

Метаданные в хранилищах данных

Распределённые базы данных и хранилища данных

Понятие метаданных

- Метаданные позволяют пользователям понять, на какие вопросы может отвечать данная информационная система.
- С технической точки зрения метаданные – это совокупность спецификаций и данных, которая в целом дает ответы на вопросы, какова степень охвата предметной области в информационной системе, какие данные в ней представлены, какова архитектура системы и т.д.
- Метаданные содержат семантическую интерпретацию или толкование содержания элементов данных, циркулирующих в информационной системе.



Понятие метаданных

- Под метаданными понимается совокупность элементов данных и спецификаций, содержащих описание данных информационной системы и процессов их обработки.
- Метаданные (от лат. meta – цель, конечный пункт, предел, край и данные) – информация о другой информации, или данные, относящиеся к дополнительной информации о содержимом или объекте.
- Метаданные раскрывают сведения о признаках и свойствах, характеризующих какие-либо сущности, которые позволяют автоматически искать и управлять ими в больших информационных потоках.



Проектирование метаданных

Задача проектировщика хранилища данных:

- идентификации объектов хранилища данных и их атрибутов;
- идентификации источников данных;
- описании семантики данных источников и хранилища данных;
- описании алгоритмов преобразования и агрегации данных;
- описании путей доступа к данным и т.п.



Функции метаданных

Основные функции метаданных в хранилище данных:

- Представление соответствия данных источников и данных хранилища данных.
- Управление данными во времени.
- Поддержка версионности.
- Интерпретация данных в терминах бизнес-пользователей.
- Обеспечение открытости.



Функции метаданных

- **Представление соответствия данных источников и данных хранилища данных.** Как правило, это описание представляет собой фиксацию взаимосвязи атрибутов данных источника и атрибутов данных хранилища данных, правила преобразования первых во вторые, изменение в наименовании данных, в их физических характеристиках и т.д. Такая информация позволяет идентифицировать источники данных для хранилища данных, правильность данных в хранилище и их корректность.



Функции метаданных

- **Управление данными во времени.** Временной горизонт данных в хранилище данных гораздо больше, чем в системах операционной обработки данных. В хранилище данных в одно и то же время может существовать несколько схем данных, отвечающих различным периодам эволюции источников данных. Запись о таких структурных изменениях сохраняется в метаданных хранилища данных.



Функции метаданных

- **Поддержка версионности.** Метаданные должны отражать изменения внутренней структуры данных источников и, следовательно, должны сами изменяться, для того чтобы обеспечить непрерывность истории изменения структуры данных хранилища. Поддержка версионности метаданных позволяет в каждый момент времени в прошлом обеспечить правильное описание модели данных, а аналитики получают возможность знать, какие данные, когда и как попали в хранилище данных.



Функции метаданных

- **Интерпретация данных в терминах бизнес-пользователей.** Метаданные должны поддерживать в запросах понятную для пользователя терминологию, независимо от того, какие правила наименования атрибутов были использованы проектировщиком хранилища данных.




Функции метаданных

- Обеспечение открытости (доступности другим информационным системам) системы складирования данных для ее интеграции с другими аналитическими системами организации. Опрос метаданных хранилища данных другой системой позволяет последней выяснить структуру данных хранилища данных и поддерживать обмен данными между системами.



Состав метаданных

Базовые элементы метаданных хранилища данных:

- Описание таблиц, их атрибутов, ключей и т.д.
 - Описание преобразований данных.
 - История экстрагирования (поступления) данных.
 - Терминологические соответствия понятий.
 - Информация о состояниях и статистике использования данных хранилища данных.
 - Алгоритмы агрегации и суммирования данных, критерии выборки из источников, правила преобразования данных источников перед загрузкой в хранилище данных, описание взаимосвязей между объектами хранилища данных, их кардинальность и т.п.
 - Кто отвечает за содержание и актуальность различных источников данных.
-
- 

Состав метаданных

Описание преобразований данных включает в себя:

- идентификацию полей источников данных;
 - соответствие между атрибутами сущностей источников данных и атрибутами объектов хранилища данных;
 - преобразования атрибутов;
 - физические характеристики преобразований;
 - преобразования таблиц кодировки и ссылочных таблиц;
 - изменения наименований (соответствие имен источников и объектов хранилища данных);
 - изменение ключевых атрибутов;
 - значение полей по умолчанию;
 - логика (алгоритмы) формирования данных хранилища данных из нескольких источников (приоритетность источников);
 - алгоритмы трансформации данных и т.д.
-



Классификация метаданных

- **Разделяемые метаданные (Shared metadata).** Элементы разделяемых метаданных необходимы для точного определения объектов и их семантики для хранилища данных в целом.
- **Уникальные метаданные (Unique metadata).** Уникальные метаданные описывают уникальные объекты системы.



Классификация метаданных

По функциональному назначению:

- Предметно-ориентированные метаданные (Business metadata). Содержат определения сущностей предметной области в терминах пользователей, логические отображения между данными на различных уровнях их представления.
- Структурные метаданные (Structural metadata). Содержат описание структуры различных объектов данных.
- Технические метаданные (Technical metadata). Содержат определения и данные о физических объектах хранилища данных.
- Метаданные процесса обработки данных (Process metadata). Содержат информацию, связанную с процессом обработки данных.

