On-demand Data Integration for Decision-making Applications

Jānis Grabis and Jānis Kampars
Institute of Information Technology, Riga Technical University, Kalku 1, Riga, Latvia

Keywords: Data Integration, Data Services, Service Selection, Decision-making.

Abstract: Decision-making efficiency depends upon timely availability of appropriate data. On-demand data integration from web based data sources provides an attractive solution to data gathering. The two major challenges associate with on-demand data integration are selection of appropriate data services and efficient implementation of the data integration process. In this paper, it is proposed to use service's business value as a service selection criterion and to elaborate a lightweight method for XML based definition of the integration processes. The business value driven approach implies that services are selected according to their impact on quality of decisions made rather than solely according to their QoS characteristics. The data integration process definition methods partitions the data integration process into atomic data integration tasks, thus allowing for high level of data retrieval parallelization, accommodating data interdependencies and enabling error recoverability without delaying the whole data integration process. The data integration methods are evaluated using a case study, which investigates on-line decision making at a taxi company.

1 INTRODUCTION

Decision-making efficiency depends upon timely availability of appropriate data. Data warehousing is the most widely used solution for providing these data. However, this solution has limited flexibility and induces significant data processing and maintenance overhead (Zhu et al., 2004). Proliferation of web services including data services provides an alternative solution in the form of on-demand data integration (Delen and Demirkan, 2013). Two major challenges associated with the usage of the data services are: 1) selection of appropriate services; and 2) efficient implementation of the data integration process, especially, in the case of real-time decision making problems. The current work on the service selection focuses on using Quality of Service (QoS) data as selection criteria (Strunk, 2010). Among others Canfora et al., (2008) and Wang et al., (2007) use optimization techniques to maximize QoS characteristics of composite services. Jeong et al., (2009) and Tsesmetzis et al., (2008) also incorporate functional characteristics of candidate web services in their service selection models. In this paper, the service selection challenge is addressed by using the service business value as a selection criterion.

To implement the data integration process, enterprise service bus (Abrahiem, 2007) and Extract-Transform-Load (ETL) (Bhide et al., 2009) based approaches can be used though they are not primarily designed for data integration and on-demand processing purposes, respectively. ETL is not well suited for processing semi-structured data and data retrieval from remote heterogeneous web services (Yue and Wang, 2010). Wang et al., (2009) have proposed a dynamic data integration model based on SOA (Service-oriented architecture), and Frehner and Brändli (2006) use a virtual database to integrate distributed spatial data. Ali et al., (2009) propose a distributed extended xQuery technology for efficient data retrieval from heterogeneous web services. In this paper, a lightweight XML based data integration method is proposed with emphasis on specification of the data integration process to account for data interdependences and acceleration of the data integration process.

The objective of this paper is to elaborate a service's business value driven approach to on-demand data integration from remote data sources for real-time decision making purposes. The approach perceives data integration as a continuous process, where candidate services are monitored and evaluated, and the data integration solution can be updated if better services become available. The business value driven approach implies that services

Grabis J. and Kampars J..
On-demand Data Integration for Decision-making Applications.
DOI: 10.5220/0004445802010208
In *Proceedings of the 15th International Conference on Enterprise Information Systems* (ICEIS-2013), pages 201-208
ISBN: 978-989-8565-59-4

are selected according to their impact on quality of decisions made rather than solely according to their QoS characteristics. The on demand integration implies that data from external sources are retrieved for every decision making case. Methods for accelerating the data integration and improving data quality are also used. The approach proposed is evaluated using a case study, which investigated on-line decision-making at one of the leading Latvian taxi . A major attention is devoted to comparison of actual data accumulated by the company and data given by web services. The comparison allows to assess accuracy of decision-making and to adjust the decision-making results to reduce the impact of errors.

The main contributions of the paper are 1) the method for using business value as a service selection criterion; 2) the method for definition and execution of the data integration process; and 3) assessment for accuracy of mapping web services. The data integration process definition method partitions the data integration process into atomic data integration tasks, thus allowing for high level of data retrieval parallelization, accommodating data interdependencies and enabling error recoverability without delaying the whole data integration process.

The rest of the paper is organized as follows. Section 2 describes the on-demand data integration approach along with the methods used for service selection and specification of the integration process. An application of the approach is demonstrated in Section 3, and Section 4 concludes.

2 INTEGRATION APPROACH

The data integration approach proposed in the paper consists of design and execution cycles, and selection of appropriate services and specification of the data integration process are key methods of the approach.

2.1 Overview

The data integration objective is to gather the necessary data for real-time decision making. The data are gathered from distributed source, are not stored locally and are used immediately for the current decision-making case. The data integration is split in two phase, namely, the design phase and the execution phase (Figure 1). The design phase defines data integration problem and the data integration process. It also includes identification of appropriate data sources (i.e., different types of web services) and selects services the best suited for the decision-making problem. Data are actually retrieved from the data sources and integrated together during the execution phase. The data integration process is executed for every decision-making case. Methods for speeding-up data integration and for addressing data quality issues are used during the execution phase.

2.2 Service Selection

Identification and selection of appropriate services is of major importance for on-demand applications. In this paper, the services are selected according to their business value, i.e., rather than using evaluation criteria like QoS and similar, the services are selected according to their impact on quality of decisions made. This quality is measured by the cost of using services.

Each candidate service i is characterized by a set of attributes $\mathbf{x}_i = (x_{i1}, \dots, x_{iN})$, where $N = |\mathbf{x}_i|$. There is a cost associated with the jth attribute, and it is denotes by a_j. The total cost of using the ith service is expressed as

$$C_i = \sum_{j=1}^{N} a_j x_{ij}. \quad (1)$$

In order to select a service or services providing the best business value, they are selected to minimize the total cost of using all web services

$$Z = \sum_{i=1}^{M} C_i Y_i \rightarrow \min, \quad (2)$$

where Y_i is one if the service is selected and zero if service is not selected. Bonders et al. (2011) show that both functional and non-functional selection criteria can be expressed in terms of costs. For instance, the response time characteristic of web services can be expressed in term of costs as a cost of employees' time wasted to wait for the web service response.

In the case of additional constraints and requirements, the minimization problem can be solved using mathematical programming or other optimization methods. If other constraints are not considered, services can be selected by ranking.

2.3 Data Retrieval

The data are integrated from multiple heterogeneous data sources that are mostly controlled by third parties and whose interfaces are dynamically changing. The data retrieval process consists of multiple steps. The main data integration challenges are to minimize data integration time and to ensure high data quality. There are multiple interdependen-

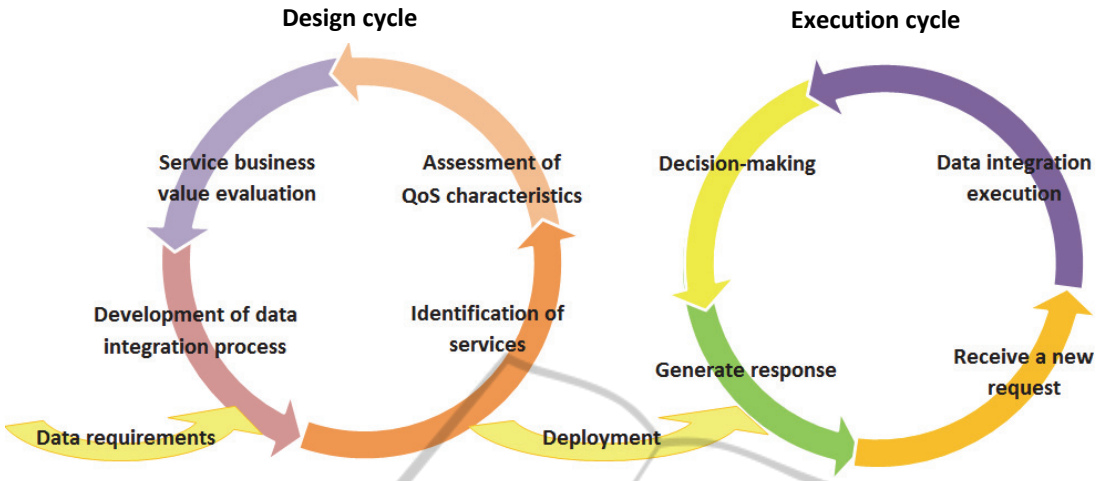

Figure 1: The design and execution cycles of the on-demand data integration approach.

cies among the steps of the data integration process, and the data retrieved should be transformed according to formatting requirements determined by decision-making needs. The data integration process starts with an XML document, which contains initial data for initialization of the data integration process and defines a required data structure. The final result is the XML document populated with all data required for decision making. The XML format is chosen as it has become a de-facto standard in electronic data interchange. Other semi structured data formats like JSON can be easily converted to XML if necessary.

In order to define the data integration process, three types of data integration tasks (DIT) are identified:

- Simple task (ST) – used to create loops for processing multiple XML nodes.
- Transformation task (TT) – used for XSLT based XML node transformation to provide conformance with abstract data retrieval operation input data models.
- Operational tasks (OT) – used to identify and query the corresponding data source based on the specified abstract data retrieval operation and input data model (execution of abstract data retrieval operation).

During the design phase, the process is composed using these three types of DIT and relationships between DIT are established (Figure 2). During the data integration process execution, DIT exchange with data integration requests (DIR). The exchange is controlled by a broker that monitors the status of the individual DIR and ensures conformance to DIR interdependencies defined in the data integration process. DIR contain a global XPath address of the element used as an input data for the recipient DIT. DIR statuses are defined as:

- Idle – DIR cannot be sent to the recipient due to a blocking dependency. Some of the data required for DIR execution are still unavailable.
- Created – the broker has received a DIR from a certain DIT and has created DIR for its child DIT.
- Sent – the broker has sent the DIR to the recipient DIT (only possible if there are no blocking dependencies).
- Processing – DIT has received the inbound DIR and it is being processed (e.g. a request to a data source is being made).
- Finished – DIT has finished processing of the inbound DIR and changes have been saved to the XML document. The outbound data integration task is created and sent to the broker.
- Error – there has been an error while executing DIR.

There are two types of the DIT relations (Figure 2):

- Succession – DIT A is successor of DIT B if it receives a DIR originating from DIT B.
- Dependency – DIT E depends on DIT D if it can process an inbound DIR only when DIT B has finished dependent DIR. In the example, DIR with relative XPath address /S/Q2 is idled until related DIR originating from the common parent of both DIT have been finished.

During the design phase data retrieval operations are defined as abstract data operations, and they are bound to actual physical data services in the performing the execution phase. In case of a QoS or Quality of Data (QoD) issues the abstract data operation is bound to an alternative physical data

service. If no alternative services exist, the status of the corresponding DIT is changed to "Error".

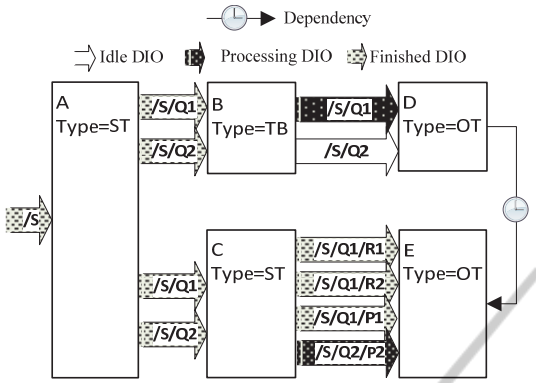

Figure 2: DIR execution.

3 CASE STUDY

A case study is used to evaluate the proposed on demand data integration approach. It focuses on service selection and definition of the data integration process.

3.1 Case Description

A taxi company's call center operations are investigated in the case study. The taxi company receives customer calls for taxi services. The decision-making problem is to find an available taxi and to quote time and cost of the service. To prepare the quote, the taxi company uses multiple services including a service for locating the current location of taxies and mapping services for geocoding and routing. Figure 3 shows the data retrieval and decision-making process at the taxi call center. As the decision-making result, a taxi to serve the customer request is determined and the customer is informed about waiting time for the taxi to arrive and about expected fare. It is assumed that a taxi driver individually decided on the actual route she chooses. The business value of the selected services is determined by time and distance the taxi driver travels to pick-up the customer at origin, and the cost of using the routing service is calculated as

$$C_i = a_1 x_{i1} + a_2 x_{i2}, \quad (3)$$

where a_1 =0.5 is the taxi travel cost by distance, $a_2 = 6$ is the taxi travel cost by time, x_{i1} is the travel distance estimate given by the ith service and x_{i2} is the travel time estimate given by the ith service.

As identified in several previous studies (e.g., Ehmke et al., 2012), mapping services overestimate or underestimate travel time and distance because of local traffic conditions and other factors. Given that the actual travel time and distance data are available to the taxi company, an error of the mapping services can be assessed, and an adjustment coefficient is introduced to account for the error for each attribute:

$$C_i^* = a_1 b_{i1} x_{i1} + a_2 b_{i2} x_{i2}. \quad (4)$$

The adjustment coefficient is specific to the particular service and it is calculated as

$$b_{ij} = R^{-1} \sum_{r=1}^{R} \frac{x_{0j}^r}{x_{ij}^r}, \quad (5)$$

where index 0 refers to the actual data accumulated by the taxi company and r refers to one particular taxi route for which actual travel data as well as those given by mapping services are available

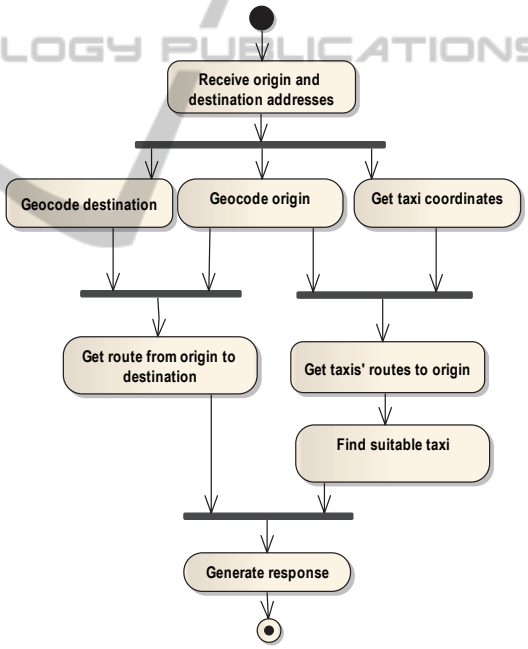

Figure 3: The business process of gathering information and quote preparation in response to the customer call.

Three types of web services are required to implement the decision-making process:

1. Geocoding service, which positions the customer request on the map;
2. Routing service, which determines the route between two geocoded locations;
3. Taxi tracking service, which identifies current location of taxis.

Multiple geocoding services are available. They

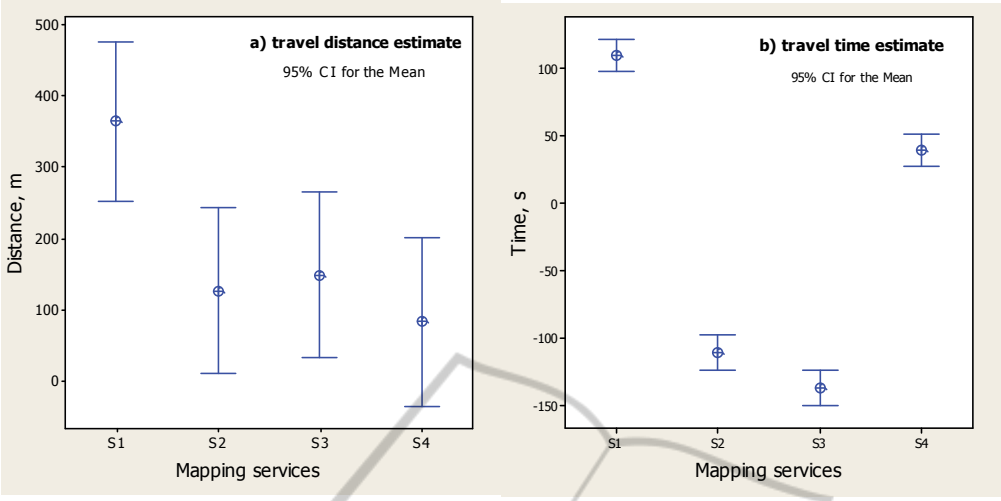

Figure 4: The interval plot of the difference between actual and estimated (a) travel distance and (b) travel time.

functionality differs by their ability to evaluate different types of customer requests, for instance, calls by address, point of interest or intersection. As a business value oriented criterion is readily available, it is decided to use all services simultaneously, and the final result is selected according to the confidence level returned by service provider. In cases when a geocoding service is unable to geocode the location, alternative services are queried. Only one taxi tracking service is available. The business values driven selection is performed for the routing service, and four different services are evaluated. These services are denoted as S_1, S_2, S_3 and S_4.

The QoS criteria are not used in the case because a related study by Bonders et al., (2011) has found that they have minor impact on service selection results for this kind of application.

3.2 Service Selection Results

In order to select appropriate services according to their business value, the taxi company's historic service data are used and compared with data given by the candidate services. The historic data contains a list of customer requests received and actual travel time and distance between different points of origin and destination. A thousand pairs of origin and destination are used. For all these pairs, the actual data are known, and time and distance given by four publicly available mapping web services are also determined.

Figure 4 shows an interval plot of differences between the actual values and the estimates given by the mapping services. It can be observed that more significant variations among services are for estimates of the travel time. Additionally, two of the services underestimate the travel time while two remaining services overestimate travel time. Values of the adjustment coefficient are reported in Table 1. These values show that the travel distance estimates are relatively more accurate than the travel time estimates.

Table 2 represents the service selection results. It compares the selection results if the cost of using the service is determined by (1) the travel distance only (i.e., $a_2 = 0$ in Eq. 4), (2) the unadjusted travel time and distance cost calculated using Eq. 4, and (3) the adjusted travel time and distance calculated using Eq. 5. The unadjusted criteria imply that S_1 is the most appropriate service. However, the adjusted criterion implies that S_2 is the best service to be used. Using the adjusted criterion is important, otherwise a service giving inaccurate underestimated travel distance and time values would be preferred over services giving more realistic estimates.

Table 1: Values of travel distance and time adjustment coefficients for each service.

	S_1	S_2	S_3	S_4
b_{i1}	1.14	1.09	1.11	1.09
b_{i2}	1.51	0.86	0.82	1.20

Table 2: The total cost of using candidate services according to the cost criterion used.

Cost criterion used	S_1	S_2	S_3	S_4
C_i, $a_2 = 0$	2597	2709	2698	2729
C_i	3176	3636	3667	3420
C_i^*	3836	**3747**	3792	3805

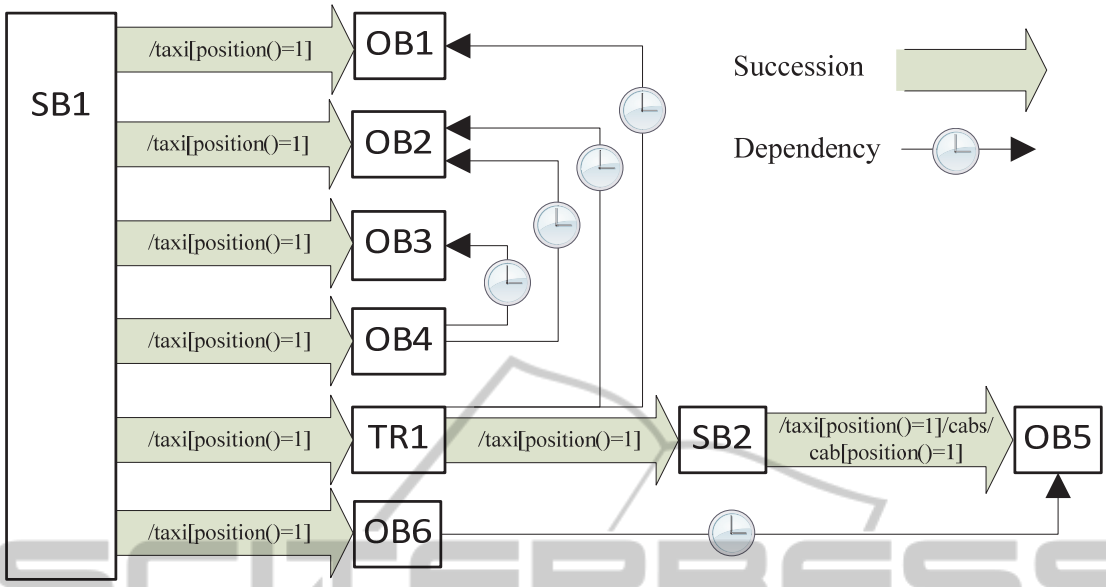

Figure 5: The data integration process for the taxi service case.

3.3 Process Definition Results

The data integration process is defined to gather the data necessary for decision-making (Figure 5). It starts at ST1, which reads the root node of the input data XML and sends it to all successor tasks. OB1 performs an abstract data integration operation in order to retrieve the list of available taxis, which is then saved to the XML document. OB2 and OB3 geocode client origin and destination, respectively. OB4 is responsible for calculating the route from the origin to the destination. OB4 depends on OB2 and OB3 as both pairs of coordinates are required for routing. TR1 transforms the XML document and adds client origin coordinates to each taxi XML node. The transformation can be performed only when the available taxi list is retrieved and origin is geocoded (dependency on OB1 and OB2). After the transformation SB2 creates a loop to process all taxi nodes and sends the corresponding DIR to OB5. OB5 calculates route from taxi location to the customer's origin for all of the available taxis. When routes are calculated the most suitable taxi is chosen in OB6.

The structure of input and output XML documents is shown in Figure 6. The input document contains the origin (street intersection) and destination (address) of the customer. After completion of the data integration process, the origin and destination nodes have been updated with corresponding coordinates retrieved by OB2 and OB3. A route between the origin and the destination has been calculated by OB4 and the results have been saved in two child nodes of the root element – distance and duration. A new taxis node has been created by OB1. Each of its child nodes has been updated with client origin coordinates (TR1), duration and distance of the route to client (OB5) and fitness (OB6). The name of the chosen taxi is saved in the chosenTaxi node.

Abstract data integration operations geocoding (OB2, OB3 in Fig. 5) and routing (OB4, OB5 in Figure 5) are mapped to publically available spatial data processing web services. None of the available web services is able to provide all of the required functionality (geocoding of address, point of interest, street intersection and routing). The approach used allows combining functionality from all services and adding new ones without altering the data integration logic. The decision of which geocoding service to use is made during late binding based on the input data model and performed abstract data integration operation. If multiple routing services have nearly equal business value, load balancing can be used to reduce data integration time. The availability of multiple alternatives also minimizes the risk of process blocking error in case of QoS or QoD issues. If the queried web service is not able to return the result due to data quality problems, a request to an alternative web service is performed. This way the data integration solution is able to provide better data quality than each individual web service.

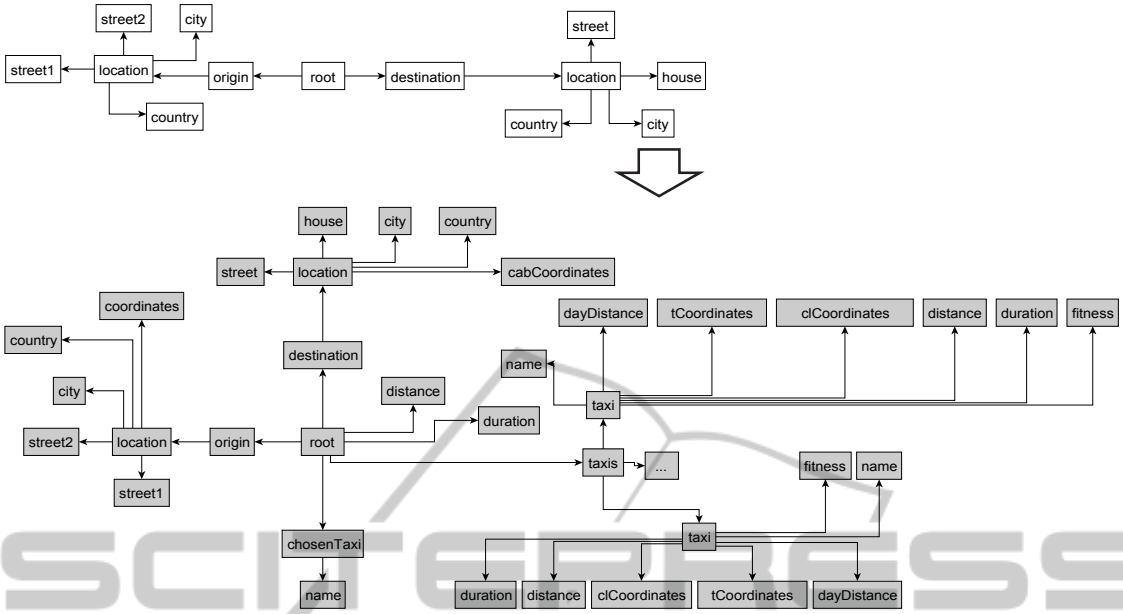

Figure 6: Input and output XML documents.

4 CONCLUSIONS

The business value driven on-demand data integration approach has been elaborated. This approach is suitable for decision-making applications, where decisions are made on-line and impact of the data quality provided by data services on decisions made can be readily quantified. In the sample case study presented in the paper, it can be readily quantified that services giving inaccurate or sub-optimal routing estimates would directly affect business performance (e.g., costs) of the service consumer.

The comparison of actual data and estimates given by the services allows to determine accuracy of travel time and distance quoted to customers. Although the taxi call center operators do not quote these errors o customers, these could be provided on company's web site or mobile apps.

In the case presented, selection of only type of services was considered and QoS characteristics were ignored (all services analyzed have very high level of the QoS chacteristics) though additional parameters and constraints can be accommodated by the proposed approach. It that case, optimization techniques like genetic algorithms should be used for service selection.

ACKNOWLEDGEMENTS

Dissemination of this research has been funded in part by the ERDF project „The development of international cooperation, projects and capacities in science and technology at Riga Technical University" Nr. 2DP/2.1.1.2.0/10/APIA/VIAA/003.

REFERENCES

Abrahiem, R., 2007. A new generation of middleware solutions for a near-real-time data warehousing architecture. In: *2007 IEEE International Conference on Electro/Information Technology,* 192-197.

Ali, M. I., Pichler, R., Truong, H. L., Dustdar, S., 2009. On using distributed extended xquery for web data sources as services. In: *9th International Conference on Web Engineering*, 497-500.

Bhide, M., Agarwal, M. K., Bar-Or, A., Padmanabhan, S., Mittapalli, S. K. , Venkatachaliah, G., 2009. XPEDIA: XML processing for data integration. In *Proceedings of the VLDB Endowment*, 2, 1330-1341.

Bonders, M., Grabis, J. , Kampars, J., 2011. Combining Functional and Nonfunctional Attributes for Cost Driven Web Service Selection. In *Frontiers in Artificial Intelligence and Applications,* 224, 227-239.

Canfora, G., Di Penta, M., Esposito, R. & Villani, M.L., 2008. A framework for QoS-aware binding and re-binding of composite web services. *Journal of Systems and Software*, 81, 1754-1769.

Delen, D., Demirkan, H., 2013. Data, information and

analytics as services. *Decision Support Systems*, In Press

Ehmke, J. F., Meisel, S., Mattfeld, D. C., 2012. Floating car based travel times for city logistics. *Transportation Research Part C: Emerging Technologies*, 21, 338-352.

Frehner, M. , Brändli, M., 2006. Virtual database: Spatial analysis in a Web-based data management system for distributed ecological data. *Environmental Modelling and Software*, 21, 1544-1554.

Jeong, B., Cho, H., Lee, C., 2009. On the functional quality of service (FQoS) to discover and compose interoperable web services. *Expert Systems with Applications*, 36, 5411-5418.

Strunk, A., 2010. QoS-aware service composition: A survey. In: *8th European Conference on Web Services, ECOWS*, 67-74.

Tsesmetzis, D., Roussaki, I. , Sykas, E., 2008. QoS-aware service evaluation and selection. *European Journal of Operational Research*, 191, 1101-1112.

Wang, J., Yu, A., Zhang, X., Qu, L., 2009. A dynamic data integration model based on SOA. In: *Second ISECS International Colloquium on Computing, Communication, Control, and Management,* 196-199.

Wang, H.C., Chang, S. L., Tsung, H. H., 2007. Combining subjective and objective QoS factors for personalized web service selection. *Expert Systems with Applications*, 32, 571–584.

Yang, K., Steele, R., 2008. A system for service-oriented data aggregation. *International Journal of Services and Standards*, 4, 119-140.

Yue, G, Wang, J., 2010. The design and implementation of XML semi-structured data extraction and loading into the data Warehouse, International Forum on Information Technology and Applications, 30-33.Zhu, F., Turner, M., Kotsiopoulos, I., Bennett, K., Russell, M., Budgen, D., Brereton, P., Keane, J., Layzell, P., Rigby, M., Xu, T., 2004. Dynamic data integration using web services. In: *IEEE International Conference on Web Services*, 262-269.