ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

Рабочая программа дисциплины (модуля)
«Data Engineering (Инженерия данных)»

Направление подготовки: *02.04.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Программа двух дипломов НИУ ВШЭ и ЦУ «Искусственный интеллект»*

Квалификация (степень) выпускника: *магистр*

Форма обучения: *очная*

Срок освоения программы: *2 года*

Год набора: *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B80C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2024

Содержание

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Data Engineering (Инженерия данных)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – магистратура по специальности 02.04.01 Математика и компьютерные науки, профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Искусственный интеллект», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 810 от 23.08.2017 года.

Изучение дисциплины (модуля) «Data Engineering (Инженерия данных)» способствует созданию надежной инфраструктуры для эффективного анализа и обработки данных, что является ключевым для принятия обоснованных решений в области продуктовой аналитики. Освоение этих технологий позволяет улучшить управление данными и увеличить производительность аналитических процессов.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки магистратуры по направлению 02.04.01 Математика и компьютерные науки, профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Искусственный интеллект» входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) изучается на 1 или 2 курсе, в 1, 2, 3 или 4 семестре на выбор.

Цель изучения дисциплины (модуля): формирование системного понимания архитектуры платформ данных и практических навыков проектирования, разработки и оптимизации распределённых потоков обработки и хранения данных с использованием современных инструментов инженерии данных.

Задачи изучения дисциплины (модуля):

— формирование знаний по темам: основные концепции работы платформ данных; распределенные системы хранения, основные компоненты и методы работы с объектным хранилищем (S3); основные концепции и принципы проектирования автоматизированных потоков данных (пайпланов); распределенная обработка данных с помощью Apache Spark; особенности механизмов хранения данных в ClickHouse и использование внешних источников; архитектура системы планирования потоков данных в Apache Airflow; принципы интеграции данных через брокера сообщений на примере Kafka; базовые принципы оптимизации и проверки качества данных; роль инструментов DE в создании платформы данных; частые сценарии использования инструментов в контексте платформы данных;

— освоение умений: работать с файлами в хранилище S3 через стандартные библиотеки и в сочетании с Apache Iceberg; разрабатывать ELT-процессы с использованием Apache Airflow; обрабатывать неструктурированные и структурированные данные в Apache Spark; организовывать хранение данных в распределенном режиме в кластере ClickHouse; использовать внешние источники из ClickHouse; разрабатывать запросы с учетом специфики кластера ClickHouse; отправлять и получать данные через брокер сообщений Apache Kafka; проектировать основные компоненты платформы данных; оптимизировать запросы в Apache Spark; применять базовые принципы проверки качества данных;

— приобретение навыков построения автоматизированных потоков данных с использованием Airflow, организации слоя хранения структурированных данных в Data Lakehouse, разработки витрин данных в аналитической СУБД Clickhouse.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-6.	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1.	Знает основные методы самооценки и анализа своей деятельности, а также принципы управления временем и целеполагания
		УК-6.2.	Умеет ставить реалистичные и достижимые цели, определять приоритеты в своей деятельности, а также разрабатывать и внедрять планы по совершенствованию своих навыков и компетенций на основе полученной самооценки
		УК-6.3.	Имеет практический опыт применения методов самооценки в своей профессиональной деятельности, включая участие в тренингах, семинарах и проектах, направленных на развитие личной эффективности и профессионального роста
ОПК-2.	Способен создавать и исследовать новые математические модели в естественных науках, совершенствовать и разрабатывать концепции, теории и методы	ОПК-2.1.	Знает основные математические модели и методы, используемые в естественных науках, включая статистическое моделирование, дифференциальные уравнения и численные методы, а также современные подходы к исследованию и анализу данных
		ОПК-2.2.	Умеет разрабатывать и адаптировать математические модели для решения конкретных проблем в естественных науках, проводить их анализ и верификацию, а также интерпретировать полученные результаты в контексте научных исследований
		ОПК-2.3.	Имеет практический опыт создания и исследования математических моделей в рамках научных проектов или

			исследований, включая участие в публикациях, конференциях или коллаборациях, где были разработаны и апробированы новые концепции и методы
ПК-3.	Способен решать задачи профессиональной деятельности, формулировать результат, увидеть следствия полученного результата	ПК-3.1.	Знает основные принципы и методы решения задач профессиональной деятельности, а также способы формулирования и представления результатов, включая анализ последствий и их значимость в контексте проекта
		ПК-3.2.	Умеет применять математические и компьютерные методы для решения конкретных задач, формулировать четкие и обоснованные результаты, а также анализировать их последствия для дальнейших действий и решений
		ПК-3.3.	Имеет практический опыт в решении профессиональных задач, включая участие в проектах, где были получены результаты и проанализированы их следствия, что способствовало принятию обоснованных решений
ПК-4.	Способен публично представлять собственные и известные научные результаты	ПК-4.1.	Знает основные принципы эффективного публичного выступления, методы визуализации данных и основные требования к научным презентациям, включая структуру и содержание
		ПК-4.2.	Умеет четко и логично формулировать свои научные результаты, адаптируя их для различных аудиторий, а также использовать визуальные средства для улучшения восприятия информации
		ПК-4.3.	Имеет практический опыт участия в научных конференциях, семинарах или других мероприятиях, где успешно представлял свои и известные научные результаты, получая обратную связь и взаимодействуя с аудиторией

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	Введение в инженерию данных	1	1		6	Домашние задания
2	Организация потоков данных	2	2		12	Домашние задания
3	Распределенная обработка данных	3	3		17	Домашние задания
4	Потоковая обработка данных с использованием брокера сообщений	1	1		6	Домашние задания, Контрольная работа
5	Повторение. Интеграция данных в Data Lakehouse	1	1		6	Домашние задания
6	Создание витрин в аналитических базах данных	3	3		17	Домашние задания
7	Оптимизация запросов в распределенных системах	1	1		6	Домашние задания
8	Качество данных	1	1		6	Домашние задания
9	Повторение. Построение витрины данных поверх Data Lakehouse	1	1		6	Домашние задания
	Зачет			4		
	Итого:	***14***	***14***	***4***	***82***	
Объем дисциплины (модуля) (в ак. ч.)		***114***				
Объем дисциплины (модуля) (в зач. ед.)		***3***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Введение в инженерию данных	Введение в инженерию данных. Профессия и основные концепции
2	Организация потоков данных	Оркестрация потоков данных с помощью Airflow Автоматизация процессов ETL с файловыми и реляционными источниками данных
3	Распределенная обработка данных	Использование Spark SQL для чтения и записи файлов в хранилище S3 Обработка структурированных данных и Apache Iceberg Интеграция Spark и Airflow. Обработка неструктурированных данных

4	Потоковая обработка данных с использованием брокера сообщений	Использование Apache Kafka в качестве источника данных. Потоковая обработка в Spark
5	Повторение. Интеграция данных в Data Lakehouse	Повторение. Интеграция данных в Data Lakehouse
6	Создание витрин в аналитических базах данных	Механизмы хранения и обработки данных в кластере Clickhouse Разработка запросов и обращение к внешним источникам в Clickhouse Построение агрегированных витрин с помощью материализованных представлений Clickhouse
7	Оптимизация запросов в распределенных системах	Оптимизация запросов в Apache Spark
8	Качество данных	Введение в контроль качества данных
9	Повторение. Построение витрины данных поверх Data Lakehouse	Повторение. Построение витрины данных поверх Data Lakehouse

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Карау, Х. Изучаем Spark. Молниеносный анализ данных : практическое руководство / Х. Карау, Э. Конвински, П. Венделл, З. Матей ; пер. с англ. - 2-е изд. - Москва : ДМК Пресс, 2023. - 305 с. - ISBN 978-5-89818-320-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102607.

2. Харенслак, Б. Apache Airflow и конвейеры обработки данных : практическое руководство / Б. Харенслак, Д. дэ Руйтер ; пер. с англ. Д. А. Беликова. - Москва : ДМК Пресс, 2022. - 502 с. - ISBN 978-5-97060-970-5. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2155905.

3. Apache Kafka. Потоковая обработка и анализ данных : практическое руководство / Г. Шапира, Т. Палино, Р. Сиварам, К. Петти. - 2-е изд. - Санкт-Петербург : Питер, 2023. - 512 с. - (Серия «Бестселлеры O'Reilly»). - ISBN 978-5-4461-2288-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2123357.

Дополнительная литература:

1. Карпова, И. П. Базы данных : учебное пособие / И. П. Карпова. - Санкт-Петербург : Питер, 2021. - 240 с. - (Серия «Учебное пособие»). - ISBN 978-5-4461-9681-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/1857026.

2. Кэмпбелл, Л. Базы данных. Инжиниринг надежности : практическое руководство / Л. Кэмпбелл, Ч. Мейджорс. - Санкт-Петербург : Питер, 2020. - 304 с. - (Серия «Бестселлеры O'Reilly»). - ISBN 978-5-4461-1310-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/1756135.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

— столами и стульями;

— компьютерной техникой;

— механическими калькуляторами;

— специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		

КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Data Engineering (Инженерия данных)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, практические занятия, домашние задания, контрольная работа, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Практическое занятие – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к практическому занятию рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы

избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Контрольная работа – письменная работа с набором задач, которые нужно решить за ограниченное время.

Цель контрольной работы – получить специальные знания по одной или нескольким темам дисциплины и продемонстрировать навыки их практического применения.

Бонусные баллы — это оценки, которые студенты могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет студентам улучшить общую оценку по дисциплине (модулю) и стимулирует углубленное изучение материала.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Data Engineering (Инженерия данных)»

Оценивание уровня учебных достижений обучающихся по дисциплине (модулю) осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***зачета***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Зачтено	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами
9	Отлично	Зачтено	
8	Отлично	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Зачтено	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	Зачтено	
3	Не сдан	Не зачтено	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «Data Engineering (Инженерия данных)» оценивается следующим образом:

Активность	Вес	Описание
Домашние задания	60%	Набор задач по темам недели

Активность	Вес	Описание
Контрольная работа	20%	Письменная работа с набором задач, которые нужно решить за ограниченное время
Зачет	20%	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

В рамках изучения дисциплины (модуля) возможно получение бонусных баллов.

Формула расчёта итоговой оценки по дисциплине (модулю) «Data Engineering (Инженерия данных)»: «0,6 × среднее за домашние задания + 0,2 × контрольная работа + 0,2 × зачет».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

1. Установить соединение с Hive.
2. Просмотреть список имеющихся баз.
3. Создать свою базу Hive твой_логин и переключиться на неё.
4. Изучить схему JSON и извлечь информацию о составе и типах данных.
5. Создать таблицу — приёмник для исходных данных о доставках JSON.
6. Создать таблицу данных о доставках в формате CSV для удобства последующего анализа.
7. Создать таблицу — приёмник для исходных данных о покупках JSON.
8. Создать таблицу данных о покупках в формате CSV для удобства последующего анализа.
9. Соединить с помощью запроса HQL полученные в п. 6 и п. 8 таблицы и вычислить следующие метрики:
    - количество заказов;
    - количество доставок;
    - количество доставленных единиц товара;
    - общую сумму заказов (поле cost_item).

В результате выполнения этой задачи мы получили данные в плоском виде в Hive.

1. Установи соединение с Hive.

Параметры подключения: укажи адрес хоста, порт, имя пользователя и пароль.

Создаём Курсор (объект, который позволяет выполнять SQL-запросы и управлять результатами).

Домашнее задание 2.

Задача 1. Создание и настройка сессии spark

Создай сессию spark в контуре Hadoop, добавив параметр сессии executor.memory = 2Gb (config("spark.executor.memory", "2g")).

Задача 2 (5 баллов). Чтение и запись

Прочитай из каталога /tmp/delivery_data_sample/ два JSON-файла (один файл с данными о покупках, второй файл с данными о доставках) с максимальной датой загрузки в табличном представлении. Подставь название файла вместо filename.

Задача 3 (5 баллов). Преобразование данных в spark

1. Создай временные представления purchases и deliveries.
Используй .createOrReplaceTempView.
2. Соедини оба представления по ключу с помощью PySpark DataFrame API.
Используй .join.
3. Посчитай количество записей результата соединения.
4. Соедини оба представления по ключу с помощью SparkSQL.
Используй spark.sql.
5. Посчитай количество записей результата соединения и убедись, что оно совпадает с пунктом 3.

Домашнее задание 3.

Задача 1 (3 балла). Выделить сущности в данных

Создай сессию Spark (воспользуйся кодом по созданию сессии Spark из предыдущего домашнего задания).
Создай временное представление над данными о покупках.
Выведи 10 строк данных, используя .show.
Выведи список столбцов с помощью метода .columns.
Создай временное представление над данными о доставках.
Выведи 10 строк данных, используя .show.
Выведи список столбцов с помощью метода .columns.

Задача 2 (2 балла). Нарисовать диаграмму данных

Нарисуй диаграмму полученных сущностей с атрибутами, укажи типы данных (примерные) и связи между таблицами по ключам.

Задача 3 (4 балла). Создать слой справочников и фактов

Напиши запрос Spark SQL и выбери уникальные значения атрибутов для всех сущностей.

Задача 4 (1 балл). Скопировать полученные таблицы справочников и фактов в GP

Установи соединение с Greenplum. Скопируй созданные датафреймы в таблицы Greenplum.

Примерные задания для контрольной работы

Контрольная работа

1. Даны 3 сценария:
a) ежедневная выгрузка продаж из OLTP в витрину для BI;
b) обнаружение мошенничества по транзакциям в реальном времени;
c) хранение логов приложений для последующего анализа.
Для каждого сценария:
- определите тип обработки (batch/streaming, near-real-time);
- предложите целевую архитектуру (источники → транспорт → обработка → хранилище → потребители);
- укажите 3 ключевых нефункциональных требования (SLA, качество, стоимость, безопасность и т.п.).
2. Напишите (псевдо)код DAG в Airflow для пайплайна:
extract → validate → load → notify, где:
- validate запускается только при успешном extract;

- notify запускается всегда (даже при ошибке), но в сообщении должен быть статус выполнения;

- предусмотрите расписание раз в день и ретраи.

Объясните, какие операторы/механизмы Airflow вы используете и почему (например, TriggerRule, retries).

3. DAG запускается ежедневно в 02:00 и обрабатывает данные за предыдущий день.

Ответьте:

- как использовать execution_date/logical date и шаблоны (templating), чтобы DAG всегда брал нужный день;

- что произойдет при включенном catchup=True, если DAG простаивал 10 дней;

- какие меры нужны, чтобы backfill не привел к дубликатам данных (идемпотентная загрузка).

4. Есть CSV-файлы в папке /data/events/YYYY-MM-DD/*.csv. Поля: event_id, user_id, event_ts, event_type, amount.

Нужно загрузить в реляционную БД (таблица fact_events).

- предложите стратегию инкрементальной загрузки (по дате/по watermark/по event_id);

- опишите шаги: дедупликация, валидация, обработка ошибок;

- напишите SQL DDL для fact_events с ключами/индексами, чтобы поддержать идемпотентные upsert/merge.

5. Источник — таблица orders в PostgreSQL (миллионы строк), ежедневно обновляется.

Нужно строить витрину в аналитической БД.

- какие подходы инкрементального извлечения возможны (timestamp-колонка, триггеры, CDC через WAL и т.п.);

- как обеспечить согласованность (snapshot, повторная доставка, exactly-once/at-least-once);

- какие метрики качества данных вы бы мониторили.

6. Напишите Spark SQL (или PySpark + SQL) для:

- чтения parquet из s3://bucket/raw/events/;

- фильтрации по event_date='2026-02-01';

- агрегации суммы amount по event_type;

- записи результата в s3://bucket/mart/events_daily/ в формате parquet с партиционированием по event_date.

Дополнительно: объясните, зачем партиционирование и как оно влияет на стоимость/скорость запросов.

7. Дано: на S3 лежит очень много мелких файлов (по 5–20 KB), чтение в Spark стало медленным.

- объясните причину деградации (overhead, list/GET, small files problem);

- предложите минимум 3 способа исправления (coalesce/repartition, compaction, оптимизация партиций, использование Iceberg/Delta/Hudi, настройка spark.sql.files.maxPartitionBytes и т.п.);

- как бы вы проверили улучшение (метрики/замеры).

8. Есть таблица customer_events (Iceberg) в S3. Требования:

- поддерживать upsert по event_id;

- хранить историю изменений и уметь откатываться;

- добавлять новые колонки без даунтайма.

Ответьте:

- какие возможности Iceberg это обеспечивают (snapshots, schema evolution, partition spec evolution);

- приведите пример DDL/SQL-команд: создание таблицы, добавление колонки, чтение данных “на момент времени” (time travel).

9. Нужно ежедневно обрабатывать JSON-логи (неструктурированные/полуструктурированные) из S3, извлекать поля и писать результат в Iceberg.

- предложите архитектуру DAG в Airflow: какие таски, какие зависимости, где запускать Spark (SparkSubmit/Kubernetes/EMR и т.п.);
- опишите обработку “плохих” записей (corrupt records), схему валидации и quarantine-путь;
- как обеспечить повторяемость и идемпотентность при перезапусках DAG.

10. Есть топик Kafka payments, сообщения в JSON: payment_id, user_id, ts, amount, currency.

Требуется:

- читать поток в Spark Structured Streaming;
- делать дедупликацию по payment_id в окне 1 день;
- считать сумму amount по currency в 10-минутных окна с watermark 30 минут;
- писать результат в хранилище (например, S3/Iceberg) и обеспечить отказоустойчивость.

Задание: напишите пример кода (PySpark) и объясните:

- что такое checkpoint location и зачем он нужен;
- чем отличается at-least-once от exactly-once на практике в этой схеме;
- какие настройки/подходы важны для стабильности (retries, maxOffsetsPerTrigger, watermark, backpressure).

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Для достижения поставленных перед вами целей, вам необходимо развернуть DWH. DWH - база данных, обеспечивающая хранение больших объёмов данных (как правило, сотни терабайт) и их обработку аналитическими запросами. Как расшифровывается DWH? Расшифровка содержит 2 слова, написанных латиницей.	Data Warehouse/ Data WareHouse/ data warehouse/ DataWareHouse/ DataWarehouse/	УК-6
2.	Напишите аббревиатуру (латинскими буквами) профиля нагрузки позволяющего оперативно получать в структурированном виде определённый срез из большого массива данных для их последующего анализа. Как правило его противопоставлением является OLTP.	OLAP /olap	ОПК-2
3.	Напишите название базы данных (латиницей), которая определяется следующим образом: MPP Shared Nothing-кластер, состоящий из большого числа баз данных PostgreSQL, функционирующих на большом числе серверов. Работу кластера координирует специальный экземпляр PostgreSQL — Master. Мастер-экземпляр работает на выделенном хосте, который называется мастер-хост.	Greenplum /Green plum /greenplum /green plum	ПК-3

4.	Как расшифровывается аббревиатура процесса ETL при обработке данных? Варианты ответа: 1. Export, Transfer, Load 2. Extract, Transform, Load 3. Extract, Transfer, Land В ответе укажи порядковый номер варианта ответа одной цифрой (1, 2 или 3)	2 / второй / вариант 2/ вариант №2 / №2	ПК-4
5.	Назовите метод самооценки навыков в работе с Kafka.	Тестирование	УК-6
6.	Укажите способ определения приоритетов в оптимизации качества данных.	Метрики	УК-6
7.	Назовите технику совершенствования деятельности в проектировании хранилищ данных.	Итерации	УК-6
8.	Укажите элемент самооценки в автоматизации потоков данных.	Мониторинг	УК-6
9.	Назовите тип модели для анализа временных рядов в данных.	Авторегрессия	ОПК-2
10.	Укажите метод создания модели для оптимизации хранилищ данных.	Нормализация	ОПК-2
11.	Назовите принцип проектирования масштабируемых архитектур.	Распределённость	ОПК-2
12.	Укажите технику исследования новых концепций в обработке больших данных.	Экспериментирование	ОПК-2
13.	Назовите инструмент для пакетной обработки данных.	Spark	ПК-3
14.	Укажите метод решения задач с использованием ClickHouse.	SQL-запросы	ПК-3
15.	Назовите способ проектирования витрин данных.	Звёздная схема	ПК-3
16.	Укажите технику для автоматизации ETL-процессов.	Airflow	ПК-3
17.	Назовите формат презентации результатов анализа данных.	Дашборд	ПК-4
18.	Укажите способ представления научных результатов в инженерии данных.	Доклад	ПК-4
19.	Назовите элемент эффективной презентации кейса.	Визуализация	ПК-4
20.	Укажите метод публичного выступления с результатами проекта.	Презентация	ПК-4