ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

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

Рабочая программа дисциплины (модуля)
«Инженерия данных»

Направление подготовки: *38.03.05 Бизнес-информатика*

Направленность (профиль) подготовки: *Бизнес-аналитика*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2024

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 7
5. **Учебно-методическое обеспечение** 9
6. **Материально-техническое обеспечение** 9
7. **Методические и оценочные материалы** 11

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Инженерия данных» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 38.03.05 «Бизнес-информатика», профиль «Бизнес-аналитика», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 838 от 29.07.2020 года.

Изучение дисциплины (модуля) «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 бакалавриата по направлению 38.03.05 «Бизнес-информатика», профиль «Бизнес-аналитика» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) изучается на 3 курсе в 5 семестре. Доступна к изучению после успешного освоения одной из дисциплин (модулей): «Разработка на Python. Основной», «Разработка на Python. Углубленный», «Разработка на Python. Профессиональный» или «SQL и базы данных».

Цель изучения дисциплины (модуля): формирование навыков сбора, анализа и интерпретации данных для оптимизации продуктовых решений и повышения их эффективности на рынке.

Задачи изучения дисциплины (модуля):

— освоить методы интеграции и взаимодействия различных компонентов распределённых систем хранения и обработки данных;

— научиться проектировать эффективные схемы данных с учётом требований производительности и масштабируемости;

— освоить подходы к автоматизации процессов передачи и обработки данных в корпоративных системах;

— развить навыки контроля качества данных и внедрения механизмов мониторинга в потоках данных;

— приобрести умения оптимизации вычислительных ресурсов для повышения эффективности работы систем обработки данных;

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— 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);

— 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.

уметь:

— обрабатывать файлы в хранилище s3 через стандартные библиотеки;

— разрабатывать elt-процессы;

— обрабатывать неструктурированные и структурированные данные;

— организовывать хранение и обработку данных в субд с поддержкой массивно-параллельной обработки данных;

— использовать внешние источники в запросах;

— обрабатывать потоковые данные с применением брокера сообщений;

— оптимизировать запросы;

— встраивать проверки качества данных в потоки данных.

владеть:

— навыком построения автоматизированных потоков данных;

— практикой использования data lakehouse на начальном уровне;

— способами построения витрин данных над data lakehouse;

— инструментами распределенной обработка данных.

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-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области аналитики, основные принципы критической оценки источников информации и их релевантности.
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
ОПК-2.	Способен проводить исследование и анализ рынка информационных систем и информационно-коммуникационных технологий, выбирать рациональные решения для управления бизнесом	ОПК-2.1.	Знает основные тенденции и характеристики рынка информационных систем и информационно-коммуникационных технологий
		ОПК-2.2.	Умеет проводить исследование и анализ рыночной информации для оценки потребностей бизнеса и выбора оптимальных решений
		ОПК-2.3.	Имеет практический опыт в разработке и внедрении стратегий управления бизнесом на основе анализа рынка информационных технологий
ПК-2.	Способен использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования для решения задач профессиональной деятельности	ПК-2.1.	Знает основные математические методы и инструментальные средства, применяемые для обработки и анализа информации
		ПК-2.2.	Умеет эффективно использовать математический аппарат для систематизации данных и решения профессиональных задач
		ПК-2.3.	Имеет практический опыт работы с инструментами анализа информации в рамках исследовательских проектов

ПК-3.	Способен готовить научно-технические отчеты, презентации, научные публикации по результатам выполненных исследований	ПК-3.1.	Знает требования и стандарты оформления научно-технических отчетов, презентаций и публикаций
		ПК-3.2.	Умеет структурировать и представлять результаты исследований в ясной и доступной форме
		ПК-3.3.	Имеет практический опыт подготовки и публикации научных материалов, отражающих результаты выполненных исследований
ПК-8	Способен под руководством специалиста более высокой категории осуществлять планирование и организацию проектной деятельности на основе стандартов управления проектами	ПК-8.1.	Знает принципы и стандарты управления проектами
		ПК-8.2.	Умеет разрабатывать планы и организовывать проектную деятельность в соответствии с установленными стандартами
		ПК-8.3.	Имеет практический опыт участия в проектной работе, включая планирование и координацию задач

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	2	2		8	Домашние задания Тест
2	Организация потоков данных	4	4		16	Домашние задания Тест
3	Распределённая обработка данных	6	6		26	Домашние задания Тест
4	Потоковая обработка данных	2	2		8	Домашние задания Тест
5	Повторение. Интеграция данных в Data Lakehouse	2		2	8	Домашние задания Тест Контрольная работа
6	Создание витрин в аналитических БД	6	6		26	Домашние задания Тест
7	Оптимизация запросов в распределённых системах	2	2		8	Домашние задания Тест
8	Качество данных	2	2		8	Домашние задания Тест
9	Платформы данных	4	4		18	Домашние задания Тест
	Экзамен			4		
	Итого:	***30***	***28***	***6***	***126***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

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	Распределённая обработка данных	Использование Spark SQL для чтения и записи файлов в S3 Обработка структурированных данных и Apache Iceberg Интеграция Spark и Airflow. Обработка неструктурированных данных
4	Потоковая обработка данных	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 БД	Механизмы хранения и обработки данных в кластере ClickHouse Разработка запросов и обращение к внешним источникам в ClickHouse Построение агрегированных витрин с помощью материализованных представлений
7	Оптимизация запросов в распределённых системах	Оптимизация запросов в Apache Spark
8	Качество данных	Введение в контроль качества данных
9	Платформы данных	Компоненты платформ данных. Задачи инженерии данных в платформе данных

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.

Дополнительная литература:

1. 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.

2. Лесковец, Ю. Анализ больших наборов данных : практическое руководство / Д. Дж. Ульман, Ю. Лесковец, А. Раджараман ; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. - ISBN 978-5-89818-304-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102592.

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) «Инженерия данных» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, тест, 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.

Семинар — 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 семинару 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.

Тест – особая форма проверки знаний. Проводится после освоения одной или нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины.

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины, понимать смысл научных категорий и уметь их использовать в профессиональной лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет

преподавателю быстро проверить уровень понимания студентами важных методологических категорий.

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) «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 ***экзамена***, 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, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	**Пятибалльная оценка**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. 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 исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	
8	Отлично	
7	Хорошо	Студент обладает знаниями предмета почти в полном

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
6	Хорошо	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.
5	Удовлетворительно	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	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

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

Активность	Вес	Количество	Описание
Тест	10%	13	Форма проверки знаний студентов через ответы на вопросы с выбором из нескольких вариантов ответов
Домашние задания	40%	12	В ходе дисциплины (модуля) будет предложено 8 домашних заданий, которые являются этапами единого проекта. Каждая домашняя работа оценивается по 10-балльной шкале
Контрольная работа	20%	1	Письменная работа с набором задач, которые нужно решить за ограниченное время
Экзамен	30%	1	Проект – исследовательская работа по дисциплине (модулю) и презентация результатов

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

Формула расчёта итоговой оценки по дисциплине (модулю) «Инженерия данных»: «0,1 × среднее за тесты + 0,4 × среднее за домашние задания + 0,2 × контрольная работа + 0,3 × экзамен».

Электронный документ

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.

Примерное задание для контрольной работы

Тема: Проектирование и обработка данных

Задание:

Вы работаете с данными интернет-магазина. Есть три источника:

- **users.csv** — id, name, email, registration_date
- **orders.csv** — order_id, user_id, order_date, total_sum
- **products.csv** — product_id, name, category, price

Часть 1. Моделирование данных (30 баллов)

1. Постройте ER-диаграмму (схему) хранилища данных. Укажите суррогатные ключи, связи.
2. Объясните, почему вы выбрали звездообразную или снежинковую схему.
3. Определите, какие таблицы — измерения (dimension), какие — факты (fact).

Часть 2. SQL (40 баллов)

Напишите SQL-запросы:

1. Найти 5 пользователей с наибольшей суммой заказов.
2. Посчитать количество заказов по категориям товаров за март 2024.
3. Найти пользователей, зарегистрированных в 2023, но не сделавших ни одного заказа.
4. Добавить столбец **is_premium** (1/0) в таблицу **users** — 1, если сумма заказов > 10000.

Часть 3. ETL и качество данных (30 баллов)

1. Опишите шаги ETL-процесса для загрузки этих данных в хранилище.
2. Какие проблемы с качеством данных могут быть в **orders.csv**? Приведите 3 примера.
3. Как проверить, что данные загружены без потерь?

Примерные задания для теста

№	Вопрос	Варианты	Ответ
1	Какой тип данных чаще всего используется для ID в data warehouse?	INT, VARCHAR, UUID, BIGINT	BIGINT / INT
2	Что означает ETL?	Extract, Transform, Load	☑
3	Какая схема содержит центральную fact-таблицу и surrounding dimension-таблицы?	Звездообразная	☑
4	Какой инструмент НЕ используется для оркестрации ETL?	Apache Airflow, Luigi, Jenkins, Pandas	Jenkins

№	Вопрос	Варианты	Ответ
5	Что такое суррогатный ключ?	Искусственный ID, не зависящий от бизнес-данных	☑
6	Какой формат данных лучше для быстрой аналитики в хранилище?	Parquet, CSV, JSON, TXT	Parquet
7	Какой SQL-оператор используется для агрегации?	GROUP BY	☑
8	Что проверяет тест "row count match" в ETL?	Соответствие количества строк до и после	☑
9	Какой уровень нормализации чаще используется в хранилищах данных?	Ненормализованные схемы (звезда)	☑
10	Какой инструмент подходит для потоковой обработки данных?	Apache Kafka	☑
11	Что делает пайплайн при встрече дубликата в данных?	Удаляет / оставляет / логирует	Логирует (или удаляет по правилу)
12	Как называется таблица с историческими изменениями (типа "SCD Type 2")?	Slowly Changing Dimension	☑
13	Какой SQL-запрос объединяет данные без дубликатов?	UNION	☑
14	Что такое data pipeline?	Последовательность шагов обработки данных	☑
15	Какой сервис AWS используется для хранилища данных?	Redshift	☑
16	Что означает idempotency в ETL?	Повторный запуск не создаёт дубли	☑
17	Какой формат поддерживает схему и сжатие?	Avro, Parquet	☑
18	Что делает валидация данных на этапе ingest?	Проверяет структуру и типы	☑
19	Какой инструмент используется для мониторинга пайплайнов?	Prometheus, Grafana, Airflow	☑ (любой)
20	Что такое "data lake"?	Хранилище сырых данных в любом формате	☑

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Какой метод используется для определения круга задач при создании аналитической платформы данных?	Декомпозиция цели на измеримые задачи (ETL/ELT пайплайны)	ПК-8
2.	Какой подход применяется для выбора способа обработки данных (Spark vs SQL) с учётом ресурсов и сроков?	Оценка трудозатрат, приоритизация методов и альтернатив	ПК-8
3.	Какие ограничения необходимо учитывать при сборе и хранении пользовательских данных в Data Lake?	Правовые нормы (152-ФЗ), этические требования и политики доступа	УК-1
4.	Какой принцип используется для обоснования выбора упрощённого решения (Batch vs Stream) при нехватке времени?	Баланс между точностью результата и затратами на его получение	УК-1
5.	Какой критерий определяет целесообразность реализации масштабного проекта по внедрению Data Lakehouse?	Соотношение ожидаемой пользы и затрат (cost-benefit analysis)	ОПК-2
6.	Какой показатель используется для оценки эффективности работы аналитической системы (ClickHouse)?	Производительность запросов / Время отклика / ROI инфраструктуры	ПК-2
7.	Какой принцип лежит в основе обоснования оптимизации стоимости запросов в распределённой системе?	Анализ вычислительной сложности и потребления ресурсов кластера	ПК-2
8.	Какой метод применяется для оценки устойчивости потоковой обработки данных (Kafka)?	Анализ задержек, пропускной способности и отказоустойчивости	ПК-2
9.	Какой подход используется для сравнения альтернативных стратегий интеграции данных (ETL vs ELT)?	Сравнительный анализ производительности и гибкости схем	УК-1
10.	Какой показатель отражает долгосрочную ценность данных для бизнеса (Data Retention)?	Статистическая значимость / Полезность для моделей / LTV данных	ОПК-2
11.	Какое понятие описывает затраты на вычислительные ресурсы для обработки одного запроса?	Cost Per Query / Вычислительная стоимость единицы данных	ПК-2
12.	Какой принцип позволяет снизить стоимость обработки данных без потери качества в Spark?	Минимизация shuffle / Оптимизация партиционирования / Кэширование	ПК-2
13.	Какой эффект наблюдается при снижении средних издержек при росте объёма данных в кластере?	Эффект масштаба (economies of scale) / Снижение удельной стоимости	ОПК-2
14.	Какой метод используется для определения минимального объёма данных для окупаемости хранения?	Анализ безубыточности (break-even analysis) / Точка окупаемости инфраструктуры	ОПК-2
15.	Какой инструмент применяется для прогнозирования роста объёмов данных в	Финансовое моделирование /	ПК-2

	платформе?	Трендовый анализ / Прогнозирование нагрузок	
16.	Какой российский закон регулирует обработку персональных данных в информационных системах?	Федеральный закон №152-ФЗ «О персональных данных»	УК-1
17.	Какой общий принцип должен соблюдаться при использовании данных пользователей в Kafka?	Законность, добросовестность, безопасность передачи и хранения	УК-1
18.	Какой документ требуется для легального сбора персональной информации в пайплайне?	Согласие субъекта персональных данных / Политика обработки данных	ПК-3
19.	Какие меры необходимо предпринять при передаче данных третьим лицам через S3?	Шифрование / Договор / Обеспечение уровня защиты по закону	ПК-2
20.	Какое право имеет пользователь при автоматизированном принятии решений на основе его данных в системе?	Право на объяснение решения / Доступ к данным / Исправление данных	УК-1