

ФОРМАЛЬНОЕ ОПИСАНИЕ МЕХАНИЗМА ТЕМПОРАЛЬНОГО ХРАНЕНИЯ ДАННЫХ

Д.В. Спандерашвили

В статье в краткой форме дано формальное описание механизма темпорального хранения данных. Сформулированы предпосылки применения темпорально-многомерного хранилища данных, вводится формальное описание темпорально-многомерного хранилища данных, определяется механизм формирования соответствия между структурными версиями, описан механизм применения матриц трансформации.

Предпосылки применения темпорально-многомерного хранилища данных

Время является одним из измерений в хранилище данных, которое позволяет сравнивать данные в различные промежутки времени. Однако измерения изменяются во времени – объединяются и распадаются территориальные образования, появляются и исчезают продукты и услуги, развиваются организационные структуры. В силу того, что в простой многомерной модели данных измерения аксиоматично считаются прямоугольными, подобные изменения нельзя представить адекватно, что накладывает строгие ограничения на достоверность запросов OLAP, охватывающих несколько промежутков времени.

Несмотря на то, что в хранилищах данных время обычно явно представлено в качестве измерения, с модификациями в измерениях, например, если образуется новое подразделение или отдел, такие системы не справляются.

Причина такого непродуктивного свойства технологии многомерных хранилищ данных заключается в имплицитной аксиоме, что измерения ортогональны. Ортогональность по отношению ко времени означает, что другие измерения должны быть инвариантами по времени. Такое предположение «по умолчанию» не позволяет в необходимой мере обрабатывать изменения многомерных данных.

Очевидно, что для обеспечения верности результатов запросов OLAP необходимо должным образом учитывать изменения в многомерных данных¹. Бизнес-структуры, территориальные структуры услуг (для телекоммуникационных компаний в связи с динамикой внедрения новых услуг) подвержены высоко динамичным изменениям. Для сравнения данных за несколько периодов, вычисления тенденций, вычисления контрольных величин по данным предыдущих периодов необходимо правильно обработать изменения в многомерных данных. Иначе возможны столкновения с бессмысленными цифрами и неправильными выводами, влекущими за собой неверные решения.

Необходимы следующие расширения многомерной модели данных.

Временное расширение – многомерные данные должны быть маркированными во времени с тем, чтобы они представляли достоверное время.

Структурные версии – если существуют временные штампы, то появляется необходимость в том, чтобы наша система работала с разными версиями структур. Мы называем версию структуры структурной версией.

Функции трансформации: нашей системе необходимо поддерживать функцию трансформирования данных из одной структурной версии в другую.

Формальное описание темпорально-многомерной модели данных

Многомерное представление данных состоит из набора измерений. Эти измерения определяют n -мерный куб данных.

Ячейка в таком n -мерном кубе содержит значение и определяется вектором v :

$$V = (D_1i, D_2i, D_3i), \quad (1)$$

где D_1i, D_2i, D_3i – элементы измерений D_1, D_2, D_3 .

Для поддержки темпоральности в многомерной структуре данных необходимо ввести следующие понятия².

Хрон. Обычно одно из измерений в модели определяет время. Мельчайший член измерения «время» является хроном Q . Временная ось определяется через последовательность

хронов. Хрон можно определить как «неделимый временной промежуток, фиксированной, минимальной продолжительности».

Временной интервал. Измерение является последовательностью членов (вершин графа) и иерархических связей между ними (ребра графа), таким образом, полученный граф является деревом, представляющим иерархическую структуру измерения. Чтобы надлежащим образом передать время актуальности такой системы, каждая вершина и ребро графа должны иметь временной интервал $[T_S, T_E]$, определяющий время валидности ребра или вершины T_S является началом временного интервала, а T_E его окончанием.

Темпорально многомерная система состоит из следующих элементов.

1. Количество измерений – N . Если рассматривать факты как измерение, то количество измерений увеличивается на 1 и будет $N+1$.

2. Набор измерений:

$$D = \{D_1, \dots, D_N, F\}, \quad (2)$$

где F – измерение, представляющее набор анализируемых фактов; D_i – все оставшиеся измерения, включая измерение «время».

3. Измерение D_i , определяется как

$$D_i = \langle D_{id}, Key, UDAs, [T_S, T_E] \rangle, \quad (3)$$

где D_{id} , – уникальный идентификатор каждого измерения, который не может быть изменен; Key – уникальный в пределах временного интервала $[T_S, T_E]$ идентификатор измерения, заданный пользователем; $UDAs$ (User Defined Attributes) – набор атрибутов, определенных пользователем; $[T_S, T_E]$ – время валидности данного члена измерения.

4. Количество членов измерений – M .

5. Набор членов измерений:

$$DM = DM_{D_1} \cup \dots \cup DM_{D_N} \cup DM_F = \{DM_{D_1}, \dots, DM_{D_N}, DM_F\}, \quad (4)$$

где DM_F – набор всех фактов; DM_{D_i} – набор всех элементов, принадлежащих измерению D_i .

6. Член измерения $D_i - DM_i$ определяется как:

$$DM_i = \langle DM_{id}, Key, D_i, UDAs, [T_S, T_E] \rangle, \quad (5)$$

где DM_{id} , – уникальный идентификатор каждого члена измерения, который не может быть изменен; Key – уникальный в пределах временного интервала $[T_S, T_E]$ идентификатор члена измерения, заданный пользователем; D_i – идентификатор измерения, к которому принадлежит данный член измерения; $UDAs$ (User Defined Attributes) – набор атрибутов, определенных пользователем; $[T_S, T_E]$ – время валидности данного члена измерения.

7. Набор иерархических связей:

$$H = \{H_0 \dots H_i \dots\}, \quad (6)$$

где H_i – иерархическая связь.

8. Иерархическая связь H_i определяется как

$$H_i = \langle DM_{id}^C, DM_{id}^P, Level, [T_S, T_E] \rangle, \quad (7)$$

где DM_{id}^C – идентификатор члена измерения, который является дочерним в связи; DM_{id}^P – идентификатор элемента измерения, который является родителем для элемента с идентификатором DM_{id}^C или пустым множеством, если у элемента с идентификатором DM_{id}^C нет родителя; $Level \in \{0..L\}$, где L – число иерархических уровней в измерении и $Level$ – номер уровня, на котором располагается элемент с идентификатором DM_{id}^C , все «листья» (элементы, не имеющие потомков) в дереве иерархии расположены на уровне 0; $[T_S, T_E]$ – время валидности данной иерархической связи между элементами измерения.

9. Структурная версия SV_i – это представление многомерной структуры, которая действительна в заданном временном интервале $[T_S, T_E]$. Формально структурная версия задается следующим образом:

$$SV_i = \langle SV_{id}, [T_S, T_E], \{DM_{D_1, SV_{id}}, \dots, DM_{D_n, SV_{id}}, DM_{F, SV_{id}}\}, H_{SV_{id}} \rangle, \quad (8)$$

где SV_{id} – уникальный идентификатор структурной версии; $[T_S, T_E]$ – временной интервал в течение которого данная структурная версия действительна; $DM_{D_1, SV_{id}}$ – набор элементов измерения D_1 , которые действительны в каждой точке временного интервала описываемой структурной версии; $DM_{F, SV_{id}}$ – набор фактов, которые действительны в каждой точке временного интервала описываемой структурной версии; $H_{SV_{id}}$ – набор иерархических связей, которые действительны в каждой точке временного интервала описываемой структурной версии. На концептуальном уровне (не на уровне реализации) можно сказать, что каждой структурной версии соответствует свой куб данных.

10. Темпорально хранилище данных DWH есть непустое конечное множество структурных версий:

$$DWH = \{SV_1, \dots, SV_n\}. \quad (9)$$

Определение соответствий между структурными версиями

Для полноты темпоральной модели следует определить функции соответствий между структурными версиями.

Структурная версия – это представление темпорального хранилища данных, действительного во временном интервале $[T_S, T_E]$. Следует различать два понятия: структура структурной версии (набор всех действительных членов измерений в совокупности с иерархическими связями) и данные структурной версии (куб, определяемый соответствием структурной версии со значением домена).

При составлении запросов к хранилищу данных пользователь выбирает обычно начальную структурную версию, которую будет использовать. Данные же, возвращенные запросом, могут порождать обращение к нескольким (различным) структурным версиям куба. Поэтому необходимо задание функций отображения между структурными версиями.

Функция отображения определяется как

$$MapF(SV_i, SV_j, DM_{id}, DM'_{id}, \{M_{id}^1, \dots, M_{id}^n\}, w), \quad (10)$$

где SV_i и SV_j определяют структурные версии; DM_{id} и DM'_{id} определяют элементы измерения, принадлежащие, соответственно, SV_i и SV_j ; $\{M_{id}^1, \dots, M_{id}^n\}$ – непустой, конечный набор идентификаторов фактов, w – коэффициент трансформации для проецирования данных из одной структурной версии в другую.

Функции отображения могут определять соответствия как между смежными ($SV_1 - SV_2$), так и между несмежными структурными версиями ($SV_1 - SV_3$).

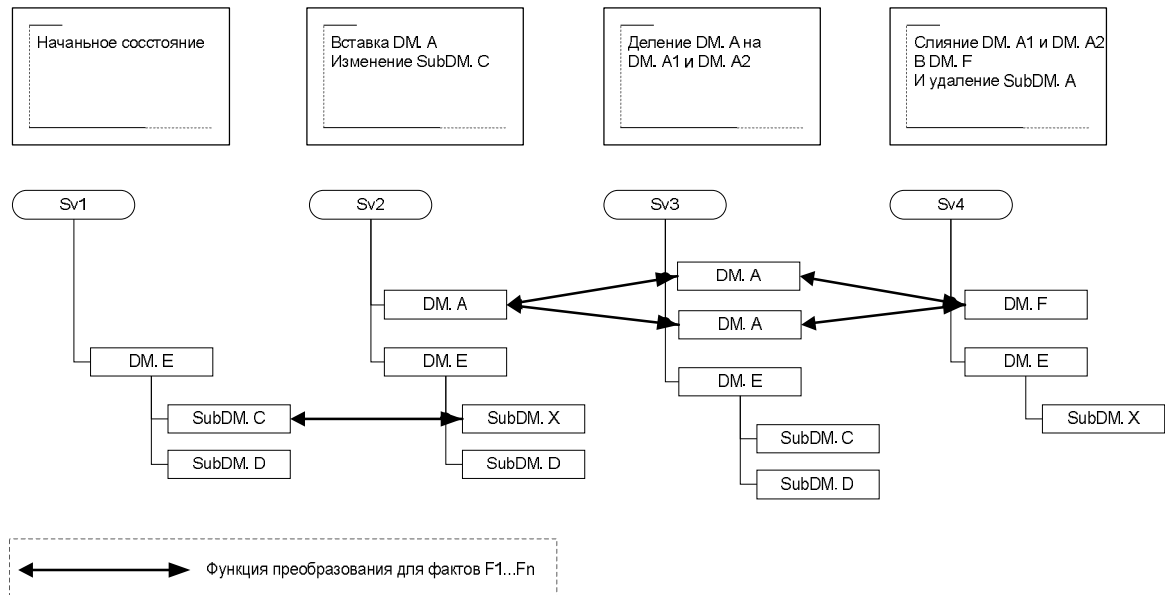


Рис. Структурные преобразования в измерении

Для отображения изменений, представленных на рисунке, необходимо определение функции отображения (приведенные ниже):

$$MapF(SV_1, SV_2, DM_C, DM_X, \{Fact\}, 1)$$

$$MapF(SV_2, SV_1, DM_X, DM_C, \{Fact\}, 1)$$

$$MapF(SV_2, SV_3, DM_A, DM_{A1}, \{Fact\}, 0.3)$$

$$MapF(SV_2, SV_3, DM_A, DM_{A2}, \{Fact\}, 0.7)$$

$$MapF(SV_3, SV_2, DM_{A1}, DM_X, \{Fact\}, 1)$$

$$MapF(SV_3, SV_2, DM_{A2}, DM_X, \{Fact\}, 1) \text{ и т.д.}$$

На функциях видно, как значения факта «Fact» делятся при переходе от структурной версии SV_2 к структурной версии SV_3 в соотношении 30 % и 70 % соответственно. При движении в обратном направлении это означает, что при переходе от SV_3 к SV_2 значение фактов, соотношенных с элементами измерений A_1 и A_2 в SV_3 , в сумме составляет значение фактов, соотношенных с измерением A в SV_2 .

Матрицы трансформации

На концептуальном уровне можно представить каждый многомерный куб и связи между элементами измерений различных структурных версий посредством матриц.

Пусть SV_i – структурная версия с N измерениями. Каждое измерение D_n состоит из набора $D_N^{L_0}$, который представляет все элементы нулевого уровня этого измерения. Эту структурную версию можно представить как $DM_1^{L_0} \times DM_2^{L_0} \times \dots \times DM_N^{L_0}$ матрицу. Так, например, структурная версия одномерного куба будет представлять вектор, двухмерного – матрицу, трехмерного куба – матрицу, элементы которой являются векторами, и т.д.

Пусть SV_1 и SV_2 – две структурные версии, для которых определим матрицу трансформации $T(SV_1, SV_2, D_i, F)$ для каждого измерения D_i и каждого факта F. Тогда $T(d_i, d_j)$ – число, представляющее коэффициент трансформации w для отображения факта F элемента измерения d_i структурной версии SV_1 в факт элемента измерения d_j структурной версии SV_2 .

Так, для представления следующих функций трансформации:

$$1. MapF(Sv_1, Sv_2, a_1, a_{11}, Fact, 0.3)$$

2. MapF(Sv₁, Sv₂, a₁, a₁₂, Fact, 0.7)
3. MapF(Sv₁, Sv₂, a₂, a₂, Fact, 1)
4. MapF(Sv₁, Sv₂, a₃, a₃, Fact, 1)
5. MapF(Sv₁, Sv₂, b₁₂, b₁, Fact, 1)
6. MapF(Sv₁, Sv₂, b₁₂, b₂, Fact, 1)
7. MapF(Sv₁, Sv₂, b₃, b₃, Fact, 1)

можно представить двумя матрицами трансформации (11) (12):

$$T(SV_1, SV_2, A, A') = \begin{matrix} & \times & a_{11} & a_{12} & a_2 & a_3 \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} & & \begin{pmatrix} 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix} \quad (11)$$

$$T(SV_1, SV_2, B, B') = \begin{matrix} & x & b_1 & b_2 & b_3 \\ \begin{matrix} b_{12} \\ b_3 \end{matrix} & & \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{matrix} \quad (12)$$

Итак, матрицу трансформации можно определить следующим образом:

$$T(SV_i, SV_j, D_{id}, D'_{id}), \quad (13)$$

где SV_i – структурная версия, в которой находятся трансформируемые данные; SV_j – структурная версия, структуру которой необходимо преобразовать; D_{id} – набор элементов измерения исходной структурной версии; D'_{id} – набор элементов целевой структурной версии.

Элемент матрицы (14) трансформации определяет коэффициент трансформации факта элемента исходного измерения в факт элемента целевого измерения.

$$t(SV_i, SV_j, DM_{D_{id}}, DM_{D'_{id}}) = w, \quad (14)$$

где SV_i – структурная версия, в которой находятся трансформируемые данные; SV_j – структурная версия, структуру которой необходимо преобразовать; $DM_{D_{id}}$ – элемент измерения D_i ; $DM_{D'_{id}}$ – элемент измерения D'_i ; w – коэффициент трансформации преобразования элементов исходного измерения в целевое.

¹ **Спандерашвили Д.В.** Механизмы отслеживания изменений в многомерных структурах данных // Инфокоммуникационные технологии в науке, производстве и образовании: Вторая международная научно-техническая конференция. Ставрополь: СКГТУ, 2006. Ч. 1. С. 160–162.

² **Спандерашвили Д.В.** Формальное описание модели и алгоритмы трансформации темпорального хранилища данных // Южно-Российский вестник геологии, географии и глобальной энергии. Астрахань: Издательский дом «Астраханский университет», 2006. № 7 (20). С. 136–142.