

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

Факультет инновационных технологий

УТВЕРЖДАЮ:

Декан

 С. В. Шидловский

«27» августа 2022 г.

Рабочая программа дисциплины

Анализ больших данных

по направлению подготовки

27.03.02 Управление качеством

Направленность (профиль) подготовки:

Управление качеством в производственно-технологических системах

Форма обучения

Очная

Квалификация

Бакалавр

Год приема

2022

Код дисциплины в учебном плане: Б1.В.03

СОГЛАСОВАНО:

Руководитель ОПОП

 В.И. Сырянкин

Председатель УМК

 О.В. Вусович

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-7 – Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности.

ОПК-8 – Способен осуществлять критический анализ и обобщение профессиональной информации в рамках управления качеством продукции, процессов, услуг.

ПК-1 – Способен анализировать причины снижения качества продукции (работ, услуг) и разрабатывать предложения по их устранению.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-7.1 Понимает принцип работы современных информационных технологий.

ИОПК-7.2 Знает и способен применять современные программные платформы в области профессиональной деятельности.

ИОПК-8.1 Владеет методами сбора и анализа информации в области управления качеством продукции, процессов, услуг.

ИОПК-8.2 Владеет методами оценки профессиональной информации.

ИПК-1.1 Умеет собирать, систематизировать и анализировать данные по показателям качества, характеризующим разрабатываемую и выпускаемую продукцию (работы, услуги), в том числе, с использованием средств и технологий цифровизации.

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:

приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;

применение статистических и математических методов для анализа больших объемов информации;

приобретение практических навыков работы с VM Cloudera Hadoop.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, является обязательной для изучения.

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

Четвертый семестр, зачет.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

Успешное овладение данной дисциплиной предполагает предварительные теоретические знания и практические навыки, полученные в дисциплине «Информатика и программирование».

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 2 з.е., 72 часа, из которых:

– лекции: 16 ч.;

– практические занятия: 16 ч.;

в том числе практическая подготовка 8 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

Hadoop Apache. 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. 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.

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.

8.1. Примерный перечень лабораторных работ/ семинарских занятий

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, выполнения домашних и практических заданий, и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет в четвертом семестре проводится в письменной форме по билетам. Билет содержит три теоретических вопроса. Продолжительность зачета 1,5 часа.

Примерный перечень теоретических вопросов

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 приведите 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. Поискковые 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 Person 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.

Результаты зачета определяются оценками «зачтено», «не зачтено».

Для аттестации обучающихся на соответствие их персональных достижений создан фонд оценочных средств по дисциплине, включающий оценочные и методические материалы, позволяющие оценивать знания, умения, навыки и уровень приобретенных компетенций.

Типовые контрольные задания, используемые для оценки результатов обучения и характеризующие этапы формирования соответствующих компетенций, представлены в фонде оценочных средств.

Каждая лабораторная работа выполняется в соответствии с методическими рекомендациями, приложенными к конкретной лабораторной работе.

Критерии оценивания

| Оценка | Характеристика ответа |
|------------|--|
| Зачтено | Работа выполнена полностью. Студент владеет теоретическим материалом, отсутствуют ошибки при описании теории, формулирует собственные, самостоятельные, обоснованные, аргументированные суждения, допуская незначительные ошибки на дополнительные вопросы |
| Не зачтено | Работа выполнена полностью. Студент не владеет теоретическим материалом, допуская грубые ошибки, испытывает затруднения в формулировке собственных суждений, не способен ответить на дополнительные вопросы |

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» – <https://moodle.tsu.ru/course/view.php?id=34401>

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

в) План практических занятий по дисциплине.

г) Методические указания по проведению лабораторных работ.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Кабанова Т.В., Марголис Н.Ю. Прикладная статистика: учебно-методическое пособие. – Томск: Издательство ТГУ, 2007. – 104 с.

– Гмурман В.Е. Теория вероятностей и математическая статистика: учебное пособие для бакалавров. - М.: Юрайт: ИД Юрайт, 2012. – 479 с.

– Просто о больших данных : пер. с англ. / Джудит Гурвиц, Алан Ньюджент, Ферн Халпер, Марсия Кауфман ; Сбербанк. - Москва : Эксмо, 2015. - 393, [2] с.: ил. - (Библиотека Сбербанка ; т. 58:) . URL: <http://sun.tsu.ru/limit/2016/000553356/000553356.pdf>.

– Макшанов А. В. Большие данные. Big Data / Макшанов А. В., Журавлев А. Е., Тындыкарь Л. Н.. - Санкт-Петербург: Лань, 2021. - 188 с.. URL: <https://e.lanbook.com/book/165835>.

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

– Дадян Э. Г. Методы, модели, средства хранения и обработки данных : учебник : [для бакалавров и магистрантов всех специальностей, аспирантов] / Э. Г. Дадян, Ю. А. Зеленков ; Финансовый ун-т при Правит. Рос. Фед.. - Москва : Вузовский учебник [и др.], 2017. - 167, [1] с.: рис., табл.

б) дополнительная литература:

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

– Л. Nadoor в действии. / Чак Л.. - Москва : ДМК Пресс. - 424 с.

– Воронов . В. Data Mining - технологии обработки больших данных : учебное пособие / В. И. Воронов, Л. И. Воронова, В. А. Усачев. - Москва : Московский технический университет связи и информатики, 2018. - 47 с.

– Воронов . В. Data Mining - технологии обработки больших данных : учебное пособие / В. И. Воронов, Л. И. Воронова, В. А. Усачев. - Москва : Московский технический университет связи и информатики, 2018. - 47 с.

в) ресурсы сети Интернет:

– ЭБС «Лань» <https://e.lanbook.com/>.

– ЭБС «Консультант студента» <https://www.studentlibrary.ru/>.

– ЭБС «Юрайт» <https://urait.ru/>.

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

– Статьи по Big Data [Электронный ресурс]. URL: <http://vital.lib.tsu.ru/vital/access/manager/Repository?query=Big+data&queryType=vitalDismax&x=0&y=0> (дата обращения: 11.07.2023).

– ВІ-решений и проектов [Электронный ресурс]. URL: <http://www.tadviser.ru/index.php/BI> (дата обращения: 11.07.2023).

– Большие данные (Big Data) [Электронный ресурс]. URL: [http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_\(Big_Data\)](http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data)) (дата обращения: 11.07.2023).

– In-Memory Computing [Электронный ресурс]. URL: [http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:In-Memory_Computing_\(%D0%92%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2_%D0%BE%D0%BF%D0%B5%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B9_%D0%BF%D0%B0%D0%BC%D1%8F%D1%82%D0%B8\)](http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:In-Memory_Computing_(%D0%92%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2_%D0%BE%D0%BF%D0%B5%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B9_%D0%BF%D0%B0%D0%BC%D1%8F%D1%82%D0%B8)) (дата обращения: 11.07.2023).

13. Перечень информационных технологий

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.

Браузер Google Chrome/Opera/Firefox для работы в электронном курсе Moodle.

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, 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.

Для проведения практических занятий групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации необходима аудитория, оснащенная оборудованием и техническими средствами обучения: компьютер преподавателя (ноутбук), персональные студенческие компьютеры с подключением к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду НИ ТГУ.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечивающие доступ к электронной образовательной среде НИ ТГУ.

15. Информация о разработчиках

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