

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

Обработка больших данных с помощью Spark

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 <i>Содержание компетенций*</i>	Перечень планируемых результатов обучения по дисциплине**
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	З-УК-1 Знать: методы системного и критического анализа; методики разработки стратегии действий для выявления и решения проблемной ситуации У-УК-1 Уметь: применять методы системного подхода и критического анализа проблемных ситуаций; разрабатывать стратегию действий, принимать конкретные решения для ее реализации В-УК-1 Владеть: методологией системного и критического анализа проблемных ситуаций; методиками постановки цели, определения способов ее достижения, разработки стратегий действий.
ПК-3	способен развивать инновационный потенциал новых научных и научно-технологических разработок	З-ПК-3 Знать основы планирования и организации научных исследований в профессиональной области; методику постановки задач по решению теоретических и прикладных исследовательских проблем; методы и средства научных исследований в профессиональной области, правила и принципы научной этики, методы математического моделирования. У-ПК-3 Уметь оценивать и развивать инновационный потенциал новых научных и научно-технологических разработок, осуществлять постановку задач по решению теоретических и прикладных исследовательских проблем; составить план научных исследований; выдвинуть гипотезы по направлению исследований и соотнести их с полученными результатами; организовать свою научно-исследовательскую работу; определять методы и средства научных исследований для решения конкретных задач в своей предметной области; оценивать результаты исследований, использовать методы математического моделирования В-ПК-3 Владеть навыками постановки

		задач по решению теоретических и прикладных исследовательских проблем; навыками выбора и использования методов и средств научных исследований задач в своей предметной области; навыками методами работы с литературными источниками; методами анализа результатов научных исследований; методами обобщения результатов научных исследований для развития инновационного потенциала новых научных и научно- технологических разработок
--	--	--

2. Место дисциплины в структуре ООП магистратуры

Дисциплина реализуется в рамках вариативной части.

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

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. Объем 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:			
<i>лекции</i>		32	32
<i>практические занятия</i>		32	32
<i>лабораторные занятия</i>			
Промежуточная аттестация			
В том числе:			
<i>зачет</i>			
<i>экзамен</i>		36	36
Самостоятельная работа обучающихся (всего)		152	152
В том числе:			
<i>проработка учебного (теоретического) материала</i>		38	38
<i>выполнение индивидуальных заданий</i>		38	38

<i>подготовка ко всем видам контрольных испытаний текущего контроля успеваемости (в течение семестра)</i>		38	38
<i>подготовка ко всем видам контрольных испытаний промежуточной аттестации (по окончании семестра)</i>		38	38
<i>Всего (часы):</i>		252	252
<i>Всего (зачетные единицы):</i>		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 трудоем- 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 Apache Spark. Архитектура. Сравнение с Hadoop.	18	4	4		20	
1.2.	Настройка и установка. Pandas, или PySpark в Jupiter Notebook? SparkSession	18	4	4	-	20	
1.3.	Введение в функциональное программирование	17	4	4	-	18	
1.4.	Spark core. RDD. DataFrame	17	4	4	-	18	
1.5.	Spark SQL. Spark UDF.	17	4	4		18	
1.6.	Spark Streaming. Spark Structured Streaming.	17	4	4	-	18	
1.7.	Работа с внешними данными.	17	4	4	-	18	
1.8.	Проектное задание. Обработка данных из различных источников на Spark.	20	4	4		22	

4.2. Содержание дисциплины, структурированное по разделам (темам)

Лекционный курс

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Введение в Apache Spark. Архитектура. Сравнение с Hadoop.	Основы Apache Spark и его назначение. Архитектура Spark: компоненты и их взаимодействие (Driver, Executor, Cluster Manager). Сравнение Spark и Hadoop: преимущества Spark (in-memory вычисления) и отличие от Hadoop MapReduce.
1.2.	Настройка и установка. Pandas, или PySpark в Jupiter Notebook? SparkSession	Установка Apache Spark и настройка окружения. Использование PySpark в Jupiter Notebook для работы с данными. Создание и настройка SparkSession для работы с DataFrame и SQL.
1.3.	Введение в функциональное программирование	Основы функционального программирования (чистые функции, неизменяемость, функции высшего порядка). Применение функциональных концепций в обработке данных в Spark.
1.4.	Spark core. RDD. DataFrame	Как работает RDD. Обертка в виде DataFrame. Промежуточные/терминальные операции.
1.5.	Spark SQL. Spark UDF.	Использование Spark SQL для обработки данных в Spark. Создание и использование пользовательских функций (UDF) в Spark SQL и DataFrame.
1.6.	Spark Streaming. Spark Structured Streaming.	Введение в потоковую обработку данных с использованием Spark Streaming. Различия между Spark Streaming и Spark Structured Streaming. Работа с непрерывной потоковой обработкой данных (Structured Streaming).
1.7.	Работа с внешними данными.	Чтение и запись данных в различных форматах (CSV, JSON, Parquet, Avro) и из разных источников (HDFS, S3, базы данных). Интеграция Spark с внешними хранилищами данных (Cassandra, HBase).
1.8.	Проектное задание. Обработка данных из различных источников на Spark.	Разработка проекта по обработке данных из нескольких источников (файлы, базы данных, потоки). Очистка, агрегация и анализ данных с использованием Spark. Оптимизация работы с данными и применение методов

		машинного обучения.
--	--	---------------------

Практические/семинарские занятия

№	Наименование раздела /темы дисциплины	Содержание
1.1.	Введение в Apache Spark. Архитектура. Сравнение с Hadoop.	Основы Apache Spark и его назначение. Архитектура Spark: компоненты и их взаимодействие (Driver, Executor, Cluster Manager). Сравнение Spark и Hadoop: преимущества Spark (in-memory вычисления) и отличие от Hadoop MapReduce.
1.2.	Настройка и установка. Pandas, или PySpark в Jupiter Notebook? SparkSession	Установка Apache Spark и настройка окружения. Использование PySpark в Jupiter Notebook для работы с данными. Создание и настройка SparkSession для работы с DataFrame и SQL.
1.3.	Введение в функциональное программирование	Основы функционального программирования (чистые функции, неизменяемость, функции высшего порядка). Применение функциональных концепций в обработке данных в Spark.
1.4.	Spark core. RDD. DataFrame	Как работает RDD. Обертка в виде DataFrame. Промежуточные/терминальные операции.
1.5.	Spark SQL. Spark UDF.	Использование Spark SQL для обработки данных в Spark. Создание и использование пользовательских функций (UDF) в Spark SQL и DataFrame.
1.6.	Spark Streaming. Spark Structured Streaming.	Введение в потоковую обработку данных с использованием Spark Streaming. Различия между Spark Streaming и Spark Structured Streaming. Работа с непрерывной потоковой обработкой данных (Structured Streaming).
1.7.	Работа с внешними данными.	Чтение и запись данных в различных форматах (CSV, JSON, Parquet, Avro) и из разных источников (HDFS, S3, базы данных). Интеграция Spark с внешними хранилищами данных (Cassandra, HBase).
1.8.	Проектное задание. Обработка данных из различных источников на Spark.	Разработка проекта по обработке данных из нескольких источников (файлы, базы данных, потоки). Очистка, агрегация и анализ данных с использованием Spark.

		Оптимизация работы с данными и применение методов машинного обучения.
--	--	---

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	функциональное программирование	УК-1	Контрольная работа № 1
1.4.	Spark core. RDD. DataFrame	ПК-3	Контрольная работа № 1
1.7.	Работа с внешними данными	УК-1	Контрольная работа № 2
1.8.	Обработка данных из различных источников на Spark	ПК-3	Контрольная работа № 2

6.2. Типовые контрольные задания или иные материалы

6.2.1 Экзамен

В билете два теоретических вопроса и один практический

Теоретические вопросы билета:

Теоретические вопросы билета:

1. Какова архитектура Apache Spark и чем она отличается от Hadoop?
2. Какие шаги необходимо предпринять для настройки и установки Apache Spark? Как использовать Pandalas или PySpark в Jupyter Notebook?
3. Что такое функциональное программирование и как оно может быть применено в Spark?
4. Объясните концепцию RDD и DataFrame в Spark. В чем их отличия и преимущества?
5. Как работает Spark SQL и какие возможности он предоставляет для обработки данных?
6. Что такое пользовательские функции (UDF) в Spark и как их можно использовать?
7. В чем разница между Spark Streaming и Spark Structured Streaming? Когда стоит использовать каждую из них?
8. Что такое GraphX и GraphFrame, и каковы их основные применения?
9. Как осуществляется работа с внешними данными в Apache Spark?

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. Данные об активности пользователей (на сайте/где-нибудь еще) в формате .csv (user_id, session_duration, activity_start_date). Данные будут сгенерированы скриптом.

Нужно агрегировать по общему времени, проведенным пользователем за последние полгода (к примеру), и затем найти самых неактивных пользователей (к примеру, если пользователь провел на сайте менее 10 минут за полгода).

6.2.2. Контрольная работа № 2

Датасет. Данные об оценках учеников в формате .csv (student_id, subject, grade(0-100)). Датасет будет сгенерированным по нормальному распределению со смещением. udf:

```
def classify_grade(avg_grade):  
if avg_grade >= 85: return "High"  
elif avg_grade >= 70: return "Medium"  
else: return "Low"
```

Используя заранее написанную функцию, применить её с помощью df.withColumn(), классифицировать студентов по трем уровням успеваемости ('high', 'medium', 'low')

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

а) основная учебная литература:

1. С. Г. Дьяков — "Анализ данных с помощью Apache Spark" — 2020 — 320 с.
2. А. И. Хватов — "Apache Spark для профессионалов" — 2019 — 400 с.
3. Е. В. Неверов — "Spark и Hadoop. Обработка больших данных" — 2020 — 350 с.
4. А. А. Баранов — "Параллельные вычисления в Apache Spark" — 2022 — 240 с.
5. Билл Чианг — "Learning Spark: Lightning-Fast Data Analytics" — 2020 — 336 с.
6. Элиас Д. и С. Хан — "Apache Spark for Data Science" — 2021 — 300 с.

б) дополнительная учебная литература:

1. С. А. Рубцов — "Spark в действии" — 2021 — 380 с.
2. В. К. Степанов — "Функциональное программирование в Apache Spark" — 2021 — 300 с.
3. Н. С. Ермолаев — "Построение распределенных приложений с Apache Spark" — 2022 — 290 с.
4. Д. Н. Лапшин — "Проектирование приложений на Apache Spark" — 2020 — 310 с.

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; пометить 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 удастся 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 больших данных с использованием Apache Flink
1.2	Параллельные вычисления в Python с использованием Dask
1.3	Анализ потоковых данных с использованием Apache Kafka
1.4	Машинное обучение в Apache Spark с использованием MLlib
1.5	Облачные платформы для обработки данных: AWS, Google Cloud, Azure
1.6	Оптимизация производительности Spark-приложений
1.7	Хранение и управление данными в Hadoop HDFS
1.8	Сравнение архитектур обработки данных: Spark, Flink и Storm
1.9	Интеграция Spark с базами данных: PostgreSQL, MongoDB, Cassandra
1.10	Визуализация данных после обработки в Spark: инструменты и подходы

Вопросы и задания для самоконтроля по всем темам:

1. Какова основная архитектура Apache Spark и какие компоненты она включает?
2. Какие шаги необходимы для настройки и установки Apache Spark, и как запустить SparkSession в Jupyter Notebook?
3. Каковы ключевые принципы функционального программирования и как они применяются в контексте Apache Spark?
4. В чем разница между RDD и DataFrame в Spark, и когда следует использовать каждую из этих структур?
5. Как использовать Spark SQL для выполнения запросов к данным, и какие преимущества он предоставляет по сравнению с обычным SQL?
6. Как создаются и используются пользовательские функции (UDF) в Spark SQL?
7. Что такое Spark Streaming и Spark Structured Streaming, и каковы их основные применения?
8. Каковы особенности GraphX и GraphFrame, и в каких сценариях их следует использовать?

9. Какие методы работы с внешними данными поддерживаются в Apache Spark?
10. Как можно организовать проект по обработке данных из различных источников на Apache Spark, и какие инструменты могут быть использованы?

12.3. Краткий терминологический словарь

Apache Spark	открытая распределённая платформа для обработки больших данных, которая позволяет выполнять параллельные вычисления и предоставляет богатый набор инструментов для анализа данных.
Архитектура	структура и организация компонентов системы, определяющие взаимодействие между ними и их функции, в случае Apache Spark это включает такие элементы, как драйвер, рабочие узлы и кластер.
Hadoop	фреймворк для хранения и обработки больших данных, который использует распределённое хранение (HDFS) и распределённые вычисления (MapReduce).
Pandas	библиотека Python для анализа и обработки данных, которая предоставляет удобные структуры данных и инструменты для манипуляции и анализа.
PySpark	интерфейс для работы с Apache Spark на языке Python, который позволяет использовать функциональность Spark через синтаксис Python.
RDD	(Resilient Distributed Dataset) основная структура данных в Apache Spark, представляющая собой распределённый набор данных, который может быть обработан параллельно.