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

Озера данных (data lake) и Hadoop
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:

с	**Результаты освоения ООП** ***Содержание компетенций****	**Перечень планируемых результатов обучения по дисциплине****
ПК-5	способен четко формулировать цели и задачи научно-прикладных проектов, разрабатывать концептуальные и теоретические модели решаемых задач	З-ПК-5 Знать основные цели и задачи научно-прикладных проектов, разрабатывать концептуальные и теоретические модели решаемых задач. У-ПК-5 Уметь четко формулировать цели и задачи научно- прикладных проектов, разрабатывать концептуальные и теоретические модели решаемых задач В-ПК-5 Владеть навыками разработки теоретических моделей решаемых задач.
ПК-6	способен к проектированию и разработке наукоемкого программного обеспечения на основе технического задания	З-ПК-6 Знать основные цели и задачи проектирования и разработки наукоемкого программного обеспечения на основе технического задания. У-ПК-6 Уметь разрабатывать наукоемкое программное обеспечение на основе технического задания. В-ПК-6 Владеть навыками разработки и проектирования наукоемкого программного обеспечения на основе технического задания.

2. Место дисциплины в структуре ООП магистратуры

Дисциплина реализуется в рамках вариативной части.

Для освоения дисциплины необходимы компетенции, сформированные в рамках изучения следующих дисциплин: Clickhouse и хранилища данных DWH, Обработка больших данных с помощью Spark

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.	Введение в Big Data, Data Lake vs Data Warehouse и Data Virtualization	18	4	4		20	
1.2.	Основы HDFS. Развертывание и использование HDFS с	18	4	4	-	20	

	помощью Docker						
1.3.	Работа с файловой системой HDFS	17	4	4	-	18	
1.4.	Знакомство с парадигмой MapReduce. Hadoop Mapreduce. Hadoop streaming.	17	4	4	-	18	
1.5.	Основы YARN	17	4	4		18	
1.6.	Основы HBase. Знакомство с Pig/Hive	17	4	4	-	18	
1.7.	Hadoop MapReduce* - больше практических задач на MapReduce, базовые задачи на Pig/Hive	17	4	4	-	18	
1.8.	Итоговый проект	20	4	4		22	

4.2. Содержание дисциплины, структурированное по разделам (темам)

Лекционный курс

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Введение в Big Data, Data Lake vs Data Warehouse и Data Virtualization	Основы концепции Big Data: определение, особенности и проблемы Различия между Data Lake и Data Warehouse: когда использовать каждую из технологий Принципы Data Virtualization и её роль в интеграции данных
1.2.	Основы HDFS. Развертывание и использование HDFS с помощью Docker	Введение в Hadoop Distributed File System (HDFS) Архитектура HDFS и основные компоненты Развертывание HDFS в контейнерах Docker Практические шаги по настройке HDFS в Docker
1.3.	Работа с файловой системой HDFS	Основные команды для работы с HDFS: hdfs dfs Загрузка, чтение и удаление файлов в HDFS Управление правами доступа и проверка состояния файловой системы
1.4.	Знакомство с парадигмой MapReduce. Hadoop Mapreduce. Hadoop streaming.	Принципы парадигмы MapReduce для обработки больших данных Основы Hadoop MapReduce: структура и работа с задачами Map и Reduce Использование Hadoop Streaming для выполнения MapReduce задач с использованием языков, таких как Python или Ruby
1.5.	Основы YARN	Введение в Yet Another Resource Negotiator (YARN) и его роль в Hadoop Основные компоненты YARN: ResourceManager, NodeManager, ApplicationMaster

		Ресурсное управление и распределение нагрузки с помощью YARN
1.6.	Основы HBase. Знакомство с Pig/Hive	Введение в HBase: работа с неструктурированными данными, распределённая база данных Основы работы с HBase: создание таблиц, чтение и запись данных Знакомство с Pig: язык скриптов для обработки данных в Hadoop Знакомство с Hive: SQL-подобный интерфейс для обработки больших данных
1.7.	Hadoop MapReduce* - больше практических задач на MapReduce, базовые задачи на Pig/Hive	Решение практических задач с использованием Hadoop MapReduce Основы оптимизации MapReduce задач Применение Pig и Hive для обработки данных в Hadoop: примеры и практическое применение
1.8.	Итоговый проект	Итоговый проект

Практические/семинарские занятия

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Введение в Big Data, Data Lake vs Data Warehouse и Data Virtualization	Основы концепции Big Data: определение, особенности и проблемы Различия между Data Lake и Data Warehouse: когда использовать каждую из технологий Принципы Data Virtualization и её роль в интеграции данных
1.2.	Основы HDFS. Развертывание и использование HDFS с помощью Docker	Введение в Hadoop Distributed File System (HDFS) Архитектура HDFS и основные компоненты Развертывание HDFS в контейнерах Docker Практические шаги по настройке HDFS в Docker
1.3.	Работа с файловой системой HDFS	Основные команды для работы с HDFS: hdfs dfs Загрузка, чтение и удаление файлов в HDFS Управление правами доступа и проверка состояния файловой системы
1.4.	Знакомство с парадигмой MapReduce. Hadoop Mapreduce. Hadoop streaming.	Принципы парадигмы MapReduce для обработки больших данных Основы Hadoop MapReduce: структура и работа с задачами Map и Reduce Использование Hadoop Streaming для выполнения MapReduce задач с использованием языков, таких как Python или Ruby
1.5.	Основы YARN	Введение в Yet Another Resource Negotiator (YARN) и его роль в Hadoop Основные компоненты YARN: ResourceManager,

		NodeManager, ApplicationMaster Ресурсное управление и распределение нагрузки с помощью YARN
1.6.	Основы HBase. Знакомство с Pig/Hive	Введение в HBase: работа с неструктурированными данными, распределённая база данных Основы работы с HBase: создание таблиц, чтение и запись данных Знакомство с Pig: язык скриптов для обработки данных в Hadoop Знакомство с Hive: SQL-подобный интерфейс для обработки больших данных
1.7.	Hadoop MapReduce* - больше практических задач на MapReduce, базовые задачи на Pig/Hive	Решение практических задач с использованием Hadoop MapReduce Основы оптимизации MapReduce задач Применение Pig и Hive для обработки данных в Hadoop: примеры и практическое применение
1.8.	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	Работа с файловой системой HDFS	ПК-5	Контрольная работа № 1
1.4.	Hadoop Mapreduce. Hadoop streaming.	ПК-6	Контрольная работа № 1
1.6.	Основы HBase	ПК-5	Контрольная работа № 2
1.7.	Базовые задачи на Pig/Hive	ПК-6	Контрольная работа № 2

6.2. Типовые контрольные задания или иные материалы

6.2.1. Зачет

В билете два теоретических вопроса и один практический

Теоретические вопросы билета:

1. Что такое Big Data и какие основные характеристики данных, относящихся к этому понятию?
2. В чем разница между Data Lake и Data Warehouse, и в каких сценариях лучше использовать каждую из этих архитектур?
3. Что такое Data Virtualization и какие преимущества она предоставляет для работы с данными?
4. Каковы основные компоненты HDFS и как они взаимодействуют между собой?
5. Какие шаги необходимо предпринять для развертывания и использования HDFS с помощью Docker?
6. Как осуществлять работу с файловой системой HDFS и какие команды для этого используются?
7. Что такое парадигма MapReduce и как она применяется в Hadoop?
8. Какова роль Hadoop MapReduce в обработке больших данных и какие основные этапы этого процесса?
9. Что такое Hadoop Streaming и как он позволяет использовать языки программирования, отличные от Java, в MapReduce?
10. Какова основная архитектура YARN и какие функции она выполняет в экосистеме Hadoop?
11. Что такое HBase и какие преимущества она предоставляет по сравнению с традиционными реляционными базами данных?
12. Каковы основные функции Pig и Hive, и как они помогают в обработке данных в Hadoop?

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. Зачет считается сданным при оценке не ниже 25 баллов.

6.2.2. Контрольная работа № 1

Датасет. Любые логи, у которых установлена дата записи.
Архивирование старых данных (сказать, что разные папки могут быть примонтированы на разных дисках. Упомянуть, за создатель Перов Виктор чем такое нужно).
Необходимо написать программу, которая автоматически архивирует данные старше определенной даты в HDFS в отдельную директорию.
В директории /logs/ хранятся файлы логов. Программа должна переместить все файлы старше одного года в директорию /archive/logs.

6.2.2. Контрольная работа № 2

Мониторинг использования дискового пространства в HDFS.
Нужно написать скрипт для мониторинга HDFS (с помощью утилиты hdfs dfs -df). Может быть выводить графики. (* - выводить алерты куда-нибудь).
Возможно, проверка состояния HDFS.

```
fs = hdfs.connect(host='localhost', port=9870)
# Проверка состояния HDFS
def check_hdfs_status():
status = fs.df()
print(status)
```

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
Зачет (100 баллов)	Контрольная работа № 1	30
	Контрольная работа № 2	30
	Ответы на билет	40

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

1. А. И. Костров — "Hadoop. Разработка приложений" — 2020 — 400 с.
2. С. Г. Дьяков — "Big Data: архитектура и технологии" — 2021 — 350 с.
3. Е. В. Неверов — "Apache Hadoop и MapReduce" — 2019 — 320 с.
4. Д. Н. Лапшин — "Hadoop. Учебное пособие" — 2020 — 310 с.
5. С. А. Рубцов — "YARN: Основы и практика" — 2021 — 240 с.
6. А. И. К. — "Виртуализация данных: концепции и технологии" — 2022 — 330 с.

б) дополнительная учебная литература:

1. Н. С. Ермолаев — "Hadoop для аналитиков: от HDFS до HBase" — 2020 — 300 с.
2. А. А. Баранов — "Pig и Hive: Применение в Big Data" — 2019 — 280 с.
3. В. К. Степанов — "Основы работы с HDFS" — 2021 — 270 с.
4. И. В. Поляков — "Введение в Big Data и Data Lake" — 2022 — 290с.

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 зачету	При подготовке к зачету 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 зачете.

№	Тема и часть, изучаемая (осваиваемая) самостоятельно
1.1	Обработка потоковых данных с использованием Apache Kafka
1.2	Введение в машинное обучение с использованием Apache Spark
1.3	Эффективные методы хранения данных: NoSQL vs. реляционные базы данных
1.4	Интеграция Hadoop с облачными платформами (AWS, Google Cloud, Azure)
1.5	Оптимизация производительности MapReduce задач
1.6	Модели данных в Data Lake и Data Warehouse
1.7	Сравнение технологий ETL и ELT в обработке данных
1.8	Использование Apache NiFi для управления потоками данных
1.9	Парадигмы работы с графами: GraphX и Neo4j
1.10	Роль искусственного интеллекта в обработке больших данных

Вопросы и задания для самоконтроля по всем темам:

1. Что такое Big Data и какие ключевые характеристики определяют данные как "большие"?
2. В чем основные отличия между Data Lake и Data Warehouse?
3. Какие преимущества и недостатки имеет Data Virtualization в контексте работы с большими данными?
4. Каковы ключевые компоненты HDFS и как они обеспечивают надежность и доступность данных?
5. Как установить и настроить HDFS с помощью Docker? Какие основные команды используются для работы с HDFS?
6. В чем заключается парадигма MapReduce и какие этапы обработки данных в этой модели?
7. Как работает Hadoop MapReduce и каковы его основные функции?
8. Что такое Hadoop Streaming и какие языки программирования могут использоваться с ним?
9. Каковы основные функции YARN и как она управляет ресурсами в экосистеме Hadoop?
10. Что такое HBase и в чем её основные преимущества по сравнению с реляционными базами данных?
11. Каковы основные функции и преимущества использования Apache Pig и Apache Hive для обработки данных?

12.3. Краткий терминологический словарь

Data Lake	это хранилище данных, которое позволяет сохранять структурированные, полуструктурированные и неструктурированные данные в их исходном виде, обеспечивая гибкость для будущего анализа.
HDFS	(Hadoop Distributed File System) – распределенная файловая система, используемая в экосистеме Hadoop, обеспечивающая хранение больших объемов данных с высокой надежностью и доступностью.
MapReduce	это парадигма программирования для обработки и генерации больших наборов данных с использованием параллельного и распределенного алгоритма на кластерах, состоящая из двух основных этапов: Map и Reduce.