Discussion paper: Filling the gap between business rules and technical requirements in Business Analytics: The Fact - Centered ETL approach

Antonella Longo[1], Mario Bochicchio[1], Marco Zappatore[1], Lucia Vaira[1]

[1] Set-Lab, Dept. of Engineering for Innovation, Univ. of Salento,
via per Monteroni, 73100 Lecce, Italy
{antonella.longo, mario.bochicchio, marcosalvatore.zappatore,
lucia.vaira}@unisalento.it

Abstract. In real-time Business Analytics scenarios, like Business Activity Monitoring (BAM) or Operational Intelligence, using modeling language to describe ETL processes is fundamental to provide business users with up-to-date data in order to support decision-making process and to optimize business operations. To bridge the gap between business and technical requirements and encapsulate business requirements into technical specifications, it can be very convenient to design ETL processes starting from business facts; this would effectively support business users in decision-making processes and would represent business rules and entities as soon as possible in the BI process design. Moreover a fact centered approach to ETL process design can split the traditional black-box approach into several fact-centered flows, which can be processed in parallel exploiting the advantages of distributed multi-threaded process models, typical of big data scenario. In this context we propose a control-flow- based approach to ETL process modeling, which starts from business facts identification, and represents ETL processes using BPMN notation, which is the foundation for machine-readable code. Our main contribution consists in a proposal for structuring ETL processes and related objects and in its application to a business case. The paper has been already published at Procedia Technology, Volume 16, 2014, Pages 471-480, and this version shows only minor updates

Keywords: Business Intelligence; ETL; Process & Conceptual Models; Operational Intelligence.

1 Introduction

Business Intelligence (BI) can be defined as the process of getting information about the business from available data sources [6]. It assists managers to take correct decisions based on facts given at the right time/place throughout the life of the business.

The exponential growth of data streams is a big challenge for BI and Business Analytics (BA) as decisions based on fresh information represent a competitive advantage. This challenge is addressed by Real-time BI and Operational Intelligence (OI) that provide visibility into Business Processes (BPs), streaming events, and operations as they are happening. These topics go under Big Data challenges, where the new frontier of data management must combine big data volumes with event-centric, speed up processing that allow delivering and accessing information with low latency. Conceptually similar to OI, Business Activity Monitoring (BAM [7] is an enterprise

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy. Copyright held by the author(s).

solution that allows monitoring business activity across operational systems and BPs. It refers to the aggregation, analysis, and presentation of real-time data about customers' and partners' activities inside and across organizations. BAM is the real-time and event-driven extension of BI: while BI products usually handle historical data stored into a Data Warehouse (DWH), BAM technologies provide managers with real-time business analyses obtained by operational data sources, more and more integrated with cloud-based resources for achieving faster Return of Investments and lowering costs [3]. The success of OI, BAM and BI systems depends mostly on the adequacy of the populating system: ETL (Extract-Load-Transformation) processes are, therefore, the critical feeding component of DWH/BI systems as they retrieve data from operational systems and pre-process them for further analysis [12]. An ETL system translates specific decision-making processes into system rules but there are several challenges: volatile business rules, heterogeneous operational data schemas, proprietary ETL tools, different notations, complex ETL porting processes. As of today, there is no widely adopted methodology covering the ETL development cycle, with an easy notation for user profiles, allowing to map the model on the execution environment and offering pre-implementation validation features. We believe that in order to bridge the gap between business and technical requirements and encapsulate business requirements into technical specifications, ETL processes should be designed by starting from business facts (BFs), thus supporting business users in decision-making processes and representing business rules and entities as soon as possible in the BI process design. Such a fact-centered approach to ETL process design can split the traditional black-box approach into several fact-centered flows, which can be processed in parallel exploiting the advantages of distributed multi-threaded process models, typical of big data scenario. In this paper, the advantages of control flow models and BPMN notation are applied to ETL process design, as defined in [4, 1, 5]. First, design patterns allow structuring the process and selecting required data. Second, a uniform notation for BPs and DWH modelling eases the communication between IT and business roles. Third, non-functional requirements (e.g., performance indicators, data freshness) can be specified more easily. We propose to structure the ETL design process starting from the identification of BFs and to define a reference model for representing the ETL process. The integration between the fact-centered approach and control flow models allows identifying relevant business objects and isolating related facts in order to handle them independently.

The paper is organized as follows. Section 2 analyzes existing works on ETL process conceptual models; Section 3 discusses our case study; the proposed approach and application are described in Section 4; conclusions are presented in Section 5.

2 Related Works

Several efforts have been proposed for the conceptual modeling of ETL processes, including ad hoc formalism approaches based on standard languages like UML [10] or MDA [11]. Conceptual models that use BPMN notation for ETL design have been also already proposed. The work we present in this paper is built along the lines of [1] and [4, 5]. In the former Akkaoui, Mazón and others present a conceptual model based on the BPMN standard and provide a BPMN representation for frequently used ETL design constructs. They propose a meta-model for conceptual modeling of ETL processes based on the separation between control and data processes, and on the classification of ETL objects resulting from a study of the most used commercial and open source ETL tools. The issue of ETL conceptual view is also addressed in [4, 5]; authors propose the use of BP models for a conceptual view of ETL, and show how to translate this conceptual view to a logical and a physical ETL view that can be optimized. Our work differs from those described because we propose a fact centered approach to ETL design which allows obtaining analytical up-to-date data in a timely and efficient manner, and a control flows based modeling of ETL processes, represented with BPMN notation.

3 Preliminaries and running example

A BI system is based on data sources analysis, design and implementation of the DWH that store analytical data and on related ETL procedures, according to its typical 4-level architecture (Fig. 1) [8]. The integration of heterogeneous sources is composed of ETL procedures for extracting, integrating, cleaning, validating, filtering data and then loading them into a staging area, posed in a reconciled model which creates a central data reference model for the whole enterprise and introduces a clear separation between extraction and integration problems and issues related to DWH loading process. The DWH can be directly consulted or used as a source to build data marts (i.e., data subsets/aggregations in the DWH). The concepts of interest in the decision-making process are named facts, which typically match with events that occur within the enterprise. Every fact represents a set of events, quantitatively described by measures and analysis dimensions, detailed with hierarchies of attributes. They are modeled as Dimensional Fact Models (DFMs) [8] so that the efficient consultation of integrated data is available for reports, analyses and simulations.

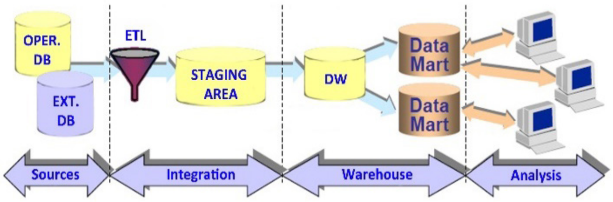

Fig. 1. Data warehousing system architecture [8]

ETL procedures require constraints checks for filtering data according to quality rules in terms of business rules and format inconsistencies. Some these rules (e.g., exception handling, data workflow patterns) can be derived from control-flow patterns. Wrong data feed a specific database that logs error types, source tables and execution dates. In our approach we propose a method to structure ETL processes based on relevant BFs, and to model them according to control-flow models.

Let us now discuss our running example: an Italian company needs a Service Level Agreement (SLA) management platform to calculate and present service levels associated with fidelity cards printing and delivery services. Operational data from several relational data sources must be mapped into hyper-cubes via an ETL process in order to produce reports and booklets for controlling the quality of services according to SLAs. Starting from the operational data sources, the system runs ETL procedures to populate the reconciled model and then the hyper-cubes. Data stored into the operational database is about fidelity cards details, and related events, which correspond to facts defined into the DWH. In this running case we describe the application of the proposed approach to fidelity cards printing, delivery and complaint events.

4 The fact-based ETL process

BFs are business artifacts meaningful to business. They can be modeled as business domain objects for transactional aims or as DFM for analytical processing. Once a BF has been defined in the domain, it must be populated starting from heterogeneous data sources. As in [1], we consider the ETL process as a combination of two perspectives. First, a *control flow process*, which manages the branching and synchronization of the flows, handles execution errors and exceptions and coordinates information exchange with external processes. Second, a *data process* that sends operational to the DWH, thus providing insights about inputs and outputs of each process element. Therefore, in Fig.2 we propose our fact population data process, which exploits the conceptual separation between quality management, data and fact population processes.

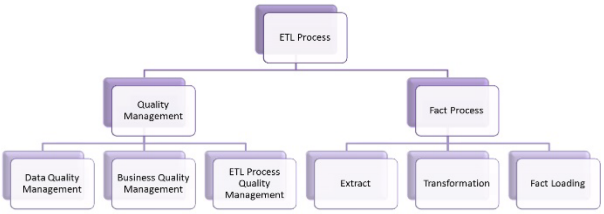

Fig. 2. ETL Process Model

The Quality Management can be further specialized as it includes data quality, ETL process and BP management; it can be considered as an event generator that checks the compliance with given constraints on business rules, data formats, integrity and ETL rules. The Quality Management handles quality checks within the whole ETL process, at different levels, via transformations exploiting syntactic and semantic rules, or business and ETL constraints. In particular, the Data Quality Management phase considers the syntactic accuracy of collected data and applies syntactic constraints to extract and load data into the reconciled model. These rules allow discarding incorrect data coming from errors that may affect the population process.

In addition to the rules on syntactic data accuracy, ETL processes need the definition and application of business rules that assess data coherence/consistency within the specific reference domain and which are defined into the Business Quality Management phase, which includes the procedures for applying business rules to data, in order to filter records out of constraints. The ETL Process Quality Management aims to filter data on constraints related to the ETL procedures (e.g., they verify correctness of data produced by ETL modules to prevent duplicated or contradicting data).

The output of the Quality Management process is the storage of filtered data into the reconciled model according to the Fact management process or into the log database that include a record for each violated constraint, with the following information: source table name, record identification, error code, error description, date check. This approach allows tracking data source errors, easily identifying reasons for records' discards and obtaining statistics such as the most violated constraint, and so on. The Fact Management Process handles data entry into data structures that can be staging area tables of reconciled models or facts inside a DWH. In both cases, the population process is constituted by a first phase of data extraction, followed by a transformation step and finally by the loading of transformed data into the specific reconciled or fact tables. In particular, the Extract activity performs the initial collection of data that will be transformed and loaded into the destination and for the integration of multiple data sources, managing topics such as format/conceptual heterogeneity, due to differences in communication protocols, data formats, schemas, vocabularies, etc. The Transformation process handles activities related to the second step of the ETL Process, which aims to obtain analytical data from operational ones, and to match input data with output data structures. Typical ETL transformation procedures include data summarization, derivation, aggregation, conversion, constants substitution, setting of values to null or default based on particular conditions, and so on. Finally, the Load module administers the population of the target table, which can be part of a reconciled database, a fact or a dimension, including the management of primary keys, integrity constraints and indexes.

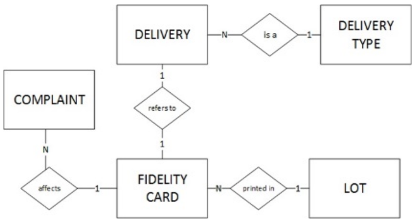

Fig. 3. Conceptual representation of reconciled model

In our model, we design an ETL process for each BF; therefore the whole ETL process can be considered as a set of fact processes and quality rules checks, which can be integrated and managed through control flow processes. We propose to design ETL starting from the definition of BFs and to split ETL process in order to isolate them within the flow. This approach: 1) allows managing facts independently, in a framework providing the integration aspects; 2) helps to structure ETL design; 3) allows business users working on their BFs even if the whole process is not completed yet. The result is the reduction of BF processing cycle, and the timely provision of analytical data. Splitting and fastening the ETL process is one of the steps for speeding BA process up enabling higher response times, which are fundamental in real-time BA systems. Furthermore, ETL process representation based on control flow models, and in particular on BPMN notation, adds information about accountability, as it allows linking process activities with the involved user profiles. Profiles typically involved in an ETL process include: 1) *Data Provider* (who assesses data readiness); 2) *Data Quality Supervisor* (who checks for data accuracy and correctness from a business point of view); 3) *Technician* (who handles the technical cycle). Roles definition, together with a BPMN-based conceptual modeling of the ETL process, improves accountability: at any time a problem occurs, managers are able to immediately know the activity that generates the problem and who is accountable for it.

Now we apply the proposed approach to our running example described in the previous section. Starting from a portion of the conceptual schema of the reconciled model, and from the DFM model corresponding to fidelity cards delivery, we illustrate the ETL process and the transformation that load the Delivery table of the reconciled model, both modeled with BPMN notation.

Fig. 3 shows the Entity-Relationship (ER) model that represents fidelity cards printing, delivery and complaints events. The printing event is modeled by the relationship between Fidelity Card and Lot entities: a printing lot includes several cards, and each card can be associated only with a lot. A delivery event is linked to each fidelity card: delivery detailed data, such as delivery date or payment, are included into Delivery entity, whereas Delivery Type entity is used to classify feasible delivery typologies, according to adopted procedures. The relationship between Fidelity Card and Complaint models complaint events: one or more complaint events can be registered on the same card, and each of them is linked to a single card.

Once the reconciled model is populated, a second step of ETL procedures elaborates data and store them into hyper-cubes. Fig. 4 (left side) shows the DFM for analytical data on fidelity cards delivery events (i.e, one of the outputs in the ETL process). Delivery events can be analyzed along many dimensions (e.g., state, card, delivery and registration date, etc.). By applying the proposed approach to ETL process that loads reconciled model tables, we represent the process in BPMN notation; Fig. 4 (right side) refers to fidelity cards printing, delivery and complaint events loading.

The process is represented as a set of complex activities, aggregated in sub-processes (each of them corresponds to a table of the reconciled model). Activity execution is modeled through the Parallel Split workflow control pattern, used when a single thread of control splits into multiple threads that can be executed in parallel [9]. In BPMN notation, the pattern is represented with a parallel gateway, to highlight the fact-centered approach that allows independently running and handling single tables' loading. The whole process ends when all the activities are concluded, according with Synchronization workflow control pattern, used when multiple parallel activities converge into one single thread of control, thus synchronizing multiple threads [9].

Each activity of the previous ETL process representation is a sub-process that can be further detailed and modeled with BPMN notation. As an example, Fig. 5 shows the BPMN representation of the fact process related to the Delivery table loading process. The fact process starts collecting operational data related to cards delivery

(Extraction phase), and this activity is in charge of the Data Provider. The process goes on with two quality controls, through which card identifier validity and delivery date existence are checked; the Data Quality Supervisor is accountable for these checks, whereas the Technician is responsible for error handling and event insertion into the appropriate table. Quality checks are modeled with the Exclusive Choice workflow control pattern, used when one of several branches is chosen, based on a decision or workflow control data [9]. Technician is also accountable for the following steps, related to lookups made in order to obtain cards and delivery codes, and to record sorting; this activity closes the Transformation phase. The last step of the fact process is the Loading phase, which inserts data into the Delivery table of the reconciled model. Activities inserted into this process are executed for each record of the data source Delivery table. The proposed approach applied to the running example allows organizing ETL process, by a conceptual separation between data quality management and facts loading procedures, and graphically representing the process, thanks to control flow models, design patterns and BPMN notation, in a fact-centered perspective, oriented to real-time BA.

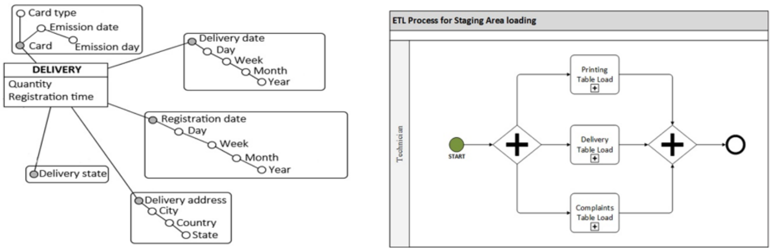

Fig. 4. DFM of cards delivery (on the left) and ETL process representation (on the right)

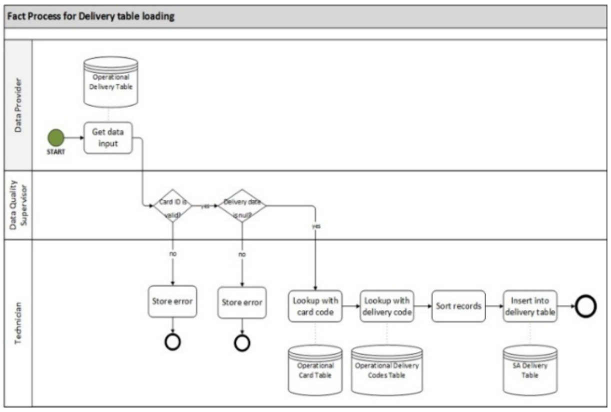

Fig. 5. Delivery Table Loading

5 Conclusions

The paper links conceptual ETL modeling with real-time BA and Big Data concepts. The changes in business requirements and needs as well as the growth of data volumes to be analyzed, lead to the introduction of new solutions to provide users with analytical data that support efficiently the decision-making process. In this context, we presented a fact centered approach to ETL design that aims at reducing ETL process granularity according to the identification of BFs so that they can be processed independently. Our proposal also structures ETL processes by distinguishing data quality management activities from fact processes, and by using control flow models, workflow patterns and the BPMN notation to support the conceptual representation of the ETL process. The proposed approach has been applied to a SLA Management platform managing service levels related to cards printing and delivery services.

Next steps will include the specification of a tool for graphically design and transform ETL process into ETL platform specific language. Moreover we are designing a wizard to support the definition of BFs, according to the business artifact theory [13].

Acknowledgement

This work is partially funded by Engineering S.p.A. and Essentia s.r.l.

References

1. El Akkaoui Z, Mazón JN, Vaisman A, Zimányi E. BPMN-Based Conceptual Modeling of ETL Processes. LNCS vol. 7448, 2012, pp 1-14.
2. Mell P, Grance T. The NIST Definition of Cloud Computing. Special Publication 800-145, National Institute of Standards and Technology, 2011.
3. Birst. Why Cloud BI? The 9 Substantial Benefits of SaaS BI. Birst, 2010.
4. Wilkinson K, Simitsis A, Dayal U, Castellanos M. Leveraging Business Process Models for ETL Design. Conceptual Modeling – ER 2010, pp 15-30.
5. Dayal U, Wilkinson K, Simitsis A, Castellanos M. Business Processes Meet Operational Business Intelligence. Bulletin of the IEEE Comp. Soc. TC on Data Engineering, 2009.
6. Azimuddin K, Karunesh S. Business Intelligence: a new dimension to business. Institute of Business Management, Pakistan Business Review, 2011.
7. Schmidt W. Business Activity Monitoring (BAM). Business Intelligence and Performance Management, Advanced Information and Knowledge Processing, 2013, pp 229-242.
8. Golfarelli M, Rizzi S. Data Warehouse. McGraw-Hill, 2006.
9. Van Der Aalst W. et al., Workflow Patterns. Distributed and Parallel Databases, 2003.
10. Luján-Mora S, Vassiliadis P, and Trujillo J. Data Mapping Diagrams for Data Warehouse Design with UML. In ER, pages 191–204, 2004.
11. Mazón JN, Trujillo J, Serrano MA, Piattini M. Applying MDA to the development of data warehouses. In DOLAP, pages 57–66, 2005.
12. Kimball R, Caserta J: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. 2004.
13. Cohn, D., & Hull, R. Business Artifacts. Bulletin of the IEEE Comp. Soc. TC on Data Engineering, vol. 32, pp. 1-7, 2009.