



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«МИРЭА – Российский технологический университет»

РТУ МИРЭА

Лекция №9

Создание витрин данных

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. **Концепция витрин данных**
2. **Виды витрин данных**
3. **Структуры витрин данных**
4. **Принципы построения витрин данных**

1. Концепция витрин данных

Витрины Данных представляют собой специальные структуры данных, которые позволяют организациям хранить, анализировать и получать доступ к различным видам информации, необходимой для принятия обоснованных бизнес-решений. Мы изучим основные компоненты витрин, их различные типы, а также процесс их реализации и преимущества использования в современных организациях.

В 1993 году Тед Кодд разработал OLAP для объединения различных транзакционных систем, но из-за низкой эффективности его использование было недолгим. Ученые начали разрабатывать новую платформу для аналитической информации, что привело к созданию специального хранилища (ХД, Data Warehouse).

С ростом объемов конфиденциальных сведений организации столкнулись с проблемами хранения и обработки, что требовало минимизации убытков и сокращения времени борьбы с неполадками.

Для решения проблемы была создана витрина статистических данных (ВД, Data Mart), содержащая часть информации из хранилища. Новая информация добавлялась в неё в периоды меньшей активности пользователей. При потере или повреждении данных в ВД запускалась процедура восстановления из хранилища.



ВД является частью хранилища и содержит информацию по определённой теме. Существуют отдельные витрины для разных специалистов. Они предназначены для хранения информации в небольших объемах и связана с определенным подразделением или направлением деятельности организации.

Витрины данных содержат подмножество информации в хранилище данных. По этой причине витрины данных обеспечивают лучшую скорость запросов для аналитиков, поскольку они, естественно, содержат меньше данных.

Размер витрины данных будет отличаться в каждой функциональной области, поскольку некоторые витрины данных требуют значительного объема хранилища. Витрины данных - это форматы, доступные только для чтения, с контролируемым обновлением, выполняемым только уполномоченным персоналом.

Хранилище данных	Витрина данных
Хранилище данных используется для хранения данных из множества предметных областей.	Витрина данных содержит данные, относящиеся к отделу, например витрину данных отдела кадров, маркетинга, финансов и т. д.
Он действует как центральное хранилище данных компании.	Это логический подраздел хранилища данных для конкретных приложений ведомства.

Он разработан с использованием схемы звезды, снежинки, галактики или созвездия фактов. Однако наиболее широко используется звездная схема.	Витрины данных используют звездообразную схему для проектирования таблиц.
Его сложно спроектировать и использовать из-за большого размера (более 100 ГБ).	Сравнительно более управляем из-за небольшого размера (менее 100 ГБ).
Предназначен для поддержки процесса принятия решений в компании.	Витрины данных предназначены для определенных групп пользователей или корпоративных подразделений.
Хранилища данных используются для хранения подробной информации в денормализованной или нормализованной форме.	Витрины данных содержат сильно денормализованные данные в обобщенной форме.
Имеет большие размеры и объединяет данные из многих источников.	Меньшие размеры для интеграции наборов данных из меньшего количества источников.
Хранилища данных являются предметно-ориентированными и зависят от времени, при этом данные существуют в течение более длительного периода времени.	Витрины данных используются для определенных областей, связанных с бизнесом, и сохраняют данные в течение более короткого периода времени.

Как правило, витрины данных состоят из следующих компонентов:

Исходные данные. Это информация, которая является отправной точкой для создания ВД. Она часто извлекается из различных источников, таких как операционные системы, внешние источники информации и хранилища сведений.

Подготовка. Этот этап включает в себя процесс преобразования и подготовки для использования в витринах. Он включает в себя операции извлечения, преобразования и загрузки данных (ETL).

Хранилище. Место, где преобразованные сведения сохраняются для последующего использования. Этот компонент включает в себя информационную модель, которая определяет, как данные организованы и структурированы для удобства доступа.

Инструменты доступа. Это приложения и инструменты, которые конечные пользователи используют для запросов и анализа данных в ВД. Они предоставляют возможность работы с данными и получения необходимой информации для принятия решений.

2. Виды витрин данных

Существует три основных типа витрин данных:

Зависимые. Эти ВД создаются путем извлечения информации напрямую из операционных систем, внешних источников или обоих.

Независимые. Эти ВД создаются без использования центрального хранилища.

Гибридные. Этот тип витрин может получать информацию как из хранилищ, так и из операционных систем.

Рассмотрим каждую из них более подробно.

Зависимая ВД

Зависимая ВД представляет собой инструмент, который используется для получения сведений организации из единого хранилища. Этот тип обеспечивает преимущества централизации данных. Если вам требуется создать одну или несколько физических ВД, вам следует настроить их как независимые витрины.

Построение зависимой ВД в хранилище можно выполнить двумя способами:

Во-первых, пользователь может получить доступ как к самой витрине данных, так и к хранилищу в целом в зависимости от конкретной потребности.

Во-вторых, доступ к сведениям может быть ограничен только через ВД.

Однако второй подход не является оптимальным, поскольку он может привести к недостаточной эффективности в использовании данных и созданию «свалки данных». Понятие «свалка данных» означает, что все данные начинаются с общего источника, но в конечном итоге они могут быть удалены или даже утеряны, что приводит к потере ценной информации.

Независимая ВД

Независимый **киоск данных** представляет собой витрину данных, которая создается без привлечения центрального хранилища данных. Этот тип ВД является отличным выбором для небольших групп внутри организации, которые нуждаются в самостоятельном доступе к данным.

Независимая ВД не связана ни с центральным хранилищем данных компании, ни с другими ВД. Данные в независимой витрине вводятся отдельно, и анализ проводится независимо от других источников данных.

Внедрение независимых ВД может противоречить целям создания централизованного хранилища данных. В первую очередь, организации требуется единое и централизованное хранилище корпоративных данных, которое может обслуживать множество пользователей с различными потребностями, требующими доступа к разнообразной информации.

Гибридный киоск данных

Гибридный киоск данных представляет собой ВД, которая объединяет входные данные из различных источников, отличных от центрального хранилища данных. Этот тип витрины данных может быть особенно полезен в случаях, когда требуется временная интеграция, например, после добавления новой группы или продукта в организацию.

Гибридный киоск данных является примером, который хорошо подходит для разнообразных сред баз данных и обеспечивает быструю реализацию в любой организации. Он также требует минимальных усилий по очистке данных.

Гибридная ВД также обладает поддержкой больших структур хранения и наилучшим образом подходит для гибких и небольших приложений, ориентированных на данные.

Разница между хранилищем и витриной данных

Использование витрин данных, озер данных и хранилищ данных зависит от ситуации и целей.

Хранилище данных — это система управления данными, которая обеспечивает анализ бизнес-данных и аналитику для всей компании. Хранилища данных обычно содержат большие объемы данных, включая исторические данные, и получают информацию из разных источников, таких как журналы приложений и транзакционные приложения. Данные в хранилище данных структурированы и используются с определенными целями.

Озеро данных позволяет организациям хранить большие объемы как структурированных, так и неструктурированных данных (например, данные из социальных сетей или о посещениях) и предоставлять к ним доступ для выполнения аналитики реального времени, глубокого анализа данных и построения сценариев использования машинного обучения. *Данные поступают в озеро данных в своей первоначальной форме без изменений.*

Основное различие между озером данных и хранилищем данных заключается в том, что озеро данных содержит большие объемы необработанных данных без предварительно определенной структуры. Организации не нужно определять заранее, как будут использоваться эти данные.

ВД — это упрощенная форма хранилища данных, которая сосредоточена на определенной теме или направлении деятельности, таких как продажи, финансы или маркетинг. По сравнению с хранилищами данных, ВД собирают данные из меньшего количества источников, таких как внутренние операционные системы, центральное хранилище данных и внешние данные.

3. Структуры витрин данных

Структура витрины данных - это предметно-ориентированная реляционная база данных, которая хранит данные в виде таблиц, т. е. строк и столбцов, к которым легче получить доступ, организовать и понять. Поля данных могут относиться к одному или нескольким объектам.

Витрины данных структурированы по многомерной схеме, которая служит основой для анализа данных пользователями базы данных. Тремя основными структурами или схемами для витрин данных являются star, snowflake и vault.

1. Звезда

Звездообразная схема - это схема, напоминающая форму звезды и состоящая из таблиц фактов, которые ссылаются на таблицы измерений в реляционной базе данных. Таблица фактов размещена в центре звезды и содержит набор показателей, относящихся к конкретному процессу.

Схема star требует меньшего количества соединений при написании запросов, поскольку нет зависимости между таблицами измерений. Процесс запроса ETL делает его чрезвычайно эффективным для доступа к большим наборам данных и навигации по ним. Указанные преимущества делают звездообразные схемы широко используемыми в большинстве систем информационных технологий.

2. Снежинка

Схема snowflake расширяет схему star дополнительными таблицами измерений, которые нормализованы для защиты целостности данных и минимизации избыточности данных. Основное преимущество схемы snowflake заключается в том, что она требует меньше места для хранения таблиц измерений.

Однако структуру snowflake сложно поддерживать из-за множества таблиц, которые необходимо заполнить и синхронизировать. Это также отрицательно влияет на производительность в результате необходимости в дополнительных таблицах измерений.

3. Хранилище

Схема хранилища позволяет пользователям проектировать гибкие корпоративные хранилища данных. Это довольно современный метод моделирования баз данных. Схема хранилища представляет собой многоуровневую структуру, которая фокусируется на гибкости и масштабируемости.

Когда следует использовать витрины данных? Диаграммы данных рекомендуются для интерактивных рабочих нагрузок данных, предназначенных для сценариев

самообслуживания. Например, если нужно работать с финансами или учетом, то можно создавать свои собственные модели данных и коллекции, которые позволяют самостоятельно решать бизнес-вопросы и получать ответы с помощью T-SQL¹ и визуальных запросов.

Кроме того, эти коллекции данных могут быть использованы для создания отчетов в (Power BI²). Datamarts рекомендуются для клиентов, которым нужна ориентация на децентрализованное владение данными и архитектура, например для пользователей, которым нужны данные в качестве продукта или для создания платформы самообслуживания данных.

Витрины данных используются для поддержки следующих сценариев:

Самообслуживание отделов — централизация небольших и средних объемов данных (примерно 100 ГБ) в полностью управляемой базе данных SQL. ВД позволяют назначить одно хранилище для самостоятельного подчиненного отчета (например, Excel, отчеты Power BI и другие), тем самым сокращая инфраструктуру в решениях самообслуживания.

Аналитика реляционных баз данных с помощью Power BI — доступ к данным ВД с помощью внешних клиентов SQL. Azure Synapse³ и другие службы и инструменты, использующие T-SQL, также могут использовать диаграммы данных в Power BI.

Создание комплексных семантических моделей — позволяет создателям Power BI разрабатывать сложные решения без зависимости от других инструментов или ИТ-команд. Datamarts избавляют от управления оркестрацией между потоками данных и семантическими моделями с помощью автоматически созданных семантических моделей, обеспечивая визуальные возможности для запроса данных и неструктурированного анализа, все поддерживаемые базой данных SQL Azure.

4. Принципы построения витрин данных

Шаги в реализации витрины данных

Внедрение ВД — это важное, но сложное занятие. Рассмотрим этапы процесса реализации:

1. Проектирование

Планирование — это начальный этап процесса внедрения ВД. Он включает в себя все этапы, начиная от постановки цели создания витрины данных и заканчивая

¹ Transact-SQL (T-SQL) — это процедурное расширение языка SQL, созданное компаниями Microsoft (для Microsoft SQL Server) и Sybase (для Sybase ASE).

В Transact-SQL добавлены следующие возможности:

- управляющие операторы;
- локальные и глобальные переменные;
- различные дополнительные функции для обработки строк, дат, математики и т. п.;
- поддержка аутентификации Microsoft Windows.

Язык Transact-SQL является ключом к использованию MS SQL Server. Все приложения, взаимодействующие с экземпляром MS SQL Server, независимо от их реализации и пользовательского интерфейса, отправляют серверу инструкции Transact-SQL.

² Power BI (Business Intelligence) — это инструмент от компании Microsoft для бизнес-аналитики. Возможности Power BI:

Объединять разные источники данных и обрабатывать их на одной платформе.

Создавать интерактивные дашборды и отчеты.

Анализировать данные и составлять прогнозы.

³ Azure Synapse Analytics — это корпоративная служба аналитики, которая ускоряет извлечение аналитических сведений в разных хранилищах данных и системах больших данных. Она сочетает в себе технологии SQL, используемые в корпоративных хранилищах данных, технологии Apache Spark для больших данных, а также Azure Data Explorer для аналитики журналов и временных рядов.

сбором информации о требованиях к ней. В конечном итоге мы создаем логический и физический дизайн для ВД.

Этап планирования включает в себя следующие шаги:

- Сбор бизнес и технических требований и определение источников данных.
- Выбор соответствующего набора данных для витрины.
- Проектирование логической и физической структуры Datamarket.

Данные могут быть организованы в соответствии с различными критериями, такими как:

- Временные интервалы.
- Отделы или функциональные подразделения компании.
- Географические местоположения.
- Любая комбинация вышеперечисленных.

Сегментация данных может осуществляться на уровне приложения или СУБД. Хотя рекомендуется проводить сегментацию на уровне приложения, так как это позволяет лучше адаптироваться к изменениям в бизнес-среде.

Какие инструменты и технологии вам понадобятся? Для начала достаточно простого набора ручки и бумаги. Однако использование инструментов, позволяющих создавать диаграммы UML или ER (Entity-Relationship model, модель «сущность — связь»), также поможет добавить метаданные в ваши логические и физические проекты.

2. Построение

Этот этап представляет собой создание физической базы данных и логических структур, что является вторым шагом в реализации витрины данных.

Этот этап включает выполнение следующей задачи – реализация физической структуры базы данных, разработанной на предыдущем этапе. Например, создание объектов схемы базы данных, таких как таблицы, индексы, представления и другие.

Для успешной реализации ВД потребуются следующие продукты и технологии:

- **Система управления реляционными базами данных (СУРБД)**, которая обеспечивает хранение, добавление и удаление данных.
- **Быстрый доступ к данным с помощью SQL-запросов**, позволяющих легко извлекать информацию в зависимости от определенных условий.
- **Защита данных**, включая возможность восстановления после сбоев и резервное копирование данных.
- **Поддержка многопользовательского доступа**, обеспечивающая одновременный доступ нескольких пользователей к данным без конфликтов и перезаписи изменений.
- **Механизмы безопасности**, позволяющие регулировать доступ пользователей к данным и операциям с ними.

3. Наполнение

Третий этап включает в себя наполнение ВД данными. Этот шаг включает выполнение следующих задач:

- Сопоставление данных из источников с целевыми данными в витрине данных.
- Извлечение данных из источников.
- Очистка и преобразование данных для соответствия целевым структурам.
- Загрузка данных в ВД.

- Создание и хранение метаданных, включая информацию о происхождении данных, их актуальности, внесенных изменениях и уровне агрегации.

Для выполнения этих задач используются инструменты **ETL (Extract, Transform, Load)**. Эти инструменты позволяют просматривать источники данных, выполнять сопоставление данных, извлекать, преобразовывать и загружать данные обратно в витрину данных.

В процессе работы инструмент создает *метаданные, связанные с происхождением данных, изменениями и уровнем агрегации.*

4. Доступ

Шаг авторизации представляет собой четвертый этап и включает использование данных: запросы данных, создание отчетов, диаграмм и их публикацию. Конечный пользователь отправляет запросы в базу данных и просматривает результаты этих запросов.

Этот этап включает выполнение следующих задач:

- Настройка метаданных, которая преобразует структуры базы данных и имена объектов в понятные бизнес-термины. Это облегчает доступ к данным для пользователей без технических навыков.
- Настройка и поддержка структур базы данных.
- Настройка API и интерфейсов, если это необходимо.

Для доступа к ВД можно использовать командную строку или графический интерфейс. Предпочтительным является графический интерфейс, так как он позволяет легко создавать графики и более удобен для пользователей, чем командная строка.

5. Управление

Этот заключительный этап процесса внедрения ВД включает в себя такие административные задачи, как:

- Постоянное управление доступом пользователей.
- Оптимизация системы и настройка для улучшения производительности.
- Добавление и управление новыми данными в витрине данных.
- Планирование сценариев восстановления и обеспечение доступности системы в случае сбоя.

Для выполнения этих задач может использоваться графический интерфейс или командная строка для управления Datamarket.

Перенос витрин данных в облако

Для повышения гибкости и использования данных при разработке общей стратегии и принятии ежедневных решений, рабочие группы и отделы стремятся к более гибким методам работы. Однако превращение постоянно растущего объема сведений в статистические показатели часто представляется сложной задачей. Финансовые директора проводят в среднем по 2,24 часа в день, анализируя информационные таблицы.

Рабочие группы часто обращаются за помощью к отделу информационных технологий (ИТ), и ИТ-специалистам приходится тратить много усилий, чтобы удовлетворить запросы пользователей и предоставить данные из различных источников в больших объемах, а также быстро реагировать на запросы.

Создание и поддержание Datamarket также может стать дополнительной нагрузкой для уже перегруженного работой отдела ИТ, поскольку им придется постоянно контролировать эти ВД и обеспечивать их безопасность.

Перенос ВД в облако может решить многие проблемы как для рабочих групп, так и для отделов ИТ, поскольку администрированием и обеспечением безопасности в облаке занимается поставщик облачных решений. Это значительно уменьшает число задач, которые нужно выполнять вручную, и снижает операционные расходы.

Преимущества использования витрины данных

Один источник с достоверной информацией

Все члены подразделения компании имеют доступ к единому набору сведений, который хранится централизованно в одном месте. Это исключает необходимость сравнения информации из различных источников. Такая организация обеспечивает надёжность информации, предоставляемых в ВД, а также основанных на них планов.

Благодаря этому сотрудники организации могут сосредоточиться на принятии решений и их последующей реализации, не тратя время на повторную проверку данных из различных источников.

Оперативное получение сведений из ВД

Сотрудники из разных отделов имеют возможность оперативно получать необходимые им данные из витрины и использовать их в сочетании с другой информацией, которая у них есть.

После настройки соответствующих параметров для взаимодействия с источником сведений можно быстро получать информацию из витрины без необходимости обращаться в информационное подразделение компании. Это увеличит производительность работы как бизнес-аналитиков, так и технических специалистов, занимающихся обслуживанием систем компании.

Оперативное принятие рабочих решений

С ускорением доступа сотрудников компании к статистической информации процесс принятия рабочих решений будет ускорен. В то время как хранилище охватывает все аспекты деятельности предприятия, Datamarket специализируется на аналитической работе конкретного отдела.

Таким образом, финансисты, рекрутеры и другие специалисты, занимающиеся решением специфических задач, могут эффективно выполнять свои обязанности.

На основе полученных сведений сотрудники компании проводят анализ различных показателей и принимают управленческие решения. Это позволяет им быстро реагировать на изменения в бизнес-среде и адаптировать стратегии в соответствии с текущей ситуацией.

Легкость и быстрота использования ВД

Настройка хранилища для использования всеми сотрудниками компании требует значительных ресурсов и времени. В отличие от этого, аналитическую витрину можно быстро настроить под нужды отдельного подразделения путем использования сокращенного объема информации.

Это позволяет ускорить процесс адаптации системы к конкретным требованиям отдела и обеспечить быстрый доступ к необходимым сведениям без необходимости внедрения сложных и дорогостоящих изменений в общую систему хранения

информации. Такой подход повышает эффективность работы сотрудников и обеспечивает оперативное принятие решений на основе актуальной информации.

ВД отличается гибкостью из за учета корпоративных потребностей

В работе отделов компании может возникнуть необходимость использовать информацию, которая была собрана и использована в предыдущих проектах. Сотрудники могут вносить изменения и дополнять информационные витрины новыми сведениями в соответствии с требованиями текущей деятельности.

Это обеспечивает гибкость в использовании витрины и позволяет адаптировать её под специфические потребности различных подразделений компании. Такой подход обеспечивает эффективное использование ресурсов и повышает общую производительность бизнеса.

ВД позволяет анализировать переходные процессы

Некоторые аналитические проекты компании требуют выполнения в короткие сроки. Например, менеджеру поручили оценить эффективность интернет-маркетинговой акции, запущенной всего две недели назад. Результаты этого анализа могут быть необходимы на предстоящем совещании с другими сотрудниками подразделения.

Важно иметь возможность быстро настроить Datamarket для выполнения конкретной задачи.

Это позволяет анализировать последнюю информацию и сведения и принимать информированные решения на основе актуальной информации. Такой гибкий подход обеспечивает эффективное использование витрины для анализа переходных процессов и оперативного реагирования на изменения в бизнес-среде.

Зачем нужна витрина данных?

Существует множество преимуществ разработки независимой архитектуры витрины данных для бизнес-пользователей, таких как:

- Уменьшая объем данных, витрина данных помогает сократить время отклика пользователя и обеспечивает быстрый доступ к часто используемым данным.
- Это легко реализовать и требует гораздо меньших затрат по сравнению с внедрением полного хранилища данных.
- Масштабируемы и гибкие, что удобно при смене моделей.
- Данные разделены в витрине данных, что позволяет лучше контролировать права на данные, т. е. тех, кто может просматривать и изменять данные.
- Данные могут храниться и организовываться на различных аппаратных или программных платформах.

Лучшие практики проектирования архитектуры витрины данных

Чтобы обеспечить эффективность и масштабируемость архитектуры витрины данных следуйте инструкциям по проектированию хранилища данных.

1. Определите область действия витрины данных

Прежде чем переходить к этапу реализации модели витрины данных предприятия, важно иметь надежный план, учитывающий все бизнес-потребности и приоритеты всех членов команды и конечных пользователей.

Начните с описания масштаба проекта, выделив все риски и ограничения. Это поможет сформулировать правильные ожидания и оценить расходы. С учетом этого

объема разработайте список основных результатов и распределите обязанности между вашей командой.

2. Обратите внимание на логическую модель витрины данных.

Логическая модель витрины данных — это теоретическая неосязаемая конструкция, которая организует данные с точки зрения логических отношений, известных как сущности и атрибуты. Сущность — это элемент данных, тогда как атрибут помогает определить эксклюзивность сущности.

При планировании архитектуры витрины данных сосредоточьтесь на потребностях пользователей. Сопоставьте исходные данные с предметно-ориентированной информацией в месте назначения. Требования конечного пользователя являются важными элементами, используемыми при разработке схемы витрины данных.

Возможно, вам придется изменить физическую реализацию логической модели данных в зависимости от параметров системы, таких как размер компьютера, количество операторов, дисковое пространство, тип сети и программное обеспечение.

3. Определите соответствующие данные

Обычно элементы данных идентифицируются на основе бизнес-требований. Однако зачастую, возможно, придется выйти за рамки запросов конечных пользователей и ожидать будущих требований.

Хороший совет — начать с **бизнес-факторов**, имеющих отношение к предметной области и имеющих решающее значение для потребителя. Например, если разрабатывается модель витрины данных для своего отдела продаж и маркетинга, ключевыми факторами могут быть клиент, местоположение, продукт, продажи и рекламные акции. Также подумайте, интересны ли вам ежемесячные, ежедневные или еженедельные записи.

Затем сформируйте список важных **полей данных** на основе потребностей, выдвинутых операторами витрин данных. Например, некоторыми полями интереса в витрине маркетинговых данных могут быть названия продуктов, характеристики продвижения, регионы и страны.

Вам также следует разделить данные на числовые показатели (называемые фактами) и описательные записи (называемые измерениями).



Источник: ксенонстек

4. Сузьте источники данных

После того как вы перечислили все измерения и факты, составляющие модель витрины данных, следующим шагом будет определение источников, которые будут «кормить» репозиторий. Эти источники могут включать базы данных, файлы Excel, файлы с разделителями и т. д.

Затем перейдите к сопоставлению измерений с таблицами поиска в вашей операционной среде, тогда как факты можно сопоставить с таблицами транзакций.

Вы также можете обнаружить, что некоторые необходимые данные невозможно сопоставить. Обычно это происходит, когда поля в исходной системе не соответствуют необходимым группам данных в витрине данных.

Например, в телекоммуникационной корпорации телефонные звонки могут быть сгруппированы по коду города, но для витрины данных требуются данные в виде почтового индекса. Сейчас сложно сопоставить эти измерения, поскольку один код города состоит из множества почтовых индексов, а почтовый индекс может включать в себя несколько кодов городов. В этой ситуации перевод данных в общий системный формат может потребовать дорогостоящей обработки.

5. Создайте звездообразную схему

При создании звездообразной схемы важно описать взаимосвязь между таблицами фактов и измерений. Это делается с помощью ключей, которые включают один или несколько столбцов, что делает строку в таблице эксклюзивной. Первичный ключ, включающий несколько столбцов, называется составным или составным ключом.

Чтобы связать факты и измерения, полезно использовать суррогатные ключи вместо первичного ключа фактической исходной таблицы. Это позволяет менеджеру витрины данных контролировать ключи внутри среда витрины данных, даже если ключи меняются в операционной системе.

Суррогатный ключ — это созданная системой серия целых чисел, которую можно включить в таблицу измерений вместе с первичным ключом. Он предлагает больше преимуществ по сравнению с первичным ключом, поскольку последний часто представляет собой длинную строку символов. Принимая во внимание, что первичный ключ включает целые числа, что сокращает время ответа на запрос.

Заключение

Со временем объем данных независимой витрины данных, вероятно, увеличится. Таким образом, при физической реализации логической модели витрины данных важно учитывать фактор масштабируемости. Чтобы удовлетворить требования к масштабируемости, рассмотрите возможность минимизации ограничений таких факторов, как размер оборудования, емкость программного обеспечения и пропускная способность системы.

Проектирование архитектуры витрины данных это сложный процесс, который включает в себя несколько трудоемких этапов, а иногда и значительные затраты. Следуя пяти рекомендациям, можно снизить вероятность ошибок и ускорить процесс проектирования.