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

На правах рукописи

УДК 004.45

ГУНДОРИНА
Елена Александровна

ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ РАБОТЫ BIG *DATA HADOOP CLUSTERS*, ИСПОЛЬЗУЯМЕТОДЫ *MACHINE LEARNING*

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1–38 80 04 Технология приборостроения

Научный руководитель
Цырельчук Игорь Николаевич
кандидат технических наук,
доцент

Минск 2015

Работа выполнена на кафедре проектирования информационно-компьютерных систем учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **Цырельчук Игорь Николаевич,** кандидат технических наук, доцент, заведующий кафедрой проектирования информационно-компьютерных системуч реждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Рецензент: **Бондарик Василий Михайлович,** кандидат технических наук, доцент, декан факультета непрерывного и дистанционного обучения учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Защита диссертации состоится «22» июня 2015 г. года в $9^{\underline{00}}$ часов на заседании Государственной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, г.Минск, ул. П.Бровки, 6, 1 уч. корп., ауд. 415, тел.: 293-20-87, e-mail: kafpiks@bsuir.by.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

Сегодня наблюдается большой интерес к технологиям класса *Big Data*, связанный с постоянным ростом данных, которыми приходится оперировать крупным компаниям. 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. Развитие и начало широкого использования этих источников послужило отправной точкой для проникновения технологий больших данных едва ли не во все сферы деятельности человека. В первую очередь, в научно-исследовательскую деятельность, в коммерческий сектор и сферу государственного управления. Накопленная информация для многих организаций является важным активом, однако обрабатывать ее и извлекать из нее пользу с каждым днем становится все сложнее и дороже.

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.

Hadoop используется для обработки большего объема данных и быстрорастущих данных, предназначен для обработки неструктурированных данных, при использовании нужно учесть, что сам по себе он не дает доступ к данным в реальном времени, при формировании запросов обрабатывается весь массив данных.

Hadoop применяют для построения глобальной аналитики, систем машинного обучения, корреляционного анализа разнообразных данных, систем статистики. *Hadoop* сам по себе не может использоваться как операционная база данных. Как правило, в корпоративной среде *Hadoop* используется совместно с реляционными базами данных. Для устранения базовых недостатков фреймворка используются дополнительные модули и внешние приложения, взаимодействие с которыми описывается ниже.

Не существует универсальных способов анализа или алгоритмов, пригодных для любых случаев и любых объемов информации. Методы анализа данных существенно отличаются друг от друга по производительности, каче-

ству результатов, удобству применения и требованиям к данным. Оптимизация может производиться на различных уровнях: оборудование, базы данных, аналитическая платформа, подготовка исходных данных, специализированные алгоритмы. Анализ большого объема данных требует особого подхода.

В связи с вышеизложенными фактами, в рамках магистерской диссертации были исследовать подходы оптимизации производительности *Big Data Hadoop Clusters*, на основе данных, предоставляемых специально разработанным программным обеспечением (агентом).

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Тема *Big Data* сегодня приобретает особую актуальность: изменяя подходы к анализу информации и способы принятия решений, этот тренд оказывает влияние на бизнес всех размеров.

Организации и раньше вкладывали значительные средства в приложения для автоматизации бизнес-процессов и улучшения операционной эффективности. Многие из этих проектов еще продолжаются, но становится совершенно ясно, что управление данными с помощью традиционных методов не обеспечивает правильной информации в правильное время и ее предоставления «правильным» сотрудникам. Это происходит по разным причинам: из-за плохой интеграции систем и невысокого качества данных, а также проблем с производительностью и масштабируемостью. Многие организации сегодня не могут справиться с трудностями получения данных, необходимых для принятия критически важного решения.

Big Data– это тренд, который влияет на бизнес всех размеров и изменяет способы, которыми компании анализируют информацию и принимают решения. Многие предприятия, которым приходится управлять притоком данных, могут получить от поступающей информации значимую ценность лишь с помощью стратегий *Big Data*. В связи с этим очень важен технологический сдвиг в плане извлечения ценности из «сырых», «неочищенных» массивов неструктурированных цифровых данных.

Развитие алгоритмов *Machine Learning* вызвано ростом возможностей современных вычислительных систем, еще более стремительным ростом объемов данных, доступных для анализа, а также постоянным расширением области применения методов машинного обучения на все более широкий класс задач обработки данных.

Технологии больших данных и аналитических методов оптимизации решений реализуются на дешевом оборудовании и бесплатном программном обеспечении, однако управление *Big Data Hadoop Clusters* является очень сложной задачей.

Использование алгоритмов *Machine Learning* для анализа производительности работы *Big Data Hadoop Clusters* является актуальным и перспективным направлением.

Степень разработанности проблемы

Исследование оптимизации производительности *Big Data Hadoop* кластера проводилось на основе построения статистических моделей, анализирующих характеристики работы системы, рассматривающихся в работах в работах В. Майер-Шенбергера, К. Кукьераи других авторов. Изучение существующих технологий построения прогностических моделей представлено в работах А. Холмса, Ч. Ларна и других авторов.

Для изучения регрессионного анализа были изучены работы С. Армстронга, Р. Фишера и других авторов.

Для изучения кластерного анализа были рассмотрены работы И. Манделя, Б. Дюрана, М. Олдендерфера и других авторов.

Особенностью исследуемой области является то, что технологии, связанные с *Big Data*, начали активно развиваться не более 10 лет назад. Поэтому на данный момент не существует эмпирических исследований и разработок в данной сфере.

Цель и задачи исследования

Целью диссертационной работы является анализ, обоснование и внедрение алгоритмов машинного обучения для мониторинга и оптимизации производительности работы *Big Data Hadoop Clusters*.

Задачи, решаемые в диссертационной работе:

– анализ существующих платформ для работы с большими данными и их оптимизация;

– кластеризация данных со схожими характеристиками производительности, использования и обращения к ресурсам с целью оптимизировать обращение к ресурсам;

– определить зависимости параметров загрузки и конфигурации кластера от характеристик производительности *Hadoop* кластера;

– определение временных тенденций нагрузки кластера, объема данных, количества операций ввода/вывода и использования памяти для оптимизации использования памяти;

– выявление значительных изменений во времени отклика для работ, использования оперативной памяти и дискового пространства на каждом *Name Node* и *Data Node*;

– предсказание влияния увеличения количества пользователей и объема данных на производительность кластера.

Объектом исследования является *Big Data Hadoop* кластер.

Предметом исследования выступают параметры производительности, загрузки и обращения к ресурсам, характеристики конфигурации *Big Data Hadoop* кластера.

Основным методом, используемым в диссертационной работе, является статистический анализ производительности работы *Big Data Hadoop Clusters*, используя методы *Machine Learning*.

Область исследования. Содержание диссертационной работы соответствует образовательному стандарту высшего образования второй ступени (магистратуры) специальности 1–38 80 04 Технология приборостроения.

Теоретическая и методологическая основа исследования

Теоретическою основу диссертации составляют результаты исследований отечественных и зарубежных учёных в области *Big Data*, машинного обучения, обработки и анализа данных.

Обработка и анализ данных с помощью методов *Machine Learning* и построение графических зависимостей проводилось с помощью *R* и *RStudio*. В качестве архива документации и программного обеспечения, написанного на языке программирования *R*, использовался репозиторий *CRAN*.

Информационная база для литературного анализа по данной теме сформирована на основе более ранних наработок и исследований в этой области, а также ресурсов интернет.

Научная новизна диссертационной работы заключается в интеграции теоретических и экспериментальных данных для разработки подходов и механизмов для оптимизации работы *Big Data Hadoop* кластера.

Основные положения, выносимые на защиту

1. Автоматизированный процесс обработки входных данных, характеризующих производительность работы *Big Data Hadoop* кластера. Алгоритм подготовки входных данных для работы с алгоритмами *Machine Learning* в зависимости от их структуры.

2. Процесс поиска оптимального метода кластеризации в рамках автономного режима работы программного модуля.

3. Возможность использования регрессионного анализа для поиска тренда данных, базирующегося на теории работы с большими данными.

4. Автоматизированный процесс поиска критических значений, выбросов.

5. Использование техник *Machine Learning* для прогнозирования изменения характеристик производительности *Big Data Hadoop* кластера.

Теоретическая значимость диссертации заключается в том, что полученные результаты позволяют более детально изучить механизмы и алгоритмы анализа данных.

Практическая значимость диссертации состоит в том, что, используя полученные результаты, можно оптимизировать производительность работы *Big Data Hadoop* кластера.

Разработанное программное обеспечение обладает высокой функциональностью: просто в использовании, наглядно, примененные шаблоны программирования обеспечивают возможность дальнейшего совершенствования.

Помимо решения чисто практических задач, разработанное программное обеспечение может использоваться для обучения студентов по дисциплинам, связанных с анализом данных.

Апробация и внедрение результатов исследования

Результаты исследований были представлены на XIX Международной научно-технической конференции «Современные средства связи» (Минск, 14-15 октября 2014 года), 51-ой научно-технической конференции аспирантов, магистрантов и студентов БГУИР (Минск, 22-25 апреля 2015 года).

Публикации

Основные положение работы и результаты диссертации изложены в четырех опубликованных работах общим объемом 5,0 страниц (авторский объем 5,0 страниц).

Структура и объем работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введе-

ния, трех глав и заключения, библиографического списка и приложений. Общий объем диссертации − 72 страницы. Работа содержит 11 рисунков. Библиографический список включает 51 наименование.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы оптимизации работы *Big Data Hadoop* кластера, определены основные направления исследования, а также описано обоснование актуальности темы диссертационной работы.

В **общей характеристике работы** сформулированы цель и задачи диссертационной работы, показана связь с научными программами и проектами, даны сведения об объекте исследования и обоснован его выбор, представлены положения, выносимые на защиту, приведены сведения о личном вкладе соискателя, апробации результатов диссертации и их опубликованность, а также структура и объем диссертации.

В первой главе были детально рассмотрены:

– основные характеристики исследуемой предметной области;

– существующие методы анализа данных, которые могут быть применены для решения поставленных задач;

– существующие аппартные решения, которые применимы для работы с большими объемами данных;

– технологии, использующиеся для работы с *Big Data*.

Данный обзор помогает глубже изучить исследуемую область, сориентироваться в существующих программных и аппаратных решениях, которые могут быть использованы для решения поставленных целей и задач.

На основе обзора литературы отечественных и иностранных авторов были проанализированы основные характеристики исследуемой предметной области *Big Data*. Проведенный обзор указывает на актуальность темы диссертационной работы и ее перспективность;

Рассмотрены существующие методы анализа данных, которые могут быть применены для решения поставленных задач. Выделены методы, которые помогут оптимизировать работу *Big Data Hadoop Cluster*. Отмечена сложность реализации и использования методов анализа данных;

Рассмотрены существующие аппаратные решения, которые активно используются в сфере *Big Data*. Выбрано оптимальное для данной диссертаци-

онной работы аппаратное решение - *Big Data Hadoop Cluster*, которое обладает большим потенциалом для работы с большими объемами данных.

Проанализированы технологии, использующиеся для работы с *Big Data*. В результате анализа существующих технологий для анализа данных и оптимизации производительности работы *Big Data Hadoop Cluster* был выбран язык программирования *R*, который имеет свободную программную среду вычислений с открытым исходным кодом.

Во второй главе сформулированы цель и задачи диссертационной работы, показана связь с научными программами и проектами, даны сведения об объекте исследования и обоснован его выбор, представлены положения, выносимые на защиту, приведены сведения о личном вкладе соискателя, апробации результатов диссертации и их опубликованность, а также, структура и объем диссертации.

Дается представление об анализируемой системе – *Big Data Hadoop* кластере. Рассказывает о методе получения и хранении анализируемых данных.

Проанализирована исследуемая система – *BigDataHadoop* кластер. Описаны главные характеристики, которые могут быть использованы для анализа производительности работы *Big Data HadoopCluster*. Предложенные характеристики являются необходимыми, но не исчерпывающими.

Рассмотрен метод получения и хранения данных, характеризирующий производительность *Big Data Hadoop* кластера. Для сбора данных, характеризирующих производительность работы *Big Data Hadoop* кластера было создано специальное программное обеспечение (агент).

Выявлены задачи решение которых, может оптимизировать работу *Big Data Hadoop* кластера. Определены характеристики, которые следует оптимизировать для оптимизации работы *Big Data Hadoop* кластера. Данные были сгруппированы для решения конкретных поставленных задач.

В **третьей главе** представлены результаты разработки программного обеспечения.

Предложен оптимальный автоматизированный механизм обработки входных данных для их дальнейшего использования машинными алгоритмами. Алгоритм имеет гибкий подход к данным, его легко модифицировать и адаптировать для любой выборки, не зависимо от количества и характеристик переменных.

Выполнено сравнение возможных вариантов проведения кластерного анализа, выбран наилучший вариант в рамках автономной работы продукта. Анализ показывает сложность реализации автономной кластеризации данных,

так как задача в общем случае должна решаться индивидуально для каждой выборки в зависимости от природы данных и исследуемой предметной области. Предложен алгоритм, который использует более 30 параметров для определения наилучшего разбиения. Данный подход является ресурсоемким, результаты его работы оправдывают затраты ресурсов своей точностью;

Выполнен регрессионный анализ, который отображает тренд данных. Реализовано два варианта, которые характеризуют тренд с разных сторон и являются дополняющими друг друга.

Предложен способ нахождения выбросов в данных. Возможно использовать двух его реализаций, одна из который использует встроенную в язык *R* функцию *getOutliers()*, а вторая – построение «ящика с усами». Возможно сравнение работы двух вышеупомянутых подходов и выбор оптимального решения.

Рассмотрены механизмы прогнозирования изменения характеристик работы *BigDataHadoopClusters* с выбором наилучшего метода для конкретной задачи. Результаты работы алгоритмов программного обучения подвергаются сравнению и анализу. Реализованные подходы легко модифицируются для решения задач, не привязываясь к структуре данных и характеристикам исследуемой области.

В **приложении** приведены распечатанные слайды презентации магистерской диссертации.

ЗАКЛЮЧЕНИЕ

1. При разработке данной диссертационной работы было рассмотрено текущее состояние вопроса анализа производительности *Big Data Hadoop* кластера с помощью алгоритмов *Machine Learning*, проанализирована специализированная литература.
2. Исследуемая сфера является актуальной в связи с увеличением потребности в обработке больших объемов данных.
3. Произведен теоретический анализ передовых методов и технологий анализа больших данных, а также существующих аппаратных решений для реализации технологий и принципов *Big Data*.
4. Проанализирована предметная область, структура данных, которые подвергались анализу. Выделены характеристики, которые влияют на производительность работы *Big Data Hadoop* кластера.

5. Результатом диссертационной работы является программный продукт, который автономно, без вмешательства человека, может решать следующие задачи:

– обработка входных данных, для дальнейшего использования алгоритмами *Machine Learning*;

– определены зависимоси параметров загрузки и конфигурации кластера от характеристик производительности;

– выявлены значительные изменения во времени отклика для работ, использования оперативной памяти и дискового пространства на каждом *Name Node* и *Data Node*;

– кластеризация данных со схожими характеристиками производительности, использования и обращения к ресурсам;

– определение временных тенденций нагрузки кластера, объема данных, количества операций ввода/вывода и использования памяти с помощью регрессионного анализа;

– поиск критических значений, выбросов во входных данных;

– прогнозирование изменения характеристик производительности *Big Data Hadoop* кластера, в частности, прогнозирование влияния увеличения количества пользователей и объема данных на производительность кластера.

6. Данный продукт может быть применен для анализа производительности работы *Big Data Hadoop* кластера, который работает с большим объемом данных, в реальном времени.

7. Данный продукт сильно привязан к реализации агента, собирающего характеристики работы *Big Data Hadoop* кластера. При некоторой настройке работы разработанного программного обеспечения возможно его внедренит в любую сферу, использующую *Big Data Hadoop* кластер, так как реализованные подходы и методики являются универсальными.

8. Результаты работы данного программного комплекса должны пройти дальнейший анализ, который поможет автономно изменять конфигурационные настройки кластера для улучшения производительности работы *Big Data Hadoop* системы, что является задачей для дальнейшего исследования.

9. Разработанное программное обеспечение может быть использовано для изучения студентами использования техник *Machine Learning* для анализа данных с помощью языка *R* и было внедрено в учебный процесс.

Список опубликованных работ

1. Гундорина, Е.А. Handling missing data in air quality problems in R / Е.А. Гундорина // Материалы XIX международной НТК «Современные средства связи» – Минск: УО ВГКС, 2014. – С. 157.

2. Гундорина, Е.А. Determining the number of clusters in cluster analysis with affinity propagation method in R / Е.А. Гундорина// Материалы 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР – Минск: УО БГУИР, 2015 (Принято, в печати).

3. Гундорина, Е.А. Determining the number of clusters in cluster analysis with k-medoids algorithm in R / Е.А. Гундорина// Материалы 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР – Минск: УО БГУИР, 2015 (Принято, в печати).

4. Гундорина, Е.А. Determining the number of clusters in cluster analysis with sum of squared error method in R / Е.А. Гундорина // Материалы 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР – Минск: УО БГУИР, 2015 (Принято, в печати).