УДК 004.6

СРАВНИТЕЛЬНЫЙ АНАЛИЗ BIG DATA И ТРАДИЦИОННОГО МЕТОДА РАБОТЫ С ДАННЫМИ

Шайхутдинов А.М.

ФГАОУ ВО Казанский (Приволжский) федеральный университет,
420008, г. Казань, ул. Кремлевская, д.18

e-mail: amir_sh00@mail.ru

поступила в редакцию 7 февраля 2017 года

Аннотация

В данной статье представлен сравнительный анализ технологии Big Data и традиционного метода работы с данными, проанализированы преимущества, предоставляемые технологиями Big Data, и обоснована необходимость их применения в деятельности коммерческих организаций. На нынешний день эта технология является одним из главных драйверов развития информационных технологий. Big Data объединяет в себе данные из различных источников и разной степени структурированности. На основе проведенного анализа был сделан вывод о том, что Big Data представляет из себя технологию извлечения информации из огромного массива данных в максимально короткие сроки с целью нахождения полезной информации и принятия эффективных управленческих решений.

Ключевые слова: Big Data, большие данные, Map-Reduce, неструктурированная информация.

Введение

Объемы информации постоянно увеличиваются, но в последнее время на IT-рынке стала обширно дискуссироваться концепция Big Data (большие данные), возникновение которой было связано с пониманием необходимости в неких качественных изменениях в подходах к хранению и использованию информации, объем которой становится с каждым днем все больше.

Основная часть

Сегодня традиционный вариант простого увеличения мощностей и ресурсов уже не работает, многие компании отмечают постоянный рост издержек на хранение, несмотря на постоянное уменьшение удельной стоимости хранения данных. Специалисты отмечают, что взрывной рост размера информации не является результатом роста числа деловых операций и, вполне возможно, объясняется неуправляемыми процессами репликации данных. Даже вендоры устройств хранения все чаще говорят о том, что круг задач управления информацией на данный момент быстро сдвигается от вопросов физического хранения данных к их использованию, что хранение данных - это средство для того, чтоб ими можно было пользоваться в подходящий момент. При всем этом тема Big Data связана напрямую с иной, уже давно обсуждаемой глобальной IT-тенденцией – с переходом к широкому внедрению облачных технологий [1].

Большие данные сосредоточены вокруг хранилищ данных, объем которых намного превышает несколько терабайт.

Таблица 1. – Классификация объемов данных.

Большие наборы данных	Огромные наборы данных	Big Data (Extremely Big Data)
1 гб - 100 гб	1 тб – 10 тб	10 тб – 100 тб (1 пб – 10 пб)

Компания Forrester определяет термин Big Data как аппаратное и программное обеспечение, которое организует, объединяет, анализирует и управляет данными, характеризующимися «4 V»: объем (Volume), разнообразие (Variety), изменчивость (Variability) и скорость (Velocity) - рисунок 1.

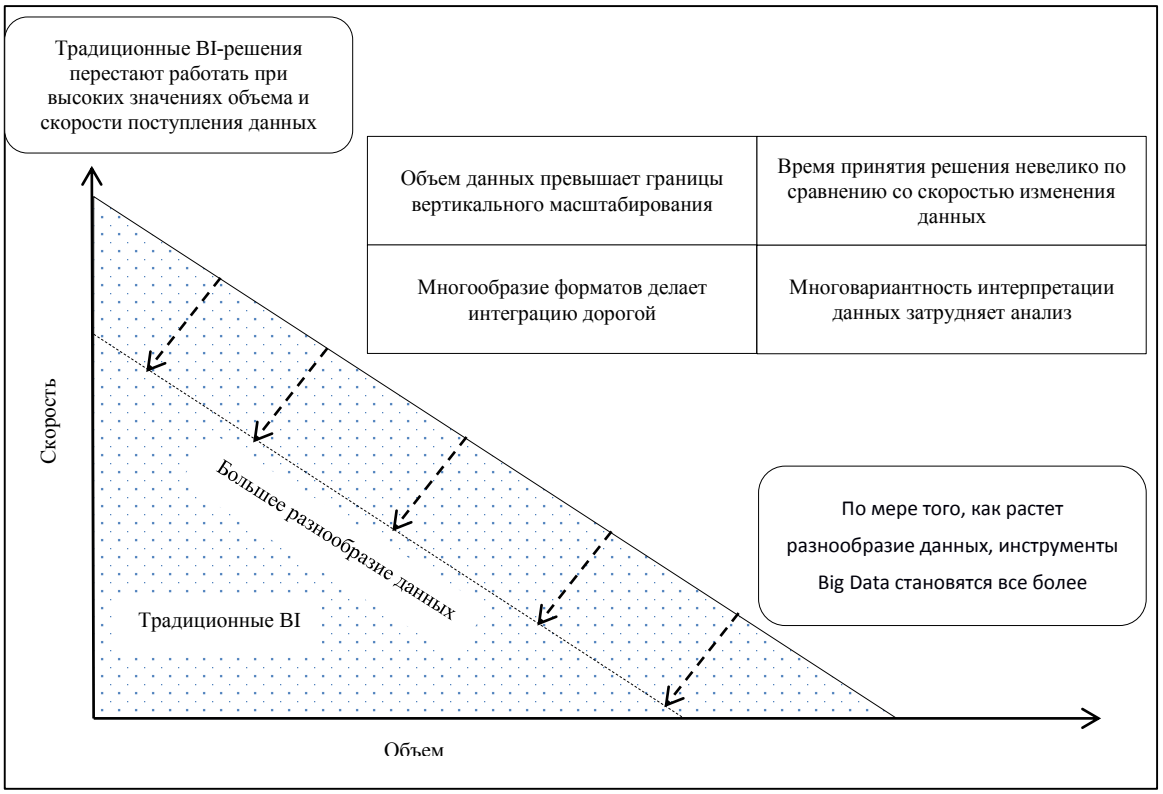

Рисунок 1. – Границы применения традиционных BI- и Big Data-технологий

Способность приложения обрабатывать огромные массивы данных, которые поступают из различных источников в разных форматах, выступает основным критерием для того, чтобы отнести его к технологии Big Data. Обычно приложения Big Data включает в себя данные из различных источников (из внутренних и внешних источников) и разной степени структурированности (слабоструктурированные, структурированные и неструктурированные). Решение многих задач требуют совместной обработки данных разных форматов - табличных данных в СУБД, иерархических данных, текстовых документов, видео, изображения, аудиофайлов и т.д.

Есть отрасли, где сбор и накопление данных осуществляется чрезвычайно интенсивно. Если рассматривать производственную сферу, к примеру, электростанции, то здесь каждую минуту либо даже каждую секунду генерируется непрерывный поток данных. Кроме того, за последние годы, внедряются технологии «smart grid», которые позволяют измерять каждую минуту либо каждую секунду потребление электроэнергии. Extremely Big Data – так классифицируют накопленные данные в приложениях, в которых данные должны храниться годами.

Растет и количество приложений Big Data среди государственных и коммерческих сегментов, где размер данных в хранилищах, может составлять больше 100 тб либо пб. Современные технологии дают возможность «отслеживать» людей и их поведение разными методами. К примеру, когда люди используют интернет, делают покупки в Интернет-магазинах, перемещаются с включенными телефонами – они оставляют след своих действий, что приводит к накоплению новой информации.

Разные методы связи, обычные телефонные звонки, загрузка информации через веб-сайты социальных сетей, таких как Вконтакте, либо обмен видео на таких веб-сайтах, как YouTube, каждый день генерируют большое количество новых данных.

Теперь, следовательно, возникают следующие вопросы, которые связаны с Big Data:

1. Где хранить большие данные и как ими управлять?
2. Как организовать неструктурированную информацию?
3. Как анализировать неструктурированную информацию?

Размер данных в сотни терабайт либо петабайт не дает возможности просто хранить и управлять ими при помощи обычных реляционных баз данных. Большая часть данных Big Data являются неструктурированными, следовательно, значимость второго вопроса не вызывает сомнений. Другими словами, третий вопрос можно сформулировать следующим образом: как на базе Big Data составлять обычные отчеты, строить и использовать углубленные прогностические модели?

Программное обеспечение, оборудование, также сервисные услуги вместе образуют комплексные платформы для хранения и анализа данных.

Таблица 2. – Big Data

<table>
<tr><td rowspan="6">Big Data</td><td rowspan="3">Программное обеспечение</td><td>ПО для организации и управления данными, обрабатывающее и готовящее все виды неструктурированных и структурированных данных для анализа (отвечают за извлечение, нормализацию, очистку и интеграцию данных)</td><td>NoSQL</td></tr>
<tr><td>ПО для аналитической обработки Big Data и выявления закономерностей</td><td>Приложения для онлайновой обработки либо оффлайновой обработки по запросу, средства определения закономерностей в данных, приложения для разных вертикальных областей</td></tr>
<tr><td>ПО для поддержки принятия решения и автоматизации его исполнения</td><td>Приложения для оптимизации торговли ценными бумагами, определения случаев мошенничества, сегментации клиентов, оптимизации цен на авиабилеты, прогнозирования погоды.</td></tr>
<tr><td rowspan="2">Технологическое оборудование</td><td>Серверы</td><td>Хранилища данных</td></tr>
<tr><td>Инфраструктурное оборудование</td><td>Средства ускорения платформ, источники бесперебойного питания, комплекты серверных консолей и др.</td></tr>
<tr><td colspan="2">Сервисные услуги</td><td>Услуги по построению архитектуры системы БД, оптимизации и обустройству инфраструктуры и обеспечению безопасности хранения данных [2].</td></tr>
</table>

Big Data обычно организуются и хранятся в распределенных файловых системах (DFS). Информацию хранят на нескольких жестких дисках, на обычных компьютерах. Так называемая «карта» (map) позволяет отслеживать, где (на каком ПК и/либо диске) хранится определенная часть информации. Для того, чтобы обеспечить отказоустойчивость и надежность, каждую часть информации обычно сохраняют несколько раз, к примеру - три раза. При помощи открытых программных продуктов для управления DFS (к примеру, Hadoop) и обычного оборудования, сравнимо просто можно реализовать в большом масштабе надежные хранилища данных [3].

Большинство собранной информации в DFS включает в себя неструктурированные данные, такие как изображения, фотографии, текст либо видео.

Обработка этих больших массивов данных может состоять из обычных операций (к примеру, обычные подсчеты, и т.д.), а может состоять из сложных, и в данном случае обработка данных будет требовать более сложные алгоритмы, которые должны быть специально разработаны для действенной работы на DFS [4].

Что касается анализа Big Data, то это большая проблема, поскольку связана с анализом неструктурированных данных.

Таблица 3. – Сравнительный анализ Big Data и традиционного подхода работы с данными

Критерии сравнения	Big Data	Традиционный подход
Возможности	- выявление скрытых зависимостей и поиск новых вопросов и ответов на основе анализа всего объема разнородных данных - прогноз - анализ текущей ситуации - анализ данных не только из внутренних, но и из внешних источников	- извлечение из «сырых» данных полезной информации и ее запись в форме, приемлемой для использования - анализ текущей ситуации
Объем информации	От петабайт (10^15 байт) до эксабайт (10^18 байт)	От гигабайт (10^9 байт) до терабайт (10^12 байт)
Способ хранения	Децентрализованный	Централизованный
Структурированность данных	Полуструктурирована, неструктурирована, структурирована	Структурирована
Модель хранения и обработки данных	Горизонтальная модель	Вертикальная модель
Взаимосвязь данных	Слабая	Сильная
Методы	- Извлечение данных - Краудсорсинг – сбор данных от большого числа источников - Консолидация данных - Визуализация - Машинное обучение - Нейронные сети - Анализ сетей - Предиктивное моделирование - Обработка сигналов и анализ временных рядов - Пространственный анализ - A/B тестирование - методы класса Data Mining: обучение ассоциативным правилам, классификация, кластерный анализ, регрессионный анализ	Обычные традиционные методы, схемы, способы анализа данных
Подходы обработки данных	NoSQL, MapReduce, Hadoop, R	SQL
Вид хранилища	Data lake (озеро данных) — хранилище больших данных в необработанном виде (распределенное хранилище, которое масштабируется по мере необходимости)	Традиционная реляционная БД
Используемые методы визуализации	Mindmap (карта мыслей), Displaying connections (Отображение связей), Treemaps, Кластерограмма, Рейтинговая диаграмма текста и т.д.	Таблицы, гистограммы, круговые диаграммы и т.д.
Решение проблем	На базе данных и моделей данных	На базе данных
Стоимость хранения данных	Низкая	Высокая
Масштабирование	+	-

Когда осуществляется анализ сотни терабайт либо петабайт данных, не представляется возможным извлечение данных в какое-либо другое место для анализа, поскольку занимает очень много времени и просит очень много трафика. Вместо этого, при анализе Big Data применяется алгоритм Map-Reduce, представляющий из себя модель для распределенных вычислений. В данном случае не данные передаются на обработку программе, а программа - данным. Таким образом, запрос представляет собой отдельную программу. Например, для того, чтобы вычислить итоговую сумму, алгоритм будет параллельно производить вычисления промежуточных сумм в каждом из узлов DFS, и потом суммировать эти промежуточные значения [5].

Хочется еще раз отметить, что к Big Data относится обработка конкретно большого размера информации, который проблемно обрабатывать традиционными методами. Приведем в таблице 3. сравнение Big Data и существующего традиционного подхода работы с данными.

В результате проведения сравнительного анализа, можно сделать вывод о том, что Big Data представляет из себя технологию извлечения информации из огромного массива данных в максимально короткие сроки с целью нахождения полезной информации и принятия эффективных управленческих решений, а традиционный подход – ПО с простым и интуитивно понятным интерфейсом, позволяющее проводить несложный анализ структурированных данных. Важно отметить, что когда мы имеем дело с Big Data, существующее «озеро» данных позволяет хранить данные из разных источников и разных форматов. Это обходится значительно дешевле традиционных хранилищ, в которые помещаются только структурированные данные.

Заключение

Проведенный анализ дал возможность сделать следующие выводы:

1. Big Data представляет из себя технологию извлечения информации из огромного массива данных в максимально короткие сроки с целью нахождения полезной информации и принятия эффективных управленческих решений, а традиционный подход – ПО с простым и интуитивно понятным интерфейсом, позволяющее проводить несложный анализ структурированных данных.

2. Big Data объединяет в себе данные из различных источников и разной степени структурированности. Способность приложения обрабатывать огромные массивы данных, которые поступают из различных источников в разных форматах, выступает основным критерием отнесения его к технологии Big Data.

Список литературы

1) Интернет-ресурс: Не пора ли вплотную заняться большими данными. www.sybase.ru/company/press/ne_pora_li_vplotnuyu_zanyatsya_bolshimi_dannymi (Дата обращения: 26.01.2017).

2) Интернет-ресурс: Аналитический обзор рынка Big Data. www.habrahabr.ru/company/moex/blog/256747 (Дата обращения: 27.01.2017).

3) Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим: И. Гайдюк, М.: Манн, Иванов и Фербер, 2014. 221 с.

4) Интернет-ресурс: Революция Big Data: Как извлечь необходимую информацию из «Больших Данных»? www.statsoft.ru/products/Enterprise/big-data.php (Дата обращения: 01.02.2017).

5) Интернет-ресурс: Big Data: проблема, технология, рынок. www.compress.ru/article.aspx?id=22725#06 (Дата обращения: 01.02.2017).