Хранилища данных. Лекция 6. Интеграция информационных ресурсов в хранилищах данных

Антон Викторович Кудинов,
доцент кафедры ВТ

Содержание

- Проблема интеграции данных
- Что такое SQL Server 2005 Integration Services
- Планирование ETL проекта для хранилища данных

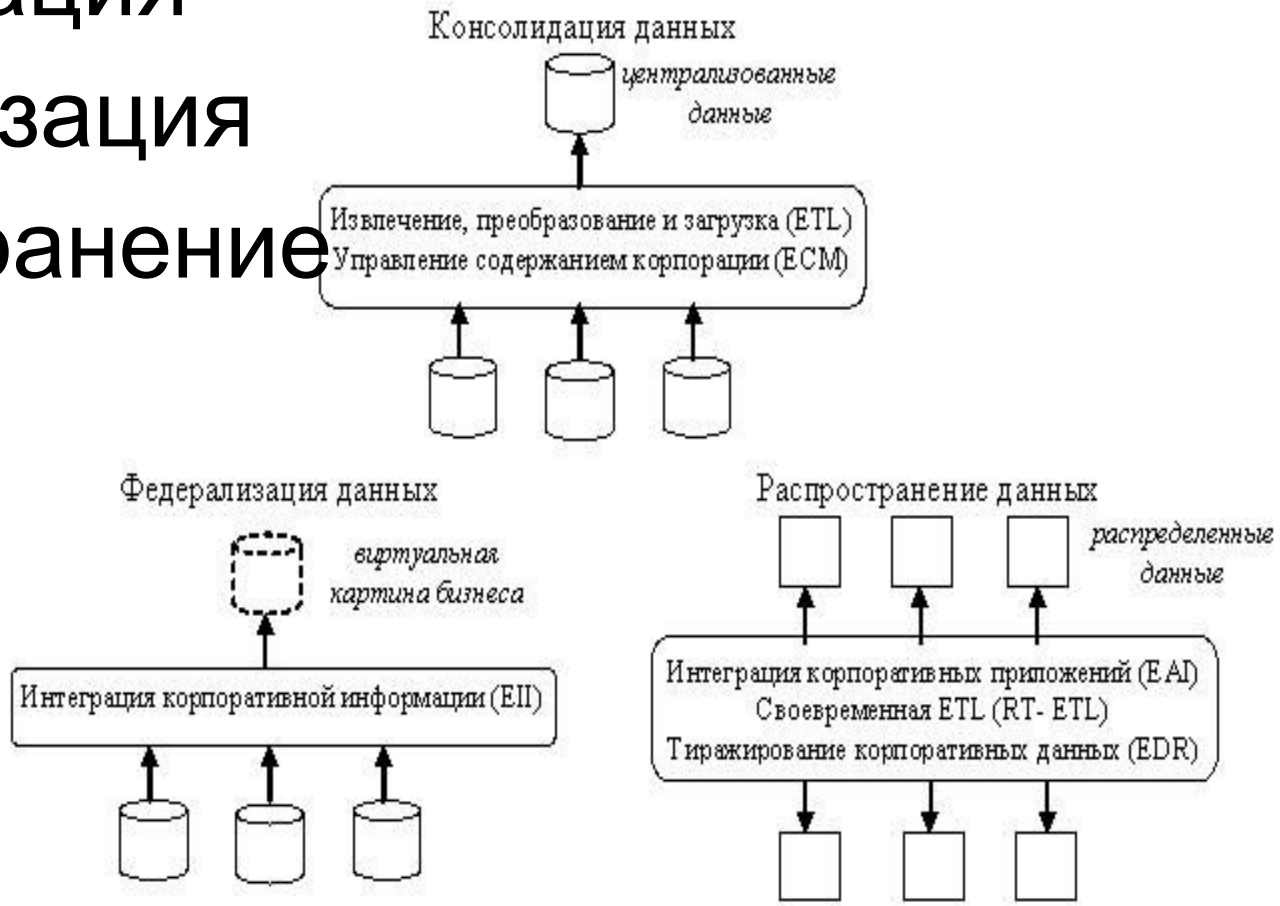
Проблема интеграции данных

ETL процессы (Extraction, Transformation, Load)
60-80% времени

- Извлечение и очистка данных
- Трансформации данных
- Загрузка данных в хранилище

Три метода интеграции данных

- Консолидация
- Федерализация
- Распространение



Консолидация данных

- Данные собираются из нескольких первичных систем и интегрируются в одно постоянное место хранения. Такое место хранения может быть использовано для подготовки отчетности и проведения анализа, как в случае хранилища данных, или как источник данных для других приложений.

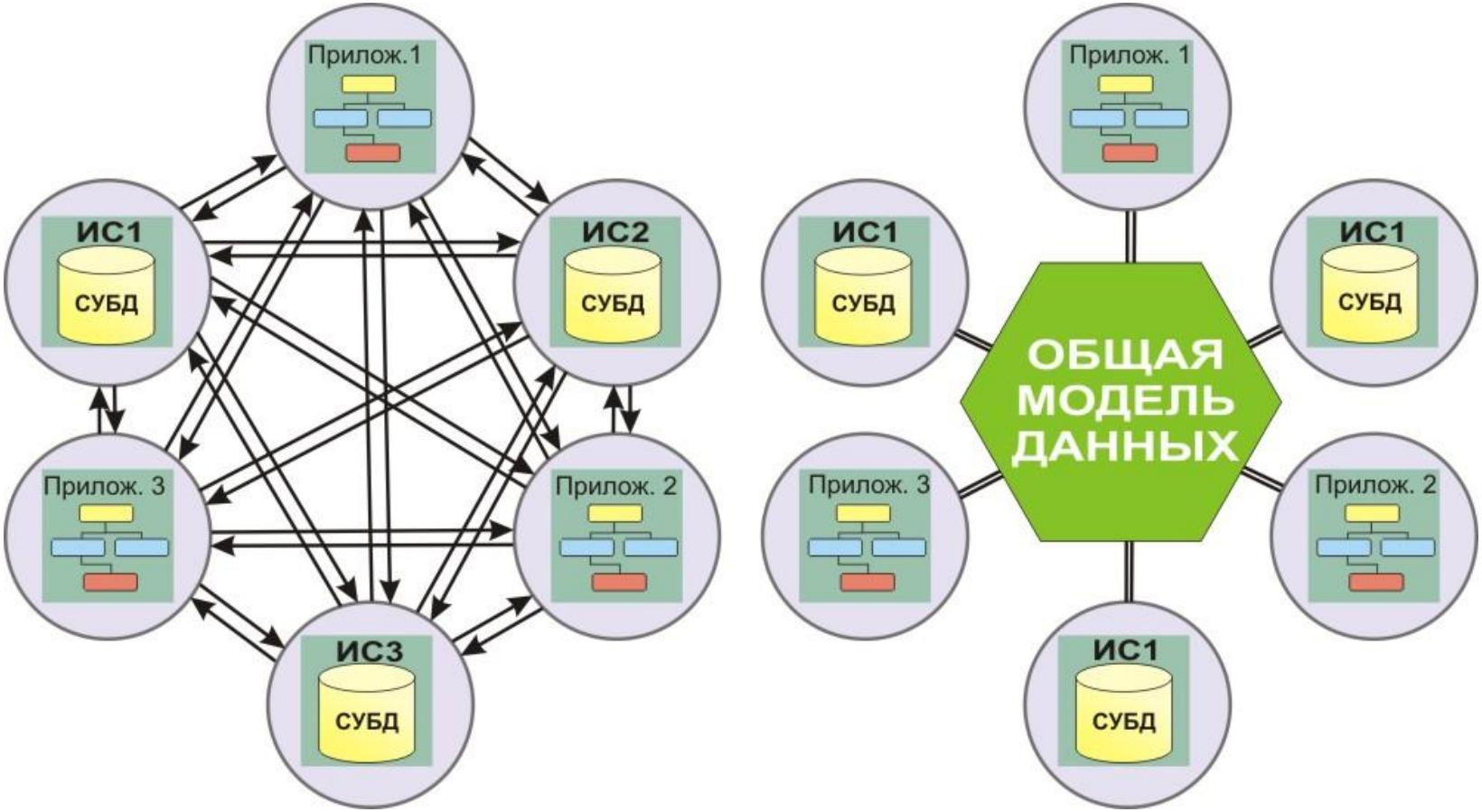
Федерализация данных

- Обеспечивает единую виртуальную картину нескольких первичных источников данных. Для получения сведений о некотором процессе, обрабатываемом в нескольких оперативных приложениях, процессор федерализации данных извлекает данные из соответствующих первичных складов данных, интегрирует их таким образом, чтобы они отвечали виртуальной картине и требованиям запроса, и отправляет результаты бизнес-приложению, от которого пришел запрос.

Распространение данных

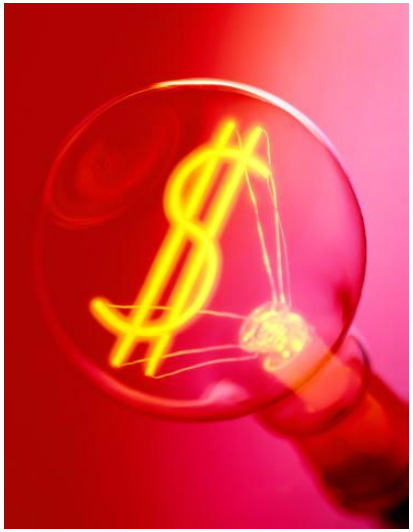
- Подразумевает их копирование из одного места в другое. Этот подход обычно используется для операций реального времени и базируется на механизмах "проталкивания", т. е. является событийно управляемым.

Интеграция на основе метамодели



Задачи при интеграции данных

- Технологические
- Организационные
- Экономические



Технологические задачи

- Гетерогенные источники данных с различным форматами
- Структурированные, полуструктурированные и неструктурированные данные
- Данные поступают в разное время
- Очень большие объемы данных
- Качество данных (пропуски, нет смысла, ошибки)
- Придание смысла данным при слиянии их из разных форматов при неполноте данных в отдельных источниках
- Преобразование данных в унифицированный формат, пригодный для бизнес-анализа

Технологические требования

- Загрузка данных в наибыстрейшее время (нет возможности «ночного» периода, 7 x 24 часа On-Line)
- Потребность загрузки данных в несколько приемников практически одновременно
- Постоянная доступность данных с минимальными задержками в актуальности данных
- Разнообразие источников данных (OLTP, OLAP, веб-сервисы, неструктурированные данные, унаследованные системы)
- Разнообразие приемников данных (порталы, персонализированные отчеты, PDA, мобильные телефоны)
- Масштабируемость и производительность

Организационные задачи

- Получение серьезной поддержки руководства компании команде по проекту интеграции данных, настоять на координации и компромиссах по выбору форматов данных и бизнес-процессов получения данных в подразделениях компании
- Определиться с единообразными технологиями для разного круга задач, так как многие подразделения используют совершенно разные системы и способы. Люди консервативны в своих привычках, не любят переучиваться. До 60% времени при получении и интеграции данных – ручной процесс

Экономические задачи

Интеграция данных – дорогостоящий процесс.

Факторы, увеличивающие стоимость проекта:

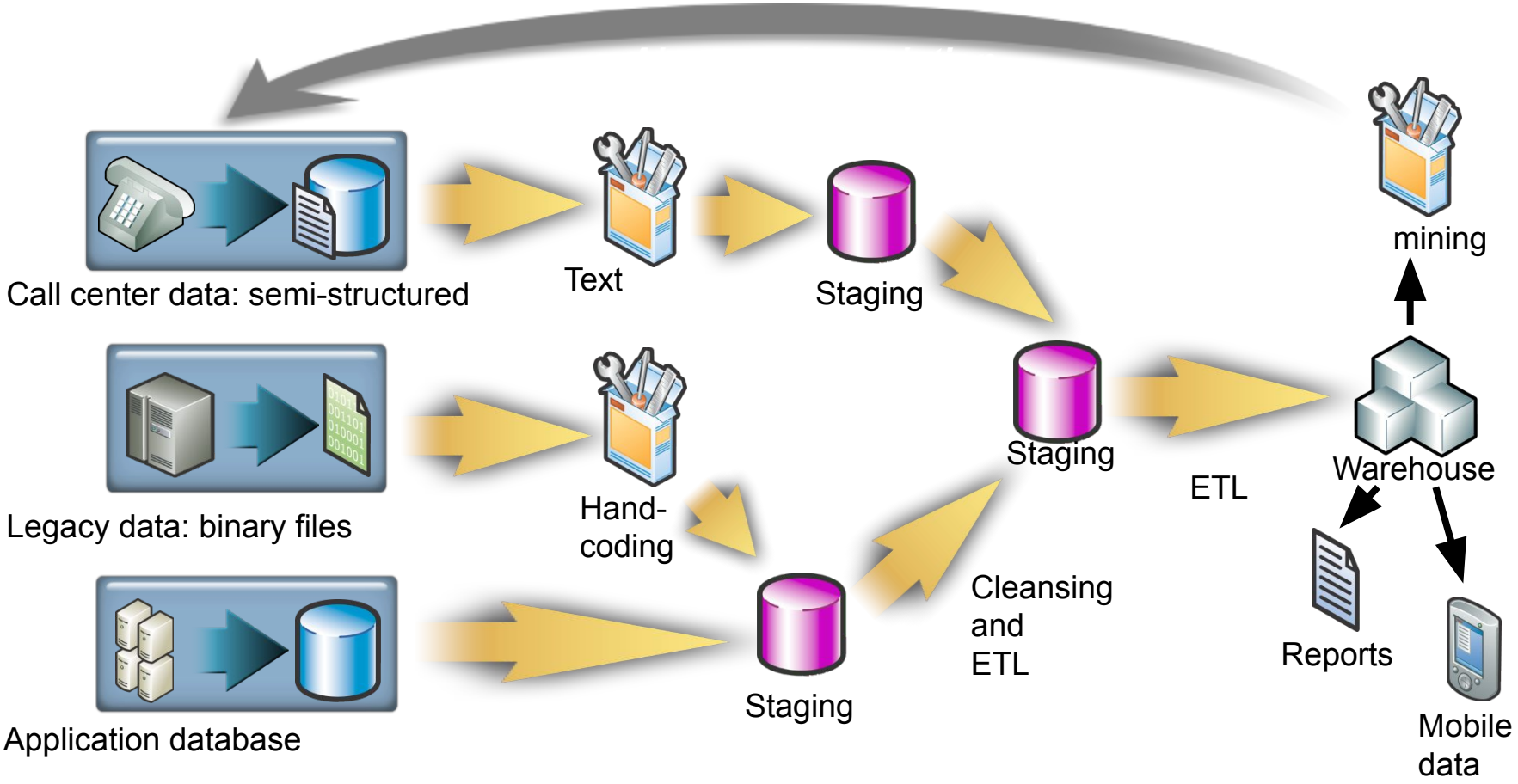
- Административные преграды, недостаток координации, недостаточная поддержка руководства
- Недостаточная функциональность имеющихся средств для ETL процессов, необходимость разработки нового ETL кода

SQL Server 2008 Integration Services

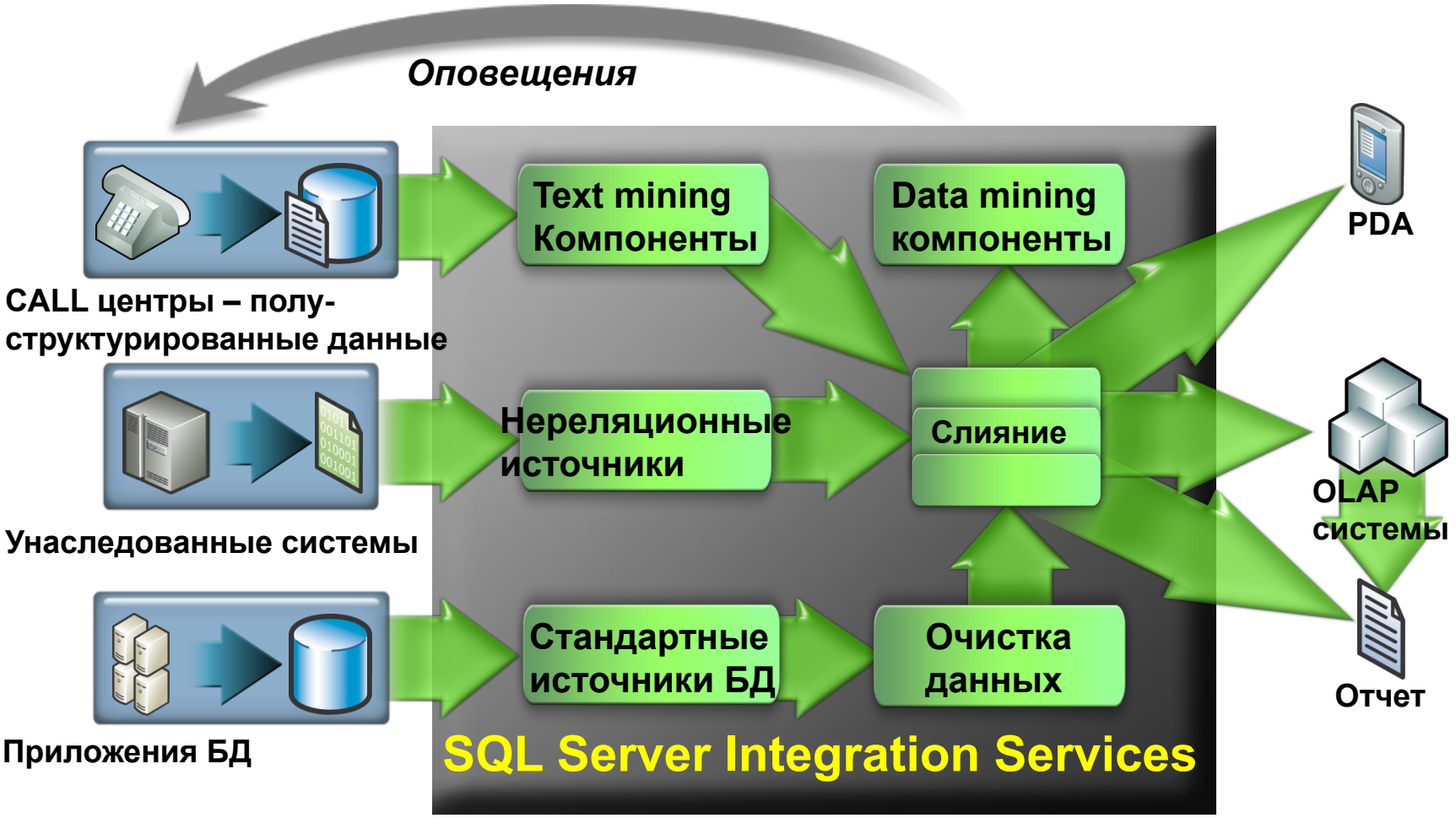
Службы Integration Services - платформа для построения высокопроизводительных решений интеграции данных и решений потока операций, включая операции извлечения, преобразования и загрузки (ETL) для хранилищ данных.

- Графические инструменты
- Мастера для построения и отладки пакетов
- Источники данных для извлечения данных
- Источники назначения для загрузки данных
- Преобразования для очистки, статистической обработки, слияния и копирования данных
- Задачи для выполнения функций потока операций
- Служба управления и администрирования пакетов
- API-интерфейсы для программирования объектной модели

До Integration Services



Integration Services 2008



Архитектура SQL Server 2008 Integration Services

Термины

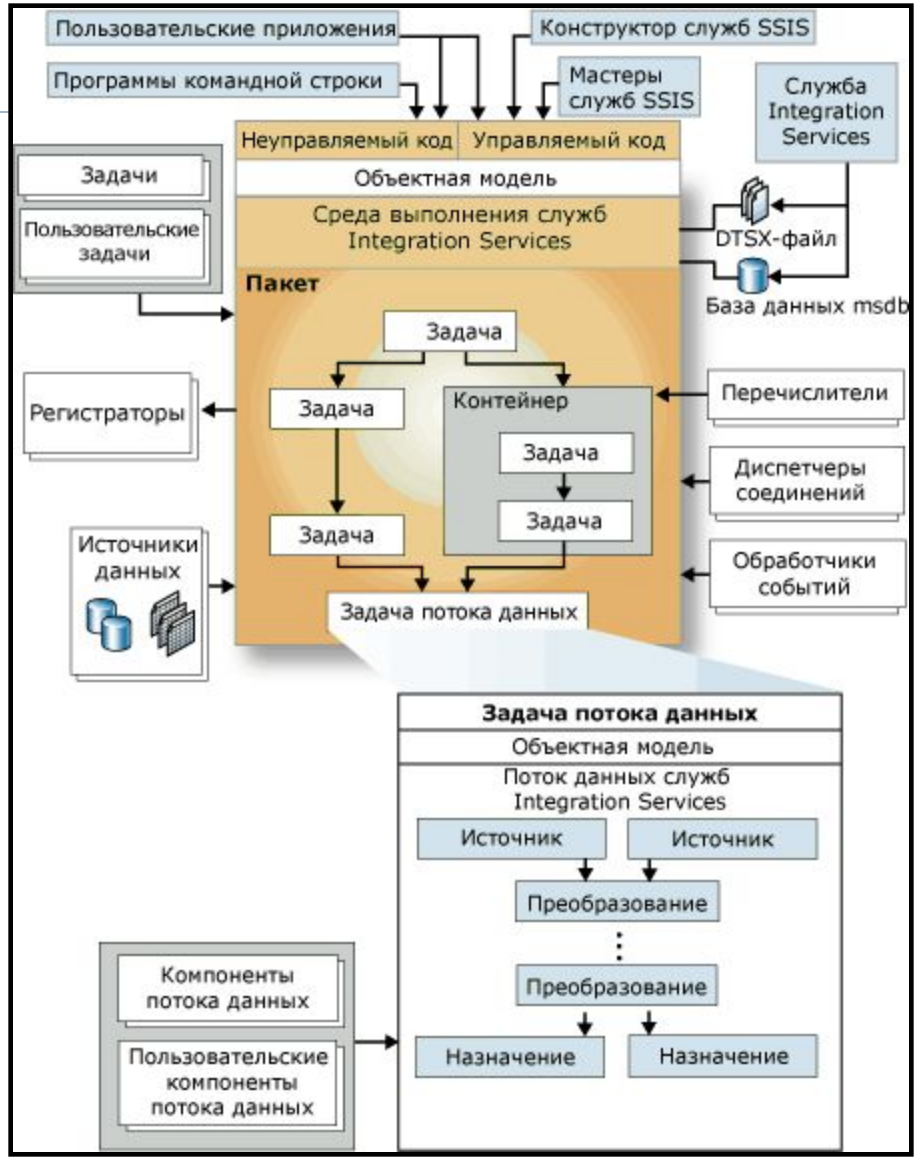
- Источник (и) - Sources
- Приёмник(и) - Destinations
- Преобразование данных - (Transformation)

- Время исполнения

- Пакет (Package)
- Задача (Task)
- Буфер (Buffer)
- Труба (pipeline) потока данных

Конструктор служб SSIS

- Поток управления (Control Flow)
- Поток данных (Data Flow)
- Обработчики событий в пакете и объектов пакета (Event Handlers)
- Просмотр содержимого пакета
- Просмотр выполнения пакета



Типовые сценарии в Integration Services

- Слияние данных из гетерогенных хранилищ данных
- Очистка, преобразование и стандартизация данных
- Заполнение хранилищ данных и витрин данных
- Встраивание бизнес-аналитики в процесс преобразования данных
- Автоматизация административных функций и загрузки данных

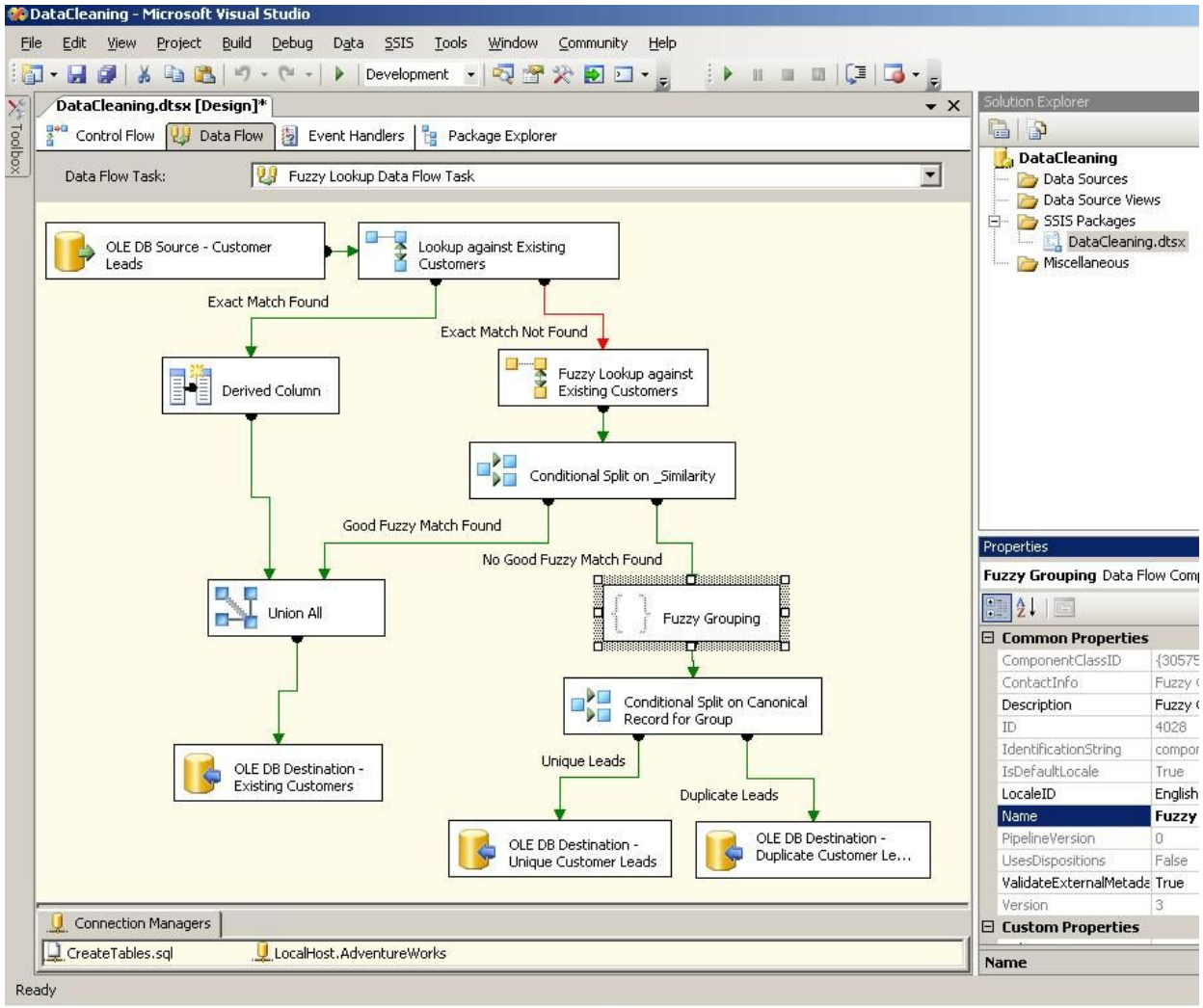
Пример: Очистка данных

Пакет SSIS

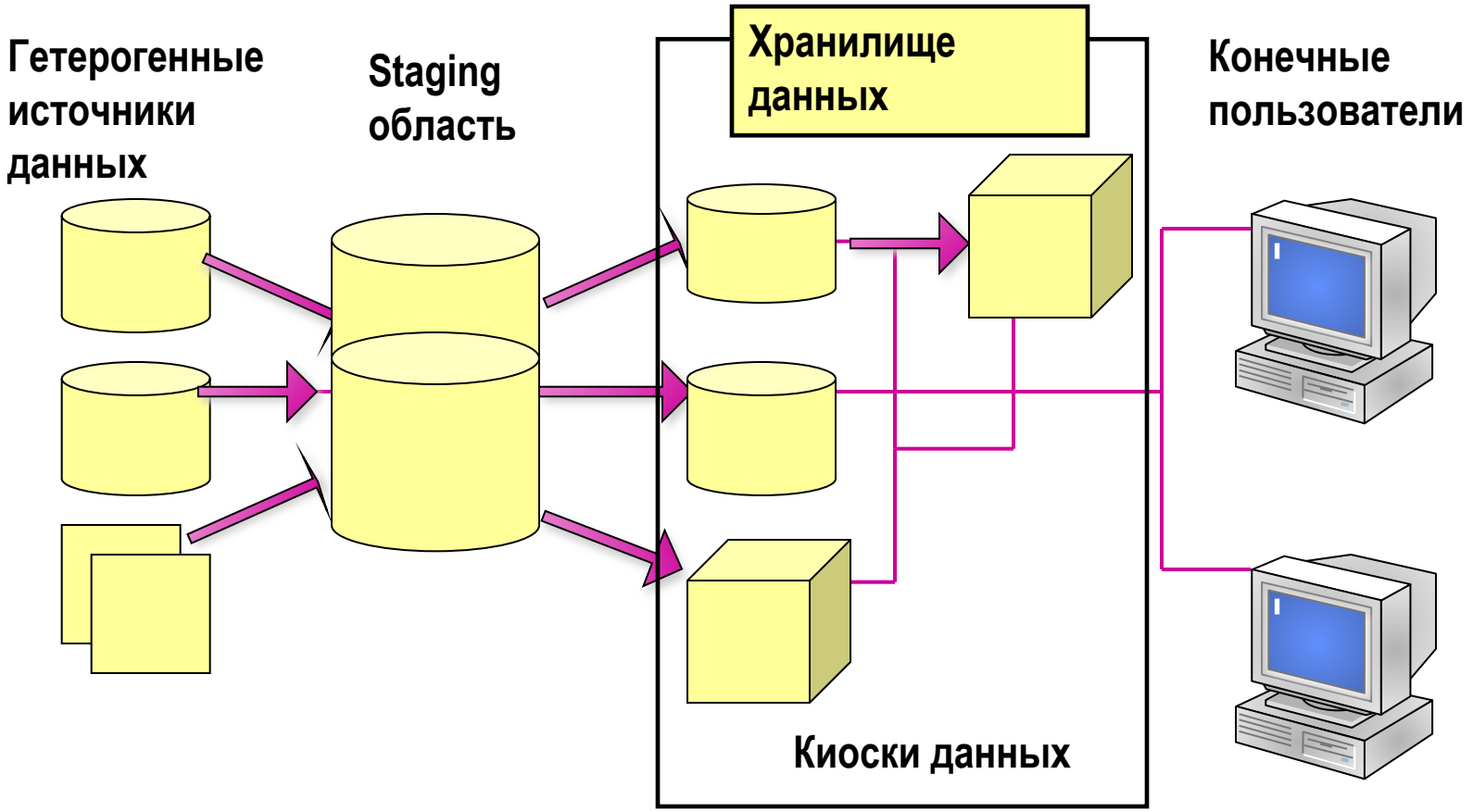
Data Cleaning Sample из Integration Services Samples.

Fussy Lookup – нестрогое соответствие новых клиентов старым записям

Fussy Grouping – нечеткий поиск фамилий дубликатов.



Планирование ETL проекта для хранилища данных



Заполнение хранилища данных в SSIS

- Источники и приемники данных
- Оценка и проверка исходных данных
- Промежуточное хранение данных (Staging storage)
- Загрузка в хранилище и киоски данных

Источники и приемники данных

- Выбрать источники данных (все форматы)
- Выбрать приемники данных (DW, Data Mart), определить структуру записываемых данных
- Определить время извлечения и записи данных (extraction and load windows), длительность извлечения и загрузки данных
- Документировать диаграмму потока данных: описать список источников, методов доступа, учетные записи, протоколы, характеристики сети

Промежуточное хранение данных (Staging storage)

В сложных ETL процессах может потребоваться промежуточное хранение данных после чтения перед загрузкой в хранилище:

- Реляционная БД
- Файлы «как есть» - raw (binary) files

После извлечения данных:

- Необходимость быстро освободить источник данных
- Выполнение ETL с заданной контрольной точки без повторного рестарта

Перед загрузкой данных:

- Асинхронное поступление данных, ожидание всех данных
- Фиксируется моментальный снимок данных на заданную дату, возможность получения отчетности по этому снимку данных
- Возможность рестарта с контрольной точки без необходимости выполнять пакет с самого начала
- Возможность провести трансформацию некоторых данных на SQL Server перед окончательной загрузкой в хранилище
- Возможность проверить и удалить невалидные данные или дубликаты после окончания трансформаций перед загрузкой

Загрузка в хранилище и киоски данных

- Загрузка измерений и мер
- Создание первичных и вторичных ключей
- Создание индексов
- Удаление временных таблиц
- Обработка измерений и секций кубов

Спасибо за внимание!

- KudinovAV@tpu.ru