Университет информационных технологий, механики и оптики

Экология

Реферат на тему

Использование BigData для решения проблем экологии

Выполнил: Мазунин К. Ю.
Группа P3217
Преподаватель: Крылова Юлия Викторовна,
кандидат географических наук,
доцент кафедры экологии и
техносферной безопасности

2017

Введение

BigData у всех на слуху уже не первый год. Хотя сама технология проявилась ещё в 2000 годах. О Big Data говорят, потому что эта методология работает. Компании принимают стратегические решения на основании результатов анализа при помощи данной технологий. BigData находит своё применение не только в коммерческом, но и в научном секторе.

Чтобы узнать, как же возможно применить BigData для решения проблем экологии были поставлены следующий цели:

- Изучение когда применимы технологии BigData
- Изучение методов обработки данных при помощи BigData
- Поиск примеров применения BigData на практике

Оглавление

1. BigData

1.1 Определение

BigData - серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.[1]

Типичный пример больших данных — это информация, поступающая с различных физических экспериментальных установок — например, с Большого адронного коллайдера, который производит огромное количество данных и делает это постоянно. Установка непрерывно выдает большие объемы данных, а ученые с их помощью решают параллельно множество задач.

Появление больших данных в публичном пространстве было связано с тем, что эти данные затронули практически всех людей, а не только научное сообщество, где подобные задачи решаются давно. В публичную сферу технологии Big Data вышли, когда речь стала идти о вполне конкретном числе — числе жителей планеты. 7 миллиардов, собирающихся в социальных сетях и других проектах, которые агрегируют людей. YouTube, Facebook, ВКонтакте, где количество людей измеряется миллиардами, а количество операций, которые они совершают одновременно, огромно. Поток данных в этом случае — это пользовательские действия. Например, данные того же хостинга YouTube, которые переливаются по сети в обе стороны. Под обработкой понимается не только интерпретация, но и возможность правильно обработать каждое из этих действий, то есть поместить его в нужное место и сделать так, чтобы эти данные каждому пользователю были доступны быстро, поскольку социальные сети не терпят ожидания.[2]

В нашей жизни все больше аппаратных средств и программ начинают генерировать большое количество данных — например, «интернет вещей».

Вещи уже сейчас генерируют огромные потоки информации. Полицейская система «Поток» отправляет со всех камер информацию и позволяет находить машины по этим данным. Все больше входят в моду фитнес-браслеты, GPS-трекеры и другие вещи, обслуживающие задачи человека и бизнеса.[2]

1.2 Большие данные

Для того, чтобы понять, что приходится работать с большим данными выделяют три признака[3]:

- **Volume**: действительно большие (хотя размер зависит от доступных ресурсов для их обработки).
- **Variety**: слабо структурированные и разнородные.
- **Velocity**: обрабатывать надо очень быстро (причем и результаты часто нужны оперативно, если речь об онлайновых сервисах).

2. Работа с BigData

2.1 Принципы

Исходя из определения **Big Data**, можно сформулировать основные принципы работы с такими данными[4]:

1. **Горизонтальная масштабируемость**. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжило работать.

2. **Отказоустойчивость**. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин. Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий.

3. **Локальность данных.** В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.

2.2 MapReduce

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах.[4]

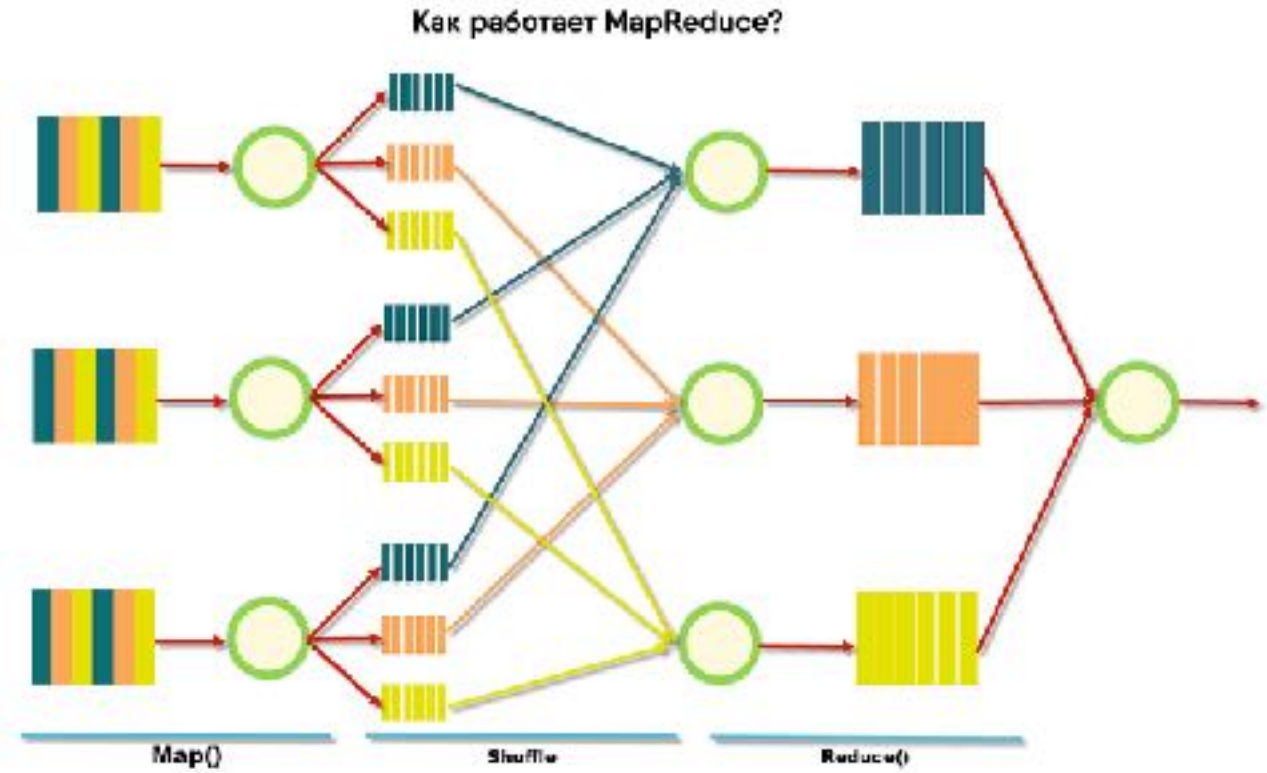

Рис. 1

На Рис.1 представлена схема работы MapReduce.

MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 стадии[4]:

1. **Стадия Map**. На этой стадии данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи.

Функция map() примененная к одной входной записи и выдаёт множество пар ключ-значение. Множество – т.е. может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение. Что будет находится в ключе и в значении – решать пользователю, но ключ –

очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce.
2. **Стадия Shuffle**. Проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти корзины послужат входом для reduce.
3. **Стадия Reduce**. Каждая «корзина» со значениями, сформированная на стадии shuffle, попадает на вход функции reduce().

Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины». Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

Особенности MapReduce:
1) Все запуски функции **map** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.
2) Все запуски функции **reduce** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.
3) **Shuffle** внутри себя представляет параллельную сортировку, поэтому также может работать на разных машинах кластера. Пункты 1-3 позволяют выполнить принцип горизонтальной масштабируемости.
4) Функция **map**, как правило, применяется на той же машине, на которой хранятся данные – это позволяет снизить передачу данных по сети (принцип локальности данных).
5) **MapReduce** – это всегда полное сканирование данных, никаких индексов нет. Это означает, что MapReduce плохо применим, когда ответ требуется очень быстро.

2.3 Hadoop

Hadoop – это проект с открытым исходным кодом, находящийся под управлением Apache Software Foundation. Hadoop используется для надежных, масштабируемых и распределенных вычислений, но может также применяться и как хранилище файлов общего назначения, способное вместить петабайты данных. Многие компании используют Hadoop в исследовательских и производственных целях[5].

Hadoop состоит из двух ключевых компонентов:
- Распределенная файловая система Hadoop **HDFS**, которая отвечает за хранение данных на кластере Hadoop;
- Система **MapReduce**, предназначенная для вычислений и обработки больших объемов данных на кластере.

На основе этих ключевых компонентов создано несколько подпроектов, таких как Pig, Hive, HBase и т.д.

2.3.1 HDFS

HDFS - это основная система хранения данных, используемая приложениями Hadoop. HDFS многократно копирует блоки данных и распределяет эти копии по вычислительным узлам кластера, тем самым обеспечивая высокую надежность и скорость вычислений[5]:
- Данные распределяются по нескольким машинам во время загрузки.
- HDFS оптимизирована больше для выполнения потоковых считываний файлов, нежели для нерегулярных, произвольных считываний.
- Файлы в системе HDFS пишутся однократно и внесение никаких произвольных записей в файлы не допускается.
- Приложения могут считывать и писать файлы Распределенной файловой системы напрямую через программный интерфейс Java.

Самый простой способ думать про **HDFS** – это представить обычную файловую систему, только больше. Обычная ФС, по большому счёту, состоит из таблицы файловых дескрипторов и области данных. В HDFS вместо таблицы используется специальный сервер – сервер имён (NameNode), а данные разбросаны по серверам данных (DataNode). В остальном отличий не так много: данные разбиты на блоки (обычно по 64Мб или 128Мб), для каждого файла сервер имён хранит его путь, список блоков и их реплик.

2.3.2 MapReduce

MapReduce – это модель программирования и каркас для написания приложений, предназначенных для высокоскоростной обработки больших объемов данных на больших параллельных кластерах вычислительных узлов[5]:
- обеспечивает автоматическое распараллеливание и распределение задач;
- имеет встроенные механизмы сохранения устойчивости и работоспособности при сбое отдельных элементов;
- обеспечивает чистый уровень абстракции для программистов.

Реализует модель обработки данных из главы 2.2

3. Проблемы BigData

3.1 Хранение и управление

Чем больше объем накопленных данных, тем требовательнее система хранения и управления этими данными. Вам придется покупать дорогостоящее оборудование или смирится с недостатками хранения данных в облаке. Вам понадобятся специалисты, способные предусмотреть возможные проблемы при анализе больших объемов данных, которые смогут организовать все нюансы таким образом, чтобы вы реально эффективно использовали данные[6].

3.2 Предвзятость

Предвзятость — еще одна из серьезных проблем в Big data. Довольно легко сделать конкретный вывод, если в вашем распоряжении результаты одного или двух исследований, но если их становится значительно больше, появляется довольно большой простор для маневра, который позволяет изменить общий смысл результатов, изменив представление данных. Поэтому очень важно позаботиться о том, чтобы на результаты исследований не влияло мнение какой-либо из заинтересованных сторон[6].

3.3 Шум

Чем больше у вас данных, тем сложнее выделить именно то, что необходимо вам в текущий момент. Конечно, природа этой проблемы напрямую связана со спецификой big data и вообще data mining, но ее не стоит упускать из виду[6].

3.4 Корректность

Специфика Big data в том, что анализ проводится на основе алгоритма, лишенного свободы действия и не имеющего возможность учесть ряд факторов. Кроме того, высокая сложность алгоритма значительно повышает риск того, что какой-то фактор будет упущен из виду. Представьте, что вам предстоит проехать по загруженной трассе, как вдруг навигатор подсказывает, что есть объезд. Вы направляетесь туда, а оказывается, что это строящаяся дорога.

Стоит отметить, что найденная корреляция не всегда может говорить о реальной взаимосвязи между явлениями: так например, в США была обнаружена корреляция между долей браузера Microsoft на рынке и числом совершенных убийств. К этой проблеме больших данных стоит

отнестись особенно серьезно, так как она ставит под угрозу целесообразность всех решений, принятых на основе анализа собранных данных.

Еще одно проявление этой проблемы в Big Data: если вы знаете алгоритм работы, вы легко можете обмануть систему. В ходе испытаний системы, проверяющей сочинения, студенты начали писать сложные и длинные предложения, так как они заметили, что система использует этот фактор, как один из критериев. В итоге качество работ упало, а оценки поднялись[6].

4. OpenData

OpenData - концепция, отражающая идею о том, что определённые данные должны быть свободно доступны для машиночитаемого использования и дальнейшей републикации без ограничений авторского права, патентов и других механизмов контроля. Освободить данные от ограничений авторского права можно с помощью свободных лицензий, таких как лицензий Creative Commons. Если какой-либо набор данных не является общественным достоянием, либо не связан лицензией, дающей права на свободное повторное использование, то такой набор данных не считается открытым, даже если он выложен в машиночитаемом виде в Интернет[7].

Популярные государственные источники данных:
- data.gov
- science.gov
- data.gov.uk
- data.gov.in

Открытые данные предоставляют возможность любому человеку проводить исследования на их основе. Это применимо и к экологии.

5. Применение в экологии

5.1 Умный город

Все мусорные ведра и контейнеры в городе снабжены датчиком, показывающим уровень их заполненности. Мусоровоз вывозит мусор не по расписанию ежедневно, а по необходимости. Все транспортные средства снабжены датчиками загрязнения воздуха, позволяющими в реальном времени определять потенциально проблемные места города. В мире уже ведется апробация проекта с мусорными контейнерами. Например, в Нью-Йорке, Женеве, Дублине установлены BigBelly Solar — высокотехнологичные урны, самостоятельно прессующие мусор и упаковывающие его. Ориентировочная стоимость одной урны — $4 000. Благодаря использованию урн, в Филадельфии удалось сократить количество мусоросборочных рейсов с 17 до 2.

Например, в Чикаго планировали внедрить датчики измерения температуры, влажности, уровня загрязнённости воздуха, тепла, параметров ветра. Приблизительная стоимость одного такого датчика — $1 000. Установка обойдется в $215-425. Датчики будут использоваться для изучения окружающей среды в городе и для своевременного решения возникающих проблем.

Как мы видим, уже многие мегаполисы активно применяют технологии больших данных для реализации крупных проектов в городской среде. В будущем вся модернизация городской инфраструктуры будет проходить с помощью больших данных, которые помогут учесть все экономические, культурные и социальные потребности горожан и эффективно использовать ресурсы.

5.2 Браконьеры

Из костей индийских тигров готовят снадобье, чрезвычайно популярное у некоторых суеверных китайцев. Добывают запрещенный товар хорошо обученные браконьеры, знакомые с каждым ручейком и каждым камнем в зоне обитания редких животных.
Поймать преступников было чрезвычайно сложно, пока активисты и власти не обратились к современным технологиям. Проанализировав данные за 43 года из 605 районов, ученые смогли определить горячие точки, в которые наиболее вероятно заглянут браконьеры.

5.3 Зелёный город

В Нью-Йорке растет около 2,5 млн деревьев. Жители города очень любят зеленые насаждения, но, как оказалось, без должного ухода те не могут ответить им взаимностью в полной мере: за период с 2009 по 2010 год только в Центральном парке из-за падающих веток пострадали четыре человека. Среди них был и 37-летний разработчик Google.
Впрочем, не бывает безвыходных ситуаций. Проблему ухода, обрезки, удобрения и выбора оптимального времени для корчевания решили с помощью анализа больших данных.

5.4 Большие данные на страже экологии

В поддержку инициативы правительства США (White House Climate Data Initiative) корпорация EMC и компания Pivotal совместно с благотворительной организацией Earthwatch Institute, а также исследовательским институтом Schoodic Institute объявили о запуске в национальном парке «Акадия» программы «Большие данные на страже экологии» (Big Data vs. Climate Change: EMC & Citizen Scientists Team Up).

Целями программы являются:
- Эффективное накопление и хранение данных, предоставленных организацией Earthwatch Institute, а также собранных учеными-любителями на порталах eBird, iNaturalist, HawkWatch, National Phenology Network и National Park Service, за счет использования озер данных (Data Lake);
- Анализ и интерактивная визуализация данных, адаптация данных для изучения международным академическим обществом;
- Задействование большего количества ученых-любителей;
- Исследование взаимосвязи «природа-климат»;
- Разработка инструментов для поддержки программ парка «Акадия».

Кэтрин Уинклер, старший вице-президент и директор по устойчивому развитию корпорации EMC, по этому поводу говорит: «Многим из нас наука о данных и климатология кажутся сложными и абстрактными. Мы надеемся, что средства и платформы для удобного интерактивного анализа и визуализации данных позволят не только лучше понять, как изменение климата влияет на природу, но и сделает климатологию и науку о данных более наглядными и понятными».

5.5 Тропические леса

В ходе совместного проекта HP и Conservation International, получившего название Earth Insights, в тропических лесах 16 стран мира были размещены камеры и климатические датчики, собирающие данные о животных, растительности, температуре, атмосферных осадках, влажности и т.д. Управлять тремя терабайтами информации, включая более 1,4 млн фотографий и более трех млн климатических показателей, помогает платформа HP HAVEn.
Технологии Big Data помогают в режиме реального времени получать информацию о численности популяций редких животных, чтобы защитить их от исчезновения. Результаты будут передаваться администрации подконтрольных лесных районов для разработки политик охоты и лесозаготовки, не нарушающих природный баланс.

Выводы

В ходе исследовательской работы было выяснено:

- BigData возможно применить тогда, когда необходимо эффективно обработать поток больших данных(глава 1).
- BigData имеет такие решения для хранения и обработки больших данные, как MapReduce(глава 2.1), Hadoop(глава 2.2)
- Уже имеется богатый опыт применения BigData для решения проблем экологии не только в области защиты животных(глава 5.2), но и так же при решение проблем экологической чистоты больших мегаполисов(глава 5.3).

Список использованной литературы

1. Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung. Big Data. Related Technologies, Challenges, and Future Prospects. — Spinger, 2014. — 100 p. — ISBN 978-3-319-06244-0. — DOI:10.1007/978-3-319-06245-7
2. James Manyika et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, June, 2011. McKinsey (9 August 2011)
3. The Four V´s of Big Data (англ.). IBM (2011)
4. Виктор Майер-Шенбергер, Кеннет Кукьер. Большие данные. Революция, которая изменит то, как мы живём, работаем и мыслим = Big Data. A Revolution That Will Transform How We Live, Work, and Think / пер. с англ. Инны Гайдюк. — М.: Манн, Иванов, Фербер, 2014. — 240 с. — ISBN 987-5-91657-936-9.
5. Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future - EWeek (15 August 2011)
6. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data - Gartner (27 June 2011)
7. Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data". The Semantic Web. Lecture Notes in Computer Science. 4825. p. 722. doi: 10.1007/978-3-540-76298-0_52. ISBN 978-3-540-76297-3.