Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт экономики и менеджмента

УТВЕРЖДАЮ:
Директор Института
экономики и менеджмента
_____ Е.В. Нехода
«_20_» ____04_____ 20_23_ г.

Рабочая программа дисциплины

Большие данные и аналитика

по направлению подготовки

38.04.01 Экономика

Направленность (профиль) подготовки:

«Экономика»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2023

СОГЛАСОВАНО:
Руководитель ОП
_____Н.А. Скрыльникова

Председатель УМК
_____М.В. Герман

Томск – 2023

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций:
– ПК-2 – Способен разрабатывать стратегии управления изменениями в организации.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:
ПК-2.1. Определяет цели и задачи стратегических изменений в организации.

2. Задачи освоения дисциплины

Целью освоения дисциплины «Анализ больших данных» является теоретическая и практическая подготовка студентов к работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут при сборе и анализе больших объемов структурированной или неструктурированной информации, при разработке моделей данных и получении новых знаний.
Задачи освоения дисциплины:

- Освоить понятийный аппарат и инструментарий хранения и анализа больших данных.
- Научиться применять понятийный аппарат больших данных для решения практических задач профессиональной деятельности.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, является обязательной для изучения. Относится к профессиональному модулю «Бизнес аналитика»;

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 2, зачет.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Премодуль (Бизнес-аналитика)», «Премодуль (Вероятностные и статистические методы в бизнес-аналитике)», «Python и R для анализа данных», «Эконометрика».

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
– лекции: 8 ч.;
– практические занятия: 20 ч.;
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Большие данные (введение).

Понятие больших данных. Большие данные в бизнесе. Источники больших данных. Задачи больших данных.

Тема 2. Методики анализа больших данных

Методики анализа больших данных. Визуализация больших данных. Сервисы визуализации больших данных.

Тема 3. Инструменты Больших данных

Инструмент Hadoop Apache и работа с ним. Работа с виртуальной машиной Cloudera.

Тема 4. Технологии хранения и обработки больших данных

Технологии хранения данных. Технологии обработки данных. Файловая система HDFS.

Тема 5. Вычислительное ядро Hadoop

MapReduce. YARN. Решение MapReduce задачи.

Тема 6. Скрипты Pig

Высокоуровневая платформа **Pig**. Использование вычислительного механизма Pig

Тема 7. Базы данных Hadoop.

Базы данных Hadoop SQL и NoSQL. Инструмент SQL Hive.

Тема 8. Озеро данных

Озеро данных. Корпоративное хранилище.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр. Вклад результатов текущего контроля в итоговой оценке по дисциплине составляет – 50 баллов (50%).

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет во втором семестре проводится в комбинированной форме по билетам. Билет содержит теоретический вопрос и задачу, которую необходимо решить соответствующими программными продуктами. Продолжительность зачета 1,5 часа.

Ответ на теоретический вопрос, проверяющий ПК-2.1, дается в развернутой форме.

Результаты зачета определяются в соответствии с балльно-рейтинговой системой – максимум 50 баллов за зачет (50%):

Критерии выставления баллов за зачет:

Баллы	Характеристика
50 баллов	Дан полный и развернутый ответ на вопрос. Задача решена верно и дана обоснованная интерпретация полученных результатов.
20 баллов	Дан неполный или фрагментарный ответ на вопрос. Задача решена верно, но интерпретация полученных результатов не убедительна.

0 баллов	Не дан ответ на вопрос. Задача решена неверно

Итоговая оценка по дисциплине складывается из результатов текущего контроля (50%) и результатов промежуточной аттестации (50%) и составляет максимум 100 баллов.

Механизм перевода результатов балльно-рейтинговой системы в двухбальную шкалу:

Баллы	Итоговая оценка
70-100 баллов	«Зачтено»
Менее 70 баллов	«Не зачтено»

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=16604

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

Примерный перечень теоретических вопросов, проверяющих ПК-2.1:

Понятие «больших данных» (Big Data). Определение и история.

2. Большие данные (Big Data) в бизнесе. Аспекты, функции, задачи.
3. Источники больших данных по отраслям. Преимущества работы с большими данными.
4. Опишите методики анализа больших данных.
5. Методика визуализации. Опишите суть методики, варианты интерпретации, виды и приведите примеры.
6. Дайте характеристику Big Data на мировом рынке.
7. Охарактеризуйте Big Data в России.
8. Подходы и инструменты хранения Больших данных, альтернативные Hadoop HDFS. Опишите доступные инструменты, альтернативные Hadoop HDFS, и основные сценарии их использования.
9. Поясните принцип локальности данных. Какой подход дает больше гибкости и универсальности при работе с данными?
10. Файловая система Hadoop. Принципы хранения данных. Какая используется модель?
11. YARN. Что это за инструмент и для чего используется?
12. Инструмент Hadoop для распределения вычислительных ресурсов кластера. Поясните принцип работы.
13. Опишите концепцию MapReduce. Сформулируйте задачу обработки данных, которую можно решить, используя только Map функцию.
14. Pig. Что это за инструмент и для чего используется?
15. Hive. Что это за инструмент и для чего используется?
16. Большие данные в бизнесе. Перечислите группы и приветите примеры.
17. Большие данные в маркетинге. Преимущества, задачи и цели.
18. Опишите пример задачи относящейся к проблематике Больших данных
19. Дайте определение Data Mining. Какие методики анализа используются в Data Mining?
20. Построение аналитических моделей в памяти
21. Безопасность хранения и использования больших данных.

22. Какие существуют СУБД для хранения Больших данных? В чем отличие реляционных от нереляционных?

23. Решения класса SQL и NoSQL над Hadoop. В чем отличие и особенности?

24. Основные описательные статистики.

25. Поисковые системы в больших данных. Их роль, задачи. Использование ИИ.

26. Методы анализа и обработки данных. Перечислите методы и их особенности.

27. Модели распределенных вычислений MapReduce и Percona server. Перечислите их особенности и преимущества.

28. Озеро данных. Для чего используется? Концепция озера данных?

29. Приведите определение понятию "шкала" в статистических методах анализа данных их различия, информативность и количество допустимых математических действий.

30. Определите различия между параметрическими, непараметрическими и номинальными методами.

31. Опишите основную идею корреляционного анализа. Приведите примеры.

32. Опишите основную идею регрессионного анализа. Приведите примеры.

33. Опишите основную идею дисперсионного анализа. Приведите примеры.

34. Опишите основную идею кластерного анализа. Приведите примеры.

35. Дискриминантный анализ: модель и общая процедура выполнения.

36. Приведите примеры факторного анализа. Цели. Приведите примеры.

37. Программные средства анализа данных. Преимущества и недостатки ПО.

Темы и содержание практических работ

Практическая работа №1 Terminal

1. Выделите и опишите основные преимущества развёртывания кластера Hadoop в «облаке». Составьте краткий отчет.

2. Скачайте образ виртуальной машины Cloudera QuickStart (скачать) предоставленный спонсорами для образовательных целей.

3. Установите и запустите виртуальную машину Cloudera QuickStart. Составьте краткий отчет.

4. Создайте в HDFS рабочую папку "lab1".

5. Произведите загрузку в HDFS всех файлов из архива data_lab1.zip в созданную ранее директорию. Выведите на экран первые 15 строчек файла.

6. Изучите код mkdir.java из вложения hdfs_mkdir.zip. Используя скомпилированный jar-пакет hdfs_client.jar с помощью команды «hadoop jar hdfs_client.jar mkdir [Directory_Path]» создайте рабочую директорию lab1_files. Опишите вывод работы jar-пакета при его корректном и некорректном использовании, а также в случаях, когда директория уже существует.

Практическая работа №2 MapReduce

1. Запустите скомпилированный WordCount.jar пакет используя YARN.

2. Запустите python скрипты mapper.py и reducer.py в виде hadoop-streaming задачи для данных приложенных в архиве.

3. Опишите каким образом необходимо изменить код WordCount.java, чтобы скомпилированный пакет можно было запускать с аргументами входная и выходная директория?

4. Опишите каким образом необходимо изменить код WordCount.java, чтобы результат подсчета частот ошибочно показывал удвоенные значения. Предложите 2 варианта правок: для этапа Map и для этапа Reduce.

Практическая работа №3 Pig Latin

1. Произведите обработку файла 2018.txt или 2019.txt из архива, data_lab3.zip с помощью скрипта Pig latin:
    1. Произведите загрузку.
    2. Извлеките первые 30 строк файла.
    3. Выведите их на экран.
    4. Произведите группировку по признаку DATE.
    5. Произведите анализ усреднения по выделенным группам.
    6. Произведите сортировку результатов.
    7. Выведите на экран 10 строк результата.
2. Повторите операции для файлов из архива lab3_variant.zip согласно вашему варианту. Совместно с усреднением используйте также агрегирующие функции минимума и максимума.

Практическая работа №4 SQL Hive

1. Произведите обработку файла 2018.txt, из архива приложенного к заданию, с помощью скрипта Pig latin:

- Создайте новую базу данных
- Создайте схему реляционной таблицы
- Произведите загрузку данных
- Извлеките первые 10 строк файла
- Создайте view с группировкой по признаку DATE и анализом усреднения по выделенным группам
- Сделайте запрос к view и произведите сортировку результатов
- Сохраните результат в таблице

2. Повторите операции для других файлов архива, согласно своему варианту. Совместно с усреднением используйте также агрегирующие функции минимума и максимума.

Практическая работа №5. Итоговое задание

В этом задании вам нужно продемонстрировать умение использования компонент Hadoop:

- HDFS
- MapReduce
- Pig
- Hive

1. Выберите набор табличных данных и сохраните его (например, с помощью MS Excel) в текстовый формат (CSV). Это могут быть данные, связанные с Вашей деятельностью, открытые данные, модельные данные.
2. Произведите исследование по плану:

- Выберите вариант инфраструктуры Hadoop
- Произведите загрузку данных
- Проведите обработку данных средствами Hadoop
- Предложите варианты по обогащению (расширению набора признаков) имеющихся данных.

Приложите краткий отчет, содержащий описание данных, проблематику их накопления и обработки, шаги исследования (по плану выше), вывод.

в) Методические указания по организации самостоятельной работы студентов:
Самостоятельная работа магистрантов включает в себя:

- самостоятельную подготовку к занятиям по заявленным темам курса в соответствии с содержанием дисциплины и литературой. Контроль выполнения производится на занятиях в блиц-опросах;
- самостоятельную работу в аудитории при ответах на вопросы, решении задач. Контроль выполнения осуществляется сразу же при оценке полученных результатов.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Макшанов А. В. Большие данные. Big Data : учебник для вузов / Макшанов А. В.,Журавлев А. Е.,Тындыкарь Л. Н.. - Санкт-Петербург : Лань, 2022. - 188 с.. URL: https://e.lanbook.com/book/198599.

– Просто о больших данных : пер. с англ. / Джудит Гурвиц, Алан Ньюджент, Ферн Халпер, Марсия Кауфман ; Сбербанк. - Москва : Эксмо, 2016. - 393, [2] с.: ил. - (Библиотека Сбербанка ;т. 58:)

– Ын А. Теоретический минимум по Big Data : все, что нужно знать о больших данных / Анналин Ын, Кеннет Су. - Санкт-Петербург [и др.] : Питер, 2019. - 205 с.

– Дейтел П. Д. Python : искусственный интеллект, большие данные и облачные вычисления / Пол Дейтел, Харви Дейтел. - Санкт-Петербург [и др.] : Питер, 2020. - 861 с.: табл., ил.

б) дополнительная литература:

– Wadkar S. Pro Apache Hadoop / by Sameer Wadkar, Madhu Siddalingaiah. // Springer eBooks. URL: http://dx.doi.org/10.1007/978-1-4302-4864-4

– Дадян Э. Г. Методы, модели, средства хранения и обработки данных : учебник / Э.Г. Дадян, Ю. А. Зеленков; Финансовый университет при Правительстве Российской Федерации. - Москва : Вузовский учебник, 2022. - 168 с.. URL:http://znanium.com/catalog/document?id=384994

– Методы и модели исследования сложных систем и обработки больших данных: монография / И. Ю. Парамонов, В. А. Смагин, Н. Е. Косых, А. Д. Хомоненко. - Санкт-Петербург : Лань, 2020. - 236 с.. URL: https://e.lanbook.com/book/126938

– Замятин А. В. Интеллектуальный анализ данных : учебное пособие : [для студентов университетов и втузов] / А. В. Замятин ; Нац. исслед. Том. гос. ун-т. - Томск : Издательский Дом Томского государственного университета, 2020. - 193 с.: ил., табл.. URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000722107

в) ресурсы сети Интернет:

– Apache Hadoop - https://hadoop.apache.org

– Apache Flume - https://flume.apache.org

– Apache Hadoop Ecosystem - https://www.cloudera.com/products/open-source/apache-hadoop.html

– Официальный сайт Федеральной службы государственной статистики РФ - https://rosstat.gov.ru/

– Официальный сайт Всемирного банка - www.worldbank.org

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

Для проведения лекционных и практических занятий необходимо лицензионное обеспечение: Операционная система Windows 7-10 или Linux, Adobe Acrobat Connect. Офисный пакет Microsoft Office 2013 или OpenOffice.

Для проведения практических занятий необходимо лицензионное программное обеспечение: ОС Windows 7-10 или Linux, свободно-распространяемый программный продукт виртуализации для операционных систем Microsoft Windows или Linux Oracle VM VirtualBox.

– Cloudera CDH;

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения практических занятий, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Погуда Алексей Андреевич, доцент кафедры информационного обеспечения инновационной деятельности факультета инновационных технологий, кандидат технических наук.