Проектирование хранилищ данных

ФИО преподавателя: Смирнов Михаил Вячеславович

e-mail: smirnovmgupi@gmail.com

Лекция 5

Процедуры ETL (extract, transform, load)

Вопросы лекции

- Определение и области применения ETL-процессов.
- Основные шаги процесса ETL.
- Понятие «автоматизированный пайплайн».
- Процесс идентификации ресурсов.
- Основные задачи трансформации данных.
- Формирование требований к процессу загрузки данных в хранилище.
- Техники загрузки данных.
- Ошибки, выявляемые при валидации данных после загрузки.

Общее определение ETL-процессов данных

Представим себе ритейлера (например бытовой техники), у которого есть как онлайн, так и оффлайн-магазины. Очевидно, что back-end данных о продажах для этих двух вариантов будет разный (оффлайн вариант еще наверное суровый древний legacy).

При получении аналитической отчетности необходимо *совместить данные* этих двух систем в одном большом хранилище. Собственно, решение этой задачи и есть ETL в общем виде.

Современные области применения ETL

- ***Облачная миграция.*** Процесс переноса данных и приложений в облако называют облачной миграцией. Она помогает сэкономить деньги, сделать приложения более масштабируемыми и защитить данные.
- ***Хранилище данных.*** Хранилище данных — база данных, куда передают данные из различных источников, чтобы их можно было совместно анализировать в коммерческих целях.
- ***Машинное обучение.*** Машинное обучение — метод анализа данных, который автоматизирует построение аналитических моделей.
- ***Интеграция маркетинговых данных.*** Маркетинговая интеграция включает в себя перемещение всех маркетинговых данных — о клиентах, продажах, из социальных сетей и веб-аналитики — в одно место, чтобы вы могли проанализировать их. ETL используют для объединения маркетинговых данных.

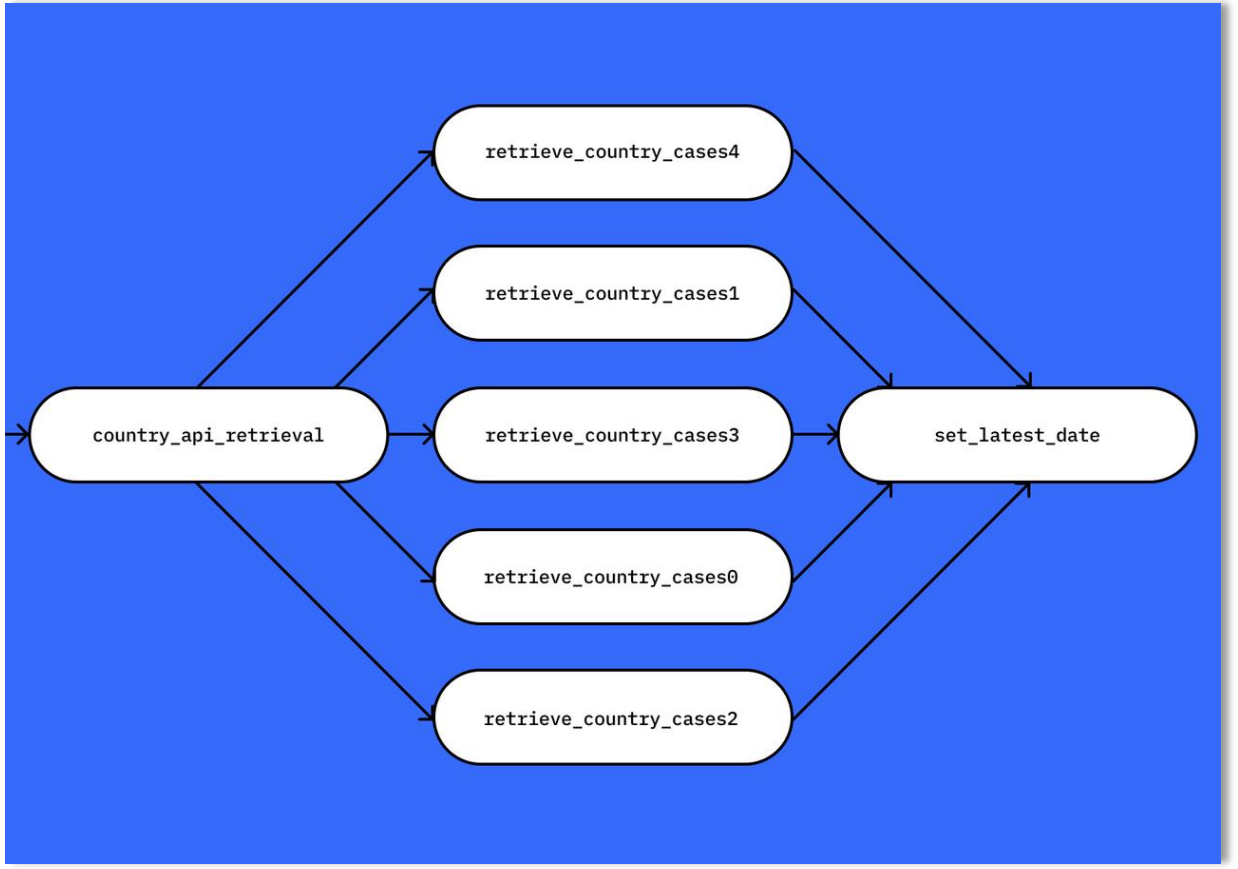
Современные области применения ETL

- ***Интеграция данных IoT.*** То есть данных, собранных различными датчиками, в том числе встроенными в оборудование. ETL помогает перенести данные от разных IoT в одно место, чтобы вы могли сделать их подробный анализ.
- ***Репликация базы данных*** — данные из исходных баз данных копируют в облачное хранилище. Это может быть одноразовая операция или постоянный процесс, когда ваши данные обновляются в облаке сразу же после обновления в исходной базе.
- ***Бизнес-аналитика.*** Бизнес-аналитика — процесс анализа данных, позволяющий руководителям, менеджерам и другим заинтересованным сторонам принимать обоснованные бизнес-решения.

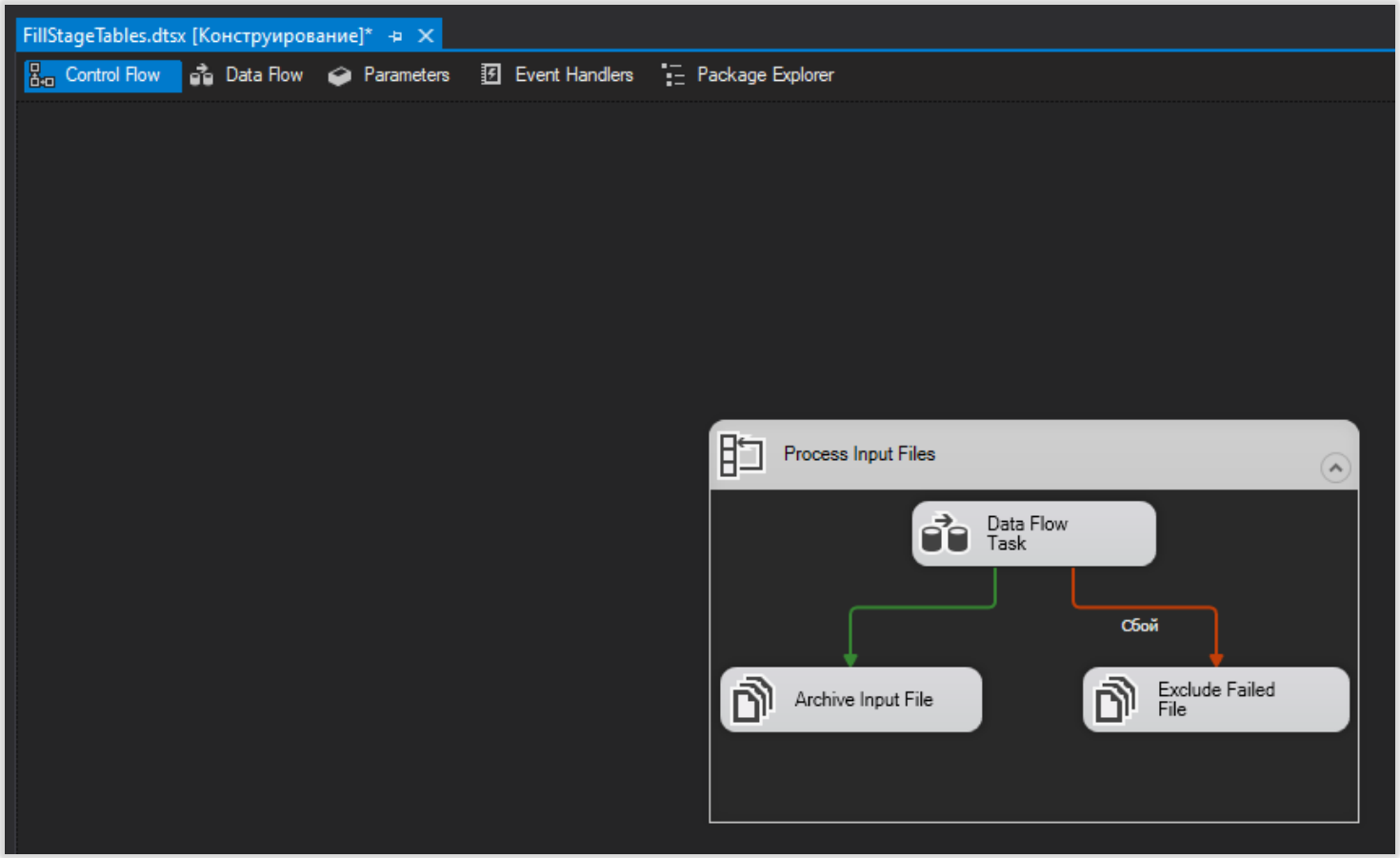
Типовая последовательность шагов процесса ETL

- Определение данных, которые следует переместить в хранилище.
- Определение всех внешних и внутренних источников данных.
- Подготовка mapping для всех элементов данных из источников.
- Формирование правил извлечения, трансформации и загрузки данных (см. файл template_ETL_processes.doc).
- План агрегации данных в таблицах (где это необходимо).
- Написание кода ETL процедур (интерфейс или, например Python).
- ETL для таблиц измерений.
- ETL для таблиц фактов.

Результат ETL – автоматизированные «пайплайны»



Пайплайн SSIS



Процесс идентификации ресурсов (extract)

- Составление простого списка метрик или фактов для всех таблиц фактов в хранилище.
- Составление простого списка атрибутов измерений для всех измерений.
- Для каждого элемента из обоих списков подобрать актуальный источник данных из имеющихся на входе (для одного элемента может быть несколько источников, выбирать доверенный).
- Если в одно значение хранилища приходят объединенные данные из нескольких источников (дата или id), сформировать правила объединения. В обратном случае, сформировать правила разбиения (операторы SQL, внимание!).
- Определить значения «по умолчанию».

Основные задачи трансформации данных

- ***Выбор*** – формирование массивов данных из источников (целиком или фрагментарно), для переноса в хранилище.
- ***Разделение/соединение данных*** (часто – соединение, редко - разделение).
- ***Конверсия*** – большой набор задач, применяемых к данным в массивах с целью стандартизации данных из разных источников, а также повышения «понятности» и «полезности» данных для пользователей.
- ***Агрегация*** – добавление свойства «гранулярности» к данным в массивах.
- ***Насыщение*** – изменение порядка, упрощение столбцов с данными в таблицах с целью более удобного их представления.

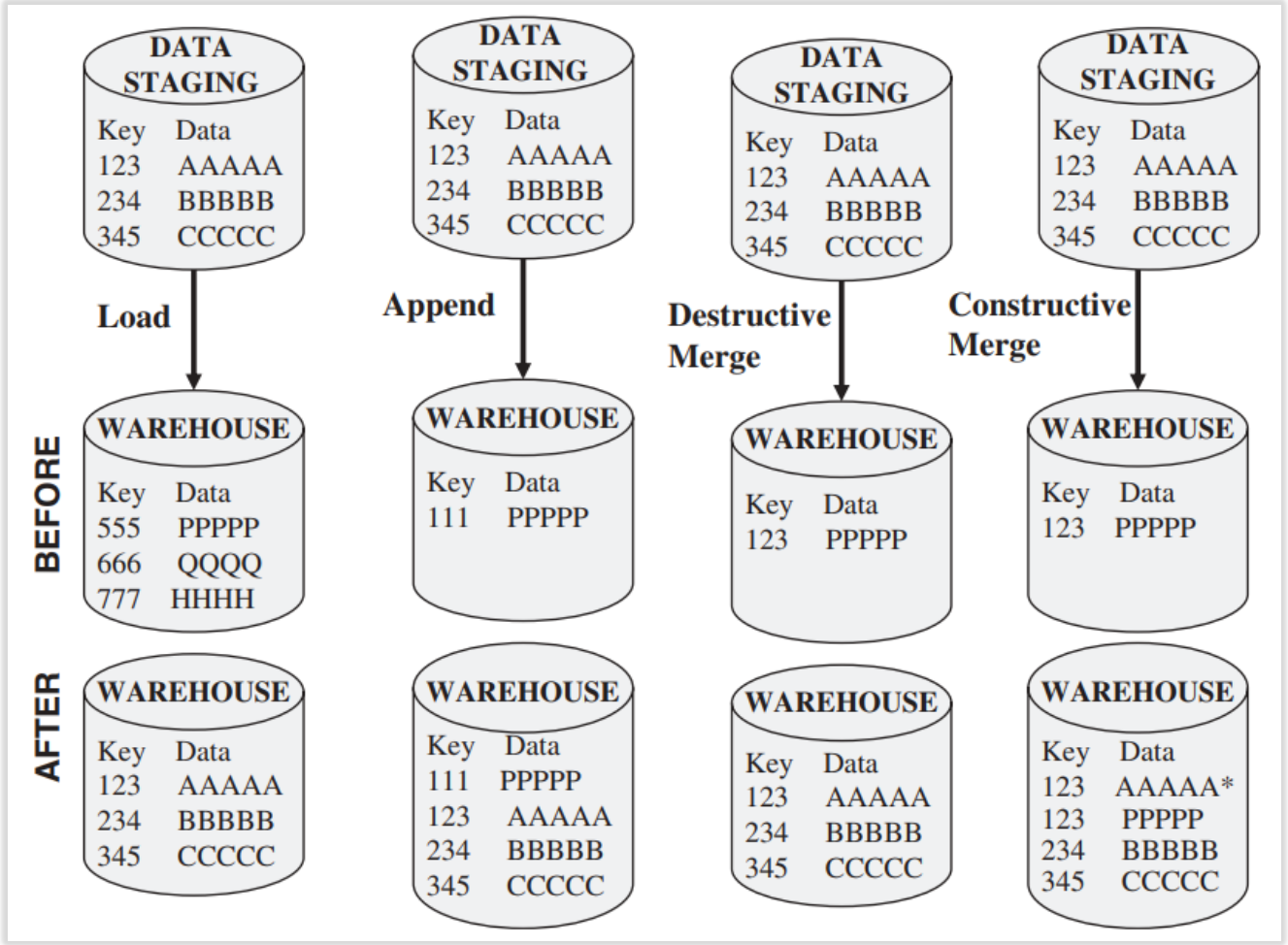
Требования к процессу загрузки данных (Load)

- Учет требований бизнеса по длительности итераций процессов загрузки данных в хранилище (загрузка в течение недели готовых в исходных системах наборов данных в 40 итераций по 1 часу в нерабочее время…).
- Проектирование технических справочников в случае загрузки данных «набегающей волной» (обновление одних и тех же наборов данных в течение последовательных периодов).
- Обеспечение «версионности» наборов данных за счет создания технических справочников для ведения конфигураций и обеспечения поточности загрузок.
- Деление пакетов данных по количеству исполнителей, ответственных за их заполнение (выделение «быстрых» и «медленных» пакетов данных для существенной оптимизации времени загрузки).

Техники загрузки данных в хранилище

- Простая загрузка. Обычная загрузка в пустые измерения и таблицы хранилища. Если измерения и таблицы не пустые, сперва они очищаются, затем грузятся данные.
- Загрузка добавлением. Грузятся только те данные, дубликатов которых еще нет в хранилище. Все спорные ситуации до дублирующийся данным решаются функциями или администратором.
- «Разрушительное слияние». Добавляет новое, заменяет старое.
- «Созидательное слияние». Добавляет новое, дополняет старое.

Техники загрузки данных в хранилище



Возможные ошибки при валидации данных. Перечисление и текст.

Внутри поля	По отношению к другим полям	Совместимость форматов при передаче между системами
Не из списка разрешенных значений. Отсутствие обязательных значений. Несоответствие формату (например, «Все договора должны нумероваться «ДГВхххх..»).	Не из списка разрешенных значений для связанного элемента. Отсутствие обязательных элементов для связанного элемента. Несоответствие формату для связанного элемента (например, «Для продукта «АИС» все договора должны нумероваться «АИСхххх..»).	Символы допустимые в одном формате, недопустимы в другом. Разная кодировка. Отсутствие маппинга при добавлении/изменении меты данных. Устаревшие значения (не из списка разрешенных в целевой системе).

Возможные ошибки при валидации данных. Числа и порядки.

Внутри поля	По отношению к другим полям	Совместимость форматов при передаче между системами
Не является числом. Не находится в границах разрешенного интервала значений. Пропущено порядковое значение (сбой в передаче пакета с данными).	Элементу «А» присвоен неправильный порядковый номер. Разницы за счет разных правил округления значений (например, в 1С и SAP может не «сойтись» рассчитанный НДС).	Переполнение (по ограничению). Потеря точности и знаков. Несовместимость форматов при конвертации в тип, отличный от числового.

Возможные ошибки при валидации данных. Даты и периоды.

Внутри поля	По отношению к другим полям	Совместимость форматов при передаче между системами
	День недели не соответствует дате. Сумма единиц времени не соответствует действительности, из-за разницы рабочие/не рабочие/праздничные/сокращенные дни.	Несовместимость формата даты при передаче текстом (например: ISO 8601 в UnixTime, или разные форматы в ISO 8601). Ошибка точки отсчета и точности при передаче числом (например: TimeStamp в DateTime).

Самостоятельное изучение и кейсы

- Ponniah – Data warehousing. Fundamentals for IT Professionals, стр. 284-308.
- Что такое ETL: как справиться с анализом big data, https://mcs.mail.ru/blog/chto-takoe-etl-ili-kak-spravitsya-s-analizom-big-data
- Как ETL-процессы помогают анализировать большие данные, https://practicum.yandex.ru/blog/chto-takoe-etl/
- Основные функции ETL-систем, https://habr.com/ru/post/248231/
- Apache Airflow: делаем ETL проще, https://habr.com/ru/post/512386/

Спасибо за внимание!