

**ФГБОУ ВО НОВОСИБИРСКИЙ ГАУ**


**Кафедра ветеринарной генетики и биотехнологии**

Рег. № Зоогб и 04-19  
«07» 10 2022 г.

**УТВЕРЖДЕН**

на заседании кафедры

Протокол от «10» 10 2022 г. №2  
Заведующий кафедрой

  
(подпись)

Е.В. Камалдинов

**ФОНД  
ОЦЕНОЧНЫХ СРЕДСТВ**

Б1.В.03 Big Data в животноводстве

36.04.02 Зоотехния

Прикладная биоинформатика

Новосибирск 2022

## Паспорт фонда оценочных средств

| №п/<br>п | Контролируемые разделы<br>(темы) дисциплины*       | Код<br>контролируемо<br>й компетенции<br>(или ее части) | Наименование<br>оценочного средства |
|----------|--|---|-------------------------------------|
| 1        | Введение в теорию больших данных                   | <i>ПК-2; ПК-3</i>                                       | Коллоквиум,<br>собеседование        |
| 2        | Введение в анализ данных                           |   |                                     |
| 3        | Python и R, как языки для анализа данных           |   |                                     |
| 4        | OLAP (OnLine Analytical Processing)                |   |                                     |
| 5        | Интернет вещей (IoT) и BigData                     |   |                                     |
| 6        | Компьютерное зрение (получение, обработка, анализ) |   |                                     |
| 7        | Моделирование в Python                             |   |                                     |
| 8        | Hadoop   |   |                                     |
| 9        | NoSQL базы данных                                  |   |                                     |
|          | Контрольная работа                                 | ПК-2, ПК-3  | Задания к контрольной работе        |
|          | Промежуточная форма (экзамен)                      | ПК-2, ПК-3  | Вопросы для экзамена                |

# **ТЕКУЩИЙ КОНТРОЛЬ УСПЕВАЕМОСТИ**

## **Вопросы к коллоквиуму собеседованию**

### **Тема 1. Введение в теорию больших данных**

1. Какие технологии и методы искусственного интеллекта используются для обработки больших данных?
2. В чем заключается принцип работы с большими данными и какие основные этапы этого процесса?
3. Какие преимущества и недостатки использования BigData технологий? Приведите примеры использования BigData в различных отраслях.
4. Какие методы анализа больших данных применяются в сельском хозяйстве и какие практические выгоды они могут принести?
5. Какие характеристики определяют данные как "большие данные" и какие проблемы могут возникнуть при работе с ними?
6. Какова экосистема больших данных и какие компоненты включает в себя данная система?
7. Какие алгоритмы машинного обучения чаще всего используются для анализа больших данных?
8. Какие технологии обработки потоков данных широко применяются в области анализа больших данных?
9. Какие методы хранения и обработки больших данных используются для обеспечения высокой доступности и отказоустойчивости?
10. Какие проблемы безопасности могут возникнуть при работе с большими данными и как их можно предотвратить?
11. Какие инструменты для визуализации больших данных существуют на рынке и какие преимущества они предоставляют аналитикам?
12. Какие способы сбора данных из различных источников используются при работе с большими данными?
13. Какие метрики используются для оценки качества анализа больших данных?
14. Какие тренды в области Big Data наблюдаются на современном рынке и какие новые технологии ожидается увидеть в будущем?
15. Какие практические примеры успешного применения Big Data технологий вы можете привести из своего опыта работы или известных кейсов?

### **Тема 2. Введение в анализ данных**

1. Какие основные определения встречаются в анализе данных и как они взаимосвязаны между собой?
2. В чем отличие между структурированными, неструктурированными и полуструктурированными данными? Приведите примеры каждого типа данных.
3. Каков жизненный цикл данных и какие основные этапы включает в себя этот цикл?

4. Как определить, являются ли данные структурированными или неструктурированными с помощью анализа данных?
5. Какие методы структуризации данных чаще всего применяются в процессе анализа информации?
6. Какие инструменты и технологии используются для обработки и анализа структурированных данных?
7. Какие преимущества и недостатки неструктурированных данных можно выделить в контексте анализа информации?
8. Какие типы полуструктурированных данных достаточно распространены и как они могут быть использованы в анализе данных?
9. Какова роль аналитики данных в жизненном цикле данных и какие функции они выполняют на различных этапах?
10. Какие методы сбора и хранения данных используются на этапе жизненного цикла анализа данных?
11. Как осуществляется защита данных на различных этапах и насколько важна безопасность информации в процессе анализа данных?
12. Какие алгоритмы анализа данных применяются для поиска закономерностей и паттернов в массивах информации?
13. Какова роль визуализации данных в процессе анализа информации и какие инструменты можно использовать для визуализации результатов?
14. Каким образом проводится интерпретация результатов анализа данных и какие выводы могут быть сделаны на основе полученных данных?
15. Как организовать процесс хранения и документирования данных в рамках жизненного цикла анализа информации?

### **Тема 3. Python и R, как языки для анализа данных**

1. Какие инструменты для анализа данных используются в Python от R?
2. Какие основные архитектурные черты отличают Python от R?
3. Какие методы статистического анализа доступны в R и Python?
4. Как установить и подключить библиотеки для анализа данных в R?
5. Как установить и подключить библиотеки для анализа данных в Python?
6. Что такое SciPy и для чего он используется?
7. Что такое Pandas и какие функции она предоставляет для работы с данными?
8. Что такое StatsModels и в чем заключается ее роль в анализе данных?
9. Какие возможности предоставляет библиотека Matplotlib для визуализации данных?
10. Что такое Seaborn и в чем отличие от Matplotlib?
11. Для чего используется Plotly в анализе данных?
12. Что такое Scikit-Learn и какие алгоритмы машинного обучения доступны в этой библиотеке?
13. Какие возможности предоставляет Keras для разработки нейронных

сетей?

14. Какие преимущества предоставляет использование Python в сравнении с R для анализа данных?

15. Какие преимущества предоставляет использование R в сравнении с Python для анализа данных?

#### **Тема 4. OLAP (OnLine Analytical Processing)**

1. Какая роль играет OLAP в многомерном анализе данных?

2. Что такое многомерная модель данных и какие преимущества она предоставляет для анализа данных?

3. Каков принцип работы OLAP-систем в контексте аналитической обработки данных?

4. Какие типы хранилищ данных существуют в OLAP (многомерные, реляционные, гибридные) и как они отличаются друг от друга?

5. Какие требования обычно предъявляются к OLAP-системам?

6. Какие преимущества предоставляют OLAP-системы по сравнению с другими методами анализа данных?

7. Какие основные компоненты входят в структуру OLAP-системы?

8. Какие техники мультимедийного отображения данных используются в OLAP-системах?

9. Какие методы агрегации данных поддерживаются в OLAP-системах?

10. Какие основные виды запросов могут быть выражены в языке OLAP-запросов?

11. Какие инструменты и технологии используются для визуализации данных в OLAP-системах?

12. Какие методы анализа данных поддерживаются в OLAP-системах (дробление, свертка, дробно-свертывающий анализ)?

13. Каким образом OLAP-системы помогают бизнесу принимать более обоснованные решения?

14. Какие специфические функции агрегирования данных (среднее, максимальное, минимальное и т. д.) предоставляют OLAP-системы?

15. Какие проблемы могут возникнуть при проектировании и реализации OLAP-систем? Как их можно избежать или решить?

#### **Тема 5. Интернет вещей (IoT) и BigData**

1. Какова структура и компоненты экосистемы IoT?

2. Какие примеры использования IoT в сельском хозяйстве можно привести?

3. Какова типичная архитектура IoT-системы и какие роли играют в ней устройства, сенсоры и облака?

4. Каким образом происходит сбор и обработка малых данных в сети Интернета вещей?

5. Как Big Data используется в сфере IoT и в каких областях это принесло наибольшую пользу?

6. Какие вызовы и проблемы могут возникнуть при интеграции IoT и

Big Data технологий?

7. Каким образом IoT устройства собирают и передают данные в облако для дальнейшего анализа?
8. Какие инструменты и технологии Big Data используются для обработки и анализа данных, собранных с IoT устройств?
9. Какие меры безопасности должны быть приняты при работе с данными, собранными из IoT устройств и Big Data систем?
10. Какие преимущества можно получить от использования аналитики Big Data в IoT для улучшения бизнес-процессов?
11. Какие тенденции и направления развития связывания IoT и Big Data можно наблюдать на рынке в настоящее время?
12. Какие технологические ограничения могут возникнуть при работе с большими объемами данных из IoT устройств?
13. Какова роль облачных платформ в обработке и хранении данных, полученных из IoT устройств?
14. Каким образом данные, полученные из IoT сенсоров, могут быть использованы для принятия управленческих решений на основе аналитики Big Data?
15. Каковы примеры успешного использования технологий IoT и Big Data в агропромышленном комплексе?

#### **Тема 6. Компьютерное зрение (получение, обработка, анализ)**

1. Какие задачи решаются с помощью компьютерного зрения?
2. Как компьютерное зрение применяется в сельском хозяйстве? Приведите примеры.
3. Чем отличается машинное зрение от компьютерного зрения?
4. Как происходит подготовка изображения для анализа в компьютерном зрении?
5. Что такое цветовые пространства и как они используются при обработке изображений?
6. Что такое детектирование и сегментация при анализе изображений?
7. Чем отличаются нормализация и бинаризация изображений?
8. Как можно использовать библиотеку OpenCV в языке программирования Python для работы с изображениями?
9. Как происходит коррекция геометрии и моделирование шума при анализе изображений?
10. Какие подходы существуют к шумоподавлению в компьютерном зрении?
11. Как можно реализовать алгоритмы поиска похожих изображений на практике?
12. В чем состоит роль компьютерного зрения в автоматизации процессов промышленных предприятий?
13. Какие технологии помогают улучшить точность распознавания объектов на изображениях?
14. Каким образом компьютерное зрение используется для

распознавания лиц и биометрической идентификации?

15. Какие вызовы и проблемы сталкиваются при реализации систем компьютерного зрения в реальных условиях?

### **Тема 7. Моделирование в Python**

1. Что такое регрессия и какие виды регрессии существуют?
2. Как работает линейная регрессия в моделировании данных в Python?
3. Чем отличается логистическая регрессия от линейной регрессии?
4. Что такое градиентный спуск и как он используется при моделировании данных?
5. Какая роль полиномиальной регрессии в моделировании данных?
6. Что такое регуляризация и как она помогает улучшить производительность моделей?
7. В чем заключается процесс настройки гиперпараметров модели?
8. Какие методы уменьшения размерности данных используются в моделировании?
9. Какая связь между регуляризацией и решением проблемы переобучения моделей?
10. Каким образом можно использовать методы регуляризации для борьбы с переобучением?
11. Как происходит подбор оптимальных гиперпараметров модели в Python?
12. Какие алгоритмы регуляризации в Python позволяют улучшить качество моделей?
13. Как влияет уменьшение размерности на производительность моделей машинного обучения?
14. Как выбрать наиболее подходящий метод градиентного спуска для конкретной задачи моделирования?
15. Какие библиотеки и инструменты можно использовать в Python для реализации моделей с регуляризацией и уменьшением размерности?

### **Тема 8. Hadoop**

1. . Что такое Big Data и какая роль у MapReduce в обработке больших объемов данных?
2. Какие основные компоненты входят в экосистему Hadoop?
3. Чем отличается распределенная файловая система HDFS от традиционных файловых систем?
4. Как работает фреймворк MapReduce и какие задачи он выполняет в обработке данных?
5. Какие алгоритмы используются в MapReduce для распределенной обработки данных?
6. Каким образом можно использовать MapReduce для работы с графами и решения графовых задач?
7. Что такое Pig и каким образом этот инструмент используется в

контексте Hadoop?

8. Что такое Hive и в чем состоит его применение при разработке процедурной логики для анализа данных?

9. Как Hadoop помогает организациям обрабатывать и хранить большие объемы данных?

10. Каким образом данные хранятся и организуются в HDFS?

11. Как происходит работа MapReduce в случае сбоя узла в кластере Hadoop?

12. Какие альтернативы используются для MapReduce в технологии обработки больших данных?

13. Каким образом можно улучшить производительность алгоритмов MapReduce?

14. Какие специальные возможности позволяют графы оптимизировать процессы обработки данных в MapReduce?

15. Какие преимущества и недостатки существуют у инструментов Pig и Hive по сравнению с нативным MapReduce?

## Тема 9. NoSQL базы данных

1. Чем отличается NoSQL база данных от реляционной базы данных?
2. Какие преимущества предлагает использование NoSQL баз данных?
3. Какие ограничения есть у NoSQL баз данных?
4. Какая основная цель использования MongoDB?
5. Как установить MongoDB на компьютер?
6. Как создать новую базу данных в MongoDB?
7. Какие типы данных поддерживает MongoDB?
8. Как создать новую коллекцию и добавить документы в MongoDB?
9. Как выполнить запрос и выбрать данные из коллекции MongoDB?
10. Как обновить данные в коллекции MongoDB?
11. Как удалить данные из коллекции MongoDB?
12. Как объединить несколько запросов в один запрос в MongoDB?
13. Как провести поиск на совпадение в тексте в MongoDB?
14. Как обеспечить моментальную обработку данных в MongoDB?
15. Каким образом можно оптимизировать производительность базы данных MongoDB?

## Контрольная работа

### Тема 1.

1. Создайте новую БД при помощи языка программирования SQL
2. В созданной БД создайте таблицу, включающую столбцы с типами данных:

| Название столбца | Тип данных | Описание                            |
|------------------|------------|-------------------------------------|
| id               | int        | Порядковый номер записи начиная с 1 |
| short_name       | varchar30  | Короткая кличка                     |
| name             | varchar70  | Полная клика                        |



|                 |           |   |
|-----------------|-----------|---|
| register_number | varchar20 | Международный номер                         |
| source          | varchar50 | Источник (CDN, локальная база хоз-ва и т.д) |
| birth_data      | data      | Дата рождения                               |

3. Присвойте первичный ключ столбцу *id*
4. Присвойте столбцу *id* автоинкремент
5. Создайте в таблице 50 записей о рандомных быках как минимум их 3-х источников

## Тема 2.

1. Создайте таблицу согласно заданию 1.5
2. К таблице добавьте столбец с записями о породе (char2)
3. Из таблицы извлеките строки, в которых номер животного содержит «USA»
4. Замените все записи «840» на «USA»
5. Выведите информацию о самом старшем животного

## Тема 3.

1. Создайте таблицу согласно заданию 1.5
2. Отсортируйте записи по короткой кличке
3. Выведите информацию о 8 последних животных
4. Выведите информацию о животных с 20-й по 34-ю строк
5. Выведете дату рождения и короткую кличку для животных, номера которых содержат «HOCAN».

## Тема 4.

1. Создайте таблицу согласно заданию 1.5
2. Создайте таблицу:

| Название столбца | Тип данных | Описание   |
|------------------|------------|--|
| id               | int        | Порядковый номер записи начиная с 1              |
| name             | varchar70  | Полная клика                                     |
| Register_number  | varchar20  | Международный номер                              |
| milk             | int        | Геномная оценка быка по прибавке к удою в фунтах |
| fat              | num(1.2)   | Геномная оценка быка по прибавке к жиру %        |
| protein          | num(1.2)   | Геномная оценка быка по прибавке к белку %       |

3. Объедините таблицу с таблицей из темы 1 используя в качестве ключа номер животного
4. Выведете информацию о 3 животных с наибольшей прибавкой по жиру, номера которых содержат «HOUSA ».
5. Суммируйте данные столбцов milk, fat, protein

## Тема 5.

1. Сгенерировать большой массив данных используя данные каталогов быков AltaGenetics, ST Genetics, Semex и WWS. Преобразовать в объект *tibble*. Установить пакет *purrr*. Записать массив данных по частям в несколько файлов.
2. Сформировать репрезентативная выборку размера не превышающей 1/10 от общего объёма.
3. При помощи функций пакета *data.table* произвести загрузку и выгрузку файлов. Вычислить среднее время задержки прибытия (*mean\_arr\_delay*) и отправления (*mean\_dep\_delay*) для различных перевозчиков (*carrier*).
4. Установить и загрузить библиотеки *sqldf* и *nycflights13*. Ознакомиться со структурой набора данных *flights*. Вычислить количество наблюдений для всех перевозчиков *carrier* в таблице *flights*. Отобразить в консоли значения полей *dep\_time*, *dep\_delay*, *arr\_time*, *carrier*, *tailnum* из таблицы *flights* (первые и последние 5 строк).
4. Сгенерировать *data.frame* с тремя столбцами и 100 строками. Преобразовать данные из широкого в длинный формат. Установить пакет *reshape2*

## Тема 6.

1. Сгенерировать вектор (массив, таблица данных) и добавить в него элементы NA. Очистить данные с использованием функции *is.na()*.
2. Сгенерировать таблицу данных с числовыми и текстовые столбцами. Очистить данные с функции *complete.cases()*.
3. Сгенерировать два числовых набора данных, добавить в них выбросы. С использованием функции *boxplot* обнаружить выбросы и удалить их.
4. Сгенерируйте таблицу данных, в которой дублируются строки. Удалите строки с использованием функций *unique()*, *duplicated()*. Сравните результаты.
5. Сгенерировать числовую таблицу данных с пропусками. С использованием функции *preProcess* из пакета *caret* заполнить пропуски предсказанными значениями (среднее, медиана).

## Тема 7.

1. Установить пакет CARET, выполнить команду *names(getModelInfo())*, ознакомиться со списком доступных методов выбора признаков. Выполните графический разведочный анализ данных с использование функции *featurePlot()* для набора данных из справочного файла пакета CARET:  

```
x <- matrix(rnorm(50*5),ncol=5)
y <- factor(rep(c("A", "B"), 25))
```

Сохранить полученные графики в \*.jpg файлы. Сделать выводы.

2. С использованием функций из пакета *Fselector* определить важность признаков для решения задачи классификации. Использовать набор *data(iris)*. Сделать выводы.

3. Установите пакет *Boruta* и проведите выбор признаков для набора данных *data("Ozone")*. Построить график *boxplot*, сделать выводы.

4. В таблице (тема 5) осуществите поиск животных с повторяющимися короткими кличками используя функции пакета *data.table*.

5. Загрузите таблицу (тема 1) используя возможности пакета *RmySQL*. Объедините данные с таблицей (тема 12) используя полную кличку быка в качестве ключа

## Тема 8.

1. Выполните классификацию k-ближайших соседей с использованием функции *knn()* из пакета *class* на наборе данных *iris*. Проведите нормализацию данных, разделите выборку на обучающую и тестовую.

2. Оцените построенную модель с использованием функции *CrossTable()* из пакета *gmodels*. Постройте матрицу ошибок и диагональную оценку качества прогноза (*diagonal mark quality prediction*).

3. Рассмотрите пример реализации метода опорных векторов с использованием функции *svm()* из пакета *e1071*. Постройте линейный классификатор для прогнозирования. Для подбора параметров модели выполните перекрестную проверку с делением исходной выборки на 10 равных частей (*cross=10*).

4. Выполните расчет главных компонент с использованием пакета *vegan()* и его функции *rda()*. Постройте ординационную диаграмму методом PCA [3, с. 49] и сделайте выводы.

5. Выполните кластерный анализ по данным таблицы *iris*

## Тема 9.

1. С использованием функции *discretize()* из пакета *arules* выполните преобразование непрерывной переменной в категориальную [1] различными методами: «interval» (равная ширина интервала), «frequency» (равная частота), «cluster» (кластеризация) и «fixed» (категории задают границы интервалов). Используйте набор данных *iris*.

2. С использованием пакета *discretization* выполните дискретизацию с использованием алгоритмов Chi2 и CAIM. Используйте набор данных *iris*.

3. Выполните классификацию k-ближайших соседей с использованием функции *knn()* из пакета *class* на наборе данных *iris*. Проведите нормализацию данных, разделите выборку на обучающую и тестовую. Оцените построенную модель с использованием функции *CrossTable()* из пакета *gmodels*. Постройте матрицу ошибок и диагональную оценку качества прогноза (*diagonal mark quality prediction*).

4. Рассмотрите пример реализации метода опорных векторов с

использованием функции `svm()` из пакета `e1071`. Постройте линейный классификатор для прогнозирования. Для подбора параметров модели выполните перекрестную проверку с делением исходной выборки на 10 равных частей (`cross=10`) [3, с.172].

5. Выполните расчет главных компонент с использованием пакета `vegan()` и его функции `rda`. Постройте ординационную диаграмму методом PCA.

## Тема 10.

1. Выполните команду `library(tidyverse)`. Используя функцию `gather` преобразуйте таблицу `table4a` в

|                    |             |               |
|--------------------|-------------|---------------|
| <i>Afghanistan</i> | <i>1999</i> | <i>745</i>    |
| <i>Brazil</i>      | <i>1999</i> | <i>37737</i>  |
| <i>China</i>       | <i>1999</i> | <i>212258</i> |
| <i>Afghanistan</i> | <i>2000</i> | <i>2666</i>   |
| <i>Brazil</i>      | <i>2000</i> | <i>80488</i>  |
| <i>China</i>       | <i>2000</i> | <i>21376</i>  |

2. Используя команду `filter` (библиотека `tidyverse`) найдите все рейсы, которые:

- С задержкой прибытия на два или более часа
  - Выполнялись в Хьюстон (IAH или HOU)
  - Выполнялись авиакомпаниями United, American или Delta
  - Вылетели летом (июль, август и сентябрь)
  - Прибыли с опозданием более чем на два часа, но не улетели поздно
  - Были задержаны как минимум на час, но провели в полете более 30 минут
  - Вылетели между полуночью и 6 часами утра (включительно))
3. Используя команду `between` определите на скольких рейсах отсутствует `dep_time`, `dep_delay`, `arr_time`, and `arr_delay` from?
4. Преобразуйте `dep_time` и `sched_dep_time` в минуты, проранжируйте по возрастанию, представьте результаты описательной статистики и соответствие нормальному распределению при помощи критерия Андерсона-Дарлинга.
5. С помощью функции `lag()` исследуйте, как задержка рейса связана с задержкой непосредственно предшествующего рейса.

## Критерии оценки:

- оценка «зачтено» выставляется студенту, если содержание ответов на вопросы в контрольной работе в целом соответствует теме задания, продемонстрировано знание фактического материала и уверенное владение понятиями - терминологическим аппаратом дисциплины, отсутствуют ошибки в употреблении терминов, ответы четко структурированы и выстроены в заданной логике, работа выполнена аккуратно.

- оценка «не зачтено» выставляется студенту, если содержание ответов в контрольной работе не соответствует теме задания или соответствует ему в очень малой степени, продемонстрировано крайне низкое знание фактического материала и слабое владение понятийно - терминологическим аппаратом дисциплины, присутствуют многочисленные ошибки в употреблении терминов, ответ представляет собой сплошной текст без структурирования, нарушена заданная логика, работа выполнена неаккуратно.

## **ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ**

### **Список вопросов для подготовки к экзамену** (Компетенции ПК-2; ПК-3)

1. Какие технологии и методы используются в области искусственного интеллекта для обработки больших данных?
2. Какие принципы лежат в основе работы с большими данными и какие инструменты используются для их анализа?
3. Какие преимущества и недостатки существуют при работе с Big Data? Приведите примеры успешного применения Big Data.
4. Каким образом технологии Big Data могут быть применены в сельском хозяйстве для повышения производительности и улучшения качества продукции?
5. Какие характеристики определяют большие данные и какова структура экосистемы для обработки и анализа данных в рамках Big Data?
6. Что такое структурированные данные в контексте анализа данных и какие примеры можно привести?
7. Какие основные характеристики имеют неструктурированные данные и почему их анализ является вызовом для специалистов по обработке данных?
8. Что такое полуструктурированные данные и какие преимущества они могут предоставить при анализе данных?
9. Какие этапы включает жизненный цикл данных и почему важно учитывать каждый из них при анализе данных?
10. Какие методы и инструменты можно использовать для эффективной работы с данными на различных этапах их жизненного цикла?
11. В чем заключается основное назначение инструментов для анализа данных и как они помогают специалистам в области Data Science?
12. Какие архитектурные черты характеризуют языки программирования Python и R с точки зрения анализа данных и машинного обучения?
13. Какие возможности предоставляют статистические методы анализа данных в R и Python, и в чем основные отличия между ними?
14. Какие популярные библиотеки для анализа данных существуют в R и Python, и как их правильно установить и подключить для использования в проектах?

15. Каким образом библиотеки SciPy, Pandas, StatsModels, Matplotlib, Seaborn, Plotly, Scikit-Learn, Keras используются для анализа данных и машинного обучения в языке программирования Python?

16. Как связаны OLAP и многомерный анализ данных? Какие преимущества предоставляет такой подход в сравнении с другими методами анализа данных?

17. В чем заключается многомерная модель данных при работе с OLAP системами и как она себя проявляет в анализе информации?

18. Разъясните принцип работы OLAP-систем. Какие основные компоненты включает в себя система OLAP и как они взаимодействуют между собой?

19. Какие типы хранилищ данных (многомерные, реляционные, гибридные) могут быть использованы в OLAP-системах? В чем различия между ними и какой тип хранилища предпочтительнее для OLAP?

20. Какие требования выдвигаются к OLAP-системам? Какие преимущества обладают OLAP-системы по сравнению с другими методами анализа данных и почему их использование становится все более популярным в современном бизнесе?

21. Какие элементы включает в себя экосистема Интернета вещей (IoT) и как они взаимодействуют между собой для обеспечения функционирования системы?

22. В каких областях сельского хозяйства можно использовать технологии Интернета вещей (IoT) для улучшения производительности и оптимизации процессов?

23. Какая архитектура лежит в основе систем Интернета вещей (IoT) и какие компоненты включает в себя каждый уровень этой архитектуры?

24. Чем отличается сбор и обработка малых данных в системах Интернета вещей (IoT) от работы с большими данными, и какие инструменты используются для этого процесса?

25. Какова роль совместного использования технологий Интернета вещей (IoT) и Big Data в современном мире? Какие выгоды могут быть получены от интеграции этих двух технологий для бизнеса и общества в целом?

26. Какие задачи могут решаться с использованием компьютерного зрения и какие применения данной технологии наиболее распространены в современном мире?

27. Какие примеры успешного применения компьютерного зрения в сельском хозяйстве существуют и какие конкретные преимущества они принесли в данной отрасли?

28. Чем отличается машинное зрение от обычного компьютерного зрения и какие задачи могут быть решены с его помощью?

29. Какие этапы включают в себя подготовка, обработка и анализ изображений при работе с компьютерным зрением, и какие техники обработки иногда применяются для улучшения качества изображений?

30. Что такое цветовые пространства и какие существуют основные

цветовые модели? Какое значение они имеют при обработке изображений и анализе данных в рамках компьютерного зрения?

31. В чем суть линейной регрессии и какие задачи она помогает решать в моделировании данных? Какие инструменты предоставляет Python для работы с линейной регрессией?

32. Что такое градиентный спуск и в каких случаях он применяется при построении моделей в Python? Какие основные параметры и методы выбора шага градиентного спуска могут повлиять на качество модели?

33. Чем отличается полиномиальная регрессия от линейной? Какие преимущества может принести использование полиномиальной регрессии при анализе данных с помощью Python?

34. Как работает регуляризация в контексте моделирования данных? Какие методы регуляризации существуют и как они помогают уменьшить переобучение модели? Как можно настроить гиперпараметры регуляризации в Python?

35. Какие способы уменьшения размерности данных существуют в Python и какие методы могут быть применены для улучшения производительности моделей? Какие преимущества может принести уменьшение размерности при работе с большими объемами данных?

36. Чем отличается NoSQL база данных от реляционной базы данных?

37. Как установить MongoDB на свой компьютер и создать новую базу данных?

38. Как создать новую коллекцию и добавить в нее документы в MongoDB?

39. Как с помощью программирования добавить данные в коллекцию MongoDB?

40. Как написать запрос для выборки данных из коллекции MongoDB по определенным критериям?

41. Какие принципы лежат в основе работы с Big Data и почему MapReduce является важной технологией в этой области?

42. Какие основные компоненты включает в себя экосистема Hadoop и как они взаимодействуют между собой?

43. Чем характеризуется HDFS как распределенная файловая система и какие преимущества она предоставляет при работе с большими объемами данных?

44. В чем состоит основной принцип работы MapReduce и какие задачи он позволяет эффективно решать?

45. Какие алгоритмы используются в MapReduce для обработки данных и какие особенности имеют алгоритмы работы с графами в этом контексте?

46. Чем отличаются инструменты Pig и Hive от MapReduce и какие задачи они предназначены помогать решать в области обработки данных?

## **Порядок аттестации студентов по дисциплине**

Для аттестации студентов по дисциплине используется традиционная система контроля и оценки успеваемости обучающихся.

Условием допуска к экзамену является посещение не менее 50% академических часов в рамках контактной работы. Для получения оценки «отлично» необходимо правильно решить практическую задачу с использованием ЭВМ и ответить на два теоретических вопроса, «хорошо» - решить практическую задачу и ответить на один теоретический вопрос, «удовлетворительно» - решить практическую задачу. При отсутствии решения практической задачи выставляется отметка «удовлетворительно».

### **Шкала оценивания:**

Критерии оценивания устного ответа на экзаменационные вопросы:

«5» (отлично) — дан полный, развёрнутый ответ на поставленный вопрос, показана совокупность осознанных знаний об объекте, доказательно раскрыты основные положения темы; в ответе прослеживается чёткая структура, логическая последовательность, отражающая сущность раскрываемых понятий, теорий, явлений. Знание об объекте демонстрируется на фоне понимания его в системе данной науки, и междисциплинарных связей. Ответ изложен литературным языком в терминах науки, показана способность быстро реагировать на уточняющие вопросы;

«4» (хорошо) — дан полный, развёрнутый ответ на поставленный вопрос, показано умение выделить существенные и несущественные признаки, причинно-следственные связи. Ответ чётко структурирован, логичен, изложен в терминах науки, Однако допущены незначительные ошибки ли недочёты, исправленные студентом с помощью «наводящих» вопросов;

«3» (удовлетворительно) — дан неполный ответ, логика и последовательность изложения имеют существенные нарушения. Допущены грубые ошибки при определении сущности раскрываемых понятий, теорий явлений, в следствии непонимания студентом их существенных и несущественных признаков и связей. В ответе отсутствуют выводы. Умение раскрыть конкретные проявления обобщённых знаний не показано. Речевое оформление требует поправок, коррекций.

«2» (неудовлетворительно) — студент демонстрирует незнание теоретических основ предмета, не умеет выделять аргументированные выводы и приводить примеры, показывает слабое владение монологической речью, не владеет терминологией, проявляет отсутствие логичности и последовательности изложения, делает ошибки, которые не может исправить даже при коррекции преподавателем, отказывается отвечать на занятия.



## ЗАДАНИЯ НА УРОВЕНЬ СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ

### Компетенция ПК-2

#### Задания закрытого типа

1. Данные у которых нет строго зафиксированного формата (Например: текстовые документы, PDF, изображения и видеозаписи являются:

- а. квазиструктурированными
- б. неструктурированными**
- в. полуструктурированными
- г. структурированными

2. Какое из нижеперечисленных понятий не относится к перечню необходимых критериев для создания проекта, связанного с большими данными?

- а. Географическое положение**
- б. Производительность
- в. Гибкость анализа
- г. Скорость принятия решения

3. Выберите верную зависимость структурированности информации от ее объема

- а. чем больше объем, тем более структурирована информация
- б. чем больше объем информации, тем менее она структурирована**
- в. чем меньше объем, тем менее структурирована информация
- г. они не зависят друг от друга

4. Что означает термин Big Data в информационных технологиях?

- а. комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов**
- б. представление времени, дня, месяца и года в качестве значения количества миллисекунд, прошедших с начала нашей эры
- в. файлы с большим количеством данных
- г. комплексный набор методов для создания файлов большого объема

#### Задания открытого типа

1. Какие принципы лежат в основе работы с Big Data и почему MapReduce является важной технологией в этой области?
2. В чем заключается основное назначение инструментов для анализа данных и как они помогают специалистам в области Data Science?
3. В каких областях сельского хозяйства можно использовать технологии Интернета вещей (IoT) для улучшения производительности и оптимизации процессов?
4. Какие примеры успешного применения компьютерного зрения в сельском хозяйстве существуют и какие конкретные преимущества они принесли в данной отрасли?

### Компетенция ПК-3

#### Задания закрытого типа

1. Вертикальное масштабирование...
  - а) Требуется изменений в прикладных программах, работающих на таких системах**
  - б) Не требуется никаких изменений в прикладных программах, работающих на таких системах
  - в) Уменьшает производительность каждого компонента БД
  - г) Увеличивает скорость загрузки данных
  
2. Для достижения какого свойства в БД типа NoSQL нет JOIN операций?
  - а) Intercepting
  - б) Concurrency**
  - в) Consistency
  - г) Capacity
  
3. Что, согласно теореме CAP (теореме Брюера), возможно обеспечить в любой реализации распределённых вычислений?
  - а) Только согласованность данных
  - б) Только доступность данных
  - в) Согласованность данных, доступность данных, устойчивость к разделению**
  - г) Не более двух свойств из трёх вышеприведённых
  
4. Выберите верное определение понятия AP-система:
  - а) Система, во всех узлах которой данные согласованы и обеспечена доступность, жертвует устойчивостью к распаду на секции
  - б) Распределённая система, в каждый момент обеспечивающая целостный результат и способная функционировать в условиях распада**
  - в) Распределённая система, отказывающаяся от целостности результата
  - г) Система, автоматически изменяющая данные алгоритма своего с целью сохранения оптимального состояния

#### Задания открытого типа

1. Что представляет из себя Big data. Основные отличительные особенности
2. Перечислите особенности работы с регулярными выражениями в языках программирования SQL и R
3. Дайте понятие структурированные данные в контексте анализа данных? Приведите примеры.

4. Какие особенности работы с *tibble* объектами при помощи набора библиотек?

**МАТРИЦА СООТВЕТСТВИЯ КРИТЕРИЕВ ОЦЕНКИ УРОВНЮ  
СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ**

| Критерии оценки                       | Уровень сформированности компетенций |
|---------------------------------------|--------------------------------------|
| <b>Оценка по пятибалльной системе</b> |                                      |
| «Отлично»                             | «Высокий уровень»                    |
| «Хорошо»                              | «Повышенный уровень»                 |
| «Удовлетворительно»                   | «Пороговый уровень»                  |
| «Неудовлетворительно»                 | «Не достаточный»                     |

**Методические материалы, определяющие процедуру оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций**

1. Положение «О балльно-рейтинговой системе аттестации студентов»: СМК ПНД 08-01-2022, введено приказом от 28.09.2011 №371-О (<http://nsau.edu.ru/file/403>: режим доступа свободный);

2. Положение «О проведении текущего контроля и промежуточной аттестации обучающихся в ФГБОУ ВО Новосибирский ГАУ»: СМК ПНД 77-01-2022, введено в действие приказом от 03.08.2015 №268а-О (<http://nsau.edu.ru/file/104821>: режим доступа свободный).

Составители:

ст. преподаватель

(должность)

01.10.2022

дата



подпись

Петров А.Ф.

ФИО

ст. преподаватель

(должность)

01.10.2022

дата



подпись

Шатохин К.С.

ФИО