ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

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

Рабочая программа дисциплины (модуля)
«Инженерия данных 1»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика»*

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 1» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 02.03.01 «Математика и компьютерные науки», профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Инженерия данных 1» важно для создания надёжных, масштабируемых и автоматизированных систем обработки данных, что является ключевым элементом современных аналитических и МО-платформ. Полученные знания и навыки позволяют эффективно работать с большими объёмами данных в распределённой среде и строить инфраструктуру, поддерживающую бизнес-аналитику и принятие решений.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 «Математика и компьютерные науки», профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) изучается на 3 и 4 курсе в 6 или 7 семестре на выбор. Доступна к изучению после успешного освоения одной из дисциплин (модулей): «Язык программирования Python. Основной», «Продвинутый Python» или «Базы данных».

Цель изучения дисциплины (модуля): освоить фундаментальные знания и практические навыки по проектированию, построению и управлению потоками данных в современных платформах с использованием распределённых систем хранения, инструментов автоматизации и технологий обработки структурированных и потоковых данных.

Задачи изучения дисциплины (модуля):

— освоить основы работы с распределёнными хранилищами данных, включая взаимодействие с объектным хранилищем s3 и обработку структурированных и неструктурированных данных;

— научиться проектировать и реализовывать автоматизированные elt-процессы с использованием современных подходов к интеграции и трансформации данных;

— изучить архитектуру и принципы работы систем оркестрации потоков данных на примере apache airflow;

— освоить методы обработки потоковых данных с применением брокера сообщений kafka и интеграции внешних источников в аналитические процессы;

— развить навыки оптимизации запросов, обеспечения качества данных и построения витрин данных на базе архитектуры лейкхаус данных (data lakehouse).

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 хранилища-озера на начальном уровне;
— способами построения витрин данных над хранилищем-озером;
— инструментами распределенной обработки 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 искусственного интеллекта, 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
УК-2.	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1.	Знает действующие правовые нормы, регулирующие деятельность в области решения задач, основные методы и подходы к определению круга задач
		УК-2.2.	Умеет определять круг задач в рамках поставленной цели, выбирать оптимальные способы решения задач, учитывая имеющиеся ресурсы и ограничения
		УК-2.3.	Имеет практический опыт применения знаний о правовых нормах и ресурсах в реальных ситуациях, разработки и реализации решений в соответствии с установленными ограничениями
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области

		ОПК-1.2.	Умеет анализировать и формулировать математические задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в понятной и доступной форме
		ОПК-1.3.	Имеет практический опыт работы над проектами или исследованиями в области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи
ПК-1.	Способен формулировать задачи с математической точностью, обосновывать утверждения строго и анализировать полученные результаты в области математики и компьютерных наук	ПК-1.1.	Знает методы и подходы к формулированию задач, а также основные принципы математического доказательства и анализа результатов.
		ПК-1.2.	Умеет корректно ставить и формулировать математические задачи, применять строгие методы доказательства и анализировать полученные результаты.
		ПК-1.3.	Имеет опыт работы с задачами в области математики и компьютерных наук, включая применение математических методов для решения практических задач
ПК-2.	Способен решать типовые задачи профессиональной деятельности в области искусственного интеллекта, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной безопасности	ПК-2.1.	Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности
		ПК-2.2.	Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования информационной безопасности
		ПК-2.3.	Имеет опыт работы с информационными ресурсами и технологиями в области искусственного интеллекта, включая соблюдение норм информационной безопасности

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			
1	Введение в инженерию данных	2	2		8	Домашнее задание Тест
2	Организация потоков данных	4	4		18	Домашнее задание Тест
3	Распределённая обработка данных	6	6		25	Домашнее задание Тест
4	Потоковая обработка данных	2	2		8	Домашнее задание Тест
5	Повторение. Интеграция данных в лейкхаус данных (data lakehouse)	2		2	8	Домашнее задание Контрольная работа
6	Создание витрин в аналитических БД	6	6		25	Домашнее задание Тест
7	Оптимизация запросов в распределённых системах	2	2		8	Домашнее задание Тест
8	Качество данных	2	2		8	Домашнее задание Тест
9	Платформы данных	4	4		18	Домашнее задание 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 лейкхаус данных (data lakehouse)	Интеграция данных в лейкхаус данных (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.

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

Дополнительная литература:

1. 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 1» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, тесты, 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 1»

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 1» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	40%	12	Набор задач по темам недели
Тест	10%	13	Форма проверки знаний студентов через ответы на вопросы с выбором из нескольких вариантов ответов
Контрольная работа	20%	1	Письменная работа с набором задач, которые нужно решить за ограниченное время
Экзамен	30%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

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

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

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

1. Объясните, в чём заключается роль инженера данных в современных аналитических и МО-платформах;
2. Опишите ключевые компоненты платформы данных и их взаимодействие на примере типичной архитектуры;
3. Сравните подходы к хранению данных;
4. Объясните, чем отличаются процессы ETL и ELT и в каких сценариях предпочтительнее каждый из них;
5. Приведите примеры задач, решаемых с помощью инженерии данных в реальных бизнес-кейсах.

Домашнее задание 2.

1. Объясните, как Apache Airflow оркестрирует выполнение задач в автоматизированных потоках данных;
2. Опишите структуру DAG в Airflow и назначение основных компонентов: операторов, сенсоров, хуков;
3. Реализуйте простой ETL-пайплайн в Airflow для загрузки данных из CSV-файла в реляционную СУБД;
4. Объясните, как Airflow обрабатывает зависимости между задачами и обеспечивает надёжность выполнения;
5. Проанализируйте преимущества автоматизации процессов ETL по сравнению с ручной обработкой данных.

Домашнее задание 3.

1. Объясните, как Spark SQL позволяет обрабатывать большие объёмы данных с использованием распределённых вычислений;
2. Опишите процесс чтения и записи данных в S3 с помощью Spark и особенности работы с форматами Parquet и JSON;
3. Объясните, как Apache Iceberg улучшает управление данными в лейкхаус данных по сравнению с традиционными подходами;
4. Реализуйте пример обработки структурированных данных в Spark: фильтрация, агрегация, соединение таблиц;
5. Опишите, как можно интегрировать Spark с Airflow для запуска распределённых задач в автоматизированном пайплайне.

Примерные задания для теста

Тест 1.

Вопрос 1.
Какой из перечисленных компонентов не является частью архитектуры лейкхаус данных (data lakehouse)?

A. Объектное хранилище
B. Табличный слой управления метаданными
C. Многоуровневая схема индексации в реальном времени
D. Поддержка ACID-транзакций

Ответ: C.

Вопрос 2.
Что означает аббревиатура ELT в контексте инженерии данных?

A. Extract, Load, Transform
B. Extract, Label, Transfer
C. Export, Load, Transfer
D. Export, Log, Transform

Ответ: A.

Вопрос 3.
Какой из следующих сценариев лучше всего подходит для использования озера данных (data lake)?

A. Хранение структурированных данных с жёсткой схемой для регулярной отчётности
B. Хранение неструктурированных данных с возможностью последующей обработки
C. Высокочастотные транзакции в банковской системе
D. Работа с OLTP-нагрузкой в реальном времени

Ответ: B.

Вопрос 4.
Чем принципиально отличается хранилище данных (Data warehouse) от озера данных (data lake)?

A. Хранилище данных (Data warehouse) хранит данные в бинарном формате, а озеро данных (data lake) — в текстовом
B. Озеро данных (data lake) требует схему до загрузки, а хранилище данных (Data warehouse) — после
C. Хранилище данных (Data warehouse) ориентирован на структурированные данные и аналитику, озеро данных (data lake) — на сырые и неструктурированные данные
D. Озеро данных (data lake) поддерживает только SQL-запросы, а хранилище данных (Data warehouse) — любые языки

Ответ: C.

Вопрос 5.
Какой из перечисленных инструментов чаще всего используется для централизованного хранения и каталогизации метаданных?

A. Apache Kafka
B. Apache Airflow
C. AWS Glue Data Catalog
D. Spark SQL

Ответ: C.

Тест 2.

Вопрос 1.
Что означает аббревиатура DAG в Apache Airflow?

A. Distributed Action Graph
B. Directed Acyclic Graph
C. Data Access Gateway
D. Dynamic Automation Grid

Ответ: B.

Вопрос 2.
Какой компонент Airflow отвечает за выполнение конкретной задачи в пайплайне?

A. Sensor
B. Hook
C. Operator
D. Trigger

Ответ: C.

Вопрос 3.
Какой из операторов используется для выполнения SQL-запроса к внешней базе данных?

A. PythonOperator
B. BashOperator

C. SqlSensor
D. PostgresOperator

Ответ: D.

Вопрос 4.

Что делает Sensor в Airflow?

A. Запускает внешние процессы
B. Проверяет наличие условия перед выполнением следующей задачи
C. Передаёт данные между DAG-ами
D. Логирует состояние выполнения

Ответ: B.

Вопрос 5.

Какой из подходов позволяет автоматизировать ETL-процесс из CSV-файла в PostgreSQL?

A. Использование BashOperator для копирования файла по сети
B. Настройка DAG с PythonOperator, читающим файл и загружающим данные через pandas и SQLAlchemy
C. Ручной запуск скрипта каждый день
D. Использование Kafka для потоковой передачи строки за строкой

Ответ: B.

Тест 3.

Вопрос 1.

Какой формат файлов наиболее эффективен для аналитических запросов в S3 при использовании Spark?

A. CSV
B. JSON
C. Parquet
D. TXT

Ответ: C.

Вопрос 2.

Что такое Apache Iceberg?

A. Система потоковой передачи данных
B. Формат сериализации сообщений
C. Высокоуровневый API для построения веб-интерфейсов
D. Формат таблицы с поддержкой ACID и масштабируемым метастором

Ответ: D.

Вопрос 3.

Как Spark SQL читает данные из S3?

A. Через встроенные HTTP-запросы без дополнительных библиотек
B. С помощью Hadoop FileSystem API и S3-коннектора
C. Только через Kafka Connect
D. Напрямую через браузер

Ответ: B.

Вопрос 4.

Какой из следующих шагов не является частью обработки структурированных данных в Spark?

A. Чтение данных из Parquet-файла
B. Фильтрация по условию
C. Выполнение JOIN между таблицами
D. Сохранение состояния в Redis после каждой строки

Ответ: D.

Вопрос 5.

Как можно запустить Spark-задачу в Airflow?

A. С помощью оператора SparkSubmitOperator
B. Через BashOperator с вызовом curl
C. С использованием EmailOperator
D. Через PythonOperator, имитирующий выполнение

Ответ: A.

Примерные задания для контрольной работы

1. Объясните, в чём заключается разница между data lake, data warehouse и хранилищем-озером (data lakehouse). Приведите пример сценария, где предпочтительнее использовать каждый из них.
2. Опишите основные этапы процесса ETL. В чём ключевое отличие ELT от ETL и какие преимущества оно даёт в современных платформах данных?
3. Что такое DAG в Apache Airflow? Опишите его структуру и объясните, как он обеспечивает надёжность и воспроизводимость выполнения потоков данных.
4. Реализуйте простой DAG в Airflow, который раз в день извлекает данные из CSV-файла в S3, преобразует их (например, удаляет дубликаты, фильтрует по условию) и загружает в таблицу PostgreSQL.
5. Объясните, как Spark SQL обрабатывает данные в распределённой среде. Какие компоненты Spark отвечают за выполнение SQL-запросов и управление кластером?
6. Напишите запрос на Spark SQL, который читает данные из Parquet-файла в S3, выполняет агрегацию по одному из полей (например, сумма по категории) и сохраняет результат в новую таблицу в формате Iceberg.
7. Что такое Apache Iceberg и как он решает проблемы управления большими наборами данных в data lake? Перечислите три ключевые особенности Iceberg.
8. Опишите, как интегрируются Spark и Airflow. В каких случаях Airflow передаёт управление Spark и как отслеживается статус выполнения распределённой задачи?
9. Объясните, как Apache Kafka используется в качестве источника потоковых данных. Какие компоненты Kafka обеспечивают надёжную доставку сообщений?
10. Реализуйте потоковую обработку данных в Spark Streaming: подключитесь к топику Kafka, прочитайте JSON-сообщения, извлеките нужные поля, выполните агрегацию по окну времени (например, 1 минута) и выведите результат в консоль.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Назовите ключевую концепцию, объединяющую хранение и обработку структурированных и неструктурированных данных.	Лейкхаус данных / Data lakehouse	УК-1
2	Укажите подход, при котором данные из разных источников объединяются в единый аналитический контур.	Интеграция данных	УК-1
3	Как называется процесс автоматизации последовательных этапов обработки данных?	Оркестрация / Пайплайн	УК-1
4	Назовите принцип, при котором архитектура системы рассматривается как совокупность взаимодействующих компонентов.	Системный подход	УК-1
5	Как называется документ, описывающий структуру данных и правила их обработки в проекте?	Дата-контракт / Data contract	УК-2
6	Назовите метод, позволяющий оптимизировать выполнение задач в распределённой системе.	Оптимизация запросов	УК-2
7	Укажите способ обеспечения надёжности потоков данных при сбоях.	Обработка ошибок / Fault	УК-2

		tolerance	
8	Как называется процесс проверки соответствия данных бизнес-требованиям?	Контроль качества данных / Data quality check	УК-2
9	Назовите язык, используемый для обработки структурированных данных в Spark и ClickHouse.	SQL	ОПК-1
10	Укажите математическую основу для агрегации данных в аналитических системах.	Реляционная алгебра	ОПК-1
11	Как называется метод преобразования данных при их загрузке в хранилище?	Трансформация / ETL	ОПК-1
12	Назовите формат файла, оптимизированный для аналитических запросов.	Parquet / ORC	ОПК-1
13	Назовите систему оркестрации, используемую для автоматизации ETL-процессов.	Apache Airflow	ПК-1
14	Как называется компонент Airflow, отвечающий за последовательность задач?	DAG	ПК-1
15	Назовите технологию для распределённой обработки больших данных.	Apache Spark	ПК-1
16	Как называется подход к обработке данных в реальном времени?	Потоковая обработка / Stream processing	ПК-1
17	Назовите принцип защиты данных при передаче между системами.	Шифрование / TLS	ПК-2
18	Как называется механизм, обеспечивающий безопасный доступ к данным в хранилище?	Аутентификация и авторизация / IAM	ПК-2
19	Назовите технологию, используемую для потоковой передачи данных между системами.	Apache Kafka	ПК-2
20	Назовите систему, обеспечивающую централизованное управление метаданными в платформе данных.	Data Catalog / AWS Glue / Databricks Unity Catalog	ПК-2