

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО
Заместитель директора

Ю.О. Соболев

	Рабочая программа дисциплины (модуля)
по дисциплине:	Инструменты Big Data
по направлению:	Прикладная математика и информатика
профиль подготовки:	Науки о данных
	центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
	центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
курс:	2
квалификация:	магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Аудиторных часов: 22 всего, в том числе:

лекции: 2 час.

семинары: 20 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 83 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составили:

П.И. Ахтямов, преподаватель

К.А. Лапин, старший методист

Программа обсуждена на заседании центра дополнительного, дополнительного профессионального и
онлайн-образования "Пуск" 01.03.2025

Аннотация

В курсе рассматриваются основные подходы к анализу больших данных (Big Data). Прежде всего, вводятся основные понятия, связанные с анализом данных, рассматривается структура данных в современной компании, основные принципы хранения и обработки информации, а также понятия, связанные с "экосистемой больших данных". Отдельные занятия посвящены подходам к организации процесса анализа, в т.ч. больших данных, в современной компании, рассматриваются основные фазы и подходы, особое внимание уделяется подготовке данных, их предобработке, выделению и синтезу ключевых характеристик (feature engineering). Рассматриваются методы хранения и организации доступа к данным в организации, а также основные технологии для анализа данных, такие как Hadoop, Apache Spark и т.п. Даются некоторые алгоритмические подходы, используемые для анализа данных, в т.ч. больших данных. Обсуждаются потоковые алгоритмы, модификации "классических" алгоритмов с учетом объемов обрабатываемых данных. Обсуждаются методы построения интеллектуальных систем основанных, в т.ч. на подходах обучения с подкреплением (reinforcement learning), вероятностных графовых моделях и т.п.

1. Цели и задачи

Цель дисциплины

- изучение базовых алгоритмов и технологий по обработке и анализу данных, изучение инструментария для обработки, в том числе "больших данных" (Big Data), для их применения в реальных проектах.

Задачи дисциплины

- приобретение студентами навыков по обработке и анализу данных, способности выбирать необходимые инструменты и алгоритмы анализа данных в зависимости от характера данных, структуры и т.п., а также потребностей организации по их анализу.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности
ОПК-3 Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	ОПК-3.2 Владеет исследовательскими методами и способен использовать их при решении новых задач, применяя знания из различных областей науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, задач, понимает и учитывает на практике границы применимости получаемых решений
ОПК-4 Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области	ОПК-4.1 Умеет применять информационно-коммуникационные технологии для поиска и анализа профессиональной информации, выделения в ней главного, структурирования, оформления и представления в виде аналитических обзоров с обоснованными выводами и рекомендациями

профессиональной деятельности	ОПК-4.3 Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ОПК-6 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области информатики и вычислительной техники, учитывая особенности и ограничения различных методов решения	ОПК-6.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-6.5 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
	ОПК-6.2 Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- подходы к организации хранилищ данных в современной компании, а также тенденциях их развития и способы взаимодействия с ними;
- традиционные и «нетрадиционные» источники данных для бизнес-анализа;
- решаемые в процессе анализа данных задачи;
- основные способы извлечения данных;
- основные подходы и методы анализа данных.

уметь:

- планировать работы по выполнению проектов, связанных с анализом, в том числе больших, данных;
- использовать инструментарий для извлечения данных из различных источников (БД, публичные web-сервисы и т.п.);
- использовать инструментарий для анализа данных (статистические пакеты и т.п.), в том числе в рамках современных парадигм обработки данных больших объемов данных (map-reduce и т.п.).

владеть:

- навыками постановки задачи анализа данных в интересах компании, способами предобработки и предварительной визуализации данных;
- навыками построения аналитических моделей и методов их оценки;
- навыками донесения результатов аналитических исследования до бизнес-спонсоров и коллег.

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

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Что такое большие данные. Хранилища больших данных	1	4		17
2	Обработка больших объемов данных при помощи Hadoop MapReduce	1	4		16
3	Аналитические функции при помощи Apache Hive		4		17

4	Обработка пакетных данных на примере Apache Spark		4		16
5	Нереляционные базы данных и построение pipeline		4		17
Итого часов		2	20		83
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 3 (Осенний)

1. Что такое большие данные. Хранилища больших данных

Основы и проблемы Big Data. Архитектуры Big Data. Распределенная файловая система HDFS — теоретические основы. Практика по HDFS

2. Обработка больших объемов данных при помощи Hadoop MapReduce

Решение распределенных задач. MapReduce. Расширение MapReduce Streaming. Глобальная сортировка в MapReduce. Комплексные операции в MapReduce.

3. Аналитические функции при помощи Apache Hive

Apache Hive: Data Definition Language. Apache Hive: Data Manipulation Language. Apache Hive: продвинутые операции.

4. Обработка пакетных данных на примере Apache Spark

Предыстория пакетной обработки данных. Основные операции по обработке пакетных данных: RDD. Разбираем задачу в Spark. Spark SQL. Очереди сообщений на примере Kafka.

5. Нереляционные базы данных и построение pipeline

Основы Apache HBase. Основы Apache Cassandra. Хранение данных при помощи Amazon S3 API. Настройка pipeline при помощи Apache Airflow

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Обучение проходит с использованием дистанционных образовательных технологий.

6. Перечень рекомендуемой литературы

Основная литература

1. Анализ данных и регрессия [Текст] : в 2-х вып. Вып.2/Ф. Мостеллер, Дж. Тьюки, -М., Финансы и статистика, 1982
2. Администрирование баз данных [Текст] : [учеб. пособие для вузов] / Дж. Уэлдон ; пер. с англ. В. И. Будзко, А. И. Прохорова. — М. : Финансы и статистика, 1984. — 207 с.

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

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

Образовательная платформа

Webinar.ru

Zoom

Google Drive

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студенту рекомендуется внимательно слушать лектора, следить за тем, что написано на доске или представлено на слайдах презентации, анализировать получаемую им информацию. В случае, если материал лекции непонятен, следует задать вопрос в отведенное для вопросов время. Студенту также рекомендуется конспектировать материал лекции в тетради, что улучшает запоминание.

При выполнении практических работ студенту рекомендуется внимательно анализировать поставленную задачу, уделяя особое внимание критериям оценки точности решения задачи. Особенное внимание следует уделять методологическим аспектам решения задач.

При ведении самостоятельной работы студенту рекомендуется внимательно подходить к изучению научных статей, обращать внимание на значимость полученного результата, на требования к обучающей выборке, на скорость работы предлагаемых алгоритмов, на результаты их сравнения с существующими. В случае, если изучаемый материал понятен не до конца, рекомендуется обращение к дополнительной литературе.

Студенту рекомендуется внимательно анализировать вопросы в экзаменационном билете. Ответ на экзаменационный билет должен быть подробным и четким, все релевантные формулы должны быть приведены и пояснены. При ответе на вопрос студент должен проявить не только умение запомнить материал, но и глубокое его понимание. Рекомендуется избегать приведения в ответе материала, не относящегося к билету.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладная математика и информатика
профиль подготовки:	Науки о данных центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск" центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
курс:	<u>2</u>
квалификация:	магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Разработчики:

П.И. Ахтямов, преподаватель
К.А. Лапин, старший методист

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности
ОПК-3 Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	ОПК-3.2 Владеет исследовательскими методами и способен использовать их при решении новых задач, применяя знания из различных областей науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, задач, понимает и учитывает на практике границы применимости получаемых решений
ОПК-4 Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности	ОПК-4.1 Умеет применять информационно-коммуникационные технологии для поиска и анализа профессиональной информации, выделения в ней главного, структурирования, оформления и представления в виде аналитических обзоров с обоснованными выводами и рекомендациями
	ОПК-4.3 Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ОПК-6 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области информатики и вычислительной техники, учитывая особенности и ограничения различных методов решения	ОПК-6.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-6.5 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
	ОПК-6.2 Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем

2. Показатели оценивания компетенций

В результате изучения дисциплины «Инструменты Big Data» обучающийся должен:

знать:

- подходы к организации хранилищ данных в современной компании, а также тенденциях их развития и способы взаимодействия с ними;
- традиционные и «нетрадиционные» источники данных для бизнес-анализа;
- решаемые в процессе анализа данных задачи;
- основные способы извлечения данных;
- основные подходы и методы анализа данных.

уметь:

- планировать работы по выполнению проектов, связанных с анализом, в том числе больших, данных;
- использовать инструментарий для извлечения данных из различных источников (БД, публичные web-сервисы и т.п.);
- использовать инструментарий для анализа данных (статистические пакеты и т.п.), в том числе в рамках современных парадигм обработки данных больших объемов данных (map-reduce и т.п.).

владеть:

- навыками постановки задачи анализа данных в интересах компании, способами предобработки и предварительной визуализации данных;
- навыками построения аналитических моделей и методов их оценки;
- навыками донесения результатов аналитических исследования до бизнес-спонсоров и коллег.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Примерный перечень вопросов для текущего контроля:

1. Ценность данных для современной компании. Понятие BigData. Понятие Data Science. Методология, хранения, обработки и анализа данных.
2. Задачи специалиста по обработке данных. Жизненный цикл проекта анализа данных. Представление результатов анализа.
3. Регрессионный анализ: линейная и логистическая регрессия.
4. Вероятностные графовые модели. Сети Байеса. Наивный Байес.
5. Деревья принятия решений. Анализ временных рядов.
6. Рекомендательные системы (recommender systems): основные понятия, история возникновения и базовые подходы к построению, неперсонализированные и персонализированные системы, современные рекомендательные системы: совместная фильтрация (пользователь-пользователь, продукт-продукт,...), фильтрация содержимого.
7. Особенности анализа потоковых данных.
8. Потоковые алгоритмы и соответствующий инструментарий.
9. Понятие кэширования данных, основные подходы и их применение.
10. Понятие хэширования, основные подходы, специальные виды хэширования (LSH).
11. Вероятностные графовые модели.
12. Основные понятия и классификация. Сети Байеса.
13. Способы задания и использования.
14. Основные структуры данных и типы запросов к сети.
15. Анализ потоков влияния.
16. Инструменты для формирования моделей. Вероятностное программирование.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Примерный перечень вопросов:

1. Цели и задачи хранения, обработки и анализа данных.
2. Современные тенденции. Структура данных в современной компании.
3. Структурированные и неструктурированные данные.
4. Понятие “Big Data”. Примеры проектов анализа данных.
5. Понятие “Data Science”.
6. Классификация и краткая характеристика используемых методов и подходов.
7. Требования к компетенциям специалиста.
8. Места специалиста по анализу данных в проектах компании.
9. Место статистики в жизненном цикле анализа данных.
10. Понятие гипотезы. Проверка гипотезы.
11. Описательная статистика при анализе данных.

12. Значимость данных Доверительный интервал. ANOVA.
13. Основы визуализации данных.
14. Исследование данных и представление результатов.
15. Признаки зашумленных данных.

Примерные экзаменационные билеты:

Билет №1:

1. Автоматическая кластеризация данных. Метода К-средних.
2. Алгоритм кластеризации. Принципы использования.

Билет №2:

1. Определение оптимального К. Оценка результатов кластеризации. Иерархическая кластеризация.
2. Системы выдачи рекомендаций (Recommender System). Правила ассоциации. Алгоритм Apriori. Оценка результатов.

Критерии оценивания

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Итоговой контроль проходит в форме экзамена на lms. На экзамен отводится 4 академических часа.