ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Приказом ректора
АНО ВО «Центральный университет»
от «02» июня 2026 г. № 0602.21

Рабочая программа дисциплины (модуля)
«Инженерия данных (Data Engineering)»
дополнительной профессиональной программы – программы профессиональной переподготовки «Прикладное машинное обучение»

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. **Тематический план** 5
3. **Содержание дисциплины (модуля)** 5
4. **Учебно-методическое обеспечение** 7
5. **Материально-техническое обеспечение** 7
6. **Методические и оценочные материалы** 9

1. Краткая характеристика дисциплины (модуля)

Изучение дисциплины (модуля) «Инженерия данных (Data Engineering)» 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.

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

Задачи изучения дисциплины (модуля):

- освоить методы интеграции и взаимодействия различных компонентов распределённых систем хранения и обработки данных;
- научиться проектировать эффективные схемы данных с учётом требований производительности и масштабируемости;
- освоить подходы к автоматизации процессов передачи и обработки данных в корпоративных системах;
- развить навыки контроля качества данных и внедрения механизмов мониторинга в потоках данных;
- приобрести умения оптимизации вычислительных ресурсов для повышения эффективности работы систем обработки данных.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основные концепции работы корпоративных хранилищ данных;
- основы распределенных систем хранения, основные компоненты и методы работы с S3;
- основные компоненты и методы работы с Apache Spark;
- основные компоненты и методы работы с ClickHouse;
- основные компоненты и методы работы с Apache Airflow;
- основные компоненты и методы работы с Kafka;
- базовые принципы оптимизации и проверки качества данных;
- основные концепции в проектировании баз данных и корпоративных хранилищ данных;
- основные концепции и принципы проектирования автоматизированных потоков данных.

уметь:

- решать задачи с использованием S3;
- решать задачи с использованием Apache Spark;
- решать задачи с использованием ClickHouse;
- решать задачи с использованием Apache Airflow;
- решать задачи с использованием Kafka;
- проектировать схемы хранения данных в базах данных и корпоративных хранилищах данных;
- применять базовые принципы оптимизации используемых ресурсов;
- применять базовые принципы проверки качества данных.

владеть:

- навыком построения автоматизированных потоков данных с использованием: S3, Kafka, Apache Spark, ClickHouse, Apache Airflow;
- навыками разработки и внедрения автоматизированного контроля качества данных;
- навыком оптимизации используемых ресурсов в автоматизированных потоках данных.

2. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	Введение в инженерию данных	3	3		15	Домашние задания Тесты
2	Организация потоков данных	4	4		15	Домашние задания Тесты
3	Распределённая обработка данных	5	4		12	Домашние задания Тесты
4	Потоковая обработка данных	3	3		12	Домашние задания Тесты
5	Интеграция данных в Data Lakehouse	3	3	2	15	Контрольная работа
6	Создание витрин в аналитических БД	5	4		15	Домашние задания Тесты
7	Оптимизация запросов в распределённых системах	3	3		15	Домашние задания Тесты
8	Качество данных	2	2		15	Домашние задания Тесты
9	Платформы данных	2	2		12	Домашние задания Тесты
	Экзамен			4		
	Итого:	***30***	***28***	***6***	***126***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

3. 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	Интеграция данных в Data Lakehouse	Интеграция данных в Data Lakehouse.
6	Создание витрин в аналитических БД	Механизмы хранения и обработки данных в кластере ClickHouse. Разработка запросов и обращение к внешним источникам в ClickHouse. Построение агрегированных витрин с помощью материализованных представлений.
7	Оптимизация запросов в распределённых системах	Оптимизация запросов в Apache Spark.
8	Качество данных	Введение в контроль качества данных.
9	Платформы данных	Компоненты платформ данных. Задачи инженерии данных в платформе данных.

4. Учебно-методическое обеспечение

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

Каждый слушатель 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.

Слушателям 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.

5. 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

6. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Инженерия данных (Data Engineering)» 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 слушатели 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 слушателями 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 слушатели могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет слушателям улучшить общую оценку по дисциплине (модулю) и стимулирует углубленное изучение материала.

Самостоятельная работа – работа слушателей, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы слушатели взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи слушателя 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)

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***.

Для оценивания текущего контроля успеваемости и промежуточной аттестации

используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Зачтено	Слушатель 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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	Зачтено	
8	Отлично	Зачтено	
7	Хорошо	Зачтено	Слушатель 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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Слушатель 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. Слушатель способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
3	Не сдан	Не зачтено	Слушатель не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «Инженерия данных (Data Engineering)» оценивается следующим образом, если иное не указано в силлабусе дисциплины:

Активность	Вес	Описание	Активность
Тест	10%	13	Форма проверки знаний слушателей через ответы на вопросы с выбором из нескольких вариантов ответов
Домашние задания	40%	12	Набор задач по темам недели
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 (Data Engineering)»: «0,1 × среднее за тесты + 0,4 × среднее за домашние задания + 0,2 × контрольная работа 0,3 × экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные задания для теста

Тест 1. Основы инженерии данных и оркестрация потоков

1. Какая из перечисленных задач НЕ относится к основным задачам инженера данных?
    1. Проектирование ETL-процессов
    2. Настройка оркестрации рабочих потоков
    3. Построение прогнозных моделей машинного обучения
    4. Обеспечение качества и надежности данных

    Правильный ответ: 3
2. Какой компонент Apache Airflow отвечает за выполнение конкретной задачи в DAG?
    1. Scheduler
    2. Executor
    3. Operator
    4. Worker

    Правильный ответ: 3
3. Что означает буква «T» в аббревиатуре ETL?
    1. Transfer
    2. Transform
    3. Transport
    4. Translate

    Правильный ответ: 2

4. Какой из перечисленных источников данных чаще всего используется в пакетной обработке?
    1. Apache Kafka
    2. S3
    3. RabbitMQ
    4. WebSocket

    Правильный ответ: 2
5. Какой формат данных наиболее эффективен для хранения больших объёмов структурированных данных в Data Lake?
    1. CSV
    2. JSON
    3. Parquet
    4. XML

    Правильный ответ: 3

Тест 2. Распределённая и потоковая обработка данных

1. Какой фреймворк используется для распределённой обработки больших данных?
    1. Pandas
    2. NumPy
    3. Spark
    4. Scikit-learn

    Правильный ответ: 3
2. Какой сервис чаще всего используется для потоковой передачи данных в реальном времени?
    1. S3
    2. HDFS
    3. Kafka
    4. PostgreSQL

    Правильный ответ: 3
3. Какой формат данных поддерживает схему и оптимизирован для аналитических запросов в Apache Spark?
    1. JSON
    2. Avro
    3. Parquet
    4. ORC

    Правильный ответ: 3
4. Какой компонент ClickHouse позволяет автоматически обновлять агрегированные данные при вставке новых записей?
    1. Обычное представление
    2. Материализованное представление
    3. Внешняя таблица
    4. Временная таблица

    Правильный ответ: 2
5. Что такое Data Lakehouse?
    1. Хранилище только структурированных данных
    2. Гибрид Data Lake и Data Warehouse с поддержкой ACID и аналитики

3. Инструмент визуализации данных
4. База данных для транзакций

Правильный ответ: 2

Примерные домашние задания

Домашнее задание 1.

1. Установить соединение с Hive.
2. Просмотреть список имеющихся баз.
3. Создать свою базу Hive твой_логин и переключиться на неё.
4. Изучить схему JSON и извлечь информацию о составе и типах данных.
5. Создать таблицу — приёмник для исходных данных о доставках JSON.
6. Создать таблицу данных о доставках в формате CSV для удобства последующего анализа.
7. Создать таблицу — приёмник для исходных данных о покупках JSON.
8. Создать таблицу данных о покупках в формате CSV для удобства последующего анализа.
9. Соединить с помощью запроса HQL полученные в п. 6 и п. 8 таблицы и вычислить следующие метрики:

— количество заказов;

— количество доставок;

— количество доставленных единиц товара;

— общую сумму заказов (поле cost_item).

В результате выполнения этой задачи мы получили данные в плоском виде в Hive.

1. Установи соединение с Hive.

Параметры подключения: укажи адрес хоста, порт, имя пользователя и пароль.

Создаём Курсор (объект, который позволяет выполнять SQL-запросы и управлять результатами).

Домашнее задание 2.

Задача 1. Создание и настройка сессии spark

Создай сессию spark в контуре Hadoop, добавив параметр сессии executor.memory = 2Gb (config("spark.executor.memory", "2g")).

Задача 2 (5 баллов). Чтение и запись

Прочитай из каталога /tmp/delivery_data_sample/ два JSON-файла (один файл с данными о покупках, второй файл с данными о доставках) с максимальной датой загрузки в табличном представлении. Подставь название файла вместо filename.

Задача 3 (5 баллов). Преобразование данных в spark

1. Создай временные представления purchases и deliveries.

Используй.createOrReplaceTempView.

2. Соедини оба представления по ключу с помощью PySpark DataFrame API.

Используй.join.

3. Посчитай количество записей результата соединения.
4. Соедини оба представления по ключу с помощью SparkSQL.

Используй spark.sql.

5. Посчитай количество записей результата соединения и убедись, что оно совпадает с пунктом 3.

Домашнее задание 3.

Задача 1 (3 балла). Выделить сущности в данных

Создай сессию Spark (воспользуйся кодом по созданию сессии Spark из предыдущего домашнего задания).

Создай временное представление над данными о покупках.

Выведи 10 строк данных, используя.show.

Выведи список столбцов с помощью метода.columns.

Создай временное представление над данными о доставках.

Выведи 10 строк данных, используя.show.

Выведи список столбцов с помощью метода.columns.

Задача 2 (2 балла). Нарисовать диаграмму данных

Нарисуй диаграмму полученных сущностей с атрибутами, укажи типы данных (примерные) и связи между таблицами по ключам.

Задача 3 (4 балла). Создать слой справочников и фактов

Напиши запрос Spark SQL и выбери уникальные значения атрибутов для всех сущностей.

Задача 4 (1 балл). Скопировать полученные таблицы справочников и фактов в GP

Установи соединение с Greenplum. Скопируй созданные датафреймы в таблицы Greenplum.

Примерные задания для контрольной работы

Задание 1.

Напишите SQL-запрос для создания материализованного представления в ClickHouse, которое агрегирует количество заказов по регионам из таблицы **orders** (поля: region, order_id).

Ожидаемое решение:

```
CREATE MATERIALIZED VIEW orders_by_region
ENGINE = SummingMergeTree()
ORDER BY region
AS SELECT region, count(*) AS order_count
FROM orders
GROUP BY region;
```

Задание 2.

Опишите шаги ETL-процесса для загрузки данных из CSV-файла в S3 в таблицу в ClickHouse с использованием Spark и Airflow.

Ожидаемое решение:

- Использовать Airflow DAG с PythonOperator.
- Прочитать файл из S3 с помощью Spark (format: csv).
- Преобразовать данные (очистка, приведение типов).
- Записать в ClickHouse через JDBC-соединение.
- Добавить проверку качества данных.

Задание 3.

Напишите краткое описание роли Apache Kafka в потоковой обработке данных.

Ожидаемое решение:

Kafka — это распределённая система потоков, которая позволяет публиковать и подписываться на потоки записей, хранить их с отказоустойчивостью и обрабатывать в реальном времени. Используется как буфер между источниками и обработчиками данных.