BIG DATA

студент, Яковлев Владимир Сергеевич
научный руководитель: Агеев Евгений Юрьевич
Томский Государственный университет Систем Управления и Радиоэлектроники;

В сущности, понятие больших данных подразумевает работу с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности работы, создания новых продуктов и повышения конкурентоспособности.

Почему вообще данные становятся большими?

Источников больших данных в современном мире великое множество. В их качестве могут выступать непрерывно поступающие данные с измерительных устройств, события от радиочастотных идентификаторов, потоки сообщений из социальных сетей, метеорологические данные, данные дистанционного зондирования земли, потоки данных о местонахождении абонентов сетей сотовой связи, устройств аудио и видеорегистрации. Собственно, массовое распространение перечисленных выше технологий и принципиально новых моделей использования различно рода устройств и интернет-сервисов послужило отправной точкой для проникновения больших данных едва ли не во все сферы деятельности человека. В первую очередь, научно-исследовательскую деятельность, коммерческий сектор и государственное управление.

Примеры:

-Каждый месяц в сети Facebook выкладывается в открытый доступ 30 млрд новых источников информации.

-Ежегодно объемы хранимой информации вырастают на 40%, в то время как глобальные затраты на ИТ растут всего на 5%.

-По состоянию на апрель 2011 года в библиотеке Конгресса США хранилось 235 терабайт данных.

- Примеру, датчики, установленные на авиадвигателе, генерируют около 10 Тб за полчаса. Примерно такие же потоки характерны для буровых установок и нефтеперерабатывающих комплексов. [1]

Задачи, связанные с Big Data.

Существуют три типа задач связанных с Big Data:

1. Хранение и управление

Объем данных в сотни терабайт или петабайт не позволяет легко хранить и управлять ими с помощью традиционных реляционных баз данных.

2. Неструктурированная информация

Большинство всех данных Big Data являются неструктурированными. Т.е. как можно организовать текст, видео, изображения, и т.д.

3. Анализ Big Data

Как анализировать неструктурированную информацию? Как на основе Big Data составлять простые отчеты, строить и внедрять углубленные прогностические модели?

Хранение и управление Big Data

Big Data обычно хранятся и организуются в распределенных файловых системах.

В общих чертах, информация хранится на нескольких (иногда тысячах) жестких дисках, на стандартных компьютерах.

Так называемая «карта» (map) отслеживает, где (на каком компьютере и/или диске) хранится конкретная часть информации.

Для обеспечения отказоустойчивости и надежности, каждую часть информации обычно сохраняют несколько раз, например – трижды.

С помощью стандартного оборудования и открытых программных средств для управления этой распределенной файловой системой (например, Hadoop HDFS (Hadoop Distributed File System), сравнительно легко можно реализовать надежные хранилища данных в масштабе петабайт.

Неструктурированная информация

Большая часть собранной информации в распределенной файловой системе состоит из неструктурированных данных, таких как текст, изображения, фотографии или видео.

Это имеет свои преимущества и недостатки.

Преимущество состоит в том, что возможность хранения больших данных позволяет сохранять “все данные”, не беспокоясь о том, какая часть данных актуальна для последующего анализа и принятия решения.

Недостатком является то, что в таких случаях для извлечения полезной информации требуется последующая обработка этих огромных массивов данных.

Анализ Big Data

Это действительно большая проблема, связанная с анализом неструктурированных данных Big Data: как анализировать их с пользой. О данном вопросе написано гораздо меньше, чем о хранении данных и технологиях управления Big Data.

Технологии Анализа:

Map-Reduce. При анализе сотни терабайт или петабайт данных, не представляется возможным извлечь данные в какое-либо другое место для анализа

Процесс переноса данных по каналам на отдельный сервер или сервера (для параллельной обработки) займет слишком много времени и требует слишком большого трафика.

Вместо этого, аналитические вычисления должны быть выполнены физически близко к месту, где хранятся данные.

Алгоритм Map-Reduce представляет собой модель для распределенных вычислений. Принцип его работы заключается в следующем: происходит распределение входных данных на рабочие узлы распределенной файловой системы для предварительной обработки) и, затем, свертка уже предварительно обработанных данных

Таким образом, скажем, для вычисления итоговой суммы, алгоритм будет параллельно вычислять промежуточные суммы в каждом из узлов распределенной файловой системы, и затем суммировать эти промежуточные значения.

В Интернете доступно огромное количество информации о том, каким образом можно выполнять различные вычисления с помощью модели map-reduce, в том числе и для прогностической аналитики. [2]

R (язык программирования). *R* — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом

Особенности:

R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения. Ещё одной особенностью R являются графические возможности, заключающиеся в возможности создания качественной графики, которая может включать математические символы. [3]

Hadoop — проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и фреймворка для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Используется для реализации поисковых и контекстных механизмов многих высоконагруженных веб-сайтов, в том числе, для Yahoo! и Facebook. [4]

Итак, создание и поддержка хранилищ объемом в терабайт, петабайт и более стало возможным благодаря технологиям распределенных файловых систем.

В распределенных системах, вместо хранения данных в одной файловой системе, данные сохраняются и индексируются на нескольких (и даже тысячах) жестких дисках и серверах. Создается также «карта» (map), где содержится информация о том, где именно находятся те или иные данные.

Hadoop является одной из самых известных систем, использующих данный подход.

Чтобы обработать данные в распределенной файловой системе, необходимо проводить низкоуровневые вычисления, такие как суммирование, агрегирование и др. в месте их физического размещения в распределенной файловой системе. Создать карту (map) проведенных вычислительных алгоритмов и отслеживать локальные результаты. Затем, аккумулировать результаты (reduced). Данный подход и шаблон проведения вычислительных алгоритмов получил название Map-Reduce.

На практике, анализ Big Data редко заключается в том, чтобы вычислить статистические итоги по всем данным. Вместо этого значимость Big Data заключается в возможности разделения данных на «микро-сегменты» и с помощью методов data mining и прогностического моделирования построить большое число моделей для небольших групп наблюдений. [2]

Проекты Big Data в России и мире

В большинстве разговоров о Big Data, сосредоточенных на решении бизнес-задач, легко потерять из виду тот факт, что конечной возможностью big data является решение по-настоящему глобальных проблем, которые влияют на все мировое сообщество. Одной из них является голод.

Одной из компаний, ищущих пути решения проблем питания с привлечением больших данных, является Intel. В настоящее время компания имеет два научно-исследовательских проекта, направленных на использование больших объемами данных, чтобы решать продовольственные и сельскохозяйственные задачи в мире. Один проект исследует орошение, а другой — расположение снегов в горах Сьерра-Невада.

Проект с орошением называется Precision Farming — «Точное земледелие». Intel работает с университетом Калифорнии в Дэвисе, совершенствуя методы ирригации путем размещения датчиков в сельскохозяйственных культурах для мониторинга почвы и уровня влажности воздуха. Анализа различных датчиков позволит связать подачу воды с ее

реальной необходимостью для растений и почвы. Этот подход может уменьшить количество воды, используемой для орошения, на целых 50 процентов.

Второй проект использует данные из лаборатории, которая контролирует снежный покров в горной цепи Сьерра-Невада, соотнося его с размерами водоснабжения Калифорнии. Цель состоит в том, чтобы создать базу данных изображений, которая позволит правительству штата и фермерам предсказывать условия засухи и соответствующим образом планировать свои действия. [5]

Проекты и сервисы которые запущены в России не решают каких-то глобальных проблем. Они направлены на развитие бизнеса и анализ поведения клиентов. Рынок Big Data в России находится еще в стадии формирования.

- *Crosss*

Это персонализованный мерчандайзинг для интернет-магазинов. Сервис собирает информацию об интересах пользователей, анализирует ее и затем помогает магазинам предугадывать желания покупателей. Crosss может перестроить весь контент сайта лично для каждого пользователя (на основе его поведения), правильно выстроить выдачу товаров в каталоге, создавать персонализированные и таргетированные email-расссылки. Сервис запустился в 2012 году и с того момента услугами пользуются магазины в 9 странах, включая Великобританию.

- АлгоМост

Решает всевозможные проблемы клиентов с помощью Больших Данных. Область, в которой лежит задача, может быть, в принципе, любой, — будь то транспортная логистика, ритейл, страхование или банкинг. Позволяет искать «то, не знаю что» и интересен тем, кто нацелен на открытия.

- ZOOM TV

Новый подход к просмотру телевизора предлагает ZOOM TV. Сервис решает проблему всех любителей провести вечер в домашней обстановке — из сотен ТВ-передач выбрать что-то реально интересное. ZOOM TV берет на себя листание каналов и выдает самое интересное на основе предпочтений зрителя. Из других полезных функций — можно вернуть передачу на самое начало, перемотать скучные куски или поставить выпуск новостей на паузу.

- Opiner

Сервис открывает глаза SMM-отделу компании на то, в каком направлении им работать и как на самом деле относятся к их бренду на просторах интернета. Opiner предлагает решение для мониторинга всех социальный сетей, упоминаний о компании или физическом лице, их анализе и выявлении трендов. Так можно моментально узнать о негативных комментариях и тут же устранить проблемы, или оценить, чего от вас ждут клиенты. А еще можно узнать все о своих конкурентах и их репутации в сети — где они побеждают, где проседают и какова лояльность их клиентов. [6]

Список используемых источников:

1. *ФОРС. Интернет-журнал, № 1 | Большие данные (Big Data).* [Электронный ресурс]. Режим доступа: *www.fors.ru/upload/magazine/01/html_texts/total_big_date(2).html* (дата обращения 17.09.15);
2. *Big Data: аналитика и решения.* [Электронный ресурс]. Режим доступа: *http://statsoft.ru/products/Enterprise/big-data.php* (дата обращения 17.09.15).
3. *R(язык программирования)—Википедия* [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/R_(язык_программирования) (дата обращения 18.09.15).

4. *Hadoop—Википедия* [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/Hadoop (дата обращения 18.09.15).
5. *Как решить проблемы мирового голода с помощью big data* [Электронный ресурс]. Режим доступа: http://data247.ru/2014/06/10/kak-reshit-problemy-mirovogo-goloda-s-pomoshhyu-big-data/ (дата обращения 31.10.15).
6. *Кто делает Big Data в России|Rusbase [*Электронный ресурс]. Режим доступа: http://rusbase.com/news/10-best-big-data-startups-in-rus/ (дата обращения 1.11.15).