

Документ подписан простой электронной подписью

Информация о владельце:

ФИО: Узунов Федор Владимирович

Должность: Ректор

Дата подписания: 24.09.2026 14:09:48

Уникальный программный ключ:

fd935d10451b860e912264c0378b448457bfdb607f94388008e39877e6bcb55

**АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОРГАНИЗАЦИЯ
«ОБРАЗОВАТЕЛЬНАЯ ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ»
«УНИВЕРСИТЕТ ЭКОНОМИКИ И УПРАВЛЕНИЯ»**

Факультет экономики, управления и юриспруденции

Кафедра управления и бизнес-информатики



УТВЕРЖДАЮ

Проректор по учебно-методической работе

Г.П. Узунова / Г.П. Узунова

«02» февраля 2026 г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ

АНАЛИЗ БОЛЬШИХ ДАННЫХ

Направление подготовки

09.03.02 «Информационные системы и технологии»

Профиль

«Специалист информационных систем и технологий»

Квалификация

Бакалавр

Для всех

форм обучения

Симферополь, 2026 г.

1. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ

Средства оценивания в ходе текущего контроля:

- устные опросы в ходе лекционных и лабораторных занятий;
- отчеты по лабораторным работам;
- рефераты;
- тестирование;
- задания, выполняемые в ходе лабораторного занятия или рекомендуемые для самостоятельной работы.

Коды компетенции	Результаты освоения ОПОП	Перечень планируемых результатов обучения по дисциплине
ОПК-7	ОПК-7. Способен осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации информационных систем	ОПК-7.1. Знать: методы настройки, наладки программно-аппаратных комплексов. ОПК-7.2. Уметь: анализировать техническую документацию, производить настройку, наладку и тестирование программно-аппаратных комплексов. ОПК-7.3. Владеть: навыками проверки работоспособности программно-аппаратных комплексов

1.1 Вопросы к текущему контролю

1. **Определение:** Раскройте суть понятия «Большие данные». Перечислите классические характеристики (5V): Объем, Скорость, Разнообразие, Достоверность, Ценность.
2. **Отличия от классических БД:** В чем принципиальное отличие подходов к хранению и обработке в традиционных СУБД (SQL) и экосистемах Big Data?
3. **Жизненный цикл данных:** Опишите основные этапы жизненного цикла анализа больших данных: от сбора сырых логов до визуализации результатов.
4. **Источники данных:** Классификация источников структурированных, полуструктурированных и неструктурированных данных (внутренние системы предприятия vs внешние потоки IoT, соцсети).
5. **Проблема качества данных:** Основные типы «грязных» данных (пропуски, выбросы, дубликаты). Стратегии их очистки на этапе препроцессинга.
6. **HDFS:** Архитектура распределенной файловой системы Hadoop (HDFS). Роль *NameNode* и *DataNode*. Принцип дробления блоков (*splitting*) и репликации.
7. **NoSQL базы данных:** Сравнительный анализ типов NoSQL хранилищ: ключ-значение (*Redis*), документы (*MongoDB*), колонки (*Cassandra*), графы (*Neo4j*).
8. **MapReduce:** Объясните парадигму MapReduce. Что происходит на этапах *Map* и *Reduce*? Приведите пример задачи подсчета слов (*Word Count*).
9. **Spark vs Hadoop:** В чем преимущество вычислительного движка Apache Spark над классическим MapReduce с точки зрения использования оперативной памяти (*RDD*, *DAG*)?
10. **Lambda Architecture:** Описание слоев Лямбда-архитектуры (*Batch Layer*, *Speed Layer*, *Serving Layer*). Проблемы поддержки двух кодовых баз.

11. **Kappa Architecture:** Суть упрощенной Каппа-архитектуры как альтернативы Лямбде.
12. **Сценарии применения:** Примеры задач, требующих обработки потоков в реальном времени (фрод-мониторинг, предиктивное обслуживание оборудования).
13. **Pandas:** Назначение библиотек Series и DataFrame. Отличие методов .loc[] (label-based) от .iloc[] (position-based).
14. **Векторизация NumPy:** Почему операции в библиотеке NumPy выполняются значительно быстрее, чем циклы на чистом Python?
15. **Агрегация данных:** Использование метода .groupby() в Pandas для вычисления сводной статистики по категориям.
16. **Пропущенные значения:** Методы работы с пропусками (NaN) в Pandas: удаление строк/столбцов (dropna) против заполнения значениями (fillna, интерполяция).
17. **Feature Engineering:** Процесс создания признаков. Масштабирование (*StandardScaler*, *MinMaxScaler*) и кодирование категориальных переменных (*One-Hot Encoding*).
18. **Обучение без учителя:** Применение алгоритмов кластеризации (например, K-means или DBSCAN) для сегментации клиентов в маркетинговом анализе.
19. **Классификация и регрессия:** Выбор между алгоритмами решающих деревьев и логистической регрессией. Интерпретируемость моделей («белый ящик» против «черного ящика»).
20. **Оценка качества:** Какие метрики вы выберете для оценки модели предсказания оттока клиентов, если количество ушедших составляет менее 5% от всей выборки? Обоснуйте выбор Precision/Recall вместо Accuracy.

1.2 Темы рефератов

1. **Эволюция парадигмы Big Data:** переход от централизованных хранилищ (*Data Warehouse*) к распределенным системам озер данных (*Data Lake*).
2. **Архитектура Lambda vs Kappa:** сравнительный анализ подходов к обработке пакетных и потоковых данных в реальном времени.
3. **Роль проекта Apache Hadoop** в становлении индустрии анализа больших данных: история создания *HDFS* и *MapReduce*.
4. **CAP-теорема** и компромиссы согласованности, доступности и устойчивости к разделению в распределенных базах данных (*NoSQL*).
5. **Экономика данных:** влияние национальных программ цифровизации на развитие инфраструктуры сбора и хранения информации.
6. **Сравнительный анализ NoSQL СУБД:** документно-ориентированные (*MongoDB*), колоночные (*Cassandra*) и графовые (*Neo4j*) базы данных.
7. **Apache Spark как замена MapReduce:** преимущества вычислений в оперативной памяти (*In-memory processing*) и архитектура *DAG*.
8. **Потоковая обработка событий:** обзор фреймворков *Apache Kafka*, *Flink* и их применение в задачах мониторинга транзакций.
9. **Облачные озера данных:** архитектурные особенности платформ *Amazon S3*, *Google BigQuery* и *Azure Data Lake*.
10. **Методы очистки «грязных» данных:** стратегии работы с пропущенными значениями (*imputation*) и выбросами (*outlier detection*).
11. **Feature Engineering:** техники масштабирования, нормализации и кодирования категориальных переменных для высокоразмерных данных.
12. **Проблема снижения размерности:** метод главных компонент (*PCA*) и стохастическое вложение соседей с t-распределением (*t-SNE*).

13. **Разведочный анализ данных (EDA):** визуализация многомерных массивов и поиск скрытых корреляций перед построением моделей.
14. **Масштабируемое машинное обучение:** алгоритмы градиентного бустинга над большими таблицами (*XGBoost, LightGBM, CatBoost*).
15. **Кластерный анализ петабайтных объемов:** модификация алгоритма K-means для распределенных вычислений.
16. **Ансамблевые методы:** построение коллективов решающих правил при расчете коэффициентов нейронных сетей и деревьев решений.
17. **Оценка качества моделей:** метрики точности, полноты и F1-меры в условиях сильного дисбаланса классов (*Imbalanced data*).
18. **Интеллектуальный анализ текстов (Text Mining):** векторизация документов методами *TF-IDF* и *Word Embeddings (Word2Vec)*.
19. **Анализ социальных медиа (Social Media Analytics):** методы sentiment-анализа и тематического моделирования (*LDA*) сообщений пользователей.
20. **Предиктивная аналитика интернета вещей (IoT Predictive Maintenance):** прогнозирование отказов оборудования на основе временных рядов сенсорных данных.

1.3 Тестовые задания

Тест 1. Характеристики (5V)

Вопрос: Какая из характеристик описывает скорость генерации данных и необходимость их обработки в реальном времени?

- а) Volume (Объем)
- б) Variety (Разнообразие)
- в) Velocity (Скорость)
- г) Veracity (Достоверность)

Ответ: в) Velocity

Тест 2. Определение проблемы

Вопрос: Что является главной технической проблемой Больших данных согласно классическому определению?

- а) Высокая стоимость серверов
- б) Невозможность обработки существующих объемов традиционными средствами за приемлемое время
- в) Отсутствие квалифицированных программистов
- г) Сложность визуализации графиков

Ответ: б) Невозможность обработки...

Тест 3. Источники данных

Вопрос: Какой источник НЕ относится к типичным генераторам неструктурированных Big Data?

- а) Логи веб-серверов
- б) Посты в социальных сетях с фото и видео
- в) Таблицы Excel со строгой структурой бухгалтерского учета
- г) Поточковые данные с датчиков IoT

Ответ: в) Таблицы Excel...

Тест 4. Hadoop Ecosystem

Вопрос: Какую роль выполняет компонент *NameNode* в экосистеме *HDFS*?

- а) Хранит сами блоки данных
- б) Управляет метаданными файловой системы и пространством имен
- в) Выполняет код на языке Python
- г) Обеспечивает безопасность сети

Ответ: б) Управляет метаданными...

Тест 5. MapReduce

Вопрос: В чем заключается основная задача этапа *Map* в модели распределенных вычислений?

- а) Агрегация результатов от всех узлов
- б) Фильтрация и предварительная трансформация входных пар ключ-значение
- в) Запись итогового отчета в базу данных
- г) Сортировка жестких дисков

Ответ: б) Фильтрация и предварительная трансформация...

Тест 6. Spark vs Hadoop

Вопрос: Главное преимущество Apache Spark перед классическим MapReduce:

- а) Использование оперативной памяти (*In-memory processing*) для ускорения итерационных алгоритмов
- б) Бесплатная лицензия
- в) Работа только под операционной системой Linux
- г) Отсутствие необходимости в Java

Ответ: а) Использование оперативной памяти...

Тест 7. NoSQL базы данных

Вопрос: К какому типу NoSQL СУБД относится MongoDB?

- а) Ключ-Значение
- б) Графовая
- в) Документно-ориентированная
- г) Колоночная

Ответ: в) Документно-ориентированная

Тест 8. Pandas (Python)

Вопрос: Какой объект библиотеки Pandas представляет собой двумерную размеченную структуру данных (таблицу)? а) Series б) DataFrame в) ndarray г) Matrix

Ответ: б) DataFrame

Тест 9. Пропущенные значения

Вопрос: Какой метод в Pandas используется для удаления строк, содержащих пропущенные значения (*NaN*)?

- а) `.dropna()`
- б) `.fillna(0)`
- в) `.isnull()`
- г) `.replace()`

Ответ: а) `.dropna()`

Тест 10. Feature Engineering

Вопрос: Для чего применяется масштабирование признаков (*Feature Scaling*) в машинном обучении?

- а) Чтобы привести разные признаки к сопоставимому масштабу (например, возраст от 0 до 100 и доход от 0 до 1 млн)
- б) Чтобы удалить лишние столбцы
- в) Чтобы превратить текст в числа
- г) Чтобы увеличить объем выборки

Ответ: а) Чтобы привести разные признаки к сопоставимому масштабу...

Тест 11. Оценка моделей

Вопрос: Если в наборе данных 99% отрицательных примеров и 1% положительных (диагностика редкой болезни), какая метрика будет наиболее адекватной для оценки качества модели? а) Ассурасу (Точность) б) Precision / Recall (точность/полнота) в) Среднеквадратичная ошибка г) Коэффициент детерминации R^2

Ответ: б) Precision / Recall

Тест 12. Кластеризация

Вопрос: Какой алгоритм кластеризации требует предварительного указания количества искомых групп (*K*)?

- а) DBSCAN
- б) K-means
- в) Agglomerative Clustering
- г) Isolation Forest

Ответ: б) K-means

Тест 13. Регрессия

Вопрос: Задача предсказания цены квартиры на основе её площади и района — это пример: а) Задачи классификации

- б) Задачи регрессии
- в) Задачи кластеризации
- г) Задачи снижения размерности

Ответ: б) Задачи регрессии

Тест 14. Аналитика временных рядов

Вопрос: Как называется устойчивый долгосрочный тренд роста или падения показателя во временном ряду (например, рост населения Земли)?

- а) Цикличность
- б) Сезонность
- в) Тренд
- г) Шум

Ответ: в) Тренд

Тест 15. NLP (Обработка текста)

Вопрос: Процесс приведения словоформы к его начальной форме (например, «бегу», «бежал» -> «бежать») называется:

- а) Токенизация
- б) Стемминг
- в) Лемматизация г) Удаление стоп-слов

Ответ: в) Лемматизация

1. Установите соответствие между оператором SQL и его назначением:

Оператор	Назначение
1. SELECT	А) Извлечение данных из базы
2. INSERT	Б) Добавление новых записей в таблицу
3. UPDATE	В) Изменение существующих записей
4. DELETE	Г) Удаление записей из таблицы

Правильный ответ: 1-А, 2-Б, 3-В, 4-Г

17. Установите соответствие между типом соединения и его описанием:

Тип соединения	Описание
1. INNER JOIN	А) Возвращает только строки,

Тип соединения	Описание
	имеющие соответствие в обеих таблицах
2. LEFT JOIN	Б) Возвращает все строки из левой таблицы и соответствующие строки из правой
3. RIGHT JOIN	В) Возвращает все строки из правой таблицы и соответствующие строки из левой
4. FULL JOIN	Г) Возвращает все строки из обеих таблиц

Правильный ответ: 1-А, 2-Б, 3-В, 4-Г

18. Установите соответствие между агрегатной функцией и её действием:

Функция	Действие
1. COUNT	А) Подсчёт количества строк
2. SUM	Б) Сумма значений в столбце
3. AVG	В) Среднее арифметическое значений
4. MAX	Г) Максимальное значение в столбце

Правильный ответ: 1-А, 2-Б, 3-В, 4-Г

19. Установите соответствие между нормальной формой и её условием:

Нормальная форма	Условие
1. 1НФ	А) Все атрибуты атомарны
2. 2НФ	Б) Отсутствие частичной зависимости неключевых атрибутов от составного ключа
3. 3НФ	В) Отсутствие транзитивных зависимостей неключевых атрибутов

Правильный ответ: 1-А, 2-Б, 3-В

20. Установите соответствие между свойством ACID и его описанием:

Свойство	Описание
1. Atomicity	А) «Всё или ничего»
2. Consistency	Б) Сохранение целостности данных
3. Isolation	В) Параллельные транзакции не влияют друг на друга
4. Durability	Г) Результаты зафиксированной транзакции сохраняются при сбоях

Правильный ответ: 1-А, 2-Б, 3-В, 4-Г

21. Расположите этапы проектирования базы данных в правильном порядке:

1. Физическое проектирование
2. Концептуальное проектирование (ER-диаграмма)
3. Логическое проектирование (преобразование в реляционную модель)

(Правильный ответ: $2 \rightarrow 3 \rightarrow 1$)

22. Расположите нормальные формы в порядке увеличения строгости требований:

1. Первая нормальная форма (1НФ)
2. Вторая нормальная форма (2НФ)
3. Третья нормальная форма (3НФ)

(Правильный ответ: $1 \rightarrow 2 \rightarrow 3$)

23. Расположите разделы оператора SELECT в порядке их обработки СУБД:

1. FROM
2. WHERE
3. GROUP BY
4. HAVING
5. SELECT
6. ORDER BY

(Правильный ответ: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$)

24. Расположите команды в хронологическом порядке их появления в стандарте SQL:

1. SELECT
2. JOIN
3. WITH (CTE)

(Правильный ответ: $1 \rightarrow 2 \rightarrow 3$)

25. Расположите уровни изоляции транзакций в порядке увеличения строгости (от наименее строгого):

1. READ UNCOMMITTED
2. READ COMMITTED
3. REPEATABLE READ
4. SERIALIZABLE

(Правильный ответ: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$)

26. Какой оператор используется для удаления таблицы из базы данных?

а) DELETE

- б) DROP (*Правильный ответ: б)*
- в) REMOVE
- г) TRUNCATE

27. Какое ключевое слово используется для устранения дубликатов в результате SELECT?

- а) UNIQUE
- б) DISTINCT (*Правильный ответ: б)*
- в) DIFFERENT
- г) SINGLE

28. Какой тип данных следует выбрать для хранения больших текстовых данных (до 4 ГБ)?

- а) VARCHAR
- б) CHAR
- в) TEXT / CLOB (*Правильный ответ: в)*
- г) BLOB

29. Какой оператор используется для изменения структуры существующей таблицы?

- а) MODIFY TABLE
- б) ALTER TABLE (*Правильный ответ: б)*
- в) UPDATE TABLE
- г) CHANGE TABLE

30. Что произойдёт при выполнении запроса SELECT COUNT(*) FROM table_name;?

- а) Вернёт сумму всех числовых полей
- б) Вернёт количество всех строк в таблице, включая NULL (*Правильный ответ: б)*
- в) Вернёт количество уникальных строк
- г) Вернёт количество столбцов в таблице

1.4 Задания

Задача 1. Оптимизация запросов

Условие: Дан запрос SELECT * FROM sales WHERE YEAR(date) = 2023. Объясните, почему использование функции над колонкой (YEAR(date)) делает невозможным использование индекса по полю date. Предложите переписанный вариант запроса для эффективного поиска по индексу.

Задача 2. Проектирование Data Warehouse

Условие: Разработайте схему «Звезда» (Star Schema) для анализа продаж интернет-магазина. Определите одну таблицу фактов (продажи) и три таблицы измерений (Товары, Время, Клиенты), указав ключевые поля.

Задача 3. Работа с JSON (Полуструктурированные данные)

Условие: В таблице логов есть колонка event_data типа JSON: {"user_id": 1, "action": "click", "item": {"id": 42, "price": 100}}. Напишите SQL-запрос (используя синтаксис PostgreSQL или BigQuery), который извлекает идентификатор товара из вложенного объекта.

Задача 4. Оконные функции

Условие: Имеется таблица транзакций (customer_id, transaction_date, amount). Напишите запрос, который выведет для каждой транзакции сумму текущей покупки и нарастающий итог покупок данного клиента за всё время до текущего момента.

Задача 5. Обработка дубликатов

Условие: Опишите алгоритм удаления полных дубликатов строк в таблице объемом 1 ТБ при ограничении оперативной памяти сервера в 64 ГБ. Почему стандартный DELETE с подзапросом может не сработать?

Задача 6. Сравнение производительности

Условие: Даны два списка чисел длиной 1 миллион элементов. Реализуйте двумя способами вычисление суммы квадратов всех чисел: через цикл for и через векторизованную операцию библиотеки NumPy. Объясните разницу в скорости выполнения.

Задача 7. Группировка и агрегация

Условие: Используя библиотеку Pandas, напишите код для DataFrame df, который сгруппирует данные по столбцу City и посчитает среднее значение Salary и количество сотрудников в каждом городе.

Задача 8. Объединение таблиц (Join)

Условие: Есть датафреймы users(id, name) и orders(user_id, product). Выполните левое объединение (*left join*), чтобы вывести имена пользователей, которые НЕ совершали заказов.

Задача 9. Пропуски в данных

Условие: В числовом столбце температур датчика пропущено 10% значений случайным образом. Опишите недостатки метода заполнения средним значением (mean imputation) и предложите более сложный метод (например, интерполяцию).

Задача 10. Выбросы (Outliers)

Условие: Дан ряд данных о доходах населения. Примените метод межквартильного размаха (IQR), чтобы определить границы выбросов. Постройте Box-plot (ящик с усами) для визуализации результата.

Задача 11. Проверка гипотез

Условие: Проведено A/B тестирование двух версий сайта. Конверсия версии А — 5%, версии Б — 5.5%. Объем выборки — 100 000 человек. Сформулируйте нулевую и альтернативную гипотезы и опишите метрики для проверки статистической значимости различия.

Задача 12. Мультиколлинеарность

Условие: При построении модели линейной регрессии обнаружено, что корреляция между признаками height_in_cm и height_in_inches близка к 1. Как называется эта проблема и какой метод её устранения вы примените перед обучением модели?

Задача 13. Масштабирование признаков

Условие: Для алгоритма K-Nearest Neighbors (KNN) критически важен масштаб данных. Почему? Что произойдет, если признак «Вес» (от 0 до 200 кг) оставить без изменений, а признак «Рост» перевести в диапазон от 0 до 1?

Задача 14. Матрица ошибок (Confusion Matrix)

Условие: Модель спам-фильтра выдала следующие результаты: Истинный спам определен как спам (TP=90), нормальный email определен как спам (FP=10), спам пропущен (FN=5), нормальный email определен верно (TN=85). Рассчитайте Precision и Recall.

Задача 15. Кластеризация текста

Условие: У вас есть коллекция новостных статей. Какой алгоритм машинного обучения вы выберете для автоматического разделения их по темам (спорт, политика, технологии) без предварительной разметки учителем? Обоснуйте выбор.

Задача 16. Регуляризация

Условие: Ваша модель линейной регрессии переобучается (высокая дисперсия). Какой параметр регуляризации нужно изменить в Ridge/Lasso регрессии, чтобы уменьшить сложность модели?

Задача 17. CAP-теорема

Условие: Распределенная база данных работает в режиме AP (*Availability + Partition tolerance*). Что происходит с согласованностью данных (*Consistency*) при сетевом разрыве между узлами? Приведите пример такой СУБД.

Задача 18. MapReduce логика

Условие: Опишите словами решение задачи Word Count (подсчет слов) через Mapper и Reducer. Какие пары ключ-значение выдаст Mapper для строки «Hello World Hello»?

Задача 19. Графовые базы данных

Условие: Вам нужно найти кратчайший путь («друзей друзей») между двумя пользователями в социальной сети. Почему реляционная БД будет неэффективна для этого, и какую особенность графовой БД (*Neo4j*) вы используете?

Задача 20. Поточковая обработка (Streaming)

Условие: Датчик IoT отправляет показания каждую секунду. Почему нельзя просто записывать эти данные напрямую в классическую реляционную БД (OLTP) и какая технология (очередь сообщений) обычно ставится между датчиком и хранилищем для буферизации потока?

2. КРИТЕРИИ ОЦЕНИВАНИЯ ПРИ ПРОВЕДЕНИИ ТЕКУЩЕГО КОНТРОЛЯ

Вид контроля	Наименование работы	Наименование оценочных средств	Шкала оценивания
Текущий контроль	Вопросы для обсуждения на занятиях; Устные опросы по ранее изученному материалу; Письменные работы: рефераты, тестовые задания; Практические задания; Рефераты и доклады по темам (вопросам), вынесенным на самостоятельную работу.	Оценка выступлений на практическом (семинарском) занятии, проверка заданий, устный опрос, оценивание докладов, рефератов	отлично хорошо удовлетворительно неудовлетворительно

Критерии оценивания устных ответов обучающихся

Шкала оценивания	Характеристика оценивания
отлично	оценивается ответ, который показывает прочные знания основных процессов изучаемой предметной области, отличается глубиной и полнотой раскрытия темы; владение терминологическим аппаратом; умение объяснять сущность, явлений, процессов, событий, делать выводы и обобщения, давать аргументированные ответы, приводить примеры; свободное владение монологической речью, логичность и последовательность ответа.
хорошо	оценивается ответ, обнаруживающий прочные знания основных процессов изучаемой предметной области, отличается глубиной и полнотой раскрытия темы; владение терминологическим аппаратом; умение объяснять сущность, явлений, процессов, событий, делать выводы и обобщения, давать аргументированные ответы, приводить примеры; свободное владение монологической речью, логичность и последовательность ответа. Однако допускается одна - две неточности в ответе.
удовлетворительно	оценивается ответ, свидетельствующий в основном о знании процессов изучаемой предметной области, отличающийся недостаточной глубиной и полнотой раскрытия темы; знанием основных вопросов теории; слабо сформированными навыками анализа явлений, процессов, недостаточным умением давать аргументированные ответы и приводить примеры; недостаточно свободным владением монологической речью, логичностью и последовательностью ответа. Допускается несколько ошибок в содержании ответа.
неудовлетворительно	оценивается ответ, обнаруживающий незнание процессов изучаемой предметной области, отличающийся неглубоким раскрытием темы; незнанием основных вопросов теории, несформированными навыками анализа явлений, процессов; неумением давать аргументированные ответы, слабым владением монологической речью, отсутствием логичности и последовательности. Допускаются серьезные ошибки в содержании ответа.

Критерии оценивания работы обучающихся на семинарских занятиях

Шкала оценивания	Показатели	Критерии
------------------	------------	----------

Шкала оценивания	Показатели	Критерии
Отлично	1. Полнота выполнения практического и тестового задания (полнота ответа); 2. Своевременность выполнения задания; 3. Последовательность и рациональность выполнения	Задание решено самостоятельно. При этом составлен правильный алгоритм решения задания, в логических рассуждениях, в выборе формул и решении нет ошибок, получен верный ответ, задание решено рациональным способом. Дан правильный и исчерпывающий ответ на поставленные теоретические и тестовые вопросы, в которых обучающийся показал всестороннее системное знание программного материала, усвоение основной и дополнительной литературы, четкое владение понятийным аппаратом.
Хорошо	практического задания (логичность и четкость ответа); 4. Правильность ответов на вопросы; 5. Самостоятельность решения (владение дополнительным материалом); 6. Знание нормативно-законодательной базы и терминологии курса	Задание решено с помощью преподавателя. При этом составлен правильный алгоритм решения задания, в логическом рассуждении и решении нет существенных ошибок; правильно сделан выбор формул для решения; есть объяснение решения, но задание решено нерациональным способом или допущено не более двух несущественных ошибок, получен верный ответ. На поставленные теоретические и тестовые вопросы, при которых обучающийся показал достаточный уровень знаний основного программного материала: освоение информации лекционного курса и учебных пособий, овладение понятийным аппаратом, методикой исследований при попытке анализа различных ситуаций.
Удовлетворительно		Задание решено с подсказками преподавателя. Задание решено в общем виде. Обучающийся показал средний уровень знаний основного программного материала, но не мог убедительно аргументировать свой ответ, ошибся в использовании понятийного аппарата, показал недостаточные знания литературных источников.
Неудовлетворительно		Задание не решено. Обучающийся продемонстрировал значительные пробелы в знаниях основного программного материала, не аргументировал свой ответ, показал неудовлетворительные знания понятийного аппарата и специальной литературы.

Критерии оценивания рефератов

Средство контроля	Критерии оценивания	Шкала оценивания
Реферат	Выполнены все требования к написанию и защите реферата: обозначена проблема и обоснована её актуальность, сделан краткий анализ различных точек зрения на рассматриваемую проблему и логично изложена собственная позиция, сформулированы выводы, тема раскрыта полностью, выдержан объём, соблюдены требования к внешнему оформлению, даны правильные ответы на дополнительные вопросы.	отлично

	Реферат раскрывает поднятую проблематику в полном объеме. Основные требования к реферату и его защите выполнены, но при этом допущены недочёты. В частности, имеются неточности в изложении материала; отсутствует логическая последовательность в суждениях; не выдержан объём реферата; имеются упущения в оформлении; на дополнительные вопросы при защите даны неполные ответы. В реферате имеются неточности и предметная область выступления раскрыта не в полной мере.	хорошо
	Имеются существенные отступления от требований к реферированию. В частности: тема освещена лишь частично; допущены фактические ошибки в содержании реферата или при ответе на дополнительные вопросы; во время защиты отсутствует вывод. В реферате не в полной степени раскрыт понятийный аппарат, имеются существенные неточности в процессе формирования выводов.	удовлетворительно
	Тема реферата не раскрыта, обнаруживается существенное непонимание проблемы. Тема реферата не раскрыта или выполнена не по существу ранее поставленного вопроса. Реферат не сдан / доклад не сделан.	неудовлетворительно

Критерии оценивания тестов

Средство контроля	Критерии оценивания – процент положительных ответов	Шкала оценивания
Тестирование	90-100	отлично
	70-89	хорошо
	40-69	удовлетворительно
	< 39	неудовлетворительно

3. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Средства оценивания в ходе промежуточной аттестации:

- вопросы к зачету с оценкой;
- практические задания зачета.

Коды компетенции	Результаты освоения ОПОП	Перечень планируемых результатов обучения по дисциплине

ОПК-1	ОПК-7. Способен осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации информационных систем	ОПК-7.1. Знать: методы настройки, наладки программно-аппаратных комплексов. ОПК-7.2. Уметь: анализировать техническую документацию, производить настройку, наладку и тестирование программно-аппаратных комплексов. ОПК-7.3. Владеть: навыками проверки работоспособности программно-аппаратных комплексов
-------	---	--

3.1 Вопросы к зачету с оценкой

1. Определение: Дайте определение понятия «Большие данные». В чем принципиальное отличие от традиционных баз данных?
2. Характеристики (5V): Раскройте суть пяти основных характеристик Больших данных: Volume, Velocity, Variety, Veracity, Value.
3. Источники данных: Приведите классификацию источников данных (внутренние системы предприятия, интернет вещей — IoT, социальные медиа).
4. Жизненный цикл данных: Опишите основные этапы работы с большими данными: генерация, сбор, хранение, очистка, анализ, визуализация.
5. Проблематика: Почему возникла необходимость в новых технологиях обработки? Что такое «узкое место» (*bottleneck*) традиционной архитектуры при работе с петабайтами данных?
6. Data Lake vs Data Warehouse: Сравните концепции озера данных и хранилища данных.
7. Экосистема Hadoop: Назовите ключевые компоненты экосистемы Apache Hadoop и их назначение.
8. HDFS: Объясните архитектуру распределенной файловой системы *HDFS*. Роль *NameNode* и *DataNode*.
9. MapReduce: Опишите парадигму программирования MapReduce на примере задачи подсчета слов (*Word Count*).
10. NoSQL СУБД: Классификация нереляционных баз данных (ключ-значение, документы, колонки, графы). Примеры каждой категории.
11. CAP-теорема: Сформулируйте теорему CAP. Какой тип согласованности обычно жертвуется в системах класса *BigTable/DynamoDB*?
12. Apache Spark: В чем преимущество вычислительного движка Spark перед классическим MapReduce?
13. RDD и DataFrame: Понятие устойчивых распределенных коллекций (*RDD*) и датафреймов в Spark.
14. Поточковая обработка: Различие между пакетной (*batch*) и потоковой (*streaming*) обработкой. Микропакетная модель Spark Structured Streaming.
15. Kafka: Какую роль играет *Apache Kafka* в архитектуре современных аналитических систем?
16. Библиотека Pandas: Основные структуры данных (*Series*, *DataFrame*). Отличие label-based индексации (*loc*) от позиционной (*iloc*).
17. Пропущенные значения: Методы поиска и обработки пропусков в Pandas (*isnull()*, *dropna()*, *fillna()*).
18. Группировка и агрегация: Использование метода *.groupby()* для вычисления сводной статистики.

19. Векторизация NumPy: Почему операции над массивами NumPy выполняются быстрее, чем циклы Python?
20. Feature Engineering: Процесс создания признаков. Масштабирование и нормализация данных.
21. Обучение без учителя (Кластеризация): Суть алгоритма K-means. Проблема выбора числа кластеров (K).
22. Линейные модели: Градиентный спуск как метод оптимизации функции потерь.
23. Деревья решений и ансамбли: Принцип работы алгоритма Random Forest и Gradient Boosting.
24. Оценка качества: Матрица ошибок (*Confusion Matrix*). Когда использовать метрику Recall вместо Accuracy?
25. Переобучение (*Overfitting*): Причины возникновения и методы борьбы (регуляризация L1/L2, кросс-валидация).
26. NLP (Текстовый анализ): Этапы предобработки текста (токенизация, лемматизация, удаление стоп-слов).
27. Тематическое моделирование (*LDA*): Цель применения латентного размещения Дирихле для анализа больших корпусов текстов.
28. Рекомендательные системы: Разница между коллаборативной фильтрацией и контентной фильтрацией.
29. A/B тестирование: Статистическая значимость результатов эксперимента. Понятия нулевой гипотезы и p-value.
30. Визуализация многомерных данных: Применение методов снижения размерности (PCA, t-SNE) для построения графиков.

3.2 Практические задания к зачету с оценкой

Задача 1. Оптимизация запросов

Условие: Дан запрос `SELECT * FROM sales WHERE YEAR(date) = 2023`. Объясните, почему использование функции над колонкой (`YEAR(date)`) делает невозможным использование индекса по полю `date`. Предложите переписанный вариант запроса для эффективного поиска по индексу.

Задача 2. Проектирование Data Warehouse

Условие: Разработайте схему «Звезда» (*Star Schema*) для анализа продаж интернет-магазина. Определите одну таблицу фактов (продажи) и три таблицы измерений (Товары, Время, Клиенты), указав ключевые поля.

Задача 3. Работа с JSON (Полуструктурированные данные)

Условие: В таблице логов есть колонка `event_data` типа JSON: `{"user_id": 1, "action": "click", "item": {"id": 42, "price": 100}}`. Напишите SQL-запрос (используя синтаксис PostgreSQL или BigQuery), который извлекает идентификатор товара из вложенного объекта.

Задача 4. Оконные функции

Условие: Имеется таблица транзакций (`customer_id`, `transaction_date`, `amount`). Напишите запрос, который выведет для каждой транзакции сумму текущей покупки и нарастающий итог покупок данного клиента за всё время до текущего момента.

Задача 5. Обработка дубликатов

Условие: Опишите алгоритм удаления полных дубликатов строк в таблице объемом 1 ТБ при ограничении оперативной памяти сервера в 64 ГБ. Почему стандартный DELETE с подзапросом может не сработать?

Задача 6. Сравнение производительности

Условие: Даны два списка чисел длиной 1 миллион элементов. Реализуйте двумя способами вычисление суммы квадратов всех чисел: через цикл `for` и через векторизованную операцию библиотеки NumPy. Объясните разницу в скорости выполнения.

Задача 7. Группировка и агрегация

Условие: Используя библиотеку Pandas, напишите код для DataFrame df, который сгруппирует данные по столбцу City и посчитает среднее значение Salary и количество сотрудников в каждом городе.

Задача 8. Объединение таблиц (Join)

Условие: Есть датафреймы users(id, name) и orders(user_id, product). Выполните левое объединение (*left join*), чтобы вывести имена пользователей, которые НЕ совершали заказов.

Задача 9. Пропуски в данных

Условие: В числовом столбце температур датчика пропущено 10% значений случайным образом. Опишите недостатки метода заполнения средним значением (mean imputation) и предложите более сложный метод (например, интерполяцию).

Задача 10. Выбросы (Outliers)

Условие: Дан ряд данных о доходах населения. Примените метод межквартильного размаха (*IQR*), чтобы определить границы выбросов. Постройте Box-plot (ящик с усами) для визуализации результата.

Задача 11. Проверка гипотез

Условие: Проведено А/В тестирование двух версий сайта. Конверсия версии А — 5%, версии В — 5.5%. Объем выборки — 100 000 человек. Сформулируйте нулевую и альтернативную гипотезы и опишите метрики для проверки статистической значимости различия.

Задача 12. Мультиколлинеарность

Условие: При построении модели линейной регрессии обнаружено, что корреляция между признаками height_in_cm и height_in_inches близка к 1. Как называется эта проблема и какой метод её устранения вы примените перед обучением модели?

Задача 13. Масштабирование признаков

Условие: Для алгоритма K-Nearest Neighbors (*KNN*) критически важен масштаб данных. Почему? Что произойдет, если признак «Вес» (от 0 до 200 кг) оставить без изменений, а признак «Рост» перевести в диапазон от 0 до 1?

Задача 14. Матрица ошибок (Confusion Matrix)

Условие: Модель спам-фильтра выдала следующие результаты: Истинный спам определен как спам (*TP*=90), нормальный email определен как спам (*FP*=10), спам пропущен (*FN*=5), нормальный email определен верно (*TN*=85). Рассчитайте Precision и Recall.

Задача 15. Кластеризация текста

Условие: У вас есть коллекция новостных статей. Какой алгоритм машинного обучения вы выберете для автоматического разделения их по темам (спорт, политика, технологии) без предварительной разметки учителем? Обоснуйте выбор.

Задача 16. Регуляризация

Условие: Ваша модель линейной регрессии переобучается (высокая дисперсия). Какой параметр регуляризации нужно изменить в Ridge/Lasso регрессии, чтобы уменьшить сложность модели?

Задача 17. CAP-теорема

Условие: Распределенная база данных работает в режиме AP (*Availability + Partition tolerance*). Что происходит с согласованностью данных (*Consistency*) при сетевом разрыве между узлами? Приведите пример такой СУБД.

Задача 18. MapReduce логика

Условие: Опишите словами решение задачи Word Count (подсчет слов) через Mapper и Reducer. Какие пары ключ-значение выдаст Mapper для строки «Hello World Hello»?

Задача 19. Графовые базы данных

Условие: Вам нужно найти кратчайший путь («друзей друзей») между двумя пользователями в социальной сети. Почему реляционная БД будет неэффективна для этого, и какую особенность графовой БД (*Neo4j*) вы используете?

Задача 20. Потокковая обработка (Streaming)

Условие: Датчик IoT отправляет показания каждую секунду. Почему нельзя просто записывать эти данные напрямую в классическую реляционную БД (OLTP) и какая технология (очередь сообщений) обычно ставится между датчиком и хранилищем для буферизации потока?

4. КРИТЕРИИ ОЦЕНИВАНИЯ ПРИ ПРОВЕДЕНИИ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Шкала оценивания уровня сформированности компетенций (по пятибалльной системе) зачет с оценкой

Формируемые уровни освоения компетенций	Критерии оценивания	Шкала оценивания
Высокий уровень	Изложено правильное понимание вопроса, четко и самостоятельно дан исчерпывающий ответ, содержание раскрыто полно, профессионально, грамотно. Обучающимся усвоена взаимосвязь основных понятий дисциплины в их значении для приобретаемой профессии. Отражает успешное и систематическое применение навыков и умений по данной дисциплине в соответствии с ФГОС.	отлично
Базовый уровень	Изложено правильное понимание вопроса, дано достаточно подробное описание предмета ответа, приведены и раскрыты в тезисной форме основные понятия, относящиеся к предмету ответа. Ответ отражает полное знание учебно-программного материала, систематический характер знаний по дисциплине, а также наличие базового уровня овладения практическими умениями и навыками по данной дисциплине в соответствии с ФГОС	хорошо
Пороговый уровень	Ответ отражает теоретические знания основного учебно-программного материала в объеме, необходимом для дальнейшей учебы и предстоящей работы по профессии. Данная оценка может быть выставлена обучающемуся, допустившему неточности в ответе, но обладающими необходимыми знаниями для их устранения под руководством преподавателя, отмечен начальный уровень овладения практическими умениями и навыками по данной дисциплине в соответствии с ФГОС	удовлетворительно

Неудовлетворительный уровень	При ответе обучающегося обнаружено отсутствие знаний, умений и навыков и/или фрагментарные знания основного учебно-программного материала.	неудовлетворительно
------------------------------	--	---------------------

Текущий контроль и промежуточная аттестация осуществляются в соответствии с «Положением о текущей и промежуточной аттестации обучающихся в Автономной некоммерческой организации «Образовательная организация высшего образования» «Университет экономики и управления».

Вид промежуточной аттестации – зачет с оценкой.

Форма проведения промежуточной аттестации – письменный зачет.