

Федеральное государственное бюджетное учреждение науки
Ордена Трудового Красного Знамени
Институт солнечно-земной физики
Сибирского отделения Российской академии наук
(ИСЗФ СО РАН)

УТВЕРЖДАЮ:

Директор ИСЗФ СО РАН

чл.-корр. РАН _____ А.В. Медведев

«11» мая 2023 г.

Рабочая программа дисциплины

Б1.В.ДВ.1.1 Введение в технологии Больших Данных

Направление подготовки **03.04.02 Физика**

Направленность (профиль): **Физика солнечно-земных связей**

Квалификация выпускника: **МАГИСТР**

Тип профессиональной деятельности: **научно-исследовательский,
педагогический**

Форма обучения: **очная**

Иркутск 2023

Рабочая программа составлена на основании Федерального государственного образовательного стандарта высшего образования по направлению подготовки 03.04.02 Физика (уровень магистратуры), утвержденного приказом Минобрнауки России от 07.08.2020 № 914

РАБОЧУЮ ПРОГРАММУ разработал кандидат физико-математических наук	О.И. Бернгардт
---	----------------

1. Место и роль дисциплины (модуля) в структуре ОПОП

Дисциплина «Введение в технологии Больших Данных» относится к вариативной части блока 1 «Дисциплины (модули) по выбору» основной образовательной программы по направленности (профилю) подготовки Физика солнечно-земных связей направления подготовки 03.04.02 Физика.

Предшествующие дисциплины, на которые данная дисциплина опирается: Компьютерные технологии.

2. Цели и задачи дисциплины (модуля)

Целью дисциплины «Введение в технологии Больших Данных» является формирование у обучающихся комплекса теоретических знаний и практических навыков по работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут в понимании принципов работы различного программного обеспечения в этой области, а также помогут при сборе, разработке методов анализа и самом анализе структурированной или неструктурированной информации существенных объемов

Задачами дисциплины «Введение в технологии Больших Данных» является:

- получение начальных знаний о предмете Big Data
- получение начальных знаний и умений по получению Big Data
- получение начальных знаний и умений по простейшей обработке Big Data
- получение начальных знаний и умений по созданию систем обработки Big Data и использованию BigData в системах принятия решений
- получение начальных знаний и умений по созданию систем обработки Big Data и использованию BigData в системах реального времени

3. Требования к результатам освоения дисциплины (модуля)

Процесс изучения дисциплины «Введение в технологии Больших Данных» направлен на формирование следующих компетенций в соответствии с ОПОП по направлению подготовки 03.04.02 Физика:

Компетенции	Индикаторы достижения компетенции	Планируемые результаты обучения по дисциплине
ОПК-3. Способен применять знания в области информационных технологий, использовать современные компьютерные сети, программные продукты и ресурсы информационно-телекоммуникационной сети "Интернет" (далее - сеть "Интернет") для решения задач профессиональной деятельности, в том числе находящихся за пределами профильной подготовки	ИД 1. Сбор и систематизация научно-исследовательской информации о рассматриваемом объекте или явлении с использованием информационных технологий в рамках задач предметной области	Знать основные понятия, подходы и алгоритмы в приложении к задачам Big Data; Уметь применять основные понятия, подходы и алгоритмы в приложении к задачам Big Data; Владеть подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи
	ИД 2. Критическая оценка достоверности полученной научно-исследовательской информации о рассматриваемом объекте или явлении;	Знать достоинства и недостатки современных подходов и алгоритмы в приложении к задачам Big Data, теоретические основы их работы; Уметь анализировать эффективность подходов и алгоритмы в приложении к конкретным задачам Big Data; Владеть умением определить наиболее эффективные подходы и алгоритмы в приложении к задачам Big Data, аргументировать их

		выбор Знать особенности реализации современных программно-аппаратные решения в приложении к задачам Big Data; Уметь применять современные программно-аппаратные решения в приложении к задачам Big Data; Владеть современными программно-аппаратными решениями для решения практических задач Big Data
ИД 4. Применение на практике методов и алгоритмов разработки программного обеспечения для решения проблем в рамках научно-исследовательских задач в том числе задач обработки наблюдательных данных.		
ПКА-2. Способен проводить научные исследования в области физики солнечно-земных связей, используя необходимые знания теоретических и экспериментальных разделов физики	ИД 3. Использует современные теоретические и экспериментальные методы, включая методы обработки и анализа данных, при проведении научных исследований и реализации научных проектов в области физики солнечно-земных связей	Знать возможности современных программно-аппаратных решений в приложении к задачам Big Data; Владеть умением проводить аргументированный выбор основных подходов и алгоритмов в приложении к решению конкретных задач Big Data;

4. Объем дисциплины (модуля) и виды учебной работы

Общая трудоемкость дисциплины составляет 3 зачетных единиц, 108 часов.

Вид учебной работы	Всего часов / зачетных единиц
Аудиторные занятия (всего)	36/1
В том числе:	
Лекции	36/1
Лабораторные работы	
Практические занятия	36/1
Самостоятельная работа (всего)	36/1
Вид промежуточной аттестации (зачет)	
Контактная работа (всего)	72/2
Общая трудоёмкость (часы/зачетные единицы)	108/3

Содержание дисциплины

5.1. Содержание разделов и тем дисциплины (модуля).

Раздел 1. Введение

Тема 1. Проверка базовых знаний и вводная лекция

Введение в Big Data и машинное обучение. Объем данных, скорость данных, различность данных, качество данных, значимость данных. Получение данных. Хранение данных. Доступ к данным. Обработка данных. Проверка гипотез и выявление скрытых зависимостей. Параллельные и облачные вычисления. Машинное обучение. Искусственные нейронные сети. Входной тест.

Тема 2. Введение в Питон

Элементы языка Питон: базовые операторы, структуры и функции. Работа с файлами. Работа с модулями. Создание модулей.

Раздел 2. Дата майнинг

Тема 3. Протоколы и форматы интернета

Структура сервер-клиент. Работа с сокетами. Telnet. Организация протокола HTTP, команды, авторизация. Организация протокола FTP, команды, авторизация. Шифрование. Организация протокола SSH, авторизация. Протокол HTTPS.

Тема 4. Протоколы и форматы баз данных

Структура реляционной базы данных на примере MySQL. SQL - команды. Создание базы данных. Добавление, удаление и редактирование записей. Поиск и выбор записей. Объединения таблиц. Индексы.

Тема 5. Подготовка к анализу текстов

Команды и элементы языка regexr. Шаблоны. Контекстный поиск. Контекстный поиск и замена. Многовариантный поиск. Переменные regexr. Особенности использования на Питоне. Ассоциированные массивы - реализация и использование. Задача сбора статистики.

Тема 6. Подготовка к анализу соотношений

Графы, отношения и пары. Ребра и вершины. Ориентированные и неориентированные графы. Маршруты, пути и циклы. Связность графа. Деревья. Отображение и анализ графов.

Раздел 3. Предварительная обработка больших данных

Тема 7. Принятие простых решений

Статистические распределения, вероятность. Условная вероятность. Функция распределения и плотность распределения. Интегральная оценка вероятности события. Принятие решения на основе статистических данных. Байесовский подход. Проверка статистических гипотез. Ошибки первого и второго рода.

Тема 8. Выделение явных закономерностей

Корреляция. Регрессионная зависимость. Линейная и нелинейная регрессия. Метод наименьших квадратов и случаи его сведения к системе линейных уравнений. Метод максимального правдоподобия. Периодичность процессов и Фурье-анализ.

Тема 9. Выявление скрытых закономерностей.

Метод наименьших квадратов и случаи его сведения к анализу матриц. Матрицы: определители и ранги, собственные числа и собственные вектора. Уменьшение числа параметров. Линейно-зависимые и линейно-независимые параметры. Метод главных компонент. Матричный анализ. Сингулярное разложение.

Раздел 4. Экспертные системы

Тема 10. Линейные модели и деревья принятия решений.

Таблицы принятия решений. Деревья принятия решений. Листья и ветви. Обучение (построение) дерева. Условия ветвления. Отсечение ветвей. Глубина дерева и остановка.

Тема 11. Нейронные сети.

Искусственный нейрон. Искусственные нейронные сети. Входные, скрытые и выходные нейроны. Типы и структура нейронных сетей. Сети прямого распространения и рекуррентные сети. Архитектуры нейронных сетей. Обучение с учителем, без учителя и с подкреплением. Генетические алгоритмы.

Раздел 5. Технологии работы с большими объемами входных данных

Тема 12. Введение в нереляционные базы данных (NoSQL)

Языки разметки. XML. Структура языка XML. Элементы языка XML. JSON и XML - сходства и отличия. Способы генерации JSON документов. Технологии, использующие JSON-документы.

Тема 13. Технологии и реализации NoSQL

Серверные приложения. Хранение и доступ к большим данным. MongoDB. Структура базы данных. Команды MongoDB. Создание записей. Изменение записей. Удаление записей. Выбор записей.

Тема 14. Технологии реального времени в BigData

Параллельные вычисления и обработка данных в режиме реального времени. Кластеры и OpenMPI. Параллельные вычисления на OpenMPI. Элементы программы и управления процессами. GPU и CUDA. Параллельные вычисления на CUDA. MapReduce. Операции Map и Reduce. Реализация алгоритма сбора статистики на примере MapReduce для MongoDB. Распределенные системы. Hadoop. Распределенные файловые системы. HDFS. Распределенные базы данных. HBASE.

Тема 15. Финальный проект.

Разработка собственной системы анализа BigData на доступных наборах больших данных (например <http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free>; <https://catalog.data.gov/dataset>).

5.2. Разделы дисциплины (модуля) и виды занятий

№ п/п	Раздел	Всего часов	Аудиторные занятия				СРС
			Лекции	Лаб. занятия	Практические занятия	Семинары	
1.	Раздел 1. Введение	9	4		3		2
2.	Тема 1. Проверка базовых знаний и вводная лекция	4	2		1		1
3.	Тема 2. Введение в Питон	5	2		2		1
4.	Раздел 2. Дата майнинг	24	8		8		8
5.	Тема 3. Протоколы и форматы интернета	6	2		2		2
6.	Тема 4. Протоколы и форматы баз данных	6	2		2		2
7.	Тема 5. Подготовка к анализу текстов	6	2		2		2
8.	Тема 6. Подготовка к анализу соотношений	6	2		2		2
9.	Раздел 3. Предварительная обработка больших данных	19	6		7		6
10.	Тема 7. Принятие простых решений	6	2		2		2
11.	Тема 8. Выделение явных закономерностей	6	2		2		2

12.	Тема 9. Выявление скрытых закономерностей.	7	2		3		2
13.	Раздел 4. Экспертные системы	20	8		6		6
14.	Тема 10. Линейные модели и деревья принятия решений.	6	2		2		2
15.	Тема 11. Нейронные сети.	14	6		4		4
16.	Раздел 5. Технологии работы с большими объемами входных данных	36	10		12		14
17.	Тема 12. Введение в нереляционные базы данных (NoSQL)	6	2		2		2
18.	Тема 13. Технологии и реализации NoSQL	8	2		4		2
19.	Тема 14. Технологии реального времени в BigData	10	4		4		2
20.	Тема 15. Финальный проект	12	2		2		8
Итого (часы)		108	36		36		36
Итого (з.е.)		3	1		1		1

5.3. Разделы и темы дисциплины (модуля) и междисциплинарные связи

№ п/п	Наименование обеспечиваемых (последующих) дисциплин и практик	№ № разделов и/или тем данной дисциплины, необходимых для изучения обеспечиваемых (последующих) дисциплин
1.	Производственная практика (научно-исследовательская работа)	Все

5.4. Перечень лекционных занятий

№ п/п	№ раздела и темы дисциплины (модуля)	Наименование используемых технологий	Трудоемкость (часы)	Оценочные средства
1.	P1.T1	Тема 1. Проверка базовых знаний и вводная лекция	2	Устный опрос
2.	P1.T2.	Тема 2. Введение в Питон	2	Устный опрос
3.	P2.T3.	Тема 3. Протоколы и форматы интернета	2	Устный опрос
4.	P2.T4.	Тема 4. Протоколы и форматы баз данных	2	Устный опрос
5.	P2.T5.	Тема 5. Подготовка к анализу текстов	2	Устный опрос
6.	P2.T6.	Тема 6. Подготовка к анализу соотношений	2	Устный опрос
7.	P3.T7.	Тема 7. Принятие простых решений	2	Устный опрос
8.	P3.T8.	Тема 8. Выделение явных закономерностей	2	Устный опрос
9.	P3.T9.	Тема 9. Выявление скрытых закономерностей.	2	Устный опрос
10.	P4.T10.	Тема 10. Линейные модели и деревья принятия решений.	2	Устный опрос
11.	P4.T11.	Тема 11. Нейронные сети.	6	Устный опрос
12.	P5.T12.	Тема 12. Введение в	2	Устный опрос

		нереляционные базы данных (NoSQL)		
13.	P5.T13.	Тема 13. Технологии и реализации NoSQL	2	Устный опрос
14.	P5.T14.	Тема 14. Технологии реального времени в BigData	4	Устный опрос, Итоговое тестирование
15.	P5.T15.	Тема 15. Финальный проект	2	Доклад

5.5. Перечень семинарских, практических занятий и лабораторных работ

№ п/п	№ раздела и темы дисциплины (модуля)	Наименование семинаров, практических и лабораторных работ	Трудоемкость (часы)	Оценочные средства
1.	P1.T1	Тема 1. Проверка базовых знаний и вводная лекция	1	Практическое задание
2.	P1.T2.	Тема 2. Введение в Питон	2	Практическое задание
3.	P2.T3.	Тема 3. Протоколы и форматы интернета	2	Практическое задание
4.	P2.T4.	Тема 4. Протоколы и форматы баз данных	2	Практическое задание
5.	P2.T5.	Тема 5. Подготовка к анализу текстов	2	Практическое задание
6.	P2.T6.	Тема 6. Подготовка к анализу соотношений	2	Практическое задание
7.	P3.T7.	Тема 7. Принятие простых решений	2	Практическое задание
8.	P3.T8.	Тема 8. Выделение явных закономерностей	2	Практическое задание
9.	P3.T9.	Тема 9. Выявление скрытых закономерностей.	3	Практическое задание
10.	P4.T10.	Тема 10. Линейные модели и деревья принятия решений.	2	Практическое задание
11.	P4.T11.	Тема 11. Нейронные сети.	4	Практическое задание
12.	P5.T12.	Тема 12. Введение в нереляционные базы данных (NoSQL)	2	Практическое задание
13.	P5.T13.	Тема 13. Технологии и реализации NoSQL	4	Практическое задание
14.	P5.T14.	Тема 14. Технологии реального времени в BigData	4	Практическое задание
15.	P5.T15.	Тема 15. Финальный проект	2	Практическое задание

5.6. Тематика заданий для самостоятельной работы

Раздел	Тема	Вид самостоятельной работы	Задание	Рекомендуемая литература	Кол-во часов
1	P1.T1	Выполнение практических заданий	Решение задач по теме 1.	СИР [1-6]	1
2	P1.T2.	Выполнение практических заданий	Решение задач по теме 2.	СИР [1-6]	1
3	P2.T3.	Выполнение практических заданий	Решение задач по теме 3.	СИР [1-6]	2
4	P2.T4.	Выполнение практических заданий	Решение задач по теме 4.	СИР [1-6]	2
5	P2.T5.	Выполнение практических заданий	Решение задач по теме 5.	СИР [1-6]	2
6	P2.T6.	Выполнение практических заданий	Решение задач по теме 6.	СИР [1-6]	2
7	P3.T7.	Выполнение практических заданий	Решение задач по теме 7.	СИР [1-6]	2
8	P3.T8.	Выполнение практических заданий	Решение задач по теме 8.	СИР [1-6]	2
9	P3.T9.	Выполнение практических заданий	Решение задач по теме 9.	СИР [1-6]	2
10	P4.T10	Выполнение практических заданий	Решение задач по теме 10.	СИР [1-6]	2
11	P4.T11	Выполнение практических заданий	Решение задач по теме 11.	СИР [1-6]	4
12	P5.T12	Выполнение практических заданий	Решение задач по теме 12.	СИР [1-6]	2
13	P5.T13	Выполнение практических заданий	Решение задач по теме 13.	СИР [1-6]	2
14	P5.T14	Выполнение практических заданий	Решение задач по теме 14.	СИР [1-6]	2
15	P5.T15	Выполнение практических заданий	Выполнение проекта	СИР [1-6]	8

5.7. Методические рекомендации по организации самостоятельной работы обучающихся

Самостоятельная работа студентов в рамках изучения дисциплины «Введение в технологии больших данных» регламентируется общим графиком учебной работы, предусматривающим посещение практических занятий и регулярное выполнение заданий по ним, выполнение домашних заданий.

При организации самостоятельной работы по дисциплине «Введение в технологии больших данных» студенту следует:

1. Внимательно изучить материалы, характеризующие курс и тематику самостоятельного изучения дисциплины. Это позволит четко представить, как круг изучаемых тем, так и глубину их постижения.

2. Составить подборку литературы и источников, достаточную для изучения предлагаемых тем. Следует заметить, что данный курс является крайне современным, и доступная печатная литература в настоящее время отсутствует, поэтому необходимо использовать литературу из интернет-источников. В программе дисциплины представлены основной и дополнительный списки литературы. Они носят рекомендательный характер, это означает, что всегда есть литература, которая может не входить в данный список, но является необходимой для освоения темы.

3. Основное содержание той или иной проблемы следует уяснить, изучая учебную литературу и источники, с опорой на конспекты лекций.

4. Абсолютное большинство задач носит практический характер, и они могут быть решены студентом только с привлечением компьютерной обработки данных. Это предполагает наличие у студентов не только знания категорий и понятий, но и умения использовать их в качестве инструмента для анализа и выполнения практических задач.

5. Финалом практической работы студента является выполнение им финального проекта, решающего некую практическую задачу, связанную с тематикой курса (в паре или самостоятельно). Использование парного выполнения проекта поощряется, поскольку во-первых, связано с современными методиками эффективного (экстремального) программирования, позволяющим глубже разобраться в проблеме и используемых для ее решения методах, а во-вторых, учит эффективной командной работе.

Пояснительная записка к финальному проекту должна включать:

Постановку задачи, включающую требования к функционалу системы, к программному и аппаратному обеспечению, необходимому для ее работы, описанию исходных данных и ожидаемых результатов

Описание проекта разрабатываемой системы в виде структуры программы, идеи решения проблемы, используемых технических решений, метода отладки.

Исходный код программной реализации программной системы на выбранном языке программирования.

Описание метода проверки решения, оценочные точности (если возможно).

К отчету по финальному проекту допускаются студенты, продемонстрировавшие работу программной системы. На отчете студент должен ответить на вопросы преподавателя по алгоритмам и программной реализации предложенного решения.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Основная литература

№ п/п	Автор, название, место издания, издательство, год издания учебной и учебно-методической литературы	Количество экземпляров
1.	Джонсон, Н., Статистика и планирование эксперимента в технике и науке: Методы обработки данных [Текст] / Н. Джонсон, Ф. Лион. - М. : Мир, 1980. - 610 с.	2
2.	Водолазкий, В., Энциклопедия Perl / В. Водолазкий, В. Семериков. - СПб. : Питер, 2002. - 576 с.	2
3.	Бернгардт О.И. Введение в Большие Данные и методы машинного обучения (конспекты лекций). Часть 1. Классические методы и базовые алгоритмы	ЭБ http://irbis.iszf.irk.ru неограниченный доступ

6.2. Дополнительная литература

№ п/п	Автор, название, место издания, издательство, год издания учебной и учебно-методической литературы	Количество экземпляров
1.	Колемаев, В. А. Теория вероятностей и математическая статистика [Текст] : учебник для вузов / В. А. Колемаев, В. Н. Калинина. - 2-е изд., испр. и доп. - М. : ЮНИТИ-ДАНА, 2017. - 352 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
2.	Секей, Г. Парадоксы в теории вероятностей и математической статистике [Текст] : пер. с англ. / Г. Секей, В.М. Калинина - М.: Мир, 1990. - 240 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
3.	Доусон, Майкл. Программируем на Python : пер. с англ. / М. Доусон. - 2-е изд. - СПб : Питер, 2014. - 416 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
4.	Петрович, М. Л. Статистическое оценивание и проверка	ЭБ http://irbis.iszf.irk.ru

6.3. Профессиональные базы данных, используемые при осуществлении образовательного процесса по дисциплине:

- <http://sdrus.iszf.irk.ru/>

6.4. Информационные справочные системы, используемые при осуществлении образовательного процесса по дисциплине:

- Статьи по машинному обучению портала <https://habrahabr.ru>
- Библиотека разработчика IBM
<https://www.ibm.com/developerworks/analytics/library/>

6.5. Ресурсы информационно-телекоммуникационной сети «Интернет», необходимые для освоения дисциплины:

- Курс MIT <https://www.edx.org/course/introduction-computational-thinking-data-mitx-6-00-2x-5>
- Курс Стенфорда <https://www.youtube.com/watch?v=UzxYIbK2c7E>
- Курс TeamDEV <https://www.youtube.com/watch?v=fIZ64zHC6sU>
- Открытый курс машинного обучения. <https://habrahabr.ru/company/ods/blog/322626/>
- Открытый курс машинного обучения <http://jsman.ru/mongo-book/>
(<https://github.com/karlseguin/the-little-mongodb-book>)
- Открытый курс машинного обучения
<https://www.ibm.com/developerworks/ru/library/l-hadoop-1/index.html>

6.6. Программное обеспечение

Лицензионное и свободно распространяемое программное обеспечение, в том числе отечественного производства используемое при осуществлении образовательного процесса по дисциплине:

- Операционная система Ubuntu 18.04 (свободно распространяемое ПО)
- Офисный пакет Libre Office (свободно распространяемое ПО)
- 7-Zip (свободно распространяемое ПО)
- Adobe Acrobat Reader DC (свободно распространяемое ПО)
- Mozilla Firefox 1 (свободно распространяемое ПО)
- VLC Mediaplayer (свободно распространяемое ПО)
- K-Lite Codec Pack (свободно распространяемое ПО)
- Дистрибутив Python Anaconda (свободно распространяемое ПО)
- Набор компиляторов GCC (свободно распространяемое ПО)
- Операционная система Microsoft Windows 10 Pro
- Система ВКС VideoMost Proton

7. Образовательные технологии

В учебном процессе используются как активные, так интерактивные формы проведения занятий.

Аудиторные занятия (АЗ) проводятся в интерактивной форме с использованием мультимедийного обеспечения (ноутбук, проектор).

Практические занятия (ПЗ) включают в себя выполнение поставленных преподавателем заданий в индивидуальном и групповом порядке, заключительным этапом является выполнение финального проекта (парное или индивидуальное).

8. Практическая подготовка

Практическая подготовка обучающихся в рамках реализации данной учебной дисциплины осуществляется на практических занятиях.

9. Материально-техническое обеспечение дисциплины (модуля)

Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации	Аудитория укомплектована специализированной мебелью на 30 посадочных мест, оснащена оборудованием и техническими средствами обучения, служащими для представления учебной информации большой аудитории: <ul style="list-style-type: none">• доска магнитно-маркерная Branberg• экран для проектора Projecta• проектор BenQ MH733 1920 x 1080• ноутбук ASUS L1500CDA Windows 10 Pro• система акустическая Electro Voice EVID 6.2
Учебная аудитория для групповых и индивидуальных консультаций и самостоятельной работы	Аудитория укомплектована специализированной мебелью на 7 посадочных мест, оснащена компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде: <ul style="list-style-type: none">• персональные компьютеры Неттоп Think Center Lenovo M710Q• мониторы IIYAMA PL2283H, Dell CRHX9K2• доска магнитно-маркерная Branberg• экран для проектора Projecta• проектор BenQ MH733 1920 x 1080