

A Semantic Approach to Data Integration

Posted on September 13, 2018

No matter how perfect the relational and functional approaches to data consolidation are, in fact, they are primitive and rely on the traditional engineering worldview, which has not changed since the late 60s and separates data from content. A semantic approach to integration will allow more consciousness to work with data, which will greatly simplify the search and redistribution between applications and integration.

Free Cost Calculation

For a long time, the discussion of the topic of knowledge management (Knowledge Management, KM) was, for the most part, abstracted from practice, but in recent years, some real signs of KM can be found in the concepts of Web 2.0 and Enterprise 2 – the computer really becomes an intellectual assistant. The complete solution to this problem is still far away, but for now, the more practical and pressing is the task of assembling together information from disparate sources, creating what is called a single source of truth – a single source of truth, data integration methods.

There are two alternative approaches to data integration: old-syntactic and new-semantic. The first is based on the external similarity of the combined data, and the second is on the content. For example, if we, following the first, combine two ordinary tables, then we assume that in the field “Temperature,” the values are expressed on one scale, and if we could somehow organize the semantic storage, then we could count on the fact that in the “Temperature” field there are data and metadata, that is, a record of the physical quantity and an indication of what scale it is measured on, and could combine the arrays of data recorded both in Celsius and in Fahrenheit.

Semantic integration is based on knowledge and consideration of the nature of the data. Of course, storing data together with metadata creates additional difficulties, but it provides more convenience. The understanding of this truth came quite recently – it was hampered by the fact that many years ago, the creators of the first computers found a successful solution, they thought, separating the data from the context and giving the opportunity to work with the data in its pure form. As a result, there was a kind of computer “split consciousness”: the data – in the car and the context – in the person. While there is little data, it is not worth thinking about, but for an indefinitely long time, such a state could not be preserved, and today, it is becoming more difficult to keep in mind what’s what for. From this “split,” the most affected business analysts need to see not individual facts but the picture as a whole, so they were the first to formulate their need for information integration. So, in IT, a direction called enterprise information integration emerged (enterprise information integration, EII).

Integration Problems

Let's start with a broader concept than EII, which is the integration of the enterprise. This type of integration has two interpretations depending on the subject of integration: the enterprise itself with its subcontractors and partners or the integration process being limited to the walls of the enterprise. In the first case, integration means the creation of electronic communications between all participants in the process from production to marketing, that is, between producers, assemblers, suppliers, and consumers; these communications ensure interaction between partners along the whole chain of progress, from raw materials to commodities. The second interpretation limits the scope of enterprise integration to the boundaries of the enterprise, which is the means for uniting people, processes, systems, and technologies. Briefly, the task of the second type of integration can be defined as follows: the right people and the necessary processes should receive the necessary information in the right place and at the right time.

Of the two interpretations of integration, the first reflects the global trend towards globalization, information aspects of which are considered, for example, by Thomas Friedman in the book "Flat World: A Brief History of the Twenty-First Century," a bestselling book several years ago. The author believes that the current information period of globalization is the third one; it was preceded by the colonial period, which began with the discovery of America by Columbus, and capitalism, which stretched from 1800 to 2000. During the first period, the main actor was the state; during the second, -it was the corporation, and no, -it was individuals and relatively small teams. The specific features of each are due, inter alia, to the means of communication that existed at that time. Using modern computer and communication technologies, people can enter into a partnership or competitive relations regardless of geographical and political influences; the whole world turns into a single production space, the actions that are subject to general rules. This is the essence of the concept of a "flat world," flattened by computer and communication technologies. Integration of enterprises is one of the most important features of globalization.

The second interpretation of enterprise integration refers to two complementary technology groups: using EII, you can combine data sources, and by integrating enterprise applications (Enterprise Application Integration, EAI), processes are combined. These solutions can be called two subsystems of one organism, say, the circulatory and lymphatic. Technologies EII is focused on reporting and analytical functions, and EAI is focused on the automation of business processes. Strange, but in many works, they are put in one row and even compared, which is fundamentally wrong - EII and EAI are not alternatives; they complement each other, forming a single whole.

The term Enterprise Information Integration, proposed by the analysts of the Aberdeen Group in May 2002, does not carry any new content but was simply convenient for use as a marketing umbrella brand. He replaced the previously existing acronym, obtained from Enterprise Information Interoperability. Seven years have passed since then, but even today, EII is quite vaguely defined as technologies that provide uniform access to heterogeneous data sources without first being uploaded

to the repositories in order to represent a multitude of sources in the form of a single homogeneous system. Despite the vagueness of this definition, EII is actively commercializing, and it is already possible to talk about the emergence of an appropriate integration industry.

To clarify the meaning, put today in EII, and we will make a small digression. The appearance of EII was preceded by more than thirty years of existence of a similar sense and sound discipline – data integration (Data Integration). The term itself is defined as the combination of data from different sources in order to create a holistic view of the so-called “one image of truth.” Data integration is usually used when merging several bases in a business combination, and the means for creating a data composition are one of the most important tools. With this definition, there is nothing to argue about, but one can not agree with yet another statement: “In business management, data integration is called Enterprise Information Integration.” Thus, the identity of the two domains is postulated, and it turns out that the integration of data and the integration of corporate information are one and the same thing, which is incorrect. Let’s try to understand the differences between data integration and information integration, considering that the comparison of the two groups of technologies is also of interest because, in this context, the differences between the objects of integration are more clearly seen.

Classification And Methods Of Syntactic Integration Of Data

First of all, we will clarify the use of the term “integration” and its difference from “consolidation” in the appendix to the data. In the management of everything, the economy, there has always been and remains the need to consolidate data coming from diversified sources – the ministries and departments have been carrying out this work for centuries, accumulating various kinds of archives. In the computer age, it became possible to create automated data warehouses. Data consolidation procedures form the basis of the data warehousing guideline known as ETL (Extract-Transform-Load). Dictionaries treat consolidation as the union of separate differentiated parts into a whole. While working with data, different methods are required, but the use of simpler methods of consolidation is limited by the level of complexity of the data. Usually, consolidation is sufficient when working with data is limited to compiling reports, and there is no need to apply serious analytical methods. Actually, the development of various analytical methods has become the main reason for the increased attention to data integration. Integration is different in that it presents the user with a uniform view of heterogeneous data sources, which implies a general model and a general relation to semantics in order to provide an opportunity to access data and, if necessary, resolve conflict situations.

The task of data integration technologies is to overcome the many manifestations of heterogeneity inherent in information systems that were created and created, guided by nothing but not a unified relation to data. Systems have different functionalities, use different types of data (alphanumeric and

media, structured and unstructured), their components differ in autonomy, and have different performances. The systems are built on different hardware platforms and have different data management tools, different intermediate layer software, data models, user interfaces, and much more.

Ten years ago, a researcher from the University of Zurich, Klaus Dittrich, proposed a scheme for classifying data integration technologies from six levels (Figure 1).

Common Data Storage. It is implemented by merging data from different storage systems into one common. Today, we would combine these two levels into one and name it virtualization of storage systems.

Uniform Data Access (Unified Data Access). At this level, the logical integration of data is carried out; different applications receive a uniform vision of physically distributed data. Such virtualization of data has its undoubted merits, but homogenization of data in the process of working with them requires considerable resources.

Integration by Middleware. The software of this layer plays an intermediary role; its components are capable of performing certain functions assigned to them, and the integration task is fully solved in interaction with the applications.

Integration by Applications. Provides access to various data sources and returns generalized results to the user. The complexity of integration at this level is due to the wide variety of interfaces and data formats.

Common User Interface. Allows uniform access to data, for example, using a browser, but the data remains unintegrated and heterogeneous.

Manual Integration. The user himself combines the data, applying different types of interfaces and query languages.

The Dittrich scheme is interesting and convenient because it allows you to connect the integration of data with the integration of information – as you move upwards, simple atomic data acquires semantics, becomes accessible to human understanding, and turns into useful information presented in a convenient form. We will correlate this classification with the known technologies of data integration.

Data integration at the common data level can be implemented using repositories in which data coming from different OLTP systems is extracted, transformed, and loaded (ETL) into the data store and then used in the data warehouse. Analytics. From the point of view of data integration, a close example of integration at the lower level is the Operational Data Store (ODS), which is created to store “atomic” data for a relatively short time for real-time operation. They are smaller than the classical storage facilities designed for long-term storage and are sometimes called “warehouses with

fresh data.”

An example of a solution at the level of unified data access can be federated databases that unify, as the name implies, separate databases using the metadata device. The bases included in the federation can be geographically distributed and connected by computer networks. Sometimes, such DBMSs are called virtual because they provide access to various databases; for this purpose, the federated database management system decomposes the request for requests sent to it into the databases belonging to the federation.

An example of data integration using middleware is the Oracle Data Integration Suite, released in 2008 from Oracle Fusion Middleware. The package includes the necessary components for moving data, assessing their quality and preliminary analysis (profiling), as well as tools for working with metadata. Oracle Data Integration provided the ability to create solutions tied to the service architecture and focused on business intelligence applications.

Applications intended for data integration have been released by several companies; the most common is Altova’s MapForce product. It allows you to combine data presented in various formats, including XML, various databases, text files, spreadsheets, etc. MapForce can be considered as middleware for distributed applications, both within the enterprise and in cloud infrastructures. To interact with the user, MapForce has a convenient graphical interface that allows you to select the necessary functional filters from the library and collect the process of data conversion and integration. At the same level of the model are workflow management systems (WFMS), whose integration function is that they link individual work performed by users or applications into a single stream. Typically, WFMS systems support the modeling, execution, and management of complex processes of human interaction with applications.

A typical example of the implementation of a common user interface is a different kind of portal, for example, iGoogle or My Yahoo, “catching” data from different sources but not ensuring their uniform representation.

Integration by hand is in the pure form of the user’s work with information; this approach makes sense when no automated methods of integration are possible.

The above classification shows that there can not be any real integration of information of speech for now: everything that we see today is reduced only to the consolidation of data at one or another level of abstraction. In this case, we can dwell on the following definition: “Integration of corporate information is usually called consolidation at the corporate level of data from different sources. EII enables people (administrators, developers, end users) and applications to interpret a variety of data as data from one database or delivered by a single service. ”

Integration In The Complex

The market for data integration technologies is presented today by three groups of companies. Major manufacturers: IBM, Informatica, SAP / BusinessObjects, and Microsoft; small companies that are interested in their specific solutions: Ab Initio, iWay Software, Syncsort, DMEExpress, Embarcadero Technologies, Sunopsis; startups.

Over the past few years, Informatica has not left the position of the leader in the classification of Gartner analysts, which is understandable: the company specializes in data integration, and its activities extend both to the intracorporate level and to the level of inter-corporate interaction. The general direction of Informatica was set in 1993 when the company was created by two immigrants from India – Gaurav Dhillon and Diaz Nizamoni. Even then, they realized that manual download technologies should be countered with simple and convenient means for automated loading of data into storage. Since then, Informatica has offered the classic ETL platform of data integration Informatica PowerCenter, which allows organizations, regardless of their size, to access and integrate data from arbitrary formats from various sources. This company's flagship product includes various options, such as Metadata Manager, Data Analyzer, and Data Profiler. With Informatica Power Exchange, you can connect to various databases, and with the help of the Informatica Data Quality product, you can check and clean data, identify duplicates and relationships, correct individual rows, and so on. The platform has high levels of fault tolerance and scalability.

For interconnection with external organizations, the Informatica B2B Data Exchange is a subsystem that is used, but the external environment is unsafe, so it can not be accessed without personal control. The Informational Cloud product allows the creation of applications based on the principles of SaaS.

Towards Semantic Integration

And yet, no matter how perfect the relational and functional approaches to data consolidation are from the technical point of view, they are primitive in nature. All of them are based on the traditional engineering worldview that has evolved throughout the years of the 1960s, "the data is separate, the content is separate." Their common weakness is that data is treated as primitive, raw material, as sets of bits and bytes, and their semantics are not taken into account in any way. In addition, there is a clear definition of what is, is, and is; data is what is on the lower level in the hierarchy of data-information-knowledge. "All this is strange; how could you create such databases with such an understanding of the subject? Involuntarily, the question arises – is this the base of what?"

The first serious attempts to change the existing situation were related to the creation of systems in the 1980s consisting of multiple databases, followed by DBMSs with agents acting as "mediators agents." These works made it possible to understand that the reason for the complexity of integrating the data contained in the classical databases is the tight binding of the bases and the uniqueness of

the storage scheme used in them. These works remained at the level of academic or university research and, moreover, did not offer solutions to the problem of integrating heterogeneous data, a problem that was aggravated by the widespread of unstructured or quasi-structured data. The common cause of failure is that integration problems are not purely technical – it's not at all difficult to combine different relational databases using Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC) interfaces, but it's more difficult to integrate data from sources that have different models or, worse, having different semantics, that is, interpreting the same data in different ways. To automate the work with data, the semantics must be explicitly expressed and included in this data. Here, you can draw a parallel with how a person generalizes data in his daily life based on an understanding of the world around him, the semantics of which are already contained in it. The computer does not possess intelligence, and integration programs of knowledge about semantics do not have. Therefore, there is only one way out: data should contain descriptions of its own semantics. If this is so, then it will be possible to move to the next level of integration, which can be called semantic. Semantic integration can ensure the unification of only those data that correspond or are closest to the same entities in the surrounding world. The first attempts to create systems with semantic integration date back to the early 1990s. Then, for the first time, the concept of ontologies was applied to computer data. Ontologies are considered here as ways of formal description of concepts and their interrelations; they correspond to vocabularies containing basic concepts.

Ontologies allow you to create models that more closely correspond to reality than other methods of classification. At the same time, the use of ontologies to create queries and analyses is no more complex than traditional methods, primarily because the ontological graph or map reflects the relationships between the entities themselves and not their identifiers. Despite all these advantages, the semantic methods did not go beyond research projects until, in May 2001, Tim Berners-Lee, along with James Handler and Ora Lassila, published an article on The Semantic Web in Scientific American Magazine. In it, they uncovered the concept of the semantic web, which they developed in the W3C in 1996. From the ordinary web, the semantic web differs in that metadata is widely used, making it a universal medium for data, information, and knowledge. Since then, the Semantic Web is still in the making; whether it will be implemented and, if so, how is unclear, but the ideas developed by the W3C consortium, standards, and languages are already actively applied to corporate systems. In a sense, the story repeats; what is happening now with the Semantic Web can be compared to what it was with Web services several years ago. The service ideas, SOAP, UDDI, WSDL, and other protocols originated on the Web, but they were quickly adapted to corporate systems, and a service architecture was born. As a consequence, the services used in SOA have long been called Web services exclusively, although they were related only to the use of a common stack of standards from the Web. Gradually, services were spun off from the Web and became an independent SOA foundation.

The Semantic Web approach adds a new quality, allowing you to use data not blindly but consciously, defining and linking them in such a way as to simplify the search, automate work with them, redistribute between applications, and integrate. The way data is represented in the Semantic Web can be viewed as a new step in data management, and it is natural to take advantage of these

advantages in corporate information systems. It is easy to see that the unity of all components of the information infrastructure (SOA, databases, business processes, software) gives them a common set of terms and agreements. They link separate fragments to the overall picture; that is, they are semantically united. They already exist, but they exist implicitly. This fact was usually overlooked, so formal integration solutions proved to be complex, expensive, and often disastrous.

For the most part, semantic models are built on the basis of one of the directions in first-order logic (calculus of predicates), on so-called descriptive logic, which is a family of languages that allows formally and unambiguously to describe concepts in any subject area. Each class ("concept") can be correlated with another similar concept by adding metadata tags that point to properties, commonalities, differences, etc. Extension of models with tags allows the creation of structures that could not previously exist. In a semantic model, any information unit is represented by a graph, which simplifies its modernization; for example, the merging of the two models is reduced to combining their graphs. The information item can be represented by the Uniform Resource Identifier (URI), through which relationships can be established between two or more information units. Semantic models, which are also ontologies, can be written using the Resource Description Framework (RDF) model to represent data developed by W3C in Web Ontology Language (OWL).

Fig. 2. Semantic model of information representation

Fig. 2 shows a semantic model of information. At the bottom, the atomic level of data representation is a glossary, followed by taxonomies, representing a design paradigm, in contrast to a glossary formed by a simple hierarchy of data elements. Each content element must represent its membership in a particular taxonomy. The next level – ontology (glossary and taxonomy), is a structure that simultaneously expresses a hierarchy and a set of relations between the elements of the vocabulary in this hierarchy. The semantic set unites ontologies, taxonomies, and glossaries that are part of the corporate system.

On The Way To Information Technology

In order for computer technologies to be rightly called informational, they must at least learn how to work with information.

In the terminology of the Semantic Web, it is rather difficult to understand because most words are already used in other contexts, where they have a different, usually broader meaning. For example, ontology has always been called the section of philosophy that studies being, and in computer science, ontology is the formalization of a certain area of knowledge with the help of a conceptual scheme. Information ontologies are even simpler; they consist of instances, concepts, attributes, and relationships and must have a format that the computer can process (Figure 3).

Without delving into the Semantic Web languages, we confine ourselves to presenting a stack of existing and developed standards (Table 1). Table 2 compares the traditional and semantic

approaches to data integration.

What's Next?

With a positive outcome, it is likely to be possible to talk about a semantic enterprise (Semantic Enterprise), where information technology will be firmly linked to the semantics. Most likely, all this will be combined with enterprise architecture (enterprise architecture, EA), using metamodels such as UML, DoDAF, FEAF, ToGAF, and Zachman, but this will be in the distant future. However, already today, there are a number of products on the market in which, albeit in part, semantic integration methods are implemented. Previously, others (late 2006) released the Semantic Integrator DataXtend Semantic Integrator company Progress Software, and today, DataXtend exists as a family of products actively promoted in the market. Elements of semantic integration are in the products of the company DataFlux, which is part of SAS. Oracle Spatial 11g, part of Oracle Database 11g, is interesting in that it supports, in particular, the standards RDF, RDFS, and OWL.

Ontology Language On The Web

The new OWL language will help launch automated tools for a next-generation global network, offering accurate Web search services, intelligent software agents, and knowledge management.

Quasi-Structured Data And Report Mining

In his blog, Bill Inmon left a very curious entry: "Structured and unstructured data, building bridges," in which he writes that most of the current information systems have grown around structured data with their strict rules, formats, fields, lines, columns, and indices, but today the bulk of the accumulated data is classified as unstructured, for which there are no rules and formats.

Now, these two data worlds exist in isolation, but bridges can be built between them, and Inmon looks at these bridges through the prism of relational consciousness, believing that a large part of the unstructured can be integrated if they are structured. Inmon is convinced that this is possible if texts are rid of words that are ignored (stop words), the text is reduced differently, and words are grouped and classified. After that, you can apply the same analysis methods to the texts that are used for structured data. Inmon notes that this approach does not exclude the use of semantic methods based on ontologies, glossaries, and taxonomies.

However, there is another way to bring two worlds of data together, which is through an intermediate class of semi-structured data. Very many documents used in business can not be considered completely free of structuring, but their structures are not as stringent as databases. While reports and similar documents are intended for human readings, they are quite formal; you can find the form accepted in the workflow during the analysis; for example, some of the data contains metadata.

These features are different reports created by systems such as ERP, CRM, and HR. Recognition, parsing, and transformation of reports constitute a new direction called Report Mining – the development of reports.