Качановский Ю.П., Язынин С.А.

**Анализ механизмов односторонней синхронизации данных большого объёма
между различными СУБД**

*Липецкий государственный технический университет
(Россия, Липецк)*

doi: 10.18411/lj-05-2019-17

idsp: ljjournal-05-2019-17

Аннотация

Статья посвящена исследованию и анализу механизмов односторонней синхронизации данных. Исследованы существующие методы синхронизации, определены наиболее подходящие для выполнения конкретной прикладной задачи. Изучение механизмов односторонней синхронизации осуществлялось на построенной имитационной модели [1]. Имитационная модель учитывает особенности различных СУБД (PostgreSQL и Oracle) и реализует основные виды синхронизации. Анализ механизмов синхронизации осуществлялся на основе сравнения времени выполнения процесса репликации данных и степени использования вычислительных ресурсов.

Ключевые слова: синхронизация данных, репликация данных, односторонняя синхронизация, синхронизация между PostgreSQL и Oracle.

Abstract

The article is devoted to the study and analysis of the mechanisms of one-way data synchronization. The existing methods of synchronization are investigated, the most suitable ones to execute a concrete applied task are determined. The study of the mechanisms of one-way synchronization was carried out on the constructed simulation model. The simulation model takes into account the features of various DBMS (PostgreSQL and Oracle) and implements the main types of synchronization. The analysis of synchronization mechanisms was carried out based on comparison of the execution time of the data replication process and the degree of computing resources usage.

Keywords: data synchronization, data replication, one-way synchronization, synchronization between PostgreSQL and Oracle.

Введение

Одной из особенностей современных информационных систем является распределённое хранение данных, что часто приводит к необходимости синхронизации данных между различными частями системы. Задача синхронизации реляционных баз данных возникает при необходимости обеспечения идентичности структуры и информационного наполнения нескольких экземпляров базы данных, эксплуатируемых в организациях, подразделения которых включены в многосторонний информационный обмен. Периодическая синхронизация баз данных обеспечивает возможность совместной работы пользователей из различных подразделений или отделов с массивами данных и электронными документами [2].

Необходимость анализа механизмов односторонней синхронизации данных большого объёма между различными СУБД возникла в Липецком государственном техническом университете при разработке подсистемы расчёта книгообеспеченности и книговыдачи для единой информационной системы (ЕИС) ЛГТУ.

Сотрудники библиотеки ЛГТУ для работы с книжным фондом применяют информационную систему «Библиосфера», её база данных постоянно пополняется новыми записями, которые необходимо своевременно добавлять в ЕИС ЛГТУ. Так как функционал подсистемы расчёта книгообеспеченности и книговыдачи не предполагает изменение данных о книжном фонде библиотеки ЛГТУ, метод синхронизации должен быть односторонним.

Целью проведенного анализа механизмов односторонней синхронизации данных большого объёма между различными СУБД являлся поиск оптимального решения для переноса и синхронизации информации из библиотечной информационной системы

«Библиосфера» в БД единой информационной системы ЛГТУ. При анализе учитывались объём данных хранящихся в БД «Библиосфера», специфика их изменения и добавления новых. Также учитывалась производительность серверного оборудования, скорость каналов данных, физическое размещение баз данных ЕИС ЛГТУ и ПО «Библиосфера» и различия между СУБД PostgreSQL и СУБД Oracle.

1. Исследование различия типов данных PostgreSQL и Oracle

Перед осуществлением синхронизации данных между различными СУБД необходимо выявить их основные различия. Поскольку синхронизация будет осуществляться между PostgreSQL и Oracle были определены отличия и сходства между этими системами, в первую очередь с точки зрения имеющихся наборов типов данных. Если типы данных различаются существенно, придется создать механизм приведения типов. Во время исследования было составлено соответствие основных типов данных СУБД PostgreSQL и СУБД Oracle.

2. Исследование видов синхронизации данных в реляционных СУБД

Для переноса данных из одной БД в другую необходимо реализовать процесс репликации, т.е. приведения двух и более баз данных или их отдельных таблиц к одинаковому состоянию [3].

Для организации репликации предлагается большое количество методов синхронизации данных. Каждый из них имеет свои преимущества и недостатки, а, следовательно, и область применения. В настоящее время наиболее популярными являются два метода синхронизации в реплицируемых базах данных: по данным журнала транзакций, по текущему состоянию таблиц [4].

Синхронизация по журналу транзакций. Этот метод используется чаще всего. Результаты операций, выполняемых в базе данных (вставка, удаление и изменение записей), хранятся в каком-либо специально организованном журнале (обычно это отдельная таблица или набор таблиц в самой БД). Когда транзакция выполняется, данные о выполненных операциях помещаются в журнал, чтобы впоследствии можно было восстановить базу данных и повторить операции. Как правило, журнал заполняется с помощью триггеров, связанными с таблицами, участвующими в репликации.

Синхронизация по текущему состоянию. Синхронизация по текущему состоянию всегда выполняется с установлением соединения между синхронизируемыми БД. В связи с этим применять этот метод без какого-либо канала связи между серверами невозможно. Процесс синхронизации устанавливает соединение с обеими базами данных и сканирует все таблицы, подлежащие синхронизации, в поисках различий. Различия в синхронизированных таблицах можно определить по первичному ключу записи [5]. Средство синхронизации «выравнивает» таблицы, приводя обе базы данных к одинаковому набору данных. При односторонней синхронизации можно не производить поиск различий, допускается простое копирование данных из главной БД.

Все методы репликации можно классифицировать по двум наиболее важным критериям [6].

1) По времени проведения сеанса репликации:

1.1) Синхронная – обновление данных на других узлах происходит в одной транзакции (репликация реального времени).

1.2) Асинхронная – обновленные данные распространяются на другие узлы спустя некоторое время, а не в той же транзакции (отложенная репликация).

2) По направлению проведения:

2.1) Односторонняя – данные изменяются только в узлах-источниках, а в узлах-приемниках эти данные только хранятся и не подвергаются изменениям.

2.2) Многосторонняя – данные могут изменяться на всех узлах системы.

Односторонние методы синхронизации как правило в реализации проще чем многосторонние. Практически любой метод синхронизации может быть применён для односторонней синхронизации данных.

3. Разработка имитационной модели для анализа механизмов односторонней синхронизации

Для проведения сравнительного анализа методов односторонней синхронизации была разработана имитационная модель, позволяющая осуществлять репликацию между базой данных библиотечной информационной системы «Библиосфера» и единой информационной системой ЛГТУ.

Программное обеспечение модели написано на языке Python 2.7. Для осуществления взаимодействия с СУБД PostgreSQL используется библиотека «psycopg2», для подключения к СУБД Oracle – «cx_Oracle».

В модели было реализовано два способа односторонней синхронизации данных:

- 1) очистка таблиц в БД ЕИС ЛГТУ и заполнение их данными из БД «Библиосфера»;
- 2) выгрузка таблиц для синхронизации, определение новых записей в БД «Библиосфера» и перенос в БД ЕИС ЛГТУ.

Схема работы ПО для синхронизации представлена на рисунке 1.

Рис 1. Схема работы средства синхронизации

Реализованные методы по сути являются разновидностью «синхронизации по текущему состоянию». Отличия заключаются в способе добавления данных. В 1 способе таблица, в которую необходимо добавить новые данные, очищается и заполняется всеми записями из соответствующей таблицы в БД «Библиосфера». Во 2 способе вначале происходит выгрузка таблиц для синхронизации из БД «Библиосфера» и ЕИС ЛГТУ, затем определяются новые записи, которые будут добавлены.

Для настройки параметров синхронизации созданы два файла конфигурации. Первый файл конфигурации содержит настройки ПО для синхронизации, а второй позволяет задать таблицы для синхронизации, а также способ их обновления.

Файл настройки таблиц, подлежащих синхронизации, на данный момент содержит название таблиц и тип синхронизации, который для каждой таблицы можно задать индивидуально.

Для осуществления процесса синхронизации данных необходимо предпринять следующие действия:

- 1) в БД ЕИС ЛГТУ создать таблицы полностью повторяющие переносимые;
- 2) настроить модель для синхронизации;
- 3) запустить процесс синхронизации.

В процессе работы модель создаёт лог и записывает основную информацию, связанную с процессом синхронизации. В лог записываются время запуска и завершения обновления конкретной таблицы, количество записей в PostgreSQL и Oracle, количество обновлённых записей, количество ошибок.

4. Проведение вычислительного эксперимента

В ходе проведения вычислительного эксперимента сравнивались два способа односторонней синхронизации данных (далее способ 1 и способ 2):

- 1) очистка таблиц в БД ЕИС ЛГТУ и заполнение их данными из БД «Библиосфера»;
- 2) выгрузка таблиц для синхронизации, определение новых записей в БД «Библиосфера» и перенос в БД ЕИС ЛГТУ.

Имитационная модель для односторонней синхронизации исследовалась в виртуальной машине со следующими характеристиками:

- операционная система Windows XP x86;
- 4 Гб оперативной памяти;
- центральный процессор Intel Xeon X3440 с частотой 2.5 ГГц.

Вычислительный эксперимент проводился для разного количества записей и для разных таблиц. Время выполнения процесса синхронизации получено из лога. Измерялась также нагрузка на ПК, где запускалась модель для синхронизации, на нём также находилась БД информационной системы «Библиосфера». Результаты проведения эксперимента отображены в таблицах 2 и 3.

Таблица 1.

Результат эксперимента для таблицы «Авторы» (65632)

	Способ 1	Способ 2	Способ 2 (4 потока)
Время (секунды)	11160	312	203
Объём используемой оперативной памяти на ПК (Мб.)	90	160	162
Загрузка ЦП (%)	3	25	92
Добавлено записей	65632	1000	1000

Таблица 2.

Результат эксперимента для таблицы «Издательство» (9066)

	Способ 1	Способ 2	Способ 2 (4 потока)
Время (секунды)	1557	26	7
Объём используемой оперативной памяти на ПК (Мб.)	56	63	64
Загрузка ЦП (%)	3	25	94
Добавлено записей	9066	100	100

В зависимости от разных условий скорость синхронизации несколько отличалась. В частности, для способа 2 скорость зависела не только от общего количества записей в таблице, но и от количества новых записей, а также от количества используемых ядер центрального процессора (в режиме многозадачности). Также на время синхронизации влияет скорость сетевого соединения с сервером ЕИС ЛГТУ.

Вычислительный эксперимент позволил сделать следующие выводы. Способ синхронизации данных с полной очисткой не позволяет достаточно быстро провести синхронизацию, при этом возрастает нагрузка на сервер, где расположена БД ЕИС ЛГТУ (увеличивается количество запросов), но снижается объём требуемой оперативной памяти на ПК, где запущена программа синхронизации. Чтобы снизить нагрузку на сервер ЕИС ЛГТУ можно воспользоваться способом определения новых записей в БД «Библиосфера», но следует отметить, что данный метод значительно увеличивает объём требуемой оперативной памяти на ПК и повышает загрузку центрального процессора.

Заключение

Проанализированы разные механизмы синхронизации данных между различными СУБД. В результате вычислительного эксперимента получены численные статистические данные для различных методов односторонней синхронизации. На основе которых можно выявить наиболее быстрый способ репликации данных, а также проанализировать потребление вычислительных ресурсов. После анализа были выбраны и реализованы наиболее подходящие механизмы для решения поставленной задачи.

Результаты анализа различных механизмов синхронизации позволяют сделать вывод, что одного оптимального механизма для всех ситуаций не существует. В зависимости от объёма информации, скорости соединения между базами данных, надёжности соединения необходимо протестировать несколько механизмов и на основе выбранных критериев найти оптимальный.

***

1. Использование имитационного моделирования для определения основных характеристик загруженности производственного цеха [Текст] – Качановский Ю.П., Широков А.С. // Вести высших учебных заведений Черноземья. Липецк: ЛГТУ, 2017, №1 (47). С.66-72
2. Организация пакетной синхронизации реляционных баз данных [Текст] –Жучков Д. В. // Образовательные ресурсы и технологии. 2016. – С. 161-168.

3. Разработка подсистемы синхронизации между АИС и системой дистанционного обучения в рамках единой информационной среды «Школа информатики СГАУ» [Текст] – Кузнецов И. А. // Актуальные проблемы авиации и космонавтики. 2014. С. 317–318.
  4. Разработка системы репликации для распределенной базы данных предприятия [Текст] – Шаталова Ю. Г., Жиглов Я. В. // Символ науки. 2017. – С. 137-141.
  5. Разработка программного приложения для синхронизации базы данных справочника материалов и сортов [Текст] – Баранов Д.Е., Давыденко В.А., Цупранков А.В., Янишевская А.Г., Василевский В.Б., Соседко В.В. // Инженерный вестник Дона. 2013.
  6. Аналитический анализ обзор методов репликации и синхронизации баз данных [Текст] – Н.С. Рябков // Информационные технологии в менеджмент качество и инновационном менеджменте. 2006. № 6. С. 56-63
-