

**Федеральное государственное бюджетное учреждение науки
Институт машиноведения им. А.А. Благонравова
Российской академии наук
(ИМАШ РАН)**

УТВЕРЖДАЮ
Заместитель директора по вопросам
подготовки кадров высшей
квалификации ИМАШ РАН
к.ф.-м.н. Жданов Э.Р.
« 01 » _____ 2026 года
М.П.



**Комплект оценочных материалов
по дисциплине «Методы анализа больших данных (Big Data Analytics)»**

Направление подготовки:	02.04.03 Математическое обеспечение и администрирование информационных систем
Профиль подготовки:	Искусственный интеллект и интеллектуальная обработка данных
Квалификация:	Магистр
Форма обучения	Очная

Содержание

1. Общие положения.....	3
2. Результаты освоения дисциплины, подлежащие проверке.....	3
3. Контроль и оценка освоения учебной дисциплины.....	4
4. Оценочные материалы для проведения текущего контроля успеваемости.....	4
5. Оценочные материалы для проведения промежуточной аттестации.....	9

1. Общие положения

Комплект оценочных материалов (далее - ОМ) предназначен для контроля и оценки образовательных достижений обучающихся, освоивших программу дисциплины «Методы анализа больших данных (Big Data Analytics)».

ОМ включают контрольные материалы для проведения текущего контроля успеваемости и промежуточной аттестации в форме экзамена.

ОМ разработаны на основании:

- основной профессиональной образовательной программы по направлению подготовки 02.04.03 Математическое обеспечение и администрирование информационных систем;

- рабочей программы дисциплины «Методы анализа больших данных (Big Data Analytics)».

2. Результаты освоения дисциплины, подлежащие проверке

В результате освоения дисциплины обучающийся должен обладать следующими компетенциями

ПК-1	Создание информационных технологий нового поколения
ПК-1.1	Управление этапами жизненного цикла методологической и технологической инфраструктуры анализа больших данных в организации
ПК-1.2	Управление разработкой продуктов, услуг и решений на основе больших данных
ПК-1.3	Разработка и внедрение новых методов и технологий исследования больших данных

Код индикатора компетенции	Планируемые результаты обучения по дисциплине
ПК-1.1	<i>Знать:</i> - современные архитектуры и компоненты инфраструктуры Big Data; - этапы жизненного цикла аналитических систем; - методы мониторинга и оптимизации производительности распределённых систем. <i>Уметь:</i> - проектировать архитектуру хранилищ и вычислительных кластеров для задач анализа данных; - выбирать и внедрять инструменты для сбора, хранения и обработки данных; - организовывать процессы сопровождения и развития инфраструктуры. <i>Иметь практический опыт:</i> - развёртывания и администрирования кластера Hadoop/Spark; - настройки пайплайнов обработки данных; - мониторинга и оптимизации работы распределённых приложений.
ПК-1.2	<i>Знать:</i> - методологию управления проектами в области Big Data; - современные подходы к формированию продуктовых команд (Data Science, Data Engineering); - методы оценки эффективности аналитических продуктов. <i>Уметь:</i> - формулировать требования к аналитическим решениям; - организовывать процесс разработки и тестирования моделей на больших данных; - интегрировать аналитические сервисы в бизнес-процессы организации. <i>Иметь практический опыт:</i> - проектирования и реализации прототипа продукта на основе анализа больших данных; - презентации и защиты аналитического решения перед стейкхолдерами.

ПК-1.3	<i>Знать:</i> - современные методы машинного обучения и анализа данных для Big Data; - тенденции развития технологий искусственного интеллекта в области обработки больших объёмов информации; - принципы экспериментирования и внедрения инноваций. <i>Уметь:</i> - адаптировать существующие алгоритмы для распределённой реализации; - проводить исследовательские эксперименты с новыми методами анализа данных; - оценивать применимость новых технологий для решения практических задач. <i>Иметь практический опыт:</i> - реализации исследовательского проекта по анализу больших данных с использованием нестандартных методов или инструментов.
--------	--

3. Контроль и оценка освоения учебной дисциплины

Элемент учебной дисциплины	Текущий контроль		Промежуточная аттестация	
	Форма контроля	Проверяемые идк	Форма контроля	Проверяемые идк
Тема 1. Введение в анализ больших данных	Отчет по лабораторному практикуму	ПК-1.1 ПК-1.2 ПК-1.3		
Тема 1. Введение в анализ больших данных	Отчет по практикуму по решению задач	ПК-1.1 ПК-1.2 ПК-1.3		
Тема 2. Методы предобработки и анализа данных	Отчет по лабораторному практикуму	ПК-1.1 ПК-1.3		
Тема 2. Методы предобработки и анализа данных	Отчет по практикуму по решению задач	ПК-1.1 ПК-1.3		
Тема 3. Машинное обучение на больших данных	Отчет по лабораторному практикуму	ПК-1.1 ПК-1.3		
Тема 3. Машинное обучение на больших данных	Отчет по практикуму по решению задач	ПК-1.1 ПК-1.3		
Тема 4. Управление проектами и внедрение решений	Отчет по лабораторному практикуму	ПК-1.1 ПК-1.2		
Тема 4. Управление проектами и внедрение решений	Отчет по практикуму по решению задач	ПК-1.1 ПК-1.2		
			Экзамен	ПК-1.1 ПК-1.2 ПК-1.3

4. Оценочные материалы для проведения текущего контроля успеваемости

Показатели и критерии оценивания компетенций на различных этапах их формирования, описание шкал оценивания

Лабораторный практикум
15-13 – работа выполнена в срок, самостоятельно, правильно поняты и использованы соответствующие формулы, правильно определены соответствующие спецификации,

использована требуемая информация, правильно выполнены требуемые расчеты, правильно выбраны совместимые комплектующие, сделаны необходимые выводы, хорошо аргументированы, даны исчерпывающие ответы на все поставленные вопросы;

12-9 – работа выполнена в срок, самостоятельно, правильно поняты и использованы соответствующие формулы, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, правильно выбраны совместимые комплектующие, необходимые выводы сделаны частично, хорошо аргументированы, даны ответы на все поставленные вопросы;

8-5 – работа выполнена в срок, в основном самостоятельно, использованы соответствующие формулы; определены соответствующие спецификации, имеются ошибки в расчетах; выбраны совместимые комплектующие необходимые, выводы сделаны частично, слабо аргументированы, даны ответы не на все вопросы;

4-1 – обучающийся подготовил работу несамостоятельно или не завершил в срок, описание спецификации содержит незначительные ошибки, выводы и ответы на вопросы отсутствуют.

Практикум по решению задач

10-8 – работа и отчет выполнены в срок, самостоятельно, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, сделаны необходимые выводы, хорошо аргументированы, даны исчерпывающие ответы на все поставленные вопросы;

7-5 – работа и отчет выполнены в срок, самостоятельно, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, необходимые выводы сделаны частично, хорошо аргументированы, даны ответы на все поставленные вопросы;

4-1 – работа и отчет выполнены в срок, в основном самостоятельно, выводы сделаны частично, слабо аргументированы, даны ответы не на все вопросы;

0 – обучающийся подготовил работу и отчет несамостоятельно и/или не завершил в срок, описание содержит ошибки, выводы и ответы на вопросы отсутствуют.

Порядок начисления баллов

Если студент набрал достаточное количество баллов ТКУ по дисциплине (более 70 баллов) в течение семестра, он может отказаться от сдачи экзамена, и результат будет выставлен автоматически по его согласию.

Типовые контрольные задания и иные материалы, необходимые для процедуры оценивания знаний, умений, навыков и(или) опыта деятельности

Типовые задания для практикумов по решению задач

Тема 1. Введение в анализ больших данных

Задача 1. Установка и настройка кластера Hadoop/Spark, загрузка и первичная обработка данных.

Цель: получить опыт развертывания распределенной среды и выполнения базовых операций с данными.

Инструменты: Docker, Hadoop (HDFS, YARN), Apache Spark (PySpark или Scala).

Ход работы:

- развернуть кластер Hadoop (например, с помощью Docker Compose или облачного сервиса) на локальной машине или виртуальных машинах.
- проверить работоспособность кластера (запустить базовые команды HDFS, убедиться в запуске ResourceManager и NodeManager).
- загрузить в HDFS публичный датасет большого объема (например, логи веб-сервера, данные о транзакциях или набор данных с Kaggle).
- подключиться к кластеру Spark и выполнить базовые операции: чтение данных из HDFS, просмотр схемы, фильтрация строк по условию, группировка и агрегация данных (например, подсчет количества записей по категориям).
- сохранить результат обработки обратно в HDFS.

Результат: Работающий кластер, загруженный датасет и отчет с кодом на

PySpark/Scala, показывающий базовые навыки работы с экосистемой.

Тема 2. Методы предобработки и анализа данных

Задача 1. Реализация ETL-процесса с помощью Spark и построение интерактивных дашбордов

Цель: создать сквозной пайплайн обработки данных от загрузки до визуализации.

Инструменты: Apache Spark, Apache Airflow (или аналог), BI-система (например, Apache Superset, Tableau, Power BI).

Ход работы:

- выбрать исходный "сырой" датасет (например, JSON-файлы с логами или CSV с некорректными типами данных).

- написать Spark-скрипт для ETL-процесса: Extract: Загрузка данных из источника.

Transform: Очистка (удаление пропусков, дубликатов), исправление типов данных, стандартизация форматов (даты, строки), обогащение данных (например, вычисление новых признаков).

Load: Запись очищенных данных в аналитическое хранилище (например, Parquet-файлы в S3/HDFS или таблицу в ClickHouse).

- настроить расписание выполнения ETL-процесса с помощью Airflow.

- подключить BI-систему к хранилищу очищенных данных и построить интерактивный дашборд с ключевыми метриками и графиками.

Результат: Работающий ETL-пайплайн по расписанию и опубликованный интерактивный дашборд.

Тема 3. Машинное обучение на больших данных

Задача 1. Реализация задачи классификации на Spark MLlib

Цель: научиться строить и оценивать модели машинного обучения в распределенной среде.

Инструменты: PySpark или Scala, Spark MLlib.

Ход работы:

- выбрать задачу бинарной классификации (например, предсказание оттока клиентов или спама). Загрузить размеченный датасет большого размера в Spark DataFrame.

- выполнить предварительную обработку признаков: кодирование категориальных переменных (StringIndexer, OneHotEncoder), масштабирование числовых признаков (StandardScaler), сбор всех признаков в один вектор (VectorAssembler).

- разбить данные на обучающую и тестовую выборки.

- обучить модель логистической регрессии (или случайного леса) на обучающей выборке с помощью Spark MLlib.

- оценить качество модели на тестовой выборке с помощью метрик (Area Under ROC, Accuracy, F1-Score) для несбалансированных классов.

- провести тюнинг гиперпараметров модели с использованием CrossValidator.

Результат: Обученная модель с оценкой качества и отчет о проведенном эксперименте по тюнингу.

Тема 4. Управление проектами и внедрение решений

Задача 1. Проектирование и презентация прототипа аналитического решения для бизнес-кейса

Цель: отработать навыки сбора требований и проектирования end-to-end решения для бизнес-задачи.

Формат: Командная работа (3-5 человек) с финальной презентацией перед "заказчиком" (преподавателем).

Ход работы:

- команда получает описание бизнес-кейса (например, "Разработать систему персональных рекомендаций для онлайн-кинотеатра" или "Создать систему мониторинга

мошеннических транзакций").

Этап анализа:

- сбор и формализация требований к системе (Functional/Non-functional requirements).

- определение ключевых метрик успеха (KPI) для проекта.

Этап проектирования:

- разработка архитектуры решения: выбор источников данных, описание ETL-пайплайна, выбор хранилища (Data Lake/Warehouse), описание модуля машинного обучения.

- создание диаграммы архитектуры (например, в draw.io или Miro) и описание технологического стека (Tech Stack).

- описание этапов внедрения модели в продакшн (MLOps): версионирование моделей (MLflow), мониторинг качества.

- подготовка презентации (10-15 минут) и демонстрация прототипа архитектуры/кода ETL-пайплайна.

Результат: Презентация проекта перед комиссией и готовый пакет проектной документации (требования, архитектура).

Типовые задания для лабораторных практикумов

Тема 1. Введение в анализ больших данных

Задание 1. Сравнительный анализ производительности HDFS и локальной файловой системы

Цель: оценить накладные расходы и преимущества распределенной файловой системы при работе с большими объемами данных.

Инструменты: Hadoop (HDFS), Python (библиотека time).

Ход работы: Подготовить синтетический датасет большого размера (например, 10 ГБ) с помощью Python-скрипта.

Измерить время выполнения операции копирования этого файла в локальную файловую систему (cp/shutil).

Измерить время выполнения операции копирования этого же файла в HDFS (hdfs dfs -put).

Измерить время чтения файла целиком из локальной ФС и из HDFS.

Проанализировать результаты: объяснить разницу во времени, учитывая механизмы репликации и сетевой передачи данных в HDFS.

Результат: Таблица с замерами времени и аналитический отчет, объясняющий полученные цифры.

Тема 2. Методы предобработки и анализа данных

Задание 1. Влияние качества данных на результаты агрегации

Цель: продемонстрировать на практике, как пропущенные значения и выбросы искажают итоговые метрики.

Инструменты: Python (Pandas), Jupyter Notebook.

Ход работы:

- создать небольшой синтетический датасет (например, данные о продажах магазина за неделю).

- рассчитать базовые метрики: среднее значение чека, общую выручку.

- искусственно внести в данные проблемы: Удалить часть значений в столбце "цена" (пропуски).

- добавить аномально высокие значения в столбец "сумма чека" (выбросы).

- пересчитать те же метрики на "испорченном" датасете.

- применить методы предобработки: заполнить пропуски медианой/заменой по группе, отфильтровать выбросы с помощью межквартильного размаха (IQR).

- пересчитать метрики снова и сравнить все три результата.

Результат: Визуальное сравнение метрик до и после очистки данных с выводами о важности этапа предобработки.

Тема 3. Машинное обучение на больших данных

Задание 1. Сравнение производительности алгоритмов MLlib на разных объемах данных

Цель: оценить масштабируемость алгоритмов машинного обучения в Spark.

Инструменты: PySpark или Scala.

Ход работы:

- подготовить один и тот же датасет в нескольких размерах: малый (100 МБ), средний (1 ГБ), большой (10 ГБ). Это можно сделать путем репликации малого датасета или генерации синтетических данных.
- выбрать два разных алгоритма из MLlib (например, Logistic Regression и Random Forest) для решения одной задачи классификации.
- для каждого алгоритма и каждого размера датасета измерить время обучения модели (используя встроенные счетчики времени).
- построить графики зависимости времени обучения от объема данных для каждого алгоритма.
- проанализировать результаты: какой алгоритм масштабируется лучше? Как время обучения соотносится с ростом объема данных?

Результат: Графики производительности и аналитический отчет о масштабируемости выбранных алгоритмов.

Тема 4. Управление проектами и внедрение решений

Задание 1. Проектирование CI/CD пайплайна для ML-модели

Цель: получить представление об автоматизации жизненного цикла модели от разработки до деплоя.

Инструменты: GitHub Actions / GitLab CI / Jenkins, Docker, MLflow (или аналог).

Ход работы:

Задача — описать этапы пайплайна для гипотетического проекта. Описать этапы пайплайна: Build: Сборка Docker-образа с кодом модели и зависимостями по коммиту в ветку develop.

Test: Запуск модульных тестов кода модели на синтетических данных внутри контейнера.

Train: при слиянии в main — автоматический запуск обучения модели на выделенном сервере/кластере Spark с логированием параметров и метрик в MLflow.

Deploy: при успешном прохождении тестов — автоматическое развертывание новой версии модели как REST-сервиса через Docker-контейнеризатор (например, Docker Compose, Kubernetes) с версионированием через MLflow.

Нарисовать схему этого пайплайна в виде диаграммы последовательности или блок-схемы.

Написать конфигурационный файл для системы CI/CD (например, .github/workflows/ml-pipeline.yml), описывающий эти шаги декларативно.

Описать стратегию отката (rollback) в случае деградации качества модели после деплоя.

Результат: Схема CI/CD пайплайна и конфигурационный файл для его реализации.

5. Оценочные материалы для проведения промежуточной аттестации

Промежуточная аттестация по дисциплине проводится в форме экзамена.

Экзамен представляет собой выполнение обучающимся заданий билета, включающего в себя:

Задание №1 – теоретический вопрос на знание базовых понятий предметной области дисциплины, а также позволяющий оценить степень владения обучающегося принципами

предметной области дисциплины, понимание их особенностей и взаимосвязи между ними.

Задание №2 – задание на анализ ситуации из предметной области дисциплины и выявление способности обучающегося выбирать и применять соответствующие принципы и методы решения практических проблем, близких к профессиональной деятельности.

Задания №3 – задания на проверку умений и навыков, полученных в результате освоения дисциплины.

Выполнение обучающимся заданий билета оценивается по следующей балльной шкале:

Задание 1: 0-30 баллов

Задание 2: 0-30 баллов

Задание 3: 0-40 баллов

-90 и более (отлично) – ответ правильный, логически выстроен, приведены необходимые выкладки, использована профессиональная лексика. Задания решены правильно. Обучающийся правильно интерпретирует полученный результат.

-70 и более (хорошо)– ответ в целом правильный, логически выстроен, приведены необходимые выкладки, использована профессиональная лексика. Ход решения задания правильный, ответ неверный. Обучающийся в целом правильно интерпретирует полученный результат.

-50 и более (удовлетворительно)– ответ в основном правильный, логически выстроен, приведены не все необходимые выкладки, использована профессиональная лексика. Задания решены частично.

-менее 50 (неудовлетворительно)– ответы на теоретическую часть неправильные или неполные. Задания не решены.

Типовые задания для проведения промежуточной аттестации обучающихся

Задания 1 типа

1. Дайте определение понятию Big Data. Перечислите и охарактеризуйте ключевые свойства больших данных (модель 5V или 7V).

2. Назовите основные источники генерации больших данных в современном мире и приведите примеры для каждого источника.

3. Сформулируйте основные задачи, которые решает анализ больших данных в бизнесе и науке.

4. Опишите классическую архитектуру распределённых систем обработки данных. Каковы функции её основных компонентов?

5. Сравните экосистемы Hadoop и Spark: архитектурные различия, модели вычислений (MapReduce vs. RDD/DAG), основные компоненты.

6. Что такое HDFS? Опишите её архитектуру, принципы хранения данных (блочное хранение, репликация) и основные демоны.

7. Дайте классификацию NoSQL-баз данных. В чём принципиальное отличие key-value, document, column-family и graph баз?

8. Перечислите основные этапы установки и настройки кластера Hadoop/Spark. Какие параметры критически важны для производительности?

9. Какие существуют методы и инструменты для загрузки данных в кластер (Data Ingestion) и их первичной обработки?

10. Раскройте этапы процесса предобработки данных: очистка, нормализация, трансформация. Какие проблемы решает каждый этап?

11. Что такое интеграция данных? Опишите методы слияния данных из гетерогенных источников.

12. Для чего применяются методы выборки (sampling) при работе с большими данными? Назовите основные типы выборок.

13. Какова роль агрегации в анализе данных? Приведите примеры агрегирующих функций.

14. Какие инструменты и библиотеки используются для визуализации больших данных? С какими трудностями сталкиваются при визуализации?
15. Опишите процесс ETL (Extract, Transform, Load). Как реализуются этапы трансформации и загрузки данных с использованием Apache Spark?
16. Что такое интерактивный дашборд? Какие компоненты необходимы для его построения на базе стека Big Data?
17. В чём заключаются особенности реализации алгоритмов машинного обучения в распределённой среде по сравнению с локальной?
18. Перечислите основные алгоритмы, реализованные в библиотеке Spark MLlib для задач классификации, регрессии и кластеризации.
19. Опишите принципы работы рекомендательных систем на больших данных. Какие алгоритмы используются для их построения в Spark MLlib?
20. Каковы особенности оценки качества моделей машинного обучения на больших выборках? Какие метрики используются для задач классификации и регрессии?
21. Опишите жизненный цикл проекта Big Data: от постановки бизнес-задачи до внедрения решения в эксплуатацию.
22. Что такое DataOps / DevOps для Data Science? Как эти методологии применяются для организации командной работы над проектами?
23. Каковы этапы внедрения обученной модели машинного обучения в промышленную эксплуатацию? Что такое MLOps?
24. Из каких этапов состоит проектирование прототипа аналитического решения для бизнес-кейса? Каковы критерии успешности прототипа?
25. Какова структура презентации аналитического решения для стейкхолдеров? На какие аспекты следует обратить особое внимание при защите проекта?

Задания 2 типа

1. В крупной торговой сети ежедневно накапливается 5 ТБ логов транзакций из кассовых систем. Руководству необходимо анализировать динамику продаж в режиме, близком к реальному времени. Какую архитектуру обработки данных (стриминг или батч) вы выберете и какие компоненты экосистемы (Kafka, Spark Streaming, Flink) внедрите для решения задачи?
2. Оператор связи фиксирует резкий рост оттока абонентов. У вас есть доступ к данным о звонках, использовании трафика и обращениях в поддержку. Опишите последовательность шагов (от постановки задачи до выбора признаков) для построения предиктивной модели оттока с использованием Spark.
3. Банку необходимо в течение 24 часов выявлять мошеннические транзакции из потока в 10 000 операций в секунду. Как вы спроектируете ETL-конвейер и какой подход к обучению модели (онлайн-обучение или переобучение на батчах) предложите для минимизации задержки?
4. На заводе установлено 1000 IoT-датчиков, генерирующих телеметрию. Руководство хочет внедрить предиктивное обслуживание оборудования. Какие методы интеграции данных (batch vs. streaming) и алгоритмы машинного обучения (классификация аномалий, регрессия) вы примените?
5. У вас есть база данных клиентов (CRM) и сырые логи их поведения на сайте (кликаут-стрим). Перед построением модели сегментации вы обнаружили, что данные содержат дубликаты и пропуски. Опишите ваш план действий по очистке и интеграции этих данных с помощью Spark.
6. Транспортная компания хочет оптимизировать маршруты доставки на основе данных GPS-трекеров за последние 3 года. Данные хранятся в распределённой файловой системе. Как вы построите витрину данных для аналитиков и какие методы агрегации будете использовать для подготовки датасета?
7. Исследовательскому центру нужно объединить данные из электронных медицинских карт (структурированные данные) с неструктурированными текстами

заклучений врачей и снимками МРТ. Какую архитектуру хранилища (Data Lake) и методы предобработки данных вы предложите?

8. Маркетплейс хочет внедрить систему рекомендаций «С этим товаром покупают». У вас есть история просмотров и покупок пользователей. Какую задачу машинного обучения (коллаборативная фильтрация, контентная) вы выберете и как реализуете её на Spark MLlib?

9. Компания хочет выявить скрытые факторы, влияющие на увольнение ценных сотрудников. У вас есть данные о производительности, обучении и обратной связи. Как вы сформулируете задачу (классификация или кластеризация) и какие признаки создадите для модели?

10. Необходимо провести мониторинг репутации бренда в реальном времени по потоку постов пользователей. Как вы построите пайплайн: от сбора данных через API до визуализации тональности на дашборде?

11. Энергосбытовая компания получает данные со «умных» счётчиков ежечасно. Данные поступают с задержкой и не всегда полны. Какие методы импутации (заполнения пропусков) данных и алгоритмы обнаружения аномалий потребления вы примените?

12. Для анализа загруженности дорог необходимо объединить данные с камер видеонаблюдения, GPS-трекеров такси и датчиков на дорогах. Опишите процесс ETL для создания единого датасета о трафике.

13. Авиакомпаниям нужно спрогнозировать спрос на билеты по 500 направлениям на год вперёд с учётом сезонности и цен конкурентов. Какую модель регрессии вы выберете в Spark MLlib и как подготовите временные ряды?

14. Стриминговый сервис фиксирует высокий процент отказа от просмотра видео после первых 30 секунд. Какие метрики вы соберёте из логов плеера, как их агрегируете и какую гипотезу проверите с помощью A/B-тестирования?

15. Система безопасности генерирует миллионы событий логов в день (логи брандмауэра, IDS). Вам нужно автоматически классифицировать их по уровню угрозы. Как вы решите проблему несбалансированности классов при обучении модели?

16. Фермерское хозяйство собирает данные с дронов (мультиспектральные снимки полей) и датчиков влажности почвы. Задача — построить карту предписаний для внесения удобрений. Какой тип NoSQL-хранилища оптимален для хранения геоданных?

17. Разработчик онлайн-игры хочет сегментировать игроков по стилю игры для персонализации предложений в магазине. У вас есть сырые ивенты (события) из игры. Как вы трансформируете их в признаки для кластеризации?

18. Агентство недвижимости имеет базу объявлений (структурированные данные) и архив отсканированных документов на право собственности (неструктурированные). Как реализовать полнотекстовый поиск по документам с помощью инструментов Big Data?

19. Сеть ресторанов внедряет систему прогнозирования спроса на ингредиенты на неделю вперёд на основе данных о бронях столов и погоде. Опишите процесс Feature Engineering для этой задачи.

20. На месторождении необходимо анализировать телеметрию с буровых установок для предотвращения аварий (Predictive Maintenance). Данные поступают с высокой скоростью, но имеют шум. Какой алгоритм фильтрации шума вы примените перед обучением модели?

21. Служба поддержки получает тысячи тикетов в день на разных языках через чат-боты и почту. Необходимо автоматически маршрутизировать обращения по темам и срочности. Какую модель классификации вы построите и как оцените её качество?

22. Стартапу необходимо быстро развернуть прототип аналитического решения для проверки гипотезы о рынке, имея ограниченный бюджет на инфраструктуру. Выберете ли вы локальный кластер Hadoop или облачное решение? Обоснуйте выбор.

23. Банку нужно рассчитать кредитный скоринг для малого бизнеса, используя не только финансовые отчёты, но и данные из открытых источников (новости, соцсети). Как обеспечить легитимность использования внешних данных при интеграции?

24. -университет хочет выявить студентов из «группы риска» (вероятность отчисления). Какие поведенческие признаки из LMS (системы управления обучением) будут наиболее информативными для модели классификации?

25. Вы — руководитель проекта по внедрению DWH (корпоративного хранилища данных). Бизнес-заказчик требует результат через месяц, а объём сырых данных — 50 ТБ, требующих сложной очистки. Как вы спланируете работу команды, используя принципы Agile/DevOps, чтобы показать ценность на раннем этапе?

Задания 3 типа

1. Настройка инфраструктуры и загрузка данных

Разверните локальный (standalone) кластер Spark. Загрузите в память Spark набор данных (например, логи веб-сервера или датасет из репозитория UCI). Выполните базовые операции: просмотр схемы, вывод первых строк, подсчёт количества записей.

2. Реализация ETL-процесса

Имеется набор CSV-файлов с данными о продажах, содержащий пропуски, дубликаты и некорректные значения. Используя PySpark или Scala, реализуйте ETL-пайплайн: очистите данные, нормализуйте форматы дат и числовых значений, выполните трансформацию признаков и сохраните итоговый датасет в формате Parquet в HDFS.

3. Анализ данных и агрегация

На основе загруженного датасета рассчитайте ключевые бизнес-метрики (KPI): среднее значение, медиану, количество уникальных пользователей, сумму продаж по категориям. Используйте методы агрегации Spark (groupBy, agg) и методы выборки для анализа репрезентативности данных.

4. Визуализация результатов

На основе результатов предыдущего задания постройте интерактивный дашборд (например, в Apache Superset, Tableau или Power BI), отображающий динамику продаж по времени и распределение чеков. Подготовьте краткую презентацию выводов на основе визуализаций.

5. Работа с NoSQL-хранилищем

Выберите одну из NoSQL-баз (например, MongoDB, Cassandra или HBase). Спроектируйте схему хранения для данных о пользователях интернет-магазина. Реализуйте загрузку данных из Spark DataFrame в выбранную базу и выполните несколько запросов на выборку (поиск по ID, фильтрация по атрибутам).

6. Построение модели классификации

Используя Spark MLlib, решите задачу бинарной классификации (например, предсказание оттока клиентов или мошенничества с транзакциями). Проведите разведочный анализ данных, выполните векторизацию признаков (VectorAssembler), обучите модель (например, логистическую регрессию или Random Forest) и оцените её качество с помощью метрик (Accuracy, Precision, Recall, F1-Score).

7. Кластеризация данных

Примените алгоритм кластеризации (например, K-Means) из Spark MLlib к набору данных для сегментации клиентов. Определите оптимальное число кластеров с помощью метода «локтя» (Elbow Method). Проанализируйте полученные сегменты и опишите их бизнес-характеристики.

8. Построение рекомендательной системы

Используя готовые данные о рейтингах товаров/фильмов пользователями, постройте простую рекомендательную систему на основе алгоритма collaborative filtering (ALS) в Spark MLlib. Сформируйте топ-5 рекомендаций для заданного пользователя.

9. Внедрение модели в эксплуатацию

Возьмите обученную модель из задания 6. Сохраните её на диск с помощью PipelineModel.save(). Напишите отдельный скрипт-загрузчик, который считывает новые данные, применяет к ним сохранённый пайплайн и выводит предсказания. Опишите, как бы вы организовали мониторинг качества такой модели в промышленной среде.

10. Проектирование аналитического решения

Выберите один из бизнес-кейсов (оптимизация логистики, прогноз спроса, управление рисками). Подготовьте техническое задание на разработку прототипа аналитического решения. Опишите архитектуру системы (источники данных, хранилище, инструменты обработки), этапы ETL-процесса и методы анализа/машинного обучения, которые будут применяться.