МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»

Обнинский институт атомной энергетики –

филиал федерального государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский ядерный университет «МИФИ»

(ИАТЭ НИЯУ МИФИ)

Утверждено на заседании
УМС ИАТЭ НИЯУ МИФИ
Протокол №2-8/2024 От 30.08.2024

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Clickhouse и хранилища данных DWH

Шифр, название дисциплины

01.04.02 «Прикладная математика и информатика»

Шифр, название специальности/направления подготовки

Математическое моделирование и прикладной анализ данных

Название программы магистратуры

магистр

(Квалификация (степень) выпускника)

Форма обучения: очная

г. Обнинск 2024 г.

Программа составлена в соответствии с требованиями образовательного стандарта высшего образования национального исследовательского ядерного университета «МИФИ» по направлению подготовки 01.04.02 – Прикладная математика и информатика. (квалификация (степень) магистр).

Программу составил:

_____ С.В. Ермаков, доцент, к.ф.-м.н, доцент

Рецензент:

_____ Г.Е. Деев, доцент, к.ф.-м.н, доцент

Программа рассмотрена на заседании ОИКС

(протокол № 5/7 от «30» июля от 2024 г.)

Руководитель направления подготовки 01.04.02
«Прикладная математика и информатика»

_____ Ермаков С.В.

«____»_____2024 г.

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения ООП магистратуры обучающийся должен овладеть следующими результатами обучения по дисциплине:

Коды компетенций	Результаты освоения ООП *Содержание компетенций**	Перечень планируемых результатов обучения по дисциплине**
ОПК-3	Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	З-ОПК-3 Знать основные методы и принципы математического моделирования, методы построения математических моделей типовых профессиональных задач, способы нахождения решений математических моделей и содержательной интерпретации полученных результатов. У-ОПК-3 Уметь составлять математические модели типовых профессиональных задач, находить способы их решения и профессионально интерпретировать смысл полученного результата. В-ОПК-3 Владеть методами построения математических моделей типовых профессиональных задач, способами нахождения решений математических моделей и содержательной интерпретации полученных результатов.

2. Место дисциплины в структуре ООП магистратуры

Дисциплина реализуется в рамках общенаучного модуля.

Для освоения дисциплины необходимы компетенции, сформированные в рамках изучения следующих дисциплин: SQL для анализа данных, Озера данных (data lake) и Hadoop

Дисциплина изучается на 1 курсе в 3 семестре.

3. Объем дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам занятий) и на самостоятельную работу обучающихся

Общая трудоемкость (объем) дисциплины составляет 7 зачетных единиц (з.е.), 252 академических часа.

3.1. Объём дисциплины по видам учебных занятий (в часах)

	Семестр		
	№ 1	№ 3	Всего
	Количество часов на вид работы:		
Контактная работа обучающихся с			

преподавателем			
Аудиторные занятия *(всего)*		**64**	**64**
В том числе:			
лекции		32	32
практические занятия		32	32
лабораторные занятия			
Промежуточная аттестация			
В том числе:			
зачет			
экзамен		36	**36**
Самостоятельная работа обучающихся *(всего)*		**152**	**152**
В том числе:			
проработка учебного (теоретического) материала		38	38
выполнение индивидуальных заданий		38	38
подготовка ко всем видам контрольных испытаний текущего контроля успеваемости (в течение семестра)		38	38
подготовка ко всем видам контрольных испытаний промежуточной аттестации (по окончании семестра)		38	38
Всего (часы):		**252**	**252**
Всего (зачетные единицы):		**7**	**7**

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины и трудоемкость по видам учебных занятий (в академических часах)

№ п/п	Наименование раздела /темы дисциплины	Общая трудоём-кость всего *(в часах)*	Виды учебных занятий, включая самостоятельную работу обучающихся и трудоемкость *(в часах)*				Формы текущего контроля успевае-мости
			Аудиторные учебные занятия			СРО	
			Лек	Сем/Пр	Лаб		
1.			32	32	-	152	
1.1.	Глава 1. Введение в ClickHouse	24	4	4		16	
1.2.	Глава 2. Построчные преобразования	24	4	4	-	16	
1.3.	Глава 3. Агрегация	24	4	4	-	16	
1.4.	Глава 4. Оконные функции	24	4	4	-	16	

1.5.	Глава 5. Массивы	22	3	3		16	
1.6.	Глава 6. Работа с сырыми данными	22	3	3	-	16	
1.7.	Глава 7. Соединения данных	22	3	3	-	16	
1.8.	Глава 8. Спешиал (Кейсы)	22	3	3	-	16	
1.9.	Итоговый проект	30	3	3		24	

4.2. Содержание дисциплины, структурированное по разделам (темам)

Лекционный курс

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Глава 1. Введение в ClickHouse	Введение в ClickHouse
1.2.	Глава 2. Построчные преобразования	Построчные преобразования
1.3.	Глава 3. Агрегация	Агрегация
1.4.	Глава 4. Оконные функции	Оконные функции
1.5.	Глава 5. Массивы	Массивы
1.6.	Глава 6. Работа с сырыми данными	Работа с сырыми данными
1.7.	Глава 7. Соединения данных	Соединения данных
1.8.	Глава 8. Спешиал (Кейсы)	Кейсы
1.9.	Итоговый проект	Итоговый проект

Практические/семинарские занятия

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Глава 1. Введение в ClickHouse	Введение в ClickHouse
1.2.	Глава 2. Построчные преобразования	Построчные преобразования
1.3.	Глава 3. Агрегация	Агрегация
1.4.	Глава 4. Оконные функции	Оконные функции
1.5.	Глава 5. Массивы	Массивы
1.6.	Глава 6. Работа с сырыми данными	Работа с сырыми данными
1.7.	Глава 7. Соединения данных	Соединения данных

1.8.	Глава 8. Спешиал (Кейсы)	Кейсы
1.9.	Итоговый проект	Итоговый проект

Лабораторные занятия

Не предусмотрены.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

В качестве учебно-методических материалов используется рекомендованная литература.

6. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

6.1. Паспорт фонда оценочных средств по дисциплине

№ п/п	Контролируемые разделы (темы) дисциплины (результаты по разделам)	Код контролируемой компетенции (или её части) / и ее формулировка	Наименование оценочного средства
1.3	Оконные функции	ОПК-3	Контрольная работа № 1
1.5.	Массивы	ОПК-3	Контрольная работа № 1
2.3.	Агрегация	ОПК-3	Контрольная работа № 2
2.4.	Работа с сырыми данными	ОПК-3	Контрольная работа № 2

6.2. Типовые контрольные задания или иные материалы

6.2.1. Экзамен

В экзаменационном билете два теоретических вопроса и один практический

Теоретические вопросы билета:

1. Какие особенности ClickHouse делают его подходящим для аналитической обработки больших объемов данных?
2. Каковы основные принципы работы ClickHouse и как он обеспечивает высокую производительность при обработке данных?
3. Какие преимущества ClickHouse дает при обработке данных в реальном времени и анализа больших наборов данных?
4. Что такое построчные преобразования в ClickHouse? Приведите примеры таких преобразований.
5. Как выполнить построчные преобразования с помощью SQL-запросов в ClickHouse?
6. Как работает агрегация данных в ClickHouse? Какие функции агрегации поддерживает эта СУБД?
7. В чем отличие между агрегацией в ClickHouse и агрегацией в других СУБД, например, в PostgreSQL?

8. Как использовать оконные функции в ClickHouse для выполнения аналитики по данным? Приведите примеры их применения.
9. Что такое оконные функции в ClickHouse и как они помогают анализировать данные в рамках отдельных групп?
10. Как создать и работать с массивами в ClickHouse? В чем преимущество использования массивов для хранения данных?
11. Какие типы данных используются для работы с массивами в ClickHouse и какие операции можно выполнять с ними?
12. Как ClickHouse обрабатывает "сырые" данные и что это означает с точки зрения подготовки данных для анализа?
13. Какие подходы используются для работы с неструктурированными или слабо структурированными данными в ClickHouse?
14. Как создать и выполнить SQL-запросы, работающие с сырыми данными в ClickHouse?
15. Что такое соединения данных в ClickHouse и какие виды соединений поддерживаются?
16. Как выполнить соединение таблиц в ClickHouse с использованием различных типов JOIN (INNER JOIN, LEFT JOIN и т. д.)?
17. Какие ограничения существуют при выполнении соединений в ClickHouse и как с ними бороться?
18. Как можно оптимизировать запросы в ClickHouse для работы с большими объемами данных в реальном времени?
19. Какие особенности работы с ClickHouse помогают эффективно обрабатывать и анализировать большие наборы данных с использованием разных подходов?
20. Как использовать возможности ClickHouse для реализации высокопроизводительных отчетов и дашбордов?

Критерий оценки – правильность и полнота ответа на вопросы. Оценка выставляется по шкале от 0 до 40 баллов: теоретические вопросы –30 баллов, 10 баллов– дополнительные вопросы. Экзамен считается сданным при оценке не ниже 25 баллов.

6.2.2. Контрольная работа № 1

У вас есть таблица employee_sales с данными о сотрудниках и их продажах:

employee_id	employee_name	sale_amount
1	John	500
2	Mary	700
3	Alice	800
4	Bob	600

Напишите запрос, который выводит сумму всех продаж для каждого сотрудника и добавляет колонку с процентным вкладом каждого сотрудника в общую сумму продаж (окно - по всем строкам).

6.2.2. Контрольная работа № 2

У вас есть таблица product_reviews, где каждая строка содержит список оценок продуктов:

product_id	reviews
1	[5, 4, 3, 4]
2	[4, 4, 5, 5]
3	[3, 2, 4]

Напишите SQL-запрос для вычисления средней оценки (average_review) для каждого продукта с использованием массива reviews.

б) критерии оценивания компетенций (результатов) – правильная работа кода программы, понимание алгоритма метода оптимизации, умение вывести необходимые для алгоритма формулы.

в) описание шкалы оценивания:
Каждая задача оценивается по шкале от 0 до 10 баллов.
Контрольная работа считается выполненной успешно при суммарной оценке не ниже 18 баллов.

6.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Форма аттестации	Наименование оценочного средства	Баллы
Экзамен (100 баллов)	Контрольная работа № 1	30
	Контрольная работа № 2	30
	Ответы на экзаменационный билет	40

7. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

а) основная учебная литература:

1) Распределенные данные. Алгоритмы работы современных систем хранения информации | Петров Алекс, 2021 год. Количество страниц — 288.
2) Высоконагруженные приложения. Программирование, масштабирование, поддержка" авторства Мартина Клеппмана, в 2020 год. Количество страниц — 624.

б) дополнительная учебная литература:

1) SQL для анализа данных" авторства Кэти Танимуры была издана в 2020 году. Количество страниц — 352.

8. Перечень ресурсов* информационно-телекоммуникационной сети «Интернет» (далее - сеть «Интернет»), необходимых для освоения дисциплины

-

9. Методические указания для обучающихся по освоению дисциплины

Вид учебного занятия	Организация деятельности студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначить вопросы, термины, материал, который вызывает трудности, пометить и попытаться найти ответ в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на консультации, на практическом занятии.
Практические занятия	Проработка рабочей программы, уделяя особое внимание целям и задачам, структуре и содержанию дисциплины. Работа с конспектом лекций, просмотр рекомендуемой литературы. Изучение выбранной предметной области на примерах решения задач семинарских занятий, индивидуальных домашних заданий.
Курсовая работа	Не предусмотрена
Контрольная работа	Ознакомиться с основной и дополнительной литературой, включая справочные издания, зарубежные источники, основополагающие термины. Попрактиковаться в решении аналогичных домашних задач по всем темам контрольных работ.
Лабораторная работа	Не предусмотрена.
Подготовка к экзамену	При подготовке к экзамену необходимо ориентироваться на конспекты лекций и рекомендуемую литературу.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Издательская система LaTeX для подготовки докладов, презентаций и учебного материала.

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Видеопроектор, компьютер, издательская система LaTeX для подготовки докладов, презентаций и учебного материала.

12. Иные сведения и (или) материалы

12.1. Перечень образовательных технологий, используемых при осуществлении образовательного процесса по дисциплине

Часов в интерактивной форме – 8.

В ходе практических занятий происходит публичное обсуждение каждой решаемой задачи. При этом студенты высказывают свои мнения по выбору наиболее простого способа поиска оптимального решения.

После решения домашних работ на консультациях проводится разбор допущенных студентами ошибок.

12.2. Формы организации самостоятельной работы обучающихся (темы, выносимые для самостоятельного изучения; вопросы для самоконтроля; типовые задания для самопроверки

Некоторые темы изучаются студентами самостоятельно. Для изучения используется приведённая в списке основная и дополнительная литература. Контроль освоения материала осуществляется при проверке контрольных работ, домашнего задания и на экзамене.

№	Тема и часть, изучаемая (осваиваемая) самостоятельно
1.1	Регулярные выражения
1.2	Apache Kafka. Потоковая обработка и анализ данных
1.3	Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение

Вопросы и задания для самоконтроля по всем темам:

1. Что такое ClickHouse и чем он отличается от других СУБД, таких как PostgreSQL или MySQL?
2. Какие преимущества предоставляет ClickHouse при работе с большими объемами данных в реальном времени?
3. Каковы основные принципы архитектуры ClickHouse, обеспечивающие его высокую производительность?
4. Какие особенности синтаксиса SQL в ClickHouse необходимо учитывать при написании запросов?
5. Что такое построчные преобразования в ClickHouse? Приведите примеры таких преобразований.
6. Как выполнить построчные преобразования в ClickHouse с использованием SQL? Какие функции для этого используются?
7. Что такое фильтрация данных в запросах и как с её помощью можно оптимизировать выполнение построчных преобразований в ClickHouse?
8. Как можно применить математические или строковые функции в ClickHouse для выполнения построчных преобразований?
9. Как происходит агрегация данных в ClickHouse? Перечислите и объясните основные функции агрегации.
10. Как в ClickHouse можно агрегистрировать данные с использованием группировки по ключам?

12.3. Краткий терминологический словарь

Агрегация	процесс объединения данных, например, вычисление сумм, средних значений, минимумов и максимумов по группам данных.
Уникаль	уникальные значения, которые помогают идентифицировать строки или записи в

ные идентификаторы	таблицах, часто используются для объединения таблиц
Колоночное хранение	способ хранения данных, при котором значения для каждого столбца таблицы хранятся отдельно, что ускоряет выполнение аналитических запросов в больших объемах данных