ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

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

Рабочая программа дисциплины (модуля)
«Big Data и Data Engineering (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)

Рабочая программа дисциплины (модуля) «Big Data и Data Engineering (Инженерия данных)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 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) «Big Data и 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 бакалавриата по направлению 38.03.05 Бизнес-информатика, профиль Бизнес-аналитика и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений.

Дисциплина (модуль) является выборной и доступна для изучения на 3 или 4 курсе в 6, 7 или 8 семестрах на выбор.

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 корпоративных хранилищ данных;

— основы распределенных систем хранения, основные компоненты и методы работы с S3;

— основные компоненты и методы работы с Apache Spark;

— основные компоненты и методы работы с ClickHouse;

— основные компоненты и методы работы с Apache Airflow;

— основные компоненты и методы работы с Kafka;

— базовые принципы оптимизации и проверки качества данных;

— основные концепции в проектировании баз данных и корпоративных хранилищ данных;

— основные концепции и принципы проектирования автоматизированных потоков данных.

уметь:

— решать задачи с использованием S3;

— решать задачи с использованием Apache Spark;

— решать задачи с использованием ClickHouse;

— решать задачи с использованием Apache Airflow;

— решать задачи с использованием Kafka;

— проектировать схемы хранения данных в базах данных и корпоративных хранилищах данных;

— применять базовые принципы оптимизации используемых ресурсов;

— применять базовые принципы проверки качества данных.

владеть:

— навыком построения автоматизированных потоков данных с использованием: S3, Kafka, Apache Spark, ClickHouse, Apache Airflow;

— навыками разработки и внедрения автоматизированного контроля качества данных;

— навыком оптимизации используемых ресурсов в автоматизированных потоках 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-3.	Способен управлять процессами создания и использования продуктов и услуг в сфере информационно-коммуникационных технологий, в том числе разрабатывать алгоритмы и программы для их практической реализации	ОПК-3.1.	Знает принципы управления процессами разработки и внедрения продуктов и услуг в сфере информационно-коммуникационных технологий
		ОПК-3.2.	Умеет разрабатывать алгоритмы и программы, обеспечивающие эффективное создание и использование информационных продуктов
		ОПК-3.3.	Имеет практический опыт в управлении проектами в области информационно-коммуникационных технологий, включая координацию команд и ресурсов для достижения поставленных целей
ОПК-4.	Способен понимать принципы работы информационных технологий; использовать информацию, методы и программные средства ее сбора, обработки и анализа для информационно-аналитической поддержки принятия управленческих решений	ОПК-4.1.	Знает основные принципы работы информационных технологий и их влияние на бизнес-процессы
		ОПК-4.2.	Умеет использовать методы и программные средства для сбора, обработки и анализа информации, обеспечивая качественную информационно-аналитическую поддержку
		ОПК-4.3.	Имеет практический опыт в применении аналитических инструментов для поддержки принятия управленческих решений в организациях
ПК-1.	Способен использовать основные методы естественнонаучных,	ПК-1.1.	Знает ключевые методы естественнонаучных, экономических и ИТ-дисциплин, применяемые в профессиональной деятельности

	экономических и ИТ-дисциплин в профессиональной деятельности для теоретического и экспериментального исследования	ПК-1.2.	Умеет интегрировать различные методологические подходы для проведения теоретических и экспериментальных исследований
		ПК-1.3.	Имеет практический опыт применения методов в реальных проектах для достижения научных и практических результатов
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-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	Основные концепции	4	4		16	Домашние задания
2	Файловые хранилища	3	3		16	Домашние задания
3	Распределённые системы вычисления	3	3		12	Домашние задания
4	Автоматизация ETL процессов	3	3		12	Домашние задания
5	Построение корпоративных хранилищ данных	3	3		16	Домашние задания
6	Clickhouse	3	3		16	Домашние задания
7	Оптимизация	3	3		16	Домашние задания
8	Качество данных	3	3		16	Домашние задания
9	Kafka	3	3		12	Домашние задания
	Экзамен			2		Проект
Итого:		***28***	***28***	***2***	***132***	
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	Основные концепции	Развитие подходов к хранению и обработке данных. Инструменты и инфраструктура платформ данных. Корпоративные хранилища данных.
2	Файловые хранилища	Данные и большие данные. Форматы файлов. Использование объектного хранилища в контексте платформы данных.
3	Распределённые системы вычисления	Отделение вычислительного слоя от слоя хранения данных. Пакетная обработка данных.
4	Автоматизация ETL процессов	Основные принципы разработки процессов загрузки данных. Построение потоков из озера данных в структурированные витрины данных.
5	Построение корпоративных хранилищ данных	Многослойная масштабируемая архитектура хранилища данных. Озера данных и Lakehouse. Проектирование витрин данных.
6	Clickhouse	Распределенные базы данных shared nothing. Механизмы хранения данных в Clickhouse.

		Применение Clickhouse для создания процессов обработки данных на SQL.
7	Оптимизация	Основы оптимизации в задачах извлечения данных. Ситуации, в которых оптимизация является необходимой. Примеры оптимизации распределенной обработки данных.
8	Качество данных	Измерения качества данных. Метрики качества данных. Интеграция правил проверки данных в процессы ETL.
9	Kafka	Архитектура брокера сообщений. Потоковая обработка 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. Чернышев, С. А. Основы программирования на Python : учебник для вузов / С. А. Чернышев. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 349 с. — (Высшее образование). — ISBN 978-5-534-17139-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/567821.

Дополнительная литература:

1. Гниденко, И. Г. Технологии и методы программирования : учебник для вузов / И. Г. Гниденко, Ф. Ф. Павлов, Д. Ю. Федоров. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 241 с. — (Высшее образование). — ISBN 978-5-534-18130-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/581329.

2. Казарин, О. В. Надежность и безопасность программного обеспечения : учебник для вузов / О. В. Казарин, И. Б. Шубинский. — 2-е изд. — Москва : Издательство Юрайт, 2025. — 352 с. — (Высшее образование). — ISBN 978-5-534-19386-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/580669.

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) «Big Data и Data Engineering (Инженерия данных)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, проект, 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.

Проект – исследовательская работа по дисциплине (модулю) и презентация результатов.

Для успешной подготовки к проекту: четко определите цели и задачи проекта, распределите роли и обязанности между участниками, а также установите сроки выполнения каждой части работы. Регулярно проводите встречи для обсуждения прогресса

и решения возникающих вопросов.

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) «Big Data и 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 ***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:

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 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
6	Хорошо	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		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) «Big Data и Data Engineering (Инженерия данных)» оценивается следующим образом:

Активность	Вес	Описание
Домашние задания	80%	В ходе дисциплины (модуля) будет предложено 8 домашних заданий, которые являются этапами единого проекта. Каждая домашняя работа оценивается по 10-балльной шкале
Экзамен	20%	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) «Data Engineering (Инженерия данных)»: «0,8 × среднее за домашние задания + 0,2 × экзамен».

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 проекта

Задание 1. Развитие ETL на базе Airflow

8 БАЛЛОВ

• изучил сырые данные о покупках, поступающих через приложение;

• преобразовал данные о покупках и доставках в удобный для хранения формат;

• спроектировал слои хранилища данных и выполнил нормализацию;

• автоматизировал ежедневную загрузку, преобразование и доставку данных до хранилища.

Коллегам из команды аналитиков понравился результат, и они обратились к тебе с новой задачей.

В исходных данных появилась информация о даркстерах, которые используются для быстрой доставки заказов покупателям. Разработчики предоставляют эту информацию в виде файлов в формате JSON, поступающих ежедневно в папку с исходными данными. Файл данных содержит сведения о номере склада, адресе склада и заказах, собранных на этом складе.

Аналитики хотят своевременно получать эти данные и формировать витрину с информацией о том, какое количество товара и на какую сумму было отгружено каждым даркстором на каждый день.

Для реализации этой задачи изучи новые данные, добавь их в модель данных, указав связи, и автоматизируй следующие ежедневные шаги в Airflow:

• преобразование нового файла данных из формата JSON в Parquet;

• выгрузку новой таблицы данных в хранилище Greenplum;

• формирование витрины по даркстерам в отдельной таблице Greenplum.

Витрина должна содержать следующие поля с агрегированной информацией о заказах:

• дата;

• наименование даркстора;

• количество заказов, которые были выполнены с помощью этого даркстора на эту дату;

• количество штук товаров в заказах, которые были выполнены с помощью этого даркстора на эту дату;

• общая сумма заказов, которые были выполнены с помощью этого даркстора на эту дату.

Критерии оценивания

1. Данные о даркстерах добавлены в диаграмму модели данных — 1 балл.
2. Реализовано преобразование исходных данных о даркстерах в виде кода на Python с использованием Spark — 2 балла.
3. Данные о даркстерах помещены в таблицу Greenplum — 1 балл.
4. Сформирована витрина с информацией о даркстерах в виде отдельной таблицы Greenplum — 0,3 балла за каждое поле в витрине, максимум 1,5 балла.

5. Шаги встроены в даги AirFlow и выполняются ежедневно по расписанию — 2,5 балла.

Задание 2. Презентация результатов

2 БАЛЛА

• Расскажи о проделанной работе на курсе по материалам домашних заданий, а также не забудь включить в свой доклад дарксторы из предыдущего задания.

• Сделай слайды pptx и подготовь видео защиты проекта (OBS/Loom/etc.) со скринкастом слайдов и устным объяснением на 5–7 минут.

• Презентация должна представлять из себя цельное и полное описание проекта, которое ты сможешь использовать в своём портфолио.

Критерии оценивания

1. Презентация содержит обязательную информацию — максимум 0,8 балла.
Обязательная информация:
• на каком проекте потребовалось участие дата-инженера — 0,1 балла;
• какие задачи были поставлены перед дата-инженером — 0,1 балла;
• как поступали данные, формат исходных данных, его особенности и недостатки — 0,1 балла;
• какие слои данных ты спроектировал для реализации потоков данных (добавь описание слоёв, включая форматы файлов, описание инструментов преобразования и способов хранения данных) — 0,2 балла;
• созданная тобой схему модели данных на слайде — 0,1 балла;
• описание и схема зависимостей реализованных дагов Airflow — 0,1 балла;
• вывод о достигнутых результатах — 0,1 балла.
2. Предоставлено видео защиты проекта — 1,2 балла.

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-1
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-3

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-1
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-1
5.	Назовите метод самооценки навыков в работе с Kafka.	Тестирование	УК-1
6.	Укажите способ определения приоритетов в оптимизации качества данных.	Метрики	УК-1
7.	Назовите технику совершенствования деятельности в проектировании хранилищ данных.	Итерации	УК-1
8.	Укажите элемент самооценки в автоматизации потоков данных.	Мониторинг	УК-1
9.	Назовите тип модели для анализа временных рядов в данных.	Авторегрессия	ОПК-3
10.	Укажите метод создания модели для оптимизации хранилищ данных.	Нормализация	ОПК-4
11.	Назовите принцип проектирования масштабируемых архитектур.	Распределённость	ОПК-3
12.	Укажите технику исследования новых концепций в обработке больших данных.	Экспериментирование	ОПК-4
13.	Назовите инструмент для пакетной обработки данных.	Spark	ПК-2
14.	Укажите метод решения задач с использованием ClickHouse.	SQL-запросы	ПК-2
15.	Назовите способ проектирования витрин данных.	Звёздная схема	ПК-2
16.	Укажите технику для автоматизации ETL-процессов.	Airflow	ПК-2
17.	Назовите формат презентации результатов анализа данных.	Дашборд	ПК-8
18.	Укажите способ представления научных результатов в инженерии данных.	Доклад	ПК-8
19.	Назовите элемент эффективной презентации кейса.	Визуализация	ПК-8
20.	Укажите метод публичного выступления с результатами проекта.	Презентация	ПК-8