International Journal of Information Systems and Project Management

Volume 3 | Number 3 Article 4

2015

SOA enabled ELTA: approach in designing business intelligence solutions in Era of Big Data

Viktor Dmitriyev
University of Oldenburg

Tariq Mahmoud
University of Oldenburg

Pablo Michel Marín-Ortega
University of Oldenburg

International Journal of
Information Systems and Project Management
ISSN (print):2182-7796, ISSN (online):2182-7788, ISSN (cd-rom):2182-780X
Available online at www.sciencesphere.org/ijispm

SOA enabled ELTA: approach in designing business intelligence solutions in Era of Big Data

Viktor Dmitriyev
University of Oldenburg
Ammerländer Heerstr. 114-118, Oldenburg 26129
Germany
www.shortbio.net/viktor.dmitriyev@uni-oldenburg.de

Tariq Mahmoud
University of Oldenburg
Ammerländer Heerstr. 114-118, Oldenburg 26129
Germany
www.shortbio.net/tariq.mahmoud@uni-oldenburg.de

Pablo Michel Marín-Ortega
Central University Marta Abreu of Las Villas
Carretera a Camajuaní Km. 5 y 1/2, Santa Clara, Villa Clara
Cuba
www.shortbio.net/pablomo@uclv.edu.cu

Abstract:
The current work presents a new approach for designing business intelligence solutions. In the Era of Big Data, former and robust analytical concepts and utilities need to adapt themselves to the changed market circumstances. The main focus of this work is to address the acceleration of building process of a "data-centric" Business Intelligence (BI) solution besides preparing BI solutions for Big Data utilization. This research addresses the following goals: reducing the time spent during business intelligence solution's design phase; achieving flexibility of BI solution by adding new data sources; and preparing BI solution for utilizing Big Data concepts. This research proposes an extension of the existing Extract, Load and Transform (ELT) approach to the new one Extract, Load, Transform and Analyze (ELTA) supported by service-orientation concept. Additionally, the proposed model incorporates Service-Oriented Architecture concept as a mediator for the transformation phase. On one side, such incorporation brings flexibility to the BI solution and on the other side; it reduces the complexity of the whole system by moving some responsibilities to external authorities.

Keywords:
Big Data; Business Intelligence; BI; ETL; ELT; ELTA; SOA.

DOI: 10.12821/ijispm030303

Manuscript received: 21 September 2014
Manuscript accepted: 27 December 2014

1. Introduction

Companies are following different strategies in order to be competitive on the market, show permanent growth in generating revenue, increase return on investment (ROI) [1]. According to Porter [2], the advantages can be derived from following two aspects: operational efficiency and unique value creation for customers. Both aspects involved in building such enterprise structures and designing such business processes that function in a systemic and unique way.

In order to meet the two goals, operational efficiency and unique value creation, in most cases the business models and processes should become more complex and, because of such behavior, more performance power for systems is needed. However, not only system performance is important. According to the prior experience, the budgets of such projects originally dedicated more money for the hardware and software costs. However, the situation is changing and nowadays, hardware become less expensive than human sources, and we have situation in which we see the "people *versus* hardware" is contradictory in comparison with situation existed couple of decades ago - "computers were expensive and people were cheap" [3].

Therefore, discovering new value adding business process based on business historical behavior (extracted from data) to overcome competitors is emerging. Such efforts can be achieved with the support of business intelligence (BI). According to [4] current BI implementations suffer from several shortcomings:

- Missing focus on individual needs of particular analysts, analytical team or decision makers. These users are forced to rely on standard reporting tools and predefined analytical methods that often do not respond to all individuals and very case-specific needs. They strongly depend on either IT administration or own technical skills and IT expertise;
- The lack of information on business context level, such as definitions, business goals and company strategies as well as business rules and best practices for the provided analytical data. Hence, business users have to understand the semantics of data by themselves and take decisions besides deriving strategies using additional information sources (often may lead to an escalation of efforts and costs);
- Poor alignment between business and IT department. The setup and configuration of current BI systems requires deep insight in the data to be analyzed and the intended analytical tasks. Content and data models have to be provided in advance by the IT department and it must support the whole information in the decision-making process;
- The mean time for new BI implementations is between 3 and 6 months causing implementation and support costs to deter companies from having a wider BI deployment;
- BI solutions have a strong focus on structured, enterprise-internal data but lack the capability of integrating external and/or unstructured information in an easy, (near) real-time and effective way. Consequently, a lot of useful information is never included in the analysis. Not considering this information might provide a distorted or incomplete view of the actual world and consequently, it might lead to wrong business decisions.

Current work focuses on presenting a new approach for designing BI solutions. It will address the following goals: time reduction that is spent on BI solution's design phase, flexibility achievement in BI solution by removing "data agnosticism" and preparedness of BI solution to be used with big data. This research extends the existing ELT (Extract, Load and Transform) concept to an ELTA (Extract, Load, Transform and Analyze) one.

2. Background

2.1 Business Intelligence

Business intelligence systems support and assist decision-making processes. It is also taking part in the organization of strategic plans, which are normally addressing the achievement of management effectiveness. BI is defined as "a set of methodologies, processes, architectures and technologies that transform raw data into meaningful and useful information used to enable more effective strategic tactical, and operational insights and decision-making" [5]. Effective BI systems give decision makers access to quality information, enabling them to accurately identify where the company has been, where it is now, and where it needs to be in future. Despite the immense benefits that an effective BI system can bring, numerous studies showed that the usage and adoption of BI systems remain low, particularly among smaller institutions and companies with resource constraints [5].

According to [6], each BI system should have the following basic features:

- **Data Management**: including data extraction, cleaning, integration, as well as efficient storage and maintenance of large amounts of data;
- **Data Analysis**: including information queries, report generation, and data visualization functions;
- **Knowledge Discovery**: extracting useful information (knowledge) from the rapidly growing volumes of digital data in databases.

The most important feature to succeed in building BI solutions is to perform well on the stage of Data Management. Data Management is the foundation of any BI solution. It is usually the most stressing and time-consuming part. Nowadays, there are many companies offering their own solutions [7]. However, their applications do not assure that all necessary information in the decision-making process will be available. Rather than focusing on necessary information to build good solutions, most of these providers are focusing on the technological aspects. Such behavior is not satisfying real business needs, and not supporting the fact that there is not alignment between the business and technological domains.

2.2 Big Data

Big Data is entrenched term that is well understood by industry, academia and mass media. However, there are still debates about the exact meaning of this term. Historically, the first one who mentioned and used the term Big Data with its nowadays meaning were Weiss and Indurkhya in their publication [8]. Informally, Big Data is defined as the limitation of analytics and storage capabilities of standard data processing tools like database management systems. Nowadays, the majority of people involved into the process of working with Big Data understand it through its triple "v" concept: volume, velocity and variety. Volume states the fact of data processing limitations that are coming from huge size of data. Velocity argues that data input speed is also crucial, because data is generated and inserted into data storage on high speed. Variety states that data is coming from different heterogeneous sources (social networks, sensors, transactional data, etc.) [9].

Despite Big Data is kind of buzzword, the business cannot ignore it without losing competitiveness on the market. Datameer Inc. (2013) reported that the major goals for the companies to implement big data are [10]: increase revenue; decrease costs; and increase productivity.

Data and its knowledge extraction are too different things, but they cannot be separated. When data is stored, proper analytical methods must be applied in order to get value out of it. Mainly, there are two ways that are used to implement analytics over data: SQL and MapReduce [11]. SQL proved its applicability by the long and robust history of usage (more than 40 years). While MapReduce appeared less than a decade ago, it is already one of the most popular programming models to support complex analysis over huge volumes of structure and unstructured data. Multiple researches stated that SQL was not designed for current needs, and new models and ways, like MapReduce, should deal

with analytical challenges addressed in the era of Big Data. But, [12] and [13] showed that database management systems with SQL on board were significantly faster and required less code to implement information extraction and analytical tasks. However, the process of database tuning and data loading takes more time in comparison with MapReduce.

As was mentioned before, the major feature of Big Data are increasing revenue, decreasing costs, and increasing productivity. These three features are very desirable for any BI project. In the typical architecture of BI system, it is very common to have data warehouses with the whole information needed or even several data marts together to conform to the data warehouse, in this point.

The domain where the big data can be efficiently utilized is an optimization. In particular, game industry can use big data's triple "V" vision while understanding the background process of the ongoing game and optimizing data heavy process usually maintained in such companies. Game industry, especially during the last period saw a huge growth towards online game platforms. However, despite focusing more on online games, the goals of the game companies remains the same. They always tried to increase acquisition and retention of their customers (gamers) and improve monetization policies in order to generate more profit. In addition, one of the roads to follow to reach established goals is to improve satisfactory rate of their customers (gamers). Big Data can be handy for achieving such goals, for example, one of the possible scenarios is to bring together user profiles and game event logs to better understand users' behavior and interactional models during the game. Normally, the tremendous amounts of data generated by users' interactions are simply ignored, very rarely used to generate some information-based insights, or just stored "forever" to be processed "afterwards". Understanding such models can help to create better user experience for gamers and increase revenue for companies [10].

2.3 ETL vs. ELT

In a typical BI infrastructure, data, extracted from Operational Data Sources (ODS) are firstly transformed, then cleaned and loaded into a data warehouse. Before data are loaded into a data warehouse, it is necessary to process or perform a kind of "data wrangling" with the input raw data. For example, a data warehouse typically consolidates a multitude of different ODS with different schemas and metadata behind. Hence, incoming data must be normalized and brought into a common view, transformed if needed and then loaded. Also, the ODS may contain erroneous, corrupted or missed data, so the process of cleaning and reconsolidation are needed. This pre-processing is commonly known as Extract, Transform and Load (ETL): data are first extracted from the original data source, then transformed including normalization and cleansing and finally loaded into the data warehouse [14].

While database technologies used for data warehousing had seen tremendous performance and scalability enhancements over the past decade, ETL has not been improved in scalability and performance as database technology. As a result, most BI infrastructures are increasingly experiencing a bottleneck: data cannot be easily acquired to the data warehouse with necessary actuality. Clearly, in order to provide near real-time BI, this bottleneck needs to be resolved.

Costs of data storage were always a significant factor, but they are becoming cheaper with time, and as a result, analysis can be performed over bigger amounts of data with less investment. And in changing circumstances, former (but robust) Extract, Transform and Load approach cannot be easily applied to meet all business needs, which includes a strong desire to work with big data and, as a result, new approaches and/or architectural changes are needed. Main disadvantage of ETL is that data must be firstly transformed and only then loaded. It means that on transformation phase, mass amounts of potentially valuable data are thrown out. However, to eliminate drawbacks of ETL, latest improvement of storage techniques can be used. One of the approaches that addresses such challenges is called Extract, Load and Transform [15]. The basic idea is to perform the Load process immediately after the Extract process, and apply the Transformation only after getting the data stored.

ELT, in comparison with ETL, has these four advantages:

- The flexibility in adding new data sources (extract and load parts);
- Aggregation can be applied multiple times on same raw data (transform part);
- Transformation process can be re-adopted even on legacy data; and
- Speeding-up the process of implementation (usually, most of the time during the "data wrangling" is spent on the transformation).

According to [2] the competitiveness of the enterprise strongly rely on the time needed to perform decisions (will not be at least destructive for the business) and bring more added value. To make such decisions, BI solutions became "de-facto" standard. As long as time is considered a very important factor, it is crucial to design BI solutions in shorter period. One of the shortcomings, according to [4], is that the time spent on the BI solution's implementation phase causes increasing of costs and may lead to budget extension, replanting, and overall delay. The next shortcoming is that BI solution needs to be flexible in order to reflect environmental changes and adopt them in shorted possible period. Addressing flexibility of the BI solutions in the rapidly changing world can be treated as a sustainability factor for a smoother company development.

ETL process is not addressing flexibility in terms of reflecting environmental changes and therefore, classical BI solutions need vast amount of time to be implemented. The nature of ETL process is to perform *transform* operation immediately after *extract* operation and only then initiate *load* process. Such approach makes the data inserted into a data warehouse only during the last step. In contrast, ELT allows executing firstly the extraction and loading processes over data, and then applying the transformation on the consolidated data. *Transform* can be done "on demand" and multiple times if needed. Such nature is a very important goal for fast changing business models. Moreover, transformation with ELT can be applied and re-applied taking into account changes in business requirements. Based on the above reasons, it is more preferable to adopt Extract, Load and Transform (ELT) instead of Extract, Transform and Load (ETL) in BI solutions.

2.4 Service-Orientation

This section describes the service-orientation concept enabled by Web Service technology. Service-oriented Architecture can be referred to as a software architecture model that provides services to end-user applications, executable (business) processes, or to other services by means of published and discoverable service interfaces. The OASIS SOA Reference Model group defines SOA as follows: "Service Oriented Architecture (SOA) is a paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains. It provides a uniform means to offer, discover, interact with and use capabilities to produce desired effects consistent with measurable preconditions and expectations" [16].

Business functionalities in SOA can be realized and implemented in form of self-expressed and reusable building blocks called services. These services:

- Provide high level business concepts representation;
- Can be published and discovered in a distributed network; and
- Can be reused to build new (business) functions and applications.

Service can be defined as "the means by which the needs of a consumer are brought together with the capabilities of a provider" where the service provider represents "an entity (person or organization) that offers the use of capabilities by means of a service" [16].

Several SOA implementations can be found in academia and industry. One of these implementations extended the concept of service-orientation with concepts of lightweight semantic enablement using Resource Description

Framework (RDF) statements to group Web Services based on predefined criteria. This implementation is called Semantic-enabled Enterprise Service-Oriented Architecture (SESOA) [17]. More detailed information about SESOA can be found in the following research papers: [18], [19].

In this paper, we will enhance the ELTA concept with the service-orientation concept that is similar to the one applied in SESOA in which the semantic service repository assembles services based on their business area. However, to achieve that, the transformation services have to be split into two groups to be integrated within the service repository provided in SESOA. These groups will be explained more in details in section 3.3.

3. The Proposed Model

3.1 Introduction to ELTA Model

This paper defines ELTA term as follows:

(a) A process called *Extract* enables data extraction from heterogeneous sources in different formats (transactional data, machine-generated data, etc.);

(b) The *Load* process provides the ability to store data inside dedicated storage systems;

(c) The *Transform* process provides the ability to transform data from raw state, on demand and according to the needs of decision-making process;

(d) The *Analyze* phase enables business users to efficiently utilize the preprocessed data to understand enterprise behavior through implementing and trying different analysis methods and algorithms over already prepared data.

Based on the approach proposed in [20], a framework to define an Enterprise Architecture (EA) as a solution foundation is required. There are various available EA frameworks. Among them, the Zachman Framework [21] is selected as a core EA framework. Zachman EA's major idea is that the same EA can be viewed by different people involved into a project from different aspects. Importance of particular aspect can be very high for a particular person and not considered as a critical one by other person or group of persons. This is because there is a different responsibility of each particular member. Besides responsibilities, the viewpoint is also highly depended on the expertise of each member, and the framework goal is to consolidate all expertise in most optimal and understandable way. However, Zachman framework lacks in modelling for detailed EA components and relationships among them and does not provide concrete implementing method. It is valuable in the point that it presents general framework that every enterprise can use to build its own EA [22]. Besides that, "the Zachman Framework is an ontology - a theory of the existence of a structured set of essential components of an object for which explicit expression is necessary, and perhaps even mandatory for creating, operating, and changing the object (the object being an enterprise, a department, a value chain, a solution, a project, an airplane, a building, a product, a profession, or whatever)" [23]. We are considering only the first four rows of the framework, which are defined as follows: strategy model; business model; system model; and technology model.

In accordance with what previously expressed and for better understanding, the proposed model is depicted as component diagram and it contains different packages. This model can be mainly divided into three packages with out-of-package component called "Analyze Component". As shown Fig. 1, the model consists of following packages: "External Data Sources", "Enterprise Architecture" and "Big Data Processing". Each package consists of different components. The package "External Data Sources" is consolidating external sources of data needed for further processing. All possible data sources can be included in this package and the main purpose of the package is to group data sources on a logical level. The package "Enterprise Architecture" contains components related to Zachman's framework implementation and it also contains "Balanced Scorecard" (BSC) as a separate component, because the output artefacts of the BSC are directly used by the "Analyze Component". The package "Big Data Processing" consists out of three independent components: "Data Storage"; "Service Repository" and "Virtual Data Marts". Component "Data Storage" is acting like "a single point of truth" or mediating for all types of data used in one single BI solution.

This component is dedicated to store raw data, perform "extract" and "loading" processes and also interacts with the other two components of the package through means of transformation process. The goal of the "Virtual Data Marts" component is to perform transformation "on demand" and pass (publish) the transformed data to the "Analyze Component". The "Service Repository" is acting like an intermediate transformer between raw data storage and virtual data marts.

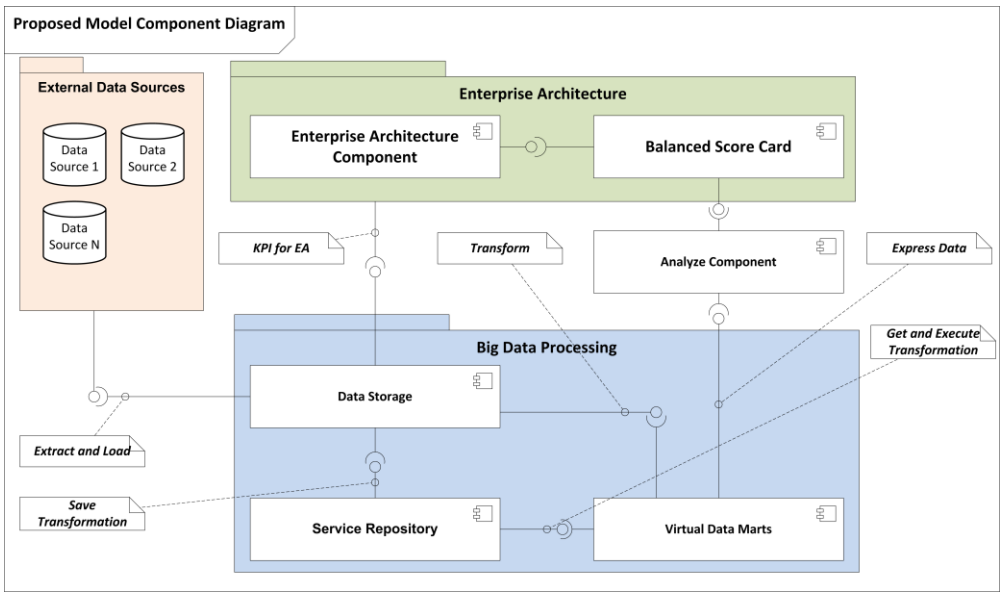

Fig. 1. Component Diagram Extract, Load, Transform and Analyze

3.2 Guideline

Guideline section consists out of seven steps: EA fulfillment; Extract and Load Processes; Management Control Tools; Transformation Process; Virtual Data Mart Layer; Develop BI System; and Analysis.

Step 1 EA fulfillment: According to the structure defined in [21], Zachman EA must be completed by rows, where each row represents a top level with respect to the one that follows in order. Nevertheless, there exists a big dependency among each of the elements of the columns. Table 1 shows the proposed dependencies between cells. The order in which cells must be fulfilled depends on the relationships and dependencies between them. The "*What*" column represents the modelling perspectives of EA. The "*How*" column represents the processes of EA. The "*Where*" column represents a location where the implementation should take place. The "*Who*" column addresses the organizational structure within the organization. The "*When*" column represents the timelines and time-related artifacts. The very first row named "Scope Contents" is more important for management rather that technical aspects. However, with each new row, the importance of the row for the technical level is increasing and the significance for the management level is decreasing.

Table 1. Proposed rules to fulfil Zachman EA

	What	*How*	*Where*	*Who*	*When*
Scope Contents	A1	B1	C1	D1	E1
Business Concepts	A2=(A1)	B2=(B1+A2)	C2=(C1+B2)	D2=(D1+B2+C2)	E2=(E1+A2+C2)
System Logic	A3=(A2+B2+C2)	B3=(B2+C2)	C3=(C2+A3+B3)	D3=(D2+C2+B3)	E3=(E2+B3+C3)
Technology Physics	A4=(A3)	B4=(B3+A4)	C4=(C3+A4+B4)	D4=(D3+A4+B4)	E4=(E3+D4)

Step 2 Extract and Load Processes: Based on the information defined in step 1, users can extract all necessary for business information from heterogeneous data sources and load it in data storage. The necessity of some particular piece of data or information is defined on the previous step of current guideline and it should be extracted from completed Zachman's EA. As long as data wrangling is not a trivial process for the business to be implemented, this step should be implemented by IT users.

Step 3 Management Control Tools: The main goal of this step is to define all necessary information for the decision-making process. The idea is to use data from data storage to create a new global indicator for the Balanced Scorecard perspective as described in [24]. This assures the reduction of the gap between strategic and tactical levels, because it is possible to know, how to link each indicators from different management levels and to improve the enterprise knowledge. Methodology as in [24] includes one step with Principal Component Analysis (PCA) [25] in order to discover the correlation among the whole indicators. This step should be performed by business users.

Step 4 Transformation Process: The main goal of this step is to properly transform all data based on the necessity of information for the decision-making process. Based on the data storage, which should be populated with data from external data sources through extracted and loaded processes and the indicators defined during step 3 of the current guideline, it is possible to know which transformations are necessary to support the entire business report requirement. This step should be implemented by IT users and stored in the service repository.

Table 2. Detailed description how to fill each cell of Zachman's EA

	What (A)	*How (B)*	*Where (C)*	*Who (D)*	*When (E)*
Scope Contents (1)	Create list of organizational entities related to the particular case	Create list of processes related to the particular case	Create list of geographical locations involved into particular case	Create list of organization units involved to the particular case	Create list of triggers and time loops involved to the particular case
Business Concepts (2)	Depict entities from the **A1** with Entity Relationship Model to demonstrate relationships	Model processes from the **B1** taking into account relationships from **A2**	Use locations from the **C1** and model them demonstrating **B2**	Create relationship model between roles from the **D1** taking into account results from **B2** and **C2**	Create events model using time elements from **E1** and taking into account **A2** and **C2**
System Logic (3)	Create data model diagram based on model from **A2** taking into account **B2** and **C2**	Describe processes verbally based on the **B2** and **C2** without referring to implementation	Describe locations verbally based on the **C2**, **A3** and **B3** without referring to implementation	Describe roles according to the types based on **D2**, **C2** and **B3** without referring to implementation	Describe events related to each other based on **E2**, **B3** and **C3** without referring to implementation
Technology Physics (4)	Specify on more detailed level data model diagram from **A3**	Describe processes using technology specific language based on **B3** and **A4**	Describe physical infrastructure components and their connectivity based on **C3**, **A4** and **B4**	Assigned roles and tasks on very detailed level based on **D3**, **A4** and **B4**	Describe events flows and states based on **E3** and **D4**

Step 5 Virtual Data Mart Layer: The main goal of this step is to define several virtual data marts in accordance with the business report requirements. In-memory approach [26] is used to accelerate creation and usage of data marts. Such solution is bringing more flexibility and unprecedented performance due to its in-memory nature. In this step, service repository is acting as a storage place of the previously designed and performed transformations. Using service repository brings an experience from older transformation activities to the newly created.

Step 6 Develop BI System: Based on the data marts, a structure is necessary to define the online analytical processing (OLAP) schema and business users' defined reports. There is a big variety of available tools for building BI solutions. One of the most popular solutions is Pentaho BI Suite [27]. Pentaho is popular due to its BI features and licensing policies. According to the authors experience, it is possible to achieve great flexibility in BI solution by combining Pentaho BI Suite with other tools like Birt Report [28].

Step 7 Analysis: The main goals of this step is to analyze most parts of the available information to support decision-making process and discover new patterns in the business by using data mining techniques, it will help in redefining the indicators in the Balanced Scorecard (in case it's necessary) and support the decision-making process. For this step, any external 3rd party tool like Weka [29], or integrated tool into the data storage component's analytical facilities, can be used.

3.3 SOA-enabled Transformations

The service repository component in the proposed model represents the component that is responsible of the management of transformation Web Service. This subsystem handles Web Service requests that are required to execute transformations coming from the processing unit component (where storing any kind of transformation or applying it during a process execution is needed). The processing unit is part of EA component in Fig. 1.

This component coordinates the storing of different types of transformation in form of Web Services. These transformation types are stored in the core database and the services that store these transformation types are published in the service repository. Furthermore, executing these transformation types is realized using another set of Web Services that are published in the service repository as well. The responses to the processing unit with the transformation services' availability and information are managed by service repository as well. This is done using the assemblage unit subcomponent that has an interface with the processing unit component to assist it in storing or executing transformations by responding with services' endpoints. It has another interface with the assemblage unit to ease the discovery and publication of storing and executing transformation services. Moreover, assemblages and Web Services information are stored in the core database via the DAO[1] interface between the assemblage unit and the core database. The internal architecture of this component is depicted in Fig. 2.

Fig. 3 provides an example on how to publish storing transformation services in the "Stored Transformations" assemblage.

As for executing the transformation services within the proposed model, the execution services are published in the "Transformation Execution" assemblage as shown in Fig. 4.

[1] DAO stands for a Data Access Object.

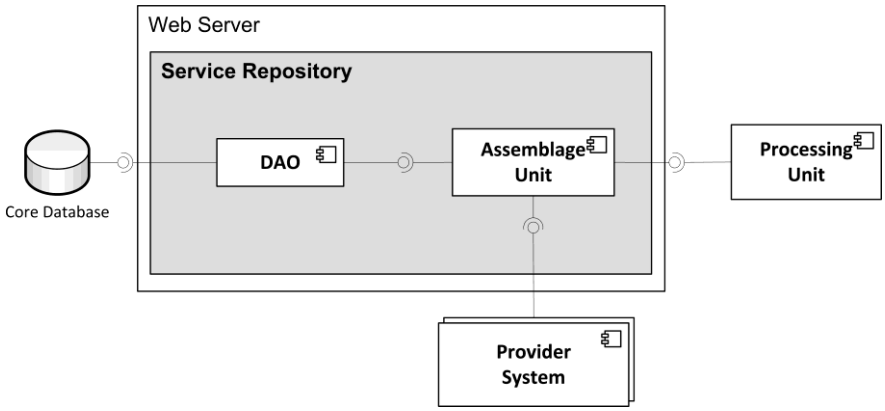

Fig. 2. Service Repository Component

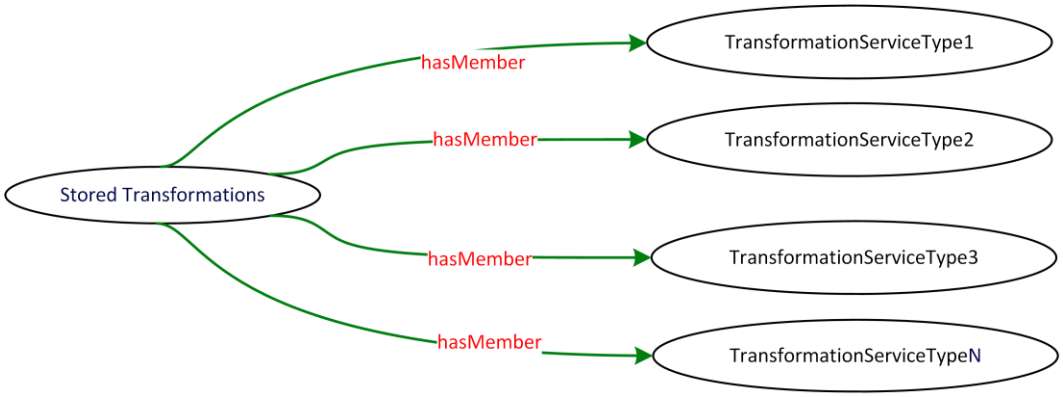

Fig. 3. Stored Transformation Services

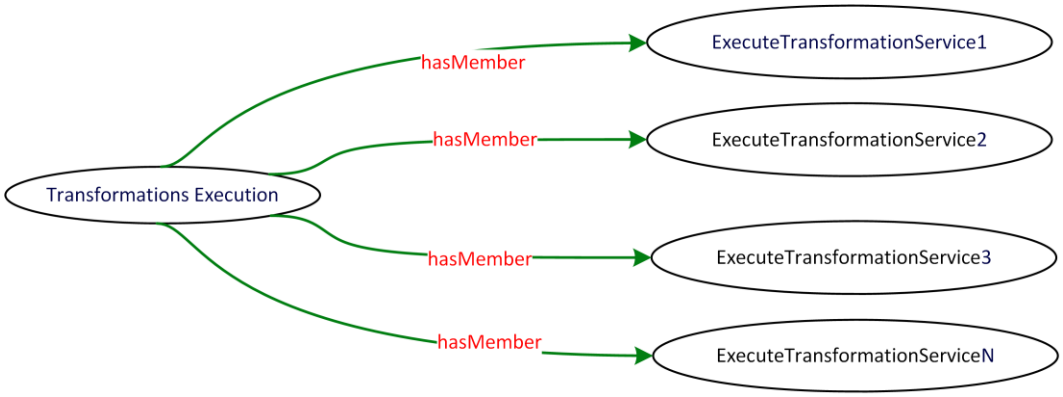

Fig. 4. Transformation Execution Services

In both transformation services, the graphs show that the relations between the transformation assemblages and services are represented using RDF[2] statements. In these statements, the transformation assemblages represent the subjects, the predicate is the "hasMember" relation, and the objects are one of the transformation storing or executing services.

4. A Banking Case Study

4.1 Description

For current case study, we are focusing on designing a new "data-centric" business service in the banking domain, and it is intentionally selected to be fully artificial and simple for understanding. The main parts that are covered by the current case study are the idea behind data-centric business product, the extraction and load processes besides transforming the big data storage's raw data and implementation of data marts after transformation. The rest parts like implementing BI system and performing complex analysis are left as "future work". While addressing the big data's 3Vs (**V**olume, **V**elocity and **V**ariety), it is important to select proper data sources. In the current case study, the following formats are used: relational, graph and log.

The description of the case study is as follows:

A bank came up with the idea to attract more clients and stimulate them to use more and more "services" offered by the bank. One way to attract new clients is by using already existing clients, however not in a way that the "other clients" will receive any kind of "spam" from bank. Rather to make it in a target-oriented style, by targeting just the proper clients. In order to meet the goal of targeting proper clients, data sources must be selected and huge amount of data needs to be analyzed. For this case, the following data sources are selected: user's social graph, user's transactional data and logs from the bank's Web server.

User social graph is used to understand new potential clients. *User transactions data* are used to understand the need of current clients and prepare good business offers. *Web server logs* are used to identify most active users and probably their interests in some products.

4.2 Application of the Proposed Model on the Case Study

Based on the guideline's *Step 1 EA fulfillment* (see Table 1 for general dependencies between different levels of the Zachman's EA and Table 2 for the detailed description of filling procedure), the next step is to perform extract and load operations.

Step 2 Extract and Load Processes is done with tools from Apache Hadoop software ecosystem [30]. In a particular case of the current case study, Apache Hadoop was selected as a staging layer for the raw data. Apache Hadoop is selected as it has its own cons and pros. The main reasons to select Hadoop as a central data-staging unit are:

1. the high scalability of its storage platform;
2. it is Fault-tolerance;
3. it is a de-facto standard for the big data world;
4. its cost effectiveness (open-source solution with active community; huge support from the major software vendors; utilization of commodity-hardware);
5. its diverse software components collections (very rich and dynamic software ecosystem) and
6. its ability to store and process variety number of formats.

[2] RDF is an official World Wide Web Consortium (W3C) Recommendation for Semantic Web data models and it stands for Resource Description Framework.

However, despite having many advantages, Apache Hadoop still remains a tool for batch processing and it needs additional workaround to be able to solve near-real time tasks. In the Apache Hadoop world, load operation can be performed in different ways. There are some standard loading routines and vendor-specific ones as well. However there are some tools which are used more often than others, for example Apache Sqoop [31] and Apache Flume [31]. Both tools have different application domains. The Apache Sqoop is designed to perform data load from structured databases. The Apache Flume was designed to mainly perform load of streaming and event based data. However in the current case study, a table and storage management layer for Apache Hadoop named Apache HCatalog [32] was used to perform data load process. Before loading data into big data storage, the extraction operation was performed. Due to the artificial nature of the case study, the extraction operation was replaced with data generation operation. Following sources of data (originally generated as csv files) were loaded into Apache Hadoop: (a) logs from Web server; (b) social graph; and (c) users' transactions.

Following the *Step 3 Management Control Tools*, meta-information about data staged in the Apache Hadoop was extracted and shared with the Enterprise Architecture level through an interface (see Fig.1) for further processing. As it was mentioned in the guideline, this step should be executed by the business user.

After having key performance indicators and understanding the necessary data needed for further processing, S*tep 4 Transformation Process* of proposed guideline advices to perform transformation. However, before executing transformation process it should be designed and stored inside repository. In particular case, we will have just tree transformation routines. Each routine transforms raw data into a structured form and prepare for further processing. For example, data that contains user's graph information are transformed into sparse-table format, data from Web server log are cleaned (groping repeated actions caused by "refresh" actions; removing information about image loading and expend fields with same IP address but different user activity), prune some of the details from transaction data.

S*tep 5 Virtual Data Mart Layer* suggests using an in-memory database to accelerate data accessibility. In current particular case study, SAP HANA is used as an in-memory database. It has multiple data process engines that meet the needs of online transaction processing (OLTP), OLAP, graph and text processing systems simultaneously [33]. As it is depicted in the diagram (see Fig. 1), the component "Virtual Data Marts" is accessing "Service Repository" component in order to get an access to the proper transformation routine. After getting from service repository the proper transformation operation and executing it, the result of the transformation is forwarded to the "Analyze Component" via a database table or a view inside SAP HANA. In current use case, the component "Virtual Data Marts" will create three tables with data from user's transactions, Web server logs and users' social graph. The structure of each table is dictated by the transformation itself. The data migration from Apache Hadoop during transformation is done with ODBC integration capability of SAP HANA. The last two steps in the guideline were not used for this use case.

5. Conclusion and Future Work

The situation with understanding benefits of using BI solution in the companies is far better than a decade ago. However, such understanding brought new challenges. For example, it is not enough for modern companies to implement a successful BI solution for making better decisions. It is also very important to make such implementations faster than the other ones in the BI market. Additional challenge to be considered in this context is the higher complexity of a particular process within companies. Nowadays, processes became tremendously complex, hard to maintain and not easy to support, extend and optimize. Such challenges should be faced and considered while offering a particular model, guideline or framework. In the upcoming era of big data, "data-centric" business services and processes of discovering new business strategies, based on historical behavior (mainly data), to achieve competitive edge over other competitors will play a huge role. Major factors succeeding in such completions is to prepare business IT solutions for the new "role-changing" requirements of the market. However, being successful cannot be achieved

only by fully applying new, modern and trending approaches and by neglecting robust and well-proved former techniques like BI. In such situation, it is more appropriate to modify or to prepare existing solutions for the market needs and benefit from the both robustness of well-proved existing techniques like BI and the promising advantages of the new approaches like big data. Our work presented ELTA (Extract, Load, Transform and Analyze) approach as one of such new approaches that can address the combination of business intelligence and big data by taking best parts from both, and in parallel, eliminating the disadvantages of business intelligence.
In future works, the proposed guideline will be enhanced with focus on the last two steps: *Step 6 Develop BI System* and *Step 7 Analysis*.

References

[1] S. Negash, "Business intelligence," *The Communications of the Association for Information Systems*, vol. 13, no. 1, p. 54, 2004.
[2] M. E. Porter, *Competitive advantage: Creating and sustaining superior performance*, 1st ed. New York City, United States: Simon and Schuster, 2008.
[3] M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem and P. Helland, "The end of an architectural era: (it's time for a complete rewrite)," in *Proceedings of the 33rd international conference on Very large data bases*, Vienna, Austria, 2007, pp. 1150–1160.
[4] H. Berthold, P. Rösch, S. Zöller, F. Wortmann, A. Carenini, S. Campbell, P. Bisson and F. Strohmaier, "An architecture for ad-hoc and collaborative business intelligence," in *Proceedings of the 2010 EDBT/ICDT Workshops*, Lausanne, Switzerland, 2010, pp. 13–18.
[5] G. M. Muriithi and J. E. Kotzé, "A conceptual framework for delivering cost effective business intelligence solutions as a service," in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, East London, South Africa, 2013, pp. 96–100.
[6] Y. Wang and Z. Liu, "Study on Port Business Intelligence System Combined with Business Performance Management," in *Proceedings of the 2009 Second International Conference on Future Information Technology and Management Engineering*, Washington, DC, USA, 2009, pp. 258–260.
[7] J. Hagerty, R. L. Sallam and J. Richardson, *Magic quadrant for business intelligence platforms*, 2012th ed. Stamford, Connecticut, United States: Gartner Inc., 2012.
[8] S. M. Weiss and N. Indurkhya, *Predictive data mining: a practical guide*, 1st ed. Burlington, Massachusetts, United States: Morgan Kaufmann Publishers Inc., 1998.
[9] D. Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety," *META Group Research Note, February*, vol. 6, 2001.
[10] S. Groschupf, F. Henze, V. Voss, E. Rosas, K. Krugler, R. Bodkin, J. Caserta and P. Shelley, *The Guide To Big Data Analytics*, 1st ed. Datameer Inc., 2013.
[11] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.
[12] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden and M. Stonebraker, "A comparison of approaches to large-scale data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, Providence, Rhode Island, USA, 2009, pp. 165–178.
[13] F. Chen and M. Hsu, "A performance comparison of parallel DBMSs and MapReduce on large-scale text analytics," in *Proceedings of the 16th International Conference on Extending Database Technology*, Genoa, Italy, 2013, pp. 613–624.
[14] F. Waas, R. Wrembel, T. Freudenreich, M. Thiele, C. Koncilia and P. Furtado, "On-Demand ELT Architecture for Right-Time BI: Extending the Vision," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 9, no. 2, pp. 21–38, 2013.
[15] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, "MAD skills: new analysis practices for big data," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1481–1492, 2009.

[16] C. M. MacKenzie, K. Laskey, F. McCabe, P. F. Brown, R. Metz and B. A. Hamilton, "Reference model for service oriented architecture 1.0," *OASIS Standard*, vol. 12, 2006.
[17] T. Mahmoud, "Lightweight Semantic-enabled Enterprise Service-Oriented Architecture," PhD. dissertation, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany, 2013.
[18] T. Mahmoud, J. Marx Gómez and T. von der Dovenmühle, "Functional Components Specification in the Semantic SOA-based Model," in *Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications*, S. Smolnik, F. Teuteberg and O. Thomas, Eds., 1st ed., Hershey, PA, United States: IGI Global, ch.14, pp. 277–291.
[19] T. Mahmoud, M. Petersen and D. Rummel, "Business Process Integration within Lightweight Semantic-Enabled Enterprise Service-Oriented Architecture," in *E-Procurement Management for Successful Electronic Government Systems*, P. O. de Pablos, J. M. Cueva Lovelle, J. E. Labra Gayo and R. Tennyson, Eds., E-Procurement Management for Successful Electronic Government Systems, ch. 12, pp. 181–192.
[20] P. Ortega, L. Ávila and J. Gómez, "Framework to Design a Business Intelligence Solution," in *ICT Innovations 2010*, M. Gusev and P. Mitrevski, Eds., vol. 83, Springer Berlin Heidelberg, pp. 348–357.
[21] J. Zachman. (2013, November 27) "The Zachman Framework: The Official Concise Definition" [Online]. Available: http://www.zachmaninternational.com/index.php/the-zachman-framework.
[22] D. Kang, J. Lee, S. Choi and K. Kim, "An ontology-based Enterprise Architecture," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1456–1464, 2010.
[23] L. A. Kappelman, *The SIM Guide to Enterprise Architecture*, 1st ed. Boca Raton, Florida, United States: CRC Press, 2009.
[24] M. Ortega, P. Michel, P. Pérez Lorences and J. Marx Gómez, "Compensatory Fuzzy Logic Uses in Business Indicators Design," in *Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support*, Mazatlan, Mexico, 2013, pp. 303–309.
[25] I. Jolliffe, *Principal component analysis*, 2nd ed. New York, United States: Springer-Verlag New York, 2002.
[26] H. Plattner and A. Zeier, *In-Memory Data Management: Technology and Applications*, 2nd ed. Springer-Verlag Berlin Heidelberg, 2012.
[27] Pentaho Corporation. (2013, December 1) "Pentaho - Business analytics and business intelligence leaders" [Online]. Available: http://www.pentaho.com/.
[28] The Eclipse Foundation. (2014, September 1) "Eclipse BIRT Project Home" [Online]. Available: http://www.eclipse.org/birt/documentation/.
[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
[30] The Apache Software Foundation. (2014, September 4) "Apache Hadoop" [Online]. Available: https://hadoop.apache.org/.
[31] The Apache Software Foundation. (2014, September 4) "Apache Sqoop" [Online]. Available: https://sqoop.apache.org/.
[32] The Apache Software Foundation and Hortonworks Inc. (2014, September 1) "Apache HCatalog" [Online]. Available: http://hortonworks.com/hadoop/hcatalog/.
[33] F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg and W. Lehner, "SAP HANA database: data management for modern business applications," *SIGMOD Rec.*, vol. 40, no. 4, pp. 45–51, 2012.

Biographical notes

Viktor Dmitriyev
M.Eng. & Tech. Viktor Dmitriyev had graduated from Kazakh-British Technical University (Kazakhstan) in 2009 with the diploma in the Computer Systems and Software Engineering. He got his degree Master of Information Systems in Engineering and Technology in 2010 from Kazakh-British Technical University. Viktor worked as a lecturer in the International IT University (Kazakhstan) for 4 years. During his work in International IT University he was a coach of the ACM ICPC team, authored couple of courses including "Fundamentals of Programming and Algorithms". Currently, he is a PhD student at the University of Oldenburg (Germany) and his primary research interests lie in the domains of big data analysis, in-memory computing and business intelligence.

www.shortbio.net/viktor.dmitriyev@uni-oldenburg.de

Tariq Mahmoud
Dr.-Eng. Tariq Mahmoud studied Information Engineering at Al-Baath University (Syria) and then completed his PhD in business information systems at Carl von Ossietzky University of Oldenburg (Germany). He is currently a research fellow at the working group of business information systems at the Carl von Ossietzky University of Oldenburg (Germany). His research interests include Semantic Enterprise SOA, Information Security, Business Intelligence, Data Warehousing and Semantic Web. His skills and experience include teaching and course building in addition to ability to deal with technology-enhanced methods of teaching.

www.shortbio.net/tariq.mahmoud@uni-oldenburg.de

Pablo Michel Marín-Ortega
M.Sc. Pablo Michel Marín-Ortega graduated from the University Central "Marta Abreu" de Las Villas, Villa Clara (Cuba) with master degree in business informatics at 2011. Currently, Pablo is a PhD Student at the University Central "Marta Abreu" de Las Villas, Villa Clara (Cuba) in a collaboration with the University of Oldenburg (Germany). His research topic is "Framework for Semantic Support to Business Intelligence Design Integrating Business and Technological Domains" and his primary research interests lie in the domains of business intelligence, business integration and semantics.

www.shortbio.net/pablomo@uclv.edu.cu