Инструментарий хранилищ данных. Управление метаданными.

Лекция №8 для студентов 4-го курса специальности «Прикладная информатика»

Вопросы

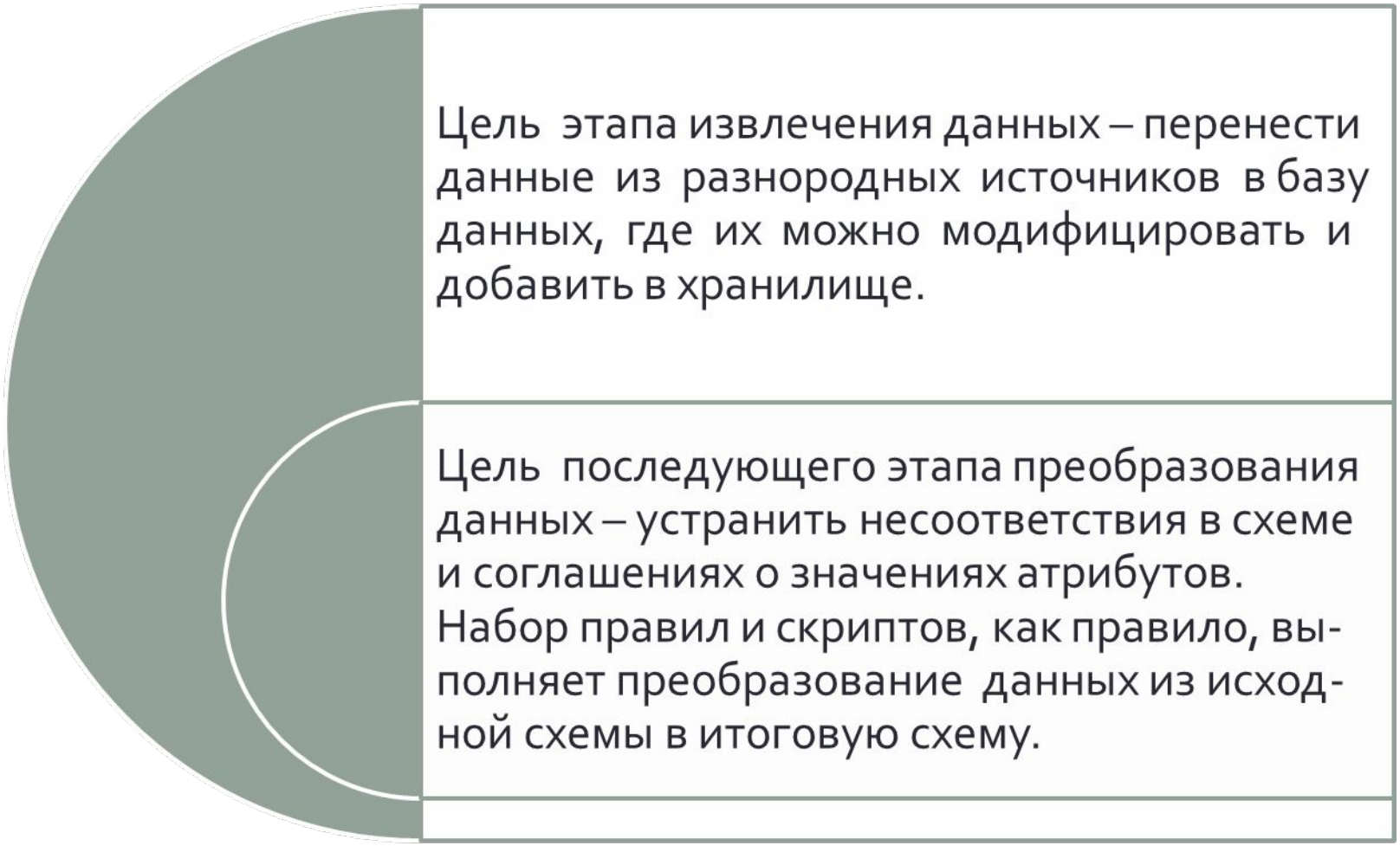
1) Инструментарий хранилищ данных.
2) Управление метаданными.

1 Инструментарий хранилищ данных

Создание хранилища данных из независимых источников данных — многоэтапный процесс, который предусматривает извлечение данных из каждого источника, преобразование их в соответствии со схемой хранилища данных, очистку, а затем загрузку в хранилище.

Data Warehousing Information Center опубликовал обширный список инструментальных средств ETL (extract, transform, load — «извлечение, преобразование, загрузка»), выполняющих эту последовательность операций.

1.1 Извлечение и преобразование



1.2 Очистка данных

Ошибки при вводе данных и различия в схемах могут привести к тому, что таблица измерений «Клиент» будет иметь несколько соответствующих кортежей для одного клиента, что приводит к неточным ответам на запросы и некорректным моделям добычи данных.

К примеру, если таблица клиентов содержит по несколько кортежей для некоторых клиентов FSC в Нью-Йорке, то Нью-Йорк может ошибочно попасть в список первых 50 стран с самым большим числом индивидуальных клиентов.

Инструменты, которые помогают определить и исправить аномалии данных, могут иметь высокую отдачу; значительное число исследований посвящено проблемам устранения дублирования и инструментам очистки данных.

1.3 Загрузка

После того, как данные извлечены и преобразованы, возможно, что их еще необходимо дополнительно обработать перед тем, как добавить в хранилище. Как правило, утилиты фоновой загрузки поддерживают такие функции, как

- проверка ограничений целостности;
- сортировка;
- суммирование,
- агрегирование и
- выполнение других вычислений для создания производных таблиц, размещаемых в хранилище;
- создание индексов и других способов доступа.

Помимо наполнения хранилища, утилита загрузки должна позволять системным администраторам проверять статус; отменять, приостанавливать и возобновлять загрузку; возобновлять работу после ошибки без потери целостности данных. Поскольку утилиты загрузки для хранилищ данных обрабатывают значительно больше данных, чем содержится в транзакционных системах, они используют разного рода алгоритмы распараллеливания.

1.4 Обновление

Обновление хранилища данных состоит в распространении обновлений на исходные данные, которые соответственным образом обновляют базовые таблицы и производные данные, материализованные представления и индексы, размещенные в хранилище. Должны быть рассмотрены два вопроса: *когда обновлять* и *как обновлять*.

Обычно хранилища данных обновляются периодически в соответствии с заранее установленным расписанием, например, ежедневно или еженедельно.

Распространять каждое обновление необходимо только в том случае, если для выполнения OLAP-запросов требуются текущие данные. Администратор должен выбрать циклы обновления таким образом, чтобы накладные расходы, вызванные обработкой больших объемов данных, не превысили расходы на выполнение утилиты инкрементальной загрузки.

2 Управление метаданными

Метаданные – информация любого рода, которая требуется для управления хранилищем данных, а управление метаданными – существенный компонент архитектуры хранения. К административным метаданным относится вся информация, которая требуется для настройки и использования хранилища данных.

Бизнес-метаданные включают в себя бизнес-термины и определения, принадлежность данных и правила оплаты услуг хранилища.

Оперативные метаданные – это информация, собранная во время работы хранилища данных, такая как происхождение перенесенных и преобразованных данных; статус использования данных; данные мониторинга.

Согласованные усилия коммерческих компаний и научных кругов привели к серьезному технологическому прогрессу в решении задач хранения данных. Это нашло отражение во множестве коммерческих продуктов, которые доступны для каждой из трех основных операций:

- пополнение хранилища данных из независимых транзакционных систем;
- хранение данных и управление ими;
- анализ данных с целью принятия обоснованных бизнес-решений.

Однако, несмотря на изобилие коммерческого инструментария, остается еще несколько важных направлений для исследования.

Очистка данных связана с интеграцией данных из неоднородных источников, проблемой, которую изучают уже много лет. На сегодняшний день основные усилия концентрируются на проблемах несогласованности данных.

Хотя очистка данных в последнее время привлекает большое внимание исследователей, предстоит еще немало сделать для создания инструментальных средств, не зависящих от предметной области, которые решают разнообразные проблемы очистки данных, связанные с разработкой хранилищ.

Большая часть исследований в области добычи данных касается разработки алгоритмов для создания более точных моделей или алгоритмов, позволяющих ускорить этот процесс.

Два других этапа процесса выявления знаний – подготовка данных и применение модели добычи данных – по большей части игнорируются.

На обоих этапах возникает несколько проблем, в частности, связанных с достижением большей гармонии между системами управления базами данных и технологией добычи данных.

В конечном итоге, новые инструментальные средства должны дать аналитикам более эффективные способы подготовки наборов данных, отвечающих конкретной цели, и более эффективные способы применения моделей к результатам произвольных SQL-запросов.