МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования «Санкт-Петербургский политехнический
университет Петра Великого»
Институт компьютерных наук и кибербезопасности
Высшая школа технологий искусственного интеллекта
Направление: 02.03.01 Математика и компьютерные науки

Отчёт по дисциплине «Образовательный форсайт»

«Наука о данных и аналитика больших объемов данных»

Студент,
группы 5130201/20102 _____ Гаар В.С.

Преподаватель,
к.т.н., доц. _____ Курочкин М.А.

«_____»_____ 2024 г.

Санкт-Петербург, 2024

Содержание

Введение

В современных условиях работы с большими объемами данных стали неотъемлемой частью практически всех отраслей экономики. Обработка данных больших масштабов требует использования специализированных инструментов и подходов. Курс "Наука о данных и аналитика больших объемов данных" охватывает ключевые аспекты работы с данными, начиная от основ анализа данных и заканчивая современными методами обработки текстов.

Цель курса — сформировать системные знания и практические навыки анализа данных, которые можно применять в разнообразных профессиональных контекстах, включая бизнес, науку и технологии.

1 Постановка задачи

В рамках курса «Образовательный форсайт» было необходимо пройти онлайн-курс «Наука о данных и аналитика больших объемов данных» на портале «Открытое образование» (`https://openedu.ru/`).

Задачей курса является предоставление учащимся комплексного представления о том, как анализировать данные, начиная от их подготовки и заканчивая визуализацией и интерпретацией результатов. В рамках курса:

- Изучался жизненный цикл аналитики данных, от сбора и очистки данных до формирования бизнес-инсайтов.
- Осваивались технологии обработки больших данных, включая Hadoop и MapReduce.
- Осуществлялась работа с инструментами визуализации и анализа текстовой информации.

На протяжении курса выполнялись задания, направленные на практическое закрепление материала, и прошли итоговый тест для оценки уровня усвоения знаний.

2 Аннотация курса и разделов

Курс состоит из семи тематических модулей:

1. **Введение в большие данные**
   На первом этапе курса учащиеся познакомились с основными концепциями и терминами, связанными с большими данными. Рассматривались основные характеристики больших данных (объем, скорость, разнообразие, достоверность и ценность) и их влияние на принятие решений.

2. **Жизненный цикл аналитики данных**
   Учащиеся изучали процесс анализа данных на уровне компаний, включая этапы консолидации, трансформации и загрузки данных (ETL). Особое внимание уделялось задачам интеграции данных из различных источников и проблемам обеспечения их качества.

3. **Высокопроизводительные вычисления**
   Этот модуль был посвящен использованию платформы Hadoop для распределенной обработки данных. Рассматривались история развития технологии, структура Hadoop Distributed File System (HDFS), и процесс MapReduce, обеспечивающий параллельное выполнение задач.

4. **Масштабирование и многоуровневое хранение данных**
   На этом этапе слушатели познакомились с основными принципами горизонтального и вертикального масштабирования, а также изучили подходы к репликации и шардингу данных. Особое внимание было уделено CAP-теореме и ее применению в распределенных системах.

5. **Визуализация данных**
   Рассматривались способы представления данных в виде графиков, диаграмм и инфографики. Учащиеся узнали, как с помощью визуализации можно ускорить процесс принятия решений и выявить скрытые закономерности в данных.

6. **Статистические методы анализа данных**
   Изучались основные статистические подходы, включая проверку гипотез и построение моделей. Учащиеся освоили критерии значимости и регрессии, а также познакомились с параметрическими и непараметрическими методами.

7. **Анализ текстов**
   Этот модуль был посвящен обработке неструктурированных данных. Учащиеся изучили методы анализа текстов, включая извлечение информации, анализ настроений и работу с текстовыми массивами данных.

3 Теоретическая часть курса

В рамках курса "Наука о данных и аналитика больших объемов данных" рассматривались ключевые темы, связанные с анализом больших данных, их обработкой и визуализацией. В теоретической части курса изучались следующие модули:

1. **Введение в большие данные**
   Основные концепции больших данных включали характеристики 3V (объем, скорость, разнообразие), дополненные признаками достоверности и ценности. Рассматривались причины возникновения больших данных, включая развитие технологий и увеличение объемов цифровой информации. Учащиеся узнали о важности использования специальных инструментов, таких как распределенные системы обработки, и о ключевых вызовах работы с большими данными, таких как интеграция и анализ разнородных источников данных

2. **Жизненный цикл аналитики данных**
   Этот модуль был посвящен процессам работы с данными в рамках полного жизненного цикла. Рассматривались следующие этапы:

   - **Сбор данных:** объединение информации из разнородных источников.
   - **Очистка данных:** устранение ошибок, пропусков и дубликатов.
   - **Трансформация и агрегирование:** создание однородных структур для последующего анализа.
   - **Загрузка и анализ:** использование ETL-процессов и бизнес-аналитики для принятия решений. Ключевой акцент был сделан на проблемах низкого качества данных и способах их решения, таких как использование стандартизированных форматов и инструментов обработки.

3. **Высокопроизводительные вычисления**
   Этот модуль рассматривал технологии Hadoop и MapReduce как основу для распределенной обработки данных. Участники узнали о ключевых преимуществах платформы Hadoop:

   - **Масштабируемость:** возможность увеличивать вычислительные мощности через добавление узлов.
   - **Отказоустойчивость:** использование репликации данных для предотвращения потерь.
   - **Эффективность:** параллельная обработка больших объемов данных. Изучалась архитектура Hadoop Distributed File System (HDFS), позволяющая обрабатывать данные размером от гигабайтов до петабайтов, и основы MapReduce для выполнения сложных аналитических задач.

4. **Масштабирование и многоуровневое хранение данных**
   В рамках модуля рассматривались принципы горизонтального и вертикального масштабирования:

   - **Горизонтальное масштабирование:** добавление новых узлов для повышения производительности.

- **Вертикальное масштабирование:** замена существующих компонентов на более мощные. Также изучались подходы репликации (Master-Slave и Peer-to-Peer) и шардинга для распределения нагрузки между серверами. Участники познакомились с CAP-теоремой, определяющей баланс между согласованностью данных, доступностью и устойчивостью к разделению в распределенных системах

5. **Визуализация данных**
   Участники изучали важность визуального представления данных для анализа и принятия решений. Рассматривались основные инструменты визуализации, включая:

   - **Графики и диаграммы:** точечные, линейные, столбчатые и круговые диаграммы.
   - **Инфографика:** сочетание текстов, графиков и изображений для наглядного представления информации.
   - **Дашборды:** интерактивные панели для мониторинга ключевых показателей. Изучались подходы к созданию визуализаций, отвечающих требованиям масштабируемости и интерпретации сложных структур данных

6. **Статистические методы анализа данных**
   Основное внимание было уделено анализу выборок и проверке гипотез. Рассматривались:

   - **Параметрические методы:** t-критерий Стьюдента, критерий Фишера.
   - **Непараметрические методы:** анализ рангов и частот.
   - **Оценка параметров:** точечная и интервальная. Изучались ключевые понятия, такие как ошибки первого и второго рода, уровень значимости, и критерии согласия для оценки достоверности данных

7. **Анализ текстов**
   Этот модуль был посвящен обработке неструктурированных данных. Рассматривались методы извлечения информации, включая:

   - **Анализ настроений:** определение общего мнения на основе текстовых данных.
   - **Категоризация и классификация текстов:** выделение ключевых тематических групп.
   - **Text Mining:** применение методов Data Mining для анализа текстов. Особое внимание уделялось использованию текстовых данных для прогнозирования и выявления закономерностей

4 Результаты аттестации по модулям

На Рис. 1 представлены результаты прохождения итоговой аттестации.

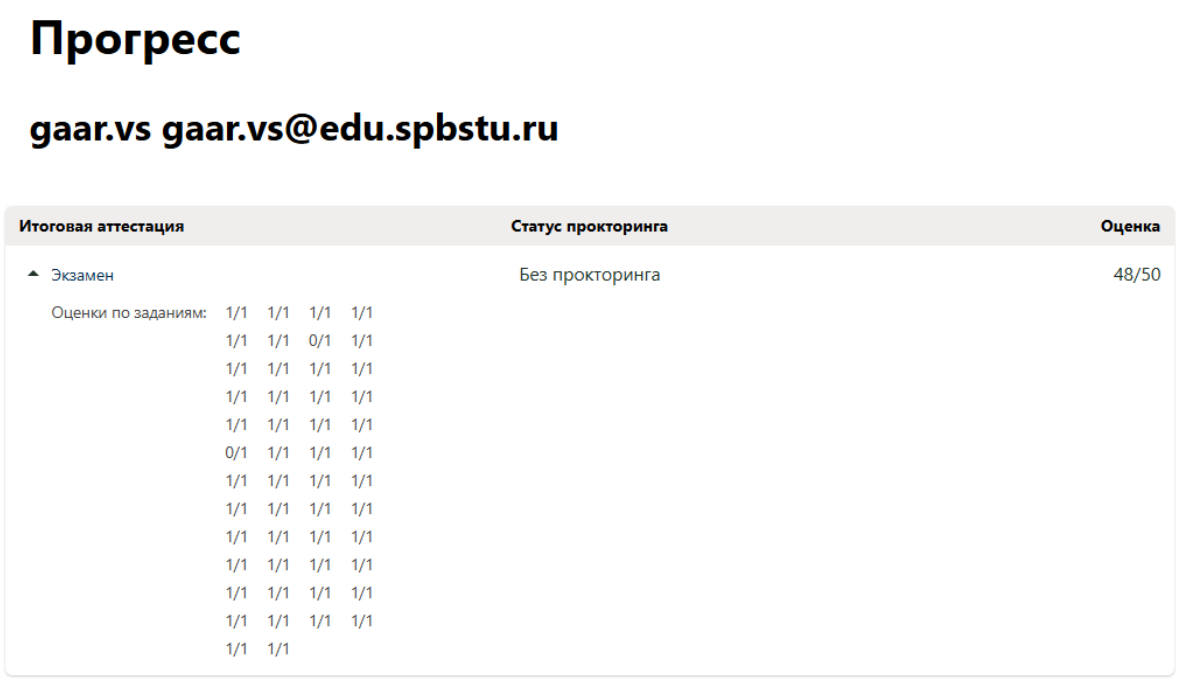

Рис. 1. Результаты прохождения итоговой аттестации

Как видно, итоговый тест был пройден на 48/50 баллов, что соответствует оценке 96% при проходном балле 50%.

Заключение

Пройдя курс "Наука о данных и аналитика больших объемов данных", я получил ценные знания и навыки, которые помогают лучше понимать и решать задачи обработки и анализа данных. Курс охватывал ключевые аспекты работы с данными, включая их сбор, очистку, анализ, визуализацию и текстовую обработку. Каждый из модулей предоставил структурированное понимание современных технологий, таких как Hadoop, NoSQL и методы Text Mining.

Особенно полезным было изучение жизненного цикла аналитики данных и принципов работы распределенных систем, что позволило понять подходы к обработке больших объемов информации. Модуль по визуализации подчеркнул важность представления результатов в наглядной форме, а статистические методы анализа данных дали инструменты для принятия обоснованных решений.

Курс имеет ряд значительных преимуществ: структурированность материала, актуальность рассматриваемых тем и знакомство с передовыми инструментами работы с данными. Однако, как и у любого онлайн-обучения, здесь не хватало взаимодействия с преподавателями и коллегами, а также большего количества практических заданий для закрепления теоретических знаний.

Несмотря на эти ограничения, курс стал важным шагом в моем профессиональном развитии. Полученные знания и навыки я планирую использовать в решении реальных задач анализа данных, что позволит принимать более обоснованные и точные решения.

Список источников

[1] OpenEdu. Наука о данных и аналитика больших объемов данных [Электронный ресурс] URL: https://apps.openedu.ru/learning/course/course-v1:spbstu+BIGDATA+fall_2024/home (дата обращения 10.12.2024).