenced by selected data modeling technique. 2 main aspects were then derived from these functionalities:

A1) Data warehouse schema should be robust, flexible and have implicit mechanisms for absorption of changes in source systems.
A2) Data warehouse schema should be optimized for querying of large amount of data.

Table 1 shows results of assessment of reviewed data modeling in context of specified aspects of agile BI according to respective literature that deals with each data modeling technique with numerical rating of their characteristics.

Aspect **Technique**	**A1**			**A2**		
3NF	Date/timestamp as a part of primary key to maintain temporal validity	**+1**	**-2**	Extensive indexing strategy is needed to obtain viable query performance (lots of joins)	**-1**	**-2**
	Parent-child complexities (complicated cascading change impacts)	**-1**				
	Difficulties in near real-time loading and troublesome query access (due to lots of table joins)	**-1**		Inmon’s CIF methodology enforces creation of enterprise-wide data model of a data warehouse so indexes could acquire substantial size and subsequently slow down queries	**-1**	
	Problematic drill-down analysis	**-1**				

DM	Slowly Changing Dimensions (SCD) data versioning mechanism – partial absorption of requirements' changes (a system of temporal validity attributes in dimensional tables that delimit historical validity of facts)	**+1**	**+3**	Schema is designed for superior query performance – reduced number of joins due to denormalization of database schema	**+1**	**-1**
	Shareable dimensional table representing occurrence of facts in time	**+1**		Proper indexing strategy is needed to maintain query performance	**-1**	
	Kimball's approach leads to creation of business process oriented data marts that offer flexibility by addressing needs a of specific business area (extension of schema by adding or altering dimensions and adding foreign keys in fact tables according to changes in the respective business process)	**+1**		Query performance can falter when dealing with dimensions/fact tables containing lots of attributes/foreign keys. (Linstedt, 2002), (Kimball and Ross, 2002)	**-1**	
DV	Implicit mechanism of extension –temporal validity of tuples in satellite entities using date/timestamp attributes	**+1**	**+2**	Increased performance over 3NF and star schema due to decomposition of the data model and reduction of redundancy was observed	**+1**	**+1**
	Link tables facilitate usage of many-to-many relationships and thus possible extensions of relationships between key business entities – allows for iterative development of data warehouse	**+1**		Decomposition facilitates easier maintenance (reduction in data loading volumes) due to implicit reuse of data in tables. (Linstedt, 2002)	**+1**	
	The model scope could be targeted on a specific business process	**+1**		Proper indexing strategy is needed to maintain query performance.	**-1**	
	Issues with drill-down analysis could arise according to model's scope and size	**-1**				
AM	High degree of normalization with high level of data reuse – the decomposition of database schema into components facilitates building and altering the data model incrementally	**+1**	**+2**	6NF leads to explosion of tables however many database systems enforce table elimination technique to speed up very complex queries using large amount of tables in the join	**+1**	**+1**
	The model's scope can be targeted and developed for respective business process	**+1**		Query performance tests exhibited greater performance of a sample anchor model over a 3NF model (Regardt, et al., 2009)	**+1**	

	Attributes and ties can be "historized" (enriched with information on temporal validity) – a tool for absorption of changes in time	**+1**		Large number of tables requires an extensive indexing strategy for proper query performance (according to model's size) – authors also suggest using a columnar data store due to existence of tables with low count of attributes	**-1**	
	Due to large normalization there can arise possible issues with drill-down analysis	**-1**				

Table 1 Specifications of data modeling techniques in context of defined agile BI aspects with numerical rating of each characteristic.

5. CONCLUSIONS

The paper dealt with the term agile Business Intelligence and its aspects in context of data warehouse modeling. There were 4 data modeling approaches (techniques) found that were indicated as mostly discussed according to results of a short research. According to respective agile BI's specifics there were 2 main aspects of agile BI derived that were selected as key in context of modeling the data warehouse. Data modeling techniques were then studied in context of these 2 aspects and their characteristics were numerically evaluated. Numerical ratings show distinct differences between each technique.

Traditional techniques of data modeling (Inmon's 3NF normalized data warehouse and Kimball's dimensional modeling) show some rigidity although being merely standardized in the field of data warehousing – issues of flexibility are addressed by changes in the schema structure which bring many anomalies and issues, like too wide dimensional tables, complicated changes propagation due to complexity of some parent-child relationships etc.

Data vault modeling and anchor modeling were specified as modern approaches for design of data warehouse. They mostly utilize normalization and/or decomposition of the data warehouse database schema into components that can be managed more flexibly and facilitate scalability of the schema. Although schema-based BI was stated as not very flexible in case of agile BI, some form of conceptualization of a data model is useful for respective project team members (developers, information architects).

In our research project we will focus on testing query performance of AM to complete list of performance references in contrast with DM.

LITERATURE

Ballard, Chuck, et al. *Dimensional Modeling: In a Business Intelligence Environment.* San Jose : IBM RedBooks, 2006.

Biere, Mike. *The New Era of Enterprise Business Intelligence: Using Analytics to Achieve a Global Competitive Advantage.* 1st ed. Boston : Pearson Education, 2011.

Forrester. 2010a. Best Practices For Breaking Through The BI Backlog. [Online] April 2010. [Cited: 24. 3. 2012.] http://resources.idgenterprise.com/original/AST-0008798_AgileBIBestPracticesForBreakingThroughTheBIBacklog.pdf.

Forrester. 2010b. Agile BI: Is It Time To Make The Move? [Online] July 2010. [Cited: 26. 3. 2012.] http://resources.idgenterprise.com/original/AST-0008799_AgileBIIsittimetomakethemove.pdf.

Howson, Cindi. *Successful Business Intelligence: Secrets of Making BI a Killer App.* New York : McGraw-Hill, 2008.

Inmon, W. H. *Building the Data Warehouse.* 4th ed. New York : Wiley, 2005.

Kimball, Ralph and Ross, Margy. *The Data Warehouse Toolkit.* 2nd ed. New York : Wiley, 2002.

Linstedt, Dan E. Data Vault Series 1 - Data Vault Overview. *The Data Administration Newsletter.* [Online] 1. 6. 2002. [Cited: 20. 3. 2012]. http://www.tdan.com/view-articles/5054/.

Patel, Alpa R. and Patel, Jayesh M. Data Modeling Techniques for Data Warehouse. *International Journal of Multidisciplinary Research.* 2012, vol. 2, issue 2, pp. 240-246.

Regardt, Olle, et al. Anchor Modeling: An Agile Modeling Technique Using the Sixth Normal Form for Structurally and Temporally Evolving Data. *ER 2009.* Lecture Notes in Computer Science, 2009, vol. 5829, issue 1, pp. 234–250.

Turban, Efraim, et al. *Decision Support and Business Intelligence Systems.* 8th ed. New Jersey : Pearson Prentice Hall, 2007.

ACKNOWLEDGEMENT

This paper was made under financial support of Student Grant Competition, research project SP2012/184 “Analýza charakteristik modelování databázových schémat datového skladu s důrazem na agilní přístup k vývoji Business Intelligence systému.”