Министерство науки и высшего образования Российской Федерации

федеральное государственное бюджетное образовательное учреждение высшего образования
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра Прикладной информатики

Рабочая программа дисциплины

Создание Big data

Основная профессиональная образовательная программа
высшего образования по направлению подготовки

09.03.03 Прикладная информатика

Направленность (профиль):

Прикладные информационные системы и технологии

Уровень:
Бакалавриат

Форма обучения
Очная

Согласовано
Руководитель ОПОП
Яготинцева **Яготинцева Н.В.**

Утверждаю
Председатель УМС **И.И. Палкин**

Рекомендована решением
Учебно-методического совета
22 10 **2019 г., протокол №** 2

Рассмотрена и утверждена на заседании кафедры
22 октября **2019 г., протокол №** 9
Зав. кафедрой **Истомин Е.П.**

Авторы-разработчики:
Попов / **Попов Н.Н.**
/ Колпенов О.Н.
Яготинцева / Яготинцева Н.В.

Санкт-Петербург 2019

1. Цель и задачи освоения дисциплины

Цель освоения дисциплины – изучение концепций Big Data и Интернета вещей, архитектуры хранения данных. Получение знаний о технологиях и методах анализа Big Data и интерпретации результатов.

Задачи:

- формирование практического навыка по работе с кластером хранения и обработки Big Data на примере Cloudera Hadoop, выбору методов и технических решений в зависимости от типа решаемой задачи обработки данных и их объема. Умение использовать инструментарий анализа данных и их визуализации.
- приобретение навыков управления проектом развития системы информационного обеспечения принятия управленческих решений от сбора данных, их хранения и обработки до представления результатов

2. Место дисциплины в структуре основной профессиональной образовательной программы

Дисциплина относится к факультативным, изучается в 6 семестре и является вспомогательной для написания ВКР.

3. Перечень планируемых результатов обучения

Процесс изучения дисциплины направлен на формирование компетенций ПК-11

Таблица 1.

Профессиональные компетенции

Задача ПД	**Объект или область знания**	**Категория профессио-нальных компетенций**	**Код и наименование профессиональной компетенции**	**Код и наименование индикатора достижения профессиональной компетенции**	**Основание (ПС, анализ опыта)**
Тип задач профессиональной деятельности проектный					
проектирование информационных систем в соответствии со спецификой профиля подготовки по видам обеспечения (программное, информационное, организационное, техническое);	Прикладные и информационные процессы; Информационные системы; Информационные технологии		**ПК-11. Способен проектировать программное обеспечение**	ИДПК-11.1. Использует существующие типовые решения и шаблоны проектирования программного обеспечения ИДПК-11.2. Применяет методы и средства проектирования программного обеспечения, структур данных, баз	06.001 Программист

				данных, программных интерфейсов ИДПК-11.3. Использует принципы и виды построения архитектуры программного обеспечения	

4. Структура и содержание дисциплины

4.1. Объем дисциплины

Объем дисциплины составляет 3 зачетные единицы, 108 академических часа.

Таблица 2.

Объем дисциплины по видам учебных занятий в академических часах

Объём дисциплины	**Очная форма обучения**
Объем дисциплины	
Контактная работа обучающихся с преподавателем (по видам аудиторных учебных занятий) – всего:	
в том числе:	**-**
лекции	**14**
лабораторные занятия	**28**
Самостоятельная работа (далее – СРС) – всего:	**66**
Вид промежуточной аттестации	**зачет**

4.2. Структура дисциплины

Таблица 3.

Структура дисциплины для очной формы обучения

№	**Тема дисциплины**	**Семестр**	**Виды учебной работы, в т.ч. самостоятельная работа студентов, час.**			**Формы текущего контроля успеваемости**	**Формируемые компетенции**	**Индикаторы достижения компетенций**
			Лекции	**Лабораторные работы**	**СРС**			
1	Источники данных	6	2	4	13	Опрос Сдача лабораторных работ	ПК-11	ИДПК-11.1 ИДПК-11.2 ИДПК-11.3
2	Основные понятия Big Data	6	3	6	13	Опрос Сдача лабораторных работ	ПК-11	ИДПК-11.1 ИДПК-11.2 ИДПК-11.3

3	Технологии работы с Big Data	6	3	6	13	Опрос Сдача лабораторных работ	ПК-11	ИДПК-11.1 ИДПК-11.2 ИДПК-11.3
4	Аналитика данных	6	3	6	13	Опрос Сдача лабораторных работ	ПК-11	ИДПК-11.1 ИДПК-11.2 ИДПК-11.3
5	Управление проектами Big Data	6	3	6	14	Опрос Сдача лабораторных работ	ПК-11	ИДПК-11.1 ИДПК-11.2 ИДПК-11.3
	ИТОГО	**-**	**14**	**28**	**66**	**-**	**-**	**-**

4.3. Содержание разделов дисциплины

Введение
Общий обзор дисциплины, специфика, области применения, терминология.

Тема 1. Источники данных
Обзор открытых источников данных. Организация сбора данных о среде с помощью концепции Интернета вещей. Взаимодействие с облачными хранилищами данных.

Тема 2. Основные понятия Big Data
Краткая история развития концепции. Виды и характеристики Big Data

Тема 3. Технологии работы с Big Data
Шардинг и репликация. Архитектуры хранения. Стек технологий Hadoop, распределенная файловая система HDFS, модель вычислений MapReduce

Тема 4. Аналитика данных
Основные понятия. Виды анализа данных. Методология исследования данных CRISP-DM

Тема 5. Управление проектами Big Data
Анализ сервисов облачных вычислений. Программные комплексы машинного обучения Apache Spark и Vowpal Wabbit. Моделирование и оценка результатов

4.4. Содержание занятий семинарского типа

Таблица 4.

Содержание лабораторных занятий для очной формы обучения

№ темы дисциплины	Тематика лабораторных занятий	Всего часов
1	Источники данных	4
2	Основные понятия Big Data	6
3	Технологии работы с Big Data	6
4	Аналитика данных	6
6	Управление проектами Big Data	6

5. Перечень учебно-методического обеспечения самостоятельной работы

обучающихся по дисциплине

Попов Н.Н., Александрова Л.В., Абрамов В.М. Инновационные технологии геоинформационного обеспечения управления данными предприятия. Режим доступа: http://elib.rshu.ru/files_books/pdf/rid_04837d21305f4a808ed637c5fda17db0.pdf

Методические рекомендации по организации самостоятельной работы обучающихся.

6. Оценочные средства для текущего контроля успеваемости и промежуточной аттестации по итогам освоения дисциплины

6.1. Текущий контроль

Текущий контроль проводиться в форме опроса и демонстрации преподавателю результатов лабораторной работы.

Примерные вопросы к опросу:

1. Опишите основные способы применения Hadoop
2. Приведите пример открытых источников гидрометеорологических данных
3. Как производится автоматизация сбора данных о среде
4. Что такое «Интернет вещей»
5. Как происходит взаимодействие с облачными хранилищами данных?
6. Назовите основные архитектуры хранения
7. Опишите стек технологий Hadoop
8. Что такое распределенная файловая система HDFS и где она используется?
9. Дайте краткое описание модели вычислений MapReduce
10. Назовите основные направления аналитики данных
11. Перечислите виды анализа данных
12. Что такое методология исследования данных CRISP-DM
13. Назовите известные вам инструменты анализа обычных данных
14. Опишите области применения различных инструментов анализа Big Data

Критерии оценивания:

Ответ засчитается, если студент владеет теоретическим материалом, формулирует собственные, самостоятельные, обоснованные, аргументированные суждения, представляет полные и развернутые ответы на вопросы.

Примерные задания на лабораторные работы:

Тема 1: Источники данных.

Цель: Получить представление об открытых источниках данных, используемых в научно-исследовательской деятельности.

Задание:

1. Изучить различные источники геоданных.
2. Оценить количественные характеристики баз данных

2. Используя возможности информационных технологий реализовать решение поставленной научной задачи.

Краткие теоретические сведения.

Любая геоинформационная система оперирует слоями данных, основанными на наборе пространственных данных. Их сбор представляет собой значительную проблему для исследователей разного уровня. В качестве подготовки к проведению исследования необходимо:

- выявить информационные потребности;
- осуществить отбор источников информации;
- осуществить оценку качества имеющихся данных.

Ход выполнения работы

1. Ознакомиться с заданием лабораторной работы и краткими теоретическими сведениями.

2. Выбрать из списка доступных проектов на сайте NASA те, которые отвечают географическим требованиям.

3. Провести анализ имеющихся данных.

4. Зарегистрироваться в системе.

5. Сделать подборку из 3-5 блоков данных.

6. Скачать данные и оценить структуру данных

В отчет по выполнению лабораторной работы включить основные шаги и описание задачи, которое можно решить с использованием выбранных данных.

В описание структуры данных необходимо включить описание формата хранения и перечень инструментов для работы с ними.

Тема 2: Основные понятия Big Data.

Цель: Ознакомится с предметной областью, относящейся к машинному обучению и большим данным.

Задание:

1. Дать определение основным терминам и понятиям, относящимся к предметной области.

2. Определить взаимосвязи прикладных и теоретических дисциплин, являющихся базовыми к машинному обучению.

Краткие теоретические сведения.

Любая современная концепция в области информационных технологий имеет бекграунд в одном из разделов базовых дисциплин. Перед началом изучения концепции необходимо:

- выявить базовые операции, проводимые в области;
- осуществить анализ необходимых знаний;
- определить понятия.

Ход выполнения работы

1. Ознакомиться с заданием лабораторной работы и краткими теоретическими сведениями.

2. Выбрать из предлагаемого списка термины, значение которых непонятно обучающемуся.

3. Обсудить с партнером по выполнению работы возможные варианты трактовки.

4. Получить разъяснения преподавателя.

5. Описать взаимосвязи концепций машинного обучения и Big Data и статистики.

В отчет по выполнению лабораторной работы включить описания основных терминов и технологий, применяемых в области машинного обучения.

Тема 3: Технологии работы с Big Data.

Цель: Ознакомится с основными технологиями работы с Big Data и машинным обучением.

Задание:

Ознакомится с применением языка программирования Python в качестве системы обработки данных и машинного обучения.

Краткие теоретические сведения.

Задача автоматизации работы с Big Data стоит особенно остро. Общее программное обеспечение зачастую не способно обрабатывать большие данные и проводить машинное обучение. В качестве основного инструмента разработчика выступает язык программирования Python с подключаемыми библиотеками, позволяющими:

- осуществить структурирование разрозненных данных;
- провести первичный анализ массива;

- выполнить действия по обработке информации, оценке ее полноты и значимости и по представлению ее в удобном виде.

Ход выполнения работы

1. Ознакомиться с заданием лабораторной работы и краткими теоретическими сведениями.

2. Ознакомится с описанием библиотек, применяемых в машинном обучении, на сайте разработчика ПО.

3. Провести анализ предложенного блока данных и установить или опровергнуть взаимосвязь между событиями.

В отчет по выполнению лабораторной работы включить результаты анализа блока данных, а также листинг программы

Тема 4: Аналитика данных.

Цель: Ознакомится с основными технологиями работы с Big Data и машинным обучением на примере программных комплексов Weka и Rubi.

Задание:

Оценить возможности программных комплексов Weka и Rubi на примере решения задачи машинного обучения, приведенного в качестве базового примера разработчиками.

Краткие теоретические сведения.

Задача машинного обучения стоит особенно остро. В последнее время все чаще специалисты предметной области рекомендуют программные комплексы Weka и Rubi, как наиболее простые и эффективные инструменты по работе с данными. С помощью них, как и с помощью языка программирования Python возможно:

- осуществить структурирование разрозненных данных;
- провести первичный анализ массива;
- выполнить действия по обработке информации, оценке ее полноты и значимости и по представлению ее в удобном виде.

Ход выполнения работы

1. Ознакомиться с заданием лабораторной работы и краткими теоретическими сведениями.

2. Ознакомится с инструкцией пользователя ПО Weka и Rubi на сайтах разработчиков.

3. Провести анализ предложенного блока данных и установить или опровергнуть взаимосвязь между событиями.

В отчет по выполнению лабораторной работы включить результаты анализа блока данных, а также листинг программы

Тема 5: Управление проектами Big Data.

Цель: Изучить типовой пример развития и управления проектом машинного обучения.

Задание:

1. Разработать структуру проекта для реализации научной задачи. Варианты производственных задач:

- формирование научной задачи;
- планирование потребности в вычислительных ресурсах;
- планирование потребности в распределенном хранении данных;
- учет и контроль источников данных;
- контроль качества информации;
- задача по выбору студента.

2. Используя возможности информационных технологий реализовать решение поставленной научной задачи.

Краткие теоретические сведения.

В любой геоинформационной системе организуются определенные процессы, чтобы: выявить информационные потребности;

- осуществить отбор источников информации;
- осуществить сбор информации;
- выполнить действия по обработке информации, оценке ее полноты и значимости и по представлению ее в удобном виде;
- вывести информацию для предоставления потребителям или передачи в другую систему;
- организовать использование информации для оценки тенденций, разработки прогнозов, оценки альтернатив решений и действий, выработки стратегии;
- организовать обратную связь — по результатам обработки данных осуществить коррекцию взаимодействия с внешней средой.

Ход выполнения работы

1. Ознакомиться с заданием лабораторной работы и краткими теоретическими сведениями.

2. Выбрать предметную область для автоматизации в рамках данной лабораторной работы.

3. Провести системный анализ предметной области, выявить научную задачу, выполнение которой нужно автоматизировать.

4. Выделить геоинформационные объекты предметной области, необходимые для решения поставленной задачи. Описать выделенные геоинформационные объекты, выбирая уровень абстрагирования с позиции решаемой задачи. Указать связи между объектами.

5. Построить датологическую модель предметной области с учетом ограничений, накладываемых СУБД и инструментальными средствами, выбранными для физической реализации базы данных.

6. Произвести физическое проектирование и реализацию базы данных и приложения для работы с ней. На данном этапе необходимо создать базу данных, разработать формы для ввода данных, отчеты и запросы, необходимые для решения поставленной задачи.

В отчет по выполнению лабораторной работы включить результаты системного анализа предметной области, инфологического и датологического проектирования, описание физической структуры программы.

В описание структуры данных необходимо включить:

- список сущностей;
- связи между сущностями;
- отношения связей и сущностей.

Критерии оценивания:

Лабораторные работы принимаются в формате зачтено/ не зачтено.

Зачтено, если задание выполнено полностью, в представленном отчете обоснованно получено правильное выполненное задание.

Не зачтено, если задания выполнены частично или не выполнено.

6.2. Промежуточная аттестация

Форма промежуточной аттестации по дисциплине – **зачет.**

Форма проведения зачета: устно по билетам

Перечень вопросов для подготовки к зачету:

ПК-11

Перечень вопросов для подготовки к зачету

1. Что такое большие данные
2. Характериситка VVV

3. Применение Hadoop
4. Пример открытых источников гидрометеорологических данных
5. Сбора данных о среде
6. Концепция «Интернет вещей»
7. Взаимодействие с облачными хранилищами данных
8. Работа с thingspeak.com
9. Виды и характеристики Big Data
10. Технологии работы с Big Data
11. Шардинг и репликация
12. Архитектуры хранения
13. Стек технологий Hadoop
14. Распределенная файловая система HDFS
15. Модель вычислений MapReduce
16. Аналитика данных
17. Виды анализа данных
18. Методология исследования данных CRISP-DM
19. Инструменты анализа обычных данных (RapidMiner, Weka, Knime)
20. Инструменты анализа Big Data (Hive, Pig)
21. Batch обработка и обработка в реальном времени
22. Анализ сервисов облачных вычислений
23. Программные комплексы машинного обучения Apache Spark и Vowpal Wabbit
24. Моделирование и оценка результатов
25. Создание слоев ГИС и их отображение на веб-портале

Зачет оценивается по двухбалльной шкале: «зачтено»/ «незачтено».

Критерии оценивания:

«Зачтено» - студент способен разрабатывать методики выполнения аналитических работ.

«Незачтено» - студент не способен разрабатывать методики выполнения аналитических работ.

7. Методические указания для обучающихся по освоению дисциплины

7.1. Методические указания к занятиям лекционного типа

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.

7.2. Методические указания к занятиям семинарского типа

Лабораторные работы

При подготовке к лабораторным работам необходимо заранее изучить методические рекомендации по его проведению. Обратить внимание на цель занятия, на основные вопросы для подготовки к занятию, на содержание темы занятия.

Лабораторное занятие проходит в виде выполнения определенного задания на компьютере с использованием специального программного обеспечения. Студент должен сдавать лабораторную работу в виде наглядной демонстрации достигнутых результатов преподавателю.

7.3. Методические указания по организации самостоятельной работы

Материал, законспектированный на лекциях, необходимо регулярно прорабатывать и дополнять сведениями из других источников литературы, представленных не только в программе дисциплины, но и в периодических изданиях.

При изучении дисциплины сначала необходимо по каждой теме прочитать рекомендованную литературу и составить краткий конспект основных положений, терминов, сведений, требующих запоминания и являющихся основополагающими в этой теме для освоения последующих тем курса. Для расширения знания по дисциплине рекомендуется использовать Интернет-ресурсы; проводить поиски в различных системах и использовать материалы сайтов, рекомендованных преподавателем.

При ответе на экзамене необходимо: продумать и четко изложить материал; дать определение основных понятий; дать краткое описание явлений; привести примеры. Ответ следует иллюстрировать схемами, рисунками и графиками.

8. Учебно-методическое и информационное обеспечение дисциплины обеспечение дисциплины

8.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. Аппаратно-программные средства геоинформационного обеспечения поддержки решений в рамках рационального природопользования / Н.Н. Попов, Л.В. Александрова, В.М. Абрамов, – СПб.: СпецЛит, 2016. - 51 с. (elib.rshu.ru/files_books/pdf/rid_f982b417571f4e62a275b6c34e00be1c.pdf)
2. Инновационные технологии геоинформационного обеспечения управления данными предприятия / Н.Н. Попов, Л.В. Александрова, В.М. Абрамов, – СПб.: СпецЛит, 2017. - 51 с. (elib.rshu.ru/files_books/pdf/rid_04837d21305f4a808ed637c5fda17db0.pdf)

Дополнительная литература

1. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс]. – Электрон. дан. – М.: ДМК Пресс, 2015. – 400 с. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=69955. – Загл. с экрана.
2. Коэльо Л.П. Построение систем машинного обучения на языке Python [Электронный ресурс] / Л.П. Коэльо, В. Ричарт. – Электрон. дан. – М.: ДМК Пресс, 2016. – 302 с. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=82818. – Загл. с экрана.

8.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

1. http://www.citforum.ru/database/case/index.shtml. (CASE - технологии. Современные методы и средства проектирования информационных систем).
2. http://books.listsoft.ru/book.asp?cod=123239&rp=1 (List SOFT. Каталог программ).

8.3. Перечень программного обеспечения

1. MS Windows
2. Google Chrome
3. VMWare

8.4. Перечень информационных справочных систем

Не используется

8.5. Перечень профессиональных баз данных
Электронно-библиотечная система elibrary

9. Материально-техническое обеспечение дисциплины

Материально-техническое обеспечение программы соответствует действующим санитарно-техническим и противопожарным правилам и нормам и обеспечивает проведение всех видов аудиторных занятий и самостоятельной работы студентов.

Учебная аудитории для проведения лекционных занятий - укомплектована проектором и компьютером, связанным с Интернетом.

Учебная лаборатория прикладных информационных технологий.

Учебная аудитория для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации - укомплектована специализированной (учебной) мебелью.

Помещение для самостоятельной работы – укомплектовано специализированной (учебной) мебелью, оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и выходом в ЭИОС.

Помещение для хранения и профилактического обслуживания учебного оборудования.

10. Особенности освоения дисциплины для инвалидов и лиц с ограниченными возможностями здоровья

Обучение обучающихся с ограниченными возможностями здоровья при необходимости осуществляется на основе адаптированной рабочей программы с использованием специальных методов обучения и дидактических материалов, составленных с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся (обучающегося).

При определении формы проведения занятий с обучающимся-инвалидом учитываются рекомендации, содержащиеся в индивидуальной программе реабилитации инвалида, относительно рекомендованных условий и видов труда.

При необходимости для обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья создаются специальные рабочие места с учетом нарушенных функций и ограничений жизнедеятельности.