2. *Безгубова Ю.К., Гучук В.В.* Технология упреждающей критериальной адаптации в мониторинге и управлении сложными научно-техническими объектами / Материалы 7-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD-2013, Москва). Т. 2. – М.: ИПУ РАН, 2013. – С. 420-422.

3. *Кроль В.М., Виха М.В.* Психофизиология. – М.: КноРус, 2014. – 512 с.

4. *Guchuk V.V.* Application of algorithms of objectifying expert clustering of Multiparameter objects in the analysis of big arrays of information//Advances in Systems Science and Applications. – 2018. – Vol 18. № 1. – P. 102-109.

5. *Меньшиков В.А., Рудаков В.Б., Сычев В.Н.* Контроль качества космических аппаратов при отработке и производстве. – М.: Машиностроение-Полет, 2009. – 400 с.

6. *Гучук В.В.* Проектирование человеко-машинного интерфейса для систем испытания сложных научно-технических объектов//Актуальные проблемы гуманитарных и естественных наук. – 2014. – №12. – С. 68-75.

7. *Нестеров В.С., Гучук В.В., Рябых В.Ю.* Технологические аспекты организации процесса многоканальной регистрации разночастотного потока данных / Труды международной научно-практической конференции «Передовые информационные технологии, средства и системы автоматизации и их внедрение на российских предприятиях». – М.: ИПУ РАН, 2011. – С. 629-637.

Мухина А.Е.

Современные вопросы оценки качества данных в IT-экосистеме

Аннотация: Описаны основные проблемы, связанные с плохим качеством данных. Определены ключевые аспекты, влияющие на качество данных, а также проанализированы существующие системы, обеспечивающие повышение качества данных в IT-экосистемах.

Ключевые слова: качество данных, ISO/TS 8000, Master Data Management (MDM) systems

Крупные компании зачастую становятся заложниками быстрого и итеративного развития внутренней IT-экосистемы. Системы постоянно развиваются, количество интеграций увеличивается, и, кажется, что цели внедрения выполняются – компания получает продукт, который поддерживает тот или иной бизнес-процесс. К сожалению, зачастую внедрение системы несет куда более разрушительное влияние, чем её отсутствие. Если архитектурные ошибки во внутренней экосистеме можно нивелировать ценой больших затрат, то ошибки, связанные с данными, которые в течение многих лет складировались в базы данных, не опираясь на те или иные паттерны в обеспечении их качества, устраняются в течение многих лет, принося с собой потери в финансовых показателях компании. Согласно статистике Гартнера [1], среднегодовые финансовые потери от плохого качества данных составляют около 15 миллионов долларов. При этом около 60% организаций не измеряют финансовые потери или возможные прибыли, связанные с качеством данных [2].

Качество данных, согласно стандартам серии ISO/TS 8000, ГОСТ Р 56214–2014/ISO/TS 8000-1:2011 – это свойство данных удовлетворять предъявляемым к ним требованиям – полнота, точность, своевременность, происхождение, эффективность, доступность, переносимость, восстанавливаемость, конфиденциальность и др. Процессы управления качеством данных (в соответствии с ГОСТ Р 56215–2014 и ISO/TS 8000-150:2011) делят на три группы [1]:

- процессы выполнения операций над данными,
- процессы непрерывного контроля качества данных
- процессы повышения качества данных.

С точки зрения управления первичные данные принято делить на 4 класса:

- Мастер-данные (master-data) – особо ценные данные, ключевые для бизнеса. Эти данные крайне редко подвержены изменениям.
- Разделяемые справочники (reference data) – справочники, которые классифицируют остальные данные. Например, товары разделяются по товарным группам. Разделяемые справочники также принято называть нормативно-справочной информацией (НСИ).

- Транзакционные данные – те данные, которые отражают любые действия с объектами мастер-данных.
- Исторические данные – срез по всем трем классам данных, который возник после завершения соответствующих бизнес-процессов.

Обеспечение качества данных каждого класса реализуется следующими подсистемами:

- Платформы для интеграции данных – интеграция с системами/сервисами, а также поддержка ETL-процессов и загрузка информации в хранилища. На текущий момент разработаны масштабируемые инструменты, которые упрощают работу с проектированием, разработкой и выполнением заданий по перемещению и преобразованию данных [3].
- Модули классификации данных – позволяют идентифицировать информацию на этапе входа в экосистему, и маскировать конфиденциальные данные, такие как номера кредитных карт, паспортные данные и т.д.
- Системы управления мастер-данными (Master Data Management, MDM), а также системы управления справочными данными (Reference Data Management, RDM) – системы, которые составляют основу для управления НСИ.

MDM-системы зачастую составляют ядро комплексных систем по повышению качества данных в системах. Они позволяют не только подготавливать полученные данных из различных точек входа или интеграций, но и верифицировать их автоматически, или с помощью специально выделенной роли – Дата Стюарта. К примеру, если при первоначальном вводе на форме отсутствует маска для ввода страны, то система может легко распознать, что «GB», введенная в поле «Страна», является сокращением от «Great Britain», и перед сохранением, привести значение к той форме, которая соответствовала бы «золотой записи». «Золотая запись – это единая, точно определённая версия всех объектов, данных в экосистеме организации. Золотую запись можно назвать «единой версией истины», под которой подразумеваются те факты, к которым пользователи данных могут обратиться, когда хотят быть уверенными, что используют правильную информацию.

MDM-системы могут определять, является ли новая запись дублем уже имеющейся в системе записи. Если процент совпадения атрибутов такой записи очень высок, то система может автоматически соединить эти записи, во избежание появления дублей в хранилище/системе. Если же процент совпадения не достигает границы, то в современных решениях запрос на проверку и верификацию такой записи отправляется «Дата Стюарту», который может вручную соединить дубли, выбрав те атрибуты, которые он хочет обновить.

Часто случается, что обе записи, которые являются дублями, активно участвуют в бизнес-процессах компании. Существует большое количество дочерних сущностей, которые нельзя потерять при дедубликации записей. Для этого в MDM-решениях предусмотрена возможность переопределить дочерние объекты на новую запись при проведении операции слияния дублей. Это значительно сокращает риски потери данных. Все функции системы по управлению мастер-данными представлены на рисунке 1.

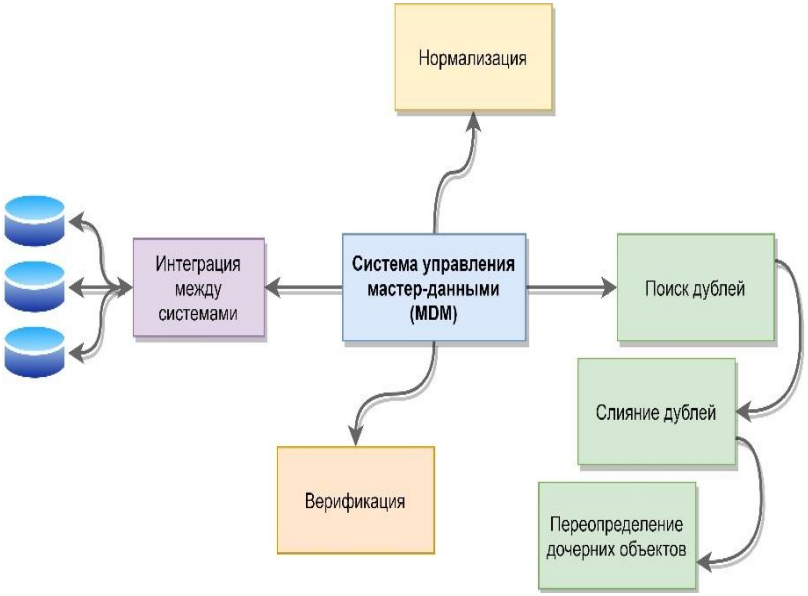

Рисунок 1 – Функции систем по управлению мастер-данными

Все качественные данные должны в конечном итоге становиться частью знаний. Это играет ключевую роль в бизнесе, так как позволяет эффективно строить интеллектуальную экосистему, а не просто автоматизировать бизнес-процессы. Существует большое количество специальных решений, которые позволят не только максимально улучшить качество уже существующих данных в экосистеме, но и предотвратить появление нерелевантной информации и далее, при расширении интеграций, наращивания объемов, хранимых данных. Очень важно задумываться о качестве данных на самых ранних этапах создания программных продуктов, так как зачастую исправить ошибки в данных гораздо сложнее и затратнее, чем предотвратить их.

Литература:

1. How to Create a Business Case for Data Quality Improvement [Электронный ресурс]. – URL: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/ (дата обращения 02.11.2020).

2. *Паклин Н.Б.* Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.

3. *Лобицын В.Н.* Повышение качества данных в контексте современных аналитических технологий//Вестник Южно-Уральского государственного университета. – 2012. – № 23 (282). – С. 83-86.