УДК 629.7:004.89 DOI: 10.18287/2223-9537-2021-11-2-170-184

Управление мастер-данными в рамках итеративного подхода

С.В. Кузнецов[1], Д.В. Кознов[2]

[1]*ООО «Юнидата», Санкт-Петербург, Россия*

[2]*Санкт-Петербургский государственный университет, Санкт-Петербург, Россия*

Аннотация

Управление мастер-данными (*Master-Data Management, MDM*) – область бизнес-информатики, нацеленная на консолидацию и централизованное управление важнейшими данными бизнес-организаций, которые распределены по разным информационным системам. Ведущие мировые ИТ-компании (*IBM, Oracle, Informatica* и др.) предлагают широкий спектр готовых продуктов по управлению мастер-данными (*MDM*-продукты). Внедрение *MDM* сопряжено с большим количеством технических и организационных трудностей. В работе рассматривается итеративная стратегия внедрения *MDM*, подразумевающая поэтапную реализацию управления мастер-данными на основе реальных нужд организации-заказчика. Вводится понятие *MDM*-решения, которое является результатом внедрения *MDM* в организацию и включает в себя адаптированный к нуждам и особенностям организации *MDM*-продукт, новые регламенты по работе с данными обученных сотрудников, налаженный процесс управления мастер-данными. Основным результатом статьи является функциональная модель управления мастер-данными, предназначенная для ранних стадий разработки *MDM*-решения в рамках итеративной стратегии. Потребности организации могут быть похожи на *MDM*, однако требуют использования других инструментов. Цель данной модели – выявить реальные потребности организации в *MDM* и установить, какие *MDM*-компоненты должны быть реализованы в первую очередь. В работе приводится описание компонент модели, представлено портфолио из шести реальных *MDM*-проектов и их анализ с позиций предложенной модели.

Ключевые слова: *управление мастер-данными, бизнес-информатика, прикладная онтология, итеративный подход, информационная система.*

Цитирование: *Кузнецов, С.В.* Управление мастер-данными в рамках итеративного подхода / С.В. Кузнецов, Д.В. Кознов // Онтология проектирования. – 2021. – Т.11, №2(40). - С.170-184. – DOI: 10.18287/2223-9537-2021-11-2-170-184.

Введение

Переход к цифровой экономике предъявляет особые требования к управлению данными в организациях [1]. При этом имеется особый вид данных, без которых трудно себе представить нормальную работу организаций: юридические данные, клиентская база, сведения о поставщиках и контрагентах и пр. Пользователи этих данных рассчитывают на их согласованность в пределах организации, т.е. они ожидают получить одну и ту же информацию об одном и том же объекте из разных источников (информационных систем, ИС) организации. Разночтения и противоречия здесь порождают проблемы: задержки, коллизии, финансовые и имиджевые потери организации [2]. Такие консолидированные данные принято называть мастер-данными (*Master Data,* далее МД)[1], а процесс их консолидации и сопровождения – управлением МД (*Master Data Management, MDM*) [1, 8].

В области *MDM* существует обширный спектр стратегий и методов, структурированных в рамках энциклопедии *DAMA-DMBOK* (*Data Management Body of Knowledge*) [2]. Существует большое количество готовых *MDM*-продуктов от таких компаний как *IBM, Oracle,*

[1] В России МД часто называют основными данными, при этом выделяют особый подвид МД – справочные данные (реестры, классификаторы, справочники и пр.), для которых существует много различных стандартов и решений [3-7].

Informatica и др. (см., например, [9]). На сегодняшний день внедрение *MDM* не является типовой, проработанной задачей, поскольку организации, особенно крупные, имеют большое количество индивидуальных черт.

На практике существуют две принципиально разные стратегии введения *MDM*: сверху-вниз и итеративная стратегия [2]. *Стратегия сверху-вниз* подразумевает следующую последовательность действий: создание стратегической *MDM*-концепции для организации, формирование требований к *MDM*-решению, внедрение и доработка существующего на рынке *MDM*-инструментария, выполнение необходимых организационно-административных работ, эксплуатация и сопровождение *MDM*-решения[2]. *Итеративная стратегия* подразумевает внедрение *MDM* для решения конкретной задачи с дальнейшим наращиванием *MDM*-функционала и/или реализацией *MDM* для других сегментов данных организации, т.е. для решения других задач[3]. Эти две стратегии коррелируют с концепцией внедрения инноваций в организацию, предложенной в [10]. Первую стратегию можно сопоставить с технологией *push*: внедрение инновации происходит на основе некоторой передовой технологии, которая должна решить различные, в том числе и не известные на данный момент проблемы организации. Вторую стратегию можно сопоставить с технологией *pull*: драйвер внедрения инноваций – сама организация, точнее, определённые её задачи. Не отвергая первой стратегии, авторы ориентируются на вторую, как менее рисковую и позволяющую достичь конкретных практических результатов в обозримые сроки.

При реализации *MDM*-проекта в рамках итеративной стратегии возникает задача перевода требований заказчика на язык *MDM*, получивший значительное развитие [2]. Если задача, имеющаяся у организации-заказчика, хорошо переводится в термины *MDM*, значит для её решения можно использовать имеющийся на рынке *MDM*-инструментарий [9], что существенно сокращает затраты на такой проект. При этом на практике оказывается, что заказчик, как правило, не владеет *MDM*-терминологией и часто под видом *MDM*-проектов пытается представить проекты иного класса или заказать реализацию *MDM*-решения «с нуля». Ошибки здесь приводят к коллизиям, растянутым срокам и денежным потерям.

В статье предлагается функциональная модель МД, которая должна помочь в первичной оценке и согласовании работ в итеративных *MDM*-проектах на самых ранних стадиях. Внимание сосредоточено на создании/развертке ИТ-инфраструктуры по поддержке *MDM*, её наладке и выполнении необходимых аналитических работ (очистка и консолидация данных, классификация и иерархизация и т.д.). В дальнейшем эти аналитические работы должны выполняться в организации на постоянной основе с помощью внедренной ИТ-инфраструктуры, и задачей-максимум является полная автоматизация этих работ. Ориентируясь на программно-аналитические аспекты *MDM*, здесь намеренно оставляются в стороне изменение бизнес-процессов и практик работы с данными в организации, обучение персонала и пр. Обычно эти вопросы лучше решаются, когда уже выполнены программно-аналитические работы по внедрению *MDM*. Кроме того, существует специальная область, в сферу которой попадают подобные вопросы – это управление данными (*Data Governance*, [11]). В статье приводятся несколько реальных индустриальных *MDM*-проектов, описанных в рамках предложенной модели.

[2] Как правило, это осуществляется посредством серии проектов, которые выполняются различными внешними компаниями.
[3] Внедрение *MDM* за пределами одного конкретного *MDM*-проекта зависит от многих условий, в том числе от оправдания ожиданий организации от данного проекта, от наличия других задач, свободных средств, заинтересованных людей и т.д. Именно такая цепочка *MDM*-проектов и даёт название всей стратегии – *итеративная стратегия* внедрения *MDM*.

1 Мастер-данные

Следуя глоссарию [8], МД являются специальным видом данных организации, описывающих её важнейшие характеристики, в том числе действующих и потенциальных клиентов, поставщиков, потребляемую/выпускаемую продукцию, офисные и производственные площади организации, платёжные реквизиты и различную информацию о счетах юридических и физических лиц, с которыми организация работает. Смысл выделения МД как отдельного объекта управления заключается в том, что в организациях существует большое количество различных ИС и/или используется значительное число внешних разнородных источников данных. В результате оказывается, что одна и та же информация выражена разными данными и в разных форматах. Более того, в разных источниках имеются разные атрибуты для одних и тех же данных, а сами данные могут противоречить друг другу. В организации имеется подмножество критически важных данных, разночтение в которых препятствует нормальному функционированию и наносит организации значительный ущерб. Именно эти данные становятся объектом специальной заботы организации, превращаясь в МД.

Для эффективного создания, использования и сопровождения МД в организации должна быть налажена специальная деятельность, посредством которой организация фокусируется на своих бизнес-процессах, вопросах качества и интеграции данных, а также на стандартизации существующих и используемых ИС [12]. *MDM* ориентирован на сбор и накопление данных из различных ИС – источников данных (ИД), консолидацию данных и их распределение (доставку) ИС – потребителям данных (ПД).

В [2] выделены следующие ключевые процессы *MDM* в организации: управление моделью данных, сбор и накопление данных, проверка, стандартизация и обогащение данных, разрешение конфликтов данных и использование данных.

MDM в организации должен быть поддержан специальным ИТ-решением, созданным и внедренным в организацию (далее – *MDM-решение*). Важнейшей составляющей *MDM*-решения является центральный репозиторий данных (хаб, от англ. *hub*). В него собираются данные, являющиеся кандидатами в МД, они надлежащим образом обрабатываются и доставляются ПД. Определены четыре варианта архитектуры хаба данных [9].

- Индексная архитектура: на хабе хранятся только соответствующие ссылки (индексы) данных; это актуально для данных, которые нельзя копировать или перемещать.
- Консолидирующая архитектура: данные регулярно загружаются в хаб, обрабатываются, и при этом обеспечивается доступ ПД к этим данными.
- Централизованная архитектура во всём подобна предыдущей, но в этом случае дополнительно хаб выполняет задачу разового ввода данных, и далее все изменения данных делаются непосредственно на хабе (ИС попадают в разряд ПД).
- Смешанная архитектура реализует сочетание консолидирующей и централизованной архитектур для различных МД организации. Если какие-то фрагменты данных организации запрещено перемещать, то для них может использоваться индексная архитектура.

Существует большое количество готового программного инструментария по созданию *MDM*-решений. Прежде всего, это такие продукты как *SAP MDG, Informatica MDM, IBM InfoSphere MDM* и др., которые ориентированы на решение стандартных задач *MDM*. Однако разнообразие практических задач столь велико, что некоторые производители (например, *Informatica MDM*, Юнидата и пр.) предлагают программные «конструкторы», которые позволяют получить *MDM*-решения для конкретных потребностей организаций.

2 *MDM*-проекты

На практике *MDM*-проекты часто принимают за обычные ИТ-проекты. И требуется определить, что задачи организации имеют ярко выраженную *MDM-специфику*, что позволяет применить готовый *MDM*-инструментарий, а также привлечь компании-интеграторы, специализирующиеся на таких проектах.

Запрос организации имеет *MDM*-специфику, когда организации необходим сбор, обогащение и консолидация данных из различных ИД, а также выдача этих данных различным ПД, использующим их далее для обеспечения жизнедеятельности организации. ИД могут находиться как в самой организации, так и вне её. Например, требуется обогатить данные о клиентах организации информацией, собранной в соцсетях. Требование наличия нескольких ПД для созданных МД является менее жёстким и иногда не выполняется. Как правило, в организации существует критический бизнес-процесс, для эффективного исполнения которого требуются качественные обогащённые и консолидированные данные из различных ИД. Например, речь может идти о процедуре проверки новых клиентов или спорных транзакций в банке.

С другой стороны, организации может быть и не нужен *MDM*-проект. Во-первых, когда речь идёт об обработке однородных данных – например, сформированных посредством ввода одним или несколькими операторами (операторный ввод). Такие проекты могут потребовать: создания логической модели данных, валидацию и очистку данных, обеспечение доступа к данным в различных режимах и т.д. Но при этом отсутствует главная задача *MDM* – консолидация данных из разных источников. Если для такой задачи операторный ввод данных заменить автоматизированным поступлением тех же данных из разных ИД, то задача приобретает *MDM*-специфику. Во-вторых, «вне юрисдикции» *MDM*-проектов лежат запросы на реализацию сложного бизнес-функционала. Такой функционал должен быть вынесен из *MDM*-решения в отдельные ИС [9], а *MDM*-проект заканчивается доставкой МД ПД.

Следует отметить, что *MDM*-специфика должна быть основной в проекте. Если этой специфики нет или она лишь частично присутствует, но является не главной, то данный ИТ-проект не является *MDM*-проектом.

3 *MDM*-решение

В результате *MDM*-проекта у организации-заказчика появляется *MDM*-решение, которое включает в себя: *MDM*-систему (программную часть решения); новые регламенты по работе с данными; обученных сотрудников организации, которые умеют использовать *MDM*-систему в соответствии с новыми регламентами; налаженный и запущенный процесс управления МД. Последний пункт важен, поскольку всё может иметься в наличии, но деятельность по *MDM* в организации отсутствует. Например, из-за нерешённых вопросов безопасности или противоречий внутренних регламентов, или в силу большой загруженности сотрудников, которые должны участвовать в *MDM*. Таким образом, *MDM*-решение является осуществляющейся в организации деятельностью, налаженной и обеспеченной всем необходимым.

MDM-система - это развёрнутое в организации готовое программное обеспечение (ПО), которое реализует основные функции *MDM* – хаб данных, консолидацию и пр. Главной частью этого ПО является базовый *MDM*-продукт (см. в [9]), а также, возможно, дополнительный набор ПО для частных вопросов (например, очистки данных). Наличие многофункционального готового ПО, которое следует лишь настроить и развернуть у организации-заказчика, существенно понижает стоимость и риски *MDM*-проекта. Однако некоторую часть

MDM-системы приходится дорабатывать в рамках *MDM*-проекта, чтобы отразить особенности организации, которые не удаётся покрыть стандартным инструментарием[4].

4 Описание модели

После первой оценки потребностей заказчика необходимо провести их детальный анализ в терминах *MDM* [8, 9]. Для этого в статье предлагается специальная функциональная модель. Она описывает типовое *MDM*-решение, включая в себя максимальный объём функциональности с тем, чтобы можно было выбрать необходимые компоненты, которые нужно реализовать в данной ситуации[5].

Для удобства использования модель представлена с помощью «метафоры» полного жизненного цикла (ЖЦ) МД и состоит из трёх пакетов (этапов): сбор данных, обработка данных и доставка данных. Пакеты содержат функциональные компоненты, каждая из которых описывает блок работ по наладке/управлению МД. Таким образом, функциональные компоненты модели включают в себя работы по наладке *MDM* и работы, выполняемые в рамках дальнейшего функционирования *MDM*-решения.

Например, при реализации одной из главных компонент модели «Консолидация данных» нужно сделать следующее:

- выполнить работы по наладке: наладить/конфигурировать/дописать ПО, поддерживающее соответствующее рабочее место аналитика, определить правила для разрешения конфликтов и выполнить первый раз консолидацию данных из ИД организации;
- осуществлять консолидацию данных в рамках дальнейшего функционирования *MDM*, поскольку далее данные из ИД будут продолжать поступать на хаб данных.

Предлагаемая модель ориентируется на внедрение новой ИТ-системы в организацию на основе готовых инструментов, которые могут быть настроены и доработаны под особенности задач организации-заказчика. Поэтому каждая компонента модели имеет *программную* и *аналитическую* части. Часть функционала компоненты выполняет соответствующее ПО, а часть – человек (аналитик). Работы по наладке компоненты целесообразно разделить на наладку/реализацию некоторого ПО и выполнение аналитических функций. Например, создание логической модели МД является аналитической функцией, а очистка данных – программно-аналитической. В последнем случае речь идёт о создании и программной реализации специальных правил очистки, которые применимы именно для этих данных, именно для этой организации, и применении этих правил, включая анализ результатов, возможно создание новых правил и т.д. При этом используется готовое ПО, но отдельные специальные правила, ориентированные на специфику данных организации, могут быть реализованы в виде дополнительного ПО, созданного в рамках *MDM*-проекта.

Основные пакеты модели.

- *Сбор данных*. В этот пакет включены компоненты, отвечающие за идентификацию данных-кандидатов в МД, так называемых «сырых» данных, а также за их дальнейший анализ и предварительную обработку. Сюда же входит доступ к различным ИД.

[4] Здесь можно привести аналогию с рынком *ERP*-решений: вначале эти решения, также как и *MDM*-решения сегодня, предназначались для больших организаций и также создавались на основе базовых *ERP*-продуктов со значительными доработками. Далее стали появляться специализированные решения для отдельных отраслей (энергетики, добывающих отраслей, логистики и пр.). Аналогично, в области *MDM* уже сейчас выделяется отдельный класс продуктов, ориентированных на работу с конкретной предметной областью (данные продукты принято называть *Product Information Management, PIM*).

[5] Предложенную модель можно также назвать идеальной моделью «*to be*» – эта терминология принята в области структурного [13] и объектно-ориентированного анализа [14], а также в области реинжиниринга бизнес-процессов [15].

- *Обработка данных***.** В этот пакет включается функционал по созданию и хранению МД в хабе данных, включая создание и поддержку логической модели данных, а также выполнение классификации, иерархизации и пр. В хаб поступают предварительно обработанные «сырые» данные, полученные из ИД. Здесь они обрабатываются, становясь МД.
- *Доставка данных***.** В этот пакет включены функциональные компоненты, отвечающие за доставку МД системам ПД. ИД и ПД могут совпадать полностью или частично. При этом оказывается важным решение вопросов разделения прав доступа к МД, а также реализация различных режимов доставки данных для ПД. Выделяют следующие режимы доставки МД: пакетный, режим реального времени и подписочный режим.

Предложенная модель ориентирована на сценарий не только одноразовой загрузки данных, а на повторяющееся обновление МД в хабе с учётом новых «сырых» данных из ИД.

Пакеты и функциональные компоненты модели представлены на рисунке 1.

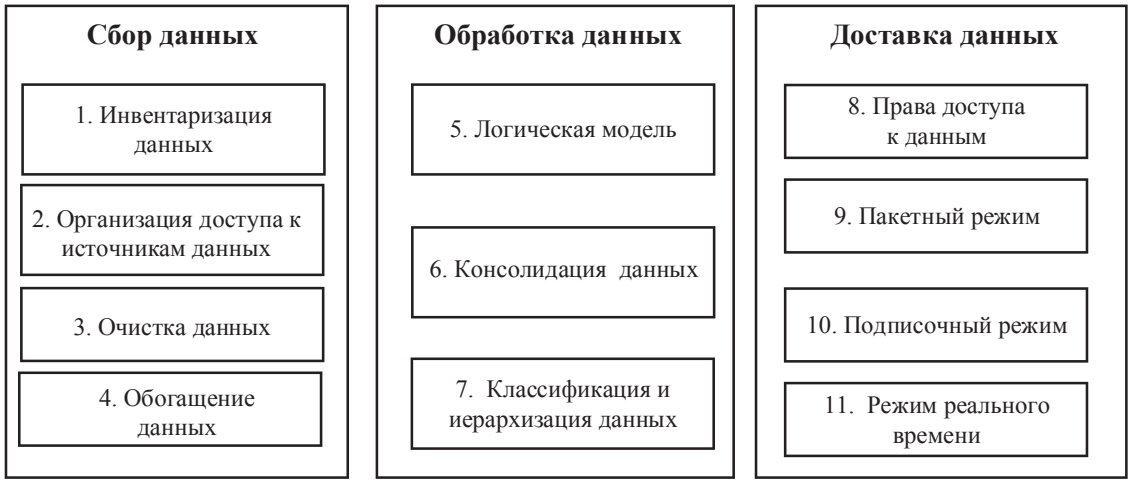

Рисунок 1 - Схема функциональной модели типового *MDM*-решения

4.1. Инвентаризация данных

В рамках этого пакета производится идентификация ИД, а также определяется, какие именно данные из этих ИД нужно преобразовывать в МД. Необходимо определить точный состав МД. Чем больше разнообразие данных, тем сложнее (и, следовательно, дороже) будет *MDM*-проект. При этом собирают лишь те атрибуты, которые будут востребованы ПД. Важно провести типизацию данных, выяснить реальную заполняемость каждого существенного атрибута, его типовые значения и пр. В этом пакете также определяется уровень доверия к различным ИД. Возможна ситуация, что некоторый ИД имеет очень низкую степень доверия, в частности, он может хранить данные, которые давно не обновлялись. Обращаться к нему следует лишь в крайнем случае. Эта функциональная компонента является преимущественно аналитической.

4.2. Организация доступа к источникам данных

Поскольку «сырые» данные, которые должны использоваться для создания МД, находятся в различных ИД организации-заказчика, то для создания *MDM*-решения необходимо организовать программный доступ к этим данным. В большинстве случаев загрузка данных является многоразовой процедурой и должна выполняться регулярно во время функционирования *MDM*-решения. Для автоматизации этой процедуры необходимо решить технические задачи, т.к. ИД часто реализованы на разных платформах и могут не иметь программных ин-

терфейсов доступа. Эта функциональная компонента является преимущественно программной. Объём работ здесь во многом зависит от того, насколько обмен данными налажен в организации (например, уже могут быть внедрены технологии передачи данных между различными ИС организации).

4.3. Очистка данных

Под очисткой данных понимается устранение ошибок в данных и нормализация данных из различных ИД перед их загрузкой в хаб. Эта компонента необходима, поскольку в противном случае будет непросто искать в данных дубликаты, а также выполнять их консолидацию. Очистка данных является трудоёмким процессом. Первичная очистка, включая нормализацию и приведение значения всех атрибутов к единому формату, не является затратной, однако она значительно упрощает дальнейшие шаги по консолидации данных и восстановлению связей между сущностями.

Эта компонента является программно-аналитической. На практике часто требуется программная доработка таких инструментов для корректной работы с различными форматами данных организации, либо реализовывать специальную логику по очистке информации. Например, ИД может хранить несколько значений в одном атрибуте, и тогда требуется разделить эти значения по соответствующим полям. Для этого нужно разработать специальное ПО, которое выполнит это разделение.

4.4. Обогащение данных

Может оказаться, что для пользователей МД требуется дополнить имеющиеся в ИД данные, например, информацией из открытых источников.

Данная компонента является программно-аналитической, поскольку требует анализа и разметки соответствующих данных из ИД, а также реализации программного доступа к публичным источникам для обогащения данных.

4.5. Логическая модель

Эта компонента предназначена для создания и сопровождения логической модели МД. Такая модель должна отражать структуру консолидированных данных со всеми атрибутами, собранными из различных ИД организации. Модель необходима для дальнейшей обработки МД, а также их доставки ПД. Одним из важнейших шагов при создании логической модели является восстановление/обнаружение различных связей в данных, которые отсутствовали в ИД, но появляются при консолидации.

Деятельность по созданию логической модели является аналитической. Она должна быть поддержана соответствующим ПО для создания такой модели, средствами визуализации, перечислением атрибутов и связей между сущностями, а также программной связи созданной мастер-модели данных с соответствующим отражением её элементов в ИД и/или ПД. При этом некоторые аспекты этого инструментария требуется дорабатывать под конкретный проект: например, в качестве ИД и ПД может выступать старая ИС, в которой модель данных жёстко задана (типичный случай – *ERP*-система), и тогда доставка новых значений для существующих атрибутов будет требовать специальной программной реализации.

4.6. Консолидация данных

Эта функциональная компонента отвечает за загрузку данных из разных ИД на хаб и выполнение консолидации реальных данных в соответствии с созданной логической моделью.

Процесс загрузки производится автоматически, с использованием соответствующих инструментов. При его выполнении возникают конфликты, которые разрешаются следующими способами.

- «Вручную» – эксперт предметной области разрешает конфликт; этот способ применяется для критических данных (например, юридических), где ошибки недопустимы и поэтому автоматические алгоритмы разрешения конфликтов неприемлемы.
- Семантический (онтологический) подход, который применяется для данных, которые хранятся в виде онтологий. Если фрагмент данных из ИД попадает с другими фрагментами в одну онтологию, то эти фрагменты являются консолидированными.
- Методы искусственного интеллекта (ИИ), в частности, методы машинного обучения, которые обучаются на типичных ситуациях с тем, чтобы разрешать возникающие в процессе консолидации конфликты автоматически.
- Смешанные стратегии, например, с помощью алгоритмов ИИ экспертам представляется на одобрение предварительные варианты разрешения конфликтов. Такой подход может снизить трудоёмкость процедуры разрешения конфликтов без снижения качества.

Загрузка данных из ИД может осуществляться одноразово, например, в случае централизованной архитектуры хаба данных или при наличии ИД, которые прекратили свою работу, но содержат ценные данные. Иначе помимо первичной загрузки требуется организовать регулярное обновление МД.

Данная компонента является программно-аналитической. Программной частью является доработка ПО деятельности аналитика по консолидации данных для работы со специфическими данными, а также для реализации уникальных правил консолидации и правил разрешения конфликтов. Если используются алгоритмы ИИ, то они должны быть адаптированы под конкретную задачу. Например, это могут быть обновляемые или самонастраиваемые правила для разбора конфликтов данных при консолидации.

4.7. Классификация и иерархизация данных

Организация нуждается в максимальной систематизации и упорядочении МД (например, основных активов, поставщиков, клиентов). Требуется разделить МД на группы, выделить значимые атрибуты для каждой группы и т.д. Необходимо связать данные организации с различными внешними классификаторами – государственными стандартами, отраслевыми классификаторами и т.д., а во многих случаях также иерархизировать данные. В качестве примера можно привести ситуацию, когда новый заказчик организации оказывается в том же организационном сегменте большой корпорации, что и предыдущий; в таком случае нет необходимости получать новое разрешение у службы безопасности, что экономит время и ресурсы. Деятельность по классификации и иерархизации данных производится над неструктурированными данными и может быть связана с обогащением данных.

Данная компонента является аналитической в части создания правил иерархизации и классификации. Она должна быть поддержана соответствующим ПО, которое позволяет эти правила отлаживать на малом количестве данных и далее применять их к полному объёму данных. Для подсказок аналитику и создания полуавтоматической классификации данных всё больше используют различные методы ИИ, в частности, методы машинного обучения.

4.8. Права доступа к данным

ПД могут находиться в различных бизнес-контурах организации и иметь разные права доступа к данным. В рамках этой компоненты требуется определить стратегию прав доступа к МД и выполнить её реализацию. Здесь необходимо полагаться на существующие в органи-

зации роли и связанные с ними права, взаимодействуя со службой безопасности организации.

Данная компонента включает работы, которые не являются трудоёмкими и оказываются преимущественно аналитическими: политика разграничения прав доступа к МД реализуется средствами администрирования ИС. Однако создание соответствующей спецификации (какие данные и кому должны быть доступны) является ответственной работой, требующей глубокого знания данных и бизнес-процессов, а также структуры организации.

4.9. Пакетный режим

Эта компонента отвечает за загрузку и обновление МД в ПД в соответствии с некоторым расписанием. Многие ПД ориентированы на получения пакетных выгрузок МД в промежуточные базы данных («витрины данных»), с которыми они работают в своём режиме. При этом каждая витрина использует, как правило, свой фрагмент мастер-модели данных. Целесообразно реализовать отдельный механизм управления такими витринами для отслеживания их своевременного обновления актуальными МД, а также для журналирования запросов к МД разными ПД. Таким образом отслеживается, какие именно МД потребляются какими ПД и в каком режиме; какие конфликты данных появляются из каких ИД и как это соотносится с потреблением данных.

Данная компонента имеет программную часть по наладке/реализации интерфейса *MDM*-системы с соответствующими ПД. Аналитическая часть заключается в определении тех ПД и тех частей МД, которые нуждаются именно в такой стратегии.

4.10. Подписочный режим

В рамках этого режима каждый ПД подписывается на определённый фрагмент МД (часть логической модели или множество сущностей и их атрибутов). Далее формируется одна или несколько очередей, куда выгружаются наиболее актуальные МД после их очередного обновления. После этого все ПД считывают свои обновления из данной очереди согласно своей подписке. Сложность реализации подписочной модели состоит в том, что необходимо либо повторно использовать существующий механизм очередей, которым уже пользуются ПД, либо доработать соответствующие ПД для использования очередей *MDM*-решения.

Данная компонента является программно-аналитической.

4.11. Режим реального времени

Данная компонента включает функционал по доставке МД ПД в режиме реального времени, т.е. непосредственно после изменения данных. Такой режим часто трудно реализовать из-за конфликтов на стороне потребителя, поскольку система ПД может временно блокировать доступ к фрагменту данных из-за выполнения некоторой операции, и это приведёт к задержкам с обновлением записи в рамках *MDM*. Например, на сайтах телекоммуникационных компаний часто есть функция проверки возможности подключения той или иной услуги по адресу. Потенциальный клиент заходит на сайт компании, вводит свой адрес (возможно с ошибками) и выбирает интересующую его услугу, например, широкополосный доступ в Интернет. Промедление с ответом сайта для такого клиента может быть опасно, т.к. клиент может уйти к конкуренту. Поэтому такая проверка адреса становится задачей поиска данного адреса (очистка от опечаток) и указанной услуги в режиме реального времени.

Данная компонента является программно-аналитической.

5 Примеры

Использование предложенной модели можно показать на примере *MDM*-проектов, выполненных при непосредственном участии авторов (см. таблицу 1).

Таблица 1 – Характеристики *MDM*-проектов

№	*Акроним*	*Название*	*Сектор индустрии*
1.	КМТР	Каталог материально-технических ресурсов	Энергетика
2.	ПК	Продуктовый каталог	Телекоммуникации
3.	КТУ	Каталог товаров и услуг	Транспорт
4.	СКБ	Сегментирование клиентской базы	Розничная торговля
5.	ЛКГ	Личный кабинет горожанина	Государственное управление
6.	ПДК	Проверка данных клиента	Энергетика и тяжелая промышленость

В описании выполненных проектов курсивом выделены компоненты функциональной модели, которые были в фокусе разработчиков.

- КМТР. Данный проект направлен на создание системы для управления данными о материально-технических ресурсах крупной организации в энергетическом секторе. Система предназначалась для решения следующих задач: обеспечить качественными данными бизнес-процессы технического обслуживания, ремонта и управления запасами; консолидировать различную информацию за счёт создания технологии стандартизации и унификации данных. В рамках проекта рассматривались данные о сырье и материалах, используемом оборудовании, запасных частях и комплектующих изделиях, необходимых для обеспечения деятельности организации. К особенностям проекта можно отнести автоматизацию сложных регламентов организации по работе с информацией, затрагивающих более десяти различных подразделений, реализацию классификатора материальных ресурсов (*классификация и иерархизация данных*), построение *логической модели данных*.
- ПК. Проект выполнялся для крупной телекоммуникационной организации и был нацелен на консолидацию информации по следующим направлениям: по продуктовым предложениям (услугам) для различных сегментов заказчиков; по проверке технических возможностей подключения услуг; по объединению финансовой информации из систем биллинга и бухгалтерской отчётности. Основной акцент был сделан на *инвентаризации данных* о продуктах компании из различных ИД, а также на создании единой *логической модели* МД с последующей *консолидацией*. Построено итоговое дерево продуктов компании с различными характеристиками, включая финансовые, для дальнейшего анализа отделом продаж и финансистами (*классификация и иерархизация данных*).
- КТУ. Основной задачей проекта была консолидация товаров и услуг, закупаемых крупной транспортной организацией. Было необходимо объединить информацию из различных классификаторов товаров и составить перечень услуг подрядчиков. Внутренним заказчиками этого *MDM*-решения стала служба закупки организации. В ходе проекта были идентифицированы по своему атрибутивному составу товары и услуги, имеющие различные цены. В результате были созданы монетарные метрики, т.е. подсчитана итоговая экономия организации по закупкам ввиду того, что требуемые товары и услуги стали закупаться по гарантированным минимальным доступным ценам автоматически. Фокус проекта был на *консолидации данных* о закупаемых товарах, *разграничении прав* и обеспечении доступа к данным в рамках *подписочного режима*.
- СКБ. Проект разрабатывался для организации, занимающейся продажей модных товаров, и предназначался для сегментирования клиентской базы и поддержки продаж в премиальном сегменте. Целью проекта было выявление клиентов из клиентской базы организации, которые активны в социальных сетях и имеют много подписчиков. Организа-

ция хотела «заручиться их лояльностью» с помощью дополнительных скидок и других мотивационных акций с целью получить больше потенциальных покупателей – подписчиков этих клиентов. В рамках проекта был сделан акцент на *обогащение* и *консолидацию данных*.

- ЛКГ. Данный проект разрабатывался для городской государственной службы управления с целью создания «умного» личного кабинета горожанина. Требовалось выполнить интеграцию личного кабинета с многочисленными ИС федерального и регионального уровней с целью извлечения профильной информации о горожанине, например, сведений о его транспортных средствах, недвижимости, банковских счетах и пр. Важными особенностями проекта была информационная безопасность и разделение *прав доступа к данным*, а также получение МД *в режиме реального времени* и в рамках *подписочной модели*.
- ПДК. Проект разрабатывался для многопрофильной международной организации из сектора энергетики и тяжелой промышленности. Организация имеет сотни тысяч клиентов в разных странах мира, поэтому процедура создания нового клиента оказывается трудоёмкой. До создания *MDM*-решения она занимала 21 день, а после – всего лишь 8 дней. *MDM*-решение позволило автоматизировать различные проверки, поиск конечных бенефициаров юридических лиц в корпоративных иерархиях, а также реализовать централизованный ввод информации. В рамках данного проекта основной акцент был сделан на *инвентаризации данных*, создании *логической модели* МД, позволившей решить задачу поиска дубликатов юридических лиц и поиска аффилированных лиц, а также реализации доступа к МД в *режиме реального времени* с целью ускорить целевой бизнес-процесс.

Для указанных проектов в таблице 2 показано, какие функциональные компоненты были реализованы в соответствующих *MDM*-проектах. При этом использовалась шкала:

- *High* – компонента является одной из основных в проекте, т.е. она бизнес-критична или технологически сложна;
- *Med*(*ium*) – компонента важна для проекта, но не является приоритетной или трудоёмкой;
- *Low* – компонента реализована в облегчённом варианте: она либо уже существует к началу проекта и требует лишь доработки, либо полная реализация компоненты была вынесена в отдельный проект;
- N/A – данная компонента в рамках этого проекта не востребована.

Таблица 2 – Описание индустриальных *MDM*-проектов в рамках предложенной функциональной модели

Элементы функциональной модели	***Проекты***					
	КМТР	ПК	КТУ	СКБ	ЛКГ	ПДК
Сбор данных						
Инвентаризация данных	Med	High	Med	Low	Med	High
Организация доступа к источникам данных	Low	Med	Low	Low	Low	Med
Очистка данных	Med	Low	Med	Med	Med	Med
Обогащение данных	N/A	Med	Med	High	n/a	Low
Обработка данных						
Логическая модель	High	High	Med	Low	Med	High
Консолидация данных	Med	High	High	High	Med	Med
Классификация и иерархизация данных	High	High	Med	N/A	N/A	High
Доставка данных						
Права доступа к данным	Low	Low	High	Low	High	Med
Подписочный режим	Med	Low	High	Low	High	Med
Режим реального времени	Med	Low	High	Low	High	Med
Пакетный режим	Med	N/A	Med	Med	High	High

6 Сопоставление существующих и описанного подходов

За последние годы создано значительное количество стандартов и методологий управления данными в организациях [2, 16, 17] и др.

Другой вид работ сфокусирован на ПО в сфере *MDM*, см., например [18]. Предлагаемый в данной статье подход отличается от этих методологий тем, что не зависит от конкретного базового *MDM*-продукта.

В [19] предложена модель для построения *MDM* в организации, которая включает семь блоков: концепцию, стратегию, метрики, информационное управление, оргвопросы и роли, ЖЦ информации, а также инфраструктуру. Эта модель предназначается для ранних стадий внедрения *MDM*, однако она ориентирована на стратегию сверху-вниз, покрывая всю деятельность организации по внедрению *MDM*. Предложенная в статье функциональная модель предназначается для итеративной стратегии, ориентированной на удовлетворение конкретных потребностей организации, которые выражаются в терминах *MDM*.

В [20] предлагается модель для анализа ЖЦ МД в организации с целью определить недостающие виды деятельности. Основными компонентами модели являются: портфолио данных; проектирование данных и системы; управление данными, поддержка данных. Эта модель слабо связана с программной частью *MDM*-решения, а также не рассматривает уникальные задачи организации по внедрению *MDM*.

Заключение

В работе предложена функциональная модель *MDM*-решения, ориентированная на разработку таких решений в рамках итеративной стратегии. Модель предназначается для ранних стадий разработки и призвана перевести потребности организации на язык *MDM* для оценки доли *MDM*-специфики. Модель позволяет выполнить планирование функционала *MDM*-решения и перейти к выбору базового инструментария, а также созданию технического задания на разработку. В работе представлен анализ выполненных *MDM*-проектов на основе предложенной модели.

В качестве дальнейшего продолжения исследования интерес представляет детальная разработка методик оценки функционала *MDM*-проектов на ранних стадиях, а также создание детальных метрик сложности *MDM*-решений. Планируется провести перевод (отображение) функциональности типового *MDM*-решения на различные *MDM*-продукты, а также осуществить более тесную интеграцию подхода с областью управления знаниями [21, 22].

Список источников

[1] ***Khatri, V.*** Designing data governance / V. Khatri V., C.V. Brown // Communications of the ACM. – 2010. – 53(1).

[2] DAMA-DMBOK: Data Management Body of Knowledge, 2017.

[3] ***Андриченко, А.Н.*** Тенденции и состояние управления справочными данными в машиностроении / А.Н. Андриченко // Онтология проектирования. – 2012. – 2 (4). – С. 25-35.

[4] ***Немцов, Э.Ф.*** ИСУЖТ и нормативно-справочные данные / Э.Ф. Немцов // Автоматика, связь, информатика. – 2020. – № 2. – С. 15-18.

[5] ***Голубев, С.С.*** Отраслевая система государственной службы стандартных справочных данных нефтегазового комплекса / С. С. Голубев, А. Н. Лоцманов, А. Ю. Кузин, В. Г. Соловьев, А. Д. Козлов, Б. А. Григорьев // Законодательная и прикладная метрология . – 2020. – № 3. – С. 12-16.

[6] ***Янченко, Г.А.*** К вопросу о стандартизации справочных данных плотностных свойств горных пород / Г.А. Янченко // Горный информационно-аналитический бюллетень (научно-технический журнал). – 2011. – № 8. – С. 111-115.

[7] ***Чигиринский, Ю.Л.*** Методика повышения надежности справочных данных / Ю.Л. Чигиринский // Известия Волгоградского государственного технического университета. – 2011. – № 13 (86). – С. 55-61.
[8] Gartner Glossary, https://www.gartner.com/en/glossary.
[9] ***Walker, S.*** Magic Quadrant for Master Data Management. / S.Walker, A. Dayley, S. Parker, M. Hawker // Gartner. – 2021.
[10] ***Zmud, R.W.*** An Examination of 'Push-Pull' Theory Applied to Process Innovation in Knowledge Work / R.W. Zmud // Management Science. – 1984. – 30 (6). – P. 727–738.
[11] ***Ladley, J.*** Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program / J.Ladley. Academic Press; 2nd edition 2019. 350 p.
[12] ***Silvola, R.*** Managing one master data – Challenges and preconditions / R. Silvola, O. Jääskelainen, H. Kropsu-Vehkaperä, H. Haapasalo//Industrial Management & Data Systems. – 2011. – 111(1). – P. 146–162.
[13] ***Yourdon, E.*** Structured Design: Fundamentals of a Discipline of Program and Systems Design / E. Yourdon, L.L. Constantine. Yourdon Press. 1975.
[14] ***Jacobson, I.*** Object-Oriented Software Engineering / I. Jacobson. ASM press. 1992. 528 p.
[15] ***Ould, M.A.*** Business Processes: Modelling and Analysis for Re-Engineering and Improvement / M. A. Ould. Wiley, 1995
[16] CMMI Data Management Maturity Model (DMM). CMMI Institute (website). http://bit.ly/1Vev9xx.
[17] IBM Data Governance Council Maturity Model https://ibm.co/2sRfBIn
[18] Velocity Methodology. Best Practices. Informatica. 2008.
[19] ***O'Kane, B.*** The Seven Building Blocks of *MDM*: A Framework for Success / B. O'Kane, M. P. Moran. August 2016. Gartner. ID: G00311161
[20] ***Ofner, M.H.*** Management of the master data lifecycle: a framework for analysis / M. H. Ofner, K. Straub, B. Otto, H. Österle // J. Enterp. Inf. Manag. – 2013. – 26(4). – P. 472-491.
[21] ***Гаврилова, Т.А.*** Представление знаний в экспертной диагностической системе АВТАНТЕСТ / Т.А. Гаврилова // Известия Академии наук СССР. Техническая кибернетика. – 1984. – № 5. – С. 165-173.
[22] ***Гаврилова, Т.А.*** Управление знаниями: от слов к делу / Т.А. Гаврилова, Д.В. Кудрявцев // Intelligent Enterprise: RE (Корпоративные системы). – 2004. – № 12-13 (101). – С. 48.

Сведения об авторах

Кузнецов Сергей Викторович, 1979 г. рождения. Окончил с отличием математико-механический факультет Санкт-Петербургского государственного университета (2001), аспирантуру (2004). Возглавлял научно-исследовательские направления в ряде компаний в области управления данными. Возглавил российское представительство компании Informatica (2010), создал компании ООО «ТаскДата» и ООО «Юнидата» (2014), разработчика одноимённой платформы управления данными. В 2020 и 2021 году система *MDM* Юнидата была отмечена в ежегодных отчётах консалтингового агентства Gartner по *MDM*. *sergey@unidata-platform.ru*.

Кознов Дмитрий Владимирович, 1971 г. рождения. Окончил с отличием математико-механический факультет Санкт-Петербургского государственного университета (1994), к.ф.-м.н. (2000), д.т.н. (2016). Профессор кафедры системного программирования Санкт-Петербургского государственного университета. Член *ACM*. Приглашенный профессор Национального университета Сингапура (Сингапур, 2000, 2001), Технического университета Лаапентанты (Финляндия, 2015), Южно-Китайского Технологического университета (Китай, 2018). Автор более 100 работ в области программной инженерии информационных систем. Author ID (RSCI): 126261. Author ID (Scopus): 8885649400. *d.koznov@spbu.ru*.

Поступила в редакцию 23.04.2021, после рецензирования 7.05.2021. Принята к публикации 14.06.2021.

Master data management in an iterative approach

S.V. Kuznetsov[1], D.V. Koznov[2]

[1]*Unidata LLC, Saint-Petersburg, Russia*

[2]*Saint-Petersburg State University, Saint-Petersburg, Russia*

Abstract

Master Data Management (MDM) is a young area of business informatics that concerns consolidation and centralized control of highly important business data distributed over different information systems. Leading IT companies such as IBM, Oracle, Informatica and others offer a wide range of ready-made products for master data management (MDM products). MDM product deployment involves many technical and organizational complications: it is necessary to adapt these products for the specifics of the business, modify business processes, create new data policies, solve security questions, etc. A popular approach to this task is the iterative strategy of MDM deployment, which supposes a step-by-step implementation of master data management based on the real needs of the business organization. In this paper, the notion of an MDM solution is introduced, which is the result of the deployment of MDM in an organization. It includes a specifically adapted MDM product, new regulations for working with data, trained employees, and an up-and-running process of master data management. The main result of the paper is a functional model of master data management intended for the early stages of the development of an MDM solution within the iterative deployment strategy. The purpose of this model is the representation of real business needs in terms of MDM. It is important to understand which MDM components should be implemented first. The paper describes a detailed description of the model components, as well as a portfolio of six real MDM projects analyzed from the viewpoint of the proposed model.

Key words: *master data management, business informatics, applied ontology, iterative approach, information systems.*

Citation: *Kuznetsov SV, Koznov DV.* Master data management in an iterative approach [In Russian]. *Ontology of designing*. 2021; 11(2): 170-184. DOI: 10.18287/2223-9537-2021-11-2-170-184.

List of figures and tables

References

[1] ***Khatri V, Brown CV.*** Designing data governance. *Communications of the ACM*. 2010; 53(1).
[2] DAMA-DMBOK: Data Management Body of Knowledge, 2017.
[3] ***Andrichenko AN.*** Tendencies and condition in the field of reference data management in the engineering industry [In Russian]. *Ontology of designing*. 2012; 2 (4): 25-35.
[4] ***Nemtsov EF.*** Reference data in the intelligent railway transport management system [In Russian]. *Automation, communication and Informatic*. 2020; 2: 15-18.
[5] ***Golubev SS, Lotsmanov AN, Kuzin AY, Soloviev VG, Kozlov AD, Grigoriev BA.*** The branch system of the National Standard Reference Data Service for the Oil and Gas Complex [In Russian]. *Legislative and applied metrology*. 2020; 3: 12-16.
[6] ***Yanchenko GA.*** On the standard reference data density properties of rocks [In Russian]. *Mining informational and analytical bulletin*. 2011; 8: 111-115.
[7] ***Chigirinsky YL.*** Methodology for improving the reliability of reference data [In Russian]. *Izvestia Volgogradskogo universitets*. 2011; 13 (86): 55-61.
[8] Gartner Glossary, https://www.gartner.com/en/glossary.
[9] ***Walker S, Dayley A, Parker S, Hawker M.*** Magic Quadrant for Master Data Management.Gartner. 2021.
[10] ***Zmud RW.*** An Examination of 'Push-Pull' Theory Applied to Process Innovation in Knowledge Work. *Management Science*. 1984; 30 (6): 727–738.
[11] ***Ladley J.*** Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program. Academic Press; 2nd edition 2019. 350 p.
[12] ***Silvola R, Jääskelainen O, Kropsu-Vehkaperä H, Haapasalo H.*** Managing one master data – Challenges and preconditions. *Industrial Management & Data Systems.* 2011; 111(1): 146–162.

[13] ***Yourdon E, Constantine LL.*** Structured Design: Fundamentals of a Discipline of Program and Systems Design. Yourdon Press. 1975.
[14] ***Jacobson I.*** Object-Oriented Software Engineering. ASM press. 1992. 528 p.
[15] ***Ould MA.*** Business Processes: Modelling and Analysis for Re-Engineering and Improvement. Wiley, 1995
[16] CMMI Data Management Maturity Model (DMM). CMMI Institute (website). http://bit.ly/1Vev9xx.
[17] IBM Data Governance Council Maturity Model https://ibm.co/2sRfBIn
[18] Velocity Methodology. Best Practices. Informatica. 2008.
[19] ***O'Kane B, Moran MP.*** The Seven Building Blocks of *MDM*: A Framework for Success. August 2016. Gartner. ID: G00311161
[20] ***Ofner MH, Straub K, Otto B, Österle H.*** Management of the master data lifecycle: a framework for analysis. *J. Enterp. Inf. Manag.* 2013; 26(4): 472-491.
[21] ***Gavrilova TA.*** Knowledge presentation in expert system ABTAHTECT [in Russian]. Izvestia Acadimii Nauk SSSR. 1984; 5: 165-173.
[22] ***Gavrilova TA, Kudryvtsiv DV.*** Knowledge management: from words to business [in Russian]. *Intelligent Enterprise: RE*. 2004; 12-13(101): 48.

About the authors

Sergey Viktorovich Kuznetsov (b. 1979) graduated with honors from the Faculty of Mathematics and Mechanics at Saint Petersburg State University (2001). He was the head of R&D departments in several data management startups, and the chief of the Russian representative office of Informatica (2010). Founded TaskData, LLC and Unidata, LLC, concentrated on development of data management platform (Unidata platform, released in 2014). Unidata master data management system was acknowledged by Gartner in their prominent annual Magic Quadrant reports on *MDM* (2020, 2021). *sergey@unidata-platform.ru*

Dmitry Vladimirovich Koznov (b. 1971) graduated with honors from the Faculty of Mathematics and Mechanics at Saint Petersburg State University (1994), received PhD degree (2000), and a Doctor of Technical Sciences degree (2016). Professor of the Software Engineering Department at Saint Petersburg State University since 2016. Member of the ACM. Visiting professor of the National University of Singapore (Singapore, 2000, 2001), Lappeenranta University of Technology (Finland, 2015), and South China University of Technology (China, 2018). Author of over 100 papers in the field of software engineering of information systems, including more than 30 Scopus-indexed articles. Author ID (RSCI): 126261. Author ID (Scopus): 8885649400. *d.koznov@spbu.ru*.

Received April 23, 2021. Revised May 7, 2021. Accepted June 14, 2021.