УДК 004.1

ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ ХРАНИЛИЩ ДАННЫХ

Мозохин А.Е., Сахаров И.Е. к.т.н.

В работе описаны современные методы и стандарты анализа производительности хранилищ данных, примеры тестов, представлена математическая модель с учетом коэффициента масштабирования.

Ключевые слова: оценка, тест производительности хранилищ данных, коэффициент загрузки, время загрузки базы данных.

UDC 004.1

EVALUATE THE PERFORMANCE OF THE DATA WAREHOUSE

Mozohin A.E., Sakharov I.E.

In modern techniques and standards of performance analysis data warehousing test examples, the mathematical model with the scaling factor.

Keywords: assessment, test performance data warehousing, load factor, load time database.

Согласно Федеральному закону РФ от 27 июля 2006 года № 149-ФЗ «Об информации, информационных технологиях и о защите информации»: «Информационная система – совокупность содержащейся в базах данных информации и обеспечивающих ее обработку информационных технологий и технических средств». Основу информационной системы составляет вычислительная система, включающая такие компоненты, как кабельная сеть и активное сетевое оборудование, компьютерное и периферийное оборудование, системы хранилищ данных и различное программное обеспечение.

Стандартной операцией при любом внедрении или изменении существующей информационной системы является оценка необходимого быстродействия системы и планирование необходимых ресурсов, как программных, так и аппаратных, для ее реализации. В настоящее время не существует абсолютно точного решения этой задачи, и если, несмотря на ее сложность и сто-

имость такой алгоритм будет предложен каким либо производителем, то даже небольшие изменения в аппаратной части, версии программного обеспечения, конфигурации системы или количестве или стандартном поведении пользователей, приведут к появлению значительных ошибок.

Если отбросить метод оценки производительности системы основанный на опыте лица принимающего решения или системного интегратора, то большинство существующих методов основывается на том или ином типе тестирования.

Использование современных хранилищ данных характеризуется двумя основными параметрами: производительность (скорость загрузки данных и выполнения расчётов) и скорость получения отчётов. Эффективность принятия решений одинаково зависит от обоих параметров.

За последние несколько лет производительность хранилищ данных стала существенно падать. Сегодня сложные инструменты позволяют интегрировать данные практически в режиме реального времени, все больше и больше пользователей применяют удобные в обращении средства бизнес-аналитики (Business Intelligence). Объёмы данных, загружаемых в хранилища, растут с невероятной скоростью.

Самым авторитетной организацией проводящей универсальные интегральные тесты является TPC (Transaction Processing Performance Council – Совет по Обработке Транзакций). TPC является независимой некоммерческой организацией, созданной для исследования обработки транзакций и производительности систем управления базами данных (СУБД) и распространения объективной и воспроизводимой информации о производительности в тестах TPC для компьютерной индустрии.

В настоящее время тесты TPC включают не только техническую спецификацию, но и формализованную процедуру проведения тестов, представления результатов, а также обязательный аудит результатов независимой аудиторской компанией. Можно не соглашаться с техническими аспектами тестов и их соответствием реальным условиям эксплуатации бизнес прило-

жений, но TPC является наиболее авторитетной в мире организацией по тестированию производительности и ее результаты действительно независимы и воспроизводимы.

TPC использует следующие параметры для оценки производительности хранилища данных:

1. Время загрузки базы данных.
2. Время использования запросов.
3. Время сопровождения данных (обновление и изменения данных).

Существуют стандарты оценки производительности. Наиболее широкое распространение получил стандарт TPC-H. Тест TPC-H оценивает производительность Систем Поддержки Принятия Решений. Основная метрика – QphH (TPC-H Composite Query-Per-Hour Performance) публикуется совместно с указанием размера базы данных – TPC-H QphH@размер, изменяемый в пределах от 1 до 10 тыс. Гбайт. Вторая метрика «цена/производительность» (TPC-H Price/Performance Metric) измеряемая в долл./QphH также относится к размеру базы данных TPC-H Price-per-QphH@размер. На сегодняшний день количество уже протестированных систем невелико.

Однако современные тенденции развития хранилищ данных привели к тому, что архитектуры систем включают гораздо большее количество таблиц (чем это предусмотрено в стандарте TPC-H), а также ушли от чистой третьей нормальной формы к разновидностям схем типа – «звезда».

Следующим этапом развития стало появление стандарта TPC-DS. TPC-DS отводит главное место возможности системы включать новые данные по мере роста объемов исходных данных. Очевидно, что включение теста операцией по работе с данными (трансформацией на основе SQL) позволяет говорить о TPC-DS как о первом отраслевом стандарте для оценки ETL-процессов (**E**xtraction – извлечение данных из внешних источников в понятном формате, **T**ransformation – преобразование структуры исходных данных в структуры, удобные для построения аналитической системы, **L**oading – загрузка данных в хранилище). Так, в отличие от стандарта TPC-H в расчёте

основного параметра оценки производительности системы по TPC-DS (Query-per-Hour Performance Metric at Scale Factor – показатель производительности запросов, выполненных в час, при заданном коэффициенте масштабирования) учитывается как время создания базы данных (загрузки данных), так и время на сопровождение данных (процессы ETL):

$$QphDS@SF = SF * 360\left(\frac{198*S}{T_{QR1}+T_{DM}+T_{QR2}+0{,}01*S*T_{Load}}\right) \quad (1)$$

где T_{QR1} – время на выполнение запросов первой серии.

T_{QR2} – время на выполнение запросов второй серии.

T_{DM} – время на выполнение операций по сопровождению данных.

T_{Load} – время на тестирование загрузки базы данных.

S – число потоков, которые выполняются при оценке производительности.

SF – коэффициент масштабирования.

Проведение тестирования системы в соответствии с TPC-DS предусматривает выполнение следующих действий:

1. Тестирование загрузки базы данных.
2. Тестирование производительности.

Под тестированием загрузки базы данных понимается создание тестовой базы. Тестирование производительности состоит из выполнения двух серий запросов и сопровождения данных.

Тесты TPC на сегодняшний день являются лучшими тестами для оценки производительности систем, построенных по клиент-серверным технологиям. Данные тесты позволяют проводить анализ предлагаемых производителями технологий и решений и сравнивать их между собой, сравнения различных архитектур и продуктов.

Список использованной литературы

1. Елашкин М. Оценка производительности программно-аппаратных решений. Проблема выбора. – Режим доступа // www.elashkin.com.

2. Hagerty J., Sallam R.L., Richardson J. Magic Quadrant for Business Intelligence Platforms // Business Intelligence facil. 2012. №5.

3. Nambiar R., Poess M. The Making of TPCDS. – Режим доступа // www.tpc.org.

4. Сахаров И.Е., Мозохин А.Е. Оценка производительности информационных систем при работе с хранилищами данных // Научно-Технический вестник Поволжья. 2012. №6.