Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.657

Гродель
Юлия Валентиновна

Обработка больших массивов данных интернет-ресурсов

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Сечко В. В.
к. т. н., доцент

Минск 2014

КРАТКОЕ ВВЕДЕНИЕ

Среди наиболее обсуждаемых тем в ИТ–изданиях в последнее время выделяется феномен Big Data, или проблема «Больших данных». Проблема хранения и обработки большого объема данных стояла всегда, но с развитием ИТ она стала беспокоить не только ряд крупнейших корпораций, но и гораздо более широкий круг компаний. Что же стало предпосылкой к ее появлению?

В первую очередь, возросло число генераторов данных, причем весьма большого объема – это социальные сети разных видов, данные электронной почты, Twitter, Wiki-проекты. Кроме того, огромные объемы данных могут генерироваться датчиками различных типов – Call Data Records (CDR) сотовых операторов, телеметрические данные, информация с камер видеонаблюдения и т.п. Во-вторых, значительное уменьшение стоимости хранения привело к тому, что многие компании могут позволить себе следовать парадигме «данные слишком ценны, чтобы их уничтожать».

Но главная проблема заключается в том, что кроме количества данных изменился и их характер. Основной объем этих данных – неструктурированная информация, поэтому ее хранение и обработка в классических системах, как правило, малоэффективна.

Сегодня Big Data не просто модный термин – для многих организаций это стало насущной проблемой, требующей немедленного решения.

Для параллельных вычислений над большими наборами данных в компьютерных кластерах существует модель распределенных вычислений MapReduce. Парадигма MapReduce позволяет решать множество задач, связанных с анализом и обработкой больших объемов данных, за приемлемое время благодаря высокому параллелизму. Кроме того, данный подход устойчив к сбоям узлов и позволяет динамически распределять Map-и Reduce-подзадачи по узлам кластера, принимая во внимание фактическое распределение данных по узлам кластера.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является анализ основных задач и методик Big Data. Изучение модели распределенных вычислений MapReduce. Разработка статистической оптимизации для MapReduce программ, исследование применимости и актуальности данной оптимизации.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить основные понятия и границы Big Data. Ознакомиться с методиками анализа больших данных. Изучить модель распределенных вычислений MapReduce.
2. Предложить и разработать алгоритм оптимизации MapReduce программ.
3. Провести экспериментальные исследования.

Объектом исследования являются большие массивы данных.

Предметом исследования является модель распределенных вычислений MapReduce.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с научно-техническими заданиями и планами работ кафедры “Программное обеспечение информационных технологий” по теме «Разработать модели, методы, алгоритмы для оценки параметров, повышения надежности и качества функционирования аппаратно-программных средств систем и сетей сложной конфигурации и внедрить в современные обучающие комплексы » (ГБ № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя В. В. Сечко заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

По теме публикации опубликовано 2 печатные работы в сборниках трудов и материалов международных конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлена общая характеристика проблемы больших данных, проанализированы задачи, связанные с Big Data и описаны методики анализа массивов данных. Также подробно разобрана модель распределенных вычислений MapReduce и её проблемы. Во второй главе предложена и разработана статистическая оптимизация для MapReduce программ, описаны алгоритм и программная реализация. Третья глава посвящена экспериментальным исследованиям, которые позволили сделать вывод об актуальности оптимизации.

Общий объем составляет 66 страниц, из которых основного текста 49 страниц, 22 рисунка на 5 страницах, список использованных источников из 27 наименований на 2 страницах и 2 приложения на 10 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы.

В **первой главе** определены основные задачи, связанные с Big Data. Это хранение и управление, неструктурированная информация и анализ больших массивов данных. Рассмотрены проблемы, связанные с каждой и задач, и описаны возможные решения. В главе также описаны различные методики анализа данных. Рассмотрены следующие подходы: A/B-тестирование, сетевой анализ, кластеризация, краудсорсинг и обработка естественного языка. Данный список не претендует на полноту, однако в нем отражены наиболее востребованные в различных отраслях подходы. Работа над созданием новых и совершенствованием существующих методик не прекращается. Кроме того, некоторые из перечисленных выше методик вовсе не обязательно применимы исключительно к большим данным и могут с успехом использоваться для меньших по объему массивов (например, A/B-тестирование).

Для параллельных вычислений над большими наборами данных в компьютерных кластерах была описана модель распределенных вычислений MapReduce. Приведена и детально рассмотрена схема алгоритма данной модели. В качестве открытой реализации рассмотрен проект Hadoop, описаны его области применения и ограничения.

В главе была рассмотрена задача модификации алгоритма MapReduce и кратко описаны существующие пути её решения.

На некоторых MapReduce задачах может наблюдаться неудачное распределение промежуточных ключей, и как результат различное по времени выполнение заданий и самих программ. Исходя из этого во **второй главе** была поставлена задача оптимизации алгоритма MapReduce.

Также была приведена простейшая реализация модели MapReduce и подробно рассмотрен существующий алгоритм на примере открытого проекта Hadoop.

Для решения задачи оптимизации было принято решение использовать другой алгоритм распределения ключей, отличный от существующего. Был проведен анализ существующих алгоритмов упаковки в контейнеры. Выбор был сделан в пользу алгоритма – First Fit Descending. Для более оптимального распределения ключей алгоритм был модифицирован. В главе приводится подробное описание данного алгоритма, а также его блок-схема. Описаны общие проблемы и проведен предварительный эксперимент по анализу разработанного алгоритма.

В главе был описан этап сбора статистических данных. Было принято решение, на каком этапе алгоритма MapReduce будет собираться статистика, где и в каком виде статистика будет храниться.

В **третьей главе** были описаны проведенные экспериментальные исследования версии с оптимизацией и без. Эксперименты были проведены для трех программ: Word count, задачей которой является подсчет количества вхождения слова в данный текст, First Character для подсчета количества слов, начинающихся с каждой из букв, и программы анализа логов для подсчета средней длины сессии пользователей.

Из экспериментов можно заметить, что оптимизированная версия действительно лучше распределяет промежуточные ключи по reduce машинам, но при очень простых операциях (типа сложения) итоговое время исполнения стадии Reduce и, как следствие, MapReduce программ не влияет. В случае же неравномерно распределенных промежуточных ключей и долго исполняющейся стадии Reduce время работы может сократиться до 20-50%.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проанализированы основные задачи, связанные с BigData. Рассмотрены проблемы, связанные с каждой и задач, и описаны возможные решения. В ходе работы сделан вывод, что инструменты и технологии, существовавшие до сегодняшнего времени, не способны справиться с проблемой больших данных.

2. Рассмотрены различные методики анализа массивов данных. Сделан вывод, что чем более объемный и несистематический массив подвергается анализу, тем более точные и релевантные данные удастся получить на выходе.

3. Для параллельных вычислений над большими наборами данных в компьютерных кластерах была описана модель распределенных вычислений MapReduce. Была рассмотрена задача модификации алгоритма MapReduce и кратко описаны существующие пути её решения.

4. Для практического изучения модели распределенных вычислений был приведен пример реализации MapReduce. А также описаны основные этапы существующего алгоритма на примере открытого проекта Hadoop.

5. Предложена и разработана оптимизация для MapReduce программ. Из ряда исследований сделан вывод: для MapReduce программ с легкой стадией Reduce, на которой производится мало операций, например, когда функция Reduce просто суммирует переданные значения, результаты оптимизации не заметны. Такое поведение легко объяснимо и ожидаемо, так как данная оптимизация направлена на стадию Reduce. В примерах же, где стадия Reduce выполняется достаточно долго, и сами промежуточные значения распределены неравномерно по ключам, можно ожидать заметного улучшения скорости исполнения MapReduce программ.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую ба-зу для последующего анализа и улучшения алгоритма MapReduce под конкретные задачи.

2. Разработанная статистическая оптимизация для MapReduce программ может быть использована для улучшения временных показателей в обработке больших массивов данных.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Гродель Ю. В. Проблема Big Data и NoSQL подход к её решению [Текст] / Ю. В. Гродель, Д. А. Лагун // Наука, образование, общество: тенденции и перспективы: Сборник научных трудов по материалам Международной научно-практической конференции 28 ноября 2014 г.: в 5 частях. Часть III. М.: «АР-Консалт», 2014 г.- 154 с

2. Гродель Ю. В. Big Data и модель распределенных вычислений MapReduce [Текст] / Ю. В. Гродель // Приоритетные направления развития науки и образования : материалы III междунар. науч.–практ. конф. (Чебоксары, 04 дек. 2014 г.) / редкол.: О. Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2014. – С. 159–160. – ISBN 978-5-906626-52-3.