**Практическая работа 2:**

**Определение элементов модели хранилища данных**

* **Тема**: Определение элементов модели ХД на основе анализа требований.
* **Задачи**:
	+ Определение измерений;
	+ Определение показателей (метрик);
	+ Определение фактов;
	+ Документирование проделанной работы.

*Анализ требований* является профессиональной задачей проектировщика ХД. При анализе требований проектировщик ХД имеет в качестве входных данных неформальные требования пользователей, модели данных для источников данных, в частности, *модель данных* *БД* *OLTP* системы, и возможно, существующую *многомерную модель* ХД, В результате анализа требований проектировщик ХД должен получить эскиз *многомерной модели* для ХД и определения элементов *многомерной модели*.

**Анализ требований проектировщиком ХД**

* В процессе анализа требований проектировщиком ХД должны быть решены следующие задачи:
	+ Определены кандидаты в показатели, факты, измерения и *иерархии измерений*.
	+ Определен уровень структуризации данных (*гранулированности*).
	+ Построен эскиз *многомерной модели*.
	+ Создан словарь предметной области для элементов модели (репозиторий метаданных).
	+ Построен эскиз *многомерной модели* ХД.

**Ключевые исходные данные**

* Корпоративная ER модель OLTP системы.
* Типовые запросы пользователей.
* Описание семантики предметной области.

**Рекомендации при решении задачи идентификации элементов модели**

* **Ориентация на запросы к ХД**. Исходя из имеющихся запросов, сначала определяются показатели, затем измерения, связанные с показателями, а затем факты.
* **Ориентация на предметную область ХД**. Исходя из элементов предметной области, сначала определяются факты, затем измерения, а потом показатели.
* **Ориентация на источники данных**. Модели данных источников данных используются как основа для выбора элементов *многомерной модели*: сначала определяются измерения, затем показатели, а потом факты.

На первом шаге проектировщик должен определить показатели, измерения с возможными иерархиями и факты. *Определение* кандидатов в элементы модели является не *формализуемой задачей*. При идентификации элементов модели могут быть использованы следующие рекомендации:

**Алгоритм действий проектировщика ХД**

* Определение кандидатов в измерения, их атрибутов и *потенциальных ключей*.
* Определение кандидатов в показатели (метрики).
* Определение кандидатов в факты из бизнес – требований.
* Определения кандидатов в факты исходя из семантического анализа предметной области.
* Документирование проделанной работы.

**Работа с корпоративной ER моделью: определение кандидатов в измерения 1**

* Определение кандидатов в измерения, их атрибутов и *потенциальных ключей*.
* Первый шаг состоит в том, чтобы вывести из модели все связанные сущности и все сущности подтипы.
	+ Это отношения N:M и подтипы.
* В ER модели нашей Компании подтипы – это сущности Product Component, Inventory, Order Line, Order, Retail Store и Corporate Sales Office.

В качестве примера мы будем использовать ER модель, которую рассматривали на прошлом семинарском занятии.

**Работа с корпоративной ER моделью: удаление подтипов**



В качестве примера мы будем использовать ER модель, которую рассматривали на прошлом семинарском занятии.

**Работа с корпоративной ER моделью: определение кандидатов в измерения 2**

* Сворачивание сущностей в конце каждого N:M отношения.
* Для каждой сущности, в том числе и выведенных подтипов, должны быть рассмотрены атрибуты, которые могли бы быть полезны в новом измерении.
* Сформировать из ключей сущностей *потенциальные ключи* измерений.
* Например, свертка Salesperson, Sales *Outlet* и Sales Region в Sales превращается в отношение Sales – Customer.

В качестве примера мы будем использовать ER модель, которую рассматривали на прошлом семинарском занятии.

**Работа с корпоративной ER моделью: кандидаты в измерения для ХД**



В качестве примера мы будем использовать ER модель, которую рассматривали на прошлом семинарском занятии.

**Бизнес -требования пользователей**

1. Какова величина среднего остатка продукции на складе и уровень запасов, при котором подается заказ, за текущий месяц для каждой модели на каждом заводе.
2. Какова величина суммарных затрат и суммарной прибыли по каждой модели, проданной сегодня, и просуммированной по отделу сбыта, типу отдела сбыта, области и по структуре продаж Компании?
3. Какова величина суммарных затрат и суммарной прибыли для каждой модели, проданной сегодня, и просуммированной по заводам и по областям?
4. Какой процент моделей получили скидки, и какие из них были проданы по факту со скидкой (в процентах) по складам (store) для всех продаж на этой неделе? В этом месяце?
5. Для каждой модели, проданной в текущем месяце, какой был процент продаж с розничной торговли, с оптовой торговли по безналичному расчету, с оптовой торговли через продавцов?
6. Какие модели и какого типа продукция не продавалась в течение последнего месяца? В течение последней недели?
7. Какие пять моделей, проданных за последний месяц, принесли наибольшую прибыль? По продажам за квартал? По суммарным затратам?
8. Какие отделы сбыта не имели продаж в течение последнего месяца для каждой модели в каждом из пяти топ-списков?
9. Какие продавцы не имели ни одной записи о продажах за последний месяц для каждой модели в каждом из трех списков 5 моделей?

Обозначим запросы через Q1, Q2, … Q9.

**Определение показателей: Запросы и показатели**

* Кандидаты в показатели можно определить исходя из бизнес - запросов. Они соответствуют элементам данных, которые пользователи используют в запросах для вычисления эффективности и поведения бизнес-процессов и объектов деятельности.
* Хорошими кандидатами в показатели являются числовые атрибуты, при определении которых используется агрегатные функции.

**Пример**

* Если потенциальные пользователи Компании часто используют запросы типа 1, 2 и 3:
	+ Определить Суммарные затраты (Total Cost)
	+ Определить средний остаток продукции на сладе (Average *Quantity* on Hand)
	+ Определить Суммарная прибыль (Total Revenue),
* то кандидатами на показатели могут быть
	+ Средний остаток продукции на сладе (Average *Quantity* on Hand)
	+ Суммарные затраты (Total Cost)
	+ Суммарная прибыль (Total Revenue),

**Определение уровня структуризации показателей**

* Уровень структуризации показателей (Measure *Granularities*) в *многомерной модели* может быть определен, как наименьший уровень детализации, используемый для представления показателя в модели.
* Например, значение атрибута Остаток продукции на складе может сохраняться в ХД за день, за месяц или квартал. Этот атрибут может также рассматриваться по продукции или категории продукции или в единицах упаковки.

**Определение уровня структуризации показателей**

* Показатель обычно связан с несколькими измерениями. Уровень структуризации показателя определяется комбинацией детализации каждого из этих измерений. Различные показатели могут иметь одинаковый уровень структуризации.
* Например, показатели Суммарные затраты (Total Cost) и Суммарная прибыль (Total Revenue) связаны с операцией продажи, и поэтому они имеют одинаковый уровень структуризации. Такие показатели являются претендентами на другой элемент *многомерной модели* - факты.
* Определение правильного уровня структуризации показателей в модели ХД является одной из главных задач проектировщика ХД. Уровень структуризации показателей определяет уровни, на которых пользователь может решать свои аналитические задачи.
* Например, 5 уровней структуризации данных обеспечивает выполнение анализа информации на 5 уровнях ее детализации. Ясно, что при увеличении уровня структурированности данных объем сохраняемой в ХД информации увеличивается, что влияет как на размер ХД, так и на производительность обработки запросов.

**Измерения и показатели**

* Измерения требуются для интерпретации показателей.
* Например, интерпретация такого показателя как остаток на складе требует знания, к какому виду продукции этот остаток относится, где этот остаток размещается (на складе какого завода) и период времени, с которым значение этого показателя связывается.
* Такая постановка вопроса приводит в определению следующих измерений для этого показателя: Продукт (Product), Завод (Manufacturing) и Время (Time).
* Если в качестве показателя взять показатель Суммарная прибыль (Total Revenue), анализируемый в запросе 2, то для него следует определить такие измерения как Продажи (Sales), Продукт (Product) и Время (Time).
* Для показателя запроса 3 Суммарные затраты (Total Cost) в качестве измерений следует взять Продукт (Product), Завод (Manufacturing) и Время (Time).

После определения уровня структуризации данных следует уточнить *список* кандидатов в измерения.

Таким образом, измерения являются координатами многомерного пространства изменения показателей, в которых последние получают свою интерпретацию в рамках *предметной области*. Как видно из выше сказанного, методом определения кандидатов в измерения является *анализ* контекста запросов к ХД.

**Измерения, показатели и запросы**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sales** | **Manufacturing** | **Product** |
| ProductID | Numeric | 5 |  |
| 1 |  | X | X |
| 2 | X |  | X |
| 3 |  | X | X |
| 4 | X |  |  |
| 5 | X |  | X |
| 6 |  |  | X |
| 7 |  |  | X |
| 8 | X |  | X |
| 9 | X |  | X |

На следующем шаге проектировщик должен объединить всех претендентов на измерения и исследовать их с точки зрения всех известных запросов к ХД. Для нашего базового учебного примера проектировщик может идентифицировать четыре измерения: Продажи (Sales), Продукт (Product), Завод (*Manufacturing*) и Время (*Time*).

Хорошим элементом анализа при построении модели в рамках нашего базового учебного примера является построение матрицы соответствия бизнес-требований, измерений и показателей. Бизнес-требования определяются списком главных вопросов. Таким образом, имеем первую итерацию матрициы:

**Запросы и показатели**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| average *quantity* on hand | X |  |  |  |  |  |  |  |  |
| Total Cost |  | X | X |  |  |  | X |  |  |
| Total Revenue |  | X |  |  |  |  | X |  |  |
| *Percentage* of models eligible for discount |  |  |  | X |  |  | X |  |  |
| *Percentage* of models eligible for discount |  |  |  | X |  |  |  |  |  |
| *Percentage* of models sold througth retail *outlet* |  |  |  |  | X |  |  |  |  |
| *Percentage* of models sold through sales office order desk |  |  |  |  | X |  |  |  |  |
| *Percentage* of models sold through sales office salesperson |  |  |  |  | X |  |  |  |  |

Средний *остаток* продукции на сладе (*average* *quantity* on *hand*)

Суммарные *затраты* (*Total* Cost)

Суммарная *прибыль* (*Total* Revenue)

*Процент* моделей со скидкой (*percentage* of models eligible for discount)

*Процент* моделей со скидкой (*percentage* of models eligible for discount) проданных по факту

*Процент* моделей проданных через розничную торговлю (*percentage* of models sold througth retail *outlet*)

*Процент* моделей проданных через оптовую. торговлю (*percentage* of models sold through sales office order desk) по безналичному расчету

*Процент* моделей проданных через оптовую. торговлю (*percentage* of models sold through sales office salesperson) продавцами

**Промежуточный итог**

* Более естественным подходом к идентификации измерений является исследование свойств потенциальных показателей в контексте бизнес-процессов и бизнес-правил. Определение потенциальных измерений в контексте бизнес - требований с этой точки зрения является первым этапом решения задачи выбора измерений.
* Например, показатель *Прибыль с продажи* связан с операцией продажи продукции, которая вовлекает в бизнес-процесс покупателя этой продукции, и, естественным образом приводит к еще одному измерению Покупатель (Customer).
* К решению задачи выбора измерений мы вернемся позже, после обсуждения решения других задач, которые непосредственно повлияют на окончательную картину многомерного пространства.

**Определение кандидатов в факты**

* Для извлечения фактов следует снова обратиться к бизнес -требованиям.
* Рассмотрим вопрос 1. Анализ этого запроса позволяет выделить такой факт как Запас (Inventory). В этом запросе вычисляется средний остаток продукции на складе заданного завода для данной модели в течение заданного периода времени. Факт Запас интерпретируется в трех измерениях: Время, Продукт и Завод.
* Рассмотрим вопрос 2 и 3 нашего учебного примера. Оба вопроса связаны с фактом Продажи (Sales). Этот факт объединяет два параметра - Суммарные затраты и Суммарная прибыль, которые являются независимыми параметрами и интерпретируются в трех измерениях: Время, Продукт и Завод.

Поскольку параметры вместе с измерениями определяют факты, то после предварительного определения параметров и измерений можно перейти к определению кандидатов в факты.

**Классификация фактов**

* Факты могут генерироваться бизнес-операциями или событиями хозяйственной деятельности организации. Так, например, с фактом Продажа связаны следующие операции и события: что продано, кто продал, где и когда было продано, как много было продано, какие скидки были использованы и т.д.
* Факты могут представлять собой состояния объектов предметной области (сферы деятельности организации). Например, состояние запаса - это ответы на вопросы: что храниться, где хранится, как много хранилось в течение данного периода и т.д.
* Факты могут также представлять изменение состояния объектов хозяйственной деятельности организации (предметной области ХД). Изменение запаса представляет собой перемещение определенного количества запаса из одного места хранения в другой или во внешнюю организацию.

Факты являются центральным элементом *многомерной модели*. Проектировщику ХД очень важно правильно выбрать факты, соответствующие задачам анализа данной *предметной области*. Однако выбор фактов на основе предварительно определенных параметром и измерений носит технологических характер.

Факты следует определять исходя из семантики *предметной области* и перспектив деятельности организации (стратегического бизнес-плана организации). При выборе фактов с такой позиции проектировщик ХД может исходить из следующей их классификации:

**Правила отбора фактов**

1. Каждый факт должен существовать с точки зрения хозяйственной деятельности организации.
2. Связанные с хозяйственной деятельностью организации факты генерируются в результате операций хозяйственной деятельности организации, событий хозяйственной деятельности организации, объектами хозяйственной деятельности, чье состояние интересно с точки зрения анализа, объектами хозяйственной деятельности, изменение состояние которых интересно с точки зрения анализа
3. Каждый факт должен иметь уникальную идентификацию в сфере хозяйственной деятельности организации.
4. Уровень структуризации основных измерений каждого факта

Проектировщик ХД должен уметь выбирать связанные с хозяйственной деятельностью организации факты, чтобы затем построить *многомерную модель*. Отбирая факты для построения *многомерной модели* ХД, проектировщику следует придерживаться следующих правил:

**Выбор фактов на основе семантики ПрО**

* Продажи являются примером факта, представляющего операцию хозяйственной деятельности или событие хозяйственной деятельности организации.
* Если руководство хочет оценить продуктивность продаж, то оно использует показатели Суммарные затраты или Суммарная прибыль, которые связаны с событием продажи товара. Факт продажи представляет "что-то, что произошло".
* Событие или операция хозяйственной деятельности может принадлежать как внешней, так и самой организации.

Проектировщик ХД должен уметь выбирать связанные с хозяйственной деятельностью организации факты, чтобы затем построить *многомерную модель*. Отбирая факты для построения *многомерной модели* ХД, проектировщику следует придерживаться следующих правил:

**Выбор фактов на основе семантики ПрО**

* Поиск фактов для ХД в рамках хозяйственной деятельности организации приводит к идентификации дополнительных показателей.
* Например, если факт Продажи определяется через призму операции продажи товара, то Количество (*Quantity* Sold) товара проданного в результате операции следует добавить к показателям, которые нужно сохранять в ХД. Количество проданного товара проходит через ряд связанных бизнес - операций, начиная с оформления счета покупателя до получения подтверждения о получении товара покупателем.

**Выбор фактов на основе семантики ПрО**

* Рассмотрим факт Продажа. Операция хозяйственной деятельности предполагает, что произошло изменение в предметной области. Обычно руководители хотят знать, какой эффект произошел в результате этих изменений, и, следовательно, этот эффект нужно как-то оценить. Поэтому, факт, который генерируется в результате операции хозяйственной деятельности, непосредственно связан с некоторой мерой эффекта этой деятельности, т.е. некоторым числовым показателем.

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.

**Выбор фактов на основе семантики ПрО**

* Если руководителей интересует, когда произошла продажа, то факт продажи интерпретируется как событие хозяйственной деятельности. Обычно, для фактов, генерируемых в результате событий хозяйственной деятельности, не ставят в соответствие какой либо показатель, который измеряет эффект от этого события. Факт Продажа означает, что в некоторый момент времени определенный продукт с определенного склада и определенного завода был продан. Факты, связанные с событиями, называются не измеряемыми фактами (factless facts или measureless facts).

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.

**Выбор фактов на основе семантики ПрО**

* Факт Уровень запаса (Inventory) представляет состояние объекта хозяйственной деятельности - запасом товара на определенном заводе в заданный период времени. Состояние такого объекта как запас можно оценить с помощью показателя Средний остаток продукции. Заметим, что в случае фактов отражающих состояние измерение Время интерпретируются в терминах период времени, а показатель, измеряющий значение состояния объекта, представляет собой численное значение в течение этого периода времени.

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.

**Выбор фактов на основе семантики ПрО**

* Одна из основных проблем для фактов, представляющих изменение состояние, - это выбор адекватного показателя. Например, если для заданного периода времени частота смены состояния достаточно высока, то показатель Средний остаток продукции становится для факта Уровень запаса не слишком подходящим, поскольку перестает отражать динамику изменения запаса.
* При решении этой проблемы можно поступать несколькими способами. Оставляя уровень структуризации измерения времени тем же самым, можно выбрать другой показатель, например, Остаток продукции на сладе. Можно пойти по пути увеличения уровня структурированности представления факта Уровень запаса, сохраняя в ХД значение показателя не за месяц, а за день. Принимая такое решение, проектировщик ХД должен задать себе вопрос - позволяет ли источник данных получать значения показателя на таком уровне структуризации. Можно посмотреть на проблему совсем с другой стороны, сохранять в ХД не состояние запаса, а изменение состояния запаса. Последний показатель будет лучше отражать динамику управления запасом.

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.

**Выбор фактов на основе семантики ПрО**

* Рассмотрим показатель Средний остаток товара для факта Уровень запаса продукции. Пусть уровень структуризации измерения Время увеличится до одного дня, но частота изменения состояния объекта по-прежнему останется высокой. С этой точки зрения проектировщик должен увеличить уровень структуризации измерения Время до часа, чтобы ХД правильно отражало динамику изменения запаса.
* Обратим внимание, что в такой постановке задачи аналитиков интересует анализ поведения объекта хозяйственной деятельности, в данном случае, изменение запаса продукции. Тогда можно рассмотреть факт Уровень запаса с иной точки зрения, т.е. изменить его интерпретацию: отслеживать изменения состояния объекта и сохранять в модели показатель Количество перемещенного запаса (*Quantity* Moved).
* Окончательную точку в решении проблемы выбора показателей фактов, описывающих состояние или изменение состояния объекта хозяйственной деятельности проектировщик ХД может решить, либо обращаясь к потенциальным пользователям (с чем предпочитают работать пользователи - с состояниями объектов или изменениями состояния), либо опираясь на методы моделирования времени в ХД (о чем пойдет речь далее).

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.

**Выводы**

* Проектировщик ХД после анализа требований должен иметь следующие документы:
	+ Перечень кандидатов в измерения и их анализ;
	+ Перечень кандидатов в показатели (метрики) и их анализ;
	+ Перечень кандидатов в факты, их классификацию и анализ.

В чем состоит различие между фактами, генерируемыми операциями хозяйственной деятельности, и событиями хозяйственной деятельности.