**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»**

**Журнал научно-исследовательской работы**

Студент  **Бобряков Александр Сергеевич**

##### Институт №8 Информационные технологии и прикладная математика

###### Кафедра 810Б «Информационные технологии в моделировании и управлении»

##### Учебная группа М8О-203М-19

Направление подготовки (специальность)

**02.04.02 «Фундаментальная информатика и информационные технологии»**

Руководитель практики от МАИ

**Ревизников Д.Л. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Подпись руководителя

Студент

**Бобряков А.С.**  « 01 » сентября 2020г.

Подпись

**Москва 2020**

1. **Индивидуальное задание студенту**

Анализ существующих подходов к решению задачи оптимизации Big Data приложений.

Анализ научных публикаций с выявлением сильных и слабых сторон подходов оптимизации.

Оценка релевантности темы оптимизации Big Data приложений.

1. **План выполнения индивидуального задания**

|  |  |  |
| --- | --- | --- |
| № пп | Наименование раздела или этапа | Срок выполнения |
| 1 | Начальный обзор темы оптимизации Big Data приложений | 20.09.2020 |
| 2 | Поиск различных публикаций и научных статей | 15.10.2020 |
| 3 | Перевод статей | 25.11.2020 |
| 4 | Анализ существующих подходов из статей | 10.12.2020 |
| 5 | Описание проведенных исследований в дипломной работе | 23.12.2020 |

Студент

**Бобряков А.С.**  « 01 » сентября 2020г.

1. **Отчет студента**

В ходе научно-исследовательской работы были проанализированы существующие подходы к оптимизации Big Data приложений, опубликованных в открытых источниках и научных статьях. В них были выявлены как достоинства, так и слабые стороны, которые необходимо учитывать в реализации собственного подхода алгоритма оптимизации приложений.

Основная статья, покрывающая тему научно-исследовательской практики:

1. *A Survey on Automatic Parameter Tuning for Big Data Processing Systems, 2020. Herodotos Herodotou, Yuxing Chen, Jiaheng Lu.*

В работе приведен анализ подходов, которые решают проблемы, связанные с автоматической настройкой параметров приложений и систем обработки больших данных (Hadoop, Spark, Apache Storm, Heron, Flink и другие). К таким проблемам можно отнести существование огромного количества параметров конфигураций, которые могут влиять на производительность системы неочевидным и сложным образом, а также неструктурированный характер входных данных. Эти проблемы ведут к непредсказуемому времени выполнения задач, а также к нестабильным нагрузкам на аппаратуру.

В статье представлено исследование нескольких подходов, направленных на решение вышеупомянутых проблем, либо для точного прогнозирования производительности приложения при различных настройках параметров, либо для нахождения почти оптимальных настроек параметров для различных сценариев.

Подходы делятся на шесть широких категорий: *на основе правил, моделирование затрат, на основе моделирования, экспериментально, машинное обучение, и адаптивный тюнинг.*

Каждый из шести подходов превосходит один или более аспектов, имеющих свои уникальные сценарии применения. Начиная с незнания настройки параметров, экспериментальный метод является наиболее доступным. По мере накопления опыта можно создавать правила или функции затрат для эффективного повышения производительности системы. По мере увеличения сложности систем моделирование небольших компонентов помогает лучше понять характеристики системы. При попытке настроить очень сложные системы и приложения с большим пространством параметров могут быть полезны методы машинного обучения с соответствующими обучающими данными, поскольку они обычно игнорируют внутренние компоненты системы. Наконец, для специальных и длительных рабочих нагрузок адаптивный подход – лучший вариант.

Так как по мере увеличения размера и сложности систем аналитики больших данных растет и потребность в автоматическом обеспечении хорошей и надежной производительности системы, способной справляться с постоянно растущими темпами производства данных, авторы ставят ряд открытых проблем:

1. *неоднородность* – проблема, вызванная наличием аппаратных узлов с различными емкостями и количествами ядер/памяти;
2. *облачная среда* – проблемы, связанные с несколькими арендаторами, накладными расходами и взаимодействием производительности при виртуализации;
3. *аналитика в реальном времени* – проблемы, вызванные предоставлением аналитики на лету, так как обычно такие системы используют несколько других систем обработки больших данных.

Углубленное понимание существующих подходов, представленное в этой статье, имеет ключевое значение для решения этих новых проблем эффективным и действенным способом для достижения конечной цели разработки действительно самонастраивающихся и самоконфигурируемых систем.

1. *Using Machine Learning to Optimize Parallelism in Big Data Applications, 2017. Alvaro Brandon Hernandeza, Maria S. Pereza, Smrati Gupta, Victor Muntes-Mulero.*

В этой статье предлагается метод на основе машинного обучения, который рекомендует оптимальные параметры для распараллеливания задач в рабочих нагрузках с большими данными. Путем мониторинга и сбора метрик на уровне системы и приложения, находятся статистические корреляции, которые позволяют характеризовать и прогнозировать влияние различных настроек параллелизма на производительность. Эти прогнозы используются, чтобы рекомендовать пользователям оптимальную конфигурацию перед запуском их рабочих нагрузок в кластере, избегая возможных сбоев, снижения производительности и нерационального использования ресурсов. Метод оценивается с помощью теста из 15 приложений Spark на стенде Grid5000. Наблюдается прирост производительности до 51% при использовании рекомендуемых настроек параллелизма. Модель также поддается интерпретации и может дать пользователю представление о том, как различные показатели и параметры влияют на производительность.

Результаты этой статьи показывают, что можно точно предсказать время выполнения приложения на основе больших данных с разными размерами файлов и настройками параллелизма с использованием правильных моделей. В качестве моделей в статье рассматривались: Bayesian Ridge, Linear Regression, SGD Regressor, Lasso, Gradient Boosting Regressor, Support Vector Regression, MLPRegressor.

1. *Learning-based Automatic Parameter Tuning for Big Data Analytics Frameworks, 2018. Liang Bao, Xin Liu, Weizhao Chen.*

В статье представлена собственная реализация системы автоматической настройки параметров, которая направлена на оптимизацию времени выполнения приложений на фреймворках аналитики больших данных.

Система изначально конструирует небольшой испытательный стенд из большой производственной системы, чтобы он мог генерировать релевантные образцы данных для лучшего обучения прогнозируемой модели в заданном временном ограничении. Кроме того, такой алгоритм создает набор образцов, которые могут обеспечить широкий охват многомерного пространства параметром конфигурации.

Алгоритм не является стандартным в области машинного обучения. Он базируется на выборке гиперкуба LHS (latin hypercube sampling) для дальнейшей генерации эффективных выборок в многомерном пространстве параметров, которые используются для выбора перспективных конфигураций в ограниченном пространстве, предлагаемых существующими лучшими конфигурациями.

1. *Quality Assurance for Big Data Application– Issues, Challenges, and Needs, 2016. Chuanqi Tao, Jerry Gao.*

В данной работе авторы проводят исследования о том, каким образом необходимо осуществлять валидацию приложений обработки больших данных для обеспечения качества работы всей системы, возникающие проблемы и потребности при этом. Также обсуждаются факторы качества и проводится сравнение между традиционным тестированием и тестированием приложений обработки больших данных.

Авторы отмечают основной процесс тестирования, состоящий из следующих шагов:

1. Функциональное тестирование систем больших данных для алгоритмов, возможности обучения, специфичных для предметной области функций:
2. Функциональное тестирование систем больших данных для целостности системы, безопасности надежности;
3. Тестирование визуализации, набора данных, удобства использования;
4. Тестирование, связанное с временными аспектами: тестирование работы в реальном времени, тестирование в течение срока службы и т.д.
5. *Machine Learning and Big Data Processing: A Technological Perspective and Review, 2018. Roheet Bhatnagar.*

Данная статья является обзорной и раскрывает роль алгоритмов и методов на основе машинного обучения в обработке и аналитике больших данных. Автор представляет обзор проблем, связанных с основными подходами машинного обучения: избыточность данных, зашумленность, неоднородный характер, дискретность, маркировка, несбалансированность, представление функций и их выбор.

Все эти проблемы необходимо учитывать при создании системы автоконфигурирования приложений анализа и обработки больших данных. В целом документ нацелен на предоставление как текущих практик, так и будущих направлений исследований в области обработки больших данных с использованием методов машинного обучения

1. **Вывод**

Таким образом, были рассмотрены публикации, так или иначе относящиеся к теме выпускной квалификационно работе.

Основным наблюдениям является то, что статей по данной проблематике достаточно мало, а существующие опубликованы совсем недавно. Основная причина заключается в том, что системы обработки больших данных усложняются в процессе роста вычислительных мощностей. Вследствие этого появляется все больше настраиваемых параметров и конфигураций, необходимых для стабильной и оптимальной работы.

Исследование методов автоматической настройки параметров – многообещающий, но достаточно сложный подход оптимизации производительности системы. Сейчас самим пользователям приходится сталкиваться с проблемой управления таких сложных систем, но для этого не хватает инструментов для упрощения конфигурирования, а также глубоких знаний по всем многочисленным параметрам, их корреляций и зависимостям. Например, системы Hadoop и Spark имеют более сотни параметров, которые можно настраивать, однако их настройка является непосильной задачей для человека, вследствие чего обычно принимают настройку по умолчанию, что совершенно не является оптимальной для текущей задачи.

Вторым важным наблюдениям является то, что рекомендуется использовать стандартные методы машинного обучения, вместо каких-либо нейросетей, так как базовые алгоритмы вполне справлялись с задачами, описываемых в статьях.

Поэтому основной задачей является реализация такой системы оптимизации выполнения Big Data приложений, которая будет учитывать все ключевые аспекты конфигурирования и предоставлять удобный интерфейс пользователю, либо будет полностью автономной. При реализации такой системы необходимо учесть недостатки существующих подходов, описанных в научных публикациях, и предоставить их решение.

Студент

**Бобряков А.С.**  « 04 » января 2021г.