Цифровизация экономики и проблемы интеграции баз данных*

А.В. Соловьев, И.В. Туманова

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. В статье описаны проблемы интеграции баз данных. Выполнен анализ методов и подходов к интеграции баз данных. Рассмотрена применимость различных подходов к интеграции баз данных на уровне данных, программных и пользовательских интерфейсов, на функционально-прикладном и организационном уровне, на уровне корпоративных информационных систем, и веб-сервисов. Проанализированы современные интегрирующие модели данных, средства интеграции и стандарты, позволяющие разрабатывать системы интеграции данных. Сделан вывод о важности интеграции данных в условиях цифровизации экономики для повышения надежности, конкурентоспособности и эффективности бизнес-процессов в условиях интеграции цифрового взаимодействия предприятий. Обоснован вывод о повышении эффективности экономики в целом на основании накопления данных о человеческом капитале, т.к. экономический рост и развитие страны существенно зависят от человеческого капитала.

Ключевые слова: *интеграция баз данных, метаданные, системы управления базами данных, цифровизация, методы интеграции, большие данные, человеческий капитал.*

DOI: 10.14357/20790279200406

Введение

В настоящее время предложены многочисленные концепции цифровизации экономики, происходит повсеместное внедрение новых цифровых платформ, позволяющих существенно повысить эффективность реализации деловых процессов предприятий и организаций. Однако по мере интеграции деловых процессов организаций в рамках цифрового взаимодействия, объединения предприятий в холдинги, все острее встает проблема объединения накопленных этими организациями цифровых информационных ресурсов. Переход всех организаций на единую цифровую платформу крайне сложен из-за высоких финансовых затрат, а также вследствие сложности выбрать однозначное предпочтение в пользу единственной цифровой платформы из-за наличия большого количества альтернатив. Даже если бы такая платформа, обладающая неоспоримыми преимуществами перед всеми другими вдруг возникла, то сложность объединения цифровых информационных ресурсов все равно бы снизилась незначительно из-за разной организации структур и форматов цифровых информационных ресурсов, исторически сложившихся в рамках деятельности тех или иных организаций.

Основой информационных ресурсов организаций и предприятий, безусловно, являются базы данных. В таких условиях проблема объединения информационных ресурсов будет представлять собой задачу интеграции разнородных баз данных, находящихся в различных узлах гетерогенной информационной системы, с целью создания единого представления данных информационных ресурсов для различных пользователей разных предприятий и организации совместной работы с интегрированными наборами данных.

Дополнительной сложностью задачи интеграции баз данных является тот факт, что к настоящему времени каждая организация имеет свою достаточно продолжительную историю развития цифровых информационных ресурсов и накопленные в них данные, зачастую, могут быть отнесены к классу «больших данных» как в части накопленных объемов, так и многообразия.

Решение же поставленной задачи открывает большие перспективы для многих отраслей цифро-

* Работа выполнена при частичной финансовой поддержке РФФИ в рамках научного проекта № 19-29-07271.

вой экономики как в плане интеграции информационных ресурсов предприятий, так и для решения важных социальных задач, таких как инвестиции в человеческий капитал. Ведь экономический рост и развитие страны существенно зависят от человеческого капитала. Под человеческим капиталом мы понимаем знания, навыки и здоровье, которые люди аккумулируют в течение своей жизни, что позволяет им реализовывать свой потенциал в качестве полезных членов общества. В этой связи решение рассматриваемой в статье задачи может также рассматриваться в качестве инвестиции в человеческий капитал. В настоящее время различные государственные ведомства, а так же частные компании и предприятия, как правило, обладают данными о человеческом капитале. Однако только совокупность этих данных может дать полную картину состояния человеческого капитала в целом, а, следовательно, и ставить цели и задачи развития экономики исходя из объективных данных. Конечно, важным и даже можно сказать, болезненным вопросом является защита базы данных о человеческом капитале от несанкционированного доступа и использования в преступных целях, но решение задачи безопасности данных должно быть неотъемлемой частью вообще всех критически важных данных в цифровой экономике в целом.

Данная статья посвящена анализу существующих проблем, методов и подходов интеграции баз данных, а также их основным особенностям. Приведены области возможного практического применения интегрированных баз данных, а также сделан вывод о пригодности тех или иных методов на практике.

1. Проблемы интеграции баз данных

Как было сказано выше, даже гипотетический переход всех организаций в рамках цифровизации экономики на единую цифровую платформу не решит проблем, связанных с объединением баз данных (БД) по-разному организованных и имеющих разную структуру. Иными словами, при интеграции любых баз данных придется решать проблемы несоответствия схем данных и самих данных разных БД [1].

Проблемы несоответствия схем данных можно кратко представить следующим списком:

– проблема неоднородности схем данных. Проблема возникает, когда в разных БД используются различные модели данных (реляционные, объектно-реляционные, документно-ориентированные, сетевые и др.);
– проблема несоответствия имен атрибутов данных. Проблема возникает, когда в различных схемах БД используются разные наименования атрибутов данных, что приводит к необходимости составления словарей омонимии и синонимии в именовании данных;
– проблема несоответствия структуры данных. Проблема возникает, когда одни и те же объекты из реального мира в разных БД представлены разными структурами данных;
– проблема несоответствия семантики данных. Проблема возникает, когда для моделирования в конкретных БД разные объекты из реального мира могут представляться похожими, аналогичные объекты наоборот представлены различными в силу специфики той или иной организации и их взгляду на объект (например, в БД силовых структур человек может быть представлен с точки зрения его отношений с действующим законодательством, а в банковской системе – с точки зрения финансовых операций).

В свою очередь задача несоответствия структуры и семантики данных порождает следующие проблемы:

– проблема несоответствия типов данных. Возникает, когда, например, в одной БД атрибут представлен числовым значением, в другом — строковым;
– проблема несоответствия в единицах измерения, например, в одной БД трудоемкость представлена в человеко-днях, в другой — в человеко-часах;
– проблема несоответствия допустимых значений, например, оценки успеваемости могут быть представлены разными шкалами: 5-, 10-, 100-бальные;
– проблема использования словарей, например, данные в одной БД представлены строкой, в другой – значениями из некоторого справочника;
– проблема представления структур данных, например, адрес может быть представлен в одном случае строкой, в другом – набором отдельных значений (индекс, страна, город, улица, дом и т.д.);
– проблемы одинаковых наименований разных сущностей. например, в одной БД «программист» это должность, в другой – квалификация или роль в системе;
– проблема несовпадения данных. В разных БД один и тот же объект описан разными значениями атрибутов, например, в силу несвоевременного обновления данных (например, в разных БД у автомобиля разные владельцы вследствие наличия непроверенной информации).

Данные проблемы могут привести к тому, что интеграция БД потребует вмешательство человека

к их разрешению даже несмотря на наличие достаточно проверенных методов решения проблем несоответствия данных (см. [2]).

Тем самым, можно сделать вывод, что в общем случае для преодоления перечисленных проблем, при решении задачи интеграции БД неизбежно возникнет необходимость решения следующих частных задач:
– разработка архитектуры системы интеграции БД;
– разработка интегрирующей модели данных БД;
– разработка методов отображения моделей данных и построение отображений в интегрирующую модель для конкретных моделей, поддерживаемых отдельными БД;
– разработка методов интеграции метаданных БД (схемы данных, индексы, справчоники и др.);
– решение проблем неоднородности БД, в том числе, разработка методологии и механизмов семантической интеграции БД.

2. Анализ методов и подходов к интеграции БД

В начале 80-х годов XX века появились первые разработки систем обеспечения совместимости разнородных БД [3]. Затем в США был разработан подход, согласно которому стало возможно загружать данные из разнородных БД в единую схему представления данных [4]. С применением такого подхода был разработан метод интеграции БД, на основании которого были созданы несколько тысячи демографических БД США.

Однако существенным недостатком данного подхода было то, что он оказался пригоден только для редкообновляемых данных. Для данных, для которых требуется непрерывное повторное выполнение процессов извлечения, преобразования, загрузки в единую БД для синхронизации, такой подход оказался слишком затратным.

Другой подход, появившийся относительно недавно, основан на обеспечении единого интерфейса запросов для доступа к данным интегрируемых БД в реальном времени. В основе реализации подхода - обеспечение обобщенной схемы запросов к разнородным БД, что позволяет извлекать информацию непосредственно из исходных БД. Подход основан на сопоставлении между обобщенной схемой запросов и схемой интегрируемых БД. Результатом сопоставления является преобразование запроса обобщенной схемы в запросы, которые соответствуют схемам данных исходных БД [5].

Подход определяет методы сопоставления:
– структур объектов обобщенной схемы запроса со структурами объектов исходных БД (метод Global As View (GAV));
– структур объектов исходных БД со структурами объектов обобщенной схемы запроса (метод Local As View (LAV)).

Метод LAV требует более сложных, по сравнению с GAV, преобразований для выполнения запроса по обобщенной схеме, но упрощает добавление новых БД в обобщенную схему запросов.

При программной реализации метода GAV следует учитывать, что обобщенная схема запроса будет всегда ограничена по сравнению с исходными БД [6]. Для систем, построенных на основе реализации метода LAV данная проблема снимается, поэтому LAV-системы могут быть предназначены для запроса неполных данных и их последующей интеграции.

При разработке LAV-систем, разработчик сначала проектирует обобщенную схему запросов, а затем интегрирует в нее схемы исходных БД. Основная сложность разработки – добавление новой схемы БД для обработчика запросов.

Обычно запросы для БД представляются в конъюнктивной нормальной форме [7]. При этом необходимо отметить, что в общем случае задача преобразования запроса является NP-полной [8]. Однако, если пространство преобразования запросов относительно невелико, то это не представляет проблемы даже для систем интеграции с сотнями БД. Если же это условие не соблюдается, сложность разработки LAV-системы многократно возрастает.

Несмотря на это, разработка алгоритмов преобразования запросов в последнее время ведется очень интенсивно. В частности, разработанный алгоритм GQR [9] в настоящее время является крайне востребованным для LAV-систем интеграции БД.

Тем самым, из анализа современных подходов и методов интеграции БД можно сделать вывод о том, что наиболее гибким и подходящим для интеграции информационных ресурсов предприятий в условиях цифровизации отдельных отраслей или экономики в целом является метод LAV в силу его большей универсальности и проработанности на уровне алгоритмов преобразования запросов к данных разнородных БД.

3. Анализ уровней интеграции БД

В настоящее время при рассмотрении проблемы интеграции БД научное сообщество выделяет несколько уровней [10–12]:
– уровень данных;

– уровень программных и пользовательских интерфейсов;
– функционально-прикладной и организационный уровень;
– уровень корпоративных информационных систем (КИС);
– уровень веб-сервисов.

Интеграция БД на каждом из перечисленных уровней обладает своими особенностями и проблемами, которые рассмотрим в данном разделе.

3.1. Интеграция БД на уровне данных

Важной проблемой интеграции БД является обилие форматов и типов данных. Данные могут быть: неструктурированные, частично-структурированные, жестко-структурированные. Второй не менее важной проблемой интеграции БД является лавинообразное нарастание объемов информационных ресурсов крупных предприятий и организаций.

Такая ситуация создает множество проблем с структурированием, обработкой, анализом, хранением, архивированием и представлением пользователю разнородных данных интегрируемых БД.

Для решения описанной проблемы в настоящее время используют:
– стандартные языки запросов и протоколы (например, SQL, JDBC, OLE DB и др.);
– надмножество словарей метаданных;
– технологии хранилищ данных.

Однако все эти способы не лишены недостатков, так стандартные языки запросов и интерфейсы, хоть и стандартизованы, мало пригодны при интеграции слабоструктурированных и неструктурированных данных. Использование словарей метаданных требуют четкого понимания структуры метаданных разнородных объектов и их сложного описания, что само по себе является крайне трудозатратной деятельностью. Реализация хранилищ данных требует много технических, финансовых ресурсов для своей реализации и под силу только крупным компаниям. Вследствие этого, ни один из используемых способов не является полностью самодостаточным, способным решить проблему интеграции данных.

3.2. Интеграция БД на уровне программных и пользовательских интерфейсов

В современной научной литературе определены следующие основные методы интеграции на уровне программных и пользовательских интерфейсов:
– «лоскутная интеграция». Этот метод подразумевает объединение разрозненных программных приложения, написанных в разное время разными разработчиками. Все приложения объединяются по принципу «каждый с каждым». Это сильно усложняет взаимодействие и использование как унаследованных, так и встроенных систем предприятия. Этот подход приемлем для небольшого количества интегрируемых приложений, если что не так, то метод практически не позволяет строить качественно новые запросы к интегрируемым данным БД;
– использование открытых программных интерфейсов (Open Application Programming Interface). Такой подход решает проблему интеграции на уровне интерфейсов. В качестве примера стандартизации этого подхода можно привести унифицированные интерфейсы на базе семейства международных стандартов POSIX. Степень интегрируемости можно характеризовать некоторой метрикой, которую можно вычислить, перемножив показатель «качества» и «показатель открытости» программного интерфейса. Показателями качества программного интерфейса могут быть: совместимость, надежность, переносимость, понятность, удобство использования;
– выделение слоя обработки данных от слоя визуализации. Еще одним подходом к интеграции БД на этом уровне является отделение слоя обработки данных от слоя их визуализации. Программный доступ к прикладным функциям приложения при этом выполняется в виде хорошо документированного программного интерфейса. Данный подход наиболее востребован и популярен последнее время для решения задачи интеграции БД.

3.3. Интеграция БД на функционально-прикладном и организационном уровне

Интеграция БД на этом уровне подразумевает объединение однотипных функций в, так называемые, макрофункции. Такой подход чаще всего предполагает перестройку организационных структур и бизнес-процессов организаций, а, следовательно и их информационного взаимодействия.

Преимуществом интеграции БД на данном уровне является то, что бизнес-процессы становятся более управляемыми и менее затратными. Как следствие снижается количество ошибок обработки данных. Однако такая интеграция БД может содержать и риски рассогласования бизнес-процессов и снижения их эффективности вследствие неверного объединения процессов. Интеграция БД на данном уровне оправдана, когда организации внедряют единую корпоративную информационную систему (КИС), что, в частности, требует приведения бизнес-процессов к определенному стандарту.

3.4. Интеграция БД на уровне корпоративных информационных систем

Интеграция БД на этом уровне предполагает создание КИС, т.е. интеграцию информационных систем организаций на базе интеграции их бизнес-процессов, т.е. создание КИС уровня холдинга, группы предприятий, отрасли и т.д.

Архитектура КИС в таком случае строится из набора слабосвязанных гетерогенных сервисов и понимается как единая парадигма организаций с использованием распределенного множества функций, которые могут контролироваться различными пользователями. Базовыми понятиями в такой архитектуре являются «информационная услуга» и «композитное приложение».

В плане объединения информационных ресурсов предприятий такой подход наиболее эффективен, но затратен, т.к. требует выработки единых стандартов управления, реализации и контроля бизнес-процессов, затрат на создание единой информационной инфраструктуры.

3.5. Интеграция БД на уровне веб-сервисов

Подход к интеграции на этом уровне является самым современным по времени. В основе его лежит обеспечение стандартного интерфейса доступа к данным всех БД объединяемых организаций с помощью веб-сервисов. Подход удобен тем, что пользователь, используя стандартный протокол SOAP и обычный веб-браузер, может одновременно использовать корпоративные приложения разных организаций, доступ к которым осуществляется и контролируется соответствующими веб-сервисами.

Преимуществами интеграции БД на уровне веб-сервисов являются:

- возможность осуществлять оперативное управление данными группы организаций;
- возможность осуществлять планомерное развитие общекорпоративной информационной системы, интегрируя в нее функциональные компоненты, исходя из приоритетов развития организаций;
- возможность при необходимости заменить любой функциональный компонент другим, более соответствующим текущим бизнес-потребностям;
- возможность инвестировать в интеграцию данных не сразу, а поэтапно, на каждом этапе, соотнося вложенные средства с полученным бизнес-эффектом;
- возможность сохранять инвестиции в имеющиеся информационные системы;
- упрощение обучения персонала.

Интеграция на этом уровне представляет собой наибольший интерес в рамках рассматриваемой задачи интеграции БД, т.к. с одной стороны на лицо сокращение затрат на интеграцию за счет выделения отдельного уровня сервисов, который относительно не сильно затрагивает уровень информационных систем отдельных организаций, с другой, возможность контролировать обмен данными между организациями с помощью модулей контроля доступа и решений распределенных реестров.

4. Модели и средства интеграции БД

4.1. Интегрирующие модели данных

В связи со стремительной цифровизацией в последние годы создан ряд моделей, обеспечивающий представление как структурированных, так и слабоструктурированных данных [13, 14].

К подобным моделям можно отнести также объединение реляционной и объектной моделей данных, доведенного, в частности, до стандартизации в рамках SQL [15, 16], объединение модели XML-данных и реляционной, также стандартизированной [17, 18].

Тем не менее, создание единой модели, включающей все особенности различных моделей данных вряд ли возможна, ввиду ее крайней сложности. Поэтому применение интегрирующих моделей позволяет упростить задачу интеграции БД, но не решает ее в общем случае.

4.2. Средства семантической интеграции БД

Выше была рассмотрена проблема несоответствия семантики объединяемых данных. В настоящее время наиболее распространен подход к решению проблемы семантической интеграции БД на основе использования семантических медиаторов (посредников). Исследованию проблем и методов семантической интеграции данных посвящены, например, работы [19, 20].

Как правило, семантические посредники разрабатываются для конкретной узкой предметной области. Механизмы посредников опираются на унифицированные метаописания интегрируемых источников данных. Недостатком такого подхода является необходимость существования интегрирующей модели данных с развитыми возможностями моделирования семантики данных.

В последние годы появился ряд публикаций (см. [21, 22]), посвященных решению проблемы семантической интеграции БД, в которых для представления обобщенной схемы данных, предлагается использовать аппарат дескриптивных логик, реализованный в языке описания онтологий OWL.

Достоинствами подхода являются:
- использование высокоуровневой семантической модели данных;
- возможность описания интегрируемых данных в терминах онтологии, которая является концептуальной моделью данных.

Недостатком является дороговизна реализации данных моделей на практике.

4.3. Стандартизация интеграции БД

При проектировании систем интеграции данных, безусловно, необходимо опираться на существующие стандарты. Это позволит существенно снизить риски создания системы, которую потом невозможно будет использовать повторно. Среди появившихся в последние годы стандартов, необходимо выделить:
- стандарты баз данных ISO/IEC SQL, ISO/IEC SQL/MED;
- стандарт объектных данных ODMG;
- стандарты языка моделирования UML консорциума ODMG;
- стандарты платформы XML консорциума W3C;
- стандарт Дублинского ядра для описания метаданных консорциума OCLC.

Стандарты важны для определения унифицированной модели метаданных интегрируемых БД, а также для разработки единого интерфейса доступа к интегрированным данным.

Некоторые стандарты, например, стандарты XML и CORBA, позволяют «погрузить» интегрированные данные в некоторую полезную инфраструктуру и использовать ее функциональность для доступа к данным. В качестве примера здесь можно привести интеграцию информационных ресурсов электронных библиотек, использующих стандарты Open Archives Initiative (http://www.openarchives.org/).

5. Возможное практическое применение интеграции БД

Практическое применение интеграции БД возможно во всех отраслях будущей цифровой экономики, образовании, создании единых информационных сервисов на подобие ГосУслуг.

Например, крайне важной задачей является развития человеческого капитала в Российской Федерации. Также различные государственные ведомства и организации, частные компании и предприятия, как правило, обладают большими наборами данных, хранящимися в собственных специализированных базах данных. Актуальность задачи интеграции БД определяется необходимостью интегрировать эти данные в единую гетерогенную распределенную информационную систему для повышения конкурентоспособности в рамках цифровизации экономики, повышения эффективности бизнес-процессов, проведения аналитических и статистических исследований, в том числе, например, отслеживания параметров состояния человеческого капитала. Интеграция баз данных о человеческом капитале важно для экономики в целом и может быть рассмотрена как инвестиция в ее развитие, особенно в условиях цифровизации. Инвестиции в человеческий капитал повышают производительность и конкурентоспособность экономики. Производительность же людских ресурсов зависит от наличия доступного здравоохранения, образования, материальных активов, от стабильной и эффективно управляемой экономики. В свою очередь, здоровые и образованные люди могут больше зарабатывать и инвестировать в экономику страны.

Заключение

В статье были проанализированы подходы к интеграции, методы и уровни интеграции БД, изучены проблемы, возникающие в процессе интеграции баз данных. В качестве подхода для построения систем интеграции в условиях исторических и технических особенностей развития информационных систем организаций и ведомств был выбран подход Loval As View (LAV). Этот подход является наиболее подходящим в условиях независимого развития информационных систем организаций, необходимости интегрировать разнородные базы данных, необходимости подключения большого числа разнородных источников исходных баз данных в условиях цифровизации экономики.

В дальнейших исследованиях планируется разработка описания информационных объектов БД о человеческом капитале, описание информационных потоков данных, методов интеграции БД.

Литература

1. *William Kent*. Solving Domain Mismatch and Schema Mismatch Problems with an Object-Oriented Database Programming Language. Proceedings of the International Conference on Very Large Data Bases. 1991.
2. *Erhard Rahm, Philip A. Bernstein*. A Survey of Approaches to Automatic Schema Matching. VLDB JOURNAL. 2001.
3. *John Miles Smith*; et al. “Multibase: integrating heterogeneous distributed database systems”.

AFIPS ‘81 Proceedings of the May 4–7, 1981, National Computer Conference. P. 487–499.
4. *Steven Ruggles, J. David Hacker, and Matthew Sobek* (1995). “Order out of Chaos: The Integrated Public Use Microdata Series”. Historical Methods. 28. P. 33–39.
5. *Shubhra S. Ray*; et al. “Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast” (PDF). IEEE Transactions on Biomedical Engineering. 56 (2): 229–236. CiteSeerX 10.1.1.150.7928. 2009.
6. *Alagić Suad, Bernstein, Philip A.* Database Programming Languages. Lecture Notes in Computer Science. 2002. 2397. P. 228–246.
7. *Jeffrey D. Ullman*. “Information Integration Using Logical Views”. ICDT 1997. P. 19–40.
8. *Alon Y. Halevy*. “Answering queries using views: A survey” (PDF). The VLDB Journal. 2001. P. 270–294.
9. *George Konstantinidis et al*. “Scalable Query Rewriting: A Graph-based Approach” (PDF). in Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD’11. June 12-16. 2011. Athens. Greece.
10. *Когаловский М. Р.* Перспективные технологии информационных систем. М.: ДМК Пресс. 2003. 288 с.
11. *Танянский С.С.* Семантическая модель предметной области в задачах интеграции неоднородных информационных систем // Вестник ХНТУ. 2005. №21. С. 52-59.
12. *Понамаренко Л. А., Танянский С. С., Филатов В. А.* Интеграция информационных систем при частичном отображении моделей данных // Проблемы системного подхода в экономике, 2008. Вып.26. С.33-44.
13. *Christophides V., Cluet S., Simeon J.* Semistructured and structured integration reconciled. http://www.rocq.inria.fr/verso/Jerome.Simeon/YAT/.
14. *Lahiri T., Abiteboul S., Widom J. Ozone*: Integrating structures and semistructured data. http://www-db.stanford.edu/pub/papers/ozone.ps.
15. *Stonebraker M.* et al. Third-Generation Data Base System Manifesto. Proc. IFIP WG 2.6 Conference on Object Oriented Databases, Windermere, England, July 1990. (Стоунбрекер М. и др. Системы баз данных третьего поколения: Манифест//СУБД. 1995. №2. С. 143–158).
16. *Эйзенберг Э., Мелтон Дж.* SQL:1999, ранее известный как SQL3. Пер. с англ. Открытые системы. 1999. № 1. С. 52-57.
17. *Eisenberg A., Melton J.* SQL/XML is Making Good Progress. SIGMOD Record, Vol. 31. № 2. June 2002.
18. ISO/IEC 9075.14:200x(E). Information technology – Database language – SQL – Part 14: XML Related Specification (SQL/XML). 2001.06.18. Working Draft. http://www.sqlx.org/.
19. *Wiederhold G.* Mediators in the Architecture of Future Information Systems. IEEE Computer 1992. 25:3. P. 38-49.
20. *Kalinichenko L. A.*, *Briukhov D.O., Skvortsov N.A., Zakharov V.N.* Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. Вторая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Протвино. 2000. С. 78-90.
21. *Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R.* Conceptual Modeling for Data Integration. http://www.inf.unibz.it/~calvanese/papers/calv-etal-book-mylopoulos-2009.pdf
22. *Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R., Ruzzi M.* Using OWL in Data Integration. Chapter 14. http://www.dis.uniroma1.it/~rosati/publications/Calvanese-etalSWIMBook-09.pdf

Соловьев Александр Владимирович. Федеральный исследовательский центр «Информатика и управление» РАН. Главный научный сотрудник, доктор технических наук. Количество печатных работ: 120. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, математическое моделирование, долговременное хранение электронных документов. E-mail: soloviev@isa.ru

Туманова Ирина Владимировна. Федеральный исследовательский центр «Информатика и управление» РАН. Ведущий программист. Количество печатных работ: 10. Область научных интересов: системное программирование, информационные технологии, электронный документооборот, электронный архив. E-mail: tumanova-irin@mail.ru

Digitalization of the economy and problems of database integration

A.V. Solovyev, I.V. Tumanova
Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. This article describes the problems of database integration. The analysis of methods and approaches to database integration is carried out. The applicability of various approaches to the integration of databases at various levels is considered: at the data level, at the level of program and user interfaces, at the functional-application and organizational level, at the level of corporate information systems, at the level of web services. Analyzed modern integrating data models, integration tools and standards that allow the development of data integration systems. The conclusion is made about the importance of data integration in the context of the digitalization of the economy to improve the reliability, competitiveness and efficiency of business processes in the context of the integration of digital interaction of enterprises. It is also concluded that an increase in the efficiency of the economy as a whole is based on the accumulation of data on human capital, since the economic growth and development of the country significantly depend on human capital.
Keywords: *database integration, metadata, database management systems, digitalization, integration methods, big data, human capital*
DOI: 10.14357/20790279200406

References

1. *William Kent*. Solving Domain Mismatch and Schema Mismatch Problems with an Object-Oriented Database Programming Language. Proceedings of the International Conference on Very Large Data Bases (1991).
2. *Erhard Rahm, Philip A. Bernstein*. A Survey of Approaches to Automatic Schema Matching. VLDB JOURNAL (2001).
3. *John Miles Smith*; et al. (1982). "Multibase: integrating heterogeneous distributed database systems". AFIPS '81 Proceedings of the May 4–7, 1981, National Computer Conference. pp. 487–499.
4. *Steven Ruggles, J. David Hacker, and Matthew Sobek* (1995). "Order out of Chaos: The Integrated Public Use Microdata Series". Historical Methods. 28. pp. 33–39.
5. *Shubhra S. Ray*; et al. (2009). "Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast" (PDF). IEEE Transactions on Biomedical Engineering. 56 (2): 229–236. CiteSeerX 10.1.1.150.7928. doi: 10.1109/TBME.2008.2005955. PMID 19272921.
6. *Alagić, Suad; Bernstein, Philip A.* (2002). Database Programming Languages. Lecture Notes in Computer Science. 2397. pp. 228–246. doi: 10.1007 / 3-540-46093-4_14. ISBN 978-3-540-44080-2.
7. *Jeffrey D. Ullman* (1997). "Information Integration Using Logical Views". ICDT 1997. pp. 19–40.
8. *Alon Y. Halevy* (2001). "Answering queries using views: A survey" (PDF). The VLDB Journal. pp. 270–294.
9. *George Konstantinidis*; et al. (2011). "Scalable Query Rewriting: A Graph-based Approach" (PDF). in Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD'11, June 12-16, 2011, Athens, Greece.
10. *Kogalovsky M.R.* Advanced technologies of information systems [Perspektivnyye tekhnologii informatsionnykh sistem] M.: DMK Press, 2003, 288 p.
11. *Tanyansky S.S.* Semantic domain model in problems of integration of heterogeneous information systems [Semanticheskaya model' predmetnoy oblasti v zadachakh integratsii neodnorodnykh informatsionnykh sistem] // Bulletin of KhNTU, 2005, №21, pp. 52-59.
12. *Ponamarenko L.A., Tanyansky S.S., Philatov V.A.* Integration of information systems with partial display of data models [Integratsiya informatsionnykh sistem pri chastichnom otobrazhenii modeley dannykh] // Problems of a systems approach in economics, 2008, Issue 26, pp. 33-44.
13. *Christophides V., Cluet S., Simeon J.* Semistructured and structured integration reconciled. http://www.rocq.inria.fr/verso/Jerome.Simeon/YAT/.
14. *Lahiri T., Abiteboul S., Widom J. Ozone*: Integrating structures and semistructured data. http://www-db.stanford.edu/pub/papers/ozone.ps.
15. *Stonebraker M.* et al. Third-Generation Data

Base System Manifesto. Proc. IFIP WG 2.6 Conference on Object Oriented Databases, Windermere, England, July 1990.
16. *Eisenberg A., Melton J.* SQL:1999, formerly known as SQL3 // Open systems. – 1999. - № 1. – pp. 52-57.
17. *Eisenberg A., Melton J.* SQL/XML is Making Good Progress. SIGMOD Record, Vol. 31, No. 2, June 2002.
18. ISO/IEC 9075.14:200x(E). Information technology – Database language – SQL – Part 14: XML Related Specification (SQL/XML). 2001.06.18. Working Draft. http://www.sqlx.org/.
19. *Wiederhold G.* Mediators in the Architecture of Future Information Systems. IEEE Computer 25:3, pp. 38-49, 1992.
20. *Kalinichenko L. A.*, *Briukhov D.O., Skvortsov N.A., Zakharov V.N.* Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. Second All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Protvino, 2000, pp. 78-90.
21. *Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R.* Conceptual Modeling for Data Integration. http://www.inf.unibz.it/~calvanese/papers/calv-etal-book-mylopoulos-2009.pdf
22. *Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R., Ruzzi M.* Using OWL in Data Integration. Chapter 14. http://www.dis.uniroma1.it/~rosati/publications/Calvanese-etalSWIMBook-09.pdf

A.V. Solovyev. Chief Researcher, Department 94, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. Moscow. Doctor of Technical Sciences. Number of publications: 120. Area of scientific interests: system analysis, database management systems, reliability theory, mathematical modeling, electronic document management, electronic archive, long-term storage of electronic documents. E-mail: soloviev@isa.ru

I.V. Tumanova. Lead programmer, Department 94, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. Moscow. Number of publications: 10. Area of scientific interests: system programming, information technology, electronic document management, electronic archive.
E-mail: tumanova-irin@mail.ru