МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования "Пермский государственный национальный исследовательский университет"

Кафедра радиоэлектроники и защиты информации

Авторы-составители: **Скляренко Максим Сергеевич**
Лунегов Игорь Владимирович

Рабочая программа дисциплины

ОБРАБОТКА И АНАЛИЗ БОЛЬШИХ ДАННЫХ

Код УМК 96225

Утверждено
Протокол №4
от «24» июня 2020 г.

Пермь, 2020

1. Наименование дисциплины

Обработка и анализ больших данных

2. Место дисциплины в структуре образовательной программы

Дисциплина входит в обязательную часть Блока « Б.1 » образовательной программы по направлениям подготовки (специальностям):

Направление: **01.03.02** Прикладная математика и информатика

направленность Инженерия программного обеспечения

3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Обработка и анализ больших данных** у обучающегося должны быть сформированы следующие компетенции:

01.03.02 Прикладная математика и информатика (направленность : Инженерия программного обеспечения)

ПК.5 Способен разрабатывать требования и проектировать программное обеспечение, в том числе интеллектуальные информационные системы

Индикаторы

ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования

4. Объем и содержание дисциплины

Направления подготовки	01.03.02 Прикладная математика и информатика (направленность: Инженерия программного обеспечения)
форма обучения	очная
№№ триместров, выделенных для изучения дисциплины	7
Объем дисциплины (з.е.)	4
Объем дисциплины (ак.час.)	144
Контактная работа с преподавателем (ак.час.), в том числе:	56
Проведение лекционных занятий	28
Проведение практических занятий, семинаров	28
Самостоятельная работа (ак.час.)	88
Формы текущего контроля	Входное тестирование (1) Защищаемое контрольное мероприятие (2) Итоговое контрольное мероприятие (2)
Формы промежуточной аттестации	Экзамен (7 триместр)

5. Аннотированное описание содержания разделов и тем дисциплины

Обработка и анализ больших данных

Основы технологий BigData

BigData
Что такое большие данные? Краткая историческая справка развития больших данных. Основные идеи организации хранения и обработки больших данных.

Основы Hadoop
Обзор Hadoop: HDFS, Hadoop, Map Reduce и др.

Настройка Hadoop
Запуск виртуальной машины Cloudera для работы с Hadoop. Изучение основ менеджера служб Cloudera Manager, Hue. Слушатели настраивают на своем ПК или в облаке инфраструктуру для последующей работы с Hadoop.

Архитектуры распределенного хранения данных

Основы распределенного хранения
Обзор Hadoop: HDFS, Hadoop, Map Reduce и др.

HDFS и MapReduce
Настройка HDFS, разработка простейшего приложения MapReduce.
Разрабатотка MapReduce- приложение по обработке файла в HDFS (на Java или Python).

Распределенные базы данных
Распределенные базы данных, архитектура Hive, архитектура HBase. Основы работы с HBase, запросы к Hive. Слушатели разрабатывают приложение-клиент для HBase.

Основы разработки приложений обработки данных в оперативной памяти на примере Apache Spark

Обзор ApacheSpark
Обработка данных в ОЗУ и Apache Spark.

Разработка под ApacheSpark
Запуск Spark-приложение на YARN, настройка подключения к Spark приложению из Python.
Разрабаотка приложения обработки данных на ApacheSpark

Обработка потоковых данных

Архитектура систем обработки потоковых данных на примере Apache Flume
Понятие потоковых данных. Архитектура распределённых систем обработки потоковых данных на примере Apache Flume. Отличия от стандартных ETL процедур.

Работа с Apache Flume
Настройка Apache Flume для обработки данных из разных источников данных, таких как web сервисы, файлы, реляционные СУБД.

Итоговый тест
итоговый тест

6. Методические указания для обучающихся по освоению дисциплины

Освоение дисциплины требует систематического изучения всех тем в той последовательности, в какой они указаны в рабочей программе.

Основными видами учебной работы являются аудиторные занятия. Их цель - расширить базовые знания обучающихся по осваиваемой дисциплине и систему теоретических ориентиров для последующего более глубокого освоения программного материала в ходе самостоятельной работы. Обучающемуся важно помнить, что контактная работа с преподавателем эффективно помогает ему овладеть программным материалом благодаря расстановке необходимых акцентов и удержанию внимания интонационными модуляциями голоса, а также подключением аудио-визуального механизма восприятия информации.

Самостоятельная работа преследует следующие цели:

– закрепление и совершенствование теоретических знаний, полученных на лекционных занятиях;

- формирование навыков подготовки текстовой составляющей информации учебного и научного назначения для размещения в различных информационных системах;

- совершенствование навыков поиска научных публикаций и образовательных ресурсов, размещенных в сети Интернет;

- самоконтроль освоения программного материала.

Обучающемуся необходимо помнить, что результаты самостоятельной работы контролируются преподавателем во время проведения мероприятий текущего контроля и учитываются при промежуточной аттестации.

Обучающимся с ОВЗ и инвалидов предоставляется возможность выбора форм проведения мероприятий текущего контроля, альтернативных формам, предусмотренным рабочей программой дисциплины. Предусматривается возможность увеличения в пределах 1 академического часа времени, отводимого на выполнение контрольных мероприятий.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации.

При проведении текущего контроля применяются оценочные средства, обеспечивающие передачу информации, от обучающегося к преподавателю, с учетом психофизиологических особенностей здоровья обучающихся.

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

При самостоятельной работе обучающимся следует использовать:

– конспекты лекций;

– литературу из перечня основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля);

– текст лекций на электронных носителях;

– ресурсы информационно-телекоммуникационной сети "Интернет", необходимые для освоения дисциплины;

– лицензионное и свободно распространяемое программное обеспечение из перечня информационных технологий, используемых при осуществлении образовательного процесса по дисциплине;

– методические указания для обучающихся по освоению дисциплины.

8. Перечень основной и дополнительной учебной литературы

Основная:

1. Петрунин Ю. Ю. Информационные технологии анализа данных. Data analysis : учебное пособие / Ю. Ю. Петрунин. — 2-е изд. — М.: КДУ, 2010. — 292 с. : ил., табл. —ISBN 978-5-98227-701-5. — Текст : электронный // Электронно-библиотечная система БиблиоТех : [сайт]. https://psu.bibliotech.ru/Reader/Book/7107

2. Маккинли, Уэс Python и анализ данных / Уэс Маккинли ; перевод А. Слинкина. — 2-е изд. — Саратов : Профобразование, 2019. — 482 c. — ISBN 978-5-4488-0046-7. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/88752.html

3. Просто о больших данных:перевод с английского/Д. Гурвиц, А. Ньюджент, Ф. Халпер, М. Кауфман.-Москва:Эксмо,2015, ISBN 978-5-699-85807-1.-400.-Глоссарий: с. 354-368. - Указ.: с. 369-391

Дополнительная:

1. Мухаметзянов, Р. Р. Основы программирования на Java : учебное пособие / Р. Р. Мухаметзянов. — Набережные Челны : Набережночелнинский государственный педагогический университет, 2017. — 114 c. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/66812.html

2. Карпов, А. С. Теоретические основы и практические подходы построения распределенных вычислительных систем : учебно-методическое пособие / А. С. Карпов. — Москва : Российский государственный университет инновационных технологий и предпринимательства, 2012. — 48 c. — ISBN 978-5-98427-047-2. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/33843

9. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

https://www.youtube.com/playlist?list=PLys0IdlMg6Xc5oJkLkoUs6pF6ij0px8Cr Видеолекции по теории распределённых вычислений

https://hadoop.apache.org/docs/stable/ Официальная документация на Hadoop

https://cwiki.apache.org/confluence/display/Hive/LanguageManual Официальная документация на Hive

https://hbase.apache.org/book.html Официальная документация на HBase

https://flume.apache.org/documentation.html Официальная документация на Apache Flume

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

Образовательный процесс по дисциплине **Обработка и анализ больших данных** предполагает использование следующего программного обеспечения и информационных справочных систем:
1) презентационные материалы (слайды по темам лекционных и практических занятий);
2) доступ в режиме on-line в Электронную библиотечную систему (ЭБС);
3) доступ в электронную информационно-образовательную среду университета;
4) интернет-сервисы и электронные ресурсы (поисковые системы, электронная почта);
Перечень необходимого лицензионного и (или) свободно распространяемого программного обеспечения
1. Проигрыватели виртуальных машин VirtualBox и VMWare Player (VMware Workstation).Пакеты офисных программ (тестовые процессоры, табличные редакторы, программы для создания презентаций и др.).
2. C++ Builder или C#, MS Visual Studio с фреймворком .net минимум версии 4.0
3. Операционная система ALT Linux;
4. Офисный пакет приложений «LibreOffice».

При освоении материала и выполнения заданий по дисциплине рекомендуется использование материалов, размещенных в Личных кабинетах обучающихся ЕТИС ПГНИУ (**student.psu.ru**).

При организации дистанционной работы и проведении занятий в режиме онлайн могут использоваться:

система видеоконференцсвязи на основе платформы BigBlueButton (https://bigbluebutton.org/).

система LMS Moodle (http://e-learn.psu.ru/), которая поддерживает возможность использования текстовых материалов и презентаций, аудио- и видеоконтент, а так же тесты, проверяемые задания, задания для совместной работы.

система тестирования Indigo (https://indigotech.ru/).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Для проведения лекционных занятий:
Аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.

Для проведения лабораторных занятий – Компьютерный класс, оснащенный персональными ЭВМ и

соответствующим программным обеспечением. Состав оборудования определен в Паспорте Компьютерного класса.

Аудитории для проведения текущего контроля;
Компьютерный класс, оснащенный персональными ЭВМ и соответствующим программным обеспечением. Состав оборудования определен в Паспорте компьютерного класса.

Аудитории для групповых (индивидуальных) консультаций;
Аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.

Аудитория для самостоятельной работы:
Аудитория оснащенная компьютерной техникой с возможностью подключения к сети «Интернет», обеспеченная доступом в электронную информационно-образовательную среду университета.
Помещения Научной библиотеки ПГНИУ

Помещения научной библиотеки ПГНИУ для обеспечения самостоятельной работы обучающихся:

1. Научно-библиографический отдел, корп.1, ауд. 142. Оборудован 3 персональными компьютера с доступом к локальной и глобальной компьютерным сетям.
2. Читальный зал гуманитарной литературы, корп. 2, ауд. 418. Оборудован 7 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
3. Читальный зал естественной литературы, корп.6, ауд. 107а. Оборудован 5 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
4. Отдел иностранной литературы, корп.2 ауд. 207. Оборудован 1 персональным компьютером с доступом к локальной и глобальной компьютерным сетям.
5. Библиотека юридического факультета, корп.9, ауд. 4. Оборудована 11 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
6. Читальный зал географического факультета, корп.8, ауд. 419. Оборудован 6 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

Все компьютеры, установленные в помещениях научной библиотеки, оснащены следующим программным обеспечением:

Операционная система ALT Linux;

Офисный пакет Libreoffice.

Справочно-правовая система «КонсультантПлюс»

Фонды оценочных средств для аттестации по дисциплине
Обработка и анализ больших данных

Планируемые результаты обучения по дисциплине для формирования компетенции. Индикаторы и критерии их оценивания

ПК.5

Способен разрабатывать требования и проектировать программное обеспечение, в том числе интеллектуальные информационные системы

Компетенция (индикатор)	Планируемые результаты обучения	Критерии оценивания результатов обучения
ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования	ЗНАТЬ: Основы технологий BigData на базе платформы Hadoop. УМЕТЬ: Создавать распределенные хранилища данных, создавать настраивать MapReduce, разрабатывать приложения ApacheSpark. ВЛАДЕТЬ НАВЫКАМИ: работы со стеком технологий BigData: HDFS, MapReduce, Apache Hive, Apache Spark и др.	**Неудовлетворител** НЕ ЗНАЕТ: Основы технологий BigData на базе платформы Hadoop. НЕ УМЕЕТ: Создавать распределенные хранилища данных, создавать настраивать MapReduce, разрабатывать приложения ApacheSpark. НЕ ВЛАДЕЕТ НАВЫКАМИ: работы со стеком технологий BigData: HDFS, MapReduce, Apache Hive, Apache Spark и др. **Удовлетворительн** ЧАСТИЧНО ЗНАЕТ: Основы технологий BigData на базе платформы Hadoop. ЧАСТИЧНО УМЕЕТ: Создавать распределенные хранилища данных, создавать настраивать MapReduce, разрабатывать приложения ApacheSpark. ЧАСТИЧНО ВЛАДЕЕТ НАВЫКАМИ: работы со стеком технологий BigData: HDFS, MapReduce, Apache Hive, Apache Spark и др. **Хорошо** ЗНАЕТ: Основы технологий BigData на базе платформы Hadoop. УМЕЕТ: Создавать распределенные хранилища данных, создавать настраивать MapReduce, разрабатывать приложения ApacheSpark. ВЛАДЕЕТ НАВЫКАМИ: работы со стеком технологий BigData: HDFS, MapReduce, Apache Hive, Apache Spark и др.

Компетенция (индикатор)	Планируемые результаты обучения	Критерии оценивания результатов обучения
		Отлично В ПОЛНОМ ОБЪЕМЕ ЗНАЕТ: Основы технологий BigData на базе платформы Hadoop. В ПОЛНОЙ МЕРЕ УМЕЕТ: Создавать распределенные хранилища данных, создавать настраивать MapReduce, разрабатывать приложения ApacheSpark. СВОБОДНО ВЛАДЕЕТ НАВЫКАМИ: работы со стеком технологий BigData: HDFS, MapReduce, Apache Hive, Apache Spark и др.

Оценочные средства текущего контроля и промежуточной аттестации

Схема доставки : Базовая

Вид мероприятия промежуточной аттестации : Экзамен
Способ проведения мероприятия промежуточной аттестации : Оценка по дисциплине в рамках промежуточной аттестации определяется на основе баллов, набранных обучающимся на контрольных мероприятиях, проводимых в течение учебного периода.
Максимальное количество баллов : 100

Конвертация баллов в отметки

«отлично» - от 81 до 100
«хорошо» - от 61 до 80
«удовлетворительно» - от 50 до 60
«неудовлетворительно» / «незачтено» менее 50 балла

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
Входной контроль	Настройка Hadoop **Входное тестирование**	знание основ БД и ООП
ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования	HDFS и MapReduce **Защищаемое контрольное мероприятие**	знание основ BigData, навыки MaPReuce, знания и владение технологией HDFS
ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования	Разработка под ApacheSpark **Итоговое контрольное мероприятие**	навыки разработки приложений ApacheSpark. ВЛАДЕТЬ НАВЫКАМИ: Apache Hive, Apache Spark
ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования	Работа с Apache Flume **Защищаемое контрольное мероприятие**	понимание принципов обработки потоковых данных.владение технологией Apache Flume.

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ПК.5.2 Проектирует используемые структуры данных и программные интерфейсы, разрабатывает алгоритмы и оценивает эффективность их использования	Итоговый тест **Итоговое контрольное мероприятие**	контролируются в форме теста знания курса

Спецификация мероприятий текущего контроля

Настройка Hadoop

Продолжительность проведения мероприятия промежуточной аттестации: **1 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **0**
Проходной балл: **0**

Показатели оценивания	Баллы
студент владеет навыками написания SQL запросов	10
студент владеет базовым синтаксисом одного из объектно-ориетированных языков	10
студент знание основы БД	5
студент понимание принципы ООП	5

HDFS и MapReduce

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**
Проходной балл: **15**

Показатели оценивания	Баллы
выполнены все требования задания 2	20
выполнены все требования задания 1	10
минимальное выполнение требований задания 2	8
минимальное выполнение требований задания 1	7

Разработка под ApacheSpark

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**
Проходной балл: **15**

Показатели оценивания	Баллы
Выполнено задание 2 в полном объеме	20
Выполнено задание 2 в минимальном объеме	10

Выполнено задание 1 в полном объеме	10
Выполнено задание 1 в минимальном объеме	5

Работа с Apache Flume

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **15**
Проходной балл: **7**

Показатели оценивания	Баллы
Задание 1 выполнено в полном объеме	15
Задание 1 выполнено не в полном объеме	7

Итоговый тест

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**
Проходной балл: **12.5**

Показатели оценивания	Баллы
вопрос 01	1
Вопрос 20	1
вопрос 3	1
вопрос 4	1
вопрос 5	1
вопрос 6	1
вопрос 7	1
вопрос 8	1
Вопрос 9	1
Вопрос 10	1
Вопрос 11	1
Вопрос 12	1
Вопрос 13	1
Вопрос 14	1
Вопрос 15	1
Вопрос 16	1
Вопрос 17	1
Вопрос 18	1
Вопрос 19	1
вопрос 02	1