ПРИМЕНЕНИЕ ТЕХНОЛОГИИ BIG DATA НА ОСНОВЕ MAPREDUCE ДЛЯ ПОВЫШЕНИЯ УРОВНЯ УСПЕВАЕМОСТИ ОБУЧАЮЩИХСЯ

Дорофеев Роман Сергеевич,
к.т.н.,
Иркутский национальный исследовательский технический университет,
г. Иркутск
Дорофеев Андрей Сергеевич,
к.т.н., доцент,
Иркутский национальный исследовательский технический университет,
г. Иркутск
Рогачева Софья Андреевна
Иркутский национальный исследовательский технический университет,
магистрант,
г. Иркутск

Высшее образование в ведущих университетах России есть и остается одним из лучших образований в мире. Ежемесячно свою квалификацию повышают десятки тысяч преподавателей, регистрируется множество патентов, изобретений, полезных моделей, заключаются десятки тысяч лицензионных договоров с предприятиями в различных отраслях производства. Миллионы студентов в стране получают наивысшую рейтинговую стипендию за свои достижения в науке и образовании. Но у всего этого есть и оборотная сторона, когда в рассмотрение не берутся студенты, которые не смогли по тем или иным причинам получить хорошие оценки и раскрыть свой потенциал в науке и были отчислены. К таким причинам можно отнести недопонимание различных моментов в читаемых дисциплинах, тем самым потерю интереса к учебе в целом, личные проблемы учащихся. Как известно, отток студентов также негативно влияет и на сами учебные заведения: чем больше отток, тем меньше прибыль и государственная финансовая поддержка. Помимо экономического фактора, процент перехода студентов на следующий курс влияет на позиции в различных рейтингах, вероятность продолжения существования программ и направлений, снижение уровня нагрузки и, как следствие, заработной платы преподавателей.

В данной статье предлагается применение технологии Big Data на основе MapReduce для улучшения уровня успеваемости обучающихся, качества образования, сохранения программ и направлений в учебных заведениях, увеличения количества студентов и выпускников. Целью работы является оперативное выявление отстающих обучающихся на ранней стадии с целью предотвращения ухудшения их уровня успеваемости и дальнейшего оттока. Повышение рейтинга учебных заведений за счет заинтересованности в успеваемости студентов, увеличение прибыли за счет сохранения коммерческих студентов и бюджетных мест. Данная система также позволит определить преподавателей, которые недостаточно уделяют внимания успеваемости студентов по своей дисциплине или имеют огрехи в качестве излагаемого материала.

Термин Big Data появился сравнительно недавно. Google Trends показывает начало активного роста употребления данного словосочетания, начиная с 2011 года. Термин ввёл редактор журнала Nature Клиффорд Линч ещё в 2008 году в спецвыпуске, посвящённом взрывному росту мировых объёмов информации. Хотя, конечно, сами большие данные существовали и ранее. Большие данные (англ. big data) — серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence. В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях — весь мировой объём данных, и вытекающих из этого трансформационных последствий.

Огромные объёмы данных обрабатываются для того, чтобы человек мог получить конкретные и нужные ему результаты для их дальнейшего эффективного применения. Фактически, Big data — это решение проблем и альтернатива традиционным системам управления данными.

Международная консалтинговая компания McKinsey & Company, специализирующаяся на решении задач, связанных со стратегическим управлением, которая в качестве консультанта сотрудничает с крупнейшими мировыми компаниями, государственными учреждениями и некоммерческими организациями, выделила следующие техники и методы анализа, применимые к Big data:

- Data Mining;
- краудсорсинг;
- смешение и интеграция данных;
- машинное обучение;
- искусственные нейронные сети;
- распознавание образов;
- прогнозная аналитика;
- имитационное моделирование;
- пространственный анализ;
- статистический анализ;
- визуализация аналитических данных.

Горизонтальная масштабируемость, которая обеспечивает обработку данных — базовый принцип обработки больших данных. Данные распределены на вычислительные узлы, а обработка происходит без деградации производительности. McKinsey включил в контекст применимости также реляционные системы управления и Business Intelligence.

Применяемые технологии:

- NoSQL;
- MapReduce;
- Hadoop;
- R;
- аппаратные решения.

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах. Модель MapReduce представлена на рисунке 1.

MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 стадии:

1. Стадия Map. На этой стадии данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи.

Функция map(), примененная к одной входной записи, выдаёт множество пар ключ-значение (может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение). Что будет находиться в ключе и в значении – решать пользователю, но ключ – очень важный параметр, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce.

2. Стадия Shuffle. Проходит незаметно для пользователя. На этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти «корзины» послужат входом для reduce.

3. Стадия Reduce. Каждая «корзина» со значениями, сформированная на стадии shuffle, попадает на вход функции reduce().

Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины». Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

Приведем несколько дополнительных фактов о MapReduce:

1) Все запуски функций map и reduce работают независимо и могут работать параллельно, в том числе на разных машинах кластера.

2) Shuffle внутри себя представляет параллельную сортировку, поэтому также может работать на разных машинах кластера. Пункты 1-2 позволяют выполнить принцип горизонтальной масштабируемости.

4) Функция map, как правило, применяется на той же машине, на которой хранятся данные – это позволяет снизить передачу данных по сети (принцип локальности данных).

5) MapReduce – это всегда полное сканирование данных, никаких индексов нет. Это означает, что MapReduce плохо применим, когда ответ требуется очень быстро.

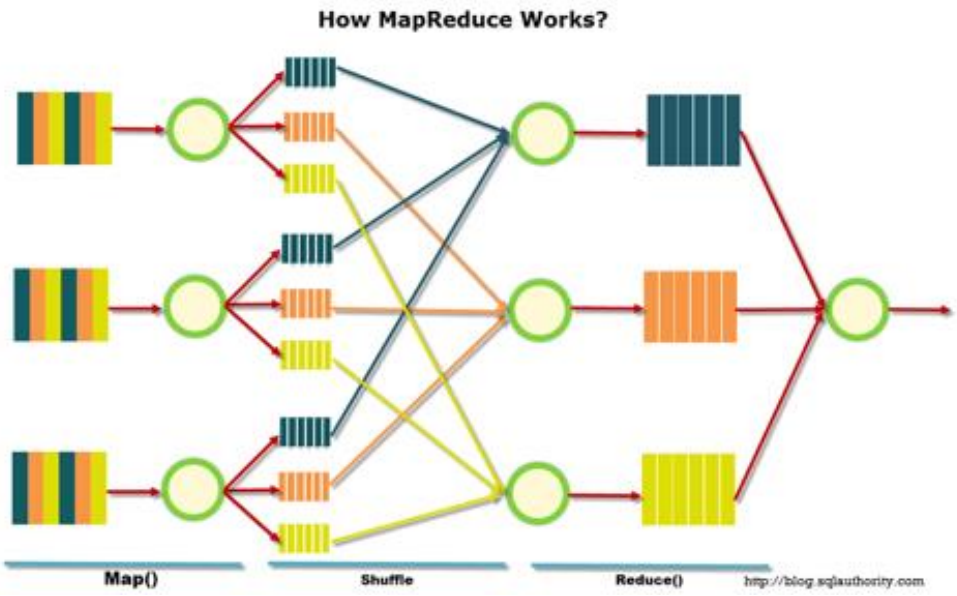

Рисунок 1. Модель MapReduce

Для решения поставленной задачи необходимо регулярно вести учет успеваемости студентов по каждой пройденной теме дисциплины или промежуточной аттестации, для чего в учебном заведении необходимо внедрить онлайн-журнал, в который каждый преподаватель сможет выставлять оценки студентам. Структура журнала может иметь следующий формат:

аименование дисциплины;
ифр группы;
амилия, имя, отчество студента;
ата промежуточной аттестации;
ема аттестации;
ценка по результатам аттестации;
амилия, имя, отчество преподавателя;

В результате ведения такого журнала на выходе получим следующий фрагмент дампа данных для дальнейшей обработки, представленный в таблице 1.

Таблица 1

ФРАГМЕНТ ДАМПА БАЗЫ ДАННЫХ ВЕДЕНИЯ ЖУРНАЛА

name_discipline	name_group	student	data	topic	result	name_teacher
Теория вероятностей	ИВ-18-1	Коганов Семен Игоревич	02.03.2019	Элементы комбинаторики	4	Абракимов Дмитрий Семенович
Математический анализ	АН-18-1	Заруба Татьяна Михайловна	02.03.2019	Свойства функций, непрерывных в точке	3	Лазарова Клементина Ивановна
Сети ЭВМ и системы телекоммуникаций	УТ-18-2	Тарасов Петр Семенович	02.03.2019	Сети кампусов	3	Головасто Виктор Григорьевич
Теория вероятностей	ИВ-18-4	Артемов Никита Андреевич	02.03.2019	Элементы комбинаторики	3	Абракимов Дмитрий Семенович
Программирование	АН-18-2	Газуба Кирилл Игоревич	02.03.2019	Циклы	4	Егорчан Надежда Павловна
Теория систем и системный анализ	ПП-18-1	Марос Людмила Геннадьевна	02.03.2019	Теория игр и принятие решений	3	Тупоров Яков Петрович

Представленный фрагмент содержит данные по студентам с пониженными оценками для лучшего восприятия решаемой задачи, естественно, по вузу данные хранятся по всем студентам со всеми результатами.

Применим MapReduce для поиска студентов, которых будем относить к неуспевающим по теме, т.е. с оценкой ниже 4. Дополнительно посчитаем количество студентов, неуспевающих по определенной дисциплине у преподавателя. В MapReduce можно задавать только пользовательские функции, ниже представлен фрагмент кода на python-like для решения поставленной задачи:

```
def map(record):
name_discipline,name_group,student,data,topic,result,name_teacher
result=int(result)
if result <4: yield name_discipline, name_group, student, topic,result,name_teacher
def reduce(name_discipline, name_group, student, topic,result,name_teacher):
yield name_discipline,name_teacher, count(student)
```

В результате выполнения будет сформирован список неуспевающих по определенным дисциплинам студентов, с которыми необходимо начать работать, предложить дополнительные занятия, репетиторство, повтор материала и др. Подсчет количества неуспевающих студентов по дисциплине и преподавателю при подавляющем количестве неуспевающих студентов относительно общего числа требует задуматься о качестве презентуемого материала и возможной переработки для повышения его усвояемости.

Данный мониторинг позволит увеличить успеваемость студентов и качество излагаемого лекционного и практического материала.

Список литературы

ациональная электронная библиотека им. Н.Э. Баймана. URL: https://ru.bmstu.wiki/MapReduce (дата обращения: 09.03.2019)

фициальный сайт проекта Apache Hadoop. URL: https://hadoop.apache. org/docs/r1.2.1/mapred_tutorial.html (дата обращения: 08.03.2019)

айт компании Rusbase — независимое издание о технологиях и бизнесе, организатор мероприятий и создатель сервисов для предпринимателей, инвесторов и корпораций. https://rb.ru/howto/chto-takoe-big-data/ (дата обращения: 10.03.2019) Официальный сайт Google AI. URL: https://ai.google/research/pubs/pub62 (дата обращения: 08.03.2019)

log SQL Server Performance Tuning Expert and an independent consultant. URL: https://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/ (дата обращения: 10.03.2019)

a
b
r
.
r
u

Н
о
в
о
с
т
н
о
й

с
а
й
т
-
б
л
о
г

/dca/blog/267361/ (дата обращения: 09.03.2019)