Облачные технологии при обработке данных большого объема

Working with big data

Ключевые слова: данные большого объема – big data; моделирование сложных систем – modeling of complex systems; облачные технологии – clouding technologies; частное облако – private cloud.

Рассмотрена специфика обработки данных, содержащих несколько миллионов записей на примерах задач моделирования инфокоммуникационных систем. Проведены эксперименты по обработке результатов практических работ около 100 студентов по дисциплине: «Моделирование инфокоммуникационных систем». В статье предлагается способ обработки данных большого объема с помощью технологии PowerPivot и с использованием облачных технологий.

The article describes methods of processing big data in excel' format, which contains millions records of the modeling samples in info-communication systems. Using clouding technologies will allow degreasing costs of information working and increase efficiency of scientific research. We suggest the PowerPivot technology implementation for big data processing in clouding.

ХОРУЖНИКОВ / KHORUZHNIKOV S.
Сергей Эдуардович
(xse@vuztc.ru)
кандидат физико-математических наук, доцент,
декан факультета инфокоммуникационных технологий,
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
директор ЦАО ИТ,
Санкт-Петербург

ЗУДИЛОВА / ZUDILOVA T.
Татьяна Викторовна
(zudilova@limtu.spb.ru)
кандидат технических наук, доцент,
заведующий кафедрой программных систем,
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
заместитель директора ЦАО ИТ,
Санкт-Петербург

ОСИПОВ / OSIPOV N.
Никита Алексеевич
(zudilova@limtu.spb.ru)
кандидат технический наук,
доцент кафедры программных средств,
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
Санкт-Петербург

ВВЕДЕНИЕ

В настоящее время большое внимание уделяется вопросам обеспечения научных исследований в условиях существенного роста экспериментальных данных. По мере роста производительности вычислительных систем исследователи получили дополнительные возможности при решении сложных задач, требующие в основном мощных аппаратных средств. Современный вычислительный эксперимент ориентирован в большей степени на данные и основан на неограниченных возможностях объема этих данных.

Для примера рассмотрим работу страховых компаний, в которых при выработке оптимальной стратегии страхования водителей собирается и оценивается разного рода статистика, в том числе – пол, возраст, стаж водителя и множество других параметров, характеризующих человека за рулем, но все это косвенные оценки, не привязанные к конкретной личности [1]. С развитием современных средств связи появилась возможность устанавливать, например, в транспортном средстве блок с регистраторами и передатчиками, который ежесекундно в режиме реального времени информирует о манере езды водителя, ускорениях при разгоне и торможении, скорости прохождения поворотов и других параметрах. Обработка информации с регистраторов позволяет получать большие объемы данных, характеризующие поведение водителя за рулем. Анализ таких данных является более сложной задачей ввиду необходимости обработки больших объемов данных.

Применение аналогичных методов и средств в других предметных областях позволяет исследова-

телям получать информацию в больших объемах, что неизбежно требует нового подхода к ее обработке и анализу. В литературе содержится описание подобных задач [1, 2], но не описаны методы их реализации на основе современных ИКТ.

Подобные исследовательские задачи решаются студентами факультета ИКТ НИУ ИТМО на практических занятиях по дисциплине «Моделирование инфокоммуникационных систем». Суммарный объем данных результатов моделирования сложных систем ввиду большого числа студентов и проведенных ими исследований превышает, как правило, несколько миллионов строк таблиц Excel. При применении стандартных подходов к обработке и анализу полученных данных требуются существенные вычислительные и временные ресурсы, которые зачастую не эффективны в решении актуальных задач моделирования ввиду отсутствия надежности, разделения ресурсов, безопасности, использования старых вычислительных алгоритмов.

В статье рассматривается новый подход к решению задач обработки больших данных с использованием технологии облачных вычислений, при котором реализуется образовательная платформа, позволяющая студентам получать качественно новый уровень знаний с использованием практически неограниченных вычислительных сервисов.

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ БОЛЬШИХ ДАННЫХ

Экспоненциальный рост объема данных кардинально изменил природу исследований. Первой парадигмой науки был эксперимент, затем появилась теория, формализующая экспериментальные знания. В XX веке новой парадигмой стало вычислительное моделирование, которое позволило проверять теории в тех областях, где экспериментирование было либо крайне дорого, либо невозможно. Созданные объемы данных привели к четвертой парадигме науки [3]: вместо того, чтобы использовать данные для проверки теории, теперь можно, опираясь на большие объемы данных, формулировать теории. Эти данные принимают форму достаточно точных моделей, которые было бы невозможно создать, основываясь на небольших наборах информации. Например, сейчас созданы точные системы автоматического перевода на основе моделей, построенных в результате байесовского статистического анализа больших совокупностей переведенных текстов.

Принято считать большими данные при наличии одного из трех признаков [2]:

- объем – количество строк в таблицах данных превышает десятки миллионов;
- разнородность данных – структурированные данные и информация из самых разнообразных источников данных хранятся совместно;
- скорость поступления данных – быстрый плотный поток информации, поступающей с датчиков.

Считается, что если присутствует один из указанных признаков, это задача из области больших данных. При этом признаки обычно оказываются взаимосвязаны, разнородные данные, как правило, имеют большие объемы, а информация с датчиков поступает с огромной скоростью. Задача обработки результатов практических работ студентов, к которым относятся лабораторные и контрольные работы, домашние задания и результаты тестирования по дисциплине «Моделирование инфокоммуникационных систем» в полной мере обладает первым из перечисленных признаков.

ДОСТОИНСТВА ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ В УЧЕБНОМ ПРОЦЕССЕ

Отметим основные преимущества и достоинства технологий облачных вычислений применительно к решаемой задаче [4, 5]:

- обеспечение при решении задач обработки и анализа больших данных высокой масштабируемости, надежности, разделения ресурсов, гибкости подключаемых сервисов, безопасности, использования старых вычислительных мощностей, а также легкости администрирования и лицензионной чистоты;
- эффективное использование учебных площадей для студентов, так как отпадает необходимость выделять отдельные и специально оборудованные помещения под традиционные компьютерные классы, что приводит к сокращению затрат, необходимых на создание и поддержание компьютерных классов;
- качественно новый уровень получения современных знаний по специальности, так как студенты получают доступ к учебным материалам и сервисам для решения задач моделирования в любое время и в любом месте, где есть Интернет;
- возможность быстро создавать аналитические отчеты и использовать их результаты в ходе учебного процесса;
- возможность осуществления для студентов интерактивного общения с преподавателем в процессе решения возникающих вопросов;
- централизованное администрирование обслуживающим персоналом и преподавателем программных и информационных ресурсов, используемых в учебном процессе.

По сравнению с персональным компьютером, вычислительная мощность, доступная студентам

Рис. 1. Обработка данных

«облачных» компьютеров, практически ограничена лишь размером «облака», т.е. общим количеством удаленных серверов. Студенты могут запускать более сложные задачи, с большим количеством необходимой памяти и места для хранения данных тогда, когда это необходимо. Возможность запуска множества копий приложения на виртуальных машинах представляет преимущества масштабируемости: количество экземпляров приложения способно практически мгновенно увеличиваться по требованию, в зависимости от нагрузок. Таким образом, можно сделать вывод о целесообразности внедрения в учебный процесс при проведении дисциплин, подобных «Моделированию инфокоммуникационных систем», технологий облачных вычислений.

РЕАЛИЗАЦИЯ ОБРАБОТКИ ДАННЫХ БОЛЬШОГО ОБЪЕМА С ПОМОЩЬЮ POWERPIVOT

Для обработки результатов практических работ студентов базовых элементов бизнес-аналитики Microsoft Excel явно недостаточно. Здесь необходимо работать с большими данными в режиме реального времени. Применение существующей надстройки PowerPivot для Excel 2010 обеспечивает возможность добавления и интеграции больших объемов данных в книги Excel. PowerPivot преодолевает существующие ограничения анализа больших объемов данных на настольном компьютере с помощью эффективных алгоритмов сжатия для загрузки даже самых больших наборов данных в память.

Выбор PowerPivot для решения задачи обработки практических работ студентов обусловлен следующими причинами. Во-первых, PowerPivot представляет собой набор приложений и сервисов, которые позволяют студентам и преподавателям самостоятельно создавать аналитические решения, поддерживает связывание между собой и расширение показателей в больших объемах данных, загруженных из гетерогенных источников (например, таких как Microsoft SQL Server, Access, Excel, SQL Azure, SSAS, Oracle, текстовые файлы) [6]. Во-вторых, на основании этих данных PowerPivot позволяет создавать таблицы и графики (PivotTables и PivotCharts), управляемые с помощью обычных и визуальных фильтров. И в-третьих, файл PowerPivot, созданный с помощью Excel, можно опубликовать на портале SharePoint, который при решении поставленной задачи использовался для реализации облачной технологии. Ниже представлены этапы обработки данных, полученных студентами при решении исследовательских задач моделирования инфокоммуникационных систем с помощью PowerPivot (рис. 1). Рассмотрим каждый этап подробнее.

Формирование исходных данных для обработки можно считать предварительным этапом. Исходными данными являются результаты лабораторных работ, тестов и домашних заданий 100 студентов, которые сводятся в отчет в таблицах Excel. Загрузка данных выполняется в окно PowerPivot, которое может содержать несколько таблиц, каждая из которых находится на отдельной вкладке. Таблицы вместе со столбцами образуют базу данных, хранящуюся в памяти. Список сданных отчетов сводится в книге Excel. Он содержит идентификатор студента (StudentID), идентификатор работы (LabrabID) и идентификатор преподавателя (InstructorID).

На этапе **преобразования данных** список представляется в виде связной таблицы PowerPivot, а затем создаются связи с таблицами Student и Instructor на основе сопоставления StudentID с полем StudentKey, a InstructorID – с полем Instructor. Для отображения данных в связанной таблице добав-

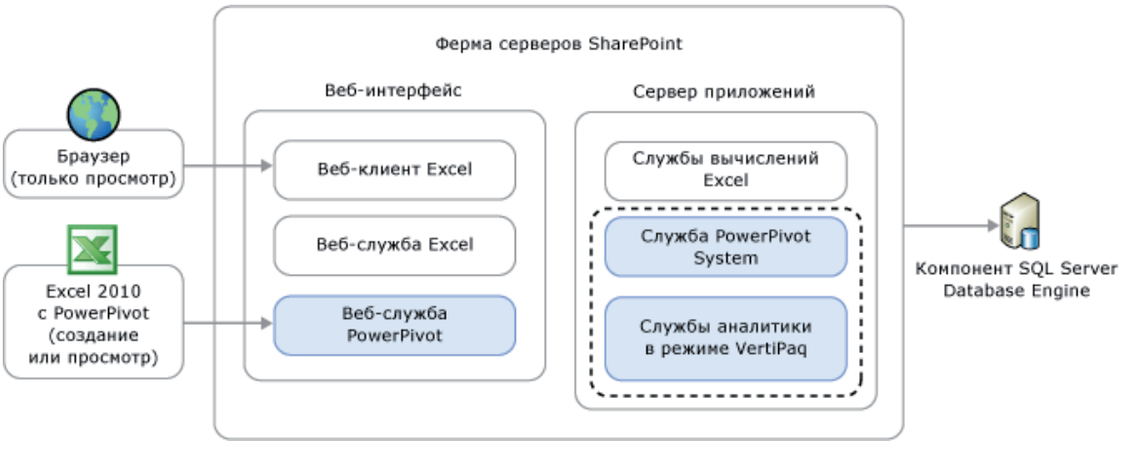

Рис. 2. Компоненты PowerPivot

ляются два новых вычисляемых столбца с помощью формулы RELATED(). Для определения количества сданных работ каждым студентом в таблице Student создается новый вычисляемый столбец, содержащий формулу, которая находит записи для каждого студента в таблице StudentLabrab и определяет количество сданных работ по каждому студенту с помощью формулы COUNTROWS().

На заключительном этапе **создания и форматирования отчетов PowerPivot** формируются отчеты для анализа данных. Отчеты PowerPivot создавались на основании построенной модели PowerPivot с помощью одного из трех инструментов: PivotTables (для создания сводных таблиц), PivotCharts (для создания сводных диаграмм) и двух функций Excel [7] «CUBE» CUBEMEMBER (возвращает элемент) и CUBEVALUE (возвращает агрегированное значение).

Для наглядности отчеты PivotTables и PivotCharts были отформатированы с помощью стандартных стилей и настроек форматирования. Отчеты PowerPivot размещены на сервере, что позволило расширить следующие функции PowerPivot:

1. Импортировать миллионы строк данных из различных источников в одну книгу Excel, создавать отношения между разнородными данными, создавать вычисляемые столбцы и показатели с помощью формул, строить сводные таблицы и сводные диаграммы с последующим анализом данных.

2. Создавать быстрые расчеты и анализировать большие объемы данных при максимально эффективном использовании возможностей многоядерных процессоров для быстрых вычислений.

Данные, с которыми студенты работают в окне PowerPivot, сохраняются в аналитической базе данных в книге Excel. Мощная система выполняет загрузку, запросы и обновление данных в базе данных. Поскольку данные PowerPivot внедряются в книгу Excel, они сразу становятся доступными для сводных таблиц, сводных диаграмм и других функций Excel, используемых для агрегатной обработки и взаимодействия с данными. Данные PowerPivot и объекты представления Excel хранятся внутри одного файла книги.

Облачные сервисы для обработки результатов данных реализованы через центр администрирования SharePoint 2010 [8]. Службы Excel включены на сервере SharePoint и студенты могут просматривать и взаимодействовать с книгами в окне браузера без необходимости установки какого-либо дополнительного клиентского ПО. Развертывание клиентских и серверных приложений PowerPivot включает в себя несколько компонентов, которые интегрируются с Excel и службами Excel в ферме SharePoint. На рисунке 2 показаны клиентские и серверные компоненты PowerPivot, позволяющие реализовать облачные технологии.

Отметим следующие преимущества PowerPivot при решении задачи обработки практических результатов данных 100 студентов по курсу «Моделирование инфокоммуникационных систем»:

- применение стандартных возможностей и функций Excel;
- практически неограниченная поддержка источников данных. PowerPivot дает возможность импортировать и комбинировать источники данных из любого расположения для анализа больших объемов данных;
- удобная информационная панель управления PowerPivot позволила отслеживать общие

приложения и управлять ими для обеспечения безопасности, высокой доступности и производительности;

• новый язык формул – язык выражений анализа данных (DAX), который расширил имеющиеся в Excel возможности работы с данными и позволил выполнять более сложное группирование, вычисления и анализ;

• публикация файла на сервере SharePoint с размещением сервисов PowerPivot для обработки данных.

Студенты имеют возможность просмат-ривать результаты обработки своих данных и оперативно реагировать на них с целью, например, повышения оценки и продолжения работ.

Таким образом, применение PowerPivot совместно с SharePoint позволило повысить производительность и эффективность работы студентов и преподавателя при обработке больших объемов аналитических данных студентов с использованием облачных технологий.

ЗАКЛЮЧЕНИЕ

Благодаря ресурсам «облака» удобные инструментальные средства, с которыми студенты и преподаватели ежедневно работают, стали во много раз более мощными. Появилась возможность для преподавателей работать с большим количеством данных и применять более сложные вычисления, используя при этом привычные инструменты. Проведенные исследования показали, что применение PowerPivot, особенно совместно с SharePoint за счет реализации доступа к данным и их обработки, повысило производительность и эффективность работы при решении аналитических задач студентами, с одной стороны, и обработки их результатов преподавателями – с другой. Возможности реализованных облачных технологий обеспечили обработку требуемого объема данных с существенным запасом, а время обработки снизилось до нескольких минут.

Поставленная задача реализована в рамках выполнения пилотного проекта «Разработка и создание сегмента корпоративной облачной инфраструктуры для формирования системы воспроизводства высококвалифицированных кадров» в Центре авторизованного обучения ИТ НИУ ИТМО.

Литература

1. *Черняк Л.* О больших данных с четырех сторон. – www.osp.ru.
2. *Дубова Н.* Большие данные – комплексный подход. – www.osp.ru.
3. *Геннон Д., Рид Д., Барга Р.* Облака: демократизация научных вычислений. – www.osp.ru.
4. *Риз Д.* Облачные вычисления. – СПб.: «БХВ – Петербург», 2011.
5. *Фингар П.* Облачные вычисления – бизнес-платформа XXI века.– М.: «Книга», 2011.
6. Справка по PowerPivot. – www.microsoft.com.
7. Функции PowerPivot. – www.microsoft.com./ru-ru/library.
8. Компоненты и средства PowerPivot. – www.microsoft.com/ru-ru/library.