The Agent Approach to the ETL Task

Valeriy P. Khranilov, Pavel V. Misevich, Pavel S. Kulyasov and Elena N. Pankratova

November 7, 2024

The Agent Approach to the ETL Task

Khranilov V.P.
Computer technologies in designing and production dept.
Nizhny Novgorod State Technical University
Nizhny Novgorod, Russian Federation
hranilov@nntu.ru

Misevich P.V.
Computational systems and technologies dept.
Nizhny Novgorod State Technical University
Nizhny Novgorod, Russian Federation
p_misevich@mail.ru

Kulyasov P.S.
Computational systems and technologies dept.
Nizhny Novgorod State Technical University
Nizhny Novgorod, Russian Federation
p.kulyasov@gmail.com

Pankratova E.N.
Foreign languages dept.
Nizhny Novgorod State Technical University
Nizhny Novgorod, Russian Federation
keibusan@gmail.com

***Abstract*—The extract transformation load (ETL) process is an important element of the database (DB) and data warehouse designing. It is proposed to use a multi-agent approach to build software to support the ETL process in the paper. The requirements for agents are formulated too. A description of the typical software architecture of a multi-agent system for supporting the ETL process is in the article. The multi-agent system is in the example. Agents support the task of cleaning up the data. The data is English words in this example. There are Russian letters in these English words. These letters are visually indistinguishable from English characters. The multi-agent system looks for these letters in the text of the test example and replaces them with English characters.**

Keywords—data consolidation, data cleaning, knowledge base, intellectual support, multi-agent system, architecture of the management system, production system

I. Introduction

The multi-agent system approach to software design of the ETL (extract, transformation, load) task is discussed in this article. This approach has been created on the bases of the generalization of the author's experience in the joint stock companies: "Russian Railways" and "Mobile Telesystems". Practice shows that the software for the ETL task functions in constantly changes environment created by the external and internal factors. Thus, the software solving the task has to be orient towards modification.

The software design process has to be based on the concept of developing flexible software. The software has to include tools for supporting ETL process during the life cycle. These tools should adapt to changes in the external-internal environment with minimal design cost.

The paper proposes to use the multi-agent approach to develop software for the ETL task. The architecture of the system and the features of the application of a production knowledge base in the agent management system are considered too.

The proposed approach is useful for consolidating poorly structured data. They are following: Word and Excel tables, Web documents, information from text files and et cetera. Usually the task of supporting ETL process is decomposed into two interrelated tasks.

The first subtask is the task of transferring data from the source to the receiver (as a rule, this is a database or a data storage).

The second subtask is data analysis and cleaning. Usually the data contain a set of errors and skips. They are the following: interactive user errors, the length of data out of the field beyond acceptable limits, two or more atomic data in one field et cetera.

Let us note that the first task is solved by a system of converter programs, which are included in the DBMS. The second task requires developing a special tool – software for supporting ETL processes. The tools analyze the data in each column of the converted table (if we work with relational databases). The aim is to identify errors and correct them (in an automatic or an interactive mode).

The article provides an example. The multi-agent software supports the ETL process. It identifies and corrects incorrect characters in the text.

II. Methodology

The ETL software has a set of tools (agents) at the conceptual level. The elements of the tools connect the ETL software to different sources, extract data from the sources and clean the data using different algorithms, transform and upload data to the database or data warehouse (target).

Let us formulate a set of requirements for the support system of ETL processes [1]. They are following:

- the system has to be modified and expanded during the life cycle;
- the system has to operate with different data formats;
- the system has to be connected to different sources;
- the system has to support various information processing algorithms for data cleaning.

It is proposed to consider the system for supporting ETL process as a multi-agent system [2]. This system consists of a set of agents and the agent management system.

Each agent-event implements a step of the ETL algorithm in the IT subject area. They are following:

- analyzing the status of the data source and connecting to it;
- identification of the fact of non-compliance of some data with restrictions;
- data correction (cleaning).

In addition, the agent has to support a set of operation. They are the following:

- logging its actions;
- interaction with the agent control system;
- interaction with other agents (in the future).

Agents solve a set of typical tasks to support the data cleaning process. They are following:

- matching the column fields to the required data type;
- matching the column fields to the maximum allowed length;
- matching the field value to the set of acceptable values;
- matching of the separator between integer and fractional parts of a number to the standard of the target;
- checking the position of the column separator positioning between the columns;
- checking the facts of conflicts between the column separator symbol and the contents of the source data fields;
- identification of fields without data and marking them in the database using a special symbol in the DB;
- identification of the facts of erasing text in the fields by replacing the text with a space character;
- searching Cyrillic characters in an English text;
- searching English characters in an Russian text;
- etc.

Each typical task is solved by the appropriate agent. For example, the algorithm for the task of looking for Cyrillic characters in English text consists of a sequence of steps. They are following:

- selecting the value of a text field and writing it into a text variable;
- decomposing this variable into characters and writing the characters to an array;
- selecting characters with identical visualization in the Russian and English alphabets (table 1);
- replacing Russian characters in the text with English characters;
- concatenation of character array elements into a text variable;
- generation of the corrected value of the text field which is checked;
- supporting logging (maintaining a Log table or file).

Note, that this typical task is very common. It happens when it is necessary to convert data from Excel files. The data usually contain a set of interactive errors made by users. Frequent reproduction of this task makes it relevant to develop a specialized agent for solving of the problem. This agent is included in the multi-agent system for supporting ETL process.

The general rule of IT technology says that if dirty data are input to software, then the output data are dirty too. If dirty data are uploaded to the database from an external source, the query will run incorrectly.

TABLE I. THE SYMBOLS OF THE RUSSIAN AND ENGLISH ALPHABETS WHICH HAVE THE SAME IMAGE

No.	English Characters	Russian Characters	Comment
1	A a	А а	
2	B b	В в	The uppercase letters visually are equal
3	C c	С с	
4	E e	Е е	
5	H h	Н н	
6	K k	К к	
7	M m	М м	The uppercase letters visually are equal
8	P p	Рр	
9	T t	Т т	The uppercase letters visually are equal
10	X x	Х х	
11	Y y	У у	The lowercase letters visually are equal

This example shows that the agents of the ETL support system have to work according to a specific algorithm. The agent has a program code which has many columns. The agents are the theoretical primitives in the concept “software ETL as a multi-agent system (MAS)”. The approach allows developing a system for supporting process of data consolidation.

Let’s adapt the technology of building multi-agent systems to the ETL subject field [3-7].

Experience shows, that if MAS does not consider the task of semantic data analysis, then the scenarios for verifying the correctness of data have a “linear continuation” or a “simple cycle” type structure. This scenario structure is allowed to use the “constructor” model to generate scenarios to support the ETL process.

Let us imagine the events in the form of vertices of a graphical object. The cause-and-effect relationships of the scenario event in the graphical object are represented by an edge coming from the vertex. Each edge has an identifier. The identifiers form a set $\{r_1, r_2, \ldots\}$. Each edge has a terminal. The terminal in the figure is represented by an arrow.

The elements of the scenario-generating set of events are shown in fig. 1. The development of ETL scenarios is carried out by filling out the terminals of the edges. Event enters into terminals. Thus, the structure of events is covered by cause-and-effect relationships.

The scenarios which are constructed from the scenario-generating set of events (fig. 1) are shown in fig. 2 and fig. 3.

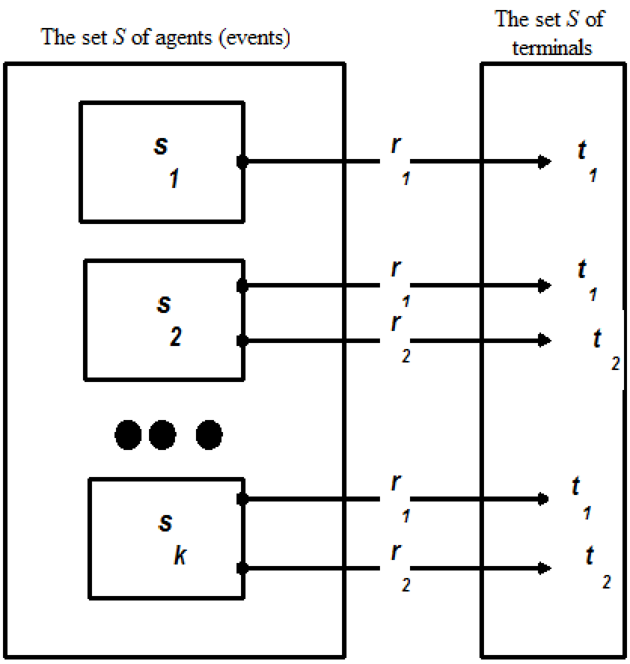

Fig. 1. The set of events (agents).

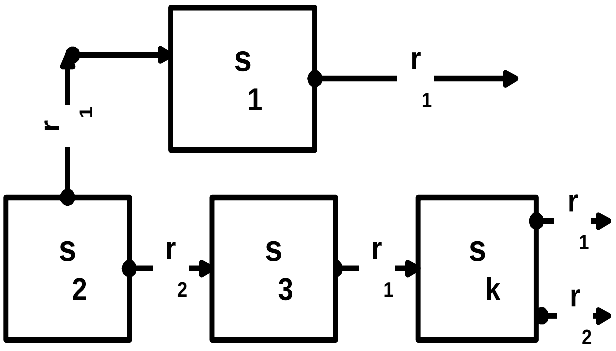

Fig. 2. The agent operation scenario.

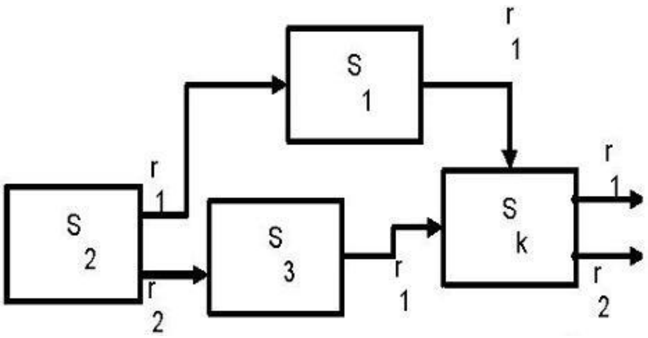

Fig. 3. The agent operation scenario 2.

In the case of using branched scenario the algorithms for checking and correcting the structure of events should be used.

III. The Software Architecture for the Support of ETL Process

Let us develop the software architecture of the ETL system. Fig. 4 shows the architecture of this agent-based system. The system contains a set of subsystems. They are following:

- an agent library;
- the agent management system;
- a logging subsystem;
- a scenario library.

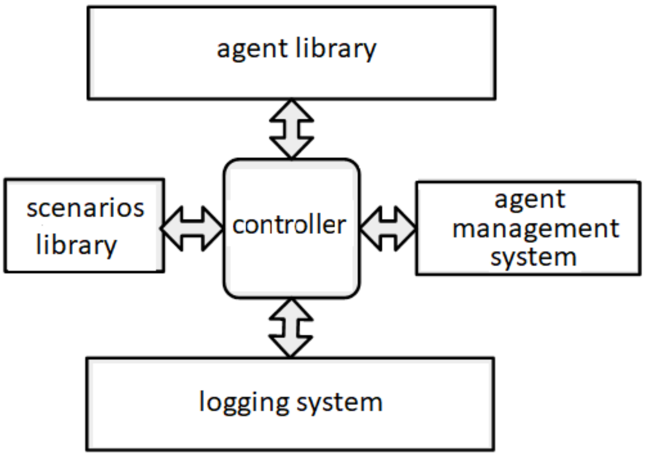

Fig. 4. The architecture of the agent system.

The agent library is a specialized agent repository. It includes new agents. The agents have a unified interface. It allows the agents to connect to the agent management system and the logging system.

The agent management system is a key subsystem in the architecture. It operates a work scenario. This scenario is described by a system of product rules. The ETL software contains an interpreter. It analyzes the situation and initiates the necessary agent to support the ETL process.

The script library contains a variety of algorithms for extracting, cleaning data and transforming formats from various sources. The script library is a specialized repository for supporting the agent management system during its lifecycle. This architecture element facilitates the process of creating and debugging scenarios for a multi-agent system.

The work logging system generates a protocol that describes which agent found errors in which field and how they were corrected. The agent is used for generating reports to analyze errors and debugging the system.

The controller in this architecture is a control system that coordinates the interaction among the subsystems described above.

Note, that the agent management system can use various technologies. The paper the use of the interpreter of the production rules. It is a management system in the MAS. The agent management system has the following architecture (fig. 5).

The interpreter analyzes data base (facts) and selects a production rule from the knowledge base to determine which event needs to be initiated in the next discrete point of time.

The knowledge base stores a description of the scenario of the production system. The scenario is used for supporting the ETL process.

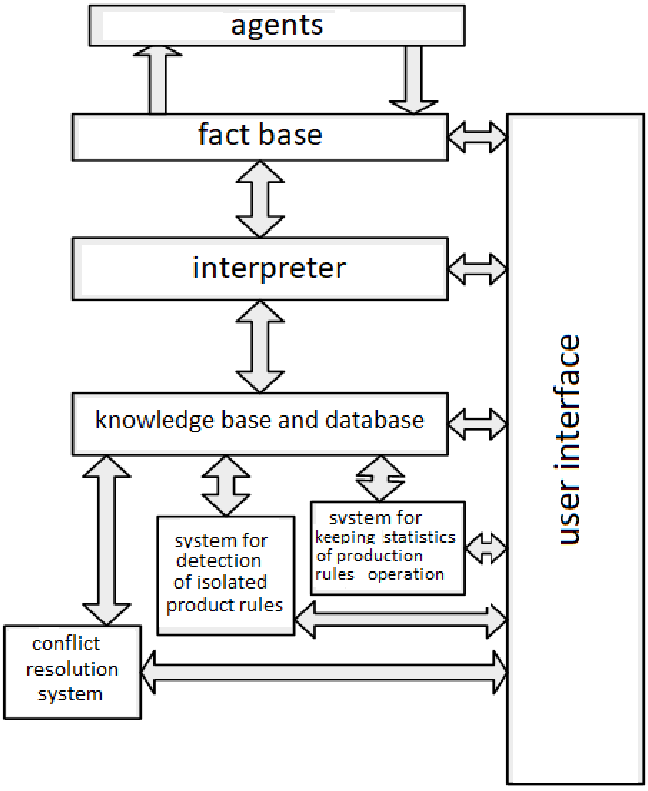
agents
fact base
interpreter
knowledge base and database
system for detection of isolated product rules
system for keeping statistics of production rules operation
conflict resolution system
user interface

Fig. 5. The agent management system architecture.

The conflict resolution system is designed to prevent the conflict sets of rules in the production knowledge base. The conflict is a situation which occurs when the knowledge base has some conflict rules. Such rules have equal left parts and different right parts.

The system for collecting statistical data is designed to accumulate information about operations of the agent management system. The purpose of collecting statistics is to identify the bottlenecks in the project.

The user interface is designed to organize an access to all system resources for IT engineers. The interface allows users to develop a script for the agent system and debug the scenario in the dialog mode.

Note, that the architecture of the multi-agent system is isomorphic to the typical architecture of an expert system. The fact orients the software to modification and expansion during its life cycle.

IV. Conclusion

The use of a multi-agent approach to create software to support the ETL process has shown same benefits. They are the following:

- reducing the cost of the developing software;
- low cost of software support during its life cycle;
- process support software has a simple and intuitive architecture;
- the agents may be used in various scenarios of the ETL process;
- the software is oriented on modification and expansion;
- the software design process can be parallelized between programmers.

The disadvantages of this approach include are the following:

- the complexity of developing the set of independent agents;
- using special tools to support the process of storing agents in a library or repository.

The operation of the multi-agent ETL process support system is based on a situational approach [8] and models [9-11].

References

[1] M.M. Singh, “Extraction transformation and loading (ETL) of data using ETL tools,” International journal for research in applied science and engineering technology, vol. 10, No. 6, pp. 4415-4420, June 2022.

[2] L.A. Gladkov and N.V. Gladkova, “Evolutionary design as a tool for developing multi-agent systems,” Proccedings of the Southern Federal University, No. 4(221), pp. 51-61, 2021.

[3] P.V. Misevich and A.E. Ermilov, “Automatization of build and support situational type monitoring systems based on skeletal code of system,” Management systems and information technologies, No. 3(69), pp. 62-65, 2017.

[4] A. E. Ermilov, “Monitoring systems development using situational approach and fuzzy logic,” IOP Conf. Ser.: Mater. Sci. Eng., 695 012016, 2019.

[5] P. V. Misevich, “The use of a logistics approach to the issues of constructing procedures for identifying and overcoming emergency situations in automated systems,” Management Systems and Information Technologies, No. 4.2(26), 2006.

[6] V.P. Khranilov, “Models in a state space for engineering applications,” XXI international conference complex systems: control and modeling problems, 2019.

[7] P.V. Misevich, “Dynamic model of functioning of an automated system,” Management Systems and Information Technologies, No. 3.1(33), 2008.

[8] M. Minskiy, “A framework for representing knowledge,” MIT AI Laboratory Memo 306, June 1974.

[9] V. P. Khranilov, P. V. Misevich, E. N. Pankratova, and A. E. Ermilov, “Models for supporting the operating scenarios during a life cycle in automated systems,” 14TH International Symposium «Intelligent Systems», INTELS’20, 2020.

[10] P. D. Basalin, K. V. Bezruk, and M. V. Radaeva, Models and methods of intellectual support of decision-making processes, Nizhniy Novgorod: Nizhniy Novgorod State University, 2011.

[11] V. A. Lazarev and P. V. Misevich, “Models of information support of systems for the maintenance of complexes of automated software testing,” Management Systems and Information Technologies, No 4(65), 2016.