МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Лекция №8
Выбор средства управления данными

Методы и средства проектирования информационно-аналитических систем
(наименование дисциплины (модуля) в соответствии с учебным планом)

Уровень	специалитет
	(бакалавриат, магистратура, специалитет)
Форма обучения	очная
	(очная, очно-заочная, заочная)
Направление(-я) подготовки	10.05.04 «Информационно-аналитические системы безопасности»
	(код(-ы) и наименование(-я))
Институт	Институт кибербезопасности и цифровых технологий (ИКБ)
	(полное и краткое наименование)
Кафедра	Информационно-аналитические системы кибербезопасности (КБ-2)
	(полное и краткое наименование кафедры, реализующей дисциплину (модуль))

Используются в данной редакции с учебного года 2023/24
(учебный год цифрами)

Проверено и согласовано «____» _____20___г. _____ _____
(подпись директора Института/Филиала с расшифровкой)

Москва 2024 г.

Учебные вопросы:

1. Понятие ETL-процесса
2. Структура ETL-процесса
3. Реализация ETL-процесса
4. Проблемы с ETL-процессами

1. Понятие ETL-процесса

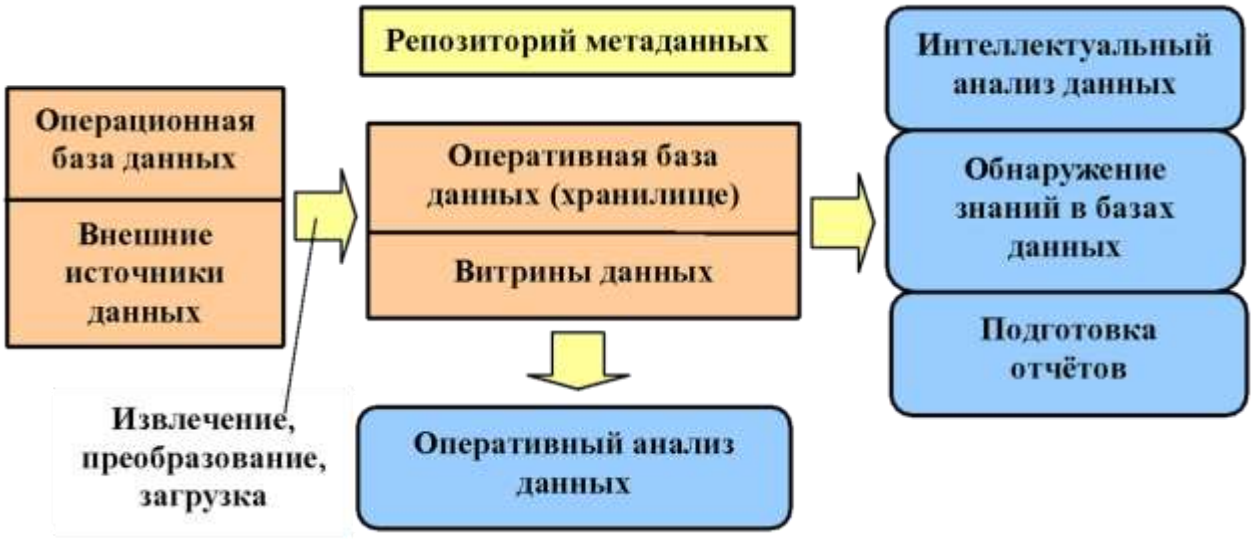


рис. Архитектура систем поддержки принятия решений

ETL-процессы – подход в управлении данными, который следует четкому алгоритму: извлечение (Extract), преобразование (Transform), загрузка (Load). Таким образом специалист сначала получает информацию, работает с ней и загружает в систему для дальнейших операций.

ETL-процессы используются в бизнес-аналитике, например, для сбора и анализа маркетинговых данных, хранящихся в разных местах. Есть у подхода свои проблемы: разные форматы информации, низкое ее качество и негибкая система, под которые нужно подстраиваться.

ETL (с англ. Extract, Transform, Load можно перевести как «извлечение, преобразование, загрузка») представляет собой процесс управления информацией, состоящий из трех этапов. На первой стадии данные извлекаются из структурированных и неструктурированных источников, после этого они трансформируются в требуемый формат и загружаются в место назначения.

Рассмотрим принцип работы ETL на примере обычного магазина. Допустим, что перед продавцом стоит задача взять изделие, завернуть его в привлекательную упаковку и отдать покупателю. На языке ETL-процессов это выглядело бы следующим образом. Продавец

извлекает товар из первоначального источника – с полки в магазине. После этого он выполняет преобразование, упаковывая изделие в подарочную бумагу. Затем продавец «загружает» продукт в пакет покупателя.

Примерно так же работают инструменты ETL. Они собирают данные из различных систем (извлечение), объединяют её с другими источниками (преобразование) и сохраняют (загружают) для последующего анализа.

К примеру, программа способна консолидировать информацию о клиенте фитнес-клуба. При этом если в отделе продаж данные вносились через одну CRM-систему, а при онлайн-покупке клиенты проходили регистрацию через другую, то с помощью ETL компания может извлечь данные из обеих платформ и трансформировать их в единую таблицу.

ETL позволяет упростить задачи, связанные с обработкой информации. Система дает возможность объединить сведения из разных источников и решить проблему переноса необработанных и распределенных данных в единый репозиторий.

Перечислим несколько основных сфер применения ETL:

- Перемещение и репликация данных. При организации этих процессов очень важна скорость. Чем быстрее предприятие сможет интегрировать данные в новые системы, тем раньше они начнут приносить пользу. Зачастую компании приходится работать с несовместимыми форматами или файлами. Трансформирование информации в ETL-процессах позволяет решить эту проблему.
- Сбор и обработка информации. Предприятия работают с разнообразными источниками данных. Рассмотрим пример использования ETL-процессов в интернет-маркетинге. Предположим, что сведения о продажах и результатах маркетинговых кампаний приходят сразу из двух сервисов. Чтобы их правильно проанализировать, нужно объединить эти данные. С помощью ETL организация может перенести информацию в требуемом формате, который будет удобен для дальнейшего использования.
- Машинное обучение (ML). Озеро данных [1] обычно содержит информацию из целого ряда источников. Однако не вся она может применяться для эффективного обучения алгоритмов. ETL позволяют находить только те сведения, которые действительно помогут процессу ML. При этом данные будут трансформированы в наиболее подходящий формат и затем загружены в озеро или БД.

[1] Data Lake (озеро данных) — это технология для получения и управления данными в разных форматах: в необработанном, неупорядоченном или, наоборот, структурированном или слабоструктурированном виде, в едином репозитории. Особенности устройства Data Lake: данные необработаны либо слабообработаны, большой срок хранения данных, есть возможность преобразования данных, поддерживаются разные схемы чтения данных.

- Аналитическое хранилище (DWH) [2] . Разработка ETL-процессов играет важнейшую роль в процессе проектирования хранилища. Дело в том, что источников данных может быть очень много, поэтому компании необходимо правильно фильтровать, очищать, объединять, разделять и сортировать информацию.
- Конвейеры данных (Data Pipelines) [3] . ETL-конвейер позволяет подготовить информацию перед аналитикой. Он помогает привести разрозненные данные к определенному стандарту. При этом разработчики освобождаются от целого ряда технических задач. За счет более тщательной предобработки данных ETL-конвейер повышает эффективность аналитики.

Плюсы ETL:

1. Систематизация данных. ETL-процессы позволяют структурировать данные и привести их в более удобный для анализа вид.

2. Автоматизация. Многие процессы, которые раньше требовали ручной обработки, теперь могут быть автоматизированы при помощи ETL-систем, что снижает риски ошибок.

3. Консолидация данных. ETL-технологии позволяют собирать данные из различных источников и объединять их в одном хранилище данных. Это упрощает анализ данных и позволяет получить более точные результаты.

4. Масштабируемость. Если ты используешь ETL-систему, то можешь легко масштабировать свою базу данных и обрабатывать большие объемы данных.

Минусы ETL:

1. Сложность разработки. ETL-процессы достаточно сложны в реализации, поэтому для их разработки нужны высококвалифицированные специалисты.

2. Стоимость. ETL-инструменты обычно стоят дорого, так что использование ETL может быть недоступно для маленьких проектов с ограниченным бюджетом.

[2] Data Warehouse (DWH) — хранилище, предназначенное для сбора и аналитической обработки исторических данных организации. Анализ помогает руководителям видеть цельную картину бизнеса и принимать решения, как развивать отдельные направления или бизнес в целом. В DWH данные из всех СУБД предприятия аккумулируют и очищают, формируя их единый источник.

[3] Data pipeline — это процесс, в котором данные собираются, обрабатываются и передаются из одной системы в другую. Это может включать различные этапы, такие как сбор, очистка, преобразование, интеграция и хранение данных. Data pipeline играет важную роль в аналитике данных, поскольку он позволяет специалистам быстро и эффективно обрабатывать большие объемы информации.

2. Структура ETL-процесса

Извлечение (Extract)

Извлечение и копирование из пула источников (к примеру, баз данных SQL и NoSQL платформ ERP[4] и CRM[5], приложений SaaS[6]) является самым первым этапом перемещения любой информации. Из-за специфики работы с некоторыми системами-источниками эта стадия зачастую является самой сложной.

Существует три вида извлечения информации:

- **Полное**. Этот способ используется в том случае, если система-источник не способна различать новые или измененные записи.
- **Частичное с уведомлениями об обновлениях**. Является самым удобным методом извлечения информации. Однако его можно использовать лишь в том случае, если в системе есть уведомления об изменениях записей. Именно они позволяют извлекать конкретные данные, без загрузки всей информации.
- **Инкрементное извлечение** или частичное извлечение без уведомлений об обновлениях. Данный метод позволяет работать только с теми записями, которые подверглись изменению.

С одной стороны, пользователю необходимо заранее определить, какие элементы данных требуется извлечь для дальнейшего преобразования и загрузки. Однако с помощью ELT-процессов информация извлекается мгновенно. Таким образом, пользователь может отложить вопрос выбора объектов для преобразования и анализа.

Преобразование (Transform)

Эта составляющая ETL-процесса представляет собой последовательность действий, которые нацелены на подготовку информации для изменения под характеристики другой системы или выполнения иных задач.

Трансформирование может состоять из следующих операций:

---

[4] ERP-системы (Enterprise Resource Planning) — это автоматизированные системы управления предприятием, которые помогают контролировать весь цикл производства. Задача ERP-систем — выстраивать процессы и вести централизованный учёт ресурсов, которые компания вкладывает в создание и сбыт продукта.

[5] CRM-система (Customer Relationship Management) — это система управления взаимоотношениями с клиентами. Она помогает грамотно структурировать бизнес-процессы, сохранять клиентскую базу, собирать историю взаимодействия с клиентами, анализировать уровень лояльности клиентов к компании и автоматизировать процессы транзакций с клиентами.

[6] SaaS (от англ. software as a service — «программное обеспечение как услуга») — одна из форм облачных вычислений, модель обслуживания, при которой подписчикам предоставляется готовое прикладное программное обеспечение, полностью обслуживаемое провайдером. В этой модели поставщик самостоятельно управляет приложением, предоставляя заказчикам доступ к функциям с клиентских устройств (как правило, через мобильное приложение или веб-браузер). Основное преимущество модели SaaS для потребителя услуги состоит в отсутствии затрат, связанных с установкой, обновлением и поддержкой работоспособности оборудования и работающего на нём программного обеспечения.

- сортировка и фильтрация информации (позволяет убрать иррелевантные (неуместные, неподходящие; не относящиеся к делу) элементы);
- устранение дубликатов и очистка;
- транслирование и преобразование;
- удаление или шифрование (повышает уровень безопасности данных);
- соединение или разделение таблиц и т.д.

Каждое из вышеперечисленных действий осуществляется вне целевой системы, во время подготовки. Реализация этих операций выполняется дата-инженерами.

К примеру, в хранилище Online Analytical Processing (OLAP[7]) можно размещать только реляционные структуры информации. Из-за этого данные необходимо предварительно трансформировать в SQL-читаемый формат. Каждое преобразование может выполняться только один раз. Из-за этого ETL является негибким процессом. Если требуется применить к уже трансформированной информации новый тип анализа, специалисту может понадобится заново модифицировать весь конвейер данных.

Загрузка (Load)

Информация загружается в целевую систему хранения. Это позволяет другим пользователям получать доступ к данным. При этом поток ETL-процесса включает в себя импорт информации (которая была заранее подготовлена и извлечена) из промежуточной БД в целевое хранилище данных или базу данных.

Для этого выполняются физические вставки определенных записей в виде новых строк таблицы хранилище. При этом используются SQL-команды или сценарий пакетной загрузки большого массива данных.

В более практичной форме это выглядит следующим образом:

Принцип работы ETL из нескольких основных шагов:

1. **Шаг загрузки**. На этом этапе данные попадают в ETL-систему. В ее основе лежит процесс Extract, однако, теперь рассматриваем происходящее «изнутри» системы, и для нас важно, как проходит сама загрузка, а не извлечение данных. При этом данные, которые попадают в систему, называются сырыми, они не обработаны и не проверены, качество данных может быть произвольным, их только сверяют по количеству строк. Если количество строк меньше, чем было в источнике, произошел сбой.

2. **Этап валидации**. На этом этапе система проводит проверку полученных данных. Это процесс валидации, при котором информация по очереди проверяется и фильтруется в

[7] OLAP (от англ. online analytical processing, «интерактивная аналитическая обработка») — технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу.Реализации технологии OLAP являются компонентами программных решений класса Business Intelligence.

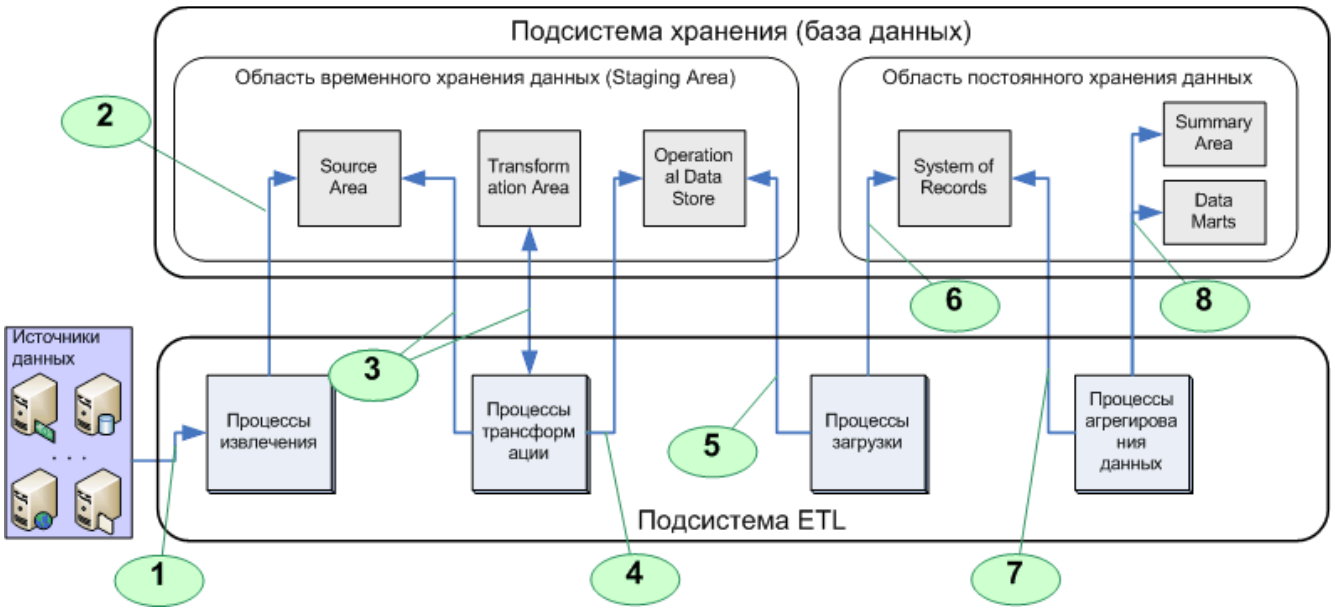
соответствии с настроенными правилами. Система анализирует полноту данных, проверяет их корректность и наличие ошибок. В конце валидации выдаются отчеты обо всех найденных ошибках. Если они есть, их нужно исправить.

3. **Этап мэппинга**. Этот этап относится к процессу Transform и призван преобразовать полученные данные в нужный формат. После прохождения валидации данные представляются в виде таблицы, к которой добавляются нужные столбцы и строки. Мэппинг может происходить с использованием различных алгоритмов в зависимости от использованного ETL-инструмента.

4. **Этап агрегации**. Этот этап, также являющийся частью процесса Transform, необходим для того, чтобы преобразованные данные можно было перенести в новое хранилище без ошибок. За счет изменения связей между данными, информация агрегируется в новую таблицу. Результат агрегации – новая таблица, в которой данные представлены в требуемом формате для нового хранилища.

5. **Выгрузка**. Этот этап реализует процесс Load, когда преобразованные и очищенные данные выгружаются из ETL-системы и отправляются в новое хранилище. Для этого используются коннекторы и различные части интерфейса ETL-системы и хранилища.

Другое описание процесса ETL:



1. Процессы извлечения данных извлекают данные из систем источников.

2. Процессы извлечения данных сохраняют извлеченные данные в интерфейсные таблицы области Source Area.

3. Процессы преобразования (трансформации) данных извлекают данные из интерфейсных таблиц (Source Area), проводят захват изменений, преобразование данных по определенным бизнес-правилам с сохранением промежуточных результатов в Transformation Area и сохраняют результат в области оперативного хранения.

4. После проведения преобразования данных данные загружаются в область оперативного хранения Operational Data Store.

5. Процессы загрузки данных производят чтение данных из области оперативного хранения.

6. Процессы загрузки данных проверяют ссылочную целостность данных и проводят их загрузку в область детальных данных (System of Records).

7. Процессы агрегации данных производят чтение детальных данных.

8. Процессы агрегации данных производят агрегацию и запись данных в Summary Area и Data Marts.

3. Реализация ETL-процесса

Настройка ETL-процесса состоит из 5 базовых этапов.

Определение задачи

Предположим, что в компании необходимо создать систему материальных поощрений для менеджеров. Чтобы это сделать, потребуется выполнить анализ информации, связанной с объемами продаж и поощрений. Компания может поставить перед специалистом задачу настройки процесса сбора и отправки нужных записей в аналитическое хранилище.

Следовательно, дата-инженеру необходимо учесть следующие параметры:

- Типы систем, которые применяются для хранения данных. К примеру, могут использоваться CRM, базы данных, документы.
- Вид таблицы-приемника, которая будет содержать итоговые данные (каков будет формат и названия колонок).
- Частота обновления данных (1 раз в сутки, раз в несколько часов или в онлайн-режиме).
- Объект обновления: информация, которая появилась за определенное время, или данные, которые уже содержались в БД.
- Проблемы, появляющиеся в информации, и способ их решения. К примеру, специалисты нередко сталкиваются с пропусками, аномалиями, тестовыми значениями, некорректными форматами.
- Способ уведомления пользователя о проблемах (скажем, если в какой-то момент в систему будет отправлено в несколько раз меньше сведений, чем раньше).

Получение доступа

Допустим, что сведения о продажах размещены в 1С, информация о работниках компании – в Гугл-таблицах, а акции и скидки – в базе данных. Каждая система имеет свои

требования к доступу. При этом у каждого из этих источников есть свои ответственные лица, которые могут открывать и закрывать доступ для других пользователей.

Специалисту потребуется:

- Связаться с ответственными лицами и запросить доступ к определенным системам и данным.
- Сформировать аккаунт для функционирования автоматического ETL-процесса. Это необходимо для того, чтобы знать, кто именно пользуется информацией.
- Создать аккаунт дата-инженера. Это позволяет быстро проверять информацию и производить отладочные работы. Как правило, специалисту не предоставляют полный доступ (к примеру, чтобы он не смог просмотреть персональные данные покупателей).
- Получить доступ к так называемому тестовому контуру. Речь идет о пробной информации, на базе которой можно скорректировать и протестировать ETL-процессы.

Проверка полученных данных (предпроцессинг)

Специалисту необходимо выполнить анализ полученных данных и понять, какие из них следует оставить, а какие убрать. К примеру, в таблицах могут находиться текстовые аккаунты менеджеров, которые не нужны для выполнения задачи. Специалист получает часть данных и изучает ее.

Если информацию нужно предобработать, то это необходимо учесть в процессе написания кода.

Написание кода ETL-процесса

Когда специалист разобрался с тем, какие именно требуются сведения, где их найти и каким образом обработать, он должен написать код, который затем станет **ETL-пайплайном** (ETL-процессом).

Инженеру данных необходимо протестировать код. При проверке он должен получить следующие результаты:

- Технически код функционирует правильно, при выполнении не появляется никаких ошибок.
- Код читаем. Иными словами, специалист правильно составил наименования параметров, учел все переносы строк и табуляцию, а также правила формирования текста.
- Информация правильно обрабатывается, в расчетах нет ошибок.

Запуск автоисполнения кода

Существуют инструменты, с помощью которых можно в автоматическом режиме запускать ETL-процесс. К примеру, *Apache Airflow*[8] или *PySpark (Python API для Apache Spark*[9]*)*. Эти фреймворки исполняют код. Пользователь может отслеживать ход работы в интерфейсе или логах.

ETL-пайплайн представляет собой список задач, которые выполняются в заранее установленной последовательности. Яркий пример – батчевый процесс[10] в Apache Airflow, где информация берется частями, а затем запускается процесс по определенному расписанию.

Стоит учесть, что при работе на распределенных системах специалист может разработать ETL-процесс таким образом, чтобы задачи выполнялись параллельно.

Есть разные способы исполнения кода: в режиме-онлайн или по определенному расписанию (к примеру, каждые 2 дня в 14:00 будет запускаться обработка сведений, полученных за прошедшие двое суток).

Реализацию ETL запроса можно описать процесс альтернативно:

Итак, **первое**, что нужно сделать, это определить, что именно мы хотим достигнуть с помощью ETL. Например, мы можем хотеть выгрузить данные из нашей базы данных, обработать их (например, провести очистку данных) и загрузить их обратно в базу данных.

Следующий шаг — **выбор инструментов**. Обычно для ETL используются специальные инструменты, такие как ***Apache Nifi***, ***Talend***, ***Apache Spark*** и т.д. Но в принципе можно использовать и стандартные инструменты СУБД, такие как PostgreSQL или Oracle.

---

[8] Apache Airflow - Открытое программное обеспечение для создания, выполнения, мониторинга и оркестровки потоков операций по обработке данных. Изначально разработано в Airbnb в октябре 2014 года. В марте 2016 года стало проектом Apache Incubator, в январе 2019 года — проектом верхнего уровня Apache Software Foundation.

[9] Apache Spark Фреймворк с открытым исходным кодом для реализации распределённой обработки данных, входящий в экосистему проектов Hadoop. В отличие от классического обработчика из ядра Hadoop, реализующего двухуровневую концепцию MapReduce с хранением промежуточных данных на накопителях, Spark работает в парадигме резидентных вычислений - обрабатывает данные в оперативной памяти, благодаря чему позволяет получать значительный выигрыш в скорости работы для некоторых классов задач, в частности, возможность многократного доступа к загруженным в память пользовательским данным делает библиотеку привлекательной для алгоритмов машинного обучения.

[10] Пакетный режим (Batch processing) — это режим работы компьютера, реализующий многозадачность (мультипрограммность), когда система обрабатывает заранее сформированный пакет заданий пользователя без вмешательства последнего в процесс обработки. При этом система сама распределяет задачи и процессы, оптимизируя загрузку ресурсов вычислительной системы и время выполнения. Наиболее эффективна пакетная обработка при решении следующих задач: обновление больших объемов информации в базах данных и системах OLTP; загрузка больших объемов данных в хранилища данных с помощью процессов ETL; обработка изображений; преобразование файлов из одного формата в другой.

Теперь мы переходим к **настройке** нашего ETL запроса. Первым этапом будет выбор источника данных. Может быть, это файл CSV, база данных или даже API. Далее мы создаем запрос на выборку данных, который нужно провести над исходными данными.

Затем мы **обрабатываем исходные данные**. Этот шаг может включать в себя **фильтрацию исходных данных**, **трансформацию данных** и **проведение очистки**. На этом этапе мы можем использовать мощный SQL или специализированные инструменты предобработки данных, такие как ***OpenRefine***[11] или ***DataWrangler***[12].

Затем **загружаем данные в целевую базу данных**. Мы можем использовать обычный SQL-запрос для вставки данных в базу данных. Важно, чтобы мы предварительно создали таблицы в базе данных и удостоверились, что соответствия столбцов верны.

И наконец, мы можем выполнить последний шаг настройки нашего ETL запроса — **обновление данных** в нашей целевой базе данных. Тут мы можем обновлять данные, которые уже существуют в нашей таблице. Можно использовать оператор ON CONFLICT для вставки новых данных или обновления существующих записей в таблице.

Таковы в общих чертах шаги, которые нужно выполнить для реализации ETL запроса. Конечно, реализация может быть несколько сложнее для больших и более сложных данных

4. Проблемы с ETL-процессами

Существует множество инструментов для работы с ETL-процессами. При их выборе необходимо учитывать задачи, стоящие перед компанией, объем обрабатываемой информации и метод их использования. Перечислим самые часто встречаемые проблемы, которые возникают при настройке процесса ETL.

Подбор оптимального способа обработки. В некоторых случаях компании иногда нужно работать с большим количеством источников и разными форматами данных. Например, с полностью и частично структурированной информацией, потоковыми данными в онлайн-режиме, плоскими файлами[13].

[11] OpenRefine (ранее известный как Google Refine) - это бесплатный и открытый инструмент для очистки, преобразования и анализа данных. Он предоставляет удобный интерфейс для работы с различными типами данных, такими как текст, числа и даты. С помощью OpenRefine можно проводить такие операции, как фильтрация, слияние, разделение и структурирование данных.

[12] Data Wrangler - это инструмент для просмотра и очистки данных, ориентированный на код, который интегрирован в VS Code и VS Code Jupyter.

[13] Плоский файл (Flat file database) – это файл, состоящий из записей одного типа и не содержащий указателей на другие записи, двумерный массив элементов данных. Файлы, которые создаются в прикладных программах пользователя, написанных на алгоритмическом языке, также относятся к этому виду организации данных. Описание логической структуры файлов и параметры размещения на машинных носителях содержатся в каждой прикладной программе обработки файлов.

Одни источники лучше конвертировать в batch-режиме, другие требуют настройки потоковой трансформации данных. Чтобы подобрать оптимальный формат обработки для каждого типа информации, специалист должен хорошо разбираться в этой теме.

Низкокачественные данные. Чтобы проанализировать данные, их нужно заранее преобразовать. При этом трансформация должна быть выполнена с максимальной точностью и в полном объеме. Если делать все вручную, то информация может быть утеряна. Кроме того, нередко возникают ошибки, связанные с дублированием данных. Однако существуют специализированные инструменты для работы с ETL, с помощью которых можно автоматизировать задачи и исключить «человеческий фактор».

Функция мониторинга позволит выявить различные проблемы. К примеру, если в систему попали данные, которые с ней несовместимы. Контроль качества также помогает выявлять дубликаты.

Проблема навигации (Navigation problem). Чтобы извлечь данные например с веб-страниц, нужно, для начала, эти страницы получить. Но современный веб устроен так, что получение страниц с интересующей информацией превращается в совершенно нетривиальную задачу. Вариантов решения множество: написание своего или использование возможностей краулера wget, сторонние библиотеки, использование результатов поисковых систем вроде гугла и т.д. На практике, огромное количество подводных камней, таких как исполнение Javascript и AJAX[14] навигация, ограничения по IP, авторизация страниц, капча, данные во флеш, отправка форм, определение логических дубликатов страниц… Короче говоря, полностью решить задачу очень затруднительно, поэтому, чаще всего, просто достигается результат, приемлемый для конкретного случая.

Проблема распознавания данных (Data extraction problem). Одна из главных проблем интеллектуального извлечения данных — распознавание.

Проблема поиска общей структуры данных (Structure synthesis problem). Следующим шагом для извлечения данных является определения их структуры. На практике очень редко встречаются ситуации, когда все элементы набора данных имеют одинаковые свойства. Одни и те же атрибуты могут иметь разное представление. Задача нетривиальная и универсального решения не существует. Для анализа WEB часто используется комбинированное решение в виде лексического анализа данных и анализа структуры DOM[15].

[14] AJAX (от англ. Asynchronous Javascript and XML — «асинхронный JavaScript и XML») — подход к построению интерактивных пользовательских интерфейсов веб-приложений, заключающийся в «фоновом» обмене данными браузера с веб-сервером. В результате при обновлении данных веб-страница не перезагружается полностью, и веб-приложения становятся быстрее и удобнее.

[15] DOM (от англ. Document Object Model — «объектная модель документа») — это не зависящий от платформы и языка программный интерфейс, позволяющий программам и скриптам получить доступ к содержимому HTML-, XHTML- и XML-документов, а также изменять содержимое, структуру и оформление таких документов.

Проблема сопоставления атрибутов извлекаемых данных (Data mapping problem). Имея некую «сетку атрибутов» данных необходимо каждый элемент из набора данных разложить по этой сетке. Происходит нормализация атрибутов и обеспечивается однородность извлекаемых данных. Например, дата представляется в разных форматах: «Сегодня 11:42», «5 часов назад» или «21 июня». Для обеспечения однородности все такие атрибуты приводятся к общему формату (дата переводится в какой-либо абсолютный формат времени, адреса и станции метро тоже могут нормализоваться).

Проблема объединения данных (Data integration problem). Информация об одном и том же элементе может находиться на разных источниках. А одни и те же данные из разных источников могут иметь разную структуру и ее придется приводить к общему виду. На этом шаге также выявляются логические дубликаты данных и отсеиваются элементы, не удовлетворяющие каким-либо исходным критериям **Плохая масштабируемость**

По мере роста организации объем информации, который она использует, будет становиться все больше. На данный момент компания может без особых проблем применять локальную БД и пакетную загрузку, однако через несколько лет этого будет недостаточно. ETL имеют неограниченные возможности масштабирования процессов и объема.

При принятии решения с использованием полученной информации компании может потребоваться в кротчайшие сроки подключить облачное хранилище. Это позволит организации оперативно обрабатывать большие объемы данных без существенных финансовых потерь.

Когда начали использоваться ETL-процессы?

ETL получил широкое распространение в 70-х годах. В том время организации начали работать с несколькими репозиториями и базами данных, что потребовало эффективной интеграции всей этой информации.

Какие ETL-системы являются самыми популярными?

Можно выделить несколько распространенных коммерческих решений:

- SAP Data Services / SAP BusinessObjects Data Integrator («Systeme, Anwendungen und Produkte in der Datenverarbeitung» / «Systems, Applications and Products in Data Processing»);
- Informatica PowerCenter;
- IBM WebSphere DataStage;
- Oracle Data Integrator;
- SAS Data Integration Server (Statistical Analysis System).

Другие инструменты ETL:

Apache Kafka. Он является одним из лидеров среди инструментов потоковой обработки данных. Он позволяет синхронизировать сообщения между различными источниками, а затем обрабатывать их в режиме реального времени. Kafka также обладает технологией масштабирования, что делает его идеальным выбором для крупных проектов. Он хорошо интегрируется с другими инструментами big data, такими как Apache Hadoop и Spark.

Apache Nifi. Этот инструмент был создан NSA (Агентством национальной безопасности США (АНБ)основан на программном обеспечении "NiagaraFiles", также разработанном АНБ), а теперь является проектом Apache Foundation. Nifi может легко обрабатывать большие объемы данных, а его графический интерфейс предоставляет очень удобный, интуитивно понятный интерфейс для определения преобразований данных. Он также позволяет автоматически масштабировать обработку данных и имеет множество различных источников данных, включая базы данных, файлы и даже API.

К категории условно бесплатных можно отнести:

Oracle Warehouse Builder

Talend Open Studio

Scriptella

<u>Кому приходится работать с ETL-системами?</u>

ETL-процессами занимаются бизнес- и дата-аналитики. Эти специалисты работают с бизнес-логикой и данными, поэтому им нередко приходится сталкиваться с разнородной информацией.

Знание ETL-процессов также может потребоваться разработчикам для реализации некоторых проектов.

Дата-инженеры выполняют проектирование, поддержку и оркестрацию (координирование работы сложных систем) платформ, предназначенных для хранения данных.

Для работы с ETL-процессами специалисту нужно хорошо знать теорию обработки и хранения данных.