

KARPOV.COURSES >>> КОНСПЕКТ



> Конспект > 3 урок > Dimensional modeling

> Dimensional modeling

> Пространственное моделирование

Витрина Данных

> Модели "Звезда", "Снежинка" и "Созвездие"

Модель "Звезда"

Модель "Снежинка"

Модель "Созвездие"

> Таблицы фактов

Виды фактов

Свойства данных в таблице фактов

Виды таблиц фактов

> Таблицы измерений

Виды измерений

> Хранение истории хранения изменений

SCD 0

SCD 1

SCD 2

SCD 3

SCD 4

SCD 6

Быстро меняющиеся измерения

> Дополнительные материалы

> Dimensional modeling

> Пространственное моделирование

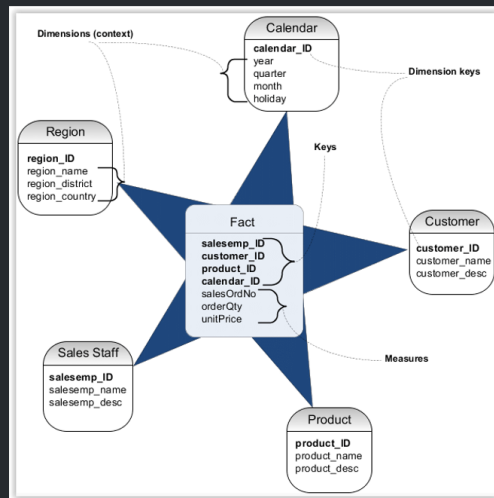
Витрина Данных

Витрина данных (Data Mart) — представляет собой срез КХД в виде массива тематической, узконаправленной информации, ориентированного, например, на пользователей одной рабочей группы или департамента.

Витрина данных, аналогично дашборду, позволяет аналитику увидеть агрегированную информацию в определенном временном или тематическом разрезе, а также сформировать отчетные данные в виде шаблонизированного документа. Витрина данных часто представлена в виде денормализованной таблицы, однако это не всегда удобно и не позволяет решать все задачи, поэтому чаще встречается таблица фактов и таблица измерений.

Таблица фактов — является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах, событиях или процессах, совокупность которых будет в дальнейшем анализироваться.

Таблица измерений (англ. dimension table) — таблица в структуре многомерной базы данных, которая содержит атрибуты событий, сохраненных в таблице фактов. Атрибуты представляют собой текстовые или иные описания, логически объединенные в одно целое.



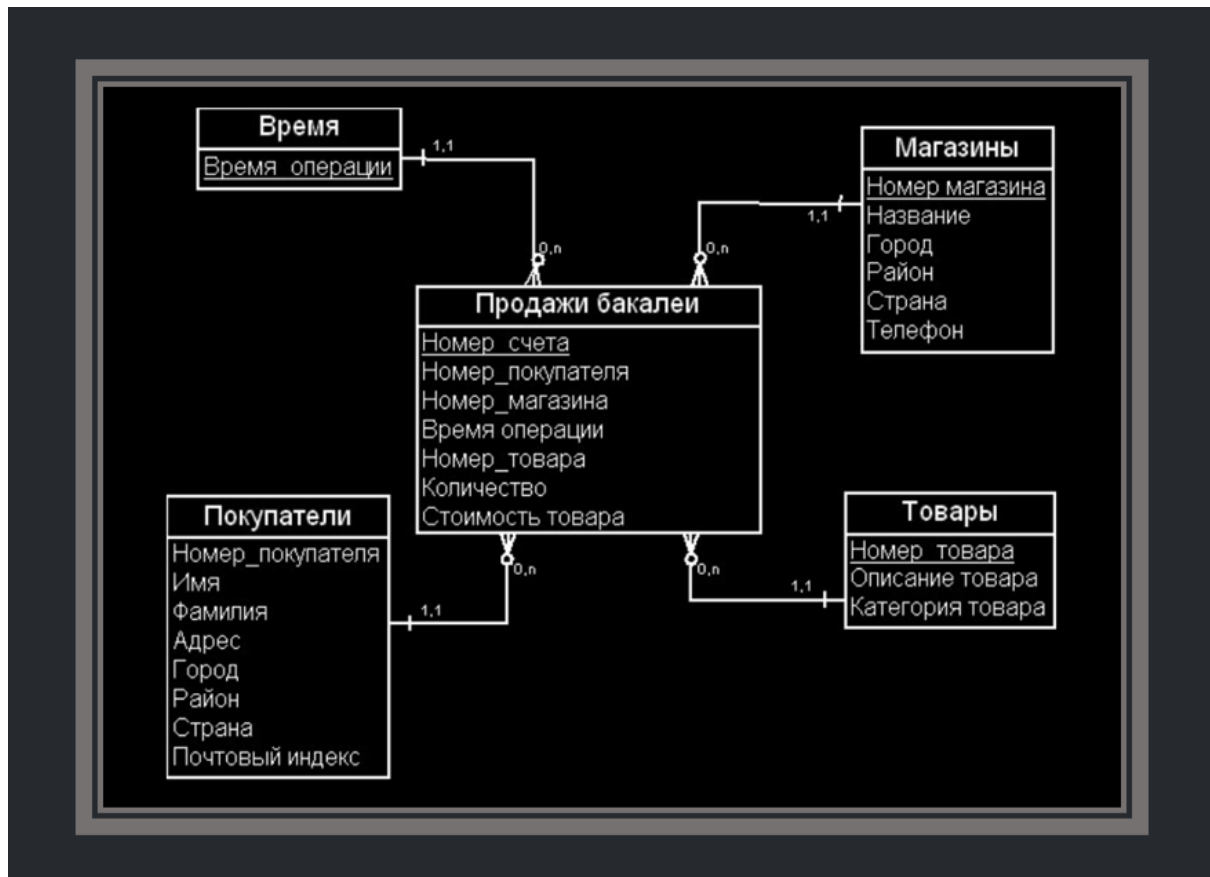
Схемы хранения данных типа «звезда». В центре находится таблица фактов и все атрибуты в ней можно разбить на два блока: идентификаторы таблицы измерений и меры (показатели) этой таблицы фактов. В таблицах измерений представлены соответствующие измерения с ключами.

> Модели "Звезда", "Снежинка" и "Созвездие"

Модель "Звезда"

Схемы «звезда» и «снежинка» — это два способа структурировать хранилище данных.

Схема типа «звезда» (пространственная модель, модель измерений и фактов, модель “сущность-связь”, dimensional model, star schema) - модель представляется двумя видами таблиц: таблицами фактов и таблицами измерений, которые описывают факты. Схема разбивает таблицу фактов на ряд денормализованных таблиц измерений. Таблица фактов содержит агрегированные данные, которые будут использоваться для составления отчетов, а таблица измерений описывает хранимые данные. Денормализованные проекты менее сложны, потому что данные сгруппированы. Таблица фактов использует только одну ссылку для присоединения к каждой таблице измерений. Более простая конструкция звездообразной схемы значительно упрощает написание сложных запросов.



Пример модели данных типа "Звезда". В центре модели таблица фактов, в таблице фактов представлены какие-либо события, например, продажи бакалеи. Таблицы измерений содержат описательные характеристики фактов и расширяют наше представление о событии (факте). На схеме видно, что связь между таблицей измерений и таблицей фактов - один ко многим.

У измерения может не быть фактов, но факт без измерения не может существовать

Модель "Снежинка"

Схема типа «снежинка» отличается тем, что использует нормализованные данные.

Нормализация означает эффективную организацию данных так, чтобы все зависимости

данных были определены, и каждая таблица содержала минимум избыточности.

Таким образом, отдельные таблицы измерений разветвляются на отдельные таблицы измерений. Схема «снежинки» использует меньше дискового пространства и лучше сохраняет целостность данных.

Основным недостатком является сложность запросов, необходимых для доступа к данным — каждый запрос должен пройти несколько соединений таблиц, чтобы получить соответствующие данные.

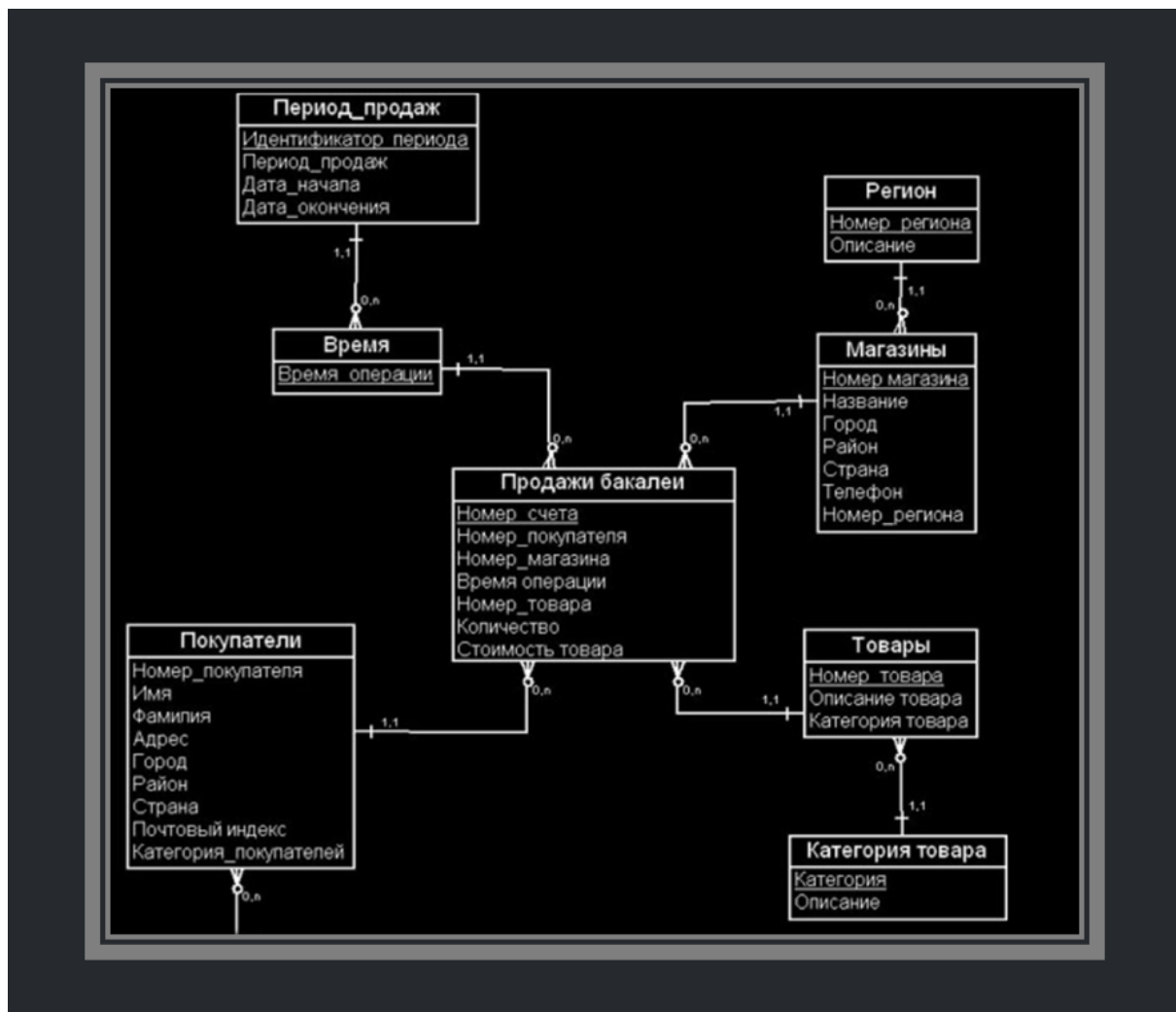
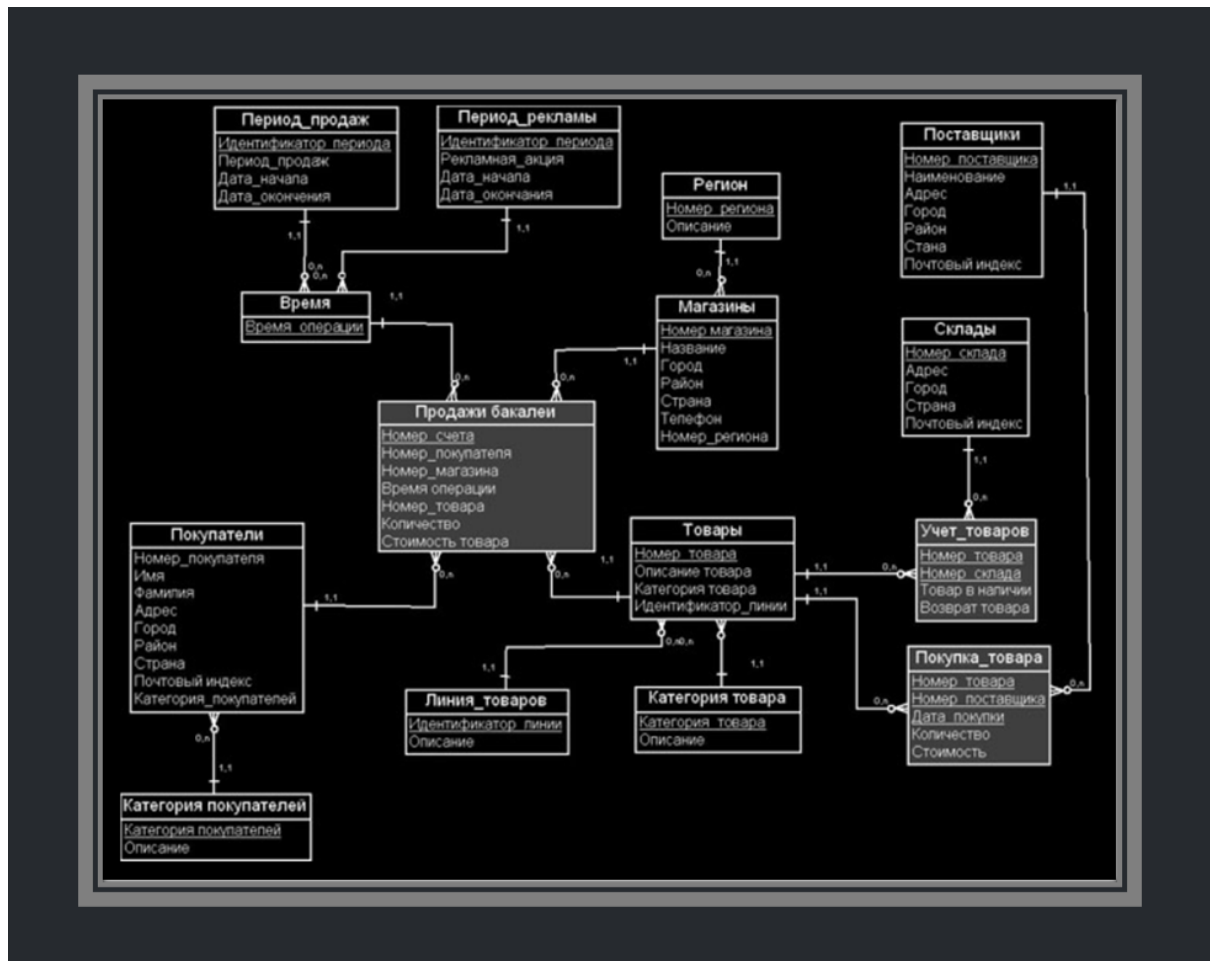


Схема "снежинка" добавляет иерархию в таблицы измерений. Например, измерение "Регион" группирует магазины по географическим регионам, измерение "Категория товара" группирует товары по категориям, измерение "Категория покупателей" группирует покупателей по категориям, а измерение "Период продаж" группирует продажи по периодам времени.

Модель "Созвездие"

Модель «Созвездие» получается из нескольких таблиц фактов, которые соединяются между собой по той или иной логике.



В "Созвездие" у нас может быть несколько таблиц фактов, а одна таблица измерений может быть связана с несколькими таблицами фактов.

> Таблицы фактов

Таблица фактов — является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах, событиях или процессах, совокупность которых будет в дальнейшем анализироваться.

Характеристики таблиц фактов:

- Таблица фактов содержит числовые параметры (метрики),
- Каждая таблица фактов имеет составной ключ, состоящий из первичных ключей таблиц измерений. Первичный ключ таблицы измерений является внешним ключом в таблице фактов.

Виды фактов

Аддитивные факты (Additive facts). Факт называется аддитивным, если его имеет смысл использовать с любыми измерениями для выполнения операций

суммирования с целью получения какого-либо значимого результата. Например, количество продаж, объем продаж и т.д.

Аддитивные факты можно суммировать по всем измерениям.

Количество продаж и суммарная прибыль – аддитивные факты.

По измерению "Время":					
Дата	Товар	Магазин	Количество продаж	Количество покупателей	Суммарная прибыль
23.01.2009	CD диск	Компьютер	10	10	1500
24.01.2009	CD диск	Компьютер	35	30	5250
25.01.2009	CD диск	Компьютер	20	15	3000
			65	55	9750
По измерению "Товар":					
Дата	Товар	Магазин	Количество продаж	Количество покупателей	Суммарная прибыль
23.01.2009	CD диск	Компьютер	10	6	1500
23.01.2009	Принтер	Компьютер	1	1	5000
23.01.2009	Сканер	Компьютер	2	2	3000
			13		9500
По измерению "Магазин":					
Дата	Товар	Магазин	Количество продаж	Количество покупателей	Суммарная прибыль
23.01.2009	CD диск	Компьютер	10	10	1500
23.01.2009	CD диск	Принтеры	10	5	1500
23.01.2009	CD диск	Оргтехника	20	7	3000
			40	22	6000

Количество продаж и суммарная прибыль могут являться аддитивными фактами

Полуаддитивные факты (Semiadditive facts). Факт называется полуаддитивным, если его имеет смысл использовать совместно с некоторыми измерениями для выполнения операций суммирования с целью получения какого-либо значимого результата. Например, числовые показатели интенсивности, такие как остаток на счете, уровень запасов на складе и т.д.;

Суммирование метрики "Количество покупателей" по измерению "Товары":					
Дата	Товар	Магазин	Количество продаж	Количество покупателей	Суммарная прибыль
23.01.2009	Бумага для факсов	Компьютер	10	6	1000
23.01.2009	Бумага для принтера	Компьютер	12	7	1320
				13	

Количество покупателей является полуаддитивным фактом, если мы просуммируем всех покупателей, то мы не получим количество уникальных покупателей, так как покупатель может совершать покупки разных товаров.

Неаддитивные факты (Non-additive facts). Факт называется неаддитивным, если его не имеет смысла использовать совместно с каким-либо измерением для выполнения операций суммирования с целью получения какого-либо значимого результата. Например, измерение комнатной температуры.

Неаддитивные факты не имеет смысла суммировать ни по каким измерениям.

Проценты и отношения величин (конверсия посетителей в покупателей) являются неаддитивными фактами. Можно хранить как отдельные параметры числитель и знаменатель отношения, когда их раздельное суммирование имеет смысл. И это будут уже аддитивные факты. К неаддитивным фактам относятся также статистические средние суммы, такие как, например, средняя температура за день. Сумма средних дневных температур за неделю не имеет никакого смысла.

Числовые меры интенсивности (Numerical Measures of Intensity). Факт называется числовой мерой интенсивности, если он, являясь неаддитивным по времени, допускает агрегацию и суммирование по некоторому числу временных периодов. Например, остаток на счете.

Свойства данных в таблице фактов

Свойства данных в таблицах фактов:

1. Числовые параметры используются для агрегации и суммирования;
2. Значения данных должны обладать свойствами аддитивности или полуаддитивности и по отношению к измерениям, для того чтобы их можно было суммировать;
3. Все данные таблицы фактов должны быть однозначно идентифицированы через ключи таблиц измерений, чтобы обеспечить доступ к ним через таблицы измерений;

Таким образом, таблицу фактов можно разделить на две части. Первая часть состоит из первичных ключей измерений, вторая — из числовых параметров функционально зависящих от ключей таблиц измерений.

Виды таблиц фактов

Транзакционная таблица фактов (Transaction facts).

В такой таблице фактов сохраняют факты, которые фиксируют определенные события (транзакции). Это факты, описывающие каждое событие бизнеса. Например, продажи товара.





Таблица фактов периодических моментальных снимков (Snapshot).

В такой таблице собирают факты, фиксирующие текущее состояние определенного направления бизнеса. Это факты, которые описывают текущее состояние определенного направления бизнеса для любой комбинации значений измерений за данный период времени. Например, продажи организации на определенную дату (ежедневно).

Таблица фактов кумулятивных моментальных снимков (Accumulated Snapshot).

В такой таблице собирают факты, фиксирующие некоторое итоговое состояние определенного направления бизнеса на текущий момент времени. Это факты, которые описывают промежуточные итоги деятельности организации по определенному направлению бизнеса для любой комбинации значений измерений за данный период времени. Например, продажи этого года на определенную дату.

Таблицы фактов

 Property	 Транзакционная таблица фактов	 Таблица фактов периодических моментальных снимков	 Таблица фактов кумулятивных моментальных снимков
<u>Определение гранулярности таблицы фактов</u>	Одна строка на бизнес-операцию	Одна строка на период	Одна строка для периода завершенного события
<u>Факты</u>	Факты связаны с операционной деятельностью	Факты связаны с периодической деятельностью	Факты связаны с деятельностью, которая имеет определенное время существования
<u>Обновления</u>	Не допускаются	Не допускаются	Допускаются
<u>Кардинальность таблицы фактов</u>	Растет быстро	Растет медленнее, чем в транзакционных таблицах	Растет быстрее, чем в таблицах фактов периодических снимков

> Таблицы измерений

Таблица измерений (англ. dimension table) — таблица в структуре многомерной базы данных, которая содержит атрибуты событий, сохраненных в таблице фактов. Атрибуты представляют собой текстовые или иные описания, логически объединенные в одно целое.

Таблица измерения имеет первичный ключ и атрибуты, описывающие факты с точки зрения некоторого направления деятельности организации.

Характеристики измерений:

1. Таблицы измерений содержат данные о детализации фактов;
2. Таблицы измерений содержат описательную информацию о числовых значениях в таблице фактов, т.е. они содержат атрибуты фактов;
3. Как правило, таблицы измерений содержат большое количество полей;
4. Таблицы измерений содержат обычно значительно меньше строк, чем таблицы фактов;
5. Атрибуты таблиц измерений обычно используются при визуализации данных в отчетах и запросах;

Виды измерений

Медленно меняющимися измерениями (Slowly Changing Dimensions) называются таблицы измерений, в которых некоторые атрибуты могут изменить свои значения по истечении некоторого периода времени, причем частота таких изменений является небольшой.

Всего существует 8 основных типов SCD, которые определяют, как история изменений может быть отражена в модели.

> **Хранение истории хранения изменений**

Медленно меняющимися измерениями (Slowly Changing Dimensions) называются таблицы измерений, в которых некоторые атрибуты могут изменить свои значения по истечении некоторого периода времени, причем частота таких изменений является небольшой.

Всего существует 8 основных типов SCD, которые определяют, как история изменений может быть отражена в модели.

SCD Type	Dimension Table Action	Impact on Fact Analysis
Type 0	No change to attribute value	Facts associated with attribute's original value
Type 1	Overwrite attribute value	Facts associated with attribute's current value
Type 2	Add new dimension row for profile with new attribute value	Facts associated with attribute value in effect when fact occurred
Type 3	Add new column to preserve attribute's current and prior values	Facts associated with both current and prior attribute alternative values
Type 4	Add mini-dimension table containing rapidly changing attributes	Facts associated with rapidly changing attributes in effect when fact occurred
Type 5	Add type 4 mini-dimension, along with overwritten type 1 mini-dimension key in base dimension	Facts associated with rapidly changing attributes in effect when fact occurred, plus current rapidly changing attribute values
Type 6	Add type 1 overwritten attributes to type 2 dimension row, and overwrite all prior dimension rows	Facts associated with attribute value in effect when fact occurred, plus current values
Type 7	Add type 2 dimension row with new attribute value, plus view limited to current rows and/or attribute values	Facts associated with attribute value in effect when fact occurred, plus current values

Основные типы SCD

SCD 0

SCD 0 — заключается в том, что данные после первого попадания в таблицу далее никогда не изменяются. Этот метод практически никем не используется, т.к. он не поддерживает версионности. Он нужен лишь как нулевая точка отсчета для методологии SCD. По сути, вообще не SCD.

Таблица, которая хранит пол родственников Дональда Дака - женский, мужской, не определено. Она также не требует ведения истории.

SCD 1

SCD 1 — это обычная перезапись старых данных новыми. В чистом виде этот метод тоже не содержит версионности и используется лишь там, где история фактически не нужна

Достоинства: Не добавляется избыточность, Очень простая структура

Недостатки: Не хранит истории

Пример: паспортные данные изменились и были перезаписаны

SCD 2

SCD 2 - для каждой версии создается отдельная запись в таблице с добавлением поля-ключевого атрибута данной версии.

Достоинства: Хранит полную и неограниченную историю версий Удобный и простой доступ к данным необходимого периода

Недостатки: Провоцирует на избыточность или заведение дополнительных таблиц для хранения изменяемых атрибутов измерения

Пример SCD2

# ID	Aa Name	# Number	Team	Date_start	Date_end
1	<u>Marc Marquez</u>	93	Honda	@November 8, 2013	@January 1, 9999
2	<u>Valentino Rossi</u>	46	Yamaha	@November 7, 2010	@January 1, 9999
3	<u>Dani Pedrosa</u>	26	Honda	@November 8, 2014	@January 11, 2018
4	<u>Jorge Lorenzo</u>	99	Ducati	@January 1, 2017	@January 1, 2019
5	<u>Jorge Lorenzo</u>	99	Honda	@February 1, 2019	@January 1, 9999

SCD 3

SCD 3 — В самой записи содержатся дополнительные поля для предыдущих значений атрибута. При получении новых данных, старые данные перезаписываются текущими значениями.

Достоинства: Небольшой объем данных, Простой и быстрый доступ к истории

Недостатки: Ограниченная история

Недостатки: Провоцирует на избыточность или заведение дополнительных таблиц для хранения изменяемых атрибутов измерения

Пример SCD3

# ID	Aa Name	# Num	Previous_team	Current_team	Date_start
1	<u>Marc Marquez</u>	93	NULL	Honda	@November 8, 2013
2	<u>Valentino Rossi</u>	46	NULL	Yamaha	@November 7, 2010
3	<u>Dani Pedrosa</u>	26	NULL	Honda	@November 8, 2014
4	<u>Jorge Lorenzo</u>	99	Ducati	Honda	@February 1, 2019

SCD 4

История изменений содержится в отдельной таблице: основная таблица всегда перезаписывается текущими данными с перенесением старых данных в другую таблицу. Обычно этот тип используют для аудита изменений или создания архивных таблиц

Достоинства: Быстрая работа с текущими версиями

Недостатки: Разделение единой сущности на разные таблицы

SCD 6

Комбинация вышеназванных методов и предназначен для ситуаций, которые они не учитывают или для большего удобства работы с данными.

Он заключается во внесении дополнительной избыточности:

Берется за основу тип SCD 2, добавляется суррогатный атрибут для альтернативного обзора версий (тип SCD 3), и перезаписываются одна или все предыдущие версии (тип SCD 1)

Пример SCD6

# Version	# ID	Aa Name	# Number	≡ Team	📅 Date_start	📅 Date_end	# Current
1	1	<u>Marc Marquez</u>	93	Honda	@November 8, 2013	@January 1, 9999	1
1	2	<u>Valentino Rossi</u>	46	Yamaha	@November 7, 2010	@January 1, 9999	1
1	3	<u>Dani Pedrosa</u>	26	Honda	@November 8, 2014	@January 11, 2018	1
1	4	<u>Jorge Lorenzo</u>	99	Ducati	@January 1, 2017	@January 1, 2019	0
2	5	<u>Jorge Lorenzo</u>	99	Honda	@February 1, 2019	@January 1, 9999	1

Быстро меняющиеся измерения

Основным приемом моделирования быстро меняющихся измерений является логическое

разбиение таблицы измерения на две или более таблицы. При этом для быстро меняющихся атрибутов часто используют дискретный диапазон изменений, чтобы сократить объем данных в таблице.

Суть приема логического разбиения состоит в следующем: создаются две сущности, одна

из которых содержит атрибуты, которые меняются медленно, а другая сущность включает в себя атрибуты, которые меняются быстро.



> Дополнительные материалы

1. Dama Dmbok - <https://www.dama.org/cpages/body-of-knowledge>