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

УДК 004.67

Козуб
Виктор Николаевич

Методы предиктивной аналитики больших объёмов данных

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 04 «Математическое моделирование, численные методы и комплексы программ»

Научный руководитель
Пилецкий И.И.
к.ф.-м.н., доцент

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

Количество данных в мире неумолимо растёт. Эти данные могут принимать разнообразные формы, быть структурированными и неструктурированными, перемещаться и оставаться на месте. При правильном подходе из этих данных можно получить ценную информацию, владея которой, можно преобразовать бизнес-процессы компании.

Объёмы данных становятся настолько большими, что традиционные методы и технологии обработки данных просто не справляются. А данные без обработки – это просто информационный шум и ничего более. Таким образом, ставится задача разработки новых средств, которые смогли бы оперировать большими объёмами данных.

В последнее время всё большую популярность набирают различные облачные платформы как сервис (PaaS). При таком подходе клиент просто загружает исходный код своего приложения в облако провайдера и указывает, какие вычислительные мощности необходимы. Всё остальное берёт на себя провайдер. Такой подход позволяет значительно снизить время на запуск программного продукта и бюджетные затраты.

В данной диссертационной работе подробно изучается возможность создания приложения для предиктивной аналитики больших объёмов данных с использованием подхода PaaS. В качестве облачной платформы используется платформа IBM Bluemix, которая содержит набор сервисов, доступных программисту для использования в своём приложении.

С помощью Bluemix был создан ряд приложений в области Big Data для практической демонстрации современного подхода к работе с большими данными. Приложения используют доступные в Bluemix сервисы для предиктивной аналитики данных.

В данной работе подробно изучается вопрос аналитики данных. Рассматриваются методы и технологии Data Mining. Особое внимание уделяется методам предиктивной аналитики больших объёмов данных как востребованному и перспективному направлению научных исследований. Рассматривается возможность организации аналитики Big Data средствами IBM Watson – суперкомпьютера, часть ресурсов которого доступна программистам в виде сервисов IBM Bluemix.

Незатронутыми остаются вопросы математического изложения рассмотренных методов и алгоритмов, которые, благодаря своей сложности и множественности подходов к проблеме, могли бы стать темой отдельной диссертационной работы и, таким образом, выходят за рамки данной работы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью данной работы является исследование методов предиктивной аналитики больших объёмов данных.

Для достижения поставленной цели в ходе работы необходимо решить следующие *задачи*:

1. Раскрыть содержание понятий «большие данные», «облачные технологии», «предиктивная аналитика».
2. Провести сравнительный анализ методов предиктивной аналитики больших объёмов данных.
3. Определить слабые и сильные стороны существующих методов предиктивной аналитики больших объёмов данных.
4. Сделать обзор популярных программных средств для предиктивной аналитики и продемонстрировать их работу.
5. Показать на примере, как и для каких целей могут использоваться методы предиктивной аналитики больших объёмов данных.

Объектом исследования являются большие данные, методы и технологии их обработки.

Предметом исследования являются методы и программное обеспечение для решения задачи предиктивной аналитики больших данных.

Основной *гипотезой*, положенной в основу данной работы, является возможность использовать различные облачные решения для предиктивной аналитики Big Data без необходимости создания соответствующей инфраструктуры на предприятии, что ускоряет выпуск программного продукта и снижает затраты на производство.

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

Работа выполнялась на базе Академического центра компетенции технологий IBM и соответствует тематике научных исследований, проводимых на кафедре информатики БГУИР.

Личный вклад соискателя

Результаты, приведённые в диссертации, получены соискателем лично. Вклад научного руководителя И.И. Пилецкого заключается в формулировке целей и задач исследования и руководстве ходом работы.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на конференции «BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий» (Минск, май 2015), на 52-й научно-технической конференции аспирантов, магистрантов и студентов БГУИР (Минск, апрель 2016), а также на конференции «BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня» (Минск, июнь 2016).

Результаты работы были использованы при подготовке тренинга по Big Data для студентов кафедры информатики.

Также на основе результатов работы будет разработан курс для обучения магистрантов кафедры информатики современным методам работы с Big Data.

Опубликованность результатов диссертации

По теме диссертации опубликовано 4 печатные работы в сборниках трудов и материалов конференций.

Структура и объём диссертации

Диссертация состоит из общей характеристики работы, введения, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложений.

Общий объем работы составляет 72 страницы, из которых основного текста – 66 страниц, 24 рисунка на 21 странице, список использованных источников из 46 наименований на 4 страницах и 2 приложения на 2 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе диссертационной работы представлен анализ предметной области – Big Data. Даны необходимые определения, выявлены основные существующие проблемы и актуальные задачи, исследован традиционный подход к работе с большими данными и соответствующее программное обеспечение.

Вторая глава посвящена исследованию современного подхода к обработке больших данных – облачным технологиям. Рассмотрены основные понятия, модели предоставления облачных услуг, свойства облачных технологий. Затронуты модели обслуживания облачных технологий: IaaS, PaaS, SaaS. Кратко рассмотрена облачная платформа IBM Bluemix, её идеология, задачи и принцип работы.

Третья глава посвящена обработке и анализу данных. Рассмотрены основные методы и алгоритмы для традиционного анализа данных, задачи и сферы применения Data Mining. Затронута тема предиктивной аналитики данных, кратко рассмотрен облачный сервис для предиктивной аналитики больших объёмов данных – IBM Watson Analytics.

В четвёртой главе подробно рассмотрены основные методы предиктивной аналитики данных: их сильные и слабые стороны, области применения, математическая аргументация. Описана практическая реализация нескольких методов в рамках тренинга по Big Data для студентов кафедры информатики БГУИР. Представлены результаты работы со студентами в ходе тренинга.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Предложен подход в организации обучения студентов технологиям для работы с большими данными и предиктивной аналитики [1-А]. В рамках подхода используется облачная платформа IBM Bluemix, которая предоставляет набор сервисов для развёртывания приложений и организации предиктивной аналитики данных [2-А]. Таким образом, нет необходимости в создании собственного облака или центра обработки данных в учебном заведении [3-А].

2. Проведена апробация предложенного подхода в рамках тренинга по технологиям больших данных, предиктивной аналитики и интернета вещей [1-А]. Тренинг проводился на базе Академического центра компетенции технологий IBM на кафедре информатики факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники. В тренинге приняли участие студенты второго курса кафедры информатики.

3. Разработано несколько приложений, использующих сервисы IBM Bluemix и возможности IBM Watson Analytics [1-А, 4-А].

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки программного обеспечения компьютерных систем для решения задач предиктивной аналитики больших данных с применением облачных вычислений. Они могут быть использованы при разработке новых и совершенствовании существующих приложений в данной области.

2. Полученные результаты формируют теоретическую и практическую базу для создания учебных курсов по обучению студентов вуза технологиям Big Data, предиктивной аналитике и концепции интернета вещей. Они могут быть использованы для модернизации и дальнейшего развития существующих в университете учебных курсов.

3. Результаты работы могут использоваться при подготовке специалистов в области больших данных в университете.

4. Результаты работы могут использоваться при подготовке преподавательского персонала для обучения студентов технологиям Big Data.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Пилецкий, И.И. Облачная платформа IBM Bluemix для тренинга по технологиям Big Data и IoT / И.И. Пилецкий, А.Е. Лещёв, В.Н. Козуб // BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня: сборник материалов междунар. научн.-практ. конф. / редкол.: М.П. Батура [и др.]. – Минск, 2016. – 200 с.

2. Александров, А.А. Исследование возможностей когнитивных сервисов IBM Watson, доступных в рамках облачной платформы IBM Bluemix / А.А. Александров, В.Н. Козуб // Компьютерные системы и сети: материалы 52-й научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2016. – с. 31.

3. Пилецкий, И.И. Виртуальная ИТ среда БГУИР для исследования Big Data и VCL / И.И. Пилецкий, А.Е. Лещёв, В.Н. Козуб // BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сборник материалов междунар. научн.-практ. конф. / редкол.: М.П. Батура [и др.]. – Минск, 2015. – 220 с.

4. Козуб, В.Н. Применение IBM Watson для определения наиболее подходящей специальности абитуриенту / В.Н. Козуб, А.А. Александров // Компьютерные системы и сети: материалы 52-й научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2016. – с. 24-26.