

Проектирование Data Warehouse (DWH) - основы

Исторические данные

Slowly Changing Dimensions (SCD) — это термин, используемый в области управления данными и хранилищ данных для обозначения различных способов управления изменениями в измерениях (часто представляющих иерархии или категории), которые изменяются со временем. Это особенно актуально для слоя Detail Data Store (DDS) в DWH, где хранятся исторические данные и где важно отслеживать изменения в долгосрочной перспективе.

Тип 0: Неизменность данных (SCD 0)

- Данные остаются неизменными после их первоначальной загрузки.
- Например, пол пользователя обычно не меняется и может быть записан как SCD тип 0.

Физически, данные просто загружаются в таблицу измерений, и никакие дальнейшие механизмы учета изменений не реализуются.

Тип 1: Перезапись (SCD 1)

- При изменении данных старое значение заменяется новым без сохранения истории.
- Если адрес электронной почты клиента изменится, старый адрес заменяется новым, и история изменений не сохраняется.

Для реализации SCD тип 1, система просто обновляет существующие записи в таблице измерений с новыми данными, обычно с помощью SQL-команды `UPDATE` .

Тип 2: Добавление новой записи (SCD 2)

- Для каждого изменения создается новая запись, таким образом сохраняется история изменений.
- Если клиент переезжает, создается новая запись с новым адресом, в то время как старая запись с предыдущим адресом сохраняется для истории.

В таблице измерений добавляются дополнительные поля для учета версионности, такие как `start_date` , `end_date` , `is_active` или `version` .

При изменении записи, старая запись помечается как неактивная (устанавливается `end_date` и `is_active = false`), и вставляется новая запись с текущими данными и пометкой активной версии.

Тип 3: Добавление нового столбца (SCD 3)

- История изменений хранится в дополнительных столбцах, обычно ограничивается одним или несколькими предыдущими значениями.
- Компания может сохранять текущий и предыдущий статус клиента (например, активный/неактивный) в двух разных столбцах в одной записи.

В таблице измерений добавляется дополнительный столбец для хранения предыдущего значения атрибута, например, `previous_address` .

При изменении данных, новое значение заносится в основной столбец, а старое значение переносится в столбец `previous_address` .

При работе с SCD часто используются ETL-инструменты, которые автоматизируют процесс обновления данных в соответствии с выбранным типом SCD. Эти инструменты могут содержать встроенные компоненты или шаблоны для обработки SCD, уменьшая ручные усилия и упрощая управление изменениями в данных.