Соловьев А.И.
Хранение и обработка больших данных

*Университет ИТМО
(Россия, Самара)*

doi:10.18411/lj-04-2018-135

idsp: 000001:lj-04-2018-135

Аннотация

В статье рассмотрены проблемы хранения и обработки больших данных, технологии, применяемые для решения этих проблем (Hadoop, MapReduce, NoSQL). Также рассмотрена возможность использования технологий обработки больших данных в различных сферах деятельности.

Ключевые слова: большие данные, Big Data, Hadoop, MapReduce, NoSQL, Bigtable.

В век огромных достижений в области информационных технологий возникла необходимость обрабатывать большие данные (Big Data) в самых различных сферах деятельности человека.

Источниками для больших данных могут формироваться из самых разных источников - социальных сетей, деловых операций, от оборудования и датчиков. При этом немаловажным фактором обработки информации является скорость накопления информации и необходимость обработки ее в реальном времени. Большие данные могут включать в себя документы, электронную почту, информацию социальных сетей, аудио- и видео-файлы, и т.д. Информация зачастую не структурирована.

Еще недавно основными пользователями больших данных были такие IT-гиганты как Yahoo и Facebook, для решения проблемы маршрутизации пользователей Web-сайтов.

В настоящее время в использовании обработанных больших данных заинтересованы предприятия торговли для бизнес-анализа и составления прогнозов, прогнозирования потребительского спроса, рекламных акций и программ, формирования и ценовой политики, а также для реакции на изменения рынка.

В банках большие данные используются для оперативного получения информации, предоставления банковских продуктов, оценки кредитоспособности банков, предотвращения нежелательных транзакций и др.

В телекоммуникации большие данные используются для изучения предпочтений пользователей, оценки рентабельности и управление лояльностью клиентов.

Интернет-бизнес использует информацию о каждой транзакции покупателей для изучения предпочтений и спроса покупателей в реальном режиме времени.

Поисковые системы используют большие данные для привлечения рекламодателей и аналитиков, для целей маркетинговых и других исследований.

Например, данные Google предоставляются из бесплатных сервисов, такие как GoogleSearch, Gmail, GoogleTalk, YouTube, Google+, GoogleDocs, Карты Google.

Использование больших данных обосновано для составления прогнозов и анализа - чем больше данных, тем более обоснованный прогноз можно сделать.

Согласно прогнозам аналитиков International Data Corporation к 2020 году общее количество информации составит около 40 зеттабайт. Наиболее массивные потоки информации составляют социальные сети и хранилища мультимедийных файлов, данных систем мониторинга и контроля.

В информационных технологиях под большими данными подразумевается ряд технических решений, методов и инструментов для обработки неструктурированных и структурированных данных огромного размера для получения воспринимаемого результата, эффективного для решения необходимого круга задач.[12]

При работе с большими данными отмечают «3V»

Объем данных (volume);

Скорость обработки данных и получения результата (velocity);

Многообразие (variety) – одновременная обработка разнообразных данных.

Благодаря образованию большого количества данных в самых различных областях человеческой деятельности стало возможным использование технологий обработки для самых разных целей - производственных, научных, государственных, коммерческих, и т.д.

Главными технологиями больших данных являются:

1. MapReduce.
2. Hadoop.
3. NoSQL. [10]

MapReduce - технология, отвечающая за вычислениями больших данных в распределенных системах. MapReduce является моделью распределенных вычислений, представленная Google, и используется в компьютерных кластерах для вычислений очень больших (несколько петабайт) наборов данных.

MapReduce - фреймворк, использующий большое количество компьютеров (сервера, ноды) для работы с распределенными задачами.

Схема работы MapReduce изображена на рисунке 1.

Данные

Перемешивание
/сортировка

Рисунок 1. Схема MapReduce

Для решения этих задач MapReduce предварительно делает обработку данных, на главный компьютер поступают обработанные данные, затем данные разделяются, и передаются для предварительной обработки на другие компьютеры (рабочие узлы).

Затем происходит сверка данных - от рабочих станций поступают обработанные данные и на их основе делается решение поставленной задачи.

Особенностью MapReduce является возможность распределено производить распределенную обработку данных. Производительность MapReduce для обработки объема порядка петабайт данных составляет всего лишь несколько часов, даже при частичных сбоях компьютерной техники. При возникновении сбоя - часть вычислений может быть передана другому компьютеру.

Принципами работы с большими данными являются:

1. Горизонтальная масштабируемость.
2. Отказоустойчивость.
3. Локальность данных.

Под горизонтальной масштабируемостью подразумевается неорганическое увеличение объема обрабатываемой информации.

Отказоустойчивость означает, что при обработке больших данных система учитывает нарушения работоспособности работающих машин и обработка данных производится корректно.

Локальность данных учитывает, что данные распределены по большому количеству машин, и по возможности осуществляет обработку данных на тех же машинах, где расположены данные. Такой подход позволяет значительно повысить производительность работы. При локализации данных снижается трафик между коммутаторами – наиболее узкое место в вычислительных процессах с использованием вычислений больших данных. Поскольку пропускная способность передачи данных по сети значительно выше, чем дисков машины, производительность работы при соблюдении принципа локальности данных значительно выше.

Apache Hadoop – одна из популярных платформ на сегодняшний день, использующих MapReduce для обработки больших данных.

Apache Hadoop использует HDFS (Hadoop Distributed File System)- файловую систему с открытым исходным кодом. [6]

Вычислительный блок в Hadoop состоит из корневого (главного) узла и подчиненных ему узлов (станций). Главный узел может выполнять роли менеджера задач, исполнителя, узла имен, узла хранения. Подчиненные узлы могут выступать в роли узла хранения или исполнителей. Менеджер задач координирует все задачи при помощи списка задач исполнителей. Исполнитель в свою очередь запускает задачи и отправляет отчет о прогрессе работы менеджеру задач. Менеджер задач ведет статистику выполнения каждой задачи. В случае, когда задача завершается некорректно, менеджер может запустить ее на другом компьютере. Все узлы относятся к кластеру HDFS.

Реализация MapReduce на платформе Hadoop организована следующим образом. Распределитель обрабатывает данные, которые по умолчанию находятся в текстовом формате и для распределителя представляют собой строки со специальным ключом.

Выполнение задачи осуществляется следующим образом. На клиентском узле (на виртуальной машине JVM Java virtual Machine) создается задача, которая запускает новую задачу менеджера задач. Менеджер задач запускает и анализирует все запущенные задачи. После этого менеджер копирует кэш с файлами и распределяет на узлы. Исполняющие узлы передают менеджеру информацию о доступной мощности и возможности запуска задачи. После анализа полученной информации менеджер назначает узлу выполнение операции или сведение. Исполнитель запускает новую JVM для выполнения задачи менеджера.

При обработке данных в MapReduce существует необходимость в определении местоположения данных. Это оказывает большое влияние на производительность, так как пропускная способность сети в большом кластере намного ниже пропускной способности дисков на вычислительном устройстве.

Платформа Hadoop предназначена для организации распределенной обработки больших объемов данных и использование модели разделения и сбора (каждая задача разделяется на более мелкие части и каждая часть набора выполняется на отдельном узле кластера).

NoSQL – модель обработки больших данных, позволяющая обрабатывать данные без ограничений реляционных данных. С ее помощью решается проблема параллельной обработки неструктурированных данных огромного размера.

NoSQL используется для описания баз данных, не использующих SQL. NoSQL позволяет поддерживать постоянную пропускную способность при неограниченном (горизонтальная масштабируемость) увеличении объема данных хранилища.

NoSQL – классификация решений:

1. Хранилища ключ-решение.
2. Хранилища на графах.
3. Колончатые решения хранилищ.
4. Документ-ориентированные хранилища.

Хранилища ключ-значение. В основе обработки лежит словарь (ассоциативный массив). Обработка данных словаря осуществляется по ключу.

Хранилища на графах используются для данных, которые можно без затруднений представить в виде графов (социальные сети). Обработка данных производится путем обхода целевого графа по ребрам с интересующими свойствами.

Документ-ориентированные хранилища объединяют множество пар «ключ-значение» (называемые документами), которые объединяются в коллекции. Обработка данных производится по ключу или по значениям атрибутов.

Колончатые хранилища сходно с реляционными данными (колонки). Количество колонок в таблице может быть неограниченным. Колонки объединяются по ключам в определенные группы (семейства), имеющие определенный набор свойств. Обработка производится по ключам.

Особенности NoSQL:

1. Высокая пропускная способность.
2. Исключение излишнего усложнения.
3. Горизонтальная масштабируемость.
4. Отказ от консистентности.

В связи с появлением огромного количества данных (больших данных) возникает необходимость не только в обработке, но и в хранении полученной информации.

В последнее время набирает популярность подход хранения, основу которого составляет модель «ключ-значение» (таблица 1).

Таблица 1

Ключ	Значение
Имя	Андрей
Фамилия	Сидоров
Возраст	23

Такой подход удобен для хранения слабоструктурированных данных (его также называют NOSQL). Рассматриваемый подход позволяет производить параллельные вычисления, обладает горизонтальным масштабированием, позволяет хранить слабоструктурированные данные. Базы данных (системы хранения) построенные по такому принципу - BigTable (Google), Amazon (Dinamo), Facebook (Cassandra).

Рассмотрим систему хранения BigTable.

Bigtable предназначен для надежного масштабирования до петабайт данных и тысяч машин. Благодаря использованию Bigtable достигнуто несколько целей: широкая масштабируемость, применимость, высокая производительность и высокая доступность данных. Bigtable используется более чем такими проектами Google, как Google Finance, Google Analytics, Orkut, Writely и Google Earth. Эти продукты предназначены для различных требовательных рабочих нагрузок, которые варьируются от задач, ориентированных на разную пропускную способность как с точки зрения размера данных, так и требований к требованиям.

Bigtable - это распределенная система хранения для управления структурированными данными, предназначенная для масштабирования до очень большого размера: петабайты данных на тысячах серверов. Несмотря на эти разнообразные требования, Bigtable это успешно высокопроизводительное и гибкое решение для всех этих продуктов Google.

Bigtable не поддерживает полную реляционную модель данных; вместо этого он предоставляет клиентам простую модель данных, которая поддерживает динамический контроль над компоновкой и форматом данных, и позволяет клиентам отслеживать общие характеристики данных, представленных в базовом хранилище.

Данные Bigtable индексируются. В целом Bigtable представляет собой распределенную, постоянную, многомерную сортированную карту. Карта индексируется ключом строки, символом столбца и пр., каждый из которых является неинтерпретированным байтом.

В современном информационном пространстве накоплено огромное количество неструктурированных данных. В силу того, что большинство информации не структурировано, обработка информации не оптимально используется. Технологии управления большими данными позволяют решить эту проблему.

Огромный объем накопленной информации и постоянно увеличивающийся поток новых данных требует совершенных методов обработки этих данных. К таким методам относят применение MapReduce и Persona Server, Hadoop, NoSQL.

Применение технологий обработки больших данных позволяет анализировать огромный объем информации в реальном состоянии без применения унификации или очистки данных.

В практической деятельности это приведет к повышению рентабельности, оперативному решению возникающих проблем и задач, поможет создать прогноз предпочтений клиентов, позволить оптимально использовать объем производств, проводить анализ кредитоспособности клиентов и прочие практические задачи.

Технологии обработки больших данных могут быть использованы для решения задач:

1. Маркетинг.
2. Прогноз рынка.
3. Принятие эффективных управленческих решений.
4. Увеличение производительности труда.
5. Прогноз рынка.
6. Логистика.
7. Мониторинг состояний фондов.
8. Оценка кредитоспособности банков.
9. Оптимизация инвестиций и другие.

1. M. Tashkova. BIG DATA – The new Challengeefacing business. Економічний вісник Донбасу № 4(46), 2016 р. 164-167
2. Shugurov I.S., Mitsyuk A.A. Applying MapReduce to conformance checking. Труды Института системного программирования РАН. 2016
3. Герасимов С.В. , Мещеряков А.В., Глотов Е.С., Попов Е.С. Обработка больших объемов данных с помощью модели вычислений Map Reduce. Труды ИСП РАН т 22.вып 6. 2015. С 315-322
4. Гусейнов А.А., Бочкова И.А. Исследование распределенной обработки данных на примере системы Hadoop. Актуальные проблемы авиации и космонавтики.Т.1. 2016
5. Мамедова Г.А., Зайналова Л.А., Меликова Р.Т, Технологии больших данных в электронном образовании. Open education v 21.№6, 2017, с 41-48
6. Массобрио Рензо, Несмачнов Серхио, Черных Андрей.и др. Применение облачных вычислений для анализа данных большого объема в умных городах. Труды Института системного программирования РАН. 2016
7. Натальченко И.А. Анализ механизмов передачи крупных массивов данных через сеть интернет с помощью технологии веб-сервиса. Инженерный вестник Дона, №2, 2015.
8. Потапов В.П., Попов С.Е., Костылев М.А. Применение массово-параллельных технологий для организации потоковой обработки радарных данных. Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016
9. Тимонин А. Ю, Бождай А. С. Использование технологий Big data для построения социального профиля человека на основе открытых источников информации/ Вестник Пензенского государственного университета 2015
10. Тимонин А.Ю., Бождай А.С. Использование технологии Big Data для построения социального профиля человека на основе открытых источников информации. Вестник Пензинского государственного университета. №2(10) 2015 с 140-145
11. Шарапов Р.В. Аппаратные средства хранения больших объемов данных. Инженерный вестник Дона, №2, 2015
12. Янишевская А.Г. Чурсин М.А. Способы хранения и обработки большого объема данных с использованием Map Reduce и Persona Server