Сравнение существующих продуктов для анализа больших данных

А.В. Елисеев, Н.А. Трушин, студенты;
Научный руководитель И.В. Горбунов, с.н.с каф. КИБЭВС,
г. Томск, ТУСУР, artyom.rus.96@gmail.com
Проект ГПО КИБЭВС-1519 - Моделирование систем защиты информации

Введение. Сегодня в информационном пространстве происходят сильные изменения, связанные с ростом объемов информации, генерируемых различными источниками. К таким источникам можно отнести, например, бизнес, государственные и научные исследования. От анализа простых данных мы переходим к анализу больших данных (Big Data). Согласно Vouchercloud, в 1992 году в день генерировалось 100 гигабайт информации в день, в 2013 году данная цифра уже составляла 28,875 гигабайт в секунду, а к 2050 году может достигнуть 50,000 гигабайт в секунду [1]. Более того, эти данные имеют тенденцию становиться все более взаимосвязанными, допустим, если взять социальные сети (Facebook, Vk, Twitter) – их пользователи так или иначе связаны с друг другом. Увеличение объемов и многогранность информации оказывает большое влияние на методы обработки и интерпретации новых знаний, а так как значительная часть информации поступает из всемирной сети и хранится в ней же, то наиболее важная задача на сегодняшний день – это определиться, в каком русле развивать бушующие технологии, чтобы анализировать Big Data. Для оперативного анализа больших объемов информации требуется отлаженное взаимодействие между потоками данных и вычислительной техникой. Сегодня существует множество различных сервисов, направленных на осуществление анализа больших данных, сравним сервисы, содержащие средства для анализа, среди них Microsoft SQL Server 2016, PENTAHO OLAP, IBM Cognos TM1 Architect.

Основная часть. Одним из продуктов, предоставляющим возможность для анализа больших данных, является Microsoft SQL Server, в новой версии которого появилось много функций, упрощающих и ускоряющих работу с большими данными. Одной из такой функции является добавление R Services в Microsoft SQL Server 2016. R это функциональный язык для обработки данных и их графического представления. Причиной данного нововведения стал тот факт, что количество анализируемых данных постоянно растет и выходит за пределы оперативной памяти. В SQL Server’е версии 2016 года, это ограничение можно обойти за счет встроенных технологий. Совсем недавно Microsoft приобрела компанию Revolution Analytics, которая как раз и занимается разработкой продуктов на языке R. Таким продуктом является R Server, для которого существуют пакеты расширений и свои версии большинства R функций, которые позволяют обрабатывать данные многопоточно, подгружая в память необходимые куски по мере необходимости. В SQL Server были добавлены все данные технологии, что позволяет выполнять код на языке R, дополнять функциональность с помощью пакетов расширений.

В SQL Server 2016 предусматривается два вида работы с R. Первый, подразумевает что приложение выполняет хранимые процедуры на SQL Server, которые могут содержать код на языке R и возвращать результат в виде таблицы, графиков, предсказаний. Второй, когда аналитик может с помощью специального API посылать команды на SQL Server, чтобы все вычисления выполнились там, а вернулись только результаты. Таким образом он может избежать долгого процесса загрузки данных на свою машину. Подводя небольшой итог, можно сказать что R Services позволяют значительно расширить функционал в области анализа больших данных [2].

Как говорилось ранее источники данных стали более разнообразными, и большинство компании сталкиваются со следующей проблемой: им приходится иметь дело с реляционными наборами данных в SQL Server и нереляционными в HDFS (Hadoop Distributed File System). Если возникает ситуация, когда необходимо совместить анализ полуструктурированных и структурированных данных, то сначала нужно будет скопировать их из одной среды в другую, а данная процедура занимает много времени и сил. В SQL Server было представлено решение

PolyBase, которое снимает проблемы разрозненности реляционных и полуструктурированных данных. Использование PolyBase и T-SQL позволяет пользователям отправлять запросы к данным HDFS так, как будто они хранятся на локальном SQL Server, что открывает массу новых возможностей для анализа [3].

Однако есть и обратная сторона медали, Microsoft SQL Server доступен только на платформе Windows и Linux в отличии от продуктов конкурентов. Поддержка сравниваемых продуктов по списку доступных операционных систем представлена в таблице 1.

Таблица №1 - Операционные системы

OLAP сервер	**Window**	**Linux**	**Unix**	**z/OS**
SQL Server 2016	Да	Да	Нет	Нет
Cognos TM1	Да	Да	Да	Да
PENTAHO OLAP	Да	Да	Да	Да

Другим продуктов, предоставляющих возможности для анализа больших данных, является PENTAHO OLAP. Данная система состоит из четырех слоев, которые работают скрыто от конечного пользователя в недрах центра обработки данных. На деле это выглядит следующим образом: уровень представлений, уровень размерностей, уровень звезда и слой хранения.

Слой представлений определяет: что пользователь видит на экране своего монитора и как пользователь может взаимодействовать с системой. Существует множество способов представления многомерных массивов данных, в том числе сводные таблицы, гистограммы, а также расширенные средства визуализации, такие как интерактивные карты и динамические графики. Все эти формы представления имеют, в общем, многомерную структуру, в которой уровень представления задает запрос, а сервер OLAP возвращает ответ. Вторым уровнем является слой размерностей. Задача данного слоя состоит в проверке и выполнении запросов многомерных выражений. Запрос оценивается в несколько этапов. Сначала вычисляются оси, затем значения ячеек в пределах данных осей. Для большей эффективности, слой размерностей посылает запросы на уровне агрегации в пакетном режиме. Трансформатор запросов позволяет приложению манипулировать существующими запросами, а не строить оператор MDX с нуля для каждого запроса. Метаданные описывают размерную модель и то, как она отображается на реляционной модели. Третий уровень – слой звезда, который отвечает за поддержание совокупного кэша. Последний уровень — это слой хранения РСУБД. Он несет ответственность за предоставление агрегированных данных ячеек и членов из таблиц измерений [4]. Сравнение доступных моделей хранения данных представлено в таблице 2.

Таблица №2 - Модель хранения данных

OLAP сервер	**MOLAP**	**ROLAP**	**HOLAP**
SQL Server 2016	Да	Да	Да
Cognos TM1	Да	Нет	Нет
PENTAHO OLAP	Нет	Да	Нет

В отличие от Microsoft SQL Server 2016, PENTAHO OLAP поддерживает только модель хранения данных ROLAP (Relational OLAP), то есть сервер хранит свои данные в реляционной базе данных. Каждая строка в таблице фактов имеет столбец для каждого измерения и меры. Необходимо хранить три вида данных: данные таблицы фактов (транзакционные записи), агрегаты и размерности. Данный выбор аргументируется тем, что в MOLAP (Multidimensional OLAP) фактические данные хранятся в многомерном формате, но, если существует больше нескольких измерений и данные разбросаны, многомерный формат не подходит. Что касается HOLAP (Hybrid OLAP), то он решает данную проблему, оставляя данные в реляционной базе данных, но сохраняет агрегаты в многомерном формате. Предварительно вычисленные

агрегаты необходимы для больших наборов данных, в противном случае некоторые запросы нельзя удовлетворить, не прочитав все содержимое таблицы фактов. В MOLAP агрегаты разбиваются на страницы и сохраняются на диске. В ROLAP агрегаты хранятся в таблицах [5].

Еще одним отличием PENTAHO OLAP является тот факт, что здесь используется MDX (multi-dimensional expressions), который похож на язык SQL запросов, но отличается структурой построения запроса. Это обуславливается тем, что зачастую запросы MDX выполняются быстрее и более четко выражают цель запроса [6].

Следующий продукт от компании IBM, Cognos TM1 – высокопроизводительная платформа для создания систем бюджетного управления, основной особенностью которой является ее ядро, которое позволяет проводить быстрый анализ данных при использовании сложных моделей и больших объемов информации, обрабатывать многомерные данные в оперативной памяти. Данные функции подходят для использования в бизнес-процессах и создания высокопроизводительных систем бюджетирования. Однако стоит отметить, что такая же функции не являются революционными так, например, технология in-memory OLTP присутствовала в SQL Server 2014 [7].

Заключение. Большие данные иметь самое разное происхождение, однако в любом случае извлечение знаний из них сопряжено с многими трудностями. В данной статье приведен обзор современных продуктов, которые позволяют справляться с данными трудностями, осуществляя анализ больших данных с большой скоростью, а также их сравнение. Подводя итог можно сказать, что в большинстве своем данные продукты похожи и выбор зависит от предпочтения аналитиков и поставленных задач, и доступного бюджета.

ЛИТЕРАТУРА

1. EVERY DAY BIG DATA STATISTICS – 2.5 QUINTILLION BYTES OF DATA CREATED DAILY [Электронный ресурс]. – Режим доступа: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/ (дата обращения: 21.11.2016).

2. SQL Server 2016: R Services. [Электронный ресурс]. – Режим доступа: http://olontsev.ru/2016/06/sql-server-2016-r-services-part-1-overview/ (дата обращения: 21.11.2016).

3. О новых функциях SQL Server 2016. [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/company/microsoft/blog/302770/ (дата обращения: 21.11.2016).

4. Layers of a Mondrian system. [Электронный ресурс]. – Режим доступа: http://mondrian.pentaho.com/documentation/architecture.php (дата обращения: 21.11.2016).

5. Storage and aggregation strategies. [Электронный ресурс]. – Режим доступа: http://mondrian.pentaho.com/documentation/architecture.php (дата обращения: 21.11.2016).

6. MDX Specification. [Электронный ресурс]. – Режим доступа: http://mondrian.pentaho.com/documentation/mdx.php (дата обращения: 21.11.2016).

7. IBM Cognos TM1. [Электронный ресурс]. – Режим доступа: http://www.bipartner.ru/software/cognostm1.html (дата обращения: 21.11.2016).