ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«25» мая 2026 г.
Протокол №2

Рабочая программа дисциплины (модуля)
«Методы и системы обработки больших данных»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Дизайн и разработка информационных продуктов*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2026*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

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

Электронный документ

Москва
2026

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 5
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) «Методы и системы обработки больших данных» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 02.03.01 Математика и компьютерные науки, профиль Дизайн и разработка информационных продуктов, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Методы и системы обработки больших данных» позволяет подготовить специалистов, способных проектировать и поддерживать промышленные системы обработки данных в условиях роста объёмов информации и высоких требований к её качеству и доступности. Полученные знания и навыки необходимы для эффективной работы в роли data engineer в современных IT-компаниях.

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 Дизайн и разработка информационных продуктов и входит в обязательную часть Блока 1.

Дисциплина (модуль) изучается на 3 курсе в 5 семестре.

Цель изучения дисциплины (модуля): освоить современные методы и архитектуры проектирования, реализации и эксплуатации систем обработки больших данных с учётом требований к надёжности, производительности и качеству данных.

Задачи изучения дисциплины (модуля):

- изучить принципы построения распределённых систем, включая партиционирование, репликацию и модели согласованности для обеспечения отказоустойчивости и масштабируемости;
- освоить batch и stream парадигмы обработки данных, включая понятия окон, watermark’ов и гарантий доставки сообщений;
- научиться проектировать и реализовывать надёжные пайплайны с использованием очередей событий и паттернов доставки, таких как consumer groups и offset management;
- разобрать архитектуры хранения и обработки данных (data lake, lakehouse, ETL/ELT) и средства оркестрации для построения эффективных систем данных;
- освоить практики обеспечения наблюдаемости и качества данных, включая настройку метрик, алертов, data lineage и проверок соответствия SLA/SLO.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- как Java обеспечивает кроссплатформенность: роль JVM, байткода, спецификаций и верификации;
- принципы распределённых систем для обработки данных: партиционирование, репликация, согласованность;
- batch и stream парадигмы: окна, watermark, exactly-once/at-least-once;
- очереди/шины событий и паттерны доставки: consumer groups, offset’ы, ретраи;
- архитектуры хранения и обработки: data lake/lakehouse, ELT/ETL, orchestration;
- наблюдаемость и качество данных: SLA/SLO, алерты, data quality checks, lineage;

уметь:

- проектировать пайплайны данных под требования по задержке/стоимости/надёжности;

— реализовывать и отлаживать batch/stream задачи, анализировать деградации и «затыки»;
— настраивать базовую наблюдаемость пайплайна: метрики, логи, алерты, ретраи;

владеть:

— навыками инженерного мышления в задачах обработки данных: воспроизводимость, тестируемость, эксплуатация;
— навыками декомпозиции пайплайна на компоненты и контракты данных;
— навыками подготовки технической документации на пайплайн и его SLA**.**

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
ОПК-1.	Способен консультировать и использовать фундаментальные знания в области математического анализа, комплексного и функционального анализа алгебры, аналитической геометрии, дифференциальной геометрии и топологии, дифференциальных уравнений, дискретной математики	ОПК-1.1.	Знает основные концепции и теории в области математического анализа и смежных дисциплин; методы и подходы, используемые в различных областях математики

	и математической логики, теории вероятностей, математической статистики и случайных процессов, численных методов, теоретической механики в профессиональной деятельности	ОПК-1.2.	Умеет применять математические методы для решения профессиональных задач
		ОПК-1.3.	Имеет практический опыт разработки и реализация математических моделей в профессиональной деятельности
ПК-4.	Способен разрабатывать программное обеспечение для решения прикладных задач в области профессиональной деятельности под руководством более опытного специалиста	ПК-4.1.	Знает основные принципы разработки программного обеспечения и методы решения прикладных задач в сфере профессиональной деятельности
		ПК-4.2.	Умеет разрабатывать программное обеспечение под руководством специалистов более высокой категории, используя современные технологии и инструменты
		ПК-4.3.	Имеет опыт участия в разработке программного обеспечения для решения прикладных задач в команде под руководством более опытных специалистов

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы					ТКУ (текущий контроль успеваемости)
		Очная форма					
		Контактная работа			Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)	Консультации			
1	Архитектуры обработки данных	2	2	1		7	Домашние задания
2	Партиционирование и масштабирование	2	2	1		7	Домашние задания
3	Очереди и события	2	2	1	6	7	Домашние задания Коллоквиум
4	Orchestration	2	2	1		7	Домашние задания
5	Stream processing	2	2	1		7	Домашние задания
6	Хранилища и форматы	2	2	1		7	Домашние задания
7	Качество данных	2	2	1		7	Домашние задания
8	Наблюдаемость	2	2	1		7	Домашние задания
9	Эффективность	2	2	2		7	Домашние задания
10	Надёжность	3	3	2		7	Домашние задания
11	Управление и безопасность	3	3	2		8	Домашние задания Проект
	Экзамен				2		
Итого:		***26***	***26***	***14***	***8***	***78***	
Объем дисциплины (модуля) (в ак. ч.)		***152***					
Объем дисциплины (модуля) (в зач. ед.)		***4***					

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Архитектуры обработки данных	Batch vs stream, требования, SLA, стоимость и задержка
2	Партиционирование и масштабирование	Шардинг, параллелизм, backpressure, skew
3	Очереди и события	Паттерны доставки, consumer groups, ретраи, DLQ
4	Orchestration	DAG, расписания, зависимости, идемпотентность, повторяемость
5	Stream processing	Окна, watermark, состояние, exactly-once, чекпоинты
6	Хранилища и форматы	Lake/lakehouse, форматы данных, схемы и эволюция
7	Качество данных	Валидации, тесты данных, контракт данных, мониторинг качества

8	Наблюдаемость	Метрики/логи/трейсы пайплайнов, алертинг, SLO
9	Эффективность	Оптимизация, стоимость вычислений/хранения, профилирование
10	Надёжность	Восстановление, повторный прогон, reprocessing, версионирование данных
11	Управление и безопасность	Доступы, классификация данных, аудит, lineage

5. Учебно-методическое обеспечение

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

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

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

Основная литература:

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

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

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

Дополнительная литература:

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

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

6. Материально-техническое обеспечение

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

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

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

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;

— специализированным оборудованием, включая демонстрационное оборудование.

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

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

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

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

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

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

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Методы и системы обработки больших данных» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, коллоквиумы, домашние задания, 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.

Консультации – структурированные встречи, на которых преподаватели предоставляют индивидуальную или групповую помощь в освоении учебного материала, обсуждении вопросов и решении проблем, возникающих в процессе обучения.

Консультации могут включать разъяснение сложных тем, подготовку к экзаменам и помощь в выполнении проектных работ, что способствует более глубокому пониманию предмета и улучшению академической успеваемости.

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.

Коллоквиум – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

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

Для успешной подготовки к проекту рекомендуется: четко определить цели и задачи проекта; составить план работы, разбив проект на этапы с указанием сроков выполнения каждого из них; использовать разнообразные источники информации и инструменты для исследования темы; регулярно проверять прогресс и вносить коррективы в план, если это необходимо.

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) «Методы и системы обработки больших данных»

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

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***экзамена***, при этом проводится оценка компетенций, сформированных по дисциплине. Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	**Пятибалльная оценка**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		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.
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 результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
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) «Методы и системы обработки больших данных» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	20%	7	Набор задач по темам недели
Коллоквиум	20%	1	Письменная работа с набором задач, которые нужно решить за ограниченное время
Проект	30%	1	Исследовательская работа по дисциплине (модулю) и презентация результатов
Экзамен	30%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Методы и системы обработки больших данных»: «0,2 × среднее за домашние задания + 0,2 × коллоквиум + 0,3 × проект + 0,3 × экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание по теме «Архитектуры обработки данных».

Разработайте архитектурное решение для обработки данных о транзакциях в онлайн-магазине, обоснуйте выбор между batch и stream подходом с учётом требований к задержке (SLA ≤ 5 мин), стоимости и надёжности, представьте сравнительную таблицу и рекомендации по компромиссам.

Домашнее задание по теме «Партиционирование и масштабирование».

Создайте симуляцию распределённой обработки данных (например, в Spark или Flink), продемонстрируйте эффект данных с перекосом (skew), измерьте производительность до и после применения стратегий выравнивания (repartitioning, salting), опишите влияние backpressure при высокой нагрузке.

Домашнее задание по теме «Stream processing».

Реализуйте стриминговое приложение (например, на Flink или Spark Streaming), которое подсчитывает количество уникальных пользователей за 5-минутные скользящие окна; настройте watermark для обработки запаздывающих событий, включите чекпоинты и продемонстрируйте гарантию exactly-once при перезапуске.

Домашнее задание по теме «Управление и безопасность».

Реализуйте систему управления доступом (RBAC) для data lake (например, через IAM + Lake Formation), классифицируйте данные (PII, sensitive), включите аудит доступа и настройте отслеживание lineage (с помощью OpenLineage или встроенных средств), постройте граф перемещения данных от источника до отчёта.

Примерные вопросы для коллоквиума.

1. В чём принципиальная разница между batch и stream обработкой? Приведите пример сценария, где использование стриминга оправдано, и объясните, как это влияет на SLA, стоимость и архитектурные решения.
2. Что такое data skew в распределённой обработке и как он влияет на производительность? Опишите методы борьбы с перекосом (skew) и их влияние на параллелизм и масштабируемость.
3. Объясните, как работает механизм consumer groups в Kafka. В каких случаях может потребоваться ретраи с экспоненциальным отступом и использование DLQ? Приведите пример обработки ошибок.
4. Что такое идемпотентность в контексте оркестрации задач? Почему она важна при построении надёжных пайплайнов? Приведите пример задачи, которая неидемпотентна, и предложите способ её исправления.
5. Как работают окна и watermark в стриминговых системах? Что произойдёт, если watermark установлен слишком агрессивно или слишком консервативно? Как это влияет на точность результатов?

6. В чём разница между data lake и lakehouse? Как форматы данных (например, Parquet, Delta Lake, Iceberg) влияют на возможности эволюции схем и ACID-транзакций?

7. Что входит в понятие «контракт данных»? Какие типы проверок (data quality checks) можно применить на разных этапах пайплайна, и как они помогают обеспечить доверие к данным?

8. Какие метрики вы бы мониторили в стриминговом пайплайне? Опишите, как вы организуете алертинг на основе SLO, и что делать, если пайплайн начинает нарушать заданные цели.

9. Как можно оптимизировать стоимость хранения и обработки данных в облаке? Приведите конкретные техники (например, сжатие, форматы, кластеризация) и объясните, как профилирование помогает выявить узкие места.

10. Что такое lineage данных и зачем он нужен? Как версионирование данных помогает при reprocessing? Приведите сценарий, когда восстановление пайплайна требует повторной обработки с учётом исторических версий.

Примеры заданий и критерии оценивания проекта.

Цель проекта:

Спроектировать и реализовать промышленный data pipeline, демонстрирующий комплексное применение современных методов и систем обработки больших данных, включая стриминговую и пакетную обработку, обеспечение качества, надёжности, наблюдаемости и эффективности. Проект должен отражать реальные практики инженерии данных и быть готов к эксплуатации в production-подобной среде.

Задачи и этапы выполнения проекта:

1. Формулировка сценария и требований
   Выберите реалистичный сценарий (например: аналитика пользовательских событий в мобильном приложении, мониторинг IoT-данных, обработка финансовых транзакций). Определите требования к пайплайну:
   - Задержка (SLA): batch (ежедневно) или stream (в реальном времени);
   - Объём данных;
   - Требования к надёжности, качеству и безопасности.
2. Проектирование архитектуры пайплайна
   Разработайте архитектуру, включающую:
   - Источники данных (симуляция или реальные);
   - Очереди событий (Kafka, Pulsar и т.п.);
   - Стриминговую и/или пакетную обработку (Flink, Spark, Kafka Streams);
   - Хранилище (data lakehouse: Delta Lake, Iceberg и т.п.);
   - Оркестратор (Airflow, Dagster);
   - Механизмы качества, наблюдаемости и безопасности.
3. Реализация пайплайна
   Реализуйте:
   - Генератор событий (в реальном времени или из файла);
   - Потребителя событий с обработкой ошибок (ретраи, DLQ);
   - Стриминговую агрегацию (например, количество событий в окнах с watermark);
   - Пакетный ETL/ELT-процесс с идемпотентностью;
   - Сохранение данных в формате, поддерживающем эволюцию схем и time travel.

4. Обеспечение качества и надёжности
    — Реализуйте контракт данных и валидации (на уровне схемы, диапазонов, уникальности);
    — Настройте версионирование и возможность reprocessing;
    — Обеспечьте восстановление после сбоев (чекпоинты, идемпотентность).
5. Настройка наблюдаемости и эффективности
    — Соберите метрики (потребление ресурсов, задержки, количество ошибок);
    — Настройте логирование и алертинг по SLO;
    — Проанализируйте стоимость и предложите оптимизации (форматы, партиционирование, сжатие).
6. Документирование и защита
   Подготовьте техническую документацию, включающую:
    — Архитектурную схему;
    — Описание компонентов и их взаимодействия;
    — SLA/SLO пайплайна;
    — Рекомендации по эксплуатации, безопасности и масштабированию. Защитите проект в формате презентации с демонстрацией ключевых функций.

Критерии оценивания.

1. Архитектурная целостность и обоснованность решений
   — Выбор технологий и архитектуры соответствует требованиям сценария.— Продемонстрировано понимание компромиссов (например, между задержкой и стоимостью, между надёжностью и сложностью).
2. Техническая реализация и качество кода
   — Пайплайн корректно работает, включает обработку ошибок, идемпотентность, валидации.
   — Код структурирован, воспроизводим, сопровождаем и соответствует best practices.
3. Обеспечение качества, надёжности и наблюдаемости
   — Реализованы механизмы контроля качества данных, алертинг, логирование, lineage.
   — Пайплайн устойчив к сбоям, поддерживает reprocessing и восстановление.
4. Глубина анализа и оптимизации
   — Проведён анализ эффективности (профилирование, стоимость), предложены и реализованы оптимизации.
   — Оценены и документированы SLO, метрики производительности.
5. Документация и презентация
   — Техническая документация полная, понятная, включает архитектуру, контракты, SLA.
   — Презентация логична, демонстрирует системное мышление и способность объяснять сложные решения.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Выберите тип обработки, подходящий для ежедневной агрегации логов: batch или stream?	batch	УК-1
2	Какой механизм в Kafka позволяет возобновить чтение с последнего прочитанного сообщения?	offset	УК-1
3	Какой подход к хранению данных позволяет хранить сырые данные без предварительной схемы?	data lake	УК-1
4	Как называется механизм в Airflow, описывающий зависимости между задачами?	DAG	УК-1
5	Какой уровень доставки сообщений допускает дублирование, но не потерю?	at-least-once	УК-1
6	Какой параметр определяет максимально допустимое время задержки обработки данных?	SLA	УК-1
7	Какой статистический метод используется для оценки аномалий в распределении данных?	z-оценка/z-score/метод межквартильного размаха/IQR	ОПК-1
8	Какой показатель позволяет оценить 95-й процентиль задержки обработки событий?	квантиль	ОПК-1
9	Какой закон распределения подходит для моделирования времени между событиями в потоке данных?	экспоненциальное распределение	ОПК-1
10	Какой метод применяется для оценки объёма выборки при тестировании качества данных?	статистическая выборка	ОПК-1
11	Какой алгоритм используется для равномерного распределения данных при наличии skew?	salting/рассоление	ОПК-1
12	Какой математический инструмент применяется для анализа корреляции между метриками в пайплайне?	коэффициент корреляции/корреляция Пирсона	ОПК-1
13	Какой формат данных поддерживает колоночное хранение и эффективное сжатие?	parquet/orc	ПК-4
14	Какой механизм в Flink гарантирует обработку каждого события ровно один раз?	чекпоинты/checkpoints	ПК-4
15	Какой компонент оркестратора отвечает за повторный запуск упавшей задачи?	retry/ретраи	ПК-4
16	Какой тип окна используется для подсчёта агрегатов каждые 10 минут по 5-минутным интервалам?	скользящее окно	ПК-4
17	Какой файловый формат поддерживает эволюцию схем и ACID-транзакции в lakehouse?	delta lake/iceberg/hudi	ПК-4
18	Какой механизм используется для хранения сообщений, которые не удалось обработать?	DLQ/dead letter queue	ПК-4