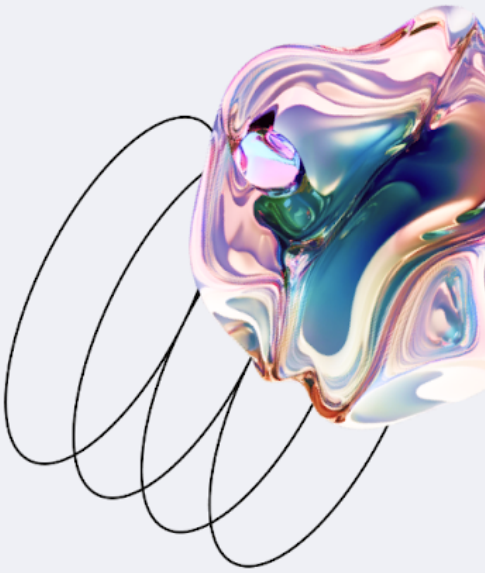
GeekBrains



Специфика применения ETL в различных предметных сферах

ETL: автоматизация подготовки данных

Оглавление

Введение

Всем привет! Сегодня у нас последняя лекция курса «ETL: автоматизация подготовки данных».

На прошлых занятиях мы разобрали, что такое Apache Airflow и как его установить. Рассмотрели типы операторов и задач, применимых в работе с Airflow. Научились строить пайплайны полного цикла ETL.

Сегодня поговорим о том, как ETL используется в различных сферах и как частичное построение ETL-системы может помочь бизнесу решать повседневные задачи.

Специфика применения ETL в различных предметных сферах

Множество компаний выбирают ETL для перемещения и обработки данных. Вот несколько примеров использования:

- **Розничная компания** может использовать ETL, чтобы извлечь данные о продажах из систем своих торговых точек, преобразовать их (очистить и стандартизировать), а затем загрузить в хранилище для анализа и составления отчетов.
- **Медицинская компания** может использовать ETL, чтобы извлечь данные о пациентах из нескольких систем электронных медицинских карт, преобразовать их (удалить повторяющиеся записи и обеспечить согласованность данных), а затем загрузить в централизованное хранилище для управления здоровьем населения.
- **Финансовая компания** может использовать ETL, чтобы извлечь торговые данные с нескольких бирж, преобразовать их (путем расчета торговой статистики и агрегирования ее по ценным бумагам), а затем загрузить в базу данных для управления рисками и составления отчетов о соответствии.
- **Социальная сеть** может использовать ETL, чтобы извлечь данные пользователей из нескольких источников — мобильных, веб- и сторонних приложений, преобразовать их (удалить конфиденциальную информацию и стандартизировать формат), а затем загрузить их в озеро данных для аналитики и машинного обучения.

Это лишь несколько примеров того, как компании используют ETL для перемещения и обработки данных. Есть множество других отраслей и вариантов использования.

ETL для науки о клинических данных / здравоохранения

Ответы на вопросы, связанные со здоровьем, требуют глубокого понимания сложной природы данных в отрасли здравоохранения, а также того, как данные из исходной системы организованы в целевой базе данных.

ETL в отрасли здравоохранения нужен для экспорта данных из одного источника (обычно из электронной медицинской карты) и для преобразования в форму, совместимую с целевой базой данных.

Электронные медицинские карты — важный источник информации в отрасли здравоохранения. Практикующим врачам и исследователям они дают возможность улучшить результаты лечения и качественнее принимать решения, связанные со здоровьем.

ETL в здравоохранении может быть как простым, так и сложным. Например:

- **простым** как объединение данных из нескольких отделений внутри клиники;
- **сложным** как интеграция данных из большого количества систем в общие модели данных (CDM): например, модели общих государственных медицинских учреждений — для обмена знаниями и исследовательской работы.

На стадии **извлечения** определять подходящие поля в исходных данных (медицинских картах или данных анализов) должны люди с экспертизой в этой области. Это нужно, чтобы корректно заполнить целевую базу данных, извлечь и обеспечить маппинг между исходными данными и целевыми элементами данных.

В контексте клинической науки интеграция данных из разрозненных источников — сложная задача, требующая нескольких итераций в процессе ETL. Могут быть трудности из-за неточного отображения, длительного выполнения запросов и проблем с качеством данных. Частая причина неверных сопоставлений — конфликты совместимости между исходными данными и целевой системой. Например, в исходной БД может быть другое представление данных, словари, термины для элементов данных и уровни детализации.

Жизненный цикл ETL

Процесс ETL состоит из нескольких итераций. Среди них спецификация ETL, извлечение и проверка данных, создание правил ETL, генерация запросов (с помощью SQL), тестирование и отладка, а также создание отчетов о качестве данных.

1. Спецификации ETL — это документ, в котором собрана информация для разработки ETL-скриптов.

2. Проверка данных — процесс обеспечения того, что данные непротиворечивые, правильные и полные.

3. Создание правил и генерация запросов включает создание правил для извлечения данных и реализацию правил использования данных (чаще всего с помощью SQL).

4. Тестирование и отладка обеспечивают точность, надежность и согласованность данных в конвейере, включая хранилище данных и этапы миграции. Измеряя эффективность процесса ETL, мы можем найти узкие места и убедиться, что процедура готова к масштабированию, если объем данных увеличится.

5. В отчетах о качестве данных указываются любые недостатки, обнаруженные в процессе ETL. Это нужно, чтобы обеспечить их целостность.

   Для оценки качества медицинских данных они считаются многомерными — для более точной оценки влияния составных частей.

Возможные проблемы с данными в здравоохранении могут быть связаны с доступностью, достоверностью, свежестью, актуальностью, полнотой, непротиворечивостью, надежностью и целостностью. Качество клинических данных — критичный вопрос, поскольку влияет на принятие врачебных решений и надежность исследований.

ETL-инструменты

Ранее мы говорили только об Apache Airflow. Но это не единственный доступный инструмент построения ETL.

Инструменты ETL — это технологические решения, предназначенные для облегчения ETL-процессов. При правильном использовании технологии ETL предлагают последовательный подход к получению, совместному использованию и хранению данных, что упрощает методы управления данными и повышает их качество.

Коммерческие компании производят и поддерживают доступные инструменты ETL со множеством функций, включая графический пользовательский интерфейс (GUI), поддержку реляционных и нереляционных баз данных и обширную документацию. Они достаточно надежные и зрелые по дизайну. Примеры: Hevo, SAS Data Management, Fivetran, Oracle Data Integrator и так далее.

Корпоративные инструменты ETL стоят дорого и нужно пройти обучение, чтобы освоить их. Можно использовать альтернативы с открытым исходным кодом — Talend Open Studio, Pentaho Data Integration, Singer или Hadoop. Но такие инструменты могут не соответствовать конкретным потребностям организации. Кроме того, поскольку технологии ETL с открытым исходным кодом часто не поддерживаются коммерческими компаниями, их обслуживание, документация, удобство использования и полезность могут различаться.

Если для компании важна гибкость, можно разработать внутреннее решение. Такой ETL-инструмент будет закрывать конкретные потребности организации, подойдет под процессы и будет соответствовать приоритетам. Для разработки можно использовать популярные языки программирования — SQL, Python и Java. Главный недостаток — нужны ресурсы для разработки, тестирования, обслуживания и обновления инструмента.

Коммерческие инструменты	**Инструменты с открытым исходным кодом**	**Кастомные инструменты**
SAS Data Management	Talend Open Studio	Внутренние решения компаний
Fivetran	Pentaho Data Integration	
Oracle Data Integrator	Singer	

Хранилище клинических данных (CDW)

Хранилище данных (DW) — один из ключевых инструментов для принятия решений. Это коллекция данных, организованная для отчетов и исследований. Она объединяет и связывает данные из разных источников, упрощает доступ к ним. Данные в хранилище объединяются и представляются в многомерной форме, что облегчает быстрое и простое отображение и анализ.

В обычных ETL-проектах извлеченные и преобразованные данные загружаются в хранилище данных (DW). В науке о клинических данных — в хранилище клинических данных (CDW).

Хранилище клинических данных — Clinical Data Warehouse (CDW) или Clinical Data Repository (CDR). Это БД в режиме реального времени, которая объединяет данные о пациенте из разных источников. Хранилища оптимизированы для извлечения данных об одном пациенте, а не о группе с общими характеристиками, поэтому они не помогут облегчить управление клиническим отделением, но помогут получить быстрый доступ ко всей информации о пациенте. Типичные данные в CDR: результаты анализов, демографические данные пациентов, информация об аптеках, радиологические отчеты, информация о госпитализации, даты выписки и перевода, коды МКБ-9, записи о ходе лечения.

CDW может стать основой для документирования, проведения, планирования и содействия клиническим исследованиям. Кроме того, CDW улучшает процесс принятия клинических решений, упрощая анализ и обработку данных.

Проблемы с ETL в здравоохранении

Частые проблемы: совместимость между исходными и целевыми данными, качество исходных данных и масштабируемость процесса ETL.

Информация в медицинских системах часто не унифицирована: поля, словари, терминология и другие параметры могут отличаться в разных организациях и разных системах. Это приводит к проблемам совместимости: не всегда целевая система может эффективно интегрировать исходные данные.

Чтобы предоставить конечным пользователям полные и правильные данные, нужно устранить проблемы с их качеством: от простых опечаток до расхождений значений, противоречивой или отсутствующей информации.

Другая проблема — масштабируемость. Данные о состоянии здоровья — это объемные данные. Разработка и поддержка инструмента, адаптированного к стремительному росту данных и при этом сохраняющего разумное время отклика, — сложная задача.

Первый шаг в решении проблем — признать, что они существуют, и что вы, скорее всего, столкнетесь с ними при создании ETL-решения.

Многие специалисты, занимающиеся ETL в сфере здравоохранения, утверждают что, отсутствие высококачественных конвейеров данных — одно из главных препятствий для широкого внедрения доступных передовых подходов. Это препятствие не присуще отрасли и не связано с практикующими врачами или пациентами.

Использование ETL в неполном цикле

На заре ИТ-эпохи любая задача загрузки, обработки и выгрузки большого объема данных из баз решалась уникальным способом. Данных было немного, баз данных — еще меньше, а трудозатраты никто не считал. Постепенно объем данных и БД рос. Все очевиднее становилась идея создать унифицированные способы решить задачу интеграции данных, а не изобретать велосипед. Из этой идеи и выросли продукты, общее назначение которых можно описать тремя словами — Extract, Transform, Load.

Сейчас на рынке множество ETL-решений. Среди них и флагманские продукты от крупных корпораций, и Open Source-решения, которые часто не уступают первым. Поэтому мало кому сегодня придет в голову организовать сбор данных с помощью процедур, написанных на коленке. Даже для компаний с ограниченным бюджетом есть бесплатные решения. А in-house разработка, скорее всего, проиграет готовым ETL-продуктам в удобстве использования, гибкости, а, возможно, и производительности.

Однако как только мы сталкиваемся с более общей задачей интеграции данных между информационными системами, не связанной с хранилищами, ETL отходит на второй план. В первую очередь начинают рассматривать разработку собственных решений или интеграцию приложений.

То есть за ETL закрепилась репутация инструмента наполнения хранилищ данных, важного и полезного, но узкоспециализированного. А ведь возможности современных ETL-продуктов гораздо шире, с их помощью можно решать задачи

интеграции распределенных систем, особенно в условиях недостаточности инфраструктуры, для организации синхронизации данных и создания единой точки ввода, обеспечения безопасности, наполнения систем данными из смежных систем в случае отсутствия соответствующих сервисов и тому подобное. Конечно, все эти задачи могут быть решены «правильным» способом – путем внедрения корпоративной шины, использования сервис-ориентированной архитектуры, модернизации аппаратной и программной инфраструктуры, но реальная жизнь всегда сложнее: где-то приходится усмирять «зоопарк» legacy-систем, где-то крайне ограничен бюджет, где-то очень сжатые сроки и так далее.

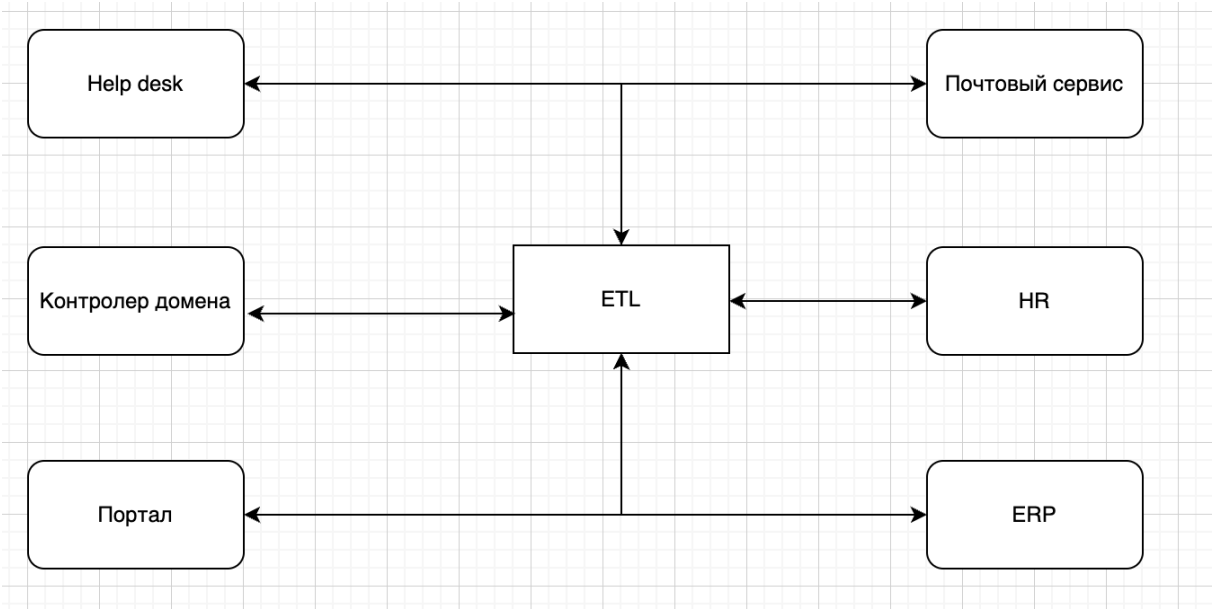
Давайте постараемся пролить свет на такую темную сторону ETL, как ее использование в прикладных задачах в отрыве от хранилищ данных. Для наглядности рассмотрим несколько примеров. Все они взяты из реальной жизни, и хотя использование именно ETL-средства могло поначалу показаться неоправданным, оно во всех случаях позволило просто, быстро, с минимальными затратами и с высокой степенью надёжности решить задачи, поставленные бизнесом перед ИТ.

История первая. Прием на работу

Работа в любой компании для нового сотрудника начинается с заведения данных в многочисленные учетные системы. Причем в небольших и средних организациях вносить эту информацию, как правило, приходится разным людям из разных подразделений. В результате возникают ситуации, когда принятый на работу сотрудник подолгу не может получить банковскую карту, потому что не был вовремя заведен в бухгалтерии, а у уволенных сотрудников есть доступ к корпоративной почте и приложениям, потому что их учетная запись в домене не заблокирована.

Представим следующую «диспозицию»: в компании используются контроллер домена, применяющий LDAP, кадровая система на базе СУБД Oracle, система бюджетирования 1С, внутренний портал со штатным расписанием, почта, Help Desk. Все эти системы не знают о существовании друг друга, при этом требуется уменьшить количество процедур ввода данных о сотруднике в учетные системы, а также обеспечить выравнивание бизнес-процессов.

В целом, это типичная задача сквозной интеграции бизнес-процессов. Она так часто встречается, что для ее решения есть отдельный класс продуктов — IdM (Identity Management, подробнее о них можно прочитать в Jet Info 1 5, 2010). Однако это практически всегда масштабные решения, внедряемые не один месяц и подразумевающие внушительные капиталовложения, а их в небольших и даже средних компаниях часто не хватает. Для налаживания взаимодействия между системами можно воспользоваться ETL, причем это не потребует значительного времени и сложной разработки. Для наглядности вся картина приведена на схеме:



Интеграция бизнес-процессов с помощью ETL

Итак, данные о новом сотруднике попадают в БД HR-системы: вводят его ФИО, должность, подразделение, дату рождения, табельный номер и добавляют фотографию. Чтобы сотрудник получил доступ к корпоративной сети и системам, учетную запись о нем нужно создать в Active Directory на контроллере домена (DC) — это задача для специалистов Help Desk.

Здесь в игру вступает ETL: заявка в системе на ввод данных нового сотрудника будет сформирована автоматически, поскольку ETL регулярно делает запрос в HR-систему для получения списка сотрудников без email-адреса (в рамках нашей компании это означает появление нового специалиста). Причем вводить массу информации, которая уже введена в штатное расписание, сотрудникам Help Desk не потребуется. Достаточно указать табельный номер, внести номер телефона подключенного сотрудника и его комнаты, а также email. Эта информация будет передана обратно в штатное расписание, в то время как подразделение и должность попадут в контроллер домена. Это обогащает обе системы нужными данными, не требуя при этом ручного ввода.

Как мы уже отмечали, компания некрупная, поэтому опрос кадровой системы и контроллера домена раз в 5 минут с целью выявления данных, требующих синхронизации, никак не скажется на их производительности. Это дает еще один плюс: в случае изменения email, рабочего телефона, должности или подразделения вся информация синхронизируется автоматически.

Нужно отметить, что до того, как данные о сотруднике начнут распространяться по системам, происходит еще одна важная проверка — достаточность бюджета подразделения для найма вакансии. Для этого используется простая выгрузка из финансовой системы, в случае если новый сотрудник не «влезает» в бюджет, заявка на ввод учетной записи в DC не формируется. В то же время создается письмо с отчетом, которое получают кадры, бухгалтерия и начальник подразделения для дальнейшего рассмотрения ситуации. Эта несложная с технической точки зрения операция позволяет выровнять вполне реальный бизнес-процесс и уменьшает количество инцидентов в бюджетировании. В реальности уже через несколько месяцев после ее введения в эксплуатацию все начальники подразделений стали гораздо внимательнее относиться к найму нового персонала и согласовывать изменение бюджета заранее, а не по факту. И, что немаловажно, это не потребовало никаких репрессивных мер со стороны высшего руководства.

Наконец, сотрудник принят и начинает трудовые будни. Но возникает вопрос: как проинформировать коллег о его номере телефона или электронном адресе, как им понять, какая у него должность, как он выглядит и когда можно праздновать его день рождения? Вся эта информация обычно представлена на внутреннем портале компании, но вводить ее туда некому, а сам портал не обладает обширными связями с остальными информационными системами. Фактически это простой сайт.

На помощь опять приходит ETL. Данные о сотруднике и его контактах попали в штатное расписание, откуда раз в день происходит выгрузка всей информации, в том числе его данных и фотографии в формате xml. Этот выгруженный файл и передается порталу, где из него строится красивый и удобный список сотрудников, а также создаются уведомления о ближайших днях рождения.

Рано или поздно приходит время расстаться с сотрудником. Обычно в таких случаях ему выдают длинный обходной лист, с которым нужно обойти все подразделения компании, чтобы отметить факт увольнения во всех учетных системах. Но и тут ETL может помочь — технология позволяет упразднить многие пункты обходного листа.

Как только в HR-систему заносятся данные о дате окончания работы сотрудника на этом месте, информация о необходимости блокировки его записи поступает контроллеру домена, его почта автоматически архивируется, а почтовый ящик блокируется. При этом также возможен полуавтоматический режим с созданием заявки на блокировку в Help Desk. Как показала практика, такой режим может потребоваться для проведения кадровых перестановок.

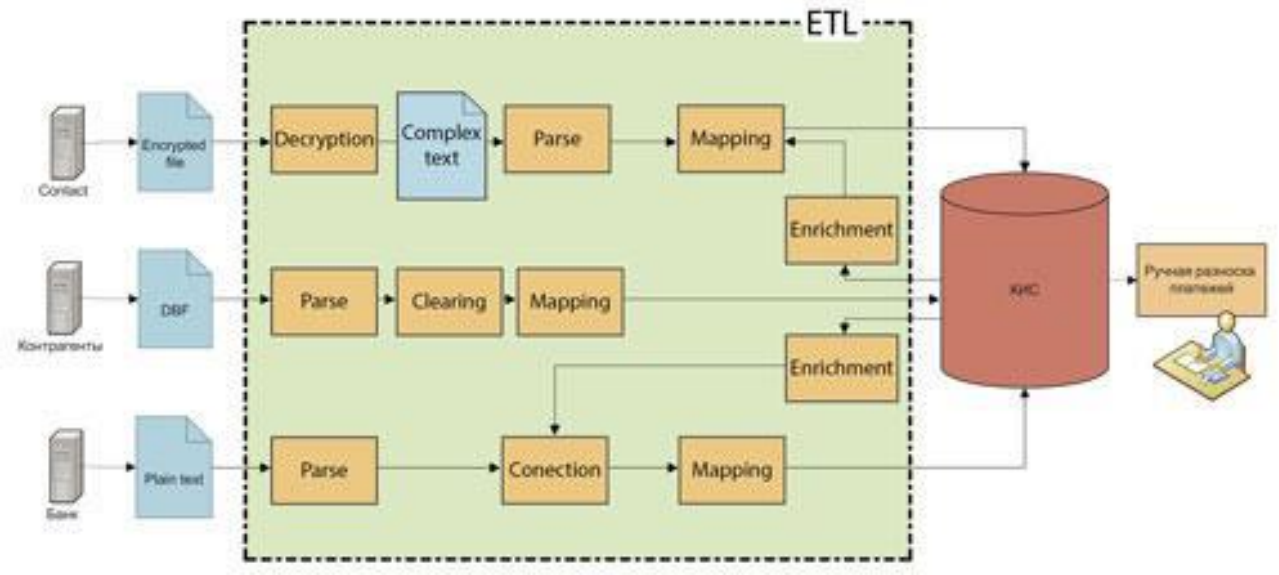
Подведем итог. Мы смогли наладить взаимодействие между пятью разнородными системами, не имеющими штатных средств для интеграции, при этом в каждом из этих взаимодействий использовались разные способы доступа и форматы данных. Среди них были и прямой доступ в БД через JDBC-драйвер или OCI, и подключение к LDAP, и разбор плоского текстового файла, и формирование xml, и отправка почты. Да, это можно было реализовать с помощью написания собственного приложения, однако использование такого количества различных источников потребовало бы много времени на разработку. Фактически мы решили задачу интеграции сквозных бизнес-процессов без применения сложных и дорогостоящих технологий, и, что самое важное, на это было потрачено минимальное количество ресурсов и времени, а эффект был весьма ощутим.

История вторая. Разноска платежей

Другая задача, с которой сталкивается большинство компаний, принимающих платежи от контрагентов через банки и платежные системы, — разноска платежей, то есть сопоставление информации, полученной в виде платежных документов, с деньгами, поступившими на расчетный счет. Часто это два независимых потока, которые сотрудники бухгалтерии или операционисты связывают вручную. При лучшем раскладе используемая корпоративная ИС имеет функционал автоматической привязки платежей.

В рассматриваемой нами компании информация о платежах поступает от платежной сети. Поскольку она содержит персональные данные, то в соответствии с 1 153-ФЗ поставляется в зашифрованном виде. Вторым потоком данных являются файлы в формате DBF, содержащие информацию о банках-контрагентах, которая требуется для геолокации платежа. И последними с минимальной задержкой в три банковских дня приходят деньги и выписка с платежами, проведенными через банк-партнер.

Нужно отметить, что в лоб эти потоки не связываются: номера документов, указанные в реестрах от платёжной системы и банка, не совпадают, а из-за особенностей работы банка дата платежа, которая значится в выписке, может не совпадать с датой реальной оплаты, которая содержится в реестре.



Организация разноски платежей

Процесс также осложняется тем, что файл реестра зашифрован, поэтому сначала нужно его расшифровать. Для этой цели платежная система предоставляет консольную программу и набор ключей шифрования. В компании был разработан несложный процесс трансформации: шифрованный файл забирают с внешнего FTP, помещают на локальный ресурс, а затем запускают преднастроенную программу для дешифровки.

На выходе специалисты получают текстовый файл сложной структуры, содержащий ФИО, телефон, паспортные данные плательщика, сумму и дату платежа, а также определенное количество дополнительных и технических данных. Немаловажно, что в последних содержится информация, идентифицирующая транзакцию, это позволяет в большинстве случаев связать платеж с данными из банковской выписки. Данные из реестра обогащаются информацией о банках-контрагентах (филиалах, подразделениях, городах и адресах отделений), после этого осуществляются их маппинг к конкретным полям таблиц корпоративных ИС и загрузка в базу. Естественно, обогащение данных происходит в рамках реляционной модели с использованием внешних ключей, кроме того, используются уже очищенные данные, загруженные в ИС.

Отметим, что загрузка информации о банках и ее очистка — отдельный ETL-процесс. После прихода банковской выписки запускается еще один процесс, задача которого состоит в сопоставлении ранее полученной информации о платежах с реально пришедшими деньгами. Так как выписки приходят из банка в текстовом формате, первый шаг трансформации — разбор файла. После этого идет процесс автоматической привязки платежей с использованием информации, ранее загруженной в корпоративную ИС из реестров платежей и банков. Нужно отметить, что в процессе привязки происходит сравнение не только ключей, идентифицирующих транзакцию (в некоторых случаях они не имеют пары), но и суммы и ФИО плательщика, а также отделения банка. Также решается задача исправления неверной даты платежа, указанной в банковской выписке, на реальную дату его совершения.

В итоге без привлечения разработчиков мы получили систему автоматической привязки платежей, при этом основные затраты были связаны с проектированием и изучением форматов файлов. На выходе происходит привязка до 90% платежей, что в условиях небольшой компании позволяет полностью снять с одного сотрудника обязанности проведения ручной привязки. По платежам, которые не привязываются автоматически, обогащение позволило существенно облегчить процедуру ручной привязки и даже повысило ее эффективность. Например, наличие телефонного номера плательщика позволяет уточнить данные о платеже лично у него, геолокация платежа дает информацию для аналитических отчетов, а также позволяет более эффективно отслеживать переводы от партнеров-брокеров.

Отметим, что описанная задача, как и в первом случае, имеет чисто прикладной характер. Традиционное решение зачастую связано с работой разработчиков. Конечно, присутствие подобных задач говорит о незрелости бизнес-процессов компании, однако на этапе запуска бизнеса в России такое не редкость. Использование ETL-средства в этом случае полностью оправдывает себя и не только позволяет автоматизировать процесс привязки платежей, но и высвобождает ресурсы, а также повышает качество, тем самым обеспечивая экономию бюджета компании.

Обе истории нельзя рассматривать как хорошие практики использования ETL, все же основной задачей этого инструмента остается интеграция данных. Но в условиях, когда функциональность продуктов ETL у большинства вендоров постоянно расширяется, использовать только 10% их потенциала неразумно. Современный ETL оставляет большое поле для экспериментов, часто позволяя решать задачи, нехарактерные для него, и делать это быстро, эффективно и с минимальным использованием навыков разработки.

Заключение

Сегодня мы рассмотрели способы применения ETL для решения бизнес-задач в различных сферах. Поговорили о возможности использования неполного цикла ETL, узнали, как это может помочь бизнесу. Обсудили, какие проблемы встречаются при внедрении ETL на примере здравоохранения.

Это лишь вершина айсберга. Важно серьезно подходить к этапу проектирования системы, понимать структуру бизнеса и особенности данных, с которыми вам придется работать в конкретной конечной системе.

Что почитать

1. Статья о том, как ETL может помочь улучшить системы здравоохранения — Dynamic-ETL: A hybrid approach for health data extraction, transformation and loading (на английском)