СТАТЬИ

УДК 378:37.047

ПОДГОТОВКА СТУДЕНТОВ К РАБОТЕ С БОЛЬШИМИ ДАННЫМИ С ПРИМЕНЕНИЕМ КЛАСТЕРА HADOOP

Абашин В.Г., Жолобова Г.Н., Горохова Р.И., Никитин П.В., Семенов А.М., Зараев Р.Э.
ФГОБУ ВО «Финансовый университет при правительстве РФ», Москва,
e-mail: rigorokhova@fa.ru

Большие данные с каждым годом занимают все большее место в самых различных областях жизни человека. Финансы, экономика, социология, образование, коммуникации – во всех этих сферах в основе получения информации лежат большие данные и их анализ. С каждым годом возникает необходимость в увеличении количества специалистов, владеющих методологией проектирования, анализа, сбора и хранения больших данных. Подготовка кадров для решения задач в данной области является приоритетной задачей системы образования. В данной статье рассматривается один из актуальных вопросов подготовки студентов к решению задач в области анализа данных и машинного обучения. Подготовка рассматривается на основе междисциплинарного подхода, реализуемого на основе метода проектов. В статье представлен проект анализа спортивных мероприятий на основе системы хранения медиафайлов, организации доступа к ним и их дальнейшей обработки. В статье описаны основные этапы обработки информации и проведено исследование свойств системы, среди которых особое внимание уделено скорости чтения/записи, максимальному объему файлов, стоимости за ТБ хранения и другие принятые метрики для систем такого рода. Результатом исследования является оценка эффективности проектного подхода по формированию умений и навыков проведения анализа данных и применения машинного обучения.

Ключевые слова: методика обучения, междисциплинарная интеграция, анализ данных, машинное обучение, метод проектов

PREPARING STUDENTS FOR BIG DATA WITH HADOOP CLUSTER

Abashin V.G., Zholobova G.N., Gorokhova R.I., Nikitin P.V., Semenov A.M., Zaraev R.E.
Financial University under the Government of the Russian Federation, Moscow,
e-mail: rigorokhova@fa.ru

Big data every year takes an increasing place in various areas of human life. Finance, economics, sociology, education, communications – in all these areas, big data and their analysis are the basis for obtaining information. Every year there is a need for an increasing number of specialists who own the methodology of designing, analyzing, collecting and storing big data. Training personnel to solve problems in this area is a priority task of the education system. This article discusses one of the topical issues of preparing students to solve problems in the field of data analysis and machine learning. Training is considered on the basis of an interdisciplinary approach, implemented on the basis of the project method. The article presents a project for the analysis of sports events based on a system for storing media files, organizing access to them and their further processing. The article describes the main stages of information processing and a study of the properties of the system, among which special attention is paid to the read / write speed, the maximum volume of files, the cost per TB of storage and other accepted metrics for systems of this kind. The result of the study is an assessment of the effectiveness of the project approach to the formation of skills in data analysis and the application of machine learning.

Keywords: methods of teaching, interdisciplinary integration, data analysis, machine learning, project method

Без информационных технологий невозможно представить современную жизнь. Устройство современного мира таково, что практически во всех сферах человеческой деятельности, будь то банки, инвестиционные, страховые, телекоммуникационные и торговые компании, медицинские учреждения, государственные органы и прочие организации, накапливается огромное количество информации. Умение эффективно использовать «цифровой след» клиента дает огромные конкурентные преимущества. Общий объем цифровой информации в 2020 г. составил около 40–44 зеттабайтов [1] и с каждым годом её объемы будут только расти. Требуется отметить, что большие данные имеют отличительной чертой гигантские объемы, значительную скорость поступления, а также многообразие самих данных. Обеспечение целостности данных, их доступности, актуальности становится нетривиальными задачами. Сбои в процессах обработки данных ведут к огромным финансовым потерям, поэтому необходимы специалисты, умеющие строить системы хранения данных, очищать и форматировать их, а также настраивать процессы обновления и приёма данных для дальнейшей работы с ними [2].

Новейшие технологии работы с большими данными получили свое развитие в больших компаниях. Так, в 2011 г. большим данным начинают уделять огромное внимание такие крупные корпорации, как Microsoft, Oracle, HP, IBM, Facebook, EMC [3]. В этом же году исследовательская компания Gartner делает вывод, что большие данные являются одним из трендов

в отрасли информационных технологий. Уже с 2013 г. большие данные получают развитие в академической среде, в университетах вводятся курсы по данной дисциплине. При подготовке специалистов данного профиля особенно важно акцентировать внимание на формирование компетенций в области построения эффективных систем аналитики больших данных, организации конвейеров доставки и преобразования нужной информации из множества разных СУБД и файлов различных форматов [4].

Цель исследования – описать методику обучения студентов технологиям обработки больших данных на основе междисциплинарной интеграции и метода проектов и проверить ее эффективность. Объектом исследования в данной работе является система анализа спортивных мероприятий, проблема которой заключается в системе хранения и доступа к большим объемам данных, поэтому предметом исследования будет являться система хранения и доступа к видеофайлам.

Материалы и методы исследования

Применение проектного подхода наиболее успешно в случае решения реально существующей актуальной задачи [5]. В частности, представленная работа посвящена описанию процесса обучения студентов технологиям обработки больших данных на примере реализации системы хранения и доступа к записям трансляций спортивных мероприятий.

При изучении данной проблемы необходимо сформировать понимание у обучающихся, что одной из главных целей обработки большого массива данных является эффективное использование всевозможных видов информации в условиях непрерывного изменения и её прироста в колоссальных объемах. Следует акцентировать внимание на том, что большие данные, в отличие от обычных, имеют другие подходы к обработке информации. На сегодняшний день существует большое количество инструментов для обработки информации:

− RDMBS (Relation Database Management System): MySQL, PostgresSQL, Oracle Database. RDMBS хорошо себя показывают для приложений, однако тяжело масштабируются и используются для аналитики. Данные читаются построчно, хранятся в жесткой структуре. Отвечает принципу атомарности, согласованности, изолированности, стойкости [6].

− Column-oriented DBMS: ClickHouse, Amazon Redshift. Используются для аналитики. Данные хранятся и читаются в колонках, хорошие механизмы сжатия, однако происходит медленная запись [7].

− NoSQL (Not Only SQL): Apache Cassandra, MongoDB, Amazon Dynamo DB. Хорошо масштабируются и могут решать расширенный набор задач. Поддерживают горизонтальную масштабируемость, репликацию [8].

− Object storages: Amazone S3, Google Cloud Storage. Важный вид хранилищ слабо структурированных данных. Отличается высокой доступностью, надежностью хранения, многомерным масштабированием (горизонтальным или вертикальным), неограниченным объемом хранения информации. Подходит для хранения текстовых, фото-, видео-, аудиоданных [9].

− Message brokers: Apache Kafka, Apache Pulsar, Amazon Web Service. Временные хранилища слабо структурированных потоковых данных [10].

Перечисленные системы зачастую являются источниками данных в схеме ETL (Extract-Transform-Load).

Обучающимся предлагается создать систему хранения и доступа к видеофайлам с записями трансляций спортивных событий. Это направление на сегодняшний день является актуальным, поскольку аналитические системы в спорте находятся на начальном этапе развития, так как они не применяют нейросетевые алгоритмы и не используют такие системы хранения информации, как Hadoop и HDFS [11]. В рамках данной работы будет реализована часть, отвечающая за хранение информации для анализа спортивных событий, в частности футбольных матчей.

Рассматриваемая система включает в себя три основополагающие части:

− Поток видеоданных в реальном времени.

− Модели машинного обучения, в которых происходит обнаружение объектов события и извлекается текстовая информация, повествующая, что происходит в данный момент времени.

− Распределенная файловая система, в которую записываются видеоданные, происходит их раскадровка, для дальнейшего обучения моделей и предоставления аналитической информации по конкретному событию.

На первом этапе видео подвергается предобработке, чтобы в модель поступала качественная информация. Данный шаг необходим для увеличения скорости работы модели обнаружения нужных объектов. На следующем этапе модель обнаруживает и классифицирует релевантные объекты события, для футбольного матча – мяч, игроки команды один, игроки команды два, судьи. После чего происходит извлечение текстовой информации о происходящих событиях на кадре. Данная информация по-

ступает в распределенную файловую систему для последующей аналитики события.

Видеоданные в момент поступления в модель также записываются в файловую систему, где будут храниться определенное время. Данный шаг необходим, так как информация о спортивном мероприятия – это довольно дорогостоящий материал, который используется для дополнительного обучения моделей, внедрения новых модификаций в систему, а также для предоставления аналитической информации о том или ином исследуемом спортивном событии.

Стоит обратить внимание обучающихся на такие характеристики хранилищ данных, как доступность и надежность данных; при этом нужно проводить необходимые расчеты.

В ходе исследования студентам будет необходимо получить характеристики работы распределенных файловых систем Hadoop и Amazon S3 с видеоданными, которые относятся к неструктурированным, и сравнить их результаты. Характеристики для исследования: скорость записи данных; скорость чтения данных; объем хранимых файлов; доступность данных.

Скорость чтения (записи) будет определяться как разность с момента передачи файла из (в) распределенную файловую систему до момента его полного получения (появления в системе):

$$read\ /\ write\ file = \frac{\sum_{i=1}^{N}\left(t_{end\_i} - t_{start\_i}\right)}{N},$$

где t_{start_i} – момент начала действия;

t_{end_i} – момент завершения действия;

N – количество итераций для одного файла.

Основной метрикой для определения результатов хранения системы является средняя скорость ввода-вывода информации:

$$Average\ IO\ rate\left(N\right) = \frac{\sum_{i=1}^{N}\frac{file\,size}{\left(t_{end\_i} - t_{start\_i}\right)}}{N},$$

где *file size* – размер объекта;

t_{start_i} – момент начала действия;

t_{end_i} – момент завершения действия;

N – количество итераций для одного файла.

Для проведения исследования студентам необходимо подобрать 5–6 объектов (видеоданных) разного объема, один из которых будет примерно равен размеру одного футбольного матча 1,5 гигабайт.

Для проведения корректного исследования необходимо произвести количество итераций не менее 10, так как на используемые метрики может влиять скорость сети Интернет. Предполагается, что приведенные метрики будут иметь отрицательную корреляцию, метрика скорости ввода-вывода информации будет выше у кластера Hadoop, так как данный инструмент может обрабатывать информацию на нескольких узлах данных (datanodes) [10].

При выполнении исследования используется язык программирования Python, распределенная файловая система Hadoop, облачная система хранения объектов Amazon Web Service Simple Cloud Storage, операционная система Ubuntu [12].

Студентам необходимо произвести работы по развертыванию кластера Hadoop как на локальной машине с использованием системы виртуализации Docker, так и на сервисе Amazon Web Service Elastic Compute Cloud (Amazon EC2). Также для тестирования и сравнения работы по хранению объектов подключиться к сервису Amazon Web Service Simple Cloud Storage (AWS S3).

На первом этапе работы необходимо развернуть на локальной машине кластер Hadoop, состоящий из namenode и трёх datanode.

После получения результатов работы кластера Hadoop на локальной машине необходимо развернуть кластер Hadoop на сервере Amazon с использованием Amazon EC2. В рассматриваемом примере кластер состоит из 4 инстансов: NameNode, DataNode1, DataNode2, DataNode3. Для приведенного в качестве примера сервера рассмотрели датацентр, находящийся в US East (Ohio) (us-easr-2).

Были созданы публичные и секретные ключи для соединения namenode и datanodes без паролей, что облегчает работу с распределенной файловой системой. Порт для работы с Hadoop из командной строки является 50070, для WEB интерфейса 50070.

Далее необходимо провести стресс-тестирование кластера Hadoop. Для этого в ходе дальнейшей работы необходимо развернуть сервис AWS S3, на котором создается пользователь с ключами для подключения к S3 из python. Получены четыре buckets, в один из которых (socker.games.coursjob) нужно разместить файлы для тестирования чтения и записи.

В приведенном примере использовались файлы следующих размеров: Soccer1: 767,1 МБ, Soccer2: 292,7 МБ, Soccer3: 61,0 МБ, Soccer4: 127,6 МБ, Soccer5: 3,9 ГБ, Soccer6: 1,5 ГБ – предполагаемый размер файла для футбольного матча, исключая перерыв на таймы. WEB интерфейс развернутого сервера для кластера Hadoop на Amazon EC2 представлен на рис. 1.

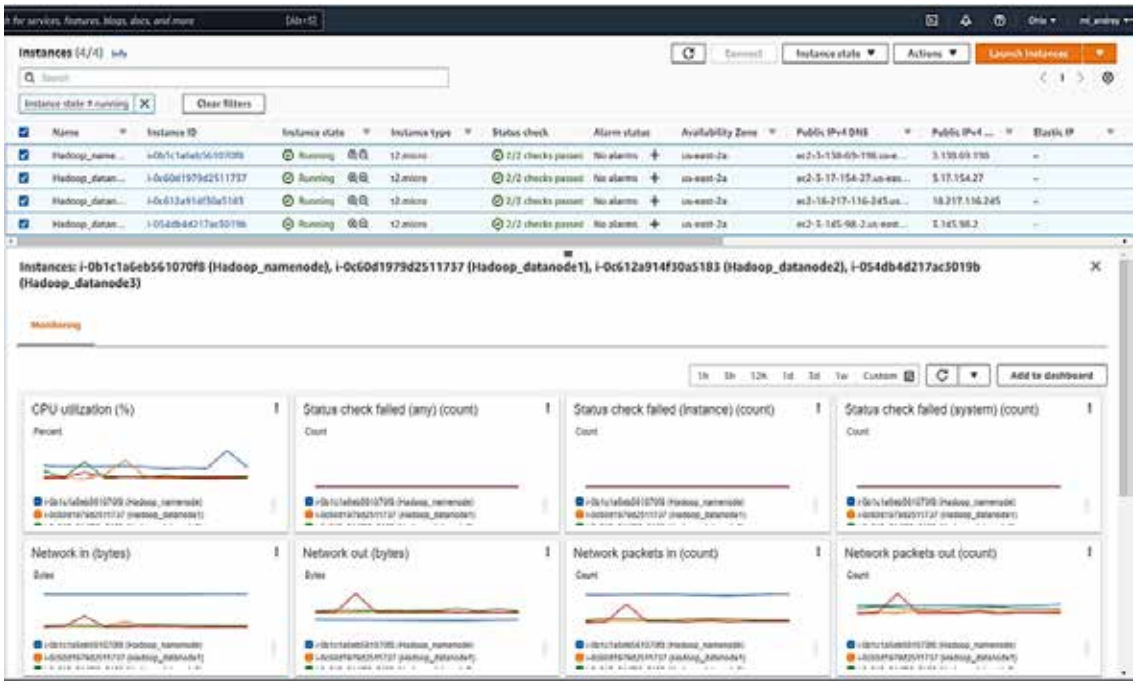

Рис. 1. WEB интерфейс развернутого сервера для кластера Hadoop на Amazon EC2

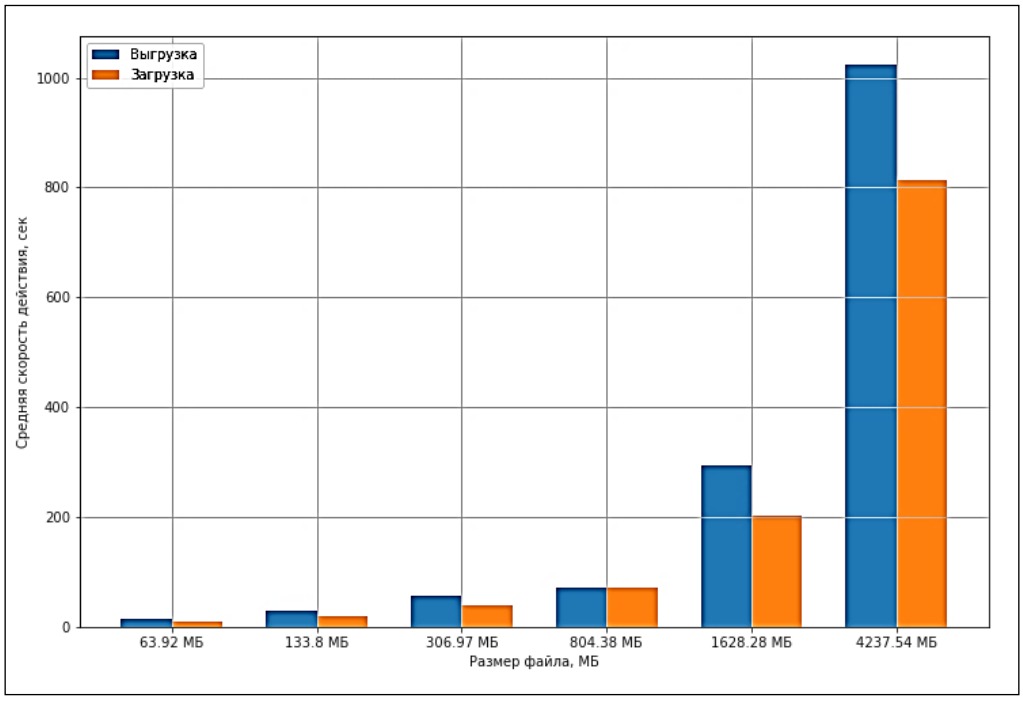

Рис. 2. Скорость выгрузки и загрузки видеообъектов в секундах

В ходе исследования студентам нужно будет сравнивать среднюю скорость чтения (выгрузки) и записи (загрузки), среднюю скорость передачи файлов в итерационном процессе. В рассматриваемой задаче, при тестировании скорости и чтения получено, что скорость выгрузки из хранилища AWS S3 превышает скорость загрузки в хранилище. Обучаемые должны понимать значение данных показателей и по полученным графикам анализировать скорость выгрузки и загрузки видеообъектов в секундах (рис. 2).

После проведённых замеров скорости чтения и записи студентам необходимо провести на основании полученных результатов сравнительный анализ S3 и Hadoop, сделать выводы, разобрать плюсы и минусы использования систем.

Например, выводы могут быть следующими: при выборе между Hadoop и AWS S3 в первую очередь стоит определить, какая задача решается. Распределенная файловая система на основе Hadoop является хорошим инструментом для хранения и распределенной обработки больших объемов данных. Следует добавить, что развертывание и поддержание трудоспособности кластера является затратным процессом. AWS S3 в свою очередь является хорошим инструментом для хранения и доступности данных. К тому же в данной системе предусмотрен механизм версионирования объектов и определения типа хранения объекта, то есть если объект нужен в данный момент, то высокий тип хранения, а если нет, то тип хранения ниже. Механизм типа хранения объектов можно выставлять ниже, он позволяет снизить денежные расходы для хранения большого количества объектов. В дальнейшем стоит хранить данные для их анализа и последующего обучения моделей, которые необходимы для определения футболистов на поле и последующего извлечения текстовой информации, в хранилище S3. Для предварительной обработки видео или кадров из него, если данный процесс будет ресурсозатратным, стоит использовать механизмы распределенных вычислений на кластере Hadoop. Видеоданные в модель для обнаружения объектов футбольного матча будут подаваться по кадрам, значит, стоит хранить видео целиком, для этого подходит сервис AWS S3. В дальнейшем добавятся новые данные, например текстовые (неструктурированный тип данных), которые нужны для обучения модели, извлекающей текстовую информацию из изображения. Для хранения текстовых данных подойдет как кластер Hadoop, так и AWS S3. В проведении последующих исследований не стоит полностью отходить от Hadoop, так как данный инструмент может пригодиться для распределенной обработки видеоданных и текстовой информации.

Заключение

Применение метода проектов и междисциплинарной интеграции показало свою эффективность при проведении занятий по дисциплине «Технологии обработки больших данных» со студентами специальности «Прикладная информатика» в Финансовом университете при Правительстве РФ.

Таким образом, данная методика позволяет студентам понять методологию работы с большими данными от анализа данных до применения машинного обучения. Эффективность доказана следующими факторами:

− у студентов сформированы умения и навыки применения методологии обработки больших данных, что они продемонстрировали при защите проектов;

− курсовые работы по дисциплине были выполнены на основе анализа реальных данных и подтвердили сформированные навыки;

− результаты защиты выпускных квалификационных работ выпускников, связанных с анализом данных и машинным обучением, отмечены высокими баллами.

Междисциплинарная интеграция и метод проектов подтвердили свою эффективность.

Список литературы

1. Гребеник В.В., Воротникова И.В. Тенденции новых цифровых технологий в развитии современного бизнеса // Вестник евразийской науки. 2018. Т. 10. № 3. С. 17.

2. Криштаносов В.Б. Цифровизация финансового сектора экономики: проблемы и перспективы // Труды БГТУ. Серия 5: Экономика и управление. 2021. № 1 (244). С. 17–40.

3. Будзко В.И., Синицын И.Н. Развитие компьютерных информационных технологий «большие данные» // Системы компьютерной математики и их приложения. 2014. № 15. С. 69–75.

4. Шевчук О.О. Эффективность построения прогнозов данных на реляционной системе SQL SERVER и распределенной файловой системе HADOOP // Международный научный журнал «Интернаука». 2017. Т. 1. № 17 (39). С. 74–76.

5. Симонова М.М., Бутырин Г.Н., Бутырина С.А. Проектный подход в процессе подготовки специалистов в современном вузе // Самоуправление. 2019. Т. 2. № 3 (116). С. 297–300.

6. Медведев Д.Н., Медведева Е.Е. Проектирование информационных систем гуманитарного профиля // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2013. № 11 (127). С. 92–98.

7. Бабенко Л.К., Басан А.С., Макаревич О.Б. Анализ производительности разработанной системы разграничения доступа в DBMS // Информационное противодействие угрозам терроризма. 2013. № 20. С. 128–133.

8. Абдышова А.Т. Семантический поиск на файлах с использованием SQL SERVER // Наука и новые технологии. 2014. № 3. С. 25–29.

9. Григорьев Ю.А., Устимов А.И. Сравнение времени выполнения запроса к хранилищу данных в среде MAPREDUCE/HADOOP и СУБД MYSQL // Информатика и системы управления. 2016. № 3 (49). С. 3–12.

10. Беляк А.А., Нестеренков С.Н. Анализ производительности технологии HADOOP // Big Data and Advanced Analytics. 2021. № 7–1. С. 343–346.

11. Гашников М.В., Глумов Н.И., Мясников Е.В., Сергеев В.В., Чернов А.В., Чичева М.А. Программная система для разработки алгоритмов обработки и анализа цифровых изображений // Компьютерная оптика. 2004. № 26. С. 113–115.

12. Абросимова М.А., Власова Л.С. Использование Python при работе с большими данными // Информационные технологии. Проблемы и решения. 2020. № 3 (12). С. 53–59.