

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ
ПРАКТИЧЕСКИХ РАБОТ ПО ДИСЦИПЛИНЕ**

Введение в большие данные

Направление подготовки 39.04.01 Социология
Направленность (профиль) – «Большие данные и цифровая аналитика»
Форма обучения – Очная
Год начала обучения – 2026 г.

Ставрополь, 2026

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	3
Тема № 1. Введение в большие данные: информация, большие данные, Big Data.....	6
Тема № 2. Основы управления большими данными.....	10
Тема № 3. Жизненный цикл данных.....	3
Тема № 4. Жизненный цикл аналитики данных.....	17
Тема № 5. Метаданные.....	20
Тема № 6. Технологии работы с большими данными.....	24
Тема № 7. Хранилища данных: эволюция и общие основы.....	27
Тема № 8. Современные технологии хранения данных: масштабирование и многоуровневое хранение данных.....	30
Тема № 9. Системы управления большими данными.....	34
Тема № 10. Аналитика больших данных и ее инструментарий.....	38
Тема № 11. Аналитика больших данных: техники обработки и анализа.....	41
Тема № 12. Архитектура системы обработки больших данных.....	44
Тема № 13. Параллельные алгоритмы для работы с данными.....	47
Тема № 14. Программные платформы и системы для больших данных.....	50
Тема № 15. Оборудование и центры обработки больших данных.....	54
Тема № 16. Визуализация данных и результатов анализа.....	59
Тема № 17. Классификация задач анализа данных.....	63
Тема № 18. Статистические методы анализа данных.....	67
СПИСОК ЛИТЕРАТУРЫ.....	71

ВВЕДЕНИЕ

За последнее десятилетие объем данных, с которыми приходится иметь дело, многократно увеличился, и в то же время стоимость хранения данных снизилась. Частные компании и исследовательские учреждения обрабатывают терабайты информации о взаимодействиях своих пользователей, бизнесе, социальных сетях, а также собирают данные с датчиков таких устройств, как мобильные телефоны и автомобили. Задача современной эпохи состоит в том, чтобы разобраться в этом море данных. Именно здесь аналитика больших данных вступает в свои права.

Big Data Analytics в основном включает в себя сбор данных из разных источников, изменение их таким образом, чтобы они стали доступными для использования аналитиками, и, наконец, предоставление продуктов данных, полезных для бизнеса.

Цель освоения дисциплины направлена на формирование универсальных и профессиональных компетенций будущего магистра по направлению подготовки 39.04.01 Социология, а также комплексное изучение концепций, технологий, процессов и инструментов, связанных с большими данными, с целью формирования глубокого понимания их природы, значения, методов управления, обработки, анализа и визуализации.

Задачи освоения дисциплины:

- раскрыть исторические корни, причины появления термина, его значение и основные характеристики;
- изучить этапы, которые проходят данные и метаданные от создания до уничтожения, а также принципы их управления;
- исследовать содержание и задачи управления, различные подходы, эволюцию организационных структур и выявить ключевые проблемы;
- ознакомиться с основными технологиями хранения, обработки и управления большими данными, а также исследовать архитектуры систем обработки и хранилищ данных;
- разобрать концепции Business Intelligence (BI), OLAP, ETL, а также различные техники, методы и инструментарий для анализа больших данных, включая параллельные алгоритмы;
- ознакомиться с существующими программными решениями, платформами и оборудованием, используемым для работы с большими данными;
- освоить принципы, типы и инструменты визуализации, а также возможности использования специализированных программных средств;
- рассмотреть различные виды интеллектуального анализа данных (Data Mining, Text Mining, Web Mining, Social Media Mining) и основные статистические методы, применяемые в этой области.

Дисциплина «Введение в большие данные», относится к обязательной части блока «Дисциплины (модули)». Ее освоение происходит в 1 семестре.

Дисциплина «Введение в большие данные» является базой для освоения

следующих дисциплин и видов образовательной деятельности: Сетевые структуры и диффузионные процессы в цифровых сообществах; Большие данные в социальной архитектуре; Управление инновациями на основе больших данных; Рекламные стратегии на основе больших данных; Подготовка к сдаче государственного экзамена; Сдача государственного экзамена; Подготовка к процедуре защиты выпускной квалификационной работы; Защита выпускной квалификационной работы.

В результате усвоения тем и разделов дисциплины «Введение в большие данные», работы на практических занятиях у студентов должны сформироваться следующие универсальные и профессиональные компетенции:

Код, формулировка компетенции	Код, формулировка индикатора	Планируемые результаты обучения по дисциплине (модулю), характеризующие этапы формирования компетенций, индикаторов
УК-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий	ИД-2 УК-1. Определяет пробелы в информации, необходимой для решения проблемной ситуации, и проектирует процессы по их устранению	анализирует проблемную ситуацию, всесторонне очерчивает информационные потребности, критически оценивает наличие или отсутствие данных и проактивно проектирует комплексные процессы для восполнения выявленных пробелов, включая оценку затрат и сроков
	ИД-3 УК-1. Критически оценивает надежность источников информации, работает с противоречивой информацией из разных источников	владеет методами критической оценки надежности источников (например, проверка авторитетности, методологии сбора, потенциальной предвзятости), умеет систематизировать противоречивую информацию, аргументированно выбирать наиболее достоверные данные или разрабатывать стратегии для их примирения.
ПК-1. Способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов	ИД-2 ПК-1. Способен использовать методы сбора и анализа социологической информации	применяет широкий спектр методов сбора и анализа социологической информации (включая продвинутое техники обработки естественного языка, кластерный анализ, статистическое моделирование), адаптируя их для решения задач, связанных с большими данными, и интерпретируя полученные результаты в социологическом контексте.
	ИД-4 ПК-1. Выявляет закономерности протекания комплексных социальных и экономических процессов и механизмы функционирования основных социальных общностей	Студент на основе анализа больших данных способен выявлять сложные, многофакторные закономерности и механизмы, характеризующие протекание комплексных социальных и экономических процессов, строить адекватные модели объяснения и прогнозирования их развития.

Методические указания выступают в качестве информационного и практического источника и могут быть использованы для очного, заочного и дистанционного обучения.

Данное пособие по структуре и содержанию соответствует программе курса, темы практических занятий соотносятся с темами лекций, а содержательная часть практических занятий дополняет и углубляет лекционный материал.

ПЛАНЫ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

ТЕМА № 1. ВВЕДЕНИЕ В БОЛЬШИЕ ДАННЫЕ: ИНФОРМАЦИЯ, БОЛЬШИЕ ДАННЫЕ, BIG DATA

План темы:

1. История и причины появления термина Big Data
2. Сущность понятия Big Data
3. Значение Big Data
4. Информация и особенности ее хранения и обработки
5. Характеристики и источники Big Data
6. Четыре основных типа данных
7. Аналитика данных
8. Задачи, решаемые Big Data

Цель: понять фундаментальные концепции информации и больших данных, их отличия и роль в современном мире.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что в современном мире объемы информации растут экспоненциально, и понимание того, что такое BIG DATA и как они отличаются от традиционных данных, является отправной точкой для извлечения ценности из этого информационного потока.

Теоретическая часть.

Понятие Big Data возникло не случайно – оно отражает трансформацию способов восприятия, накопления и анализа информации в условиях экспоненциального роста цифровых потоков. Исторически термин начал употребляться ещё в конце XX века, но настоящую популярность он получил после публикаций Google и исследований в области распределённых систем. Ключевым триггером стало увеличение объёмов данных из-за развития интернета, социальных сетей, датчиков IoT, мобильной связи и электронной коммерции. Пиковый интерес вызвало осознание того, что традиционные методы обработки данных уже не справляются с новыми масштабами.

Под Big Data понимают массивы структурированной, полуструктурированной и неструктурированной информации, которые невозможно эффективно анализировать с помощью стандартных инструментов баз данных. Это не просто большой объём – это совокупность характеристик, определяющих сложность их обработки: скорость поступления, разнообразие форматов, достоверность и необходимость в реальном времени или близком к нему. Самое главное – Big Data – это не только объём, а вызов возможностям системы хранить, обрабатывать и интерпретировать информацию в новых условиях.

Значение Big Data трудно переоценить. Оно пронизывает все сферы жизни: медицину, финансы, телекоммуникации, транспорт, образование, государственное управление. Благодаря анализу больших данных компании могут предсказывать спрос, оптимизировать цепочки поставок, повышать уровень персонализации услуг, снижать риски мошенничества. Государственные учреждения используют Big Data для моделирования городской среды, контроля за эпидемиями, прогнозирования природных катаклизмов. В медицине – для диагностики заболеваний на ранних стадиях по данным геномики и мониторинга пациентов.

Особенностью хранения и обработки информации в Big Data является переход от централизованного подхода к распределённым технологиям. Традиционные СУБД сталкиваются с ограничениями при работе с экзакбайтными объёмами. Поэтому вместо них применяются платформы вроде Hadoop, Spark, Kafka, NoSQL-базы данных. Эти решения обеспечивают отказоустойчивость, масштабируемость и возможность параллельной обработки данных без потери производительности.

Характеристики Big Data определяются шестью ключевыми параметрами: объем (Volume), скорость (Velocity), разнообразие (Variety), достоверность (Veracity), ценность (Value) и время (Time). Объёмы достигают терабайтов и петабайтов; данные поступают в режиме реального времени; они представлены в виде текста, видео, аудио, логов, изображений, показателей с датчиков. При этом важную роль играет не только сама структура, но и качество исходных данных – их достоверность и однородность. Ценность же проявляется тогда, когда из данных можно извлечь скрытые закономерности, которые помогают принимать стратегические решения.

В качестве типов данных выделяют четыре группы: структурированные (табличные данные, JSON, CSV), полуструктурированные (XML, JSON-файлы), неструктурированные (тексты, фото, видео) и метаданные (служебная информация о данных). Именно сочетание этих типов и создаёт вызов современным аналитическим системам, требуя комплексного подхода к их обработке.

Аналитика данных – это сердце Big Data. Она позволяет не просто фиксировать события, а находить взаимосвязи, выявлять тренды, прогнозировать развитие событий. Различают описательную, диагностическую, прогнозирующую и директивную аналитику. Первые две позволяют понять, что произошло и почему; последние – предсказать будущее и определить действия для его изменения.

Основные задачи, решаемые с помощью Big Data: персонализация контента, андеррайтинг страховых полисов, оптимизация маршрутов в логистике, раннее выявление заболеваний, анализ клиентского поведения, автоматизация бизнес-процессов, усиление безопасности в IT-системах. Особенно актуально применение AI/ML-алгоритмов для анализа больших массивов данных, что делает результаты более точными и адаптивными.

Вопросы для самоконтроля:

1. Что означает термин «Big Data» в информационных технологиях? Что

- является основной целью обработки Big Data?
2. Кто и в каком году впервые ввел термин «Big Data»?
 3. Какие главные характеристики Big Data?
 4. Какие данные занимают больше мировой памяти относительно остальных?
 5. Какие понятия содержит в себе принцип трех «V»?
 6. С какого года Большие данные изучаются как академический предмет в вузовских программах по науке о данных?
 7. Что является примером квази-структурированных данных?
 8. Как назывался первый суперкомпьютер, оснащенный вопросно-ответной системой искусственного интеллекта?
 9. Чем характеризуются «Большие данные»?
 10. Какие исторические события или технологические прорывы способствовали появлению термина Big Data?
 11. Почему термин «Big Data» стал актуальным именно в XXI веке?
 12. Что такое Big Data с точки зрения объёма, скорости, разнообразия и достоверности (4V)?
 13. Как отличаются обычные данные от Big Data по масштабам и сложности обработки?
 14. Назовите ключевые характеристики Big Data и объясните значение каждого из них.
 15. Какие основные типы источников порождают Big Data?
 16. Перечислите четыре основных типа данных и приведите примеры для каждого.
 17. В чём заключается различие между структурированными, полуструктурированными и неструктурированными данными?
 18. Какова роль аналитики данных в работе с Big Data?
 19. Какие бизнес-задачи могут быть решены с помощью аналитики больших данных?
 20. Почему Big Data важны для современных организаций? Приведите конкретные примеры применения.
 21. Каково соотношение информации, данных и знаний в контексте Big Data?

Практическое задание:

Проанализируйте три примера из реальной жизни (например, социальные сети, телекоммуникационные компании, ритейл), где накопление информации превращается в вызов для обработки. Укажите, какие объёмы данных, их виды (структурированные, полуструктурированные, неструктурированные) и скорость поступления характерны для каждого случая. Объясните, почему именно эти данные можно отнести к категории Big Data, опираясь на мультиаспектность (объём, разнообразие, скорость, достоверность, ценность).

Тесты:

1. Что из перечисленного является ключевым отличием больших данных от обычных данных?
 - А) Большие данные всегда структурированы
 - В) Большие данные требуют новых подходов к обработке и анализу

- С) Большие данные имеют фиксированную структуру
- Д) Большие данные не могут быть хранены в облачных системах

2. Какой из параметров НЕ относится к характеристикам Big Data («четыре V»)?

- А) Volume – объём
- В) Velocity – скорость
- С) Variability – изменчивость
- Д) Veracity – достоверность

3. Чем информация отличается от данных?

- А) Информацией называют любые цифры без контекста
- В) Информация – это данные, осмысленно интерпретированные в определённом контексте
- С) Информация всегда имеет высокую точность
- Д) Информация может существовать независимо от данных

4. Какова главная цель использования Big Data в бизнесе?

- А) Увеличить количество сотрудников
- В) Повысить эффективность принятия решений за счёт анализа скрытых паттернов
- С) Заменить все бумажные документы на электронные
- Д) Снизить затраты на аренду офисов

5. Что подразумевает термин «data-driven decision making»?

- А) Решения принимаются исключительно на основе интуиции
- В) Решения основаны на фактах, полученных из анализа данных
- С) Все решения формируются автоматически без участия человека
- Д) Использование только внешних источников информации

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 2. ОСНОВЫ УПРАВЛЕНИЯ БОЛЬШИМИ ДАННЫМИ

План темы:

1. Содержание и задачи процесса управления большими данными
2. Подходы к управлению Big Data
3. Эволюция организационных основ работы с Big Data
4. Проблемы использования Big Data

Цель: освоить базовые принципы и подходы к эффективному управлению большими объемами данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что без четких основ управления, огромные массивы данных становятся неконтролируемыми, что затрудняет их использование, снижает их ценность и порождает риски.

Теоретическая часть.

Управление большими данными – это целый комплекс мероприятий, направленных на обеспечение доступности, качества, надёжности и безопасности данных на всех этапах жизненного цикла. Его цель – создать условия, при которых Big Data может эффективно использоваться для достижения бизнес-целей, соблюдения нормативных требований и минимизации рисков. Успешное управление подразумевает не просто сбор информации, а её структурирование, классификацию, защиту, документирование и интеграцию в единую аналитическую экосистему.

На практике существуют два основных подхода к управлению: традиционный, ориентированный на хранилища данных (data warehouse), и современный – на основе платформ, поддерживающих работу с неструктурированными и быстро меняющимися массивами данных (Data Lake, Lambda/Accelerated Architecture). Второй вариант даёт больше свободы, особенно при необходимости интеграции различных источников данных без жесткой предварительной подготовки.

Эволюция подходов к работе с Big Data прошла несколько этапов. На начальной стадии доминировали централизованные системы хранения. Появление облачных сервисов и open-source решений (Hadoop, Spark) позволило перейти к гибким, масштабируемым моделям. Сегодня наблюдается переход от простого хранения к интеллектуальному управлению – внедряются системы ETL (Extract, Transform, Load), MDM (Master Data Management), а также технологии автоматического классифицирования и тэгирования данных. Особое внимание уделяется вопросам конфиденциальности: GDPR, ФЗ-152, HIPAA требуют чёткого контроля за обработкой личной информации.

Проблемы использования Big Data носят как технический, так и

организационный характер. Во-первых, высокий порог входа – требуется специализированная инфраструктура, опытные команды аналитиков и инженеры. Во-вторых, риск получения некорректных выводов из-за плохого качества данных (dirty data), несоответствия метаданных, ошибочной интерпретации. В-третьих, затруднена согласованность между отделами – каждый использует свои алгоритмы и подходы. Наконец, растёт давление на сторонние сервисы, связанные с безопасностью и этической составляющей: использование данных без согласия пользователей, дискриминация при принятии решений, персонализация на основе чувствительной информации. Все эти факторы требуют создания единой политики управления данными, включая регламентацию, обучение персонала и внутренний аудит.

Вопросы для самоконтроля:

1. Каковы цели и содержание процесса управления большими данными (Data Management)?
2. Какие основные подходы используются при управлении Big Data? Сравните традиционный подход и методы, адаптированные под Big Data.
3. Как эволюционировала архитектура управления данными с развитием цифровых технологий?
4. Какие проблемы возникают при сборе, хранении и анализе больших объёмов данных?
5. Почему стандартные системы баз данных не всегда эффективны при работе с Big Data?
6. Какие вызовы связаны с обеспечением качества, безопасности и конфиденциальности данных в Big Data?
7. Объясните, как интеграция данных из разных источников влияет на качество аналитики.
8. Какие технологии (например, Hadoop, Spark) помогают решать проблемы управления Big Data?
9. Почему важно иметь стратегию управления данными в организации?
10. Какие риски неэффективного управления большими данными могут повлиять на бизнес?

Практическое задание:

Возьмите образец данных из открытых источников (например, датасет с публичными данными о погоде или автопрокате). Постройте простую схему управления ими: укажите этапы сбора, очистки, агрегации, хранения и доступа. Опишите, какие проблемы могут возникнуть при управлении этими данными без использования системного подхода, и как использование специализированной среды (например, ETL-инструментов) может помочь решить эти проблемы.

Тесты:

1. Что входит в состав процесса управления большими данными?

- А) Только сбор данных из социальных сетей
- В) Организация доступа, контроль качества, обеспечение безопасности и целостности данных
- С) Создание графических интерфейсов для пользователей

D) Разработка мобильных приложений

2. Какой элемент управления данными обеспечивает соответствие данным нормативным требованиям?

- A) Метаданные
- B) Государственная регистрация
- C) Комплаенс-контроль
- D) Автоматизация загрузки

3. Что такое data governance?

- A) Процесс удаления устаревших файлов
- B) Набор политик, процедур и ролей, регулирующих управление информационными активами компании
- C) Методология разработки программного обеспечения
- D) Поддержка пользовательских запросов через чат-бот

4. Какой из факторов наиболее влияет на качество больших данных?

- A) Выбор языка программирования
- B) Наличие метаданных и стандартов описания
- C) Размер серверной комнаты
- D) Частота обновления веб-страниц

5. Для чего используется система управления жизненным циклом данных?

- A) Для повышения скорости интернет-соединения
- B) Для контроля доступа к базам данных
- C) Для организации этапов создания, хранения, использования и уничтожения данных
- D) Для шифрования паролей пользователей

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 3. ЖИЗНЕННЫЙ ЦИКЛ ДАННЫХ

План темы:

1. Создание данных (Data Generation/Data Capture)
2. Обслуживание данных (Data Maintenance)
3. Синтез данных (Data Synthesis)
4. Использование данных (Data Usage)
5. Публикация данных (Data Publication)
6. Архивация данных (Data Archival)
7. Уничтожение данных (Data Purging)

Цель: изучить все этапы существования данных, от их создания до утилизации, и понять их значение.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что понимание полного жизненного цикла данных позволяет оптимизировать процессы работы с ними, обеспечивать их качество, безопасность и соответствие нормативным требованиям на каждом этапе.

Теоретическая часть.

Процесс формирования данных начинается с их первоначального появления – либо через прямое измерение, регистрацию или ввод пользователями, либо за счёт автоматической сборки информации с датчиков, систем мониторинга, транзакционных баз или пользовательских интерфейсов. Ключевую роль здесь играют источники данных: IoT-устройства, формы веб-сайтов, сканеры штрихкодов, системы учета времени и другие инструменты. От качества и достоверности этих исходных данных зависит весь последующий анализ. Важно обеспечить корректность фиксации, контроль целостности и соответствие стандартам хранения.

После попадания в информационную систему данные требуют постоянного поддержания. Это включает в себя очистку от дубликатов, исправление ошибок, устранение несоответствий, актуализацию значений и обеспечение согласованности между различными источниками. Обслуживание охватывает как технические аспекты (апдейты, индексация), так и бизнес-логические правила (например, проверка возраста клиента при вводе). Эффективное обслуживание позволяет сохранять стабильность и надежность данных на всех этапах их существования.

На этом этапе происходит объединение, обобщение и преобразование разрозненных данных в единые структурированные потоки. Синтез может осуществляться путём агрегации, кросс-регистрации, консолидации нескольких источников, применения алгоритмов машинного обучения для заполнения пробелов или создания производных показателей. Особенностью

является не просто соединение, но и интерпретация – создание новых метрик, прогнозов, сценариев развития событий. Именно здесь данные перестают быть простыми фактами и становятся основой для принятия решений.

Данные активно применяются в процессах управления, оптимизации, моделирования и принятия решений. Они используются как в оперативной деятельности (например, расчёты стоимости заказа), так и в стратегических проектах (анализ рынка, прогнозирование спроса). Критически важно соблюдать ограничения доступа, контролировать права пользователей и обеспечивать соответствие политике использования. Эффективность использования определяется не только количеством запросов, но и качеством получаемых результатов – точность, своевременность, релевантность.

Этот этап предполагает официальное распространение данных внешним или внутренним пользователям. Он включает подготовку публичных отчетов, выгрузку в открытые реестры, размещение на платформах, доступных по API. При этом важно обеспечить прозрачность происхождения, указать версию данных, срок действия, методологию расчета и условия использования. Публикация должна соответствовать нормам законодательства, особенно если затрагиваются персональные или чувствительные сведения.

По достижении установленного срока хранения или после завершения их практической полезности данные переводятся в архив. Цель – сохранить информацию для будущего анализа, юридических нужд или исторических исследований без нагрузки на рабочие системы. Архивные данные обычно хранятся в менее доступном виде, с более низкой скоростью обращения, но с высокой степенью защиты и резервного копирования. Процедура архивации включает маркировку, компрессию, шифрование и организацию системы доступа.

Завершающий этап – удаление данных, которые больше не имеют значения для организации. Оно выполняется по установленному графику, по требованию закона (например, GDPR) или по инициативе владельца. Удаление должно быть полным: данные удаляются с всех уровней – серверов, резервных копий, облачных хранилищ. Особенно важна процедура, связанная с персональными данными, где требуется подтверждение полного стирания. Нарушение правил уничтожения влечёт юридическую ответственность и риск утечки информации.

Вопросы для самоконтроля:

1. Что входит в этап создания данных? Приведите примеры источников и форматов данных.
2. Какие задачи выполняет обслуживание данных? Почему оно критически важно для долгосрочного использования?
3. Что такое синтез данных? Как он отличается от простой обработки или объединения?
4. Какие факторы нужно учитывать при использовании данных в принятии решений?
5. В чём разница между публикацией данных и их распространением? Когда целесообразно делать данные доступными широкому кругу пользователей?

6. Для чего нужна архивация данных? Как определить сроки хранения и выбор формата архива?
7. Каковы правовые и этические основания для удаления (удаления/уничтожения) данных?
8. Как жизненный цикл данных влияет на безопасность, соответствие нормативным требованиям (например, GDPR)?
9. Может ли один и тот же набор данных проходить несколько этапов одновременно? Приведите пример.
10. Почему нельзя игнорировать этап уничтожения данных? Какие последствия могут быть при его пропуске?

Практическое задание:

На основе выбранного набора данных (например, история покупок клиентов в онлайн-магазине за год) воспроизведите его жизненный цикл: от формирования первоначальных записей до удаления. Для каждого этапа (создание, изменение, хранение, анализ, архивация, стирание) укажите возможные риски, связанные с безопасностью, конфиденциальностью и юридической ответственностью. Какие механизмы контроля версий и журналов событий помогут обеспечить целостность данных?

Тесты:

1. На каком этапе жизненного цикла данных происходит первичное формирование или сбор данных?
 - A) Анализ
 - B) Хранение
 - C) Сбор/генерация
 - D) Удаление
2. Что происходит на этапе «обработка» в жизненном цикле данных?
 - A) Передача данных по сети
 - B) Преобразование сырых данных в структурированный вид
 - C) Формирование отчётов для руководства
 - D) Распределение данных среди пользователей
3. Какой из этапов жизненного цикла данных связан с созданием резервных копий?
 - A) Удаление
 - B) Хранение
 - C) Анализ
 - D) Представление
4. Почему важно планировать этап утилизации данных?
 - A) Чтобы увеличить стоимость хранения
 - B) Чтобы минимизировать риски утечки информации и соответствовать законодательству
 - C) Чтобы повысить производительность CPU
 - D) Чтобы ускорить передачу данных между серверами

5. Какой этап жизненного цикла данных следует сразу после «объединения»?

- A) Представление
- B) Обработка
- C) Сбор
- D) Анализ

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 4. ЖИЗНЕННЫЙ ЦИКЛ АНАЛИТИКИ ДАННЫХ

План темы:

1. Business Intelligence (BI)
2. ETL (Extract, Transform, Load)-процесс
3. Средства BI
4. Online Analytical Processing (OLAP)
5. Инструменты анализа BI
6. Понятие жизненного цикла аналитики данных

Цель: осмыслить последовательность этапов, необходимых для получения аналитических выводов из данных, начиная от постановки задачи и заканчивая представлением результатов.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что эффективная аналитика требует структурированного подхода, и понимание её жизненного цикла гарантирует, что процесс принесет реальную пользу и будет последовательным.

Теоретическая часть.

Business Intelligence – это комплекс подходов, технологий и практик, направленных на преобразование сырой информации в понятные, наглядные и полезные бизнес-показатели. Его цель – помочь руководству принимать решения на основе фактических данных, а не догадок. BI включает в себя сбор, обработку, визуализацию и анализ данных с целью получения ключевых инсайтов. Наиболее востребованные элементы – отчеты, панели мониторинга, KPI, трендовые диаграммы. Современные BI-системы способны работать в режиме реального времени, что повышает оперативность реакций на изменения.

ETL – ключевой механизм подготовки данных перед анализом. Сначала из различных источников извлекаются (Extract) данные – это могут быть базы, файлы CSV, API. Затем они трансформируются (Transform): проводится очистка, стандартизация, агрегация, применение бизнес-правил, конвертация типов. После этого информация загружается (Load) в целевую систему – хранилище данных, аналитическую базу или OLAP-куб. От качества выполненной трансформации зависит итоговая ценность аналитического продукта. Механизм ETL может быть автоматизирован и реализован в рамках пайплайнов обработки данных.

Под средствами BI понимаются программные продукты, предназначенные для работы с большими объемами данных и построения аналитических отчетов. К ним относятся такие решения, как Tableau, Power BI, Qlik Sense, Looker, SAP BO. Эти платформы позволяют подключаться к разным источникам, строить визуальные представления, настраивать

интерактивные дашборды, задавать сложные фильтры и параметры сравнения. Главное преимущество – возможность быстро адаптировать интерфейс под потребности разных групп пользователей: от финансистов до маркетологов.

OLAP – технология многомерного анализа данных, ориентированная на быстрый поиск, свёртки и агрегацию больших массивов информации. В отличие от OLTP (оперативной обработки), где акцент на транзакциях, OLAP рассчитан на глубокий анализ. Благодаря использованию многомерных кубов (data cubes) можно легко переходить от общих показателей к детализации по регионам, периодам, категориям. OLAP-системы обеспечивают эффективную работу даже с огромными объемами данных, позволяя выполнять сложные запросы в считанные секунды.

В состав аналитических инструментов входят как программные модули (например, DAX, SQL), так и визуальные панели управления. К ним относятся: дашборды, тепловые карты, гистограммы, диаграммы Gantt, рейтинги, матрицы. Также важны функции прогнозирования, A/B-тестирования, детект-систем, выявления выбросов. Все эти средства помогают находить скрытые зависимости, выявлять аномалии, сравнивать временные ряды и строить прогнозы. Качественная аналитика

Вопросы для самоконтроля:

1. Что такое Business Intelligence (BI)? Каковы её основные цели и компоненты?
2. Объясните каждую фазу ETL-процесса: что происходит при извлечении, преобразовании и загрузке данных?
3. Какие технологии или платформы чаще всего используются для реализации ETL?
4. В чём разница между OLTP и OLAP? Кто является основным пользователем OLAP-систем?
5. Какие функции выполняют средства BI (например, Power BI, Tableau, Qlik)?
6. Что позволяет сделать многомерный анализ в OLAP? Приведите пример использования.
7. Какой порядок действий обычно используется в жизненном цикле аналитики данных?
8. Почему качественные данные перед началом анализа имеют решающее значение?
9. Как можно оценить эффективность аналитической модели после её внедрения?
10. Какие метрики используются для контроля качества аналитических процессов?

Практическое задание:

Создайте описание таблицы базы данных (например, «Заказы»), используя два типа метаданных: структурное (имена столбцов, типы данных, ограничения) и семантическое (значение полей, единицы измерения, правила заполнения). Напишите пример описания одного поля («Дата_заказа»): что оно означает, кто его вводит, какова частота изменения, влияет ли на другие

элементы системы. Как метаданные помогают в автоматизации аналитических процессов?

Тесты:

1. Что является начальным шагом в жизненном цикле аналитики данных?

- A) Интерпретация результатов
- B) Формулировка вопроса или проблемы
- C) Выбор инструмента визуализации
- D) Сбор данных вручную

2. На каком этапе проводится очистка и подготовка данных для анализа?

- A) Интерпретация
- B) Моделирование
- C) Подготовка данных
- D) Отчётность

3. Какой этап жизненного цикла аналитики данных предполагает выбор модели или метода анализа?

- A) Реализация
- B) Оценка
- C) Моделирование
- D) Визуализация

4. Что делает этап «оценка» в жизненном цикле аналитики?

- A) Обеспечивает обратную связь от пользователя
- B) Проверяет корректность и значимость полученных результатов
- C) Проводит обучение команды
- D) Удаляет временные таблицы

5. Как завершается жизненный цикл аналитики данных?

- A) Публикацией статьи в научном журнале
- B) Переходом к новому проекту
- C) Внедрением рекомендаций в бизнес-процессы
- D) Окончательным утверждением бюджета

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 5. МЕТАДААННЫЕ

План темы:

1. Понятие метаданных
2. Жизненный цикл метаданных
3. Оценка требований и анализ контента
4. Спецификация системных требований
5. Система метаданных
6. Сервис и оценка

Цель: понять природу, назначение и значение метаданных для организации, поиска и понимания данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что метаданные служат «данными о данных», делая огромные объемы информации понятными, доступными и управляемыми, что критически важно для работы с BIG DATA.

Теоретическая часть.

Понятие метаданных раскрывает суть информации об информации – это структурированное описание объектов, процессов или данных, позволяющее управлять ими эффективно, определяя их характеристики, происхождение, назначение, формат, владельца, условия использования и другие ключевые параметры. В отличие от самих данных, метаданные несут в себе информацию о том, как, где, когда и зачем были созданы эти данные. Они играют роль «картографических указателей» в информационных системах, обеспечивая понимание, поиск, интеграцию и управление данными.

Жизненный цикл метаданных охватывает все этапы существования данных – от планирования и сборки до окончательного удаления. На начальном этапе формулируются требования к метаданным, проводится анализ содержимого и определение ключевых элементов. Затем создаётся спецификация, в которой фиксируются правила построения, типы данных, связи между ними и ограничения. В процессе эксплуатации метаданные актуализируются, контролируются на соответствие изменениям в данных, а при необходимости – корректируются или заменяются. Финальная стадия – архивация или удаление, при котором сохраняется информация о предшествующем состоянии метаданных для целей аудита и восстановления.

Оценка требований и анализ контента являются ключевыми шагами перед внедрением системы метаданных. Здесь выявляются пользовательские нужды, бизнес-процессы, источники данных, уровни доступа и степень детализации. Анализ позволяет определить, какие именно атрибуты должны присутствовать в метаданных, какова их семантика, а также установить приоритетные категории – например, описательные, административные,

технические. Это позволяет избежать перегрузки и обеспечить релевантность метаданных реальным потребностям.

Спецификация системных требований к метаданным представляет собой документ, содержащий полную картину поведения системы: список обязательных и условных полей, типы данных, допустимые значения, связи между объектами, правила инициализации и проверки. Такая форма задаёт жёсткую структуру для автоматизации, стандартизирует процессы, делает возможным интеграционное взаимодействие с другими системами и поддерживает согласованность во времени и пространстве.

Система метаданных – это совокупность правил, процедур, инструментов и баз, обеспечивающих создание, хранение, обновление, использование и контроль над описательной информацией. Она может быть централизованной, децентрализованной или гибридной, в зависимости от масштабов организации и требований к безопасности. Эффективная система метаданных способствует прозрачности, снижению рисков ошибок при работе с данными, ускоряет разработку и обслуживание ИТ-систем.

Сервис и оценка метаданных – важнейший компонент цифровой зрелости компании. Работа сервиса строится вокруг мониторинга состояния метаданных, их качества, актуальности и соответствия установленным стандартам. Оценка включает проверку полноты, достоверности, унифицированности и доступности. Используются как ручные методы, так и автоматические скринеры, которые выявляют дубликаты, пробелы, несоответствия. Результаты оценки служат основой для принятия решений о модернизации, переосмыслении подходов к управлению данными и формированию внутренних регламентов.

Вопросы для самоконтроля:

1. Что такое метаданные? Приведите примеры метаданных на уровне файла, базы данных и приложения.
2. Какие виды метаданных выделяют (технические, бизнес-метаданные, операционные)? В чём их различия?
3. Опишите основные этапы жизненного цикла метаданных. Почему важно контролировать метаданные на всех этапах?
4. Какие задачи решает оценка требований и анализ контента в рамках управления метаданными?
5. Что включает спецификация системных требований с точки зрения метаданных? Приведите пример.
6. Как работает система метаданных? Как она интегрируется с другими системами управления данными?
7. Как сервисы метаданных поддерживают поиск, согласованность и качество данных в организации?
8. Почему метаданные необходимы для соответствия нормативным требованиям (GDPR, HIPAA и др.)?
9. Как метаданные способствуют документированию и воспроизводимости аналитических процессов?
10. Какие риски возникают при отсутствии централизованной системы

метаданных?

Практическое задание:

Создайте описание таблицы базы данных (например, «Заказы»), используя два типа метаданных: структурное (имена столбцов, типы данных, ограничения) и семантическое (значение полей, единицы измерения, правила заполнения). Напишите пример описания одного поля («Дата_заказа»): что оно означает, кто его вводит, какова частота изменения, влияет ли на другие элементы системы. Как метаданные помогают в автоматизации аналитических процессов?

Тесты:

1. Что такое метаданные?

- A) Данные о пользователях системы
- B) Данные, содержащие информацию о структуре, содержании и способах использования других данных
- C) Резервные копии баз данных
- D) Лог-файлы приложений

2. Какой тип метаданных описывает логическую структуру данных?

- A) Операционные метаданные
- B) Техническое метаданные
- C) Бизнес-метаданные
- D) Интеграционные метаданные

3. Какова роль метаданных в ETL-процессах?

- A) Ускоряют передачу данных между системами
- B) Определяют, какие данные нужно экспортировать из источника
- C) Содержат описание схемы, типа и происхождения данных
- D) Заменяют сами данные в хранилище

4. Какой элемент метаданных может содержать информацию о владельце данных?

- A) Источник данных
- B) Схема таблицы
- C) Описание бизнес-правил
- D) Политики доступа

5. К какому типу метаданных относятся сведения о времени последнего изменения записи?

- A) Технические метаданные
- B) Операционные метаданные
- C) Стратегические метаданные
- D) Функциональные метаданные

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3
Дополнительная: 1-3
Интернет-ресурсы: 1-10

ТЕМА № 6. ТЕХНОЛОГИИ РАБОТЫ С БОЛЬШИМИ ДАННЫМИ

План темы:

1. Становление технологии работы с большими данными
2. Современные технологии управления большими данными

Цель: ознакомиться с основными технологиями, инструментами и методами, применяемыми для эффективной обработки и анализа больших данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что современная работа с данными невозможна без специализированных, масштабируемых технологий, способных справляться с объемом, скоростью и разнообразием BIG DATA.

Теоретическая часть.

Появление технологий работы с большими данными стало ответом на вызовы цифровизации: рост объёмов информации, скорость её поступления, многообразие форматов и необходимость быстрой аналитической обработки. Изначально традиционные СУБД оказались неспособны справиться с этими нагрузками, что породило новое поколение решений, ориентированных на горизонтальное масштабирование, отказоустойчивость и высокопроизводительную обработку. Ключевой вехой стал переход от централизованных систем к распределённым платформам.

Современные технологии управления большими данными основаны на архитектуре, построенной на принципах разделения обязанностей: хранение, обработка, анализ и визуализация данных выполняются разными компонентами. Лидирующую позицию занимает экосистема Hadoop – с его распределённой файловой системой HDFS и фреймворком MapReduce для параллельной обработки. Однако сегодня активно используются более современные решения: Apache Spark, который предлагает более быстрый выполнение операций за счёт работы в оперативной памяти; Kafka – для потоковой передачи данных в режиме реального времени; Flink – для сложных событийных вычислений; а также NoSQL-решения типа Cassandra, MongoDB и Amazon DynamoDB, рассчитанные на работу с неструктурированными и полуструктурированными данными.

На фоне этих технологий возникает тренд на гибридные модели: объединение batch-обработки и stream-вычислений, применение машинного обучения прямо внутри среды больших данных через библиотеки MLlib, TensorFlow on Spark. Также широкое распространение получили облачные платформы (AWS S3, Google Cloud Storage, Azure Data Lake), которые предоставляют готовые инструменты для хранения, обработки и анализа, снижая порог входа и упрощая масштабирование. Управление большими

данными больше не сводится к выбору конкретного инструмента – это комплексный подход, сочетающий архитектуру, безопасность, политики управления данными и культуру данных.

Вопросы для самоконтроля:

1. Как развивались технологии работы с большими данными с начала 2000-х годов? Какие события стали поворотными моментами?
2. Какие ключевые технологии обеспечивают хранение, обработку и анализ Big Data? Приведите примеры (Hadoop, Spark, Kafka и др.).
3. В чём преимущество распределённых систем (например, HDFS) по сравнению с традиционными СУБД?
4. Почему Apache Spark стал популярнее MapReduce в некоторых сценариях?
5. Какие решения позволяют работать с потоковыми данными в реальном времени? Приведите примеры.
6. Какие архитектурные принципы лежат в основе современных платформ для Big Data?
7. Как технологии Big Data влияют на скорость и масштабируемость аналитики?
8. Почему облачные платформы (AWS, Azure, GCP) сыграли ключевую роль в развитии технологий Big Data?
9. Какие ограничения могут быть у высокопроизводительных решений Big Data?
10. Как выбрать подходящую технологическую платформу для проекта с большими данными?

Практическое задание:

Сравните возможности трёх технологий: Apache Spark, Apache Flink и Apache Storm. По каждому критерию (тип обработки – потоковая/пакетная, производительность, поддержка ML, сложность реализации, совместимость с хранилищами) составьте таблицу сравнения. Выберите наиболее подходящую технологию для задачи: анализ потока событий в реальном времени (например, мошенничество в банковских операциях). Обоснуйте выбор.

Тесты:

1. **Какая технология наиболее часто используется для обработки потоковых данных в реальном времени?**
 - A) Hadoop MapReduce
 - B) Apache Spark SQL
 - C) Apache Kafka
 - D) PostgreSQL JSONB
2. **Что из перечисленного является ключевым компонентом архитектуры Big Data?**
 - A) Центральный процессор высокой частоты
 - B) Распределённая файловая система (например, HDFS)
 - C) Однородная сеть подключений
 - D) Графический интерфейс пользователя

3. Какой подход позволяет обрабатывать большие объёмы данных параллельно?

- A) Монолитное программирование
- B) Нелинейная обработка
- C) Распределённые вычисления
- D) Последовательная обработка

4. Для чего применяется Apache Flink?

- A) Для хранения временных рядов
- B) Для стриминговой обработки данных с низкой задержкой
- C) Для создания веб-интерфейсов
- D) Для выполнения SQL-запросов к NoSQL-базам

5. Какой из перечисленных форматов данных эффективен для работы с большими объемами неструктурированных данных?

- A) CSV
- B) XML
- C) Avro
- D) PDF

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 7. ХРАНИЛИЩА ДАННЫХ: ЭВОЛЮЦИЯ И ОБЩИЕ ОСНОВЫ

План темы:

1. Большие данные и хранилища данных
2. Подходы к архитектуре и принципам проектирования хранилищ данных
3. Системы хранения данных

Цель: изучить развитие концепций хранилищ данных и понять базовые принципы их построения.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что надежное и эффективное хранение является краеугольным камнем работы с любыми объемами данных, и понимание его эволюции помогает выбрать оптимальные решения.

Теоретическая часть.

Развитие больших данных привело к тому, что классические реляционные базы данных стали недостаточны для полноценной поддержки аналитических запросов. Возникла потребность в специализированных системах, способных хранить огромные массивы исторических данных, быстро выполнять многомерный анализ и интегрироваться с различными источниками. Именно тогда появились хранилища данных (Data Warehouses) – централизованные хранилища, спроектированные специально для обработки сложных аналитических запросов.

Хранилище данных – это не просто место для хранения, а сложная система, объединяющая данные из различных источников: ERP, CRM, IoT-устройств, социальных сетей, внешних API. Главная цель – обеспечение единого представления данных для аналитики, агрегирования, сравнений и прогнозирования. Отличается от OLTP-систем (операционных) тем, что ориентирована на OLAP (Online Analytical Processing): поддерживает сложные запросы, длительные транзакции, работу с большим количеством строк и агрегацию по нескольким уровням.

Подходы к проектированию хранилищ данных опираются на три основные парадигмы: модель «звезда», модель «снежинка» и модель «гирлянда». В первой каждый факт хранится в центральной таблице фактов, связанной с размерами (dimension tables). Такая архитектура обеспечивает читаемость, производительность и гибкость при построении отчетов. Вторая – более нормализованная версия, снижающая избыточность, но усложняющая запросы. Третья – комбинированная структура, используемая в крупных проектах для минимизации дублирования

Вопросы для самоконтроля:

1. В чём разница между хранилищем данных (data warehouse) и операционной базой данных (OLTP)?
2. Каковы основные цели создания хранилища данных?
3. Объясните эволюцию хранилищ данных: от ручных отчётов до современных облачных data warehouses.
4. Какие подходы к проектированию хранилищ данных существуют? (Например, Kimball vs Inmon). Сравните преимущества и недостатки.
5. Какие принципы лежат в основе архитектуры хранилища данных?
6. Что такое SCD (Slowly Changing Dimensions)? Приведите примеры типов изменений.
7. Какие типы систем хранения данных вы знаете? (Relational, NoSQL, columnar storage, data lake, data mart и др.)
8. В чём разница между data warehouse и data lake? Когда какой вариант предпочтителен?
9. Как обеспечивается производительность при запросах к большим объёмам данных в хранилище?
10. Как современные хранилища данных поддерживают аналитику, машинное обучение и BI-решения?

Практическое задание:

Опишите историческую эволюцию хранилищ данных: от файловой системы до современных data warehouse и data lakes. Приведите примеры ключевых отличий между RDBMS и NoSQL-системами. Какие требования предъявляются к хранилищу данных, чтобы оно могло работать с большими объёмами и разнообразием данных? Какие преимущества даёт переход от «жесткой» структуры к более гибкой?

Тесты:

1. Чем отличается хранилище данных от операционной базы данных?

- А) Хранилище данных всегда работает в режиме онлайн-обработки транзакций
- В) Хранилище данных предназначено для аналитических запросов, а не для транзакций
- С) Хранилище данных использует только NoSQL-технологии
- Д) Хранилище данных хранит данные в реальном времени

2. Какой принцип лежит в основе архитектуры хранилища данных?

- А) Изменяемость исходных данных
- В) Непостоянство структуры
- С) Интеграция и согласованность данных из разных источников
- Д) Отсутствие исторического хранения

3. Что означает термин «консолидация данных» в контексте хранилищ?

- А) Перенос данных на внешние носители
- В) Объединение данных из различных источников в единую структуру
- С) Удаление устаревших записей
- Д) Автоматическое шифрование информации

4. Какой тип модели данных чаще всего используется в хранилищах данных?

- A) Объектно-ориентированная модель
- B) Реляционная модель с денормализацией
- C) Иерархическая модель
- D) Графовая модель

5. Что такое «слой загрузки» в ETL-процессе?

- A) Этап удаления дубликатов
- B) Процесс фильтрации данных до загрузки
- C) Шаг, где данные преобразуются и загружаются во временную зону
- D) Конечная точка хранения данных

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 8. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ХРАНЕНИЯ ДАННЫХ: МАСШТАБИРОВАНИЕ И МНОГОУРОВНЕВОЕ ХРАНЕНИЕ ДАННЫХ

План темы:

1. Архитектура корпоративной системы хранилища – DWH
2. Хранилище данных и озеро данных
3. Масштабируемость
4. Репликация
5. CAP теорема
6. MongoDB

Цель: разобраться в современных подходах к масштабируемому и многоуровневому хранению данных, обеспечивающему гибкость и экономичность.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что рост объемов данных требует решений, способных масштабироваться и оптимизировать затраты за счет размещения данных на различных уровнях хранения в зависимости от их востребованности.

Теоретическая часть.

Архитектура корпоративной системы хранилища данных (DWH) основана на принципе централизованного сбора, очистки и структурированного хранения информации из различных источников для последующего аналитического использования. Она строится по иерархической модели, где данные проходят этапы загрузки, трансформации и агрегирования. Ключевыми элементами являются слои: сырое хранилище, уровень очистки, консолидированная зона, а также уровни бизнес-показателей и отчетности. Такая архитектура обеспечивает согласованность данных, их долгосрочную доступность и возможность построения сложных отчетов.

Отличие хранилища данных от озера данных заключается в подходе к организации информации. В DWH данные предварительно структурируются, нормализуются и ориентируются под конкретные запросы, что делает их пригодными для аналитики. Озеро же данных сохраняет сырые, необработанные данные в исходном виде – включая JSON, CSV, видео, логи – позволяя хранить любые типы информации без жесткой привязки к схеме. Это даёт гибкость и позволяет использовать данные для обучения моделей ИИ, но требует дополнительных усилий на этапе подготовки.

Масштабирование в системах хранения данных реализуется как вертикально (увеличение мощности одного узла), так и горизонтально (распределённое расширение через добавление новых серверов).

Горизонтальное масштабирование становится стандартом для крупных платформ, особенно в условиях динамичного роста объёмов данных. Оно достигается за счёт шардирования, балансировки нагрузки и использования отказоустойчивых архитектур, обеспечивающих высокую производительность даже при пиковой нагрузке.

Репликация – это процесс создания копий данных на нескольких узлах или серверах. Она повышает надежность, снижает задержку доступа за счёт локальных кэшей и обеспечивает устойчивость к отказам. Различают репликацию синхронную (мгновенная доставка изменений) и асинхронную (обслуживание с задержкой). На практике чаще применяется асинхронная, поскольку она лучше совместима с сетевой задержкой и масштабируемостью.

CAP-теорема утверждает, что в распределённой системе невозможно одновременно достичь трёх свойств: согласованности (Consistency), доступности (Availability) и разделения сети (Partition tolerance). При этом всегда можно выбрать два из трёх. Например, системы типа Cassandra делают акцент на доступности и устойчивости к сбоям сети, жертвуя полной согласованностью, тогда как Oracle RAC стремятся к максимальной согласованности, теряя частично доступность при сбоях сети.

MongoDB – популярная документо-ориентированная СУБД, предназначенная для работы с большими объемами неструктурированных данных. Её преимущество – гибкая схема, быстрая запись, поддержка репликаций и шардинга. MongoDB использует модель «один документ – одна запись», что позволяет легко адаптироваться под меняющиеся требования бизнеса. Подходит для хранения данных пользовательских сессий, событий, IoT-поток и других динамических потоков информации.

Вопросы для самоконтроля:

1. Какова архитектура корпоративного хранилища данных (DWH)? Назовите ключевые компоненты и их назначение.
2. В чём разница между хранилищем данных (data warehouse) и озёрком данных (data lake)? Приведите примеры сфер применения.
3. Какие подходы используются для масштабирования систем хранения данных? Объясните горизонтальное и вертикальное масштабирование.
4. Что такое репликация данных? В каких случаях она применяется, и какие проблемы может вызвать?
5. Объясните CAP-теорему. Какие выводы из неё делают при проектировании распределённых систем?
6. Почему MongoDB считается перспективной системой хранения данных? Какие преимущества она даёт по сравнению с традиционными СУБД?
7. Какие архитектурные решения позволяют использовать «многоуровневое» хранение данных (hot/cold/archive tier)?
8. Какие технологии обеспечивают баланс между стоимостью, скоростью и надёжностью при хранении больших объёмов данных?
9. Почему выбор между data warehouse и data lake зависит от потребностей бизнеса и аналитических задач?
10. Как масштабируемость влияет на стоимость владения и время реакции

системы?

Практическое задание:

Разработайте архитектуру хранения данных для крупного интернет-магазина с миллионами пользователей. Определите уровни хранения (hot, warm, cold) и распределите данные по уровням (например, активные заказы – hot; старые истории – warm; архивные чеки – cold). Обоснуйте выбор технологии для каждого уровня (SSD, HDD, облачное хранилище AWS S3 Glacier). Рассчитайте ориентировочные затраты на хранение за год при объеме 500 ГБ ежедневных данных.

Тесты:

- 1. Какой подход обеспечивает горизонтальное масштабирование данных?**
 - A) Увеличение мощности одного сервера
 - B) Разделение данных по узлам (sharding)
 - C) Использование RAID-массивов
 - D) Упрощение схемы базы данных

- 2. Что такое многоуровневое хранение данных?**
 - A) Хранение данных в нескольких физических местах одновременно
 - B) Распределение данных по уровням скорости и стоимости (hot/cold data)
 - C) Сохранение одной версии данных на всех узлах
 - D) Хранение данных в виде графов

- 3. Какой из следующих сервисов предлагает автоматическое управление уровнем хранения?**
 - A) Amazon S3 with Intelligent Tiering
 - B) Oracle RAC
 - C) MySQL Cluster
 - D) Microsoft Access

- 4. Что означает термин «cold storage» в контексте хранения больших данных?**
 - A) Высокоэффективное кэширование
 - B) Хранение часто запрашиваемых данных
 - C) Долгосрочное хранение редко используемых данных
 - D) Резервное копирование в реальном времени

- 5. Почему важно использовать разделение данных (data partitioning)?**
 - A) Чтобы увеличить стоимость хранения
 - B) Чтобы снизить время отклика при выполнении запросов
 - C) Чтобы исключить использование SQL
 - D) Чтобы устранить необходимость в индексах

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3
Дополнительная: 1-3
Интернет-ресурсы: 1-10

ТЕМА № 9. СИСТЕМЫ УПРАВЛЕНИЯ БОЛЬШИМИ ДАННЫМИ

План темы:

1. Распределенные файловые системы
2. Распределенные фреймворки
3. Бенчмаркинг
4. Серверное программирование
5. Планирование
6. Системы развертывания
7. Интеграция данных
8. Информационная безопасность
9. Базы данных NoSQL и новые SQL базы данных

Цель: изучить архитектуру и функционал систем, предназначенных для эффективного управления всеми аспектами больших данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что для работы с BIG DATA необходимы специализированные системы, способные обрабатывать, хранить и анализировать данные в масштабе, обеспечивая при этом целостность и доступность.

Теоретическая часть.

Распределённые файловые системы (DFS), такие как Hadoop Distributed File System (HDFS), предназначены для хранения и обработки массивов данных на кластере серверов. Они обеспечивают надёжность за счёт репликации блоков данных, автоматическое восстановление при сбоях и масштабируемость. Особенность HDFS – работа с большими файлами, где каждый блок хранится на разных узлах, что позволяет эффективно обрабатывать данные параллельно.

Распределённые фреймворки, такие как Apache Spark, Flink, MapReduce, используются для выполнения вычислений над большими объёмами данных. Spark, отличающийся высокой скоростью благодаря работе в памяти, стал де-факто стандартом для ETL-процессов и машинного обучения. Фреймворки позволяют организовать обработку данных в режиме реального времени, а также реализовать сложные алгоритмы аналитики.

Бенчмаркинг служит методом сравнительной оценки производительности систем. Для больших данных используются специальные тесты: TPC-H, TPC-DS, YCSB. Они имитируют рабочую нагрузку, близкую к реальной, и помогают определить, какие решения наиболее эффективны в конкретной среде.

Серверное программирование в контексте больших данных связано с созданием масштабируемых API, микросервисов и интерфейсов

взаимодействия с системами хранения и обработки. Программирование должно учитывать особенности распределённых систем, включая управление состоянием, маршрутизацию запросов и работу с ошибками.

Планирование задач в системах больших данных – ключевой компонент управления ресурсами. Системы вроде YARN (Yet Another Resource Negotiator) распределяют вычислительные ресурсы между различными приложениями, обеспечивая балансировку нагрузки, контроль потребления памяти и CPU, а также приоритетную обработку.

Системы развёртывания, такие как Docker, Kubernetes, Ansible, Terraform, позволяют автоматизировать деплой, настройку и мониторинг распределённых компонентов. Они обеспечивают одинаковую среду исполнения во всех окружениях – от разработки до продакшена, что снижает риск ошибок при запуске.

Интеграция данных – процесс объединения информации из разных источников: CRM, ERP, соцсетей, IoT-устройств. Реализуется через ETL/ELT-процессы, потоковое преобразование, использование middleware. Критически важна согласованность формата, метаданных, временных шкал и правил качества данных.

Информационная безопасность в системах больших данных включает шифрование данных в покое и передаче, управление доступом, аудит действий пользователей, защиту от утечек и соответствие нормативным требованиям (GDPR, HIPAA). Особенно актуальна защита персональных данных при работе с большими потоками.

Базы данных NoSQL, такие как MongoDB, Cassandra, Redis, отличаются гибкостью схемы, масштабируемостью и способностью работать с неструктурированными данными. Новые SQL-системы (CockroachDB, YugabyteDB, TiDB) сочетают преимущества реляционных баз с возможностями распределённых систем: ACID-транзакции, масштабируемость, отказоустойчивость. Они становятся выбором для современных приложений, где нужна как надёжность, так и скорость обработки.

Вопросы для самоконтроля:

1. Что такое распределённые файловые системы? Приведите примеры (HDFS, Amazon S3, Google Cloud Storage).
2. Каковы основные функции распределённых фреймворков (Apache Spark, Flink, Beam)? В чём их отличие от традиционных систем обработки?
3. Что такое бенчмаркинг в контексте Big Data? Какие тесты (например, TPC-H, YCSB) используются для оценки производительности?
4. Какие задачи решает серверное программирование в системах управления большими данными?
5. Почему планирование ресурсов и нагрузки критично при работе с распределёнными системами?
6. Какие подходы используются при развёртывании больших систем (containerization, orchestration, cloud-native deployment)?
7. Как обеспечивается интеграция данных из различных источников (CRM,

- IoT, соцсети) в единую систему?
8. Какие меры защиты данных применяются в системах Big Data? Почему информационная безопасность здесь особенно важна?
 9. Чем отличаются NoSQL-базы данных от классических реляционных СУБД? Приведите примеры и области применения.
 10. Что такое «новые SQL»-базы данных (NewSQL)? Как они сочетают свойства реляционных СУБД и масштабируемость NoSQL?

Практическое задание:

Напишите подробный план установки и настройки кластера Hadoop (HDFS + YARN) с использованием Docker-контейнеров. Укажите, как организовать балансировку нагрузки между узлами, настроить политики хранения данных и контролировать доступ. Какие проблемы могут возникнуть при масштабировании кластера, и как они решаются через механизм отказоустойчивости?

Тесты:

1. Какой из перечисленных продуктов является системой управления большими данными?

- A) MySQL Workbench
- B) Apache Hive
- C) Visual Studio Code
- D) Adobe Acrobat

2. Что предоставляет Apache Oozie?

- A) Интерфейс для визуализации данных
- B) Платформа для управления рабочими нагрузками в среде Hadoop
- C) Система резервного копирования
- D) Язык для написания SQL-запросов

3. Какой механизм позволяет контролировать распределение ресурсов в кластере Hadoop?

- A) YARN
- B) ZooKeeper
- C) HBase
- D) Sqoop

4. Какой из следующих инструментов обеспечивает управление жизненным циклом данных в Big Data-системах?

- A) Airflow
- B) Logstash
- C) Prometheus
- D) Redis

5. Что делает Apache Ranger?

A) Предоставляет защиту доступа к данным через политики безопасности

- В) Обеспечивает автоматическое создание резервных копий
- С) Преобразует данные в формат JSON
- Д) Ускоряет чтение данных из SSD

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 10. АНАЛИТИКА БОЛЬШИХ ДАННЫХ И ЕЕ ИНСТРУМЕНТАРИЙ

План темы:

1. От традиционного анализа к операционному
2. Инструментарий для анализа больших данных: реляционные и нереляционные СУБД

Цель: понять возможности аналитики больших данных и ознакомиться с ключевыми инструментами, используемыми для получения ценных инсайтов.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что аналитика BIG DATA является основным драйвером принятия решений в бизнесе и науке, позволяя выявлять закономерности и прогнозировать будущие события.

Теоретическая часть.

Традиционная аналитика основана на пост-обработке исторических данных, когда информация извлекается, структурируется и анализируется после завершения процессов, часто с задержкой в часы или дни. Такие подходы ориентированы на отчётность, прогнозирование и стратегическое планирование, но не способны оперативно реагировать на изменения. Переход к операционной аналитике означает смещение фокуса с архивных исследований на непрерывную обработку и анализ потоковых данных в режиме реального времени. Это позволяет принимать управленческие решения мгновенно, ещё до того, как ситуация успела измениться. Операционная аналитика становится ключевым элементом цифровизации бизнеса – она обеспечивает прозрачность текущих процессов, снижает риск ошибок за счёт автоматического контроля, повышает адаптивность систем. В отличие от пассивного анализа прошлого, она активна, предиктивна и интегрирована прямо в рабочие процессы.

Анализ больших данных требует специализированных решений, отличающихся масштабируемостью, отказоустойчивостью и возможностью работы с разнородными данными. Реляционные базы данных (SQL), такие как PostgreSQL, Oracle и MySQL, остаются надёжным выбором для структурированной информации с жёсткими правилами целостности и поддержкой сложных JOIN-операций. Однако их использование ограничено при работе с большими объёмами неструктурированных или полуструктурированных данных. Нереляционные СУБД (NoSQL), в свою очередь, предлагают более гибкий подход. К ним относятся документоориентированные (MongoDB, Couchbase), колоночные (Apache Cassandra, Amazon DynamoDB), графовые (Neo4j) и параллельные хранилища (Elasticsearch). Они обеспечивают горизонтальное масштабирование, высокую

производительность при работе с массивами записей и позволяют хранить данные в форматах JSON, XML, бинарных блоках. Выбор между SQL и NoSQL зависит от типа задач: если нужна точная согласованность и сложная агрегация – выбирают реляционные системы; если важна скорость записи и гибкость структуры – предпочтение отдаётся NoSQL. Современные платформы часто сочетают оба подхода, используя гибридные архитектуры, где SQL служит для аналитики, а NoSQL – для сбора и временного хранения данных.

Вопросы для самоконтроля:

1. Как трансформировался подход к анализу данных — от традиционного reporting к операционному (operational analytics)?
2. В чём разница между аналитикой в режиме онлайн (real-time) и аналитикой в режиме offline?
3. Какие реляционные СУБД хорошо подходят для аналитики больших данных? Примеры: PostgreSQL, Oracle Analytics DB, Greenplum.
4. Какие нереляционные СУБД (NoSQL) используются в Big Data-аналитике? Приведите примеры: Cassandra, Redis, CouchDB.
5. Какие инструменты и платформы позволяют проводить анализ больших данных в реальном времени?
6. Какие преимущества даёт использование графовых баз данных (например, Neo4j) при анализе связей в данных?
7. Какие инструменты BI и визуализации наиболее часто интегрируются с Big Data-платформами?
8. Какие алгоритмы и методы анализа применяются в Big Data (статистический анализ, ML, NLP)?
9. Почему выбор инструментария зависит от типа данных, объема и требований к скорости анализа?
10. Как инструментарий анализа больших данных помогает принимать стратегические решения в бизнесе?

Практическое задание:

Выберите одну из задач: поиск аномалий в трафике сайта или прогнозирование оттока клиентов. Подберите соответствующие инструменты: Python (Pandas, Scikit-learn), R, Tableau, Power BI, Apache Zeppelin. Опишите, как вы будете использовать каждый инструмент на разных этапах анализа. Какие функциональные особенности этих средств делают их пригодными для работы с большими массивами данных?

Тесты:

1. Какой инструмент чаще всего используется для визуализации результатов анализа больших данных?

- A) Git
- B) Tableau
- C) Docker
- D) Vim

2. Что такое OLAP-анализ?

- A) Анализ в реальном времени
- B) Онлайн-обработка транзакций
- C) Поддержка многомерного анализа данных
- D) Обнаружение аномалий в потоках данных

3. Какой язык используется для анализа данных в среде Spark?

- A) Python
- B) Scala
- C) Java
- D) Все перечисленные

4. Какой из перечисленных инструментов предназначен для сбора и анализа событий в реальном времени?

- A) ELK Stack (Elasticsearch, Logstash, Kibana)
- B) Power BI
- C) SAS
- D) MATLAB

5. Какой подход позволяет находить скрытые закономерности в данных?

- A) Ручной анализ данных
- B) Статистическая проверка гипотез
- C) Обучение с учителем
- D) Машинное обучение и data mining

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 11. АНАЛИТИКА БОЛЬШИХ ДАННЫХ: ТЕХНИКИ ОБРАБОТКИ И АНАЛИЗА

План темы:

1. Методы анализа и обработки больших данных
2. Анализ больших данных в облаках

Цель: освоить различные техники и методы, применяемые для обработки и глубокого анализа больших массивов данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что разнообразие больших данных требует широкого спектра техник, от предварительной обработки до продвинутых статистических и машинных методов, для извлечения глубоких знаний.

Теоретическая часть.

Обработка больших данных невозможна без применения специфических методологий, учитывающих объём, скорость и разнообразие данных. Наиболее распространённые подходы включают:

- Параллельную обработку: распределение вычислений между множеством узлов через фреймворки Apache Spark, Flink или Hadoop MapReduce. Это позволяет эффективно обрабатывать терабайты и петабайты данных.

- Машинное обучение в реальном времени: применение алгоритмов классификации, регрессии, кластеризации и рекомендации на больших массивах данных с использованием библиотек TensorFlow, Scikit-learn, MLlib.

- Анализ в реальном времени (stream processing): обработка событийных потоков с помощью инструментов Kafka Streams, Kinesis, Apache Storm. Позволяет отслеживать события мгновенно, что важно для финансового мониторинга, IoT и пользовательского поведения.

- Геопространственный анализ: определение местоположения объектов, маршрутов, зон обслуживания на основе GPS-данных и картографической информации. Используется в логистике, доставке, городском управлении.

- Обработка естественного языка (NLP): анализ текстовой массы – отзывов, чатов, новостей – с целью выявления тональности, ключевых тем, персон. Применяется в маркетинге, HR, медицине.

Все эти методы работают в единой экосистеме, дополняя друг друга и обеспечивая комплексный взгляд на данные.

Облачные технологии кардинально изменили подход к анализу больших данных, сделав его доступным, гибким и экономически выгодным. Облачные провайдеры (AWS, Google Cloud, Microsoft Azure) предоставляют готовые сервисы для хранения, обработки и анализа данных: S3, BigQuery, Redshift,

Data Lake, EMR, Databricks. Преимущества использования облака – возможность мгновенной настройки вычислительных мощностей, автоматическая масштабируемость, защита данных, интеграция с AI/ML-сервисами. Благодаря серверным архитектурам, пользователям не нужно содержать собственные центры обработки данных. Кроме того, облачные платформы поддерживают гибридные модели: часть данных может оставаться на премисах, а анализ проводится в облаке. Это особенно актуально для компаний, работающих с чувствительной информацией. Также значительную роль играет экономия затрат: вместо капитальных инвестиций в оборудование платят лишь за фактически использованные ресурсы. Таким образом, облачная аналитика делает обработку больших данных универсальной, доступной даже малому бизнесу.

Вопросы для самоконтроля:

1. Какие основные методы анализа больших данных выделяют? Приведите примеры: описательная, диагностическая, прогнозная, директивная аналитика.
2. В чём разница между batch-обработкой и stream-processing? Когда какой подход более эффективен?
3. Какие технологии и платформы используются для параллельной и распределённой обработки больших данных?
4. Какие методы анализа применяются в Big Data для выявления скрытых закономерностей (например, кластеризация, регрессия, ассоциативные правила)?
5. Как машинное обучение интегрируется в процессы анализа больших данных? Приведите примеры.
6. Почему облако стало ключевой средой для анализа больших данных? Какие преимущества даёт облачная инфраструктура?
7. Какие сервисы облачных провайдеров (AWS, Azure, GCP) предназначены для анализа больших данных? Приведите примеры (Amazon Redshift, BigQuery, Azure Synapse).
8. Какие проблемы могут возникнуть при анализе данных в облаке (например, задержки, стоимость, безопасность)?
9. Что такое serverless-аналитика? Как она снижает затраты и упрощает развертывание?
10. Какие критерии следует учитывать при выборе инструмента анализа больших данных в зависимости от задачи?

Практическое задание:

Примените три различных метода анализа к одному и тому же набору данных (например, данные о продажах за 3 года): регрессия, кластеризация K-means, деревья решений. Сравните результаты по точности, интерпретируемости, скорости выполнения. Какие факторы повлияют на выбор метода в зависимости от типа данных и поставленной цели? Какие ошибки могут возникнуть при применении каждого метода?

Тесты:

1. Какой метод используется для выявления групп в данных?

- A) Регрессия
- B) Кластеризация
- C) Корреляция
- D) Суммирование значений

2. Что такое регрессия в контексте анализа данных?

- A) Метод классификации объектов
- B) Метод прогнозирования числового значения на основе входных переменных
- C) Метод визуализации зависимостей
- D) Метод определения выбросов

3. Какой из методов помогает выявить аномальные данные?

- A) PCA (главные компоненты)
- B) kNN (k ближайших соседей)
- C) Anomaly Detection (обнаружение аномалий)
- D) Clustering (кластеризация)

4. Что такое «размерность» в машинном обучении?

- A) Количество пользователей системы
- B) Количество признаков или столбцов в наборе данных
- C) Размер файла с данными
- D) Частота обновления данных

5. Какой метод часто используется для снижения размерности данных?

- A) Линейная регрессия
- B) PCA (Principal Component Analysis)
- C) SVM
- D) Naïve Bayes

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 12. АРХИТЕКТУРА СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

План темы:

1. Прием данных (Data Ingestion)
2. Сбор данных (Data Staging)
3. Анализ данных (Analysis Layer)
4. Представление результатов (Consumption Layer)

Цель: изучить принципы построения и компоненты систем, предназначенных для эффективной и масштабируемой обработки больших данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что успешная обработка BIG DATA напрямую зависит от продуманной, гибкой и масштабируемой архитектуры, способной справляться с возрастающими объемами информации.

Теоретическая часть.

Приём данных (Data Ingestion) – первый этап архитектуры обработки больших данных, отвечающий за сбор информации из различных источников. Он обеспечивает непрерывную подачу сырых данных в систему, будь то потоковые данные из IoT-устройств, социальных сетей, баз данных или файловые загрузки. Для эффективной интеграции используются специализированные инструменты, такие как Apache Kafka, Flume или AWS Kinesis, способные масштабироваться по объёму и скорости поступления. Ключевое требование к приёму – устойчивость к простоям, восстановление после отказов и поддержка разных форматов (JSON, Avro, CSV).

Этап сбора данных (Data Staging) предполагает временное хранение сырых данных перед их преобразованием. Здесь информация сохраняется в виде необработанного набора, часто в распределённых хранилищах типа Hadoop Distributed File System (HDFS), Amazon S3 или Google Cloud Storage. Такой подход позволяет отделить процесс получения от последующей очистки и аналитики, снижая нагрузку на основные вычислительные слои. В стейджинговой зоне могут применяться простые механизмы контроля версий, проверки целостности и фильтрации заведомо некорректных записей.

Анализ данных (Analysis Layer) – сердце системы, где происходят трансформация, агрегирование и выявление закономерностей. На этом уровне реализуются как простые операции (фильтрация, группировка), так и сложные модели машинного обучения, статистический анализ, обработка временных рядов. Вычисления выполняются на платформах, поддерживающих параллелизм: Apache Spark, Flink, Beam. Особое внимание уделяется оптимизации запросов, использованию индексов и кэширования, что

позволяет минимизировать время выполнения даже самых нагруженных задач.

Представление результатов (Consumption Layer) предназначено для доставки конечному пользователю или другим системам готовых аналитических решений. Результаты могут отображаться в виде дашбордов (Power BI, Tableau), экспортироваться в базы данных для дальнейшего использования, отправляться в API-интерфейсы или публиковаться в виде отчётов. Отличительная черта этого уровня – ориентированность на конкретные потребности: быстрый доступ, интерактивность, возможность фильтрации и детализации. Эффективность всей системы зависит от качества и своевременности представления данных, особенно в условиях реального времени.

Вопросы для самоконтроля:

1. Что такое приём данных (Data Ingestion)? Какие инструменты используются для сбора данных в реальном времени?
2. Какова цель этапа сбора данных (Data Staging)? Почему он важен перед аналитическим слоем?
3. Какие технологии используются на этапе анализа данных? Приведите примеры: Spark, Flink, Pig.
4. Что такое представление результатов (Consumption Layer)? Какие формы вывода данных существуют (отчёты, визуализации, API)?
5. Как архитектура данных должна быть спроектирована для обеспечения отказоустойчивости, масштабируемости и низкой задержки?
6. Как взаимодействуют различные уровни архитектуры в реальной системе (например, Kafka → Spark → BI-панель)?
7. Почему важно отделять уровень хранения от уровня обработки и представления?
8. Какие архитектурные шаблоны применяются для обработки потоковых данных (streaming architecture)?
9. Какие требования предъявляются к надёжности и точности данных на каждом уровне?
10. Как архитектура системы влияет на время выполнения аналитических задач и качество решений?

Практическое задание:

Постройте архитектурную диаграмму для системы обработки данных из социальной сети: от сбора лайков, комментариев и постов до формирования персонализированной ленты новостей. Укажите компоненты: источник данных, Kafka, Spark Streaming, база данных, система хранения, API. Опишите, как происходит передача данных между узлами, как обеспечивается надёжность и отказоустойчивость.

Тесты:

1. Какой из подходов описывает архитектуру, сочетающую batch- и stream-обработку?
 - A) Monolithic architecture
 - B) Lambda Architecture

- C) Event-driven architecture
- D) RESTful API

2. Что из нижеперечисленного является ключевым компонентом архитектуры Lambda?

- A) Только offline-обработка
- B) Offline-, real-time- и serving-слои
- C) Единая база данных для всех видов данных
- D) Отказоустойчивый центр обработки данных на одном сервере

3. Какой слой в архитектуре Карра отвечает за обработку потоковых данных?

- A) Слой хранения
- B) Слой анализа
- C) Слой обработки потоков
- D) Слой визуализации

4. Что представляет собой «serving layer» в архитектуре Lambda?

- A) Модуль хранения данных
- B) Часть, отвечающая за быстрое получение готовых результатов аналитики
- C) Компонент для мониторинга производительности
- D) Служба авторизации пользователей

5. Каково преимущество архитектуры Карра по сравнению с Lambda?

- A) Более высокая надёжность при отказах
- B) Простота реализации за счёт отказа от batch-слоя
- C) Большая скорость развертывания
- D) Возможность использовать только SQL

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 13. ПАРАЛЛЕЛЬНЫЕ АЛГОРИТМЫ ДЛЯ РАБОТЫ С ДАННЫМИ

План темы:

1. Операторы Map и Reduce
2. Оператор Reduce (свертка)
3. Оператор Map
4. Лямбда-архитектура

Цель: понять принципы работы и применения параллельных алгоритмов для ускорения обработки больших наборов данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что сложность и объем задач, связанных с BIG DATA, делают традиционные последовательные вычисления неэффективными, что делает параллельные вычисления критически важным инструментом.

Теоретическая часть.

Операторы Map и Reduce – фундаментальные элементы парадигмы обработки данных в режиме параллелизма. Они позволяют разбивать крупномасштабные задачи на мелкие, которые можно выполнять одновременно на множестве вычислительных узлов. Map отвечает за преобразование каждого элемента входного потока в новый объект, применяя заданную функцию ко всем данным. Например, преобразование строк в числа или извлечение ключевых полей из документов. Этот шаг гарантирует декомпозицию данных без зависимости между элементами, что делает его легко масштабируемым.

Оператор Reduce (свертка) объединяет результаты, полученные после применения Map, сводя их к единому значению. Он работает по принципу агрегации: суммирование, поиск максимума/минимума, подсчёт количества, конкатенация списков. Критически важно, чтобы функция свёртки была associative – порядок обработки не влиял на итоговый результат. Это свойство позволяет безопасно распределять вычисления среди узлов кластера и собирать промежуточные итоги в нескольких этапах.

Отдельный рассмотрению заслуживает сам оператор Map – он отвечает за перебор всех элементов данных и применение к каждому из них одной и той же функции. Его мощность проявляется в способности работать с большими объемами, не требуя сложных согласований состояний. Благодаря высокой степени параллелизма, Map становится основой для таких технологий, как Apache Spark, где каждый партиционированный блок данных обрабатывается независимо.

Лямбда-архитектура представляет собой многослойную модель,

сочетающую три направления обработки: скорость (stream processing), объем (batch processing) и ад-хоc-анализ (speed + batch). Она использует параллельные алгоритмы Map и Reduce в каждом слое, но с разными целями. В слое Stream используется легкий Map + Reduce для немедленной реакции на события; в Batch – более глубокая обработка с возможностью пересчета исторических данных; в Ad-hoc – комбинированные решения для ответов на нетиповые запросы. Такой подход обеспечивает отказоустойчивость, гибкость и соответствие требованиям бизнеса в реальном времени и с задержкой.

Вопросы для самоконтроля:

1. Что такое оператор Map? Как он работает в параллельных системах? Приведите пример.
2. Что такое оператор Reduce? Как он используется для агрегации данных?
3. Как работает механизм свертки (Reduce) в распределённых вычислениях? Приведите пример суммирования значений.
4. Почему комбинация Map + Reduce стала основой для многих Big Data-платформ (Hadoop, Spark)?
5. Что такое лямбда-архитектура? Как она решает проблему обработки данных в режиме реального времени и в фоновом режиме?
6. Какие три слоя содержит лямбда-архитектура? Опишите их назначение и взаимодействие.
7. В чём преимущество использования лямбда-подхода перед традиционной архитектурой?
8. Какие недостатки имеет лямбда-архитектура? Почему её иногда заменяют на Карра-архитектуру?
9. Почему параллельные алгоритмы, основанные на Map и Reduce, эффективны для обработки больших массивов данных?
10. Как можно использовать Map и Reduce для анализа логов, социальных сетей или IoT-устройств?

Практическое задание:

Реализуйте параллельный алгоритм подсчёта встречаемости слов в документах (на любом языке программирования). Используйте модель MapReduce: разделите текст на части, проведите map-операцию (подсчёт слов в каждой части), затем reduce (суммирование). Проанализируйте время выполнения на одном узле и на двух. Какие факторы снижают эффективность параллелизма?

Тесты:

1. **Что лежит в основе параллельного выполнения алгоритмов в Big Data?**
 - A) Использование одного процессора с повышенной тактовой частотой
 - B) Распределение задачи на несколько узлов или потоков
 - C) Применение функционального программирования
 - D) Создание единого массива данных
2. **Какой алгоритм широко используется для обработки больших массивов данных в параллельном режиме?**
 - A) Бинарный поиск

- В) Алгоритм Дейкстры
- С) MapReduce
- Д) Алгоритм Эйлера

3. Что такое «map» в модели MapReduce?

- А) Этап объединения результатов
- В) Этап сортировки данных
- С) Этап обработки данных с прямым применением функции к каждому элементу
- Д) Этап валидации входных данных

4. Какой из факторов не влияет на эффективность параллельных алгоритмов?

- А) Скорость соединения между узлами
- В) Наличие ошибок в коде
- С) Размер данных
- Д) Количество узлов кластера

5. Что такое «reduce» в модели MapReduce?

- А) Этап фильтрации данных
- В) Этап группировки и агрегации результатов после map
- С) Этап хранения данных
- Д) Этап визуализации результатов

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 14. ПРОГРАММНЫЕ ПЛАТФОРМЫ И СИСТЕМЫ ДЛЯ БОЛЬШИХ ДАННЫХ

План темы:

1. Системы управления потоками данных
2. Системы хранения больших данных
3. Платформы больших данных
4. Обработка данных в реальном времени
5. Системы управления большими данными
6. Аналитические платформы

Цель: ознакомиться с основными программными платформами и экосистемами, разработанными для работы с большими данными.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что существует множество специализированных платформ, предоставляющих инструменты для всего спектра задач, связанных с BIG DATA, и выбор подходящей платформы определяет эффективность работы.

Теоретическая часть.

Системы управления потоками данных предназначены для обработки непрерывных, высокоскоростных потоков информации в режиме реального времени. В отличие от традиционной обработки пакетов, где данные поступают группами с задержкой, такие системы обеспечивают непрерывную передачу и анализ данных, начиная с момента их возникновения. К ключевым характеристикам относятся низкая задержка, масштабируемость и устойчивость к отказам. Наиболее распространёнными решениями являются Apache Kafka, Apache Flink и Apache Storm. Они позволяют строить сложные потоковые архитектуры, объединяющие источники данных – IoT-устройства, серверные логи, финансовые операции или пользовательские действия – с аналитическими модулями и системами оповещения.

Хранение больших данных требует архитектур, способных эффективно работать с объёмами, измеряемыми терабайтами и петабайтами, при этом сохраняя доступность, надёжность и производительность. Традиционные базы данных не справляются с такими нагрузками из-за жёстких ограничений на структуру и масштабирование. Решением стали распределённые хранилища, основанные на принципах дублирования и разбиения данных. Наиболее популярны Hadoop Distributed File System (HDFS), Amazon S3, Google Cloud Storage и Apache Cassandra. Эти системы обеспечивают горизонтальное масштабирование, отказоустойчивое хранение и поддержку различных форматов – от неструктурированного JSON до бинарных объектов. Выбор между файловой системой, NoSQL-базой или объектным хранилищем зависит

от типа данных, частоты запросов и требований к согласованности.

Платформы больших данных представляют собой интегрированные экосистемы, объединяющие компоненты для сбора, хранения, обработки и анализа данных. Их цель – обеспечить единое пространство работы для специалистов разных профилей: инженеров, аналитиков и исследователей. Классическим примером является экосистема Hadoop, включающая HDFS, YARN, MapReduce, Pig, Hive и Spark. Современные платформы, такие как Cloudera, Hortonworks (сейчас часть Cloudera) и AWS Glue, добавляют уровень автоматизации, безопасности и мониторинга. Наряду с этими решениями, всё больше распространяется концепция data lake – централизованного хранилища, которое позволяет хранить сырые данные в любом виде и затем преобразовывать их по мере необходимости. Платформы также поддерживают интеграцию с машинным обучением, BI-инструментами и API-интерфейсами.

Обработка данных в реальном времени предполагает анализ информации сразу после её появления, без ожидания завершения пакетной загрузки. Это критически важно для сценариев, где своевременность принятия решений влияет на бизнес-результаты – от онлайн-банкинга и детектирования мошенничества до управления автопарком и мониторинга состояния оборудования. Реальная обработка возможна благодаря технологиям, работающим на основе событий: stream processing frameworks, таких как Apache Flink, Apache Spark Streaming и Apache Kafka Streams. Эти технологии используют оконные функции, фильтрацию, агрегацию и соединение потоков для извлечения значимых метрик и трендов. Для достижения низкой задержки требуется оптимизация сети, распределённой обработки и алгоритмической эффективности.

Системы управления большими данными – это программные среды, которые координируют работу всех элементов экосистемы: от сбора и хранения до обработки и представления результатов. Они решают проблемы, связанные с управлением жизненным циклом данных, политиками доступа, версионностью и целостностью. Такие системы обеспечивают централизованный контроль над процессами, что особенно важно в крупных организациях с множеством источников и потребителей данных. Примеры включают Apache Atlas (для метаданных), Apache Ranger (управление безопасностью), Amundsen (поиск и каталогизация). Успешная система управления должна поддерживать прозрачность, аудит изменений, политики шифрования и интеграцию с системами IAM. Без них невозможно обеспечить доверие к данным, соблюдение нормативных требований и эффективное сотрудничество между командами.

Аналитические платформы служат точкой входа для интерпретации больших массивов данных через визуализацию, прогнозирование и экспорт результатов. Они объединяют мощь вычислений с интуитивным интерфейсом, делая аналитику доступной не только специалистам, но и менеджерам. Ключевые возможности включают создание дашбордов, построение временных рядов, корреляционный анализ, A/B тестирование и модели машинного обучения. Лидирующие решения – Tableau, Power BI, Looker и

Qlik – интегрируются с различными источниками данных, включая SQL-базы, NoSQL-системы и cloud-хранилища. Современные платформы развивают направление self-service analytics, когда пользователи самостоятельно создают отчёты и проводят ad-hoc-анализ. Некоторые из них уже встраивают AI-поддержку: автоматическое выявление аномалий, рекомендации по параметрам отображения, прогнозирование на основе исторических данных.

Вопросы для самоконтроля:

1. Какие системы управления потоками данных (stream processing systems) наиболее распространены? Приведите примеры: Kafka, Flink, Spark Streaming.
2. В чём отличие хранилищ больших данных от традиционных СУБД? Какие типы систем хранения данных используются в Big Data?
3. Какие платформы Big Data считаются де-факто стандартами? Например, Hadoop, Spark, Snowflake, Databricks.
4. Что необходимо для обеспечения обработки данных в реальном времени? Какие технологии это обеспечивают?
5. Какие функции выполняет система управления большими данными (например, Apache Ambari, Cloudera Manager)?
6. Какие возможности предоставляют аналитические платформы (Power BI, Tableau, Looker, Superset)?
7. Как выбирается платформа для проекта Big Data? По каким критериям?
8. Почему микросервисная архитектура часто сочетается с платформами Big Data?
9. Какие преимущества даёт использование open-source решений (например, Apache projects) в Big Data?
10. Как интеграция с API и внешними системами влияет на эффективность аналитических платформ?

Практическое задание:

Задание 1. Проведите сравнительный анализ трёх систем управления базами данных для работы с большими данными: Apache Hadoop (HDFS), Apache Spark и Amazon S3. Укажите ключевые отличия в архитектуре, возможностях обработки данных, масштабируемости и поддержке инструментов ETL. Опишите, какие из них наиболее подходят для batch-обработки, а какие – для stream-аналитики.

Задание 2. Создайте простой пайплайн обработки данных с использованием Apache Airflow. Реализуйте DAG (Directed Acyclic Graph) для загрузки данных из CSV-файла, их трансформации (очистка пропущенных значений, нормализация числовых признаков) и записи в хранилище Hive. Проиллюстрируйте структуру скриптов и объясните преимущества использования описания процессов через код.

Тесты:

1. Какая из перечисленных платформ является полноценным решением для обработки больших данных на основе Hadoop?
 - A) Apache Spark
 - B) Microsoft Excel

- C) SQLite
- D) MongoDB Compass

2. Что из нижеперечисленного является частью экосистемы Hadoop?

- A) Apache Kafka
- B) Apache Spark
- C) Apache HBase
- D) Все вышеупомянутое

3. Какой проект позволяет выполнять SQL-подобные запросы к данным в Hadoop?

- A) Flume
- B) Pig
- C) Hive
- D) Sqoop

4. Для чего используется Apache Zeppelin?

- A) Для хранения данных в виде таблиц
- B) Для выполнения и визуализации аналитических заметок в интерактивном режиме
- C) Для перемещения данных между системами
- D) Для контроля версий кода

5. Какой из инструментов предназначен для интеграции данных из разных источников в Hadoop?

- A) NiFi
- B) Jupyter Notebook
- C) Elasticsearch
- D) RabbitMQ

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 15. ОБОРУДОВАНИЕ И ЦЕНТРЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

План темы:

1. Характеристики вычислительного оборудования для BIG DATA
2. Типы систем хранения данных для BIG DATA
3. Концепция и архитектура аппаратных кластеров для BIG DATA
4. Инфраструктура центров обработки данных (ЦОД) для BIG DATA
5. Масштабируемость и отказоустойчивость аппаратных решений
6. Специализированные аппаратные решения и ускорители
7. Оборудование для периферийных вычислений (Edge Computing)
8. Экономические и эксплуатационные аспекты оборудования
9. Облачная инфраструктура как решение для BIG DATA
10. Тенденции развития оборудования для BIG DATA

Цель: понять требования к аппаратному обеспечению и архитектуре центров обработки данных, необходимых для масштабируемой работы с BIG DATA.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что физическая инфраструктура, включая серверы, сети и системы хранения, играет ключевую роль в обеспечении производительности, надежности и масштабируемости систем обработки больших данных.

Теоретическая часть.

Оборудование для работы с большими данными строится вокруг ключевых требований: высокая производительность, объёмная память, скорость передачи данных и масштабируемость. В основе лежат многоядерные процессоры Intel Xeon или AMD EPYC с поддержкой множественных потоков, что позволяет эффективно распределять нагрузку при параллельной обработке. Особое внимание уделяется оперативной памяти – современные системы комплектуются до нескольких терабайтов ОЗУ, что обеспечивает загрузку массивов данных непосредственно в память без необходимости постоянного обращения к диску. Критически важна также пропускная способность шины и интерфейсов – использование PCIe Gen4/Gen5, NVMe-накопителей и сетевых инфраструктуры 10–100 Гбит/с повышает общую производительность кластера.

Платформы хранения больших данных опираются на три основных подхода: файловые системы с распределённым доступом (HDFS), объектное хранилище (S3, MinIO) и NoSQL-базы (Cassandra, DynamoDB). HDFS обеспечивает надёжное хранение больших объёмов неструктурированной информации за счёт разбиения файлов на блоки и их репликацию. Объектные

хранилища ориентированы на массовое сохранение данных с метаданными, легко масштабируются по вертикали и горизонтали. Нестандартные базы данных предлагают гибкие модели данных – от графовых до временных рядов – позволяя работать с разнородными потоками информации. Выбор зависит от типа задач: аналитика – предпочитают HDFS, хранение медиаданных – объектные системы, транзакционные данные – NoSQL.

Аппаратные кластеры для Big Data строятся по принципу распределённой обработки: множество недорогих серверов, соединённых высокоскоростной магистралью, формируют единый вычислительный ресурс. Архитектура основана на модульности: каждый узел выполняет свои функции – master-узлы управляют заданиями, worker-узлы – обрабатывают данные. Используется парадигма «распределённый кэш»: данные физически располагаются рядом с местом их обработки, снижая время передачи. Применяются технологии like YARN, Mesos, Kubernetes для управления контейнерами и ресурсами. Такие кластеры могут состоять из десятков до тысяч узлов, обеспечивая параллельную работу с петабайтами данных.

ЦОД для Big Data проектируются с учётом энергоэффективности, охлаждения и надёжности. Они включают специальные стойки с плотной установкой серверов, системы охлаждения жидкостью или воздухом, мощные электросистемы с резервным питанием. Учитывается архитектурная изоляция – отдельные зоны для хранения, вычислений и сети. Реализованы системы мониторинга нагрузки, контроля температурного режима и автоматического перераспределения трафика. ЦОДы часто размещаются в регионах с холодным климатом или близко к источникам возобновляемой энергии – это снижает эксплуатационные расходы. Важен принцип «выделенного пространства» – каждому проекту выделяется свой сегмент, исключающий конкуренцию ресурсов.

Решения для Big Data спроектированы на принципах горизонтального масштабирования: добавление новых узлов не требует замены всей инфраструктуры. Отказ одного элемента не ведёт к полному сбою благодаря репликации данных и механизму переназначения задач. Критически важна система резервного копирования, которая может реализовываться как в реальном времени (синхронная репликация), так и по расписанию (асинхронная). Для защиты от отказов применяются дублирующие каналы связи, избыточные источники питания, системы аварийного перехода. Платформы используют алгоритмы самообучения и адаптивного управления – если один узел начинает отставать, нагрузка перебрасывается на другие.

В условиях обработки огромных объемов данных всё шире применяются акселераторы – GPU, FPGA и ASIC. NVIDIA Tesla V100/V100 Tensor Core обеспечивают ускорение машинного обучения и численных расчётов. FPGA используются для оптимизации конкретных алгоритмов, таких как фильтрация данных в реальном времени. ASIC-решения, созданные под определённые задачи (например, Google TPU), позволяют достичь экстремальной скорости обработки при минимальном энергопотреблении. Эти устройства подключаются через PCI Express или специализированные

слоты, интегрируясь в архитектуру кластера. Их применение особенно актуально в AI-проектах, где требуется миллиарды операций в секунду.

Периферийные вычисления предполагают обработку данных прямо у источника генерации – в IoT-устройствах, камерах, датчиках. Это снижает задержку и нагрузку на центральный ЦОД. Аппаратные платформы Edge-устройств – это мини-серверы ARM-архитектуры, совместимые с Docker, Kubernetes, поддерживающие low-latency обработку. Используются SSD-накопители с низким потреблением, радиочастотные модули 5G/LTE, встроенные сенсоры. Такие системы способны выполнять первичную фильтрацию, классификацию и даже обучение небольших моделей. Поддерживаются протоколы MQTT, CoAP, что делает их совместимыми с облачными сервисами. Решения применяются в автономных автомобилях, смарт-фабриках, городской инфраструктуре.

Выбор оборудования для Big Data определяется стоимостью владения (ТСО) – суммарными затратами на покупку, обслуживание, электроэнергию и ремонт. Чем выше энергоёмкость, тем дороже эксплуатация. Поэтому популярны low-power серверы на ARM, использование вторичных комплектующих и оптимизация нагрузки. Расходы на охлаждение составляют до 30% от бюджета ЦОД. Значительно влияет и стоимость хранения – чем больше данных, тем выше цена за терабайт. При этом экономия достигается за счёт использования tiered storage: активные данные – на быстрых SSD, пассивные – на дешёвых HDD. Также учитываются издержки на управление – автоматизация процессов снижает трудозатраты.

Облачные провайдеры (AWS S3, Azure Blob Storage, GCP Cloud Storage) предоставляют готовые решения для хранения и обработки данных. Они предлагают бесшовную интеграцию с сервисами вычислений: EC2, EMR, Dataproc. Преимущества – нет необходимости в капитальных вложениях, возможность быстро масштабировать, высокая отказоустойчивость, защита данных. Благодаря serverless

Вопросы для самоконтроля:

1. Какие характеристики вычислительного оборудования критичны для эффективной обработки больших данных? (Производительность CPU/GPU, RAM, пропускная способность сети)
2. Какие типы систем хранения данных используются в Big Data (HDD, SSD, NVMe, объектное хранение)?
3. Как устроена архитектура аппаратного кластера для Big Data? Почему распределённость — ключевое условие?
4. Какие требования предъявляются к инфраструктуре ЦОД для работы с Big Data (охлаждение, энергообеспеченность, отказоустойчивость)?
5. Как достигается масштабируемость и отказоустойчивость в hardware-решениях?
6. Что такое GPU-ускорители, FPGA и другие специализированные устройства? Как они влияют на производительность аналитики?
7. В чём смысл Edge Computing в контексте Big Data? Как он снижает нагрузку на центральные системы?

8. Какие экономические показатели учитываются при выборе оборудования (CAPEX vs OPEX, energy efficiency, lifetime cost)?
9. Почему облачная инфраструктура (AWS, Azure, GCP) становится основой для развертывания Big Data?
10. Какие тренды в развитии аппаратного обеспечения влияют на будущее Big Data (quantum computing, neuromorphic chips)?

Практическое задание:

Задание 1. Разработайте техническое предложение по созданию мини-ЦОД (центра обработки данных) для обработки данных объёмом до 10 TB ежедневно. Укажите тип серверов, конфигурацию дисков (SSD/HDD), требования к сети (пропускная способность, протоколы), систему охлаждения и энергопитания. Обоснуйте выбор оборудования с учётом баланса производительности, стоимости и надёжности.

Задание 2. Проанализируйте влияние отказоустойчивости и распределённой архитектуры на эффективность ЦОД. Рассмотрите примеры использования RAID-массивов, кластеризации узлов и резервного копирования. Какие меры позволяют минимизировать простои при выходе из строя отдельных компонент?

Тесты:

1. Какой тип оборудования наиболее важен для повышения пропускной способности в ЦОД?

- A) ОЗУ с низкой ёмкостью
- B) Сетевые интерфейсы с высокой пропускной способностью
- C) HDD вместо SSD
- D) Низкоскоростные процессоры

2. Что такое «высота стойки» в контексте ЦОД?

- A) Количество серверов в шкафу
- B) Единица измерения плотности установки оборудования (U)
- C) Высота комнаты в метрах
- D) Уровень защиты данных

3. Какой элемент ЦОД обеспечивает бесперебойную работу оборудования?

- A) Резервное питание и кондиционирование
- B) Установка офисных столов
- C) Установка Wi-Fi-роутеров
- D) Программное обеспечение для общения

4. Что такое «цифровой двойник» в ЦОД?

- A) Копия физической инфраструктуры, отражающая её состояние в цифровом виде
- B) Специальная форма визуализации сетевых потоков
- C) Программа для расчёта затрат
- D) Устройство для хранения данных на магнитной ленте

5. Какое явление связано с перегревом серверов в ЦОД?

- A) Тепловое расслоение
- B) Thermal throttling
- C) Cloud drift
- D) Memory leak

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 16. ВИЗУАЛИЗАЦИЯ ДАННЫХ И РЕЗУЛЬТАТОВ АНАЛИЗА

План темы:

1. Типы, задачи и виды визуализации
2. Графики, диаграммы, инфографика,
3. Интерактивный сторителлинг, дашборды
4. Язык R и его возможности
5. AmazonS3
6. Дедупликация данных

Цель: освоить техники и инструменты для наглядного представления данных и результатов их анализа.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что визуализация делает сложные данные и выводы понятными для широкой аудитории, ускоряет принятие решений и помогает выявлять неочевидные закономерности.

Теоретическая часть.

Визуализация данных представляет собой способ представления информации в графической форме, позволяющий выявить скрытые закономерности, тренды, аномалии и зависимости. Основная цель – упростить восприятие сложных массивов данных, сделать анализ более наглядным и доступным. Различают несколько типов визуализаций: статические изображения, интерактивные отчёты, дашборды и элементы сторителлинга. Каждый из них решает свои задачи – от быстрого понимания общей картины до глубокого погружения в детали.

Графики и диаграммы являются фундаментом визуального анализа. Линейчатые, столбиковые, круговые, точечные, гистограммы, box-plot – каждый тип подходит для определённого типа данных и цели исследования. Например, линейчатые графики эффективны при отслеживании изменений во времени, а scatter plot помогает находить корреляции между переменными. Инфографика же объединяет данные, дизайн и нарратив, превращая информацию в зрительную историю, легко воспринимаемую аудиторией без технической подготовки.

Интерактивный сторителлинг позволяет построить повествование вокруг данных, где пользователь сам выбирает, какие параметры рассматривать, масштабировать, фильтровать. Такие подходы особенно актуальны в дашбордах – комплектах визуальных элементов, объединённых единым пространством управления. Дашборды используются как для оперативного мониторинга ключевых метрик, так и для принятия стратегических решений, обеспечивая контроль над процессами в реальном времени.

Язык программирования R предоставляет мощные инструменты для создания сложной и высококачественной визуализации через библиотеки ggplot2, lattice, plotly и другие. Благодаря модульности и расширяемости, R позволяет не только строить стандартные графики, но и реализовывать кастомные решения, включая анимацию, 3D-графики и интеграцию с веб-интерфейсами.

Хранение больших объемов данных требует надёжных решений. Amazon S3 – объектное хранилище от AWS, обеспечивающее высокую отказоустойчивость, масштабируемость и безопасность. Его можно использовать для хранения исходных данных перед их обработкой, сохранения результатов аналитики, а также для организации архива данных. Интеграция S3 с системами ETL или платформами обработки данных (например, AWS Glue, SageMaker) делает его незаменимым звеном в экосистеме аналитики.

Одна из первых стадий предобработки данных – дедупликация. Удаление дубликатов необходимо для обеспечения достоверности анализа, поскольку повторяющиеся строки могут искажать результаты: завышать средние значения, искажать частотные распределения, создавать ложные зависимости. Методы дедупликации варьируются от простого сравнения полей до использования алгоритмов семантического сходства, особенно когда речь идёт о текстовых данных. Эффективная дедупликация снижает нагрузку на системы обработки и увеличивает качество конечного результата.

Вопросы для самоконтроля:

1. Какие типы визуализации данных существуют? Приведите примеры: временные ряды, тепловые карты, деревья решений.
2. В чём разница между графиками, диаграммами и инфографикой? Как выбрать подходящий тип?
3. Что такое интерактивный сторителлинг? Как он помогает в передаче аналитического сообщения?
4. Что такое dashboards («информационные панели»)? Какие элементы должны быть включены для эффективной визуализации?
5. Как язык R (пакеты ggplot2, Shiny) используется в визуализации больших данных?
6. Как Amazon S3 используется в архитектуре визуализации данных?
7. Почему важно проводить дедупликацию данных перед визуализацией? Какие последствия могут быть при её отсутствии?
8. Как обеспечивается масштабируемость визуализации при работе с большими объёмами данных?
9. Какие инструменты позволяют строить визуализацию в реальном времени?
10. Как оценить успех визуализации: по критериям понятности, точности, воздействия на пользователя?

Практическое задание:

Задание 1. Используя данные о продажах за год (месячные показатели по регионам и категориям товаров), создайте интерактивную визуализацию с помощью Python (библиотеки Plotly или Bokeh). Включите:

- гистограмму с разбивкой по месяцам;
- тепловую карту по регионам и категориям;
- линейный график тренда по каждому товару.

Объясните выбор цветовой палитры и формата представления, чтобы обеспечить читаемость даже при высокой плотности данных.

Задание 2. Разработайте шаблон отчётности для бизнес-аналитика. Он должен содержать не менее четырёх элементов: диаграмма сегментации клиентов, карта распространённости обращений в службу поддержки, временной ряд KPI (например, средний чек), а также сводная таблица с ключевыми метриками. Подберите оптимальные типы графиков и покажите, как они помогают быстро интерпретировать информацию.

Тесты:

1. Какой инструмент наиболее популярен для построения интерактивных дашбордов?

- A) Matplotlib
- B) Plotly
- C) Pandas
- D) NumPy

2. Что из нижеперечисленного лучше всего подходит для визуализации распределения данных?

- A) Диаграмма рассеяния
- B) Гистограмма
- C) Таблица с числами
- D) Круговая диаграмма

3. Какой тип диаграммы используется для демонстрации трендов во времени?

- A) Гистограмма
- B) Линейчатая диаграмма
- C) Диаграмма Венна
- D) Box plot

4. Что означает термин «информационная перегрузка» визуализации?

- A) Невозможность сохранить данные
- B) Чрезмерное количество элементов на графике, затрудняющее восприятие
- C) Ошибка в математическом расчете
- D) Отсутствие цветового оформления

5. Какой стандарт используют современные библиотеки для визуализации?

- A) HTML/CSS/JavaScript
- B) XML/JSON
- C) SQL

D) Markdown

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 17. КЛАССИФИКАЦИЯ ЗАДАЧ АНАЛИЗА ДАННЫХ

План темы:

1. DataMining
2. Интеллектуальный анализ данных, его, отличия и задачи
3. Text Mining
4. Web Mining
5. Web Content Mining
6. Web Usage Mining
7. Social media mining
8. RapidMiner

Цель: разобраться в разнообразии задач, которые решаются с помощью анализа данных, и понять их типологию.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что четкое понимание типов задач анализа данных позволяет выбрать наиболее подходящие методы, инструменты и подходы для их эффективного решения.

Теоретическая часть.

Data Mining – это процесс извлечения скрытых, ранее неизвестных, потенциально полезных закономерностей, зависимостей и моделей из больших объёмов данных. Он представляет собой многоэтапную технологию, объединяющую методы статистики, машинного обучения, искусственного интеллекта и баз данных. Основная цель – преобразовать сырые данные в понятные и применимые знания. В основе Data Mining лежит поиск паттернов, таких как ассоциации, кластеры, классификации или прогнозирование временных рядов. Применение этого подхода распространено в маркетинге, банковской сфере, медицинских исследованиях и промышленности.

Интеллектуальный анализ данных – расширённое понятие по сравнению с Data Mining, подчёркивающее интерактивность, семантический анализ и вовлечение эксперта в процесс открытия знаний. В отличие от автоматизированного и часто шаблонного подхода Data Mining, IDA акцентирует внимание на интерпретируемости результатов, адаптивности и способности выявлять глубокие причинно-следственные связи. Задачи IDA включают обнаружение аномалий, выявление скрытых тенденций, моделирование сложных систем и поддержку принятия решений. Ключевое различие – в приоритете качества и осмысленности найденных паттернов перед их количественной значимостью.

Text Mining – направление анализа неструктурированных текстовых данных с целью извлечения информации, установления связей между концептами и формирования структурированной информации. Процесс

охватывает предобработку текста (очистка, лемматизация), определение ключевых слов, классификацию документов, извлечение сущностей, анализ тональности и построение сетей отношений. Применяется в юридических исследованиях, мониторинге общественного мнения, научной литературе и CRM-системах. Особую важность имеет сочетание естественной языковой обработки (NLP) и методов анализа данных, что позволяет переходить от «слова» к «знанию».

Web Mining – область анализа данных, извлекаемых из веб-среды, включающая три основных направления: содержимое веб-страниц, использование пользовательского поведения и структуру ссылок. Цель – получить информацию о пользователях, контенте и динамике взаимодействий. Этот тип анализа используется для персонализации сервисов, улучшения навигации, повышения эффективности рекламы и прогнозирования трендов. Важнейшие технологии – сбор данных через web crawling, фильтрация и анализ гипертекстовой структуры, применение алгоритмов кластеризации и рекомендаций.

Web Content Mining – специализированный вид Web Mining, сосредоточенный на анализе информационного содержания веб-страниц. Решает задачи извлечения смысла из текста, категоризации страниц, определения ключевых тем, выделения метаданных и построения онтологий. Используются методы NLP, классификации, кластеризации и семантического анализа. Например, системы могут группировать новостные статьи по темам, извлекать события из блогов или создавать карты знаний из множества источников. Критически важен для создания intelligent search engines, digital libraries и информационных агрегаторов.

Web Usage Mining – анализ истории взаимодействий пользователей с веб-ресурсами, основанной на лог-файлах серверов, cookie-данных, JavaScript-отчётности и других источниках поведенческой метрики. Основная задача – восстановить путь пользователя, определить популярные маршруты, выявить группы по интересам и спрогнозировать поведение. Методы включают анализ последовательностей действий (session analysis), построение графов переходов, кластеризацию пользователей по профилю активности. Практическое применение – оптимизация дизайна сайтов, улучшение UX/UI, настройка A/B-тестов и персонализация контента.

Social Media Mining – анализ данных, генерируемых в социальных платформах (Facebook, Twitter/X, Instagram, TikTok и др.), с целью выявления эмоциональной окраски, трендов, влияния пользователей и формирования общественного мнения. Обрабатывается огромный объём нетрадиционных данных: тексты, изображения, видео, хэштеги, лайки, репосты. Применяются методы sentiment analysis, topic modeling, network analysis и detection of viral content. Особенностью является высокая скорость изменения данных и необходимость учёта культурных, языковых и контекстуальных особенностей. Используется в политике, PR, брендинге, катастрофическом управлении и мониторинге кризисов.

RapidMiner – платформа для анализа данных, реализующая принципы

Data Mining, Machine Learning и Business Intelligence в едином визуальном интерфейсе. Поддерживает весь цикл работы с данными: импорт, очистка, преобразование, выбор признаков, обучение моделей, оценку и визуализацию результатов. Базируется на стандарте PMML и предлагает широкий набор алгоритмов (классификация, регрессия, кластеризация, ассоциации). Отличается открытостью, масштабируемостью и возможностями интеграции с другими системами. Широко применяется в образовании, бизнес-аналитике и разработке решений для Big Data благодаря своей доступности и мощному функционалу.

Вопросы для самоконтроля:

1. Что такое Data Mining? Какие типы задач решаются в этом направлении?
2. В чём разница между Data Mining и интеллектуальным анализом данных (Intelligent Data Analysis, IDA)?
3. Что такое Text Mining? Какие задачи решаются при анализе неструктурированных текстовых данных?
4. Что такое Web Mining? Какие три основных направления выделяют в нем?
5. Что входит в Web Content Mining? Приведите примеры (поиск по страницам, семантическая аналитика).
6. Что такое Web Usage Mining? Как он помогает понять поведение пользователей?
7. Как осуществляется анализ данных из социальных медиа (Social Media Mining)? Какие вызовы есть?
8. Что такое RapidMiner? Какие функции предоставляет эта платформа?
9. Какие технологии и алгоритмы используются в автоматизации анализа данных?
10. Как классификация задач помогает выбирать правильный инструмент и метод анализа?

Практическое задание:

Задание 1. Для каждого из перечисленных ниже сценариев определите тип задачи анализа: классификация, кластеризация, регрессия, ассоциативный анализ, детектирование аномалий.

- Предсказание вероятности оттока пользователя из мобильного приложения.
- Группировка покупателей по поведению в интернет-магазине.
- Выявление необычных паттернов в трафике сети.
- Поиск продуктов, которые часто покупаются вместе.
- Расчёт прогноза спроса на продукты на основе исторических данных.

Приведите пояснение к каждому ответу, указав, почему выбран именно этот тип задачи.

Задание 2. На основе реального набора данных (например, профили пользователей соцсети: возраст, активность, время онлайн, количество друзей) постройте дерево задач анализа. Начните с общего вопроса («Как повысить вовлечённость?»), затем разбейте его на конкретные задачи: классификацию пользователей по уровню вовлечённости, поиск факторов, влияющих на активность, кластеризацию групп пользователей для персонализированного

маркетинга.

Тесты:

1. К какой категории относится задача прогнозирования продаж на следующий месяц?

- A) Классификация
- B) Кластеризация
- C) Прогнозирование (регрессия)
- D) Обнаружение аномалий

2. Какая задача решается с помощью алгоритма K-Means?

- A) Прогнозирование цен акций
- B) Классификация клиентов по категориям
- C) Обнаружение мошенничества
- D) Группировка данных по схожести

3. Что такое задача классификации?

- A) Предсказать значение величины
- B) Присвоить объекту один из заранее известных классов
- C) Определить, есть ли в данных выбросы
- D) Найти зависимости между переменными

4. Какая задача направлена на выявление аномальных паттернов в данных?

- A) Регрессия
- B) Кластеризация
- C) Обнаружение аномалий
- D) Ассоциативный анализ

5. Что такое ассоциативный анализ?

- A) Поиск связи между двумя переменными
- B) Нахождение закономерностей в покупках (например, «купили хлеб – купили масло»)
- C) Прогнозирование будущего поведения
- D) Группировка данных по возрасту

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

ТЕМА № 18. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

План темы:

1. Статистические гипотезы и критерии
2. Машинное обучение
3. Метрический и линейных классификаторы
4. ROC-кривая
5. Кластерный анализ

Цель: изучить фундаментальные статистические методы, необходимые для количественного анализа и интерпретации данных.

Знания и умения, приобретаемые студентом в результате освоения темы, формируемые компетенции или их части: способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, определять стратегию действий (УК-1); способен к разработке моделей и методов описания и объяснения социальных явлений, и процессов (ПК-1).

Актуальность темы: связана с тем, что статистика является основой для понимания данных, выявления тенденций, проверки гипотез и построения прогнозных моделей, что критически важно для любой области, работающей с информацией.

Теоретическая часть.

Статистические гипотезы представляют собой предположения о свойствах генеральной совокупности, проверяемые по выборочным данным. Основная цель – определить, можно ли отвергнуть нулевую гипотезу (H_0) в пользу альтернативной (H_1). Для этого используется статистический критерий – функция от выборочных данных, значение которой позволяет принять решение. Критерии подразделяются на параметрические (предполагают нормальность распределения) и непараметрические (не зависят от формы распределения). Распространённые примеры: t-критерий Стьюдента, F-критерий Фишера, χ^2 -критерий Пирсона. Уровень значимости (α) задаёт вероятность ошибки первого рода – отказ от нулевой гипотезы при её истинности. Важную роль играет мощность критерия – способность правильно обнаружить альтернативную гипотезу.

Машинное обучение – направление искусственного интеллекта, изучающее алгоритмы, позволяющие строить модели на основе экспериментальных данных без явного программирования. Основные типы задач: классификация, регрессия, кластеризация. Алгоритмы делятся на обучение с учителем (обучение по размеченным данным), без учителя (выявление скрытых паттернов в необработанных данных) и с частичным контролем. Цель – минимизировать ошибку прогнозирования на новых данных. При этом важно обеспечить баланс между переобучением (слишком точное воспроизведение обучающей выборки) и недообучением (неспособность зафиксировать общие закономерности).

Метрические классификаторы основаны на сравнении объекта с элементами обучающей выборки по мере близости. Наиболее известны k-

ближайших соседей (kNN): класс определяется голосованием среди k наиболее близких к объекту образцов. Эффективность зависит от выбора метрики (Евклида, Чебышёва, Минковского) и числа соседей. Линейные классификаторы строят решающую границу как гиперплоскость, разделяющую классы. Примеры: логистическая регрессия, перцептрон, SVM с линейным ядром. Эти методы эффективны при наличии линейной разделимости или близкой к такой зависимости. Преимущество – интерпретируемость и скорость работы, но они чувствительны к шумам и неадекватны при сложных зависимостях.

ROC-кривая (Receiver Operating Characteristic) демонстрирует качество бинарного классификатора при разных порогах разделения. По оси X откладывается доля ложно положительных результатов (FPR), по оси Y – доля истинно положительных (TPR). AUC (Area Under Curve) – площадь под кривой – количественная характеристика качества: чем ближе к 1, тем лучше классификатор. Значение 0.5 соответствует случайному угадыванию. ROC полезна при анализе баланса между чувствительностью и специфичностью, особенно когда классы несбалансированы. Она позволяет выбирать оптимальный порог для максимизации эффективности в конкретном приложении.

Кластерный анализ – метод группировки объектов так, чтобы внутри кластеров они были максимально похожи, а между кластерами – максимально различались. Отличается от классификации тем, что не использует заранее заданные метки. Наиболее популярные подходы: K-means (минимизация суммарной дисперсии), иерархическая кластеризация (пошаговое слияние/расщепление кластеров), DBSCAN (на основе плотности). Выбор числа кластеров часто осуществляется по правилу elbow, Silhouette-оценке или методу силуэта. Кластеризация применяется для сегментации клиентов, анализа пользовательских поведенческих паттернов, снижения размерности данных перед обучением моделей.

Вопросы для самоконтроля:

1. Что такое статистическая гипотеза? Как проводится проверка гипотез (t-критерий, χ^2 , ANOVA)?
2. Какие основные задачи решает машинное обучение в контексте анализа данных?
3. Что такое метрические и линейные классификаторы? Приведите примеры: kNN, Logistic Regression.
4. Что такое ROC-кривая? Как интерпретируется площадь под кривой (AUC)?
5. Как работает кластерный анализ? Какие методы используются (k-means, hierarchical clustering)?
6. Как оценить качество кластеров? Какие метрики применимы (silhouette score, within-cluster sum of squares)?
7. В чём разница между supervised и unsupervised learning в контексте статистических методов?
8. Как выбрать подходящий статистический метод в зависимости от типа данных и цели исследования?

9. Почему важно учитывать предпосылки методов (нормальность, равенство дисперсий, независимость наблюдений)?
10. Как статистические методы помогают переходить от описания данных к прогнозированию и принятию решений?

Практическое задание:

Задание 1. Проведите проверку гипотезы о различии средних значений двух независимых выборок: ежемесячная выручка в двух регионах (по 50 записей в каждом). Используя t-критерий Стьюдента, рассчитайте p-value, определите, есть ли статистически значимое отличие между регионами. Объясните, что такое уровень значимости $\alpha = 0,05$ и как интерпретировать результаты.

Задание 2. Постройте корреляционную матрицу для набора переменных: доход, расходы, время на сайте, количество заказов. Определите пары, где связь может быть сильной, слабой или отсутствовать. Проанализируйте, могут ли высокие корреляции сигнализировать о мультиколлинеарности в модели регрессии. Что нужно сделать перед построением такой модели, если корреляция выше 0,8?

Тесты:

1. Какой показатель характеризует степень связи между двумя переменными?

- A) Дисперсия
- B) Коэффициент корреляции
- C) Мода
- D) Среднее арифметическое

2. Что такое p-value в статистическом тестировании гипотез?

- A) Вероятность того, что выборка случайна
- B) Вероятность получить такой же или более экстремальный результат при справедливости нулевой гипотезы
- C) Уровень доверия к результату
- D) Ошибка первого рода

3. Какой метод используется для проверки различий между средними значениями двух групп?

- A) t-критерий Стьюдента
- B) ANOVA
- C) Хи-квадрат
- D) Метод главных компонент

4. Что означает термин «дисперсионный анализ» (ANOVA)?

- A) Сравнение средних значений более чем двух групп
- B) Оценка степени вариабельности данных
- C) Проверка нормальности распределения
- D) Метод кластеризации данных

5. Какой параметр можно оценить с помощью доверительного интервала?

- A) Среднее значение генеральной совокупности
- B) Количество выборок
- C) Стандартное отклонение данных в таблице
- D) Количество строк в таблице

Список литературы, рекомендуемый к использованию по данной теме:

Основная: 1-3

Дополнительная: 1-3

Интернет-ресурсы: 1-10

СПИСОК ЛИТЕРАТУРЫ

Основная литература:

1. Параскевов, А. В. Большие данные: учебное пособие / А. В. Параскевов, А. Э. Сергеев. - Москва; Вологда: Инфра-Инженерия, 2024. - 148 с. - ISBN 978-5-9729-2120-1. - Текст: электронный. - URL: <https://znanium.ru/catalog/product/2169699> (дата обращения: 12.03.2026). - Режим доступа: по подписке.

2. Ланских, Ю. В. Введение в большие данные: учебное пособие / Ю. В. Ланских, В. Г. Ланских, К. В. Родионов. – Киров: ВятГУ, 2023. – 172 с. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/408566> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

3. Конкина, В. В. Введение в большие данные и анализ информации: учебное пособие / В. В. Конкина, А. Б. Борисенко, И. Л. Коробова. – Тамбов: ТГТУ, 2024. – 82 с. – ISBN 978-5-8265-2749-8. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/472331> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

Дополнительная литература:

1. Адлер, Ю. П. Статистическое управление процессами. «Большие данные»: учебное пособие / Ю. П. Адлер, Е. А. Черных. - Москва: Изд. Дом МИСиС, 2016. - 52 с. - ISBN 978-5-87623-969-3. - Текст: электронный. - URL: <https://znanium.com/catalog/product/1232190> (дата обращения: 12.03.2026). - Режим доступа: по подписке.

2. Буданцев, А. В. Методы анализа данных (01.03.04, 09.03.03): учебное пособие / А. В. Буданцев, Н. А. Морошкин, Н. Н. Тетерин. – Москва: РТУ МИРЭА, 2025. – 63 с. – ISBN 978-5-7339-2646-9. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/507479> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

3. Баланов, А. Н. Цифровая трансформация государственного сектора. Интеграция, управление и безопасность: учебное пособие для вузов / А. Н. Баланов. – Санкт-Петербург: Лань, 2024. – 328 с. – ISBN 978-5-507-49310-4. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/417749> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

Интернет-ресурсы:

1. Сайт Института социологии РАН – <http://www.isras.ru/socis.html>.
2. Журнал «Социологические исследования» // [Электрон. ресурс]. – Режим доступа: <http://socis.isras.ru>
3. Гуманитарный Интернет портал // [Электрон. ресурс]. – Режим доступа: www.auditorium.ru
4. Санкт-Петербургский государственный университет. Факультет социологии [Электрон. ресурс]. – Режим доступа: <http://www.soc.spb.ru/archive/>

5. «Фонд общественное мнение» (ФОМ) // [Электрон. ресурс]. – Режим доступа: <http://www.fom.ru/>
6. ВЦИОМ [Электрон. ресурс]. – Режим доступа: <https://wciom.ru/>
7. Левада-центр [Электрон. ресурс]. – Режим доступа: <http://www.levada.ru/>
8. Электронно-библиотечная система «Университетская библиотека онлайн» - www.biblioclub.ru;
9. Электронная библиотечная система «IPRbooks»;
10. База данных SCOPUS и др.

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**
**Федеральное государственное автономное образовательное учреждение
высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО
ОРГАНИЗАЦИИ И ПРОВЕДЕНИЮ САМОСТОЯТЕЛЬНОЙ РАБОТЫ
ПО ДИСЦИПЛИНЕ**

Введение в большие данные

Направление подготовки 39.04.01 Социология
Направленность (профиль) «Большие данные и цифровая аналитика»
Форма обучения – Очная
Год начала обучения – 2026 г.

Ставрополь, 2026

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	
1. САМОСТОЯТЕЛЬНАЯ РАБОТА ПРИ ПОДГОТОВКЕ К АУДИТОРНЫМ ЗАНЯТИЯМ.....	
2. ФОРМЫ ВНЕАУДИТОРНОЙ САМОСТОЯТЕЛЬНОЙ РАБОТЫ.....	
2.1. Методические рекомендации по подготовке реферата.....	
2.2. Методические рекомендации по написанию эссе.....	
2.3. Методические рекомендации по подготовке доклада.....	
2.4. Методические рекомендации по подготовке к коллоквиуму.....	
2.5. Методические рекомендации по подготовке к собеседованию.....	
2.6. Методические рекомендации по подготовке конспекта.....	
2.7. Методические рекомендации по составлению глоссария.....	
2.8. Методические рекомендации по разработке мультимедийной презентации.....	
2.9. Методические рекомендации по выполнению кейс-задания.....	
2.10. Методические рекомендации по выполнению контрольной работы.....	
2.11. Методические рекомендации по подготовке к тестированию.....	
3. СИСТЕМА КРИТЕРИЕВ И ПОКАЗАТЕЛЕЙ ОЦЕНКИ ВЫПОЛНЕНИЯ ФОРМ САМОСТОЯТЕЛЬНОЙ РАБОТЫ.....	
3.1. Оценка формы самостоятельной работы «Написание реферата».....	
3.2. Оценка формы самостоятельной работы «написание эссе».....	
3.3. Оценка формы самостоятельной работы «доклад».....	
3.4. Оценка формы самостоятельной работы «Коллоквиум».....	
3.5. Оценка формы самостоятельной работы «Собеседование».....	
3.6. Оценка формы самостоятельной работы «Конспект».....	
3.7. Оценка формы самостоятельной работы «Составление глоссария».....	
3.8. Оценка формы самостоятельной работы «Мультимедийная презентация».....	
3.9. Оценка формы самостоятельной работы «Кейс-задание».....	
3.10. Оценка формы самостоятельной работы «Контрольная работа».....	
3.11. Оценка формы самостоятельной работы «Тестирование».....	
4. ФОРМА ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ И КРИТЕРИИ ОЦЕНКИ.....	
СПИСОК ЛИТЕРАТУРЫ.....	

ВВЕДЕНИЕ

Самостоятельная работа в современном образовательном процессе рассматривается как форма организации обучения, которая способна обеспечивать самостоятельный поиск необходимой информации, творческое восприятие и осмысление учебного материала в ходе аудиторных занятий, разнообразные формы познавательной деятельности студентов на занятиях и во внеаудиторное время, развитие аналитических способностей, навыков контроля и планирования учебного времени, выработку умений и навыков рациональной организации учебного труда.

Таким образом, самостоятельная работа – форма организации образовательного процесса, стимулирующая активность, самостоятельность, познавательный интерес студентов.

Самостоятельная работа студентов всех форм и видов обучения является одним из обязательных видов образовательной деятельности, обеспечивающей реализацию требований Федеральных государственных стандартов высшего профессионального образования (ФГОС), созданных на основе Федерального закона от 29.12.2012 г. № 273-ФЗ «Об образовании в Российской Федерации».

Согласно требованиям нормативных документов, самостоятельная работа студентов является обязательным компонентом образовательного процесса, так как она обеспечивает закрепление получаемых на лекционных занятиях знаний путем приобретения навыков осмысления и расширения их содержания, навыков решения актуальных проблем формирования общекультурных и профессиональных компетенций, научно-исследовательской деятельности, подготовки к семинарам, лабораторным работам, сдаче зачетов и экзаменов.

Самостоятельная работа студентов представляет собой совокупность аудиторных и внеаудиторных занятий и работ, обеспечивающих успешное освоение образовательной программы высшего профессионального образования в соответствии с требованиями ФГОС.

Целью данных указаний является формирование у студентов навыков самостоятельного изучения теоретического материала, анализа практических ситуаций и решения задач, связанных с планированием и развитием карьеры.

Самостоятельная работа в рамках образовательного процесса в вузе решает следующие задачи:

- закрепление и расширение знаний, умений, полученных студентами во время аудиторных и внеаудиторных занятий, превращение их в стереотипы умственной и физической деятельности;
- приобретение дополнительных знаний и навыков по дисциплинам учебного плана;
- формирование и развитие знаний и навыков, связанных с научно-исследовательской деятельностью;
- развитие ориентации и установки на качественное освоение

образовательной программы;

- развитие навыков самоорганизации;
- формирование самостоятельности мышления, способности к саморазвитию, самосовершенствованию и самореализации;
- выработка навыков эффективной самостоятельной профессиональной теоретической, практической и учебно-исследовательской деятельности.

Самостоятельная работа студента – это особым образом организованная деятельность, включающая в свою структуру такие компоненты, как:

- уяснение цели и поставленной учебной задачи;
- четкое и системное планирование самостоятельной работы;
- поиск необходимой учебной и научной информации;
- освоение собственной информации и ее логическая переработка;
- использование методов исследовательской, научно-исследовательской работы для решения поставленных задач;
- выработка собственной позиции по поводу полученной задачи;
- представление, обоснование и защита полученного решения;
- проведение самоанализа и самоконтроля.

Специфическими принципами организации самостоятельной работы в рамках современного образовательного процесса являются:

- принцип интерактивности обучения (обеспечение интерактивного диалога и обратной связи, которая позволяет осуществлять контроль и коррекцию действий студента);
- принцип развития интеллектуального потенциала студента (формирование алгоритмического, наглядно-образного, теоретического стилей мышления, умений принимать оптимальные или вариативные решения в сложной ситуации, умений обрабатывать информацию);
- принцип обеспечения целостности и непрерывности дидактического цикла обучения (предоставление возможности выполнения всех звеньев дидактического цикла в пределах темы, раздела, модуля).

Самостоятельная работа студентов классифицируется: по месту организации (аудиторная и внеаудиторная); по целям организации (цели дисциплины, сформулированные и обоснованные в рабочей программе); по способу организации (индивидуальная, групповая).

Выбор формы организации самостоятельной работы студентов (индивидуальная или групповая) определяется содержанием учебной дисциплины и формой организации обучения (лекция, семинар, практическое занятие, контрольное занятие и др.).

В зависимости от формы промежуточной аттестации виды самостоятельной работы дополняются подготовкой к экзамену, зачету и процедурами текущей аттестации.

1. САМОСТОЯТЕЛЬНАЯ РАБОТА ПРИ ПОДГОТОВКЕ К АУДИТОРНЫМ ЗАНЯТИЯМ

Основные виды аудиторных занятий в вузе – лекция, практическое занятие, семинар, семинар-конференция, коллоквиум, а в рамках контрольных мероприятий – контрольная работа, зачет, экзамен. Рассмотрим подробнее особенности самостоятельной подготовки к аудиторным занятиям данных видов.

Подготовка к лекции. Необходимость самостоятельной работы по подготовке к лекции определяется тем, что изучение любой дисциплины строится по определенной логике освоения ее разделов, представленных в рабочей программе дисциплины. Чаще всего логика изучения того или иного предмета заключается в движении от рассмотрения общих научных основ к анализу конкретных процессов и факторов, определяющих функционирование и изменение этого предмета.

Следует учесть, что преподаватели нередко представляют краткие конспекты своих лекций вместе с рабочей программой или имеют авторские учебники, пособия по преподаваемому предмету. Знакомство с этими материалами позволяет заранее ознакомиться с основными положениями предстоящей лекции и активно задавать конкретные вопросы при ее изложении.

Преподаватель при чтении новой лекции обычно указывает на связь ее содержания с тем, которое было прежде изучено. Качество освоения содержания конкретной дисциплины прямо зависит от того, насколько студент сам, без внешнего принуждения формирует у себя установку на получение на лекциях новых знаний, дополняющих уже имеющиеся по данной дисциплине.

Подготовка к практическому занятию. Подготовка к практическому занятию включает следующие элементы самостоятельной деятельности: четкое представление цели и задач его проведения; выделение навыков умственной, аналитической, научной деятельности, которые станут результатом предстоящей работы. Выработка навыков осуществляется с помощью получения новой информации об изучаемых процессах и с помощью знания о том, в какой степени в данное время студент владеет методами исследовательской деятельности, которыми он станет пользоваться на практическом занятии.

Следовательно, работа на практическом занятии направлена не только на познание студентом конкретных явлений внешнего мира, но и на изменение самого себя. Второй результат очень важен, поскольку он обеспечивает формирование таких общекультурных компетенций, как способность к самоорганизации и самообразованию, способность использовать методы сбора, обработки и интерпретации комплексной информации для решения организационно-управленческих задач, в том числе находящихся за пределами непосредственной сферы деятельности студента.

Подготовка к практическому занятию нередко требует подбора

материала, данных и специальных источников, с которыми предстоит учебная работа. Студенты должны дома подготовить к занятию 3-4 примера формулировки темы исследования, представленного в монографиях, научных статьях, отчетах. Затем они самостоятельно осуществляют поиск соответствующих источников, определяют актуальность конкретного исследования процессов и явлений, выделяют основные способы доказательства авторами научных работ ценности того, чем они занимаются.

В ходе самого практического занятия студенты сначала представляют найденные ими варианты формулировки актуальности исследования, обсуждают их и обосновывают свое мнение о наилучшем варианте.

Подготовка к семинарскому занятию. Семинарское занятие является традиционной и распространенной формой организации самостоятельной работы студентов при изучении гуманитарных дисциплин.

Самостоятельная подготовка к семинару направлена:

- на развитие способности к чтению научной и иной литературы;
- на поиск дополнительной информации, позволяющей глубже разобраться в некоторых вопросах;
- на выделение при работе с разными источниками необходимой информации, которая требуется для полного ответа на вопросы плана семинарского занятия;
- на выработку умения правильно выписывать высказывания авторов из имеющихся источников информации, оформлять их по библиографическим нормам;
- на развитие умения осуществлять анализ выбранных источников информации;
- на подготовку собственного выступления по обсуждаемым вопросам;
- на формирование навыка оперативного реагирования на разные мнения, которые могут возникать при обсуждении тех или иных научных проблем.

Подготовка к семинару-конференции. Семинар-конференция проводится 1-3 раза в семестр, так как предполагает достаточно длительную самостоятельную подготовку студентов, изучающих какую-либо конкретную научную проблему. При его проведении сочетаются виды деятельности, соответствующие обычному семинарскому занятию и научной конференции, которая предусматривает организованное обсуждение докладов разных исследователей по определенному кругу проблем.

В процессе самостоятельной подготовки к семинару-конференции студенту необходимо изучить 2-3 источника (монографии, статьи), в которых раскрыты теоретические подходы к обсуждаемому вопросу и представлены материалы эмпирических исследований. Выступающий должен быть готов ответить на вопросы всех присутствующих по теме своего доклада. После каждого выступления проводится обсуждение представленных научных воззрений разных исследователей. Готовность к такой аналитической коллективной работе обеспечивается просмотром каждым студентов тех

основных работ, которые преподаватель рекомендовал прочитать к семинару-конференции.

Подготовка к зачету (в том числе к дифференцированному при отсутствии экзамена по дисциплине). Зачет является традиционной формой проверки знаний, умений, компетенций, сформированных у студентов в процессе освоения всего содержания изучаемой дисциплины. Обычный зачет отличается от экзамена только тем, что преподаватель не дифференцирует баллы, которые он выставляет по его итогам. В случае проведения дифференцированного зачета студент получает баллы, отражающие уровень его знаний, но они не указываются в зачетной книжке: в нее вписывается только слово «зачет».

Самостоятельная подготовка к зачету должна осуществляться в течение всего семестра, а не за несколько дней до его проведения. Подготовка включает следующие действия. Прежде всего нужно перечитать все лекции, а также материалы, которые готовились к семинарским и практическим занятиям в течение семестра. Затем надо соотнести эту информацию с вопросами, которые даны к зачету. Если информации недостаточно, ответы находят в предложенной преподавателем литературе. Рекомендуется делать краткие записи. Речь идет не о шпаргалке, а о формировании в сознании четкой логической схемы ответа на вопрос. Накануне зачета необходимо повторить ответы, не заглядывая в записи.

Подготовка к экзамену. Экзамен представляет собой форму контроля учебной деятельности студента, которая используется, если учебная дисциплина составляет две и более зачетных единиц, т. е. изучается более 72 часов. Оценка выявленных на экзамене знаний, умений и компетенций дифференцирована: в зачетной книжке ставится оценка «удовлетворительно», «хорошо» или «отлично».

Самостоятельная подготовка к экзамену схожа с подготовкой к зачету, особенно если он дифференцированный. Но объем учебного материала, который нужно восстановить в памяти к экзамену, вновь осмыслить и понять, значительно больше, поэтому требуется больше времени и умственных усилий. Необходимо перечитать лекции, вспомнить то, что говорилось преподавателем на семинарах и практических занятиях, а также самостоятельно полученную информацию при подготовке к ним. Важно сформировать целостное представление о содержании ответа на каждый вопрос, что предполагает знание разных научных трактовок сущности того или иного явления, процесса, умение раскрывать факторы, определяющие их противоречивость, знание имен ученых, изучавших обсуждаемую проблему. Необходимо также привести информацию о материалах эмпирических исследований, что указывает на всестороннюю подготовку студента к экзамену. Ответ, в котором присутствуют все указанные блоки информации, наверняка будет отмечен высокими баллами. Для их получения требуется ответить и на дополнительные вопросы, если экзамен проходит в устной форме.

Рекомендуется подготовку к экзамену осуществлять в два этапа. На

первом, в течение 2-3 дней, подбирается из разных источников весь материал, необходимый для развернутых ответов на все вопросы. Ответы можно записать в виде краткого конспекта. На втором этапе по памяти восстанавливается содержание того, что записано в ответах на каждый вопрос.

2. ФОРМЫ ВНЕАУДИТОРНОЙ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Самостоятельная работа является неотъемлемой частью учебного процесса и играет ключевую роль в формировании профессиональных компетенций. Данный параграф призван предоставить студентам методические рекомендации и практические советы, направленные на эффективное выполнение различных форм самостоятельной работы, способствующие более глубокому усвоению учебного материала и развитию навыков самоорганизации.

2.1. Методические рекомендации по подготовке реферата

Цель самостоятельной работы: расширение научного кругозора, овладение методами теоретического исследования, развитие самостоятельности мышления студента.

Реферат (от лат. *refere* – докладывать, сообщать) – продукт самостоятельного творческого осмысления и преобразования текста первоисточника с целью получения новых сведений и существенных данных.

Виды рефератов:

- реферат-конспект, содержащий фактическую информацию в обобщенном виде, иллюстративный материал, различные сведения о методах исследования, результатах исследования и возможностях их применения;
- реферат-резюме, содержащий только основные положения данной темы;
- реферат-обзор, составляемый на основе нескольких источников, в котором сопоставляются различные точки зрения по данному вопросу;
- реферат-доклад, содержащий объективную оценку проблемы;
- реферат – фрагмент первоисточника, составляемый в тех случаях, когда в документе-первоисточнике можно выделить часть, раздел или фрагмент, отражающие информационную сущность документа или соответствующие задаче реферирования;
- обзорный реферат, составляемый на некоторое множество документов-первоисточников и являющийся сводной характеристикой определенного содержания документов.

Написание реферата является одной из форм обучения, направленной на организацию и повышение уровня самостоятельной работы обучающихся.

Целью написания рефератов является:

- привитие обучающимся навыков библиографического поиска необходимой литературы (на бумажных носителях, в электронном виде);
- привитие навыков компактного изложения мнения авторов и своего суждения по выбранному вопросу в письменной форме;

- приобретение навыка грамотного оформления ссылок на используемые источники, правильное цитирование авторского текста.

Основные задачи обучающегося при написании реферата:

- с максимальной полнотой использовать литературу по выбранной теме;

- верно (без искажения смысла) передать авторскую позицию в своей работе.

Требование к содержанию:

- материал, использованный в реферате, должен относиться строго к выбранной теме;

- необходимо изложить основные аспекты проблемные не только грамотно, но и в соответствии с той или иной логикой (хронологической, тематической, событийной и т.д.);

- реферат должен заканчиваться подведением итогов проведенной работы.

Структура реферата:

- титульный лист: название учебного заведения; тема реферата; инициалы исполнителя (обучающегося); инициалы преподавателя; дата сдачи работы; оценка и подпись преподавателя.

- оглавление: с указанием плана работы, который должен содержать введение, название основных разделов, заключение, список использованной литературы и нумерации страниц.

- введение: в котором определяется цель и задачи исследования; его актуальность; теоретическое и практическое значение; степень разработанности выбранной темы.

- основной текст: в котором раскрывается основное содержание плана;

- заключение: где формируются доказательные выводы на основании содержания исследуемого автором материала;

- список использованной литературы и других источников; в реферате могут быть использованы приложения (фотографии, схемы, таблицы, графики и т.д.).

- Приложение создается обучающимся в том случае, если оно дополняет содержание основных проблем темы.

- Объем и технические требования, предъявляемые к выполнению реферата:

- Текст работы должен быть напечатан на ПК на одной стороне белого листа бумаги формата А4 через 1,5 интервала, шрифтом Times New Roman, размер 14.

- Каждая страница текста и приложений должна иметь поля: левое – 30мм, правое – 10 мм, верхнее и нижнее по 20 мм.

- Заголовки отделяются от основного текста пробелами в 1,5 интервала снизу, шрифт Times New Roman, размер 14, полужирное начертание.

- Нумерация страниц производится последовательно, при этом номера страниц проставляются с 2-ой страницы вверху справа.

- Отступы всех абзацев должны быть по всей работе одинаковые и соответствовать 1,25 см.
- Объем реферата составляет 15 - 25 машинописных страниц.
- Подготовленная работа сдается в установленный срок преподавателем.

При невыполнении обучающимся требований к содержанию и оформлению реферата, преподаватель имеет право вернуть работу для доработки устранения недостатков.

2.2. Методические рекомендации по написанию эссе

Цель самостоятельной работы: развитие навыков самостоятельного творческого мышления и письменного изложения собственных мыслей.

Эссе – средство, позволяющее оценить умение обучающегося письменно излагать суть поставленной проблемы, самостоятельно проводить анализ этой проблемы с использованием концепций и аналитического инструментария соответствующей дисциплины, делать выводы, обобщающие авторскую позицию по поставленной проблеме.

Классификация эссе:

- по содержанию: философские, литературно-критические, исторические, художественные, художественно-публицистические, духовно-религиозные и др.;
- по литературной форме: рецензии, лирические миниатюры, заметки, странички из дневника, письма и др.;
- различают также эссе описательные, повествовательные, рефлексивные, критические, аналитические и др.

Признаки эссе:

- Небольшой объем – от трех до семи страниц компьютерного текста; допускается эссе до десяти страниц машинописного текста.
- Конкретная тема и подчеркнуто субъективная ее трактовка.
- Свободная композиция – важная особенность эссе.
- Непринужденность повествования.
- Использование парадоксов. Эссе призвано удивить читателя, это, по мнению многих исследователей, его обязательное качество.
- Внутреннее смысловое единство.
- Ориентация на разговорную речь. В то же время необходимо избегать употребления в эссе сленга, шаблонных фраз, сокращения слов, чересчур легкомысленного тона.

Эссе должно содержать четкое изложение сути поставленной проблемы, включать самостоятельно проведенный анализ этой проблемы с использованием концепций и аналитического инструментария соответствующей дисциплины, выводы, обобщающие авторскую позицию по поставленной проблеме. В зависимости от специфики дисциплины формы эссе могут значительно дифференцироваться. В некоторых случаях это может быть анализ собранных студентом конкретных данных по изучаемой проблеме, анализ материалов из средств массовой информации, подробный

разбор предложенной преподавателем проблемы с развёрнутыми пояснениями и анализом примеров, иллюстрирующих изучаемую проблему и т.д.

Структура определяется предъявляемыми требованиями:

- 1) мысли автора по проблеме излагаются в форме кратких тезисов (Т);
- 2) мысль должна быть подкреплена доказательствами - поэтому за тезисом следуют аргументы (А).

Аргументы – это факты, явления общественной жизни, события, жизненные ситуации и жизненный опыт, научные доказательства, ссылки на мнение ученых и др. Лучше приводить два аргумента в пользу каждого тезиса: один аргумент кажется неубедительным, три аргумента могут «перегрузить» изложение, выполненное в жанре, ориентированном на краткость и образность.

Таким образом, эссе приобретает кольцевую структуру (количество тезисов и аргументов зависит от темы, избранного плана, логики развития мысли):

- вступление;
- тезис, аргументы;
- тезис, аргументы;
- тезис, аргументы;
- заключение.

Организация написания эссе:

1. Вступительная часть (1 абзац). Первое предложение - тема, которая будет рассматриваться. Но тема не должна быть абстрактной или состоять из общих слов. Мнение должно быть обязательно сформулировано (а не дан просто факт). Начать можно с формулирования собственного мнения, с вопроса, с интересного факта, с цитаты, с анекдота или истории. Из вступления должно быть понятно, о чем будет идти речь в основной части. Затем - некоторые поддерживающие предложения, содержащие разъяснения по теме. Последнее предложение - основной тезис. Он задает специфику, центральную идею всего эссе, выражает мнение или позицию автора (не должен содержать факты).

2. Основная часть. Количество абзацев в основной части зависит от числа вопросов (позиций), которые планируется обсудить. Все абзацы должны поддерживать основной тезис, который сформулирован в конце вступления. Для понимания того, что писать в основной части, можно задать вопросы к основному тезису. Ответы на них и будут абзацами в эссе.

3. Заключительная часть. Дается основной аргумент (другими словами). Возвращение к основному тезису. Комментарии по теме. Лучше начинать со слов: «в заключении, подводя итоги...». Основной тезис передается, другими словами. Финальный комментарий по предмету того, что было написано в основной части. Не следует включать новую тему в заключительную часть.

При написании эссе важно также учитывать следующие моменты:

1. Вступление и заключение должны фокусировать внимание на проблеме (во вступлении она ставится, в заключении - резюмируется мнение

автора).

2. Необходимо выделение абзацев, красных строк, установление логической связи абзацев: так достигается целостность работы.

3. Стиль изложения: эмоциональность, экспрессивность, художественность. Специалисты полагают, что должный эффект обеспечивают короткие, простые, разнообразные по интонации предложения, умелое использование «самого современного» знака препинания - тире. Впрочем, стиль отражает особенности личности, об этом тоже полезно помнить.

4. Для передачи личностного восприятия проблемы можно привести многочисленные примеры, использовать всевозможные ассоциации.

5. Эссе будет выглядеть интереснее, если в нем присутствуют непредсказуемые выводы, неожиданные повороты, интересные сцепления.

2.3. Методические рекомендации по подготовке доклада

Цель самостоятельной работы: расширение научного кругозора, овладение методами теоретического исследования, развитие самостоятельности мышления студента.

Доклад – публичное сообщение или документ, которые содержат информацию и отражают суть вопроса или исследования применительно к данной ситуации.

Виды докладов:

1. Устный доклад – читается по итогам проделанной работы и является эффективным средством разъяснения ее результатов.

2. Письменный доклад:

– краткий (до 20 страниц) – резюмирует наиболее важную информацию, полученную в ходе исследования;

– подробный (до 60 страниц) – включает не только текстовую структуру с заголовками, но и диаграммы, таблицы, рисунки, фотографии, приложения, сноски, ссылки, гиперссылки.

Титульный лист является первой страницей доклада, служит источником информации, необходимой для обработки и поиска документа.

На титульном листе приводят следующие сведения:

1) наименование учебного заведения;

2) наименование специальности (направления подготовки) ;

3) наименование группы;

4) тема доклада;

5) фамилия и инициалы студента (слушателя);

6) должность, ученое звание, фамилия и инициалы руководителя доклада;

7) место и дата составления доклада.

Содержание: в нем последовательно указываются названия пунктов доклада, страницы, с которых начинается каждый пункт.

Введение: формулируется суть исследуемой проблемы, обосновывается выбор темы, определяются ее значимость и актуальность, указываются цель и

задачи доклада.

Основная часть: может содержать несколько разделов, доказательно раскрывающих исследуемый вопрос.

Заключение: подводятся итоги или делается обобщенный вывод, отражающий содержание доклада (опираясь на поставленные задачи).

Библиографический список: приводится в алфавитном порядке, должен содержать 5-7 источников не старше 5 лет. В перечне указывается автор, наименование, место издания, издательство, год выпуска и количество страниц.

Приложения: включают таблицы, иллюстрации, рекомендации и др.

Требования к оформлению письменного доклада

Страницы текста доклада, включенные приложения, таблицы и распечатки должны соответствовать формату А4.

Доклад должен быть представлен в печатном виде. Шрифт – Times New Roman, кегль 14, полуторный интервал, выравнивание по ширине, абзацный отступ – 1,25.

Размеры полей: левое – 3 см, правое – 1,5 см, верхнее – 2 см, нижнее – 2 см.

Объем доклада: не более 5-7 страниц.

Все страницы работы должны быть пронумерованы. Номер страницы ставится внизу страницы посередине. Нумерация страниц начинается со второй страницы (титульный лист не нумеруется).

Заголовки структурных элементов доклада и разделов основной части следует располагать по центру строки без точки в конце и печатать прописными буквами (в рукописном тексте - это – заглавные буквы), не подчеркивая.

2.4. Методические рекомендации по подготовке к коллоквиуму

Цель самостоятельной работы: формирование у студента навыков анализа теоретических проблем на основе самостоятельного изучения учебной и научной литературы.

Коллоквиум – это форма проверки знаний студентов в высших учебных заведениях, которая может включать в себя дискуссии, собеседования или краткие экзамены по определённой теме или предмету. На коллоквиумах обсуждаются отдельные части, разделы, темы, вопросы изучаемого курса, а также рефераты, проекты и иные работы обучающихся.

Виды коллоквиумов:

1. Текущее занятие. Студенты отвечают на заранее озвученные преподавателем вопросы.

2. Мини-экзамен. Формат во многом зависит от специальности: студенты технических направлений обычно рассматривают разные варианты решения задач, гуманитарии обсуждают теоретические вопросы, дискутируют по актуальным проблемам той или иной отрасли науки.

3. Собрание с защитой работ. Предусматривает презентацию доклада или другого проекта. Студент или группа учащихся должны обосновать

актуальность темы, указать практическую ценность работы, рассказать о ходе исследования и описать полученные результаты.

Коллоквиум представляет собой не только одну из форм текущего контроля, но и одну из активных форм учебных занятий, проводимых как в виде беседы преподавателя со студентами, так и в виде семинара, посвященного обсуждению определенной научной темы.

Целями коллоквиума являются: выяснение у студентов знаний, их углубление (повышение) и закрепление по той или иной теме курса; формирование у студентов навыков анализа теоретических проблем на основе самостоятельного изучения учебной и научной литературы.

Основная задача коллоквиума - пробудить у студента стремление к чтению и использованию дополнительной экономической литературы.

На коллоквиум могут выноситься как проблемные (нередко спорные теоретические вопросы), так и вопросы, требующие самостоятельного изучения, а также более глубокой проработки.

На самостоятельную подготовку к коллоквиуму студенту отводится 1-3 недели. Подготовка включает в себя изучение рекомендованной литературы и составление конспекта. Коллоквиуму может предшествовать написание эссе. Коллоквиум проводится либо в форме индивидуальной беседы преподавателя со студентом, либо беседы в небольших группах (3-5 человек).

Коллоквиум оценивается в форме «зачтено», «не зачтено».

2.5. Методические рекомендации по подготовке к собеседованию

Собеседование – одна из форм проверки и оценки усвоенных знаний, получения информации о характере познавательной деятельности, уровне развития самостоятельности и активности слушателей в процессе изучения дисциплины, владении навыками коммуникации и взаимодействия.

Подготовка слушателя к собеседованию предполагает следующий алгоритм:

- по какой теме, разделу будет собеседование;
- работа с информационными ресурсами и научными источниками;
- обобщение учебной информации на основе установления внутри и межпредметных связей, интеграции научного знания и практического опыта слушателей.

Подготовка к собеседованию представляет самостоятельную работу слушателя, основное содержание которой связано с навыками работы с информационными ресурсами (учебниками, научными текстами и др.).

При подготовке к занятию слушатель самостоятельно изучает информационные источники по теме практического занятия; анализирует собственный опыт по обсуждаемым вопросам.

2.6. Методические рекомендации по подготовке конспекта

Цель самостоятельной работы: выработка умений и навыков грамотного изложения теории и практических вопросов в письменной форме в виде конспекта.

Конспект (от лат. conspectus – обзор, изложение) – 1) письменный текст,

систематически, кратко, логично и связно передающий содержание основного источника информации (статьи, книги, лекции и др.); 2) синтезирующая форма записи, которая может включать в себя план источника информации, выписки из него и его тезисы.

Виды конспектов:

- плановый конспект (план-конспект) – конспект на основе сформированного плана, состоящего из определенного количества пунктов (с заголовками) и подпунктов, соответствующих определенным частям источника информации;
- текстуальный конспект – подробная форма изложения, основанная на выписках из текста-источника и его цитировании (с логическими связями);
- произвольный конспект – конспект, включающий несколько способов работы над материалом (выписки, цитирование, план и др.);
- схематический конспект (контекст-схема) – конспект на основе плана, составленного из пунктов в виде вопросов, на которые нужно дать ответ;
- тематический конспект – разработка и освещение в конспективной форме определенного вопроса, темы;
- опорный конспект (введен В. Ф. Шаталовым) – конспект, в котором содержание источника информации закодировано с помощью графических символов, рисунков, цифр, ключевых слов и др.;
- сводный конспект – обработка нескольких текстов с целью их сопоставления, сравнения и сведения к единой конструкции;
- выборочный конспект – выбор из текста информации на определенную тему.

Формы конспектирования:

- план (простой, сложный) – форма конспектирования, которая включает анализ структуры текста, обобщение, выделение логики развития событий и их сути;
- выписки – простейшая форма конспектирования, почти дословно воспроизводящая текст;
- тезисы – форма конспектирования, которая представляет собой выводы, сделанные на основе прочитанного. Выделяют простые и осложненные тезисы (кроме основных положений, включают также второстепенные);
- цитирование – дословная выписка, которая используется, когда передать мысль автора своими словами невозможно.

Выполнение Практическое задание:

- 1) определить цель составления конспекта;
- 2) записать название текста или его части;
- 3) записать выходные данные текста (автор, место и год издания);
- 4) выделить при первичном чтении основные смысловые части текста;
- 5) выделить основные положения текста;
- 6) выделить понятия, термины, которые требуют разъяснений;
- 7) последовательно и кратко изложить своими словами существенные положения изучаемого материала;
- 8) включить в запись выводы по основным положениям, конкретным

фактам и примерам (без подробного описания);

9) использовать приемы наглядного отражения содержания (абзацы «ступеньками», различные способы подчеркивания, ручки разного цвета);

10) соблюдать правила цитирования (цитата должна быть заключена в кавычки, дана ссылка на ее источник, указана страница).

Планируемые результаты самостоятельной работы:

- способность студентов анализировать результаты научных исследований и применять их при решении конкретных образовательных и исследовательских задач;

- способность решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности.

2.7. Методические рекомендации по составлению глоссария

Цель самостоятельной работы: повысить уровень информационный культуры; приобрести новые знания; отработать необходимые навыки в предметной области учебного курса.

Глоссарий – словарь специализированных терминов и их определений, употребляемых в изучаемой дисциплине или разделе.

Статья глоссария – определение термина.

Содержание Практическое задание: сбор и систематизация понятий или терминов, объединенных общей специфической тематикой, по одному либо нескольким источникам.

Для составления глоссария по заданной теме нужно найти информацию из разных источников (сеть Internet, энциклопедии, практические пособия, учебная литература, литература по специальности), изучить ее и систематизировать.

Глоссарий составляется индивидуально.

Работа должна быть представлена на бумаге формата А4 в печатном (компьютерном) варианте.

Выполненную работу сдать к указанному сроку.

Общие требования:

Глоссарий состоит из слов, соответствующих тематике задания. Используемые слова должны быть именами существительными в именительном падеже единственного числа.

Допускается использование иностранных слов, если они подходят по теме.

Не допускаются аббревиатуры, сокращения. Все тексты должны быть отпечатаны.

Структура глоссария:

Объем работы: 5-6 листов, нумерация страниц - снизу, справа; 1 лист - титульный; 2-5 лист - толковый словарь терминов; 6 лист - список используемой литературы.

Они должны быть строго лаконичными. Не следует пространными,

делать их несущими исчерпывающими, многословными, излишне избыточную информацию.

Старайтесь подать слово с наименее известной стороны.

Просмотрите словари: возможно, в одном из них и окажется наилучшее определение. В определениях не должно быть однокоренных слов.

Планирование деятельности по составлению глоссария:

- 1) определить, с какой целью составляется глоссарий;
- 2) просмотреть и изучить материал по теме в учебнике;
- 3) продумать составные части глоссария;
- 4) изучить дополнительный материал по теме;
- 5) составить список слов;
- 6) подобрать толкование слов;
- 7) проверить орфографию текста, соответствие нумерации;
- 8) проанализировать составленный глоссарий согласно критериям оценивания;
- 9) оформить готовый глоссарий;
- 10) продумать защиту проекта-глоссария.

Алгоритм действий по составлению глоссария:

1. Для начала необходимо создать таблицу из трёх столбцов.

№ п/п	Термин	Определение

2. Далее следует составить список наиболее часто встречающихся профессиональных терминов и занести их в колонку таблицы с названием «Термин».

3. После этого необходимо найти определение термина и занести его в колонку с названием «Определение» напротив соответствующего термина.

4. Далее все слова в списке необходимо расположить в алфавитном порядке. Для этого нужно выделить 2-ю и 3-ю колонки таблицы (за исключением «шапки» таблицы) и нажать значок сортировки (рисунок 1).

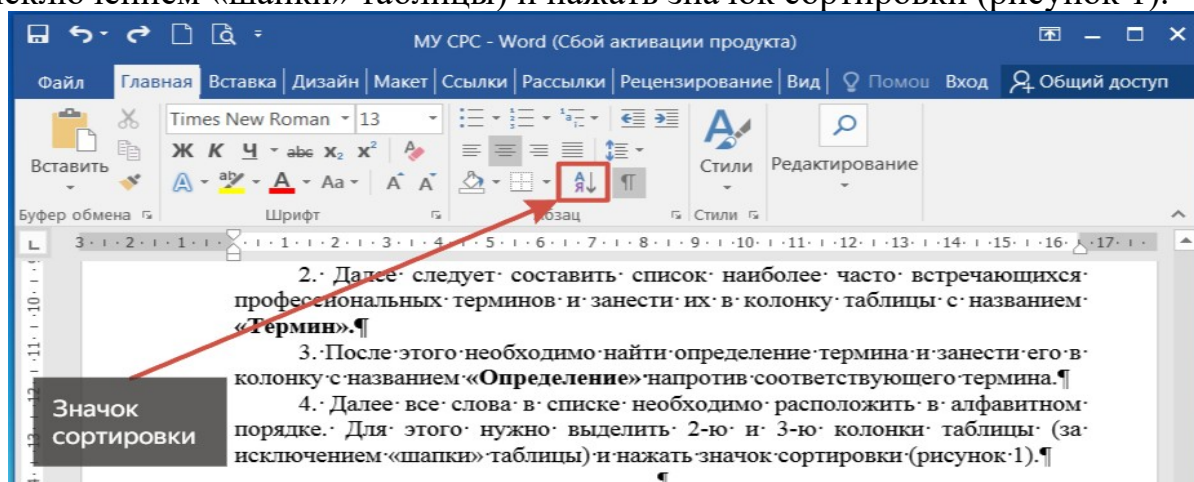


Рисунок 1 – Значок сортировки на панели инструментов
В диалоговом окне Сортировка текста (рисунок 2) в группе С начала по

выберите пункт столбцам 2 и тип Текст, а затем установите переключатель в положение по возрастанию.

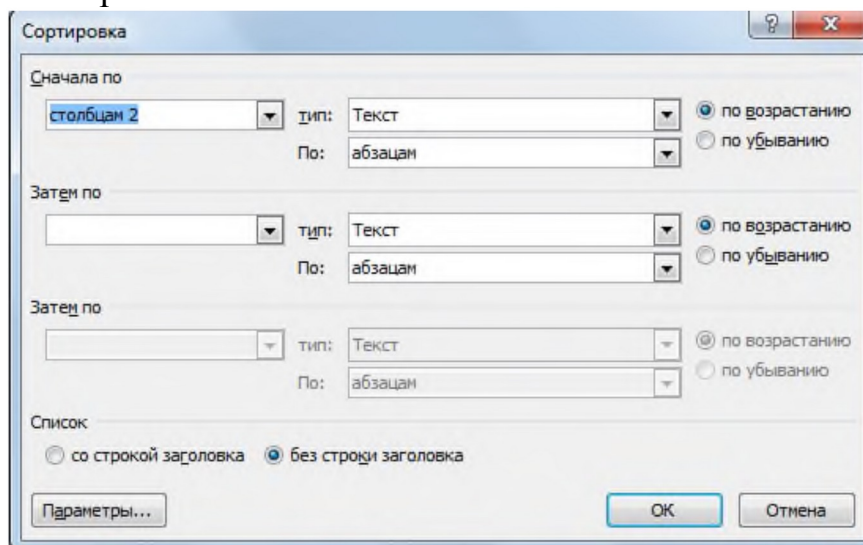


Рисунок 2 - Диалоговое окно сортировки текста

5. В завершении необходимо пронумеровать все термины.

ВНИМАНИЕ! Глоссарий оформляется на листе А4, шрифт Times New Roman, 14 пт, междустрочный интервал – 1,0.

При составлении глоссария важно придерживаться следующих правил:
отобранные термины и лексические единицы должны относиться к широкому и узкому профилю специальности;

отобранные термины и лексические единицы должны быть новыми и не дублировать ранее изученные;

отобранные термины и лексические единицы предназначены для активного усвоения и должны использоваться при чтении, восприятии на слух, говорении и письме.

Отобранные профессиональные термины необходимо выучить наизусть и сдать на занятии в устной форме. Также необходимо предоставить печатный вариант глоссария на листах формата А4 в папке-скоросшивателе.

2.8. Методические рекомендации по разработке мультимедийной презентации

Цели самостоятельной работы (варианты):

- освоение (закрепление, обобщение, систематизация) учебного материала;
- обеспечение контроля качества знаний;
- формирование специальных компетенций, обеспечивающих возможность работы с информационными технологиями;
- становление общекультурных компетенций. Мультимедийная презентация – представление содержания учебного материала, учебной задачи с использованием мультимедийных технологий.

Основные виды мультимедийной презентации:

- обучающие и тестовые презентации (позволяют знакомить с содержанием учебного материала и контролировать качество его усвоения);

- презентации электронных каталогов (дают возможность распространять большие объемы информации быстро, качественно и эффективно);

- электронные презентации и рекламные ролики (служат для создания имиджа и распространение информации об объекте);

- презентации – визитные карточки (дают представление об авторе работы);

- бытовые презентации (использование в бытовых целях фотографий и видеоизображений в электронном виде).

Мультимедийные презентации по назначению:

- презентация сопровождения образовательного процесса (является источником информации и средством привлечения внимания слушателей);

- презентация учебного или научно-исследовательского проекта (используется для привлечения внимания слушателей к основной идее или концепции развития проекта с точки зрения его возможной эффективности и результативности применения);

- презентация информационной поддержки образовательного процесса (представляет собой обновление банка литературы, контрольных и тестовых заданий, вопросов к итоговой и промежуточной аттестации);

- презентация-отчет (мультимедийное сопровождение отчета в виде нескольких фрагментов, логически связанных между собой в зависимости от структуры отчета).

Выполнение Практическое задание:

1. Этап проектирования:

- определение целей использования презентации;

- сбор необходимого материала (тексты, рисунки, схемы и др.);

- формирование структуры и логики подачи материала;

- создание папки, в которую помещен собранный материал.

2. Этап конструирования:

- выбор программы MS Power Point в меню компьютера;

- определение дизайна слайдов;

- наполнение слайдов собранной текстовой и наглядной информацией;

- включение эффектов анимации и музыкального сопровождения (при необходимости);

- установка режима показа слайдов (титульный слайд, включающий наименование кафедры, где выполнена работа, название презентации, город и год; содержательный – список слайдов презентации, сгруппированных по темам сообщения; заключительный слайд содержит выводы, пожелания, список литературы и пр.).

3. Этап моделирования – проверка и коррекция подготовленного материала, определение продолжительности его демонстрации.

Планируемые результаты самостоятельной работы:

- повышение информационной культуры студентов и обеспечение их готовности к интеграции в современное информационное пространство;

- способность решать стандартные задачи профессиональной

деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности;

- способность к критическому восприятию, обобщению, анализу профессиональной информации, постановке цели и выбору путей ее достижения;

- способность применять современные методики и технологии организации и реализации образовательного процесса на различных образовательных ступенях в различных образовательных учреждениях;

- готовность использовать индивидуальные креативные способности для оригинального решения исследовательских задач.

2.9. Методические рекомендации по выполнению кейс-задания

Цель самостоятельной работы: формирование умения анализировать в короткие сроки большой объем неупорядоченной информации, принятие решений в условиях недостаточной информации.

Кейс-задание (англ. case – случай, ситуация) – метод обучения, основанный на разборе практических проблемных ситуаций – кейсов, связанных с конкретным событием или последовательностью событий.

Виды кейсов: иллюстративные, аналитические, связанные с принятием решений.

Кейс – технологии (Case study) – технологии, основанные на комплектовании наборов (кейсов) текстовых учебно-методических материалов по какой-то выделенной теме и заданий по конкретной проблемной ситуации в ней, и передачи их обучающимся для самостоятельного изучения (с возможностью консультации у преподавателя) и решения задания с последующим коллективным обсуждением темы и вариантов для выработки наиболее рациональных и творческих предложений.

Анализ конкретных учебных ситуаций (case study) – метод обучения, предназначенный для совершенствования навыков и получения опыта в следующих областях: выявление, отбор и решение проблем; работа с информацией – осмысление значения деталей, описанных в ситуации; анализ и синтез информации и аргументов; работа с предположениями и заключениями; оценка альтернатив; принятие решений; слушание и понимание других людей – навыки групповой работы.

Цель метода case-study – научить обучающихся анализировать проблемную ситуацию - кейс, возникшую при конкретном положении дел, и выработать решение; научить работать с информационными источниками, перерабатывать ее из одной формы в другую. Итогом будет являться оценка предложенных алгоритмов и выбор лучшего в контексте поставленной проблемы.

Задачи:

Формирование и развитие информационной компетентности;

Развитие умений искать новые знания, анализировать ситуации;

Развитие навыков самоорганизации, самостоятельности,

инициативности;

Развитие умений принимать решения, аргументировать свою позицию;
Развитие умений и навыков сотрудничества.

Кейсы - учебные конкретные ситуации, специально разработанные на основе фактического материала с целью последующего разбора на уроке. В ходе разбора ситуаций обучающиеся учатся действовать в «команде», проводить анализ и принимать решения. Акцент обучения переносится на выработку готового знания самими учащимися, что очень актуально в связи с введением Федеральных государственных образовательных стандартов.

Технология работы с кейсом в учебном процессе сравнительно проста и включает в себя следующие этапы:

индивидуальная самостоятельная работы обучаемых с материалами кейса;

работа в малых группах по согласованию видения ключевой проблемы и ее решений;

презентация и экспертиза результатов малых групп на общей дискуссии.

Выполнение Практическое задание:

1) подготовить основной текст с вопросами для обсуждения:

– титульный лист с кратким запоминающимся названием кейса;

– введение, где упоминается герой (герои) кейса, рассказывается об истории вопроса, указывается время начала действия;

– основная часть, где содержится главный массив информации, внутренняя интрига, проблема;

– заключение (в нем решение проблемы, рассматриваемой в кейсе, иногда может быть не завершено);

2) подобрать приложения с подборкой различной информации, передающей общий контекст кейса (документы, публикации, фото, видео и др.);

3) предложить возможное решение проблемы. Планируемые результаты самостоятельной работы:

– способность студентов анализировать результаты научных исследований и применять их при решении конкретных исследовательских задач;

– готовность использовать индивидуальные креативные способности для оригинального решения исследовательских задач;

– способность решать нестандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий.

2.10. Методические рекомендации по выполнению контрольной работы

Цель самостоятельной работы: определить качество усвоения лекционного материала и части дисциплины, предназначенной для самостоятельного изучения.

Контрольная работа – письменная работа небольшого объема, предполагающая проверку знаний заданного к изучению материала и навыков

его практического применения. Контрольные работы могут состоять из одного или нескольких теоретических вопросов. Задание контрольной работы может быть сформулировано и в качестве одной или нескольких задач.

Задачи контрольной работы:

закрепить полученные ранее теоретические знания;

выработать навыки самостоятельной работы;

выяснить подготовленность студента к будущей практической работе.

В процессе подготовки и написания контрольной работы закрепляются, углубляются и обобщаются теоретические знания, полученные при изучении дисциплины. Появляется возможность развить творческие способности, приобрести умение обобщать отечественный и зарубежный опыт, анализировать законодательные, нормативные, научные, литературные и другие источники информации, выделять в них главное, давать критическую оценку материалам, выявлять имеющиеся проблемы, аргументировано обосновывать выводы и разрабатывать рекомендации по решению проблем.

Написание контрольной работы практикуется в учебном процессе в целях приобретения студентом необходимой профессиональной подготовки, развития умения и навыков самостоятельного научного поиска: изучения литературы по выбранной теме, анализа различных источников и точек зрения, обобщения материала, выделения главного, формулирования выводов и т. п. С помощью контрольной работы студент постигает наиболее сложные проблемы курса, учится лаконично излагать свои мысли, правильно оформлять работу.

Выполнение контрольной работы следует начинать с общего ознакомления с темой (прочтение соответствующего раздела учебника, учебного пособия, конспектов лекций).

Затем необходимо изучить другие литературные источники, рекомендованные преподавателем.

Наиболее распространенная проблема – неумение работать с библиотечными фондами. Поэтому вы должны в максимально сжатые сроки научиться самостоятельно подбирать литературу – это залог вашей успешной учебы. В каждой библиотеке, в зале каталогов, находится консультант, который всегда поможет вам сориентироваться в библиотечных фондах и правильно оформить заказ на книгу. Работая в библиотеке, учитывайте следующие факторы:

если вам необходимо подобрать литературу по конкретной теме, но вы не знаете авторов книг (монографий), используйте предметный каталог;

если вам известен автор или название книги, вы можете воспользоваться алфавитным каталогом;

в каждой библиотеке дополнительно существует картотека журнальных статей;

в настоящее время библиотеки предоставляют услуги для составления списка литературы по требуемой тематике, однако откажитесь от соблазна воспользоваться этим, пока не освоите методику поиска литературы самостоятельно;

не забывайте про ресурсы интернета.

2.11. Методические рекомендации по подготовке к тестированию

Тестирование представляет собой средство контроля усвоения учебного материала темы или раздела дисциплины.

Целью тестирования является формирование у слушателей навыков анализа теоретических проблем на основе самостоятельного изучения учебной и научной литературы.

От слушателя требуется: владение, изученным в ходе учебного процесса, материалом; знание разных точек зрения, высказанных в научной и учебной литературе по соответствующей проблеме, умение сопоставлять их между собой; наличие собственного мнения по обсуждаемым вопросам и умение его аргументировать.

Тестирование – это не только форма контроля, но и метод углубления, закрепления знаний слушателей. При подготовке к тесту не следует просто заучивать, необходимо понять логику изложенного материала. Этому немало способствует составление развернутого плана, таблиц, схем, внимательное изучение сущности вынесенных на тестирование вопросов.

Как и любая другая форма подготовки к контролю знаний, тестирование имеет ряд особенностей, знание которых помогает успешно выполнить тест:

Прежде всего, следует внимательно изучить структуру теста, оценить объем времени, выделяемого на данный тест, увидеть, какого типа задания в нем содержатся. Это поможет настроиться на работу.

Лучше начинать отвечать на те вопросы, в правильности решения которых нет сомнений, пока, не останавливаясь на тех, которые могут вызвать долгие раздумья. Это позволит сосредоточиться на выполнении более трудных вопросов.

Важно всегда внимательно читать задания до конца, не пытаясь понять условия «, по первым словам,» или выполнив подобные задания в предыдущих тестированиях.

При подготовке к тестированию слушатель самостоятельно изучает информационные источники по темам дисциплины, содержание которых отражено в тестовых заданиях; обращается за консультацией к преподавателю.

3. СИСТЕМА КРИТЕРИЕВ И ПОКАЗАТЕЛЕЙ ОЦЕНКИ ВЫПОЛНЕНИЯ ФОРМ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Порядок и особенности контрольно-оценочной деятельности

Одним из способов проверки качества организации самостоятельной работы студентов является к о н т р о л ь:

- корректирующий (осуществляется во время индивидуальных консультаций по поводу выполнения формы самостоятельной работы);
- констатирующий (по результатам выполнения специальных форм самостоятельной работы);

- самоконтроль (осуществляется самим студентом);
- текущий (ход выполнения контрольных аудиторных и прочих форм самостоятельной работы, установленных рабочей программой);
- промежуточный (оценка результата обучения как итога выполнения студентом всех форм самостоятельной работы).

Для каждого вида контроля преподаватель выбирает соответствующее оценочное средство, которым в нашем случае является форма самостоятельной работы (табл. 1).

Таблица 1 – Примерный перечень и краткая характеристика оценочных средств – форм самостоятельной работы

№ п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в рабочих программах
1	Реферат	Продукт самостоятельной работы студента, представляющий собой краткое изложение в письменном виде полученных результатов теоретического анализа определенной научной (учебно-исследовательской) темы, где автор раскрывает суть исследуемой проблемы, приводит различные точки зрения, а также собственный взгляд на нее	Темы рефератов
2	Эссе	Средство, позволяющее оценить умение обучающегося письменно излагать суть поставленной проблемы, самостоятельно проводить анализ этой проблемы с использованием концепций и аналитического инструментария соответствующей дисциплины, делать выводы, обобщающие авторскую позицию по поставленной проблеме	Тематика эссе
3	Доклад	Продукт самостоятельной работы студента, публичное выступление по представлению полученных результатов решения определенной учебно-практической, учебно-исследовательской или научной темы	Темы докладов
4	Коллоквиум	Средство контроля усвоения учебного материала темы, раздела или разделов дисциплины, организованное как учебное занятие в виде собеседования преподавателя с обучающимися.	Вопросы по темам/разделам дисциплины
5	Собеседование	Средство контроля, организованное как специальная беседа преподавателя с обучающимися на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу,	Вопросы по темам/разделам дисциплины

№ п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в рабочих программах
		теме, проблеме и т.п.	
6	Конспект	Оценочное средство, которое позволяет оценить умение обучающегося излагать материал в соответствии с планом лекции, выделять главное и второстепенное, устанавливать логическую связь между элементами темы, выделять ключевые слова и понятия	Конспект лекций
7	Глоссарий	Словарь терминов	Тематика глоссария
8	Мультимедийная презентация	Представление содержания учебного материала, учебной задачи с использованием мультимедийных технологий	Тематика презентаций
9	Кейс-задание	Проблемное задание, в котором студенту предлагается осмыслить реальную профессионально-ориентированную ситуацию, необходимую для решения данной проблемы	Задания для решения кейс-задания
10	Контрольная работа	Средство проверки умений применять полученные знания для решения задач определенного типа по теме или разделу	Комплект контрольных заданий по вариантам
11	Тест	Система стандартизированных заданий, позволяющая автоматизировать процедуру измерения уровня знаний и умений обучающегося.	Фонд тестовых заданий

3.1. Оценка формы самостоятельной работы «Написание реферата»

оценка «отлично» выставляется, если содержание реферата соответствует заявленной в названии тематике; реферат оформлен в соответствии с общими требованиями написания и техническими требованиями оформления реферата; реферат имеет чёткую композицию и структуру; в тексте реферата отсутствуют логические нарушения в представлении материала; корректно оформлены и в полном объёме представлены список использованной литературы и ссылки на использованную литературу в тексте реферата; отсутствуют орфографические, пунктуационные, грамматические, лексические, стилистические и иные ошибки в авторском тексте; реферат представляет собой самостоятельное исследование, представлен качественный анализ найденного материала, отсутствуют факты плагиата;

оценка «хорошо» выставляется, если содержание реферата соответствует заявленной в названии тематике; реферат оформлен в соответствии с общими требованиями написания реферата, но есть

погрешности в техническом оформлении; реферат имеет чёткую композицию и структуру; в тексте реферата отсутствуют логические нарушения в представлении материала; в полном объёме представлены список использованной литературы, но есть ошибки в оформлении; корректно оформлены и в полном объёме представлены ссылки на использованную литературу в тексте реферата; отсутствуют орфографические, пунктуационные, грамматические, лексические, стилистические и иные ошибки в авторском тексте; реферат представляет собой самостоятельное исследование, представлен качественный анализ найденного материала, отсутствуют факты плагиата;

оценка «удовлетворительно» выставляется, если содержание реферата соответствует заявленной в названии тематике; в целом реферат оформлен в соответствии с общими требованиями написания реферата, но есть погрешности в техническом оформлении; в целом реферат имеет чёткую композицию и структуру, но в тексте реферата есть логические нарушения в представлении материала; в полном объёме представлен список использованной литературы, лексические, стилистические и иные ошибки в авторском тексте; в целом реферат представляет собой самостоятельное исследование, представлен анализ найденного материала, отсутствуют факты плагиата;

оценка «неудовлетворительно» выставляется, если содержание реферата не соответствует заявленной в названии тематике или в реферате отмечены нарушения общих требований написания реферата; есть ошибки в техническом оформлении; есть нарушения композиции и структуры; в тексте реферата есть логические нарушения в представлении материала; не в полном объёме представлен список использованной литературы, есть ошибки в его оформлении; отсутствуют или некорректно оформлены и не в полном объёме представлены ссылки на использованную литературу в тексте реферата; есть многочисленные орфографические, пунктуационные, грамматические, лексические, стилистические и иные ошибки в авторском тексте; реферат не представляет собой самостоятельного исследования, отсутствует анализ найденного материала, текст реферата представляет собой не переработанный текст другого автора (других авторов)

3.2. Оценка формы самостоятельной работы «написание эссе»

Оценка «отлично» ставится за:

- глубоко и аргументированно раскрывающее тему, свидетельствующее об отличном знании текста произведения и других материалов, необходимых для её раскрытия, об умении целенаправленно анализировать материал, делать выводы и обобщения;
- стройное по композиции, логичное и последовательное в изложении мыслей;
- написанное правильным литературным языком и стилистически соответствующее содержанию.

Допускается незначительная неточность в содержании, один-два

речевых недочёта.

Оценка «хорошо» ставится за сочинение / эссе:

- достаточно полно и убедительно раскрывающее тему, обнаруживающее хорошее знание литературного материала и других источников по теме сочинения и умение пользоваться ими для обоснования своих мыслей, а также делать выводы и обобщения;
- логичное и последовательное изложение содержания;
- написанное правильным литературным языком, стилистически соответствующее содержанию.

Допускаются две-три неточности в содержании, незначительные отклонения от темы, а также не более трёх-четырёх речевых недочётов.

Оценка «удовлетворительно» ставится, в котором:

- в главном и основном раскрывающем тему, в целом дан верный, но односторонний или недостаточно полный ответ на тему, допущены отклонения от неё или отдельные ошибки в изложении фактического материала; обнаруживается недостаточное умение делать выводы и обобщения;
- материал излагается достаточно логично, но имеются отдельные нарушения в последовательности выражения мыслей;
- обнаруживается владение основами письменной речи;
- имеется не более четырёх недочётов в содержании и пяти речевых недочётов.

Оценка «неудовлетворительно» ставится за:

- не раскрывающее тему, не соответствующее плану, что свидетельствует о поверхностном знании либо полном непонимании текста произведения и неумении излагать свои мысли; состоит из путаного пересказа отдельных событий, без выводов и обобщений, или из общих положений, не опирающихся на текст;
- характеризуется случайным расположением материала, отсутствием связи между частями;
- отличается бедностью словаря, наличием грубых речевых ошибок.

3.3. Оценка формы самостоятельной работы «доклад»

	Не удовлетвори- тельно	Удовлетвори- тельно	Хорошо	Отлично
Степень раскрытия темы:	тема доклада не раскрыта: 4 и более замечаний.	тема доклада раскрыта частично: не более 3 замечаний;	тема доклада раскрыта частично: не более 2 замечаний;	тема доклада раскрыта;
Объем использованной научной литературы:	объем научной литературы недостаточный: 4 и более	объем научной литературы недостаточный: 3 замечания;	объем научной литературы недостаточный: не более 2	объем научной литературы достаточный;

	замечаний.		замечаний;	
Достоверность информации в докладе (точность, обоснованность, наличие ссылок на источники первичной информации):	информация имеет замечания по всем требованиям.	информация имеет замечания по двум требованиям из трех;	информация имеет замечания по одному требованию из трех;	информация точная, обоснованная, есть ссылки на источники первичной информации;
Необходимость и достаточность информации:	приведенные данные и факты не служат целям обоснования или иллюстрации определенных тезисов и положений доклада.	приведенные данные и факты служат целям обоснования или иллюстрации определенных тезисов и положений доклада частично: 3 и более замечаний;	приведенные данные и факты служат целям обоснования или иллюстрации определенных тезисов и положений доклада частично: не более 2 замечаний;	приведенные данные и факты служат целям обоснования или иллюстрации определенных тезисов и положений доклада;

3.4. Оценка формы самостоятельной работы «Коллоквиум»

оценка «зачтено» выставляется студенту, если он принимает активное участие в обсуждении вопросов, вынесенных на семинарское занятие (коллоквиум), при обсуждении опирается на литературу по теме коллоквиума, делает отсылки к авторам, приводит примеры, высказывает собственную позицию, аргументируя ее, хорошо владеет теоретическим и практическим материалом по обсуждаемой теме.

оценка «не зачтено» выставляется студенту, если он не принимает активного участия в обсуждении вопросов, вынесенных на коллоквиум, при обсуждении опирается только на собственные суждения, не используя литературу по теме коллоквиума, затрудняется с отсылками к авторам, с трудом приводит примеры (или не может их привести), высказывает собственную позицию, не аргументируя ее, плохо владеет теоретическим и практическим материалом по обсуждаемой теме.

3.5. Оценка формы самостоятельной работы «Собеседование»

Отметка «отлично» ставится, если теоретическое содержание учебного материала освоено обучающимся в полном объеме, без пробелов, необходимые практические навыки в основном сформированы, однако они могут быть недостаточными; перевод текста и задания к нему выполнены, хотя некоторые ответы могут содержать лишь незначительные ошибки.

Отметка «хорошо» ставится, если теоретическое содержание учебного материала освоено обучающимся в полном объеме, однако в процессе ответа наблюдаются ошибки, в ходе выполнения практических заданий имеются

незначительные грамматические погрешности, но в целом практические навыки сформированы; перевод текста и задания к нему выполнены, хотя некоторые ответы могут содержать лишь незначительные ошибки.

Отметка «удовлетворительно» ставится, если теоретическое содержание материала освоено частично, необходимые практические навыки работы с текстом не сформированы, большинство заданий не выполнено, либо качество их выполнения очень низкое.

Отметка «неудовлетворительно» ставится, если большинство заданий не выполнено, при дополнительной самостоятельной работе над материалом курса возможно повышение качества выполнения учебных заданий.

3.6. Оценка формы самостоятельной работы «Конспект»

	Не удовлетвори- тельно	Удовлетвори- тельно	Хорошо	Отлично
Соответствие конспекта плану содержания источника:	конспект не соответствует плану содержания.	конспект частично соответствует плану содержания: 3 и более замечаний;	конспект частично соответствует плану содержания: не более 2 замечаний;	конспект соответствует плану содержания;
Отражение в конспекте основных положений источника и наличие выводов:	основные положения не отражены, выводы не представлены.	основные положения отражены частично, выводы частично представлены;	основные положения отражены, выводы не представлены;	основные положения отражены, выводы представлены;
Ясность, лаконичность изложения:	по указанным параметрам изложение имеет 4 и более замечаний.	изложение имеет не более 3 замечаний по указанным параметрам;	изложение имеет не более 2 замечаний по указанным параметрам;	изложение ясное и лаконичное;

3.7. Оценка формы самостоятельной работы «Составление глоссария»

Оценка «отлично» ставится если выполнены все требования к оформлению глоссария: слова-термины соответствуют тематике, толкования лаконичны, точны, исчерпывающи, выдержаны все требования к его оформлению, на дополнительные вопросы при защите даны полные ответы.

Оценка «хорошо» ставится если основные требования выполнены, но при этом допущены недочёты. В частности, имеются неточности в изложении материала; не выдержан объём; имеются упущения в оформлении; на дополнительные вопросы при защите даны неполные ответы.

Оценка «удовлетворительно» ставится если имеются существенные отступления от требований к составлению глоссария. В частности, объём терминов и понятий недостаточен, допущены фактические ошибки в

толковании слов или при ответе на дополнительные вопросы; во время защиты отсутствует вывод.

Оценка «неудовлетворительно» ставится если обучающийся не представил глоссарий в печатном виде или представил термины, не относящиеся к выбранной специальности. Не готов к устному опросу.

3.8. Оценка формы самостоятельной работы «Мультимедийная презентация»

	Не удовлетвори- тельно	Удовлетвори- тельно	Хорошо	Отлично
Раскрытие темы учебной дисциплины:	тема не раскрыта: 4 и более замечаний.	тема раскрыта частично: не более 3 замечаний;	тема раскрыта частично: не более 2 замечаний;	тема раскрыта;
Подача материала (наличие, достаточность и обоснованность графического оформления: схем, рисунков, диаграмм, фотографий):	подача материала соответствует указанным параметрам частично: 4 и более замечаний.	подача материала соответствует указанным параметрам частично: не более 3 замечаний;	подача материала соответствует указанным параметрам частично: не более 2 замечаний;	подача материала полностью соответствует указанным параметрам;
Оформление презентации (соответствие дизайна всей презентации поставленной цели; единство стиля включаемых в презентацию рисунков; обоснованное использование анимационных эффектов):	презентация оформлена с замечаниями по параметру или параметрам: 4 и более замечаний.	презентация оформлена с замечаниями по параметру или параметрам: не более 3 замечаний;	презентация оформлена с замечаниями по параметру или параметрам: не более 2 замечаний;	презентация оформлена без замечаний;

3.9. Оценка формы самостоятельной работы «Кейс-задание»

	Не удовлетвори- тельно	Удовлетвори- тельно	Хорошо	Отлично
Структура кейса:	структура не соответствует требованиям.	структура соответствует требованиям частично:	структура соответствует требованиям частично:	структура соответствует требованиям;

		отсутствуют 2-3 элемента;	отсутствует 1 элемент;	
Общая валидность – установление корректности операциональных действий в рамках изучаемой темы и выстраиваемой концепции:	операциональные действия не корректны.	операциональные действия корректны частично: 3-4 замечания;	операциональные действия корректны частично: 1-2 замечания;	операциональные действия корректны;
Внутренняя валидность – установление всей полноты причинных связей:	полнота причинных связей не установлена.	полнота причинных связей установлена частично: 3 и более замечаний;	полнота причинных связей установлена частично: 1-2 замечания;	установлена вся полнота причинных связей;

3.10. Оценка формы самостоятельной работы «Контрольная работа»

Оценка «отлично» выставляется, если обучающийся выполнил контрольную работу в установленный срок; при выполнении упражнений показал высокий уровень знания лексико-грамматического и страноведческого материала по заданной тематике, проявил творческий подход при ответе на вопросы, умение глубоко анализировать проблему и делать обобщающие выводы; выполнил работу грамотно с точки зрения поставленной задачи, т.е. без ошибок и недочетов или допустил не более одного недочета.

Оценка «хорошо» выставляется, если обучающийся выполнил контрольную работу в установленный срок; при выполнении упражнений показал хороший уровень знания лексико-грамматического и страноведческого материала по заданной тематике, практически правильно сформулировал ответы на поставленные вопросы, представил общее знание информации по проблеме; выполнил работу полностью, но допустил в ней: а) не более одной негрубой ошибки и одного недочета б) или не более двух недочетов.

Оценка «удовлетворительно» выставляется, если обучающийся выполнил контрольную работу в установленный срок; показал достаточные знания по основным темам контрольной работы; выполнил не менее половины работы или допустил в ней: а) не более двух грубых ошибок, б) или не более одной грубой ошибки и одного недочета, в) или не более двух-трех негрубых ошибок, г) или одной негрубой ошибки и трех недочетов, д) или при отсутствии ошибок, но при наличии 4-5 недочетов.

Оценка «неудовлетворительно» выставляется, когда число ошибок и недочетов превосходит норму, при которой может быть выставлена оценка

«зачтено (удовлетворительно)» или, если правильно выполнено менее половины работы; если обучающийся не приступал к выполнению работы или правильно выполнил не более 10 процентов всех заданий.

3.11. Оценка формы самостоятельной работы «Тестирование»

Оценка за контроль ключевых компетенций, учащихся производится по пятибалльной системе. При выполнении заданий ставится отметка:

«5» – 85–100%;

«4» – 70–84%;

«3» – 50–69%;

«2» – менее 50%.

Основным критерием эффективности усвоения учащимися содержания учебного материала считается коэффициент усвоения учебного материала – K_y . Он определяется как отношение правильных ответов учащихся к общему количеству вопросов:

$$K_y = N/K,$$

где N – количество правильных ответов учащихся, а K – общее число вопросов. Если $K_y > 0.7$, то учебный материал считается усвоенным.

4. ФОРМА ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ И КРИТЕРИИ ОЦЕНКИ:

Зачёт – форма промежуточной аттестации, при которой уровень подготовки обучающихся оценивается на основе критерия «зачтено» или «не зачтено». Результат зачёта выставляется преподавателем в журнал учебных занятий, оценочную ведомость и в зачётную книжку обучающегося («не зачтено» в зачётную книжку не проставляется).

Критерии оценки выставления зачёта:

Оценка «зачтено» выставляется, если студент твёрдо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, владеет необходимыми знаниями, умениями и навыками.

Оценка «не зачтено» выставляется, если студент не знает значительной части программного материала, допускает существенные ошибки, неуверенно, с большими затруднениями выполняет практические работы.

Экзамен – одна из форм промежуточной аттестации, заключительная форма контроля по учебной дисциплине или междисциплинарному курсу.

Экзамены проводятся в устной форме по экзаменационным билетам.

Критерии оценки выставления экзамена:

Оценка «Отлично» выставляется, если обучающийся демонстрирует глубокие, исчерпывающие знания материала, понимание сущности и взаимосвязи рассматриваемых процессов и явлений, даёт логически последовательный, правильный, полный ответ на вопрос экзаменационного билета и дополнительные вопросы.

Оценка «Хорошо» выставляется, если обучающийся продемонстрировал твёрдые и достаточно полные знания материала, правильное понимание сущности и взаимосвязи рассматриваемых процессов и явлений, даны последовательные, правильные ответы на поставленные вопросы, допущены единичные несущественные неточности.

Оценка «Удовлетворительно» выставляется, если обучающийся демонстрирует знания и понимание основных вопросов, даётся правильный ответ на вопрос экзаменационного билета, без грубых ошибок, при ответе допущены существенные неточности.

Оценка «Неудовлетворительно» выставляется, если обучающийся не дал ответа или дал неправильный ответ на вопрос экзаменационного билета, продемонстрировал непонимание сущности предложенного вопроса, допускает грубые ошибки при ответе на вопрос.

ТЕСТОВЫЕ ЗАДАНИЯ

1. Что из перечисленного является ключевым отличием больших данных от обычных данных?

- A) Большие данные всегда структурированы
- B) Большие данные требуют новых подходов к обработке и анализу
- C) Большие данные имеют фиксированную структуру
- D) Большие данные не могут быть хранены в облачных системах

2. Какой из параметров НЕ относится к характеристикам Big Data («четыре V»)?

- A) Volume – объём
- B) Velocity – скорость
- C) Variability – изменчивость
- D) Veracity – достоверность

3. Чем информация отличается от данных?

- A) Информацией называют любые цифры без контекста
- B) Информация — это данные, осмысленно интерпретированные в определённом контексте
- C) Информация всегда имеет высокую точность
- D) Информация может существовать независимо от данных

4. Какова главная цель использования Big Data в бизнесе?

- A) Увеличить количество сотрудников
- B) Повысить эффективность принятия решений за счёт анализа скрытых паттернов
- C) Заменить все бумажные документы на электронные
- D) Снизить затраты на аренду офисов

5. Что подразумевает термин «data-driven decision making»?

- A) Решения принимаются исключительно на основе интуиции
- B) Решения основаны на фактах, полученных из анализа данных
- C) Все решения формируются автоматически без участия человека
- D) Использование только внешних источников информации

6. Что входит в состав процесса управления большими данными?

- A) Только сбор данных из социальных сетей
- B) Организация доступа, контроль качества, обеспечение безопасности и целостности данных
- C) Создание графических интерфейсов для пользователей
- D) Разработка мобильных приложений

7. Какой элемент управления данными обеспечивает соответствие данным нормативным требованиям?

- A) Метаданные
- B) Государственная регистрация
- C) Комплаенс-контроль
- D) Автоматизация загрузки

8. Что такое data governance?

- A) Процесс удаления устаревших файлов
- B) Набор политик, процедур и ролей, регулирующих управление информационными активами компании
- C) Методология разработки программного обеспечения
- D) Поддержка пользовательских запросов через чат-бот

9. Какой из факторов наиболее влияет на качество больших данных?

- A) Выбор языка программирования
- B) Наличие метаданных и стандартов описания
- C) Размер серверной комнаты
- D) Частота обновления веб-страниц

10. Для чего используется система управления жизненным циклом данных?

- A) Для повышения скорости интернет-соединения
- B) Для контроля доступа к базам данных
- C) Для организации этапов создания, хранения, использования и уничтожения данных
- D) Для шифрования паролей пользователей

11. На каком этапе жизненного цикла данных происходит первичное формирование или сбор данных?

- A) Анализ
- B) Хранение
- C) Сбор/генерация
- D) Удаление

12. Что происходит на этапе «обработка» в жизненном цикле данных?

- A) Передача данных по сети
- B) Преобразование сырых данных в структурированный вид
- C) Формирование отчётов для руководства
- D) Распределение данных среди пользователей

13. Какой из этапов жизненного цикла данных связан с созданием резервных копий?

- A) Удаление
- B) Хранение
- C) Анализ
- D) Представление

14. Почему важно планировать этап утилизации данных?

- A) Чтобы увеличить стоимость хранения
- B) Чтобы минимизировать риски утечки информации и соответствовать законодательству
- C) Чтобы повысить производительность CPU
- D) Чтобы ускорить передачу данных между серверами

Правильный ответ: B

15. Какой этап жизненного цикла данных следует сразу после «объединения»?

- A) Представление
- B) Обработка
- C) Сбор
- D) Анализ

16. Что является начальным шагом в жизненном цикле аналитики данных?

- A) Интерпретация результатов
- B) Формулировка вопроса или проблемы
- C) Выбор инструмента визуализации
- D) Сбор данных вручную

17. На каком этапе проводится очистка и подготовка данных для анализа?

- A) Интерпретация
- B) Моделирование
- C) Подготовка данных
- D) Отчётность

18. Какой этап жизненного цикла аналитики данных предполагает выбор модели или метода анализа?

- A) Реализация
- B) Оценка
- C) Моделирование
- D) Визуализация

19. Что делает этап «оценка» в жизненном цикле аналитики?

- A) Обеспечивает обратную связь от пользователя
- B) Проверяет корректность и значимость полученных результатов
- C) Проводит обучение команды
- D) Удаляет временные таблицы

20. Как завершается жизненный цикл аналитики данных?

- A) Публикацией статьи в научном журнале

- В) Переходом к новому проекту
- С) Внедрением рекомендаций в бизнес-процессы
- Д) Окончательным утверждением бюджета

21. Что такое метаданные?

- А) Данные о пользователях системы
- В) Данные, содержащие информацию о структуре, содержании и способах использования других данных
- С) Резервные копии баз данных
- Д) Лог-файлы приложений

22. Какой тип метаданных описывает логическую структуру данных?

- А) Операционные метаданные
- В) Техническое метаданные
- С) Бизнес-метаданные
- Д) Интеграционные метаданные

23. Какова роль метаданных в ETL-процессах?

- А) Ускоряют передачу данных между системами
- В) Определяют, какие данные нужно экспортировать из источника
- С) Содержат описание схемы, типа и происхождения данных
- Д) Заменяют сами данные в хранилище

24. Какой элемент метаданных может содержать информацию о владельце данных?

- А) Источник данных
- В) Схема таблицы
- С) Описание бизнес-правил
- Д) Политики доступа

25. К какому типу метаданных относятся сведения о времени последнего изменения записи?

- А) Технические метаданные
- В) Операционные метаданные
- С) Стратегические метаданные
- Д) Функциональные метаданные

26. Какая технология наиболее часто используется для обработки потоковых данных в реальном времени?

- А) Hadoop MapReduce
- В) Apache Spark SQL
- С) Apache Kafka
- Д) PostgreSQL JSONB

27. Что из перечисленного является ключевым компонентом архитектуры Big Data?

- A) Центральный процессор высокой частоты
- B) Распределённая файловая система (например, HDFS)
- C) Однородная сеть подключений
- D) Графический интерфейс пользователя

28. Какой подход позволяет обрабатывать большие объёмы данных параллельно?

- A) Монолитное программирование
- B) Нелинейная обработка
- C) Распределённые вычисления
- D) Последовательная обработка

29. Для чего применяется Apache Flink?

- A) Для хранения временных рядов
- B) Для стриминговой обработки данных с низкой задержкой
- C) Для создания веб-интерфейсов
- D) Для выполнения SQL-запросов к NoSQL-базам

30. Какой из перечисленных форматов данных эффективен для работы с большими объемами неструктурированных данных?

- A) CSV
- B) XML
- C) Avro
- D) PDF

31. Чем отличается хранилище данных от операционной базы данных?

- A) Хранилище данных всегда работает в режиме онлайн-обработки транзакций
- B) Хранилище данных предназначено для аналитических запросов, а не для транзакций
- C) Хранилище данных использует только NoSQL-технологии
- D) Хранилище данных хранит данные в реальном времени

32. Какой принцип лежит в основе архитектуры хранилища данных?

- A) Изменяемость исходных данных
- B) Непостоянство структуры
- C) Интеграция и согласованность данных из разных источников
- D) Отсутствие исторического хранения

33. Что означает термин «консолидация данных» в контексте хранилищ?

- A) Перенос данных на внешние носители
- B) Объединение данных из различных источников в единую структуру
- C) Удаление устаревших записей

D) Автоматическое шифрование информации

34. Какой тип модели данных чаще всего используется в хранилищах данных?

- A) Объектно-ориентированная модель
- B) Реляционная модель с денормализацией
- C) Иерархическая модель
- D) Графовая модель

35. Что такое «слой загрузки» в ETL-процессе?

- A) Этап удаления дубликатов
- B) Процесс фильтрации данных до загрузки
- C) Шаг, где данные преобразуются и загружаются во временную зону
- D) Конечная точка хранения данных

36. Какой подход обеспечивает горизонтальное масштабирование данных?

- A) Увеличение мощности одного сервера
- B) Разделение данных по узлам (sharding)
- C) Использование RAID-массивов
- D) Упрощение схемы базы данных

37. Что такое многоуровневое хранение данных?

- A) Хранение данных в нескольких физических местах одновременно
- B) Распределение данных по уровням скорости и стоимости (hot/cold data)
- C) Сохранение одной версии данных на всех узлах
- D) Хранение данных в виде графов

38. Какой из следующих сервисов предлагает автоматическое управление уровнем хранения?

- A) Amazon S3 with Intelligent Tiering
- B) Oracle RAC
- C) MySQL Cluster
- D) Microsoft Access

39. Что означает термин «cold storage» в контексте хранения больших данных?

- A) Высокоэффективное кэширование
- B) Хранение часто запрашиваемых данных
- C) Долгосрочное хранение редко используемых данных
- D) Резервное копирование в реальном времени

40. Почему важно использовать разделение данных (data partitioning)?

- A) Чтобы увеличить стоимость хранения
- B) Чтобы снизить время отклика при выполнении запросов

- C) Чтобы исключить использование SQL
- D) Чтобы устранить необходимость в индексах

41. Какой из перечисленных продуктов является системой управления большими данными?

- A) MySQL Workbench
- B) Apache Hive
- C) Visual Studio Code
- D) Adobe Acrobat

42. Что предоставляет Apache Oozie?

- A) Интерфейс для визуализации данных
- B) Платформа для управления рабочими нагрузками в среде Hadoop
- C) Система резервного копирования
- D) Язык для написания SQL-запросов

43. Какой механизм позволяет контролировать распределение ресурсов в кластере Hadoop?

- A) YARN
- B) ZooKeeper
- C) HBase
- D) Sqoop

44. Какой из следующих инструментов обеспечивает управление жизненным циклом данных в Big Data-системах?

- A) Airflow
- B) Logstash
- C) Prometheus
- D) Redis

45. Что делает Apache Ranger?

- A) Предоставляет защиту доступа к данным через политики безопасности
- B) Обеспечивает автоматическое создание резервных копий
- C) Преобразует данные в формат JSON
- D) Ускоряет чтение данных из SSD

46. Какой инструмент чаще всего используется для визуализации результатов анализа больших данных?

- A) Git
- B) Tableau
- C) Docker
- D) Vim

47. Что такое OLAP-анализ?

- A) Анализ в реальном времени
- B) Онлайн-обработка транзакций
- C) Поддержка многомерного анализа данных
- D) Обнаружение аномалий в потоках данных

48. Какой язык используется для анализа данных в среде Spark?

- A) Python
- B) Scala
- C) Java
- D) Все перечисленные

49. Какой из перечисленных инструментов предназначен для сбора и анализа событий в реальном времени?

- A) ELK Stack (Elasticsearch, Logstash, Kibana)
- B) Power BI
- C) SAS
- D) MATLAB

50. Какой подход позволяет находить скрытые закономерности в данных?

- A) Ручной анализ данных
- B) Статистическая проверка гипотез
- C) Обучение с учителем
- D) Машинное обучение и data mining

51. Какой метод используется для выявления групп в данных?

- A) Регрессия
- B) Кластеризация
- C) Корреляция
- D) Суммирование значений

52. Что такое регрессия в контексте анализа данных?

- A) Метод классификации объектов
- B) Метод прогнозирования числового значения на основе входных переменных
- C) Метод визуализации зависимостей
- D) Метод определения выбросов

53. Какой из методов помогает выявить аномальные данные?

- A) PCA (главные компоненты)
- B) kNN (k ближайших соседей)
- C) Anomaly Detection (обнаружение аномалий)
- D) Clustering (кластеризация)

54. Что такое «размерность» в машинном обучении?

- A) Количество пользователей системы

- B) Количество признаков или столбцов в наборе данных
- C) Размер файла с данными
- D) Частота обновления данных

55. Какой метод часто используется для снижения размерности данных?

- A) Линейная регрессия
- B) PCA (Principal Component Analysis)
- C) SVM
- D) Naïve Bayes

56. Какой из подходов описывает архитектуру, сочетающую batch- и stream-обработку?

- A) Monolithic architecture
- B) Lambda Architecture
- C) Event-driven architecture
- D) RESTful API

57. Что из нижеперечисленного является ключевым компонентом архитектуры Lambda?

- A) Только offline-обработка
- B) Offline-, real-time- и serving-слои
- C) Единая база данных для всех видов данных
- D) Отказоустойчивый центр обработки данных на одном сервере

58. Какой слой в архитектуре Карра отвечает за обработку потоковых данных?

- A) Слой хранения
- B) Слой анализа
- C) Слой обработки потоков
- D) Слой визуализации

59. Что представляет собой «serving layer» в архитектуре Lambda?

- A) Модуль хранения данных
- B) Часть, отвечающая за быстрое получение готовых результатов аналитики
- C) Компонент для мониторинга производительности
- D) Служба авторизации пользователей

60. Каково преимущество архитектуры Карра по сравнению с Lambda?

- A) Более высокая надёжность при отказах
- B) Простота реализации за счёт отказа от batch-слоя
- C) Большая скорость развертывания
- D) Возможность использовать только SQL

61. Что лежит в основе параллельного выполнения алгоритмов в Big Data?

- A) Использование одного процессора с повышенной тактовой частотой
- B) Распределение задачи на несколько узлов или потоков
- C) Применение функционального программирования
- D) Создание единого массива данных

62. Какой алгоритм широко используется для обработки больших массивов данных в параллельном режиме?

- A) Бинарный поиск
- B) Алгоритм Дейкстры
- C) MapReduce
- D) Алгоритм Эйлера

63. Что такое «map» в модели MapReduce?

- A) Этап объединения результатов
- B) Этап сортировки данных
- C) Этап обработки данных с прямым применением функции к каждому элементу
- D) Этап валидации входных данных

64. Какой из факторов не влияет на эффективность параллельных алгоритмов?

- A) Скорость соединения между узлами
- B) Наличие ошибок в коде
- C) Размер данных
- D) Количество узлов кластера

65. Что такое «reduce» в модели MapReduce?

- A) Этап фильтрации данных
- B) Этап группировки и агрегации результатов после map
- C) Этап хранения данных
- D) Этап визуализации результатов

66. Какая из перечисленных платформ является полноценным решением для обработки больших данных на основе Hadoop?

- A) Apache Spark
- B) Microsoft Excel
- C) SQLite
- D) MongoDB Compass

67. Что из нижеперечисленного является частью экосистемы Hadoop?

- A) Apache Kafka
- B) Apache Spark
- C) Apache HBase
- D) Все вышеупомянутое

68. Какой проект позволяет выполнять SQL-подобные запросы к данным в Hadoop?

- A) Flume
- B) Pig
- C) Hive
- D) Sqoop

69. Для чего используется Apache Zeppelin?

- A) Для хранения данных в виде таблиц
- B) Для выполнения и визуализации аналитических заметок в интерактивном режиме
- C) Для перемещения данных между системами
- D) Для контроля версий кода

70. Какой из инструментов предназначен для интеграции данных из разных источников в Hadoop?

- A) NiFi
- B) Jupyter Notebook
- C) Elasticsearch
- D) RabbitMQ

71. Какой тип оборудования наиболее важен для повышения пропускной способности в ЦОД?

- A) ОЗУ с низкой ёмкостью
- B) Сетевые интерфейсы с высокой пропускной способностью
- C) HDD вместо SSD
- D) Низкоскоростные процессоры

72. Что такое «высота стойки» в контексте ЦОД?

- A) Количество серверов в шкафу
- B) Единица измерения плотности установки оборудования (U)
- C) Высота комнаты в метрах
- D) Уровень защиты данных

73. Какой элемент ЦОД обеспечивает бесперебойную работу оборудования?

- A) Резервное питание и кондиционирование
- B) Установка офисных столов
- C) Установка Wi-Fi-роутеров
- D) Программное обеспечение для общения

74. Что такое «цифровой двойник» в ЦОД?

- A) Копия физической инфраструктуры, отражающая её состояние в цифровом виде

- B) Специальная форма визуализации сетевых потоков
- C) Программа для расчёта затрат
- D) Устройство для хранения данных на магнитной ленте

75. Какое явление связано с перегревом серверов в ЦОД?

- A) Тепловое расслоение
- B) Thermal throttling
- C) Cloud drift
- D) Memory leak

76. Какой инструмент наиболее популярен для построения интерактивных дашбордов?

- A) Matplotlib
- B) Plotly
- C) Pandas
- D) NumPy

77. Что из нижеперечисленного лучше всего подходит для визуализации распределения данных?

- A) Диаграмма рассеяния
- B) Гистограмма
- C) Таблица с числами
- D) Круговая диаграмма

78. Какой тип диаграммы используется для демонстрации трендов во времени?

- A) Гистограмма
- B) Линейчатая диаграмма
- C) Диаграмма Венна
- D) Box plot

79. Что означает термин «информационная перегрузка» визуализации?

- A) Невозможность сохранить данные
- B) Чрезмерное количество элементов на графике, затрудняющее восприятие
- C) Ошибка в математическом расчете
- D) Отсутствие цветового оформления

80. Какой стандарт используют современные библиотеки для визуализации?

- A) HTML/CSS/JavaScript
- B) XML/JSON
- C) SQL
- D) Markdown

81. К какой категории относится задача прогнозирования продаж на

следующий месяц?

- A) Классификация
- B) Кластеризация
- C) Прогнозирование (регрессия)
- D) Обнаружение аномалий

82. Какая задача решается с помощью алгоритма K-Means?

- A) Прогнозирование цен акций
- B) Классификация клиентов по категориям
- C) Обнаружение мошенничества
- D) Группировка данных по схожести

83. Что такое задача классификации?

- A) Предсказать значение величины
- B) Присвоить объекту один из заранее известных классов
- C) Определить, есть ли в данных выбросы
- D) Найти зависимости между переменными

84. Какая задача направлена на выявление аномальных паттернов в данных?

- A) Регрессия
- B) Кластеризация
- C) Обнаружение аномалий
- D) Ассоциативный анализ

85. Что такое ассоциативный анализ?

- A) Поиск связи между двумя переменными
- B) Нахождение закономерностей в покупках (например, «купили хлеб – купили масло»)
- C) Прогнозирование будущего поведения
- D) Группировка данных по возрасту

85. Какой показатель характеризует степень связи между двумя переменными?

- A) Дисперсия
- B) Коэффициент корреляции
- C) Мода
- D) Среднее арифметическое

86. Что такое p-value в статистическом тестировании гипотез?

- A) Вероятность того, что выборка случайна
- B) Вероятность получить такой же или более экстремальный результат при справедливости нулевой гипотезы
- C) Уровень доверия к результату
- D) Ошибка первого рода

87. Какой метод используется для проверки различий между средними значениями двух групп?

- A) t-критерий Стьюдента
- B) ANOVA
- C) Хи-квадрат
- D) Метод главных компонент

89. Что означает термин «дисперсионный анализ» (ANOVA)?

- A) Сравнение средних значений более чем двух групп
- B) Оценка степени вариабельности данных
- C) Проверка нормальности распределения
- D) Метод кластеризации данных

90. Какой параметр можно оценить с помощью доверительного интервала?

- A) Среднее значение генеральной совокупности
- B) Количество выборок
- C) Стандартное отклонение данных в таблице
- D) Количество строк в таблице

СПИСОК ЛИТЕРАТУРЫ

Основная литература:

1. Параскевов, А. В. Большие данные: учебное пособие / А. В. Параскевов, А. Э. Сергеев. - Москва; Вологда: Инфра-Инженерия, 2024. - 148 с. - ISBN 978-5-9729-2120-1. - Текст: электронный. - URL: <https://znanium.ru/catalog/product/2169699> (дата обращения: 12.03.2026). - Режим доступа: по подписке.
2. Ланских, Ю. В. Введение в большие данные: учебное пособие / Ю. В. Ланских, В. Г. Ланских, К. В. Родионов. – Киров: ВятГУ, 2023. – 172 с. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/408566> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.
3. Конкина, В. В. Введение в большие данные и анализ информации: учебное пособие / В. В. Конкина, А. Б. Борисенко, И. Л. Коробова. – Тамбов: ТГТУ, 2024. – 82 с. – ISBN 978-5-8265-2749-8. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/472331> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

Дополнительная литература:

1. Адлер, Ю. П. Статистическое управление процессами. «Большие данные»: учебное пособие / Ю. П. Адлер, Е. А. Черных. - Москва: Изд. Дом МИСиС, 2016. - 52 с. - ISBN 978-5-87623-969-3. - Текст: электронный. - URL: <https://znanium.com/catalog/product/1232190> (дата обращения: 12.03.2026). - Режим доступа: по подписке.
2. Буданцев, А. В. Методы анализа данных (01.03.04, 09.03.03): учебное пособие / А. В. Буданцев, Н. А. Морошкин, Н. Н. Тетерин. – Москва: РТУ МИРЭА, 2025. – 63 с. – ISBN 978-5-7339-2646-9. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/507479> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.
3. Баланов, А. Н. Цифровая трансформация государственного сектора. Интеграция, управление и безопасность: учебное пособие для вузов / А. Н. Баланов. – Санкт-Петербург: Лань, 2024. – 328 с. – ISBN 978-5-507-49310-4. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/417749> (дата обращения: 12.03.2026). – Режим доступа: для авториз. пользователей.

Интернет-ресурсы:

1. Сайт Института социологии РАН – <http://www.isras.ru/socis.html>.
2. Журнал «Социологические исследования» // [Электрон. ресурс]. – Режим доступа: <http://socis.isras.ru>
3. Гуманитарный Интернет портал // [Электрон. ресурс]. – Режим доступа: www.auditorium.ru
4. Санкт-Петербургский государственный университет. Факультет

- социологии [Электрон. ресурс]. – Режим доступа: <http://www.soc.pu.ru/archive/>
5. «Фонд общественное мнение» (ФОМ) // [Электрон. ресурс]. – Режим доступа: <http://www.fom.ru/>
 6. ВЦИОМ [Электрон. ресурс]. – Режим доступа: <https://wciom.ru/>
 7. Левада-центр [Электрон. ресурс]. – Режим доступа: <http://www.levada.ru/>
 8. Электронно-библиотечная система «Университетская библиотека онлайн» - www.biblioclub.ru;
 9. Электронная библиотечная система «IPRbooks»;
 10. База данных SCOPUS и др.