Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

УДК 004.62

Жук
Александр Андреевич

МЕТОДЫ И СРЕДСТВА ОЦЕНКИ ЭФФЕКТИВНОСТИ ВЫПОЛНЕНИЯ ПРИЛОЖЕНИЙ, РАБОТАЮЩИХ С БОЛЬШИМИ ДАННЫМИ

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1 – 40 80 06 – Искусственный интеллект

Научный руководитель

Бойко Игорь Михайлович
кандидат технических наук

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

Мы живем в эпоху данных. Нелегко измерить общий объем данных, хранимых в электронном виде, но по оценке IDC размер «цифровой вселенной» составляет 4,4 Зетабайта в 2013 и прогнозирует десятикратный рост к 2020 году до 44 Зетабайт. Зетабайт 10^{21} Байт или, что эквивалентно, одна тысяча Эксабайт, один миллион Петабайт или один миллиард Терабайт **?**.

Обработка таких объемов данных, используя традиционные информационные технологии, может быть крайне неэффективна, для решения подобной проблемы существует класс технологий называемый большими данными. Однако данные технологии имеют достаточно высокий порог вхождения для разработчиков, они требуют наличия глубоких знаний не только конкретных технологий, но и знания особенностей языков программирования, на которых написана данная технология. Отсутствие таких знаний может привести к проектированию неэффективных приложений, работающих с технологиями больших данных, которые будут уступать в производительности традиционным технологиям обработки данных.

Целью магистерской диссертации является повышение эффективности выполнения программ, работающих с технологиями больших данных.

Проблемы, которые необходимо решить для достижения поставленной цели:

– отсутствие средств, способных одновременно анализировать разнообразные приложения, написанные с помощью технологий больших данных, и пользоваться особенностями той или иной технологии для более качественного анализа и предоставления оптимизации для этого приложения;

– отсутствие универсальных методов оценки эффективности и оптимизации приложения, написанные с помощью технологий больших данных.

Задачами магистерской диссертации являются:

– анализ технологий в области обработки больших данных, выявление направлений повышения их эффективности;

– разработка метода анализа эффективности выполнения программ, работающих с технологиями больших данных;

– оценка эффективности предложенных решений.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Тема диссертации соответствует приоритетному направлению «Информатика и космические исследования» согласно пункту 5 перечня приоритетных направлений научных исследований Республики Беларусь на 2016–2020 гг. (Постановление Совета Министров Республики Беларусь от 12 марта 2015 г. № 190).

Целью магистерской диссертации является повышение эффективности выполнения программ, работающих с технологиями больших данных.

Проблемы, которые необходимо решить для достижения поставленной цели:

– отсутствие средств, способных одновременно анализировать разнообразные приложения, написанные с помощью технологий больших данных, и пользоваться особенностями той или иной технологии для более качественного анализа и предоставления оптимизации для этого приложения;

– отсутствие универсальных методов оценки эффективности и оптимизации приложения, написанные с помощью технологий больших данных.

Задачами магистерской диссертации являются:

– анализ технологий в области обработки больших данных, выявление направлений повышения их эффективности;

– разработка метода анализа эффективности выполнения программ, работающих с технологиями больших данных;

– оценка эффективности предложенных решений.

Опубликование результатов диссертации

По материалам выполненных исследований опубликовано 3 тезиса докладов в сборниках материалов научных конференций.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначены актуальные задачи.

Первая глава содержит:

– определение для технологий больших данных;

– определены существующие варианты архитектурных решений для технологий больших данных;

– дана спецификация конвейера обработки данных;

– поставлена задача оптимизации конвейера обработки данных.

– сформулированы требования к разрабатываемому методу.

Во **второй главе** рассмотрены существующие методы оценки эффективности приложений работающих с большими данными.

В **третьей главе** описан метод рекурсивного анализа и оптимизации абстрактного конвейера обработки данных, а также описаны общие подходы к оптимизации приложений, работающих с большими данными.

В **четвертой главе** описан эксперимент применения разработанного метода и сравнение его с уже существующего средства оптимизации.

В **выводе** описаны полученные результаты.

ЗАКЛЮЧЕНИЕ

В рамках магистерской диссертации был разработан метод рекурсивной анализа и оптимизации абстрактного конвейера обработки данных.

Была сформулирована задача оптимизации конвейера обработки данных, как абсолютная, так и условная.

Был произведен анализ существующих технологий, работающих с большими данными, анализ существующих систем анализа эффективности выполнения приложений, определены архитектурные особенности.

Были решены поставленные задачи:

– разработан метод анализа эффективности и оптимизации приложения, который является универсальным для технологий, работающих с больших данных;

– произведена оценка эффективности предложенных решений.

Метод соответствует предъявленным требованиям:

– метод является универсальным для большинства актуальных технологий больших данных;

– метод учитывает, при анализе, архитектурные особенности исследуемой технологии;

– метод дает четкое представление о качестве исследуемого приложения и определять возможные пути решения выявленных в приложении дефектов.

Разработанный метод был проверен экспериментальным путем, в результате время выполнения оптимальной реализации конвейера обработки данных составляет 9.21% от первоначальной реализации. Анализ и оптимизация при помощи разработанного метода показала лучшей результат на экспериментальном приложении, чем встроенная в технологию Apache Spark система оптимизации работы приложения Catalyst Optimizer.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

Тезисы докладов в сборниках материалов научных конференций

1. Жук, А.А. Формулировка задачи оптимизации приложений, работающих с большими данными / А.А. Жук // Информационные технологии и системы-2020 (ИТС 2020) : материалы междунар. науч. конф., Минск, 18 ноября 2020 г. / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин (гл. ред) [и др.]. — Минск, 2020. — С. 72-73.

2. Жук, А. А. Разделение данных, как универсальный метод оптимизации приложений, работающих с большими данными / Жук А. А. // Информационные технологии и управление : материалы 56-й научной конференции аспирантов, магистрантов и студентов, Минск, 21-24 апреля 2020 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2020. – С. 13.

3. Жук, А. А. Семантические средства визуализации различных видов графической информации / А. В. Бобков, А. А. Бруцкий, А. А. Жук // Информационные технологии и системы 2017 (ИТС 2017) = Information Technologies and Systems 2017 (ITS 2017) : материалы междунар. науч. конф. (Республика Беларусь, Минск, 25 октября 2017 года) / редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2017. – С. 136 - 137.