Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

Плюс 7 ФормИТ DQ on Hadoop – решение по управлению качеством данных

Решения линейки Плюс7 ФормИТ реализованы на единой промышленной платформе. Ниже приведено описание модуля по управлению качеством данных при работе в среде Hadoop – Плюс7 ФормИТ DQ on Hadoop.

Плюс7 ФормИТ – промышленная платформа интеграции данных, позволяющая эффективно решать любые стоящие перед компанией задачи в области интеграции, синхронизации данных, построения хранилищ данных, миграции данных в новые приложения, обмена информацией с контрагентами и т.д. для реализации бизнес-целей компании. Продукт Плюс7 ФормИТ зарегистрирован в реестре российского ПО: регистрационная запись №15077 от 03.10.2022.

Модуль Плюс7 ФормИТ DQ on Hadoop – компонент платформы Плюс7 ФормИТ, специализированное решение для мониторинга и управления качеством данных в рамках предприятия в среде Hadoop.

Плюс7 ФормИТ DQ on Hadoop использует мощные интеграционные возможности платформы ФормИТ и Hadoop для доступа к источникам и приемникам данных, а также для высокопроизводительной обработки потоков данных.

Возможности Плюс7 ФормИТ DQ on Hadoop

Плюс7 ФормИТ DQ on Hadoop позволяет на уровне настраиваемых бизнес-правил проводить анализ качества данных, стандартизировать и очищать данные, распознавать и выявлять дубликаты для очищенных и стандартизованных данных, осуществлять консолидацию данных, вести мониторинг и получать отчетность по качеству данных. Использование Плюс7 ФормИТ DQ on Hadoop позволяет организовать эффективную совместную работу бизнеса и IT.

Плюс7 ФормИТ DQ on Hadoop обладает следующими ключевыми возможностями:

- широкие возможности анализа данных, разбора, очистки, стандартизации, обогащения, сопоставления (выявления потенциальных дублей) данных с использованием специализированных компонентов;
- удобный и эффективный графический интерфейс разработчика, отсутствие программирования;
- профилирование данных, включая возможности drill-down по данным на каждом этапе обработки, после каждой трансформации;

- наличие визуальных средств для работы бизнес-пользователей по получению отчетности по качеству, обработке некачественных данных, работы с дублями, создания своих правил, совместной эффективной работы бизнеса и ИТ (веб-приложение ФормИТ Консультант);
- возможность подключения и использования любых словарей (ГАР, ФИО, названия компаний, префиксы и т.д.), приведения данных к эталонным словарям, проверку на вхождение/невхождение в эталонные словари;
- ведение эталонных словарей бизнес-пользователями;
- возможность ведения мониторинга и получения отчетности по качеству;
- возможности проактивного мониторинга и получения уведомлений о наступлении тех или иных событий;
- возможность обработки больших объемов данных и высокий уровень; масштабируемости;
- полная поддержка российских платформ Hadoop (Аренадата, Ростелеком и др.);
- интеграция с решениями классов DG, MDM, ETL, ERP, CRM и другими.

Плюс7 ФормИТ DQ on Hadoop работает с любыми языками. Русский язык поддерживается без ограничений. В России и странах СНГ реализован ряд успешных проектов.

Плюс7 ФормИТ DQ on Hadoop содержит в себе множество мощных и гибких возможностей для сопоставления и стандартизации данных при помощи алгоритмов вероятностного и нечёткого поиска (fuzzy logic), которые позволяют архитекторам и аналитикам определять связи между записями, обнаруживать и устранять дубликаты записей в целях проведения унификации данных.

Все функции обеспечения качества данных (очистка, стандартизация и т.д.) встраиваются в процессы интеграции данных Плюс7 ФормИТ. Используя инструмент разработки ФормИТ Разработчик, можно создать проверку качества данных непосредственно внутри интеграционного процесса (см. рис. 1). В ФормИТ Разработчик существует набор трансформаций, встроенных в Плюс7 ФормИТ, позволяющих вызывать и использовать процессы обеспечения качества данных Плюс7 ФормИТ DQ on Hadoop.

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

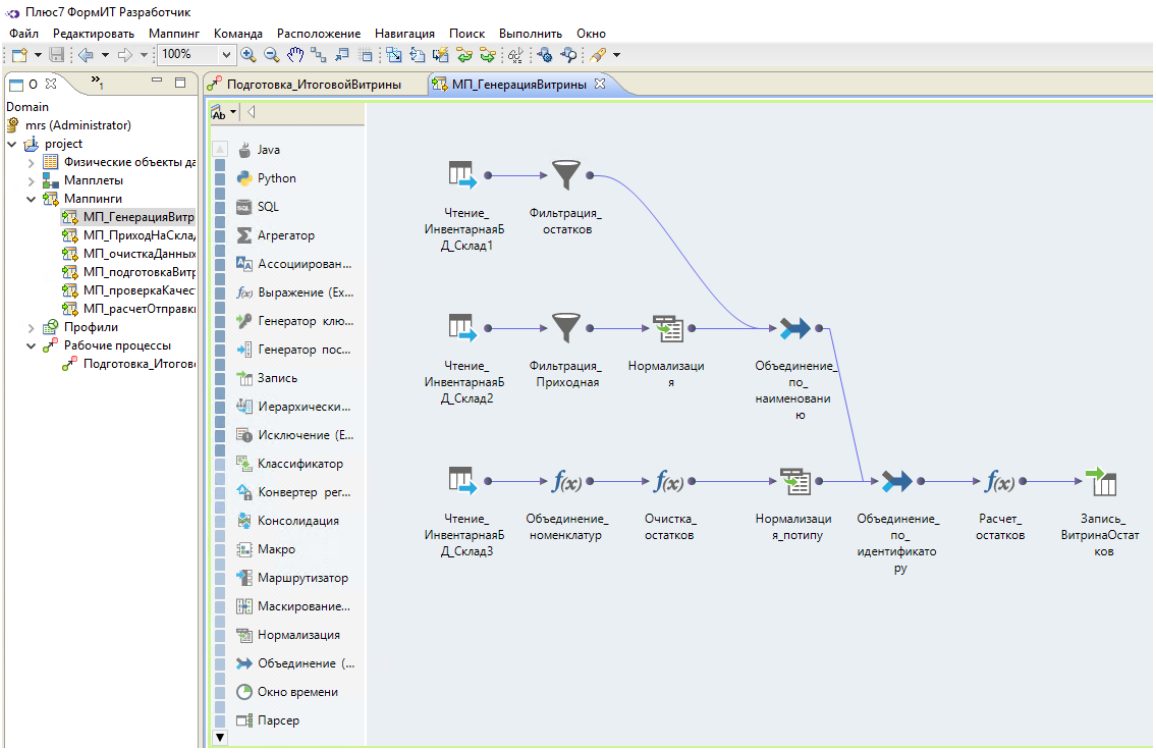

Рис. 1. Процедуры обеспечения качества данных в интерфейсе ФормИТ Разработчик

Благодаря бесшовной интеграции Плюс7 ФормИТ DQ on Hadoop получает все преимущества интеграционной платформы Плюс7 ФормИТ (включая модуль ФормИТ on Hadoop): доступ к любым источникам данных, возможно обработки данных в пакетном и онлайн режимах.

Процессы обработки качества данных могут быть встроены в общий поток загрузки данных на Hadoop. При этом всегда возникает вопрос как обрабатывать некачественные данные. Автоматическая обработка и повышение качества данных не всегда возможны. Иногда в поле отсутствует необходимое количество информации для восстановления. Также иногда невозможно бывает автоматически определить наличие дубликатов записей, что требует вмешательства эксперта.

Специально для таких случаев в составе Плюс7 ФормИТ DQ on Hadoop предусмотрено средство для ручной обработки данных (Human Task), которое позволяет выгружать по задаваемым правилам записи, не прошедшие автоматическую очистку для их ручной обработки. Ручная обработка записей производится в интерфейсе ФормИТ Консультант, представленном на рис. 2Рис. . Процесс ручной обработки данных является неотъемлемой частью общего процесса интеграции данных.

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

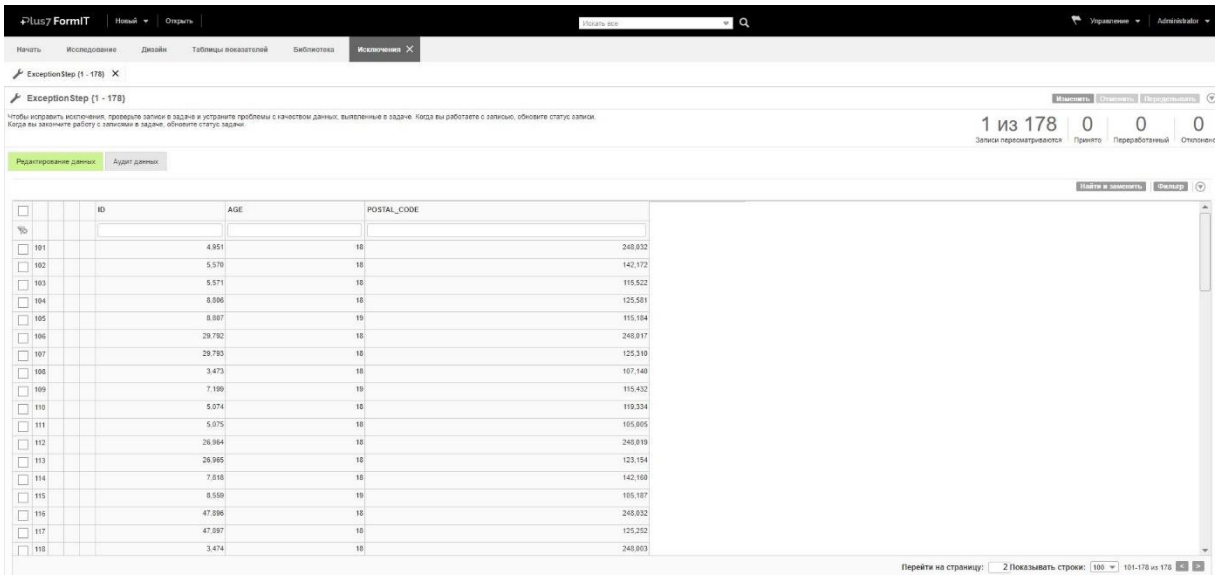

Рис. 2. Обработка некорректных данных в интерфейсе ФормИТ Консультант

Плюс7 ФормИТ DQ on Hadoop использует для обеспечения качества данных открытые текстовые словари, которые могут быть легко созданы или модифицированы собственными специалистами заказчиков. Продукт позволяет для анализа и стандартизации данных, а также для построения правил проверки качества данных одновременно использовать собственные или внешние словари.

ФормИТ DQ on Hadoop в интерфейсе ФормИТ Консультант предоставляет визуальное средство ведения словарей и референсных таблиц, позволяющее пользователям выполнять следующие основные действия с данными:

- автоматически создавать референсные таблицы и словари по существующим и новым справочникам;
- создавать референсные таблицы и словари на основе результата профилирования данных;
- обеспечивать ручное ведение справочных данных;
- автоматически вести историю изменения справочников;
- вести обработку данных на любых языках;
- выделять мастер-значения для процессов очистки данных (например, Алексей для значений Алексей, Леша, Алеша, Леха и т.д.).

Пример интерфейса ведения справочников приведен на рис. 3.

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

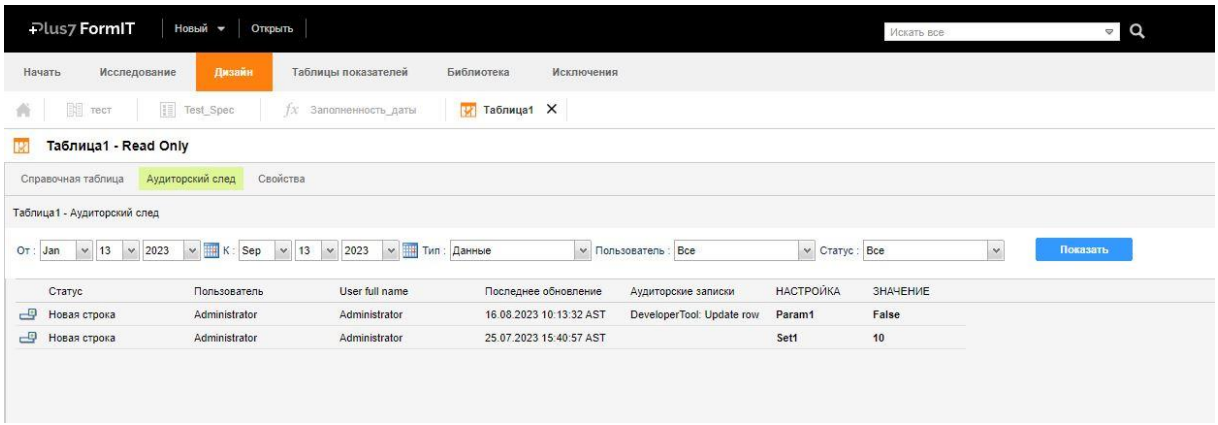

Рис. 3. Управление справочниками данных в интерфейсе ФормИТ Консультант

Благодаря всем этим возможностям Плюс7 ФормИТ DQ on Hadoop позволяет бизнес-специалистам управлять решениями по проверке и обеспечению качества данных, что, в свою очередь, позволяет серьезно снизить риски в работе бизнеса. В тоже время, это программное обеспечение повышает продуктивность разработчиков и уменьшает влияние качества данных на ресурсы при разработке и внедрении интеграционного проекта.

Использование промышленного решения позволяет в сжатые разработать максимально эффективное, гибкое решение и обеспечить возможность использования результатов в будущем во множестве других проектов. Применение промышленного решения позволяет исключить зависимость влияния количества данных на общее время выполнения проекта и минимизировать необходимость вмешательства в процесс операторов.

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

Пользовательский интерфейс ФормИТ Консультант

В составе платформы Плюс7 ФормИТ есть специализированные средства, позволяющие проводить быстрый и эффективный анализ данных в источниках.

В частности, платформа предоставляет следующие возможности:

- платформа обеспечивает профилирование, включая статистический анализ, анализ наличия шаблонов заполнения, анализ на соответствие заданным доменам данных и анализ на выявление скрытых связей между объектами данных;
- анализ данных проводится в визуальной интуитивно понятной среде без необходимости написания каких-либо скриптов вручную.

Пример результатов профилирования, которые можно получить буквально «за два клика», приведен на рис. 4.

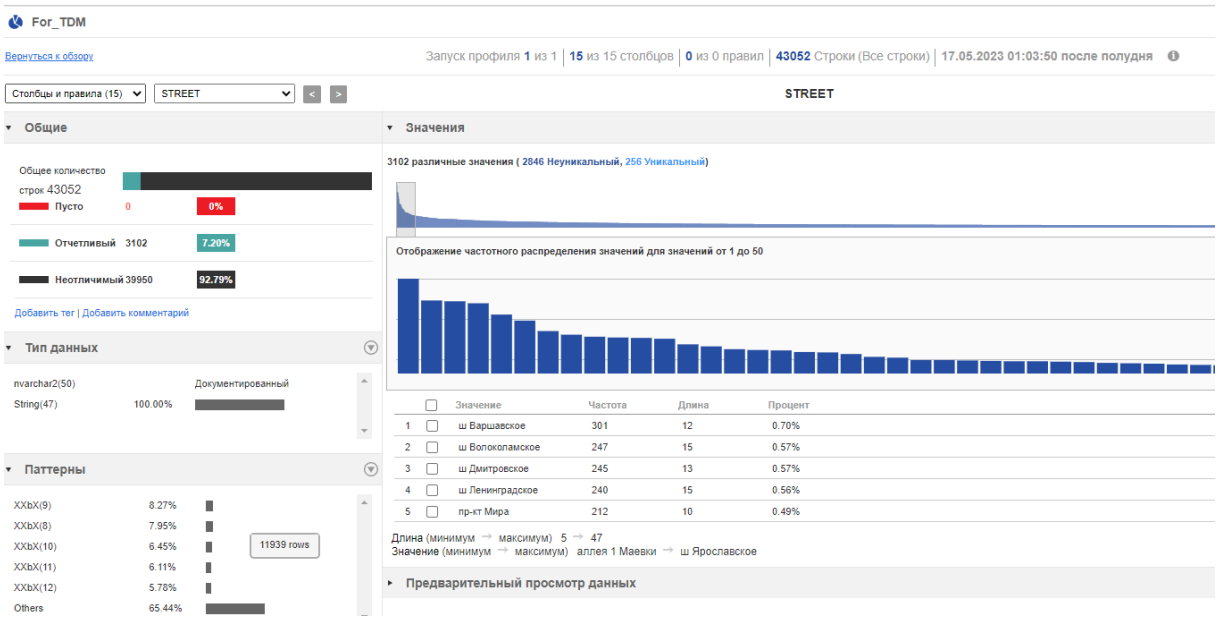

Рис. 4. Пример профилирования данных в ФормИТ Консультант

Данные средства позволяют существенно сократить время изучение и достижение необходимого уровня понимания данных в источниках данных для последующей разработки или для постановки корректной задачи по обработке данных. В конечном счете, данные средства обеспечивают сокращение сроков внедрения решения и затрат на разработку.

ФормИТ Консультант позволяет бизнес-пользователям применять готовые правила качества данных, реализованные средствами Плюс7 ФормИТ и встроенных

модулей, а также создавать свои бизнес-правила и применять их для оценки качества данных. Интерфейс для создания бизнес-правил приведен на рис. 5.

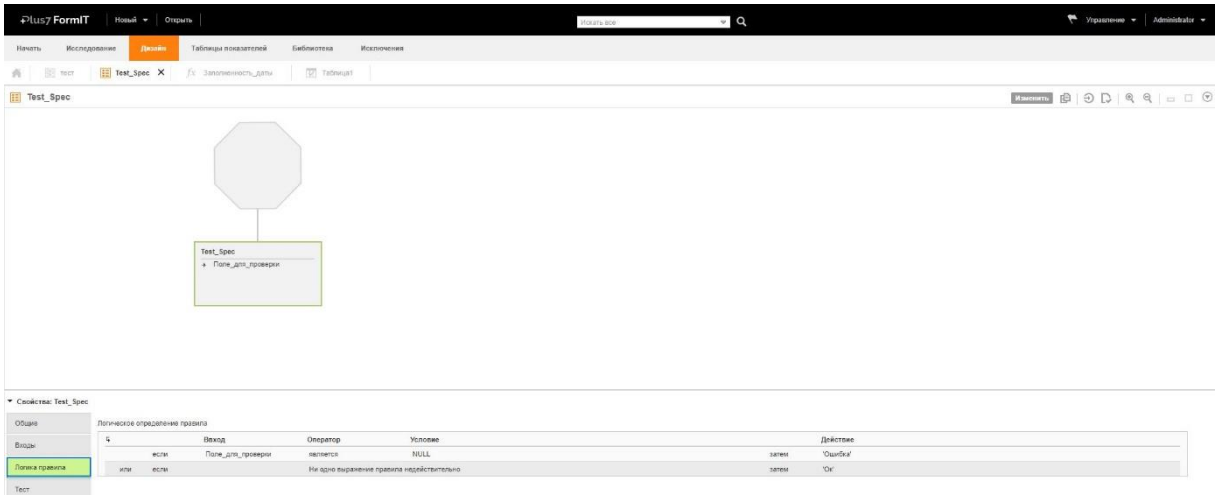

Рис. 5. Создание бизнес-правил в ФормИТ Консультант

Созданные один раз правила могут быть опубликованы и использованы разными бизнес-пользователями для оценки качества любых бизнес-данных.

Плюс7 ФормИТ DQ on Hadoop включает в себя функции мониторинга и динамического формирования отчетности о качестве данных, графические средства, которые отражают ключевые характеристики качества данных, такие как полнота, согласованность, связность, точность, целостность и отсутствие дубликатов вне зависимости от типов источников данных. Интерфейсом для предоставления данной отчетности также является ФормИТ Консультант. На рис. 6 приведен пример интерфейса мониторинга качества данных.

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

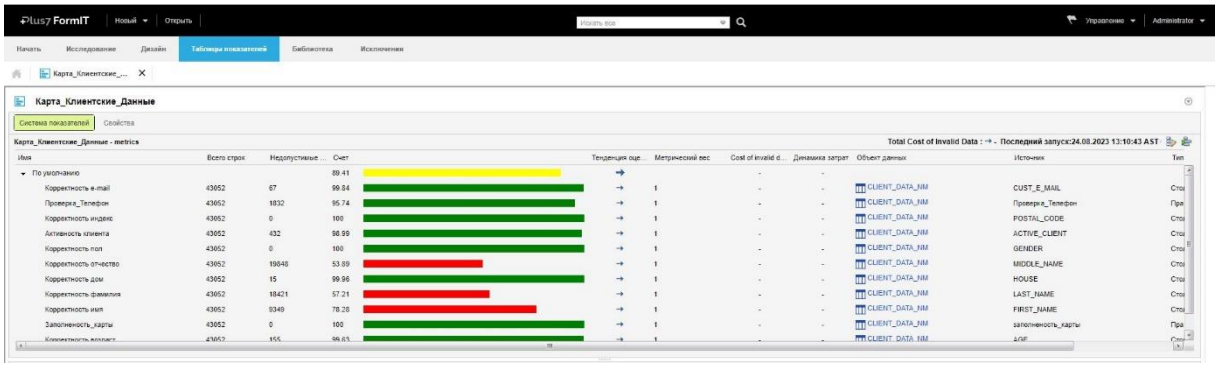

Рис. 6. Мониторинг качества данных в ФормИТ Консультант

Пользователи могут создавать графические панели, которые затем можно экспортировать в BI-системы или размещать на внутренних порталах компании, что позволит пользователям постоянно следить за качеством данных.

Ключевые преимущества Плюс7 ФормИТ DQ on Hadoop

Ключевые возможности Плюс7 ФормИТ DQ приведены ниже:

- высокая эффективность при применении в проектах по управлению НСИ, data governance, созданию тестовых сред, а также в интеграционных, миграционных проектах;
- полная поддержка российских сборок Hadoop;
- детальное и частотное профилирование любых типов данных;
- разработка правил проверки и обеспечения качества данных силами и бизнес-пользователей, и разработчиков;
- любые типы проверок и обеспечения качества данных для реализации с помощью инструмента;
- возможность повторного использования ранее разработанных правил;
- выявление дубликатов с помощью строгих, вероятностных методов и «нечеткой» логики;
- поддержка любых языков;
- онлайн-мониторинг уровня качества данных;
- высокая производительность и масштабируемость;
- возможности сетевой многосерверной обработки данных с автоматическим управлением и восстановлением работоспособности;
- визуальная среда разработки, отсутствие необходимости программирования;
- простота обучения работе с продуктом (стандартный курс – 3 дня);

Общество с ограниченной ответственностью «Дата Интегрейшн Софтвер»
125284, Россия, г. Москва, Ленинградский проспект, 31А, стр. 1, 6 этаж
ИНН 7713555858, ОКПО 77352347, ОКАТО 452777568000
Тел.: + 7 (495) 645-0201, Факс: + 7 (495) 645-0188

- простота поддержки и сопровождения – техническая поддержка оказывается в режиме 24x7;
- обеспечение прозрачности выполняемых преобразований и самодокументируемость разработок;
- возможность обработки неструктурированной и слабоструктурированной информации;
- работа в любом режиме работы – пакетном, по расписанию, в реальном времени;
- наличие серьезной экспертизы в России и странах СНГ.