Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.4

ХАДАСЕВИЧ
Александр Иванович

АНАЛИЗ ФУТБОЛЬНЫХ СОБЫТИЙ

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники
по специальности 1-40 81 04 Обработка больших объёмов информации

Научный руководитель
Хотеев Александр Леонидович
канд. физ.-мат. наук, доцент

Минск 2018

ВВЕДЕНИЕ

С каждым годом задачи обработки больших объемов данных все чаще становятся перед разработчиками. В настоящее время генерируются данные о деятельности людей и объектов в огромном количестве и растущем масштабе. Для изучения, поиска закономерностей и анализа этих данных используются специальные инструменты и методы.

Сегодня начинают терять актуальность прежние технологии обработки и анализа данных в следствие того, что объём данных стремительно увеличивается. Разумеется, когда база данных имеет сравнительно небольшой объём, а вычислительных средства справляются с их обработкой, то никаких новых проблем не возникает, но со стремительным ростом объёмов информации появляется необходимость в новых технологиях и методах её обработки. Для рассмотрения данной проблемы был введён термин больших данных – совокупность подходов, инструментов и методов обработки данных огромных объёмов, которые призваны совершать три операции:

– обрабатывать большие по сравнению со «стандартными» сценариями объемы данных;

– уметь работать с быстро поступающими данными в очень больших объемах. То есть данных не просто много, а их постоянно становится все больше и больше;

– они должны уметь работать со структурированными и плохо структурированными данными параллельно в разных аспектах.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Профессиональная футбольная статистика стала развиваться с момента создания самого футбола, постепенно охватывая все новые элементы, описывающие игровой процесс. За всю историю футбола она пережила несколько этапов своего развития – от простого подсчета забитых мячей до сложных теоретических моделей на её основе, а также определения скоростных и дистанционных показателей игроков и другой специфической информации, характеризующей качество индивидуальной и командной игры. Со временем профессиональная футбольная статистика превратилась в составную часть и непременный атрибут современной футбольной аналитики.

Важной вехой развития данного элемента аналитики футбольного матча можно считать введение математической обработки статистического материала, характеризующего действия отдельных игроков. Она учитывала не только количество совершенных технико-тактических действии, но и их позиционные и временные характеристики.

Параллельно с развитием профессионального футбола совершенствовалась и профессиональная футбольная статистика, а также методы её регистрации и аналитика на её основе. В настоящее время простой подсчет тех или иных игровых показателей футбольного матча, выполняющих группой статистов, постепенно отходит на второй план. Он часто не удовлетворяет требованию объективности, так как на него влияет субъективный человеческий фактор. На его место пришли методики фиксации действий игроков с использованием дорогостоящего видеооборудования, технических средств наблюдения и теории распознавания объектов.

Для разработки и реализации был выбран Windows Azure HDInsight – это полностью управляемая облачная служба, которая позволяет быстро, просто и без лишних затрат обрабатывать большие объемы данных. Для разработчиков Windows Azure HDInsight имеет плагин для Visual Studio, который поддерживает создание приложений. Для разработчиков Linux или Windows у HDInsight есть плагины для IntelliJ IDEA и Eclipse, двух очень популярных платформ Java IDE с открытым исходным кодом. HDInsight также поддерживает команды PowerShell, Bash и Windows, позволяющие создавать сценарии рабочих процессов. Также отличительной особенностью является возможность использования не JVM языков с платформой Hadoop.

Цель и задачи исследования

Объект исследования – задачи сбора футбольной событий и построения статистики.

Цель работы – изучение методов обработки больших объемов данных,

применение программной модели MapReduce для разработки приложения для распределенной обработки больших массивов футбольных событий

Методы исследования – изучение технологий обработки больших массивов данных, изучение документации и литературы по рассматриваемым технологиям, анализ реализаций изученных методов.

Результатом является изучение и применение парадигмы MapReduce для обработки больших объемов данных, обзор и изучение экосистемы Hadoop, реализация приложения для распределенной обработки данных на основе платформы Hadooop.

Областью применения являются системы анализа и обработки больших объемов информации, средства статистики.

Структура и объём работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, четырёх глав и заключения, библиографического списка и приложения. Общий объём диссертации – 55 страниц. Библиографический список включает 23 наименование.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы анализа больших объемов данных (хранение огромных объемов информации, обработка потоковой информации, работа с плохо структурированными данными), определены основные направления исследований, а также даётся обоснование актуальности темы диссертационной работы.

В **первом разделе** рассматриваются анализ проблемы обработки футбольных событий. Даются общие сведения об обработке больших объемов информации, обзор элементов сбора больших объемов информации, элементы профессиональной футбольной статистики, существующие аналоги. В конце раздела дается постановка задачи.

Во **втором разделе** рассматриваются технологии анализа больших данных, дается описание платформы Hadoop и программной модели MapReduce.

В **третьем разделе** приводятся реализация методов сбора футбольной статистики, разработка алгоритма построения футбольной статистики.

В **четвёртом разделе** приведены результаты разработки и тестирования программного средства.

ЗАКЛЮЧЕНИЕ

Большие данные, появившиеся как следствие движения общества по информационному пути развития, уже стали частью нашей ежедневной жизни. Почти каждый человек ежедневно генерирует информацию, которая обрабатывается и записывается на различные рода носители. Неудивительно, что правительство и бизнес, в их извечной гонке за эффективностью, крайне заинтересованы в анализе этой информации, что в свою очередь, подогревает интерес разработчиков к данной сфере.

В рамках данной работы был произведен анализ футбольных событий. Применен на практике подход MapReduce в задаче нахождения различной футбольной статистики. Был получен практический опыт работы с HDInsight Emulator и Hadoop и разработано программное средство для работы и анализа распределенных данных. Были найдены и проанализированы различные технические средства, библиотеки и фреймворки для удобной работы с платформой Hadoop. Проанализированы различные методы сбора и извлечения полезной информации из текстовых источников. Дальнейшим развитием данной темы может стать построение различных предиктивных математических моделей, для предсказания будущих результатов матчей. Также дальнейшим развитием является потоковая обработка футбольных событий, и построение в режиме реального времени футбольной статистики.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Обработка больших объёмов информации с использованием платформы Hadoop и службы облачных вычислений Microsoft Azure. Хадасевич А.И., Швец В.И., Хотеев А.А. БГУИР, 2018 г.

2-А. Обработка больших объёмов данных в автоматизированных системах управления автомобильным потоком. Швец В.И., Хадасевич А.И., Теслюк В.Н. БГУИР, 2018 г.