AgileKDD

An Agile Knowledge Discovery in Databases Process Model

Givanildo Santana do Nascimento
Federal University of Sergipe
São Cristóvão, Brazil
gsnascimento@petrobras.com.br

Adicinéia Aparecida de Oliveira
Federal University of Sergipe
São Cristóvão, Brazil
adicineia@ufs.br

***Abstract* — In a knowledge-based society, transforming data into information and knowledge to support the decision-making process is a crucial success factor for all the organizations. In this sense, the mission of Software Engineering is to build systems able to process large volumes of data, transform them into relevant knowledge and deliver them to customers to enable them to make the right decisions at the right time. However, companies still fail to determine a process model to be used in their Knowledge Discovery in Databases and Business Intelligence projects. This article introduces the AgileKDD, an agile and disciplined process for developing systems capable of discovering the knowledge hidden in databases, built on top of the Open Unified Process, KDD Process and CRISP-DM. A case study shows that AgileKDD can increase the success factor of projects whose goal is to develop Knowledge Discovery in Databases and Business Intelligence applications.**

Keywords – Knowledge Discovery in Databases; Business Intelligence; Agile Software Development; Software Process.

I. INTRODUCTION

The Organization for Economic Cooperation and Development (OECD) defined knowledge-based economies as: "economies which are directly based on the production, distribution and use of knowledge and information" [1]. In knowledge-based economies, the global competition is becoming increasingly based on the ability to transform data into information and knowledge in an effective way. Knowledge is equated with the traditional factors of production - land, capital, raw materials, energy and manpower - in the process of wealth creation. Thus, data, information and knowledge constitute key assets for all organizations working in this economic model.

Knowledge Management, Data Mining (DM), Knowledge Discovery in Databases (KDD) and, more generally, Business Intelligence (BI) are key concepts in a knowledge-based economy. BI applications have vital importance for many organizations and can help them manage, develop and share their intangible assets such as information and knowledge, improving their performance. For instance, investments made by Continental Airlines in BI had a Return on Investment (ROI) of 1000% due to increased revenue and reduced costs [2].

However, companies still face problems in determining a process model to be used to develop KDD and BI applications. As business requirements become more dynamic and uncertain, the traditional static, bureaucratic and heavy processes may not be able to deal with them. Recent researches have demonstrated that waterfall lifecycles and traditional software development processes are not successful in BI because they are unable to follow the dynamic requirement changes in a rapidly evolving environment [3]. As a software process is mandatory for KDD and BI development, one possible solution is to use an agile process, which is typically characterized by flexibility, adaptability, face-to-face communication and knowledge sharing.

This article presents AgileKDD, an agile software process designed to guide the KDD and BI applications development in a manner suitable for the current ever-changing requirement environments. The next sections are organized as follows: Section 2 describes the techniques for transforming raw data into information and knowledge. The Section 3 presents the agile software development processes. Section 4 presents the AgileKDD and a case study implemented to verify the AgileKDD applicability. Then, Section 5 presents related work, and, finally, Section 6 presents the conclusion and future work.

II. TRANSFORMING DATA INTO INFORMATION AND KNOWLEDGE

Raw data evolve into information and knowledge as they receive degrees of association, context and meaning [4]. The knowledge gained from the interpretation of data and information drives the knower to action, so knowledge is an important asset for organizations that operate in knowledge-based economies and markets. BI, as well as KDD, has the goal of transforming raw data into knowledge in order to support the decision-making process.

A. Knowledge Discovery in Databases

KDD is a nontrivial process of identifying valid, novel, potentially useful and understandable patterns in data [5]. The discovered knowledge must be correct, understandable by human users and also interesting, useful or new. In addition, the knowledge discovery method must be efficient, generic and flexible (easily changeable).

The KDD systematization effort has resulted in a variety of process models, including the KDD Process [5] and the Cross-Industry Standard Process for Data Mining (CRISP-DM) [6]. They are the most widely used in KDD projects and the most frequently cited and supported by tools. These two processes are considered the *de facto* standards in the

 ISBN: 978-1-61208-227-1

KDD area. Several other process models were derived from them. Figure 1 shows the evolution of 14 DM process models and methodologies. KDD Process can be pointed out as the initial approach and CRISP-DM as the central approach of the evolution diagram [7]. Most of the process models are based on them.

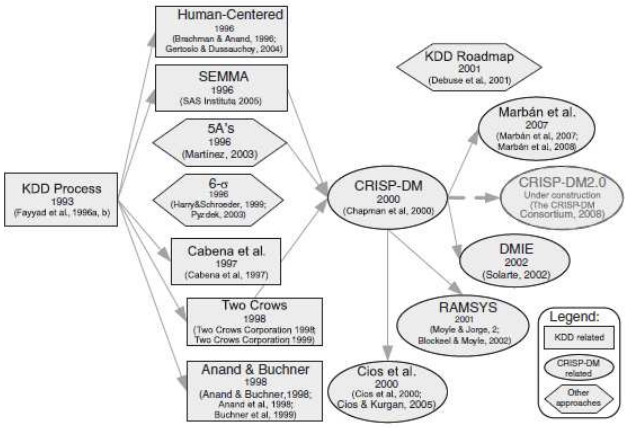

Figure 1. Evolution of data mining process models (Source: [7])

The KDD process models created between 1993 and 2008 were discussed in detail in a survey by Kurgan and Musilek [8] and then categorized by Mariscal, Marbán and Fernández [7] into three groups: (1) KDD related approaches; (2) CRISP-DM related approaches; (3) Other approaches.

Sometime later Alnoukari and El Sheikh [9] continued the older surveys done by Kurgan and Musilek [8] and Mariscal, Marbán and Fernández [7], and proposed a different categorization to the KDD process models: (1) Traditional approach; (2) Ontology-based approach; (3) Web-based approach; (4) Agile-based approach, which integrates agile processes and methodologies with traditional approaches. The main process models in this category are Adaptive Software Development – Data Mining (ASD-DM) [10] and Adaptive Software Development – Business Intelligence (ASD-BI) [1].

Thus, the knowledge discovery process models are evolving from traditional to agile processes, becoming more adaptive, flexible and human-centered [9]. However these processes still lack software engineering capabilities such as requirements management, project management and changes management.

B. Business Intelligence

Business Intelligence is an Information Technology (IT) framework vital for many organizations, especially those which have extremely large amounts of data, which can help organizations manage, develop and communicate their assets such as information and knowledge [2]. According to Mariscal, Marbán and Fernández [7], BI is a broad category of applications and technologies for gathering, storing, analyzing and providing access to data to help enterprise users make better business decisions.

The number of BI projects has grown rapidly worldwide according to Gartner Group annual reports. BI has been on the list of the top ten priorities in IT since 2005 and was at the top of this list for four consecutive years, from 2006 to 2009. In a broader sense, companies have understood that the information and knowledge provided by BI applications are essential to increase their effectiveness, support competitiveness and innovation. Thus, investments into data mining BI applications grew by 4.8% from 2005 to 2006 and by 11.2% from 2007 to 2008 [7] [11].

However, not all KDD and BI results are positive. Regardless of the priority and budgets growth, neither all the projects results were delivered [7] [12]. Many BI projects had failed to achieve their goals or were canceled because they were unable to follow the dynamic requirement changes in rapidly evolving environments. BI left the top of the list of priorities in IT and, in 2010 and 2011, dropped to the fifth position. Technologies with higher productivity, lower risk and faster ROI were prioritized instead [13].

Moreover, many companies still develop BI applications without the guidance of a software process. As any software projects, BI projects need a software process to succeed. Also, the dynamic business requirements, the needs of faster ROI and fluid communication between stakeholders and the team led to agile process as one possible solution.

III. AGILE SOFTWARE ENGINEERING PROCESSES

A software process provides an ordered sequence of activities related to the specification, design and implementation as well as validation and development of software products, transforming user expectations into software solutions [14]. According to Pressman [15], the software processes set the context in which technical methods are applied, the work artifacts (models, documents, data, reports, forms) are produced, the milestones are established, quality is assured and changes are managed.

The traditional software development processes are characterized by rigid mechanisms with a heavy documentation process, which make it difficult to adapt to a high-speed, ever-changing environment [16]. Agile approach is one answer to the software engineering chaotic situation, in which projects are exceeding their time and budget limits, requirements are not fulfilled and, consequently, leading to unsatisfied customers [17].

The Manifesto for Agile Software Development [18] defines the values introduced by the agile software processes. Based on these values, agile processes are people-oriented and have the customer satisfaction as the highest priority through the early and continuous delivery of functioning software. Agile approaches are best fit when requirements are uncertain or volatile; this can happen due to business dynamics and rapidly evolving markets. It is too difficult to practice traditional plan-oriented software development in such unstable environments [16].

Open Unified Process (OpenUP) is a variation of the Unified Process (UP) [19] that applies agile, iterative and incremental approaches within a structured lifecycle. OpenUP is a low-ceremony process that can be extended to address a broad variety of project types [20]. OpenUP has

 ISBN: 978-1-61208-227-1

compliance with the Manifesto for Agile Software Development, is minimal, complete and extensible. Moreover, it increases collaboration and continuous communication between project participants, more than formalities and comprehensive documentation [21].

The development of BI solutions must be guided by a software process. Therefore, it is mandatory to define processes that address aspects of KDD and BI, as well as the disciplines introduced by the software engineering process models. By the other hand, traditional processes are not successful in BI because they are unable to follow requirements in ever-changing environments [3]. Hence, one possible solution is to use an agile process, which is typically characterized by flexibility, adaptability, communication and knowledge sharing.

IV. AGILEKDD

AgileKDD is an agile and disciplined process for the development of KDD and BI solutions. CRISP-DM and KDD Process provide to AgileKDD the activities related to knowledge discovering. OpenUP provides the lifecycle, the phases and the disciplines, which are requirements, architecture, development, test, project management and changes management. OpenUP also adds the agile software development core values and principles, without giving up the management disciplines. The personal effort on an AgileKDD project is organized in micro-increments. They represent small work units that produce measurable steps in the project progress. The process applies intensive collaboration between the actors as the system is built incrementally. These micro-increments provide extremely short cycles of continuous feedback to identify and resolve problems before they become threats to the projects.

AgileKDD divides the projects in planned iterations with fixed time boxes, usually measured in weeks. The iterations drive the team to deliver incremental value to stakeholders in a predictable manner. Iteration plan defines what must be delivered during the iteration and the result is a demonstrable or deliverable piece of the KDD or BI solution. The AgileKDD lifecycle provides stakeholders and project team visibility and decision points at various milestones, until a working application is fully delivered to stakeholders. Figure 2 presents an overview of AgileKDD, highlighting its phases and activities.

The Inception (I) phase has the aim of developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the BI project from the customer's viewpoint. In this phase the project vision and plans are defined and agreed by all project participants. Also, in inception the target data set, or subset of variables and data samples, is selected. The knowledge discovery processes will be performed on the selected target data set. The data quality is a critical success factor for any BI project, so it is verified in Inception phase to indicate the project feasibility and quality constraints. Project management activity consists on high level project planning and governance concerns. Changes and configuration management activity is related to the version control of all the project artifacts, including documentation, sources and binaries.

The Elaboration (E) phase is responsible for the system's architecture and design, data modeling and applications integration.

Once data structures are modeled, the Construct (C) phase starts with Extract-Transform-Load (ETL) activity. ETL routines are built to extract, clean, integrate, transform and load the selected target data into databases. Also, ETL perform data cleaning to removes noise and decide on strategies for handling missing data fields. Thus, the DM techniques that best fit to the data are selected and applied to the information. DM tools search for meaningful patterns in data, including association rules, decision trees and clusters. The team can significantly aid the DM method by correctly performing the preceding steps. The reports, charts and dashboards are built to allow user information access. The verification and validation activities guarantee that the data was extracted, loaded and processed correctly, according to business objectives.

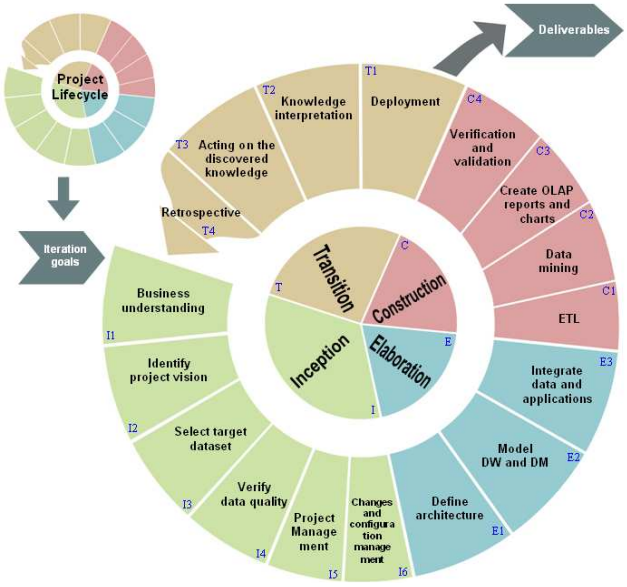

Figure 2. AgileKDD phases and lifecycle.

In Transition (T) phase the deployment of both software and knowledge takes place, the knowledge is interpreted, actions are created and the retrospective discusses lessons learnt during the project to promote continuous process improvement. Interpreting mined patterns involve visualization and storage of the extracted knowledge into knowledge bases, or simply documenting and reporting it to interested parties. This activity also includes checking for and resolving potential conflicts with previously believed knowledge. The AgileKDD process can involve significant iteration, interaction and can contain loops between any phases.

AgileKDD disciplines are the same of OpenUP: requirements, architecture, development, test, project management and configuration and changes management.

 ISBN: 978-1-61208-227-1

Table I shows the AgileKDD disciplines, their purposes and suggested work products.

During a full project cycle, most of the requirements discipline effort is concentrated in the inception phase. The architecture is the main discipline during the elaboration phase. In the same phase, the development is intensified from the definition of the system architecture and continues as the main discipline of construction phase. The tests occur mainly in verification and validation activity of construction phase. The project management discipline is concentrated predominantly in the inception phase. The configuration and change management has greater prevalence in inception and transition phases. Each discipline can be related to a set of work products created during the process phases.

TABLE I. AGILEKDD DISCIPLINES

Discipline	Purpose	Work products[a]
Requirements	Elicit, analyze, specify, validate and manage the requirements for the system being developed.	Vision document. Initial project glossary. Prototypes.
Architecture	Define an architecture for the system components.	Software architecture description. DW and DM models.
Development	Design and implement a technical solution adherent to the architecture that meets the requirements.	Software components. Integrated software increment.
Test	Validate system maturity through the design, implementation, execution and evaluation of tests.	Plan and test procedure. Test record.
Project management	Instruct, assist and support the team, helping them to deal with risks and obstacles faced when building software.	Project plan. Feasibility and risk evaluation.
Configuration and change management	Controlling changes in artifacts, ensuring a synchronized evolution of the set of artifacts that make a software system.	Work items list.

a. All the work products are optional. Only the necessary artifacts must be produced.

AgileKDD applicability has been verified by a case study in oil and gas area. The process was applied to a KDD and BI project that deals with Reservoir Evaluation data and afforded the early delivery of DM results two months after the project kickoff.

The first iteration was dedicated exclusively to the project inception. This phase aimed to identify the product requirements, to the communication with customer, project management, configuration and change management. The second iteration aimed to delivery data mining results related to Reservoir Evaluation (RA) data. The third iteration aimed to calculate the RA performance indicators and present them to users in dashboards. The fourth iteration aimed to deliver the online analytical processing (OLAP) features, including reports, graphs, and *ad hoc* exploration of the data warehouse.

It was observed that AgileKDD process was able to guide the product development since the beginning of the inception iteration to the transition phase of the last iteration performed. At the end of the case study, it was verified that some adjustments were needed in the process to improve its fitness for BI and KDD systems projects. The observations and identified adjustments needs helped to improve the process final version.

V. RELATED WORK

The main work that applies agile methodologies to KDD and BI is [1]. Alnoukari [16] discusses BI and Agile Methodologies for knowledge-based organizations in a cross-disciplinary approach. Alnoukari [22] introduces Adaptive Software Development – Business Intelligence (ASD-BI), a knowledge discovery process model based on Adaptive Software Development agile methodology. Likewise, Alnoukari, Alzoabi and Hanna [10] defined Adaptive Software Development – Data Mining (ASD-DM) Process Model. The main difference between this work and these is the fact that AgileKDD is a software process, not a methodology. As a process, AgileKDD defines what to do instead how to do KDD and BI development. Also, the process proposed by this work defines lifecycle, roles, activities, inputs and outputs regarding agile KDD and BI application development. Moreover, the process AgileKDD contains management disciplines like project, changes and requirements management, which were inherited from OpenUP.

Three surveys about DM and knowledge discovery process models and methodologies are discussed and compared by Mariscal, Marbán and Fernández [7], Kurgan and Musilek [8] and Alnoukari and El Sheikh [9]. All the process models and methodologies presented by these works focus on DM and knowledge discovery, and do not consider other BI components. As BI is more comprehensive than data mining, this work focuses on an agile process modeled to address both KDD and BI software projects, in an adaptable, flexible and systematic manner.

The objective of this work was building a software process capable of guiding KDD and BI projects in an agile and adaptive way. The cornerstone of this work was a software process, the OpenUP, created from the UP, inheriting the maturity of this process in an agile approach. Existing works relied on brand new agile methodologies, which lack of software engineering capabilities and were not scientifically proved yet.

VI. CONCLUSION AND FUTURE WORK

A software process is mandatory for KDD and BI development. However, traditional software development processes are not successful in KDD and BI because they are unable to follow the dynamic requirement changes in an ever-changing environment. Agile processes fit in KDD and BI better than traditional processes because they are characterized by flexibility, adaptability, communication and knowledge sharing.

This work presented AgileKDD, a KDD and BI process based on the Open Unified Process. AgileKDD applicability has been verified by a case study and the results indicate that software development organizations may apply AgileKDD

 ISBN: 978-1-61208-227-1

in implementing knowledge discovery projects. The process brought such benefits as more customer satisfaction through early and continuous delivery of functioning software, better communication between team members and reduced project failure risks.

The main contribution of AgileKDD is its ability to guide the BI solutions development according to the practices present in agile software development processes. AgileKDD can increase the projects success factor and customer satisfaction. The process can be used to guide BI and KDD applications projects in scenarios of continuous requirements evolving and early ROI need.

Future work can validate AgileKDD by more case studies in different areas and improve its capabilities to store the knowledge discovered in ontology bases or knowledge bases.

REFERENCES

[1] A. El Sheikh and M. Alnoukari, Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications. IGI Global, 2012. pp. 1-370.

[2] M. Alnoukari, H. Alhawasli, H. Alnafea, and Amjad Zamreek, "Business Intelligence: Body of Knowledge" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 1-13.

[3] D. Larson, "Agile Methodologies for Business Intelligence" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 101-119.

[4] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, Sixth Edition. Pearson, 2010. pp. 1200 pp.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. "From data mining to knowledge discovery: an overview" in *Proc. Advances in Knowledge Discovery and Data Mining*, 1996, pp. 1–34.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[7] G. Mariscal, O. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies". *The Knowledge Engineering Review*, vol. 25, 2010, pp. 137-166.

[8] L. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models". *The Knowledge Engineering Review*, vol. 21, 2006, pp. 1-24.

[9] M. Alnoukari and A. El Sheikh, "Knowledge Discovery Process Models: From Traditional to Agile Modeling" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 72-100.

[10] M. Alnoukari, Z. Alzoabi, and S. Hanna, "Applying adaptive software development (ASD) agile modeling on predictive data mining applications: ASD-DM Methodology" in *IEEE Proceedings of International Symposium of Information Technology*, 2008, pp. 1083–1087.

[11] M. McDonald, M. Blosch, T. Jaffarian, L. Mok, and S. Stevens, "Growing It's Contribution: The 2006 Cio Agenda". Gartner Group, 2006.

[12] Gartner Group, "Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007". 2005. Available: http://www.gartner.com/it/page.jsp?id=492112 [retrieved: Out., 2012].

[13] Gartner Group, "Gartner Executive Programs Worldwide Survey of More Than 2,000 CIOs Identifies Cloud Computing as Top Technology Priority for CIOs in 2011". 2011. Available: http://www.gartner.com/it/page.jsp?id=1526414 [retrieved: Out., 2012].

[14] I. Sommerville, Software Engineering. Addison Wesley, 2006, pp. 864.

[15] R. Pressman, Software Engineering: A Practioner's Approach. McGraw-Hill, 2005.

[16] M. Alnoukari, "Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications" in *CEPIS UPGRADE: The European Journal for the Informatics Professional*, vol. 12, pp. 56–59, 2011. Available: http://www.cepis.org/upgrade/media/III_2011_alnoukari1.pdf [retrieved: Out., 2012].

[17] Z. Alzoabi, "Agile Software: Body of Knowledge" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 14-34.

[18] K. Beck et al Manifesto for Agile Software Development. 2001. Available: http://agilemanifesto.org [retrieved: Out., 2012].

[19] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide. Addison Wesley, 1999.

[20] H. Hristov, Introduction to OpenUP. 2011. Available: http://epf.eclipse.org/wikis/openup/index.htm [retrieved: Out., 2012].

[21] S. Santos, OpenUP: Um processo ágil. 2009. Available: http://www.ibm.com/developerworks/br/rational/local/open_up/index.html [retrieved: Out., 2012].

[22] M. Alnoukari, "ASD-BI: A Knowledge Discovery Process Modeling Based on Adaptive Software Development Agile Methodology" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications.* IGI Global, 2012. pp. 1-13.

 ISBN: 978-1-61208-227-1