МИНОБРНАУКИ РОССИИ
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**
**«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**
**(ФГБОУ ВО «ВГУ»)**

УТВЕРЖДАЮ
Заведующий кафедрой
*Математических методов исследования операций*
*Азарнова Т.В.*
29.05.2023 г

# РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
## Б1.В.01 Основы Big Data

**1. Код и наименование направления подготовки / специальности:**

**38.04.05 Бизнес-информатика**

**2. Профиль подготовки / специализация/магистерская программа:**

**Информационная бизнес-аналитика**

**3. Квалификация (степень) выпускника:** магистр

**4. Форма обучения:** заочная

**5. Кафедра, отвечающая за реализацию дисциплины:** математических методов исследования операций

**6. Составители программы:** Ухлова В.В., к.ф.-м.н, доцент кафедры математических методов исследования операций

**7. Рекомендована:** НМС факультета Прикладной математики, информатики и механики № 7 от 26.05.2023

**8. Учебный год:** 2024/2025, 2025/2026 **Триместр(ы):** 5, 6

## 9. Цели и задачи учебной дисциплины

Цели дисциплины: освоение основных технологий, методов и алгоритмов работы с большими массивами данных, которые позволяют обрабатывать, анализировать, интерпретировать, оформлять и представлять профессиональному обществу результаты исследований, позволяют разрабатывать профессионально--ориентированные информационные системы с учетом возможностей современных интеллектуальных информационных технологий.

Задачами курса являются:

- знакомство с основными процессами консолидации, анализа, обработки и управления больших данных;
- изучение и совершенствование методов, алгоритмов и инструментальных средств работы с большими данными для возможности проведения аналитических исследований в рамках профессиональной деятельности;
- освоение основных навыков ведения проектов в области больших данных, в том числе, по созданию и внедрению профессионально-ориентированных информационных систем с учетом возможностей современных интеллектуальных информационных технологий.

## 10. Место учебной дисциплины в структуре ОПОП:

Дисциплина относится к обязательным дисциплинам вариативной части программы обучения. Для изучения курса необходимы знания в области ИТ-технологий, в частности, по обработке, хранению и визуализации данных.

## 11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения

| Код | Название компетенции | Код(ы) | Индикаторы(ы) | Планируемые результаты обучения |
|---|---|---|---|---|
| ПК-3 | Способен проводить обработку и анализ больших данных на базе современных языков программирования и пакетов прикладных программ моделирования | ПК-3.1 | Организует сбор данных и проводит аналитическое исследование в соответствии с согласованными требованиями | Знать: основные технологии консолидации, обработки и управления большими данными, позволяющие осуществлять поиск, сбор и хранение информации из открытых источников и специализированных баз данных; основные методологии анализа данных; алгоритмы обработки данных. основные методики исследования и испытания разработанных методов, моделей, алгоритмов, технологий и инструментальных средств по работе с данными. Уметь: осуществлять информационный поиск с использованием открытых источников информации и специализированных баз данных; |
| | | ПК-3.2 | Разрабатывает и совершенствует методы анализа массовых количественных и нечисловых данных на базе современных языков программирования и технологий управления данными | |
| ПК-4 | Способен управлять разработкой профессионально--ориентированных информационных систем с учетом возможностей современных | ПК-4.3 | Организует работы по созданию и внедрению профессионально-ориентированных информационных систем с учетом возможностей современных интеллектуальных информационных | |

| | | | | |
|---|---|---|---|---|
| | интеллектуальных информационных технологий | | технологий | использовать инструментальные средства для работы с данными, в том числе, с большими данными; проводить исследования и испытания методов, моделей, алгоритмов и инструментальных средств работы с большими данными. Владеть навыками инсталляции и настройки ПО для работы с большими данными. |

## 12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом —4/144

## Форма промежуточной аттестации *экзамен*.

## 13. Трудоемкость по видам учебной работы

| Вид учебной работы | Трудоемкость (часы) | | | | |
|---|---|---|---|---|---|
| | Всего | В том числе в интерактивной форме | По семестрам/ сессиям | | |
| | | | № сессии. 5 | № сессии 6 | ….. |
| Аудиторные занятия | | | | | |
| в том числе: лекции | 8 | 8 | - | 8 | |
| практические | | - | | | |
| лабораторные | 8 | 6 | 2 | 8 | |
| Самостоятельная работа | 119 | 54 | 65 | 119 | |
| Форма промежуточной аттестации | 9 | 0/0 | 0/9 | 0/9 | |
| Итого: | 144 | 68 | 76 | 144 | |

## 13.1. Содержание дисциплины

| № п/п | Наименование раздела дисциплины | Содержание раздела дисциплины | Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК |
|---|---|---|---|
| **1. Лекции** | | | |
| 1.1 | Понятие Data science и Big Data, область применения. | Термины и определения. Особенности технологий. Сферы применения, состояние и перспективы развития. | Основы технологий Big Data (38.04.05.) |
| 1.2 | Технологии консолидации, обработки и управления большими данными | Платформа Hadoop: архитектура и принцип работы. Организация файловой системы HDFS. Программный интерфейс Map Reduce. Система YARN. | |
| 1.3 | Основные процессы в Data science | Процессы сбора, подготовки и исследования данных. Методы моделирования данных. Визуализация данных. | |
| 2***. Лабораторные работы*** | | | |
| 2.1 | Методы работы с данными | Определение целей исследования, формирование ТЗ, выбор методов реализации. Сбор данных. | Основы технологий Big |

| | | | |
|---|---|---|---|
| | | Проверка качества данных. Очистка данных. Выбор средств хранения данных. Методы обработки и анализа данных. Инструменты управления данными. | Data(38.04.05.) |

## 13.2. Темы (разделы) дисциплины и виды занятий

| № п/п | Наименование темы (раздела) дисциплины | Виды занятий (часов) | | | | |
|---|---|---|---|---|---|---|
| | | Лекции | Практи-ческие | Лабораторные | Самостоятельная работа | Всего |
| 1 | Понятие Data science и Big Data, область применения | 2 | - | - | 6 | 8 |
| 2 | Основные процессы в Data science | 2 | - | - | 20 | 22 |
| 3 | Технологии консолидации, обработки и управления большими данными | 4 | - | - | 42 | 37 |
| 4 | Методы работы с данными | | - | 8 | 60 | 78 |
| Контроль | | | | | | 9 |
| | Итого: | 8 | - | 8 | 128 | 144 |

## 14. Методические указания для обучающихся по освоению дисциплины

Дисциплина реализуется по тематическому принципу, каждая тема представляет собой завершенный раздел курса. На первом занятии студент получает информацию для доступа к комплексу учебно-методических материалов.

Лекционные занятия посвящены рассмотрению теоретических основ дисциплины, вводятся основные понятия, изучаются базовые технологии, разбираются основные процессы работы с большими данными.

Лабораторные работы предназначены для формирования умений и навыков, закрепленных компетенциями по ОПОП. Они организовываются в виде выполнения отдельных заданий.

Самостоятельная работа студентов включает в себя проработку учебного материала лекций, разбор заданий лабораторных работ, подготовку к экзамену. Для успешного освоения дисциплины рекомендуется подробно конспектировать лекционный материал, просматривать презентации по соответствующей теме.

При использовании дистанционных образовательных технологий и электронного обучения следует выполнять все указания преподавателя по работе на LMS-платформе, своевременно подключаться к online-занятиям, соблюдать рекомендации по организации самостоятельной работы.

## 15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

а) основная литература:

| № п/п | Источник |
|---|---|
| *1* | Основы технологий Big Data [Электронный ресурс] : учебное пособие / Воронеж. гос. ун-т / В.В. Ухлова .— Электрон. текстовые дан. — Воронеж : Издательский дом ВГУ, 2020 .— Загл. с титула экрана .— Свободный доступ из интрасети ВГУ .— Текстовый файл .— |

| | |
|---|---|
| | <URL:http://www.lib.vsu.ru/ >. |
| *2* | Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2022. — 188 с. — ISBN 978-5-8114-9690-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/198599 (дата обращения: 20.03.2022). — Режим доступа: для авториз. пользователей. |
| *3* | Макшанов, А. В. Современные технологии интеллектуального анализа данных : учебное пособие для спо / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — Санкт-Петербург : Лань, 2020. — 228 с. — ISBN 978-5-8114-5451-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/149343 (дата обращения: 20.03.2022). — Режим доступа: для авториз. пользователей. |

б) дополнительная литература:

| № п/п | Источник |
|---|---|
| *4* | Макшанов, А. В. Системы поддержки принятия решений : учебное пособие для вузов / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2021. — 108 с. — ISBN 978-5-8114-8489-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/176903 (дата обращения: 05.06.2022). — Режим доступа: для авториз. пользователей. |
| *5* | Юре, Л. Анализ больших наборов данных / Л. Юре, Р. Ананд, Д. У. Джеффри ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/93571 (дата обращения: 20.03.2022). — Режим доступа: для авториз. пользователей. |

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет)*:

| № п/п | Ресурс |
|---|---|
| *6* | Электронно-библиотечная система «Лань» - Режим доступа: https://e.lanbook.com |
| *7* | Электронный каталог Научной библиотеки Воронежского государственного университета. – Режим доступа: http//:www.lib.vsu.ru. |
| *8* | Основы технологий Big Data (38.04.05, Ухлова В.В.)/ В.В. Ухлова. — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=5525 |

**16. Перечень учебно-методического обеспечения для самостоятельной работы**

Самостоятельная работа обучающегося должна включать в себя подготовку к тестированию, лабораторным занятиям и подготовку к промежуточной аттестации. Для обеспечения самостоятельной работы студентов в электронном курсе дисциплины на образовательном портале «Электронный университет ВГУ» сформирован учебно-методический комплекс, который включает в себя: программу курса, учебные пособия и справочные материалы, методические указания по выполнению лабораторных работ. Студенты получают доступ к данным материалам на первом занятии по дисциплине.

**17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):**

При реализации дисциплины используются следующие образовательные технологии: логическое построение дисциплины, обозначение теоретического и практического компонентов в учебном материале. Применяются разные типы лекций (вводная, обзорная, информационная, проблемная). Дисциплина реализуется с применением информационно-коммуникационных технологий.

Информационно-коммуникативные технологии для реализации учебной дисциплины:
- технологии синхронного и асинхронного взаимодействия студентов и преподавателя посредством служб (сервисов) по пересылке и получению электронных сообщений, в том числе, по сети Интернет;

- сервис электронной почты для оперативной связи преподавателя и студентов.

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий, для организации самостоятельной работы обучающихся используется онлайн-курс, размещенный на платформе Электронного университета ВГУ (LMS moodle), а также другие Интернет-ресурсы, приведенные в п.15в.

## 18. Материально-техническое обеспечение дисциплины:

Лекционная аудитория должна быть оборудована учебной мебелью, компьютером, мультимедийным оборудованием (проектор, экран, средства звуковоспроизведения), допускается переносное оборудование.

Лабораторные работы должны проводиться в специализированной аудитории, оснащенной учебной мебелью и персональными компьютерами с доступом в сеть Интернет (компьютерные классы, студии), мультимедийным оборудованием (проектор, экран, средства звуковоспроизведения), Число рабочих мест в аудитории должно быть таким, чтобы обеспечивалась индивидуальная работа студента на отдельном персональном компьютере.

Для самостоятельной работы необходимы компьютерные классы, помещения, оснащенные компьютерами с доступом к сети Интернет b платформе Электронного университета ВГУ (LMS moodle).

.

Программное обеспечение:
- ОС Windows 10, ОС Linux
- пакет стандартных офисных приложений для работы с документами, таблицами и т.п. (МойОфис, LibreOffice);
- ПО Adobe Reader;
- специализированное ПО (ПО MatLab);
- интернет-браузер (Mozilla Firefox).

## 19. Фонд оценочных средств:

| № п/п | Наименования раздела дисциплины | Компетенция(и) | Индикатор(ы) достижения компетенции | Оценочные средства |
|---|---|---|---|---|
| 1 | Понятие Data science и Big Data, область применения. | ПК-3 | ПК-3.2 | Контрольная работа |
| 2 | Технологии консолидации, обработки и управления большими данными. | ПК-3 | ПК-3.1, ПК-3.2 | Лабораторная работа 1, Контрольная работа |
| 3 | Основные процессы в Data science. | ПК-3 | ПК-3.1 | Лабораторная работа 2 |
| 4 | Методы работы с данными | ПК-4 | ПК-4.3 | Лабораторная работа 3 |
| Промежуточная аттестация, форма контроля - экзамен | | | | Тест |

## 20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

### 20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

- контрольные работы,
- тест,
- лабораторные работы.

**Перечень заданий контрольной работы**

Для исходного набора данных:

1) выполнить описание «идеальных» данных (тип данных, ограничения, шаблон и т.п);
2) привести варианты возможных ошибок в данных;
3) составить алгоритм повышения качества данных;
4) продемонстрировать траекторию изменения данных при использовании разработанного алгоритма;
5) составить рекомендации, позволяющие получать исходный набор данных с более высоким качеством.

Технология проведения

В качестве исходных данных студент берет любой набор из открытых источников (в формате xls/xlsx (количество записей должно быть более 50, атрибутов более 10). Если качество данных набора очень высокое, то искусственно «ухудшает» его.

Выполнение задания предусматривает использование информации из учебной и справочной литературы, а также ресурсов сети Интернет.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если работа выполнена в полном объеме (приведены все расчеты и они правильные, даны пояснения);
- оценка «хорошо» выставляется студенту, если работа выполнена полностью, но имеются незначительные ошибки;
- оценка «удовлетворительно» выставляется студенту, если работа выполнена полностью, но в представленной части много ошибок или представлена часть работы и она без ошибок;
- оценка «неудовлетворительно» выставляется студенту, если работа не выполнена.

**Перечень заданий для лабораторных работ.**

Лабораторная работа №1

Пример задания.

Выполнить расчет хранилища данных для системы офисной системы видеонаблюдения. Параметры системы видеонаблюдения: 5 камер, разрешение 2.1, 1920x1080, частота 12к/с, кодек H.264.Период хранения данных составляет 3 месяца,

Технология проведения

Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Время выполнения задания составляет 3 часа. Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если работа выполнена в полном объеме (приведены все расчеты и они правильные, даны пояснения);

- оценка «хорошо» - работа выполнена полностью, но имеются незначительные ошибки;
- оценка «удовлетворительно» - работа выполнена полностью, но в представленной части много ошибок или представлена часть работы и она без ошибок;
- оценка «неудовлетворительно» - работа не выполнена.

Лабораторная работа №2
Пример задания.

1. Обозначить бизнес-проблему.
2. Сформулировать бизнес-цели.
3. Обозначить бизнес-задачи.
4. Свести бизнес-задачу к аналитической задаче.
5. Определить потребности в ресурсах (указать источники, виды ресурсов, виды и содержание информации, которую можно получить).
6. Подобрать технологии (методы, модели, алгоритмы, инструментальные средства), позволяющие работать с определенными в п.6 ресурсами.
7. При необходимости дать рекомендации по доработке технологии (методы, модели, алгоритмы, инструментальные средства) из п.6.

Технология проведения
Предметную область студент выбирает самостоятельно, базируясь на информации из открытых источников. Время выполнения задания составляет 3 часа. Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если работа выполнена в полном объеме, полученные результаты аргументированы;
- оценка «хорошо» - работа выполнена полностью, но полученные результаты не логичны или требуют уточнения;
- оценка «удовлетворительно» - работа выполнена полностью, но имеет место большое количество ошибок или представлена часть работы и она без ошибок;
- оценка «неудовлетворительно» - работа не выполнена.

Лабораторная работа №3

Пример задания.

Установить ПО Hadoop для организации работы с большими данными. При настройке ПО реализовать автономный режим работы.

Технология проведения
Лабораторная работа выполняется в учебной лаборатории. Студенту предоставляется доступ к системным настройкам ПК. Студент проводит инсталляцию системных файлов, настройку конфигурации ПО и запускает ПО в автономном режиме. Преподаватель проверяет факт установки и готовность ПК к дальнейшей работе. По окончании лабораторной работы рекомендовано восстановление системы до первоначального состояния.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если все этапы инсталляции ПО пройдены, ПО настроено и готово к работе;
- оценка «хорошо» - если все этапы инсталляции ПО пройдены, но ПО не настроено;
- оценка «удовлетворительно» - если студент не смог пройти все этапы инсталляции ПО;
- оценка «неудовлетворительно» - работа не выполнена.

## 20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: тест.

**Тестовые задания.**

Пример компоновки вопросов теста (вопросы с вариантами ответов).
Вариант 1.

1. Приведите основные характеристики больших данных:

а) Virtualization, Volume, Variability, Vehicle;
б) Variety, Velocity, Volume, Value;
в) Verification, Volume, Velocity, Visualization;
г) Video, Value, Variety, Volume.

2. Расставьте в правильном порядке основные этапы процесса Data Science:

а) назначение цели исследования, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение данных;
б) назначение цели исследования, сбор данных, подготовка данных, моделирование данных, исследование данных, отображение данных;
в) назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных;
г) назначение цели исследования, сбор данных, подготовка данных, отображение данных, исследование данных, моделирование данных.

3. Поясните понятие:

Hadoop представляет собой…
а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах;
б) распределённую СУБД, позволяющую обрабатывать большие данные;
в) язык выполнения заданий в парадигме MapReduce;
г) распределённую файловую систему для организации хранения файлов большого объёма.

4. Принцип MapReduce состоит в том, чтобы
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки.

Технология проведения: тест состоит из 50 вопросов. Вариант теста выбирается исходя из номера зачетки (последней цифры). Время тестирования составляет 45 минут. Результаты тесты проверяются по ключу правильных ответов.

Критерии оценивания:
- оценка «отлично» выставляется студенту, если студент дал правильные ответы на 90 и более процентов заданий (тест пройден);
- оценка «хорошо» выставляется студенту, если студент дал правильные ответы менее, чем на 90 и более 80 процентов заданий (тест пройден);
- оценка «удовлетворительно» выставляется студенту, если студент дал правильные ответы менее 80 и более 50 процентов заданий (тест пройден);

- оценка «неудовлетворительно» - даны правильные ответы на менее чем на 50 процентов заданий (тест не пройден).

Для оценивания результатов обучения на экзамене используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Соотношение показателей, критериев и шкалы оценивания результатов обучения:

| Критерии оценивания компетенций | Уровень сформированности компетенций | Шкала оценок |
|---|---|---|
| По тесту получено более 90% правильных ответов. По контрольной и лабораторным работам получены оценки «отлично» и «хорошо». | Повышенный уровень | Отлично |
| По тесту получено более 80% правильных ответов. По контрольной и лабораторным работам получены оценки «отлично» и «хорошо». | Базовый уровень | Хорошо |
| По тесту получено более 50% правильных ответов. По контрольной и лабораторным работам получены оценки «отлично» или «хорошо», или «удовлетворительно». | Пороговый уровень | Удовлетвори-тельно |
| По тесту получено 50% и менее правильных ответов и/или задания контрольной и лабораторных работ не выполнены. | – | Неудовлетвори-тельно |

## 20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

ПК-3 Способен проводить обработку и анализ больших данных на базе современных языков программирования и пакетов прикладных программ моделирования

Вопросы с вариантами ответов (закрытые)

1. Какие модули входят в основной пакет ПО Hadoop для работы с большими данными:
а) Common, HDFS, MapReduce, YARN;
б) MapReduce, HDFS, YARN;
в) MapReduce, HDFS, Common, HBase;
г) Common, YARN, HBase.
Ответ: а.

2. Какие виды узлов включает в себя ядро кластера ПО Hadoop для работы с большими данными:
а) Name Node и Date Node;
б) Job Tracker и Task Tracker;
в) Master First и Master Second;
г) Master и Slave.
Ответ: г.

3. Какая концепция положена в основу системы хранения данных HDFS:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) однократной записи и многократного чтения;
в) возможности репликации данных;
г) распараллеливания процессов обработки данных.
Ответ: б.

4. Принцип MapReduce применительно к технологии обработки больших данных состоит в том, чтобы:

а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки
г) разделять узлы на те, где хранятся данные и те, на которых производятся вычисления.
Ответ: а.

5. В чем заключаются основные идеи базы данных типа NoSQL:
а) возможность применения к неструктурированным и слабоструктурированным данным, отсутствие SQL-запросов;
б) возможность работы с различными типами хранилищ, реаляционная модель данных, удобство для разработчиков;
в) нереляционная модель данных, закрытый исходный код, вертикальная масштабируемость;
г) нереляционная модель данных, открытый исходный код, хорошая горизонтальная масштабируемость.
Ответ: г.

6. Укажите основные источники больших данных:
а) мобильные устройства;
б) машинные данные и интернет;
в) интернет и социальные сети;
г) машинные данные и мобильные устройства.
Ответ: г.

7. В чем заключается научное и общественное значение больших данных:
а) возможность извлечь экономическую выгоду;
б) новые знания о мире;
в) возможность ответить на давно интересующие вопросы;
г) возможность управления будущим.
Ответ: а.

8. Какие характеристики определяют принцип «Трёх V» в отношении больших данных:
а) Virtualization, Volume, Variability;
б) Variety, Volume, Value;
в) Verification, Velocity, Visualization;
г) Volume, Variety, Velocity.
Ответ: г.

9. Выберите наиболее подходящее определение ПО Hadoop:
а) платформа для запуска приложений аналитической обработки данных;
б) ПО, предназначенное для создания и запуска распределённых приложений, обрабатывающих большие объёмы данных;
в) ПО, предназначенное для организации работы с большими данными;
г) СУБД для неструктурированных и слабоструктурированных данных.
Ответ: б.

10. Отметьте основные типы инструментальных средств платформ работы с данными BI (Business Intelligence):

а) средства сбора и представления информации и средства интеграции;
б) средства сбора информации, средства очистки информации, средства анализа;
в) средства сбора информации, средства преобразования к виду, удобному для обработки, средства моделирования;
г) средства представления информации, средства интеграции, средства анализа.
Ответ: г.

11. Дайте определение процесса ETL, используемого при обработке больших массивов данных:
а) комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных;
б) ПО для извлечения данных из реляционных БД и преобразование их к виду, удобному для хранения в БД NoSQL;
в) комплекс методов, очистки данных, извлекаемых из различных источников;
г) ПО для приведения данных из разных источников к единому формату.
Ответ: а.

12. Укажите основные фазы моделирования данных, которые имеют место быть в работе с большими данными:
а) выбор переменных, определение значимости переменных, выполнение модели, исследование результатов;
б) выбор модели, выполнение модели, оценка результатов;
в) выбор модели и переменных, выполнение модели, определение степени соответствия модели, диагностика модели;
г) выбор модели и переменных, выполнение модели, диагностика и сравнение моделей.
Ответ: г.

13. Укажите основные виды данных из предложенных классификаций, которые имеют место быть в работе с большими данными:
а) структурированные, полуструктированные, квазиструктурированные и неструктурированные;
б) структурированные, полуструктурированные и неструктурированные;
в) структурированные и неструктурированные;
г) неструктурированные, квазиструктурированные, неоструктрированные и структурированные.
Ответ: а.

14. Отметьте высказывания, наиболее полно характеризующие структурированные данные в теории больших данных:
а) не зависят от модели данных, удобны для анализа, примером является аудио файлы;
б) зависят от модели данных, имеют определенную структуру, произвольный фрагмент текста трудно подвергается расшифровке;
в) данные содержат специальные теги и иные маркёры, позволяющие отделить семантические элементы, удобны для анализа;
г) зависят от модели данных, упорядочены специальным образом, обычно такие данные хранятся в виде таблиц в реляционных базах данных, удобны для анализа.
Ответ: г.

15. Укажите типы корпоративных данных:
а) машинные данные, естественные данные, социальных сетей сотрудников и клиентов;
б) фактографические, нормативно-справочные и внутренние;
в) конфиденциальные, из открытых и условно-открытых источников;
г) открытые и закрытые.
Ответ: б.

16. Что включает в себя этап сбора данных в процессе изучения данных DS (Data Science):
а) определение источников данных, определение методов сбора данных, сбор данных, первичный анализ данных;
б) определение источников данных, формирование цепочек жизненного цикла данных и определение методов сбора данных, сбор данных, первичный анализ данных;
в) определение источников данных, формирование цепочек жизненного цикла данных и определение методов сбора данных, первичный анализ данных;
г) определение источников данных, определение методов сбора данных, сбор данных.

Ответ: б.

17. Основная задача этапа подготовки данных при реализации процесса изучения данных DS (Data Science):
а) проведение предварительного исследования данных, описание данных;
б) очистка данных и составление их описания;
в) объединение данных из разных источников и приведение их к единому формату:
г) данные разных наборов приводят к общему формату, убирают опечатки и различные ошибки ввода.
Ответ: г.

18. Укажите уровни интеграции данных при использовании технологий больших данных:
а) семантический и синтаксический;
б) физический и логический;
в) внешний и внутренний;
г) ручной и машинный.
Ответ: б.

19. Дайте определение процессу преобразования данных при реализации процесса изучения данных DS (Data Science):
а) процесс приведения данных к виду, подходящему для моделирования данных;
б) процесс очистки данных с сокращением объема файла;
в) процесс выборки из данных полезной информации;
г) процесс приведения данных к формату, пригодному для применения SQL-запросов.
Ответ: а.

20. Выберите существующие группы визуализаторов данных, используемых при работе с большими данными:
а) общего назначения, для оценки качества моделей, для интерпретации результатов анализа;
б) для оценки качества входных данных, оценки качества моделей, оценки прогнозируемых значений;
в) общего назначения, специализированные;
г) для оценки входных данных, для оценки качества моделей, для интерпретации результатов анализа.
Ответ: а.

21. Основная задача визуализации данных – это:
а) представление результатов ключевым участникам проекта и построения приложений на их основе;
б) представление результатов в графическом виде;
в) представление результатов для оценки модели;
г) представление результатов для удобства оценки результатов моделирования.
Ответ: а.

Вопросы с кратким текстовым ответом (открытые)

22. Как называются базы данных, в которых используются не только SQL-запросы (ответ запишите латинскими буквами в верхнем регистре)
Ответ: NOSQL.

23. Как называется техника масштабирования при работе с данными, которая заключается в разделении (партиционировании) базы данных на отдельные части так, чтобы каждую из них можно было вынести на отдельный сервер (ответ запишите русскими буквами в нижнем регистре)
Ответ: шардинг.

24. Приведите аббревиатуру распределенной файловой системы для ПО Hadoop (ответ запишите латинскими буквами в верхнем регистре).
Ответ: HDFS.

25. Как обозначается комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных (ответ запишите латинскими буквами в верхнем регистре)
Ответ: ETL.

ПК-4 Способен управлять разработкой профессионально--ориентированных информационных систем с учетом возможностей современных интеллектуальных информационных технологий

Вопросы с вариантами ответов (закрытые)

1. Расставьте в правильном порядке основные этапы процесса Data Science:
а) назначение цели исследования, сбор данных, подготовка данных, моделирование данных, исследование данных, отображение данных;
б) назначение цели исследования, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение данных;
в) назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных;
г) назначение цели исследования, сбор данных, подготовка данных, отображение данных, исследование данных, моделирование данных.
Ответ: б.

2. Основные методологии аналитики данных Data Science:
а) CRISP-DM, SEMMA, KDD;
б) CRISP, SEMNA, ERP;
в) KDD, SEMMA. DSM;
г) CRISP-DM, KDD.
Ответ: а.

3. Основные этапы методологии аналитики данных CRISP-DM (в порядке исполнения):
а) бизнес-анализ, анализ-данных, подготовка данных, моделирование, оценка решений, внедрение;
б) отбор данных, исследование отношений в данных, модификация данных, моделирование взаимозависимостей;
в) формирование бизнес-задачи, анализ-данных, сбор и подготовка данных, исследование данных;
г) анализ источников данных, сбор данных, построение моделей, внедрение.
Ответ: а.

4. На каком этапе методологии аналитики данных CRISP-DM применяется метод A/B-тестирования:
а) моделирование;
б) оценка полученных моделей;
в) оценка решений;
г) внедрение.
Ответ: а.

5. Перечислите системные требования к ПО Hadoop для работы с большими данными:
а) ОС Linux, поддержка Java API, кластерная топология;
б) ОС Linux, C++;
в) ОС Windows 64-разрядная, поддержка реляционных БД;
г) любая ОС, ограничений на языки программирования нет.

Ответ: а.

6. Условия реализуемости концепции MapReduce при работе с большими данными:
а) наличие распределенной файловой системы, планировщика, неиндексированное хранение данных, автоматизации распараллеливания задач на кластере;
б) поддержка репликации данных, индексированное хранение данных, наличие планировщика;
в) наличие планировщика, индексированное хранение данных;
г) наличие планировщика, неиндексированное хранение данных.
Ответ: а.

7. Принцип MapReduce при работе с большими данными состоит в том, чтобы:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки
г) разделять узлы на те, где хранятся данные и те, на которых производятся вычисления.
Ответ: а.

8. Основные механизмы реализации баз данных типа NoSQL:
а) репликация и шардинг;
б) шардинг и поддержка map/reduce;
в) репликация и поддержка map/reduce;
г) репликация и горизонтальное масштабирование.
Ответ: а.

Вопросы с кратким текстовым ответом (открытые)

9. Выберите верные утверждения относительно логики работы функции MapReduce в ПО Hadoop для работы с большими данными (выберите нужное и запишите ответ в виде последовательности цифр без пробела, например «35»):
1. Map выполняет предварительную обработку входных данных
2. Map преобразует входной набор в список пар ключ/значение
3. Map производит свёртку заранее обработанных данных
4. Reduce производит свёртку заранее обработанных данных
5. Reduce получает на выходе новый объединенный список пар ключ/значение.
Ответ: 1245.

10. Основные режимы запуска ПО Hadoop для работы с большими данными (выберите нужные варианты и запишите ответ в виде последовательности цифр без пробела, например «35»):
1. Автономный
2. Псевдораспределенный
3. Полностью распределенный
4. Локальный
5. Сетевой.
Ответ: 123.

**Критерии и шкалы оценивания заданий ФОС:**

Для оценивания выполнения заданий используется балльная шкала:

1) закрытые задания (тестовые с вариантами ответов, средний уровень сложности):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

2) открытые задания (тестовые с кратким текстовым ответом, повышенный уровень сложности):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

**Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).**