ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 025053D20009B2A5B043583E315FEF31AD
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 14.10.2024 до 14.01.2026

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: *Продуктовая аналитика*

Квалификация (степень) выпускника: *магистр*

Форма обучения: *очная*

Срок освоения программы: *2 года*

Год набора: *2024*

Москва
2024

Электронный документ

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 Продуктовая аналитика, 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 Продуктовая аналитика и входит в обязательную часть Блока 1.

Дисциплина (модуль) изучается на 1 или 2 курсе во 2, 3 или 4 семестрах на выбор.

Цель изучения дисциплины (модуля): формирование у студентов знаний и навыков по проектированию, построению и обслуживанию эффективных систем сбора, хранения и обработки больших объемов данных.

Задачи изучения дисциплины (модуля):

— формирование знаний основных концепций работы корпоративных хранилищ данных;

— формирование знаний основных распределенных систем хранения, синтаксис команд, основные компоненты и методы работы с HDFS;

— формирование знаний основных компонентов и методов работы с Apache Spark;

— формирование знаний основных компонентов и методов работы с Greenplum;

— формирование знаний основных компонентов и методов работы с Apache Airflow;

— формирование знаний базовых принципов оптимизации и проверки качества данных;

— формирование знаний основных концепций в проектировании баз данных и корпоративных хранилищ данных;

— формирование знаний основных концепций и принципов проектирования автоматизированных потоков данных;

— формирование умения решать задачи с использованием HDFS и Hive, Apache Spark, Greenplum, Apache Airflow;

— формирование умения проективровать схему хранения данных в базах данных и корпоративных хранилищах данных:

— формирование умения применять базовые принципы оптимизации используемых ресурсов;

— формирование умения применять базовые принципы проверки качества данных;

— формирование умения строить автоматизированные потоки данных с использованием: HDFS, Apache Spark, Greenplum, Apache Airflow;

— формирование умения настроить автоматизированный контроль качества данных;

— формирование умения оптимизировать количество 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-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 в области продуктовой аналитики, формулировать результаты анализа и выявлять последствия полученных данных для принятия обоснованных решений и оптимизации продуктов	ПК-3.1.	Знает методы и инструменты продуктовой аналитики
		ПК-3.2.	Умеет применять аналитические инструменты и программное обеспечение для обработки и визуализации данных, а также формулировать выводы на основе проведенного анализа
		ПК-3.3.	Имеет опыт работы над реальными проектами в области продуктовой аналитики, включая анализ пользовательского поведения и оптимизацию продуктов на основе полученных данных
ПК-5.	Способен передавать результат решенных прикладных задач в виде конкретных рекомендаций, выраженных в терминах области продуктовой аналитики	ПК-5.1.	Знает основные методы и подходы к формулированию рекомендаций на основе результатов решения прикладных задач, а также термины и концепции продуктовой аналитики
		ПК-5.2.	Умеет анализировать результаты решенных задач и формулировать четкие, конкретные рекомендации, адаптируя их к требованиям и ожиданиям целевой аудитории
		ПК-5.3.	Имеет практический опыт в разработке и представлении рекомендаций на основе анализа прикладных задач, включая участие в проектах, где результаты были успешно применены и оценены в контексте предметной области

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		28	Домашние задания
2	Распределённые хранилища	4	4		28	Домашние задания
3	Распределённые системы вычисления	4	4		28	Домашние задания
4	Проектирование баз данных	6	6		28	Домашние задания
5	Автоматизация ELT-процессов	6	6		28	Домашние задания
6	Качество данных	4	4		28	Домашние задание
	Экзамен			4		Защита проекта
	Итого:	***28***	***28***	***4***	***168***	
Объем дисциплины (модуля) (в ак. ч.)		***228***				
Объем дисциплины (модуля) (в зач. ед.)		***6***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основные концепции	Вводная. Основные концепции баз данных, DWH, data lake, lakehouse
2	Распределённые хранилища	Распределённые хранилища HDFS. Концепции Hive и MapReduce
3	Распределённые системы вычисления	Введение в Apache Spark. Основы работы с PySpark
4	Проектирование баз данных	База данных Greenplum. Проектирование базы данных. Проектирование DWH
5	Автоматизация ELT-процессов	Автоматизация ELT-процессов в Airflow. Автоматизация ELT-процессов на Pyspark. Оптимизация вычислений Pyspark
6	Качество данных	Основные концепции проверки качества данных. Автоматизация процессов проверки качества 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;

— 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 к 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, 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.

Участие в семинаре (практическом занятии) – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

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.

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 ***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 (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 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 устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует
6	Хорошо	

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

Активность	Вес	Описание
Домашние задания	80%	В ходе дисциплины (модуля) будет предложено 5 домашних заданий, которые являются этапами единого проекта. Каждая домашняя работа оценивается по 10-балльной шкале
Экзамен	20%	Защита проекта оценивается по 10-балльной шкале

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 по теме «Работа с Hive»

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 по теме «Основы работы с PySpark»

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 по теме «Проектирование DWH»

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 даркскорах, 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 даркскорам 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 даркскорах добавлены в диаграмму модели данных — 1 балл.
2. Реализовано преобразование исходных данных о даркскорах в виде кода на Python с использованием Spark — 2 балла.
3. Данные о даркскорах помещены в таблицу Greenplum — 1 балл.
4. Сформирована витрина с информацией о даркскорах 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 даркскоры 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-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-5