

Change Tracking, Change Data Capture и SSIS 2012

Отслеживание изменений в БД,
синхронизация данных, инкрементальная
загрузка в хранилище с помощью SSIS,
тонкости настройки CDC.



Олонцев Сергей
email: sergey@olontsev.ru
twitter: @SergeyOlontsev
blog: <http://olontsev.ru/>



#178 | MOSCOW 2012

Наши спонсоры



Методы отслеживания изменений

- Полная перезагрузка данных
- Для INSERT операций столбец с типом datetime
- Сторонние программные продукты
- rowversion
- MERGE
- Триггеры
- Репликация



CDC vs. Change Tracking

Change Data Capture

- Отслеживание строк, столбцов и данных, которые были изменены
- Асинхронная операция

Change Tracking

- Отслеживание какие строки и столбцы были изменены (не данные)
- Синхронная операция

Change Tracking

Для обычного приложения нам требуется постоянное соединение с базой данных.

- Но соединение не всегда доступно
 - Ненадежное соединение
 - Периодически подключаемые клиенты
 - Приложения на мобильных устройствах
- В этом случае нам требуется
 - Легкий, но функциональный кэш
 - Синхронизация с основной базой данных

Change Tracking создан для работы с несколькими клиентами и поддерживает одно или двухстороннюю синхронизацию для каждого отдельного клиента.

Настройка Change Tracking

- Сначала нужно включить на уровне БД
 - Можно указать срок хранения
`ALTER DATABASE foo SET CHANGE_TRACKING = ON
(CHANGE_RETENTION = 2 days, AUTO_CLEANUP = ON)`
 - Нужны права db_owner или выше
- Включить на каждой таблице
 - Опционально можно отслеживать, какие столбцы изменились
`ALTER TABLE tab ENABLE CHANGE_TRACKING
WITH (TRACK_COLUMNS_UPDATED = ON)`
 - Нужны права владельца таблицы или выше
- Рекомендуется использовать snapshot isolation
 - Для лучшей транзакционной целостности

Таблицы Change Tracking

- Хранение данных change tracking происходит в системных таблицах
 - `sys.change_tracking_[идентификатор таблицы]`
– на каждую таблицу, в которой идет отслеживание данных
- Таблицы с метаданными
 - `sys.change_tracking_databases`
 - `sys.change_tracking_tables`

Что отслеживается?

- Строка в таблице с изменениями содержит
 - Primary key измененной строки
 - Номер версии
 - Операция, которая была произведена (insert/update/delete)
 - Битовая маска измененных столбцов (опционально)

Получение информации об изменениях

- Используется функция CHANGETABLE для получения изменений начиная с определенной версии
- Можно отслеживать изменения для индивидуальных клиентов, используя CHANGE_TRACKING_CONTEXT
- Отслеживается номер версии
 - CHANGE_TRACKING_MIN_VALID_VERSION
 - Клиенты, у которых номер версии ниже, должны сделать полную синхронизацию
 - CHANGE_TRACKING_CURRENT_VERSION

Влияние Change Tracking

- Change Tracking добавляет некоторую **СТОИМОСТЬ**
 - Примерно аналогично добавлению индекса на таблицу
 - Snapshot isolation
- Влияет на план запроса для insert\delete\update операций
- Запрещает switching partitions

Использование Change Tracking.

ДЕМОНСТРАЦИЯ



Change Data Capture



- Возможность отслеживать операции INSERT, UPDATE и DELETE
- Не требуется вносить изменений в приложение (например, специальные триггеры)
- Не требуется вносить изменения в структуру таблиц
- Сохраняет историю изменений в той же БД
- Конфигурируемый, автоматический процесс очистки

Change Data Capture

- Change Data Capture предназначен для инкрементальной загрузки изменений в Data Warehouse
- Таблицы для изменений зеркальны оригинальным
 - Это позволяет легко применять изменения на целевой ресурс
- Используется асинхронный процесс чтения журнала транзакций
 - Минимальное влияние на текущую систему
- Опция доступна только в редакции Enterprise (в т.ч. Evaluation, Developer)

Настройка CDC

- Разрешить на уровне БД
 - `sys.sp_cdc_enable_db`
 - Требуются права уровня `sysadmin`
 - Создаются новые объекты
 - Схема `cdc` для CDC объектов
 - Пользователь `cdc`
 - Таблицы с метаданными, вспомогательные процедуры и функции
 - Data Definition Language (DDL) триггер
 - Два задания SQL Agent
 - Capture job (для сбора изменений)
 - Cleanup job (задание очистки)

Настройка CDC

- Включаем для нужных таблиц
 - `sys.sp_cdc_enable_table`
 - Требуются права `sysadmin` или `db_owner`
 - Доступ к данным может быть ограничен специальной роли
 - Создается функция для просмотра изменений
 - `cdc.fn_cdc_get_all_changes`
- Информация о CDC доступна с помощью
 - `sys.sp_cdc_help_change_data_capture`
 - Таблицы с метаданными

Суммарные изменения (Net Changes)

- Можно указать опцию “allow net changes” когда включаем CDC на таблице
 - Функция `cdc.fn_cdc_get_net_changes...` будет возвращать только одну строку на каждую измененную строку
 - Строка суммарных изменений отображает состояние строки таблицы на момент последней операции (самый старший LSN)
 - Если в таблицу в определенном интервале строка будет добавлена, а затем удалена, то в суммарных изменениях это не отобразится

Что отслеживается?

Строка в таблице изменений содержит:

- Log sequence number (LSN)
- Метадата – какой тип изменения
- Bitmask – какие столбцы были изменены
- Значения столбцов

CDC и некоторые типы данных

- Разреженные столбцы (Sparse Columns)
 - Не поддерживает отслеживание изменений при использовании набора столбцов (column set)
- Вычисляемые столбцы (Computed Columns)
 - Изменения в вычисляемых столбцах не отслеживаются. Столбец появится в таблице изменений с подходящим типом, но в нем будет содержаться значение NULL.
- Отметка времени (Timestamp)
 - Тип данных в таблице изменений будет преобразован в двоичный.
- XML
 - Изменения в индивидуальных XML-элементах не отслеживаются.
- Типы данных BLOB
 - Предыдущий снимок столбца BLOB будет сохранен только при изменении самого столбца.

CDC и изменения структуры

- Если происходит удаление столбца
 - Столбец остается в CDC таблице
 - Строки в CDC таблице с текущего момента будут иметь значение NULL в удаленном столбце
- Если происходит добавление столбца
 - Столбец не начинает автоматически отслеживаться с помощью CDC
- Можем добавить новый экземпляр отслеживания
 - Допускается до 2х экземпляров на таблицу
 - Эта схема специально реализована для переходного периода

CDC и шаблоны SSMS

- Большое количество шаблонов для работы с CDC доступно в SSMS в окне Template Explorer
- Разбито на следующие функциональные группы
 - Administration
 - Configuration
 - Enumerating Changes
 - Querying Metadata
- Шаблоны для получения изменений
 - Все изменения до текущего момента
 - Все изменения в указанном интервале (LSN)
 - Все изменения с момента последнего запроса

Использование Change Data Capture и SSIS.

ДЕМОНСТРАЦИЯ



Рекомендации для CDC

- CDC может являться причиной transaction log contention
- CDC влияет на обрезание журнала транзакций (также как и транзакционная репликация)
- Частота сбора данных и очистки настраиваются
- Необходим SQL Agent
- Если используется модель восстановления BULK_LOGGED, то все операции становятся полностью логируемыми, кроме операций с индексами
- Для восстановления или присоединения базы данных с CDC на другой сервер используйте опцию KEEP_CDC (нельзя восстановить на редакцию, отличную от Enterprise)

Рекомендации для CDC: Storage

- Планируя внедрение change data capture, необходимо учесть значительное увеличение размера журнала транзакций и количество I/O операций
- Указывайте отдельную файловую группу в процедуре `sys.sp_cdc_enable_table`
- Рекомендуется изменять файловую группу по умолчанию перед запуском `sys.sp_cdc_enable_db` таким образом, чтобы таблицы с метаданными и особенно `cdc.lsn_time_mappings` были расположены в другой файловой группе, отличной от PRIMARY

Рекомендации для CDC: Конфигурация

- Старайтесь избегать сценариев, когда строка сразу же изменяется после вставки
- Старайтесь избегать использование change data capture для отслеживания изменений в таблицах, на которых часто идут большие операции обновления данных

Рекомендации для CDC: Конфигурация

- **Параметры:**
 - Отслеживайте изменения только в тех столбцах, для которых вам это действительно нужно
 - Если вам не нужна поддержка суммарных изменений (net changes) – не используйте
 - Протестируйте, справляется ли CDC с нагрузкой на вашу систему, подберите оптимальные параметры capture job
- **Очистка:**
 - Если возможно, запускайте очистку, когда нет активности
 - Попробуйте увеличивать параметр threshold до тех пор, пока не найдете наиболее подходящее для вас значение

Какой метод выбрать?

- Change Data Capture
 - Хорошим примером потребителя данных, для которого предназначена эта технология, является приложение для извлечения, преобразования и загрузки данных (ETL) в хранилище.
 - Не отслеживает изменение данных на индивидуальной основе.
- Change Tracking
 - Хороший пример - это периодически подключаемое приложение одно или двусторонней синхронизации, которое обновляет локальное хранилище или кэш (может потребоваться реализация системы распознавания конфликтов).
 - Помогает отслеживать изменения для индивидуальных клиентов.
- Искать другое решение



<http://olontsev.ru/go/TrackingChanges>
sergey@olontsev.ru