Министерство науки и высшего образования Российской Федерации
ФГБОУ ВО «Бурятский государственный университет»
Институт математики, физики и компьютерных наук
Кафедра информационных систем и методов искусственного интеллекта

Утверждена на заседании
Ученого совета ИМФКН
«___» _____ 20__г.
Протокол №__

Рабочая программа дисциплины

Технологии сбора и обработки больших данных

Направление подготовки
01.04.02 Прикладная математика и информатика

Квалификация
магистр

Форма обучения
очная

Улан-Удэ
2023

Пояснительная записка

Цели освоения дисциплины
Дать представление о современном инструментарии обработки больших данных.

Место дисциплины в структуре образовательной программы
Дисциплина относится к вариативной части, является обязательной дисциплиной. Код дисциплины Б1.В.ОД.4

В результате освоения дисциплины студент должен:

Планируемые результаты обучения по дисциплине и индикаторы достижения компетенций.

Знать:
Общие принципы работы с большими данными; Основные концепции вычислительных технологий больших данных; Типовые задачи обработки больших данных

Уметь:
Осуществлять сбор и хранение больших данных; применять основные концепции вычислительных технологий больших данных; решать типовые задачи обработки больших данных с применением современного инструментария

Владеть:
Навыками работы со стеком технологий Hadoop; навыками работы с конкретными инструментами стека технологий Hadoop; навыками решения типовых задач

Планируемые результаты освоения образовательной программы:

- ПК-3 - Способен руководить разработкой технических спецификаций и проектированием программного обеспечения
  - ПК-3.2 - Ориентируется в возможностях существующей программно-технической архитектуры
  - ПК-3.3 - Применяет методологии и средства проектирования программного обеспечения
  - ПК-3.4 - Применяет методы и средства проектирования баз данных
  - ПК-3.1 - Применяет методы и средства разработки технических спецификаций программного обеспечения

Объем дисциплины в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем и на самостоятельную работу обучающихся

Общая трудоемкость дисциплины составляет 9 зачетные единицы, 324 часа.

№	Название разделов дисциплины	Самостоятельная работа	Лабораторная работа
Семестр 2		60	12
1	Стек технологий Hadoop и методы хранения больших данных-2	34	6
2	Стек технологий Hadoop и методы хранения больших данных-1	26	6
Семестр 3		158	22
1	Фреймворк Apache Spark и иной инструментарий обработки данных	158	22

Тематическое планирование курса

Стек технологий Hadoop и методы хранения больших данных-2
Семестр 2

Обработка данных

Самостоятельная работа. 2(0) ч. Принципы работы с Apache Hive
Самостоятельная работа. 2(0) ч. Принципы работы с Apache Pig
Самостоятельная работа. 2(0) ч. HQL (Hive query language)
Самостоятельная работа. 2(0) ч. Язык Pig Latin. Структуры данных Tuple и Bag. Базовые функции Pig Latin
Лабораторная работа. 2 ч. Язык запросов Hive query language
Самостоятельная работа. 4(0) ч. Pig Latin
Самостоятельная работа. 2(0) ч. Настройка и запуск Hadoop и Pig
Самостоятельная работа. 6(0) ч. Типы данных в Pig Latin

Иные технологии

Самостоятельная работа. 2(0) ч. Vowpal Wabbit
Лабораторная работа. 2(0) ч. Vowpal Wabbit
Самостоятельная работа. 8 ч. Vowpal Wabbit
Самостоятельная работа. 2(0) ч. Фреймворк Caffe и концепция deep learning
Лабораторная работа. 2 ч. Работа с Caffe
Самостоятельная работа. 2(0) ч. Работа с Caffe

Стек технологий Hadoop и методы хранения больших данных-1

Семестр 2

Сбор и хранение данных

Самостоятельная работа. 2(0) ч. модель вычислений Map-Reduce
Самостоятельная работа. 2(0) ч. Файловая система HDFS
Самостоятельная работа. 2(0) ч. Посредник YARN и планировщик Oozie
Самостоятельная работа. 4 ч. Планировщик Oozie
Самостоятельная работа. 2(0) ч. NoSQL базы данных
Самостоятельная работа. 2(0) ч. NoSQL база данных Cassandra
Лабораторная работа. 2(0) ч. Импорт данных
Самостоятельная работа. 2(0) ч. Импорт с помощью Apache Sqoop
Самостоятельная работа. 2(0) ч. Импорт с помощью Apache Flume
Самостоятельная работа. 4(0) ч. Импорт данных в HDFS
Самостоятельная работа. 4(0) ч. NoSQL база данных HBase
Лабораторная работа. 2 ч. Реализация Map-Reduce в Hadoop
Лабораторная работа. 2(0) ч. Файловая система HDFS

Фреймворк Apache Spark и иной инструментарий обработки данных

Семестр 3

Apache Spark

Самостоятельная работа. 2(0) ч. Модель вычислений Resilient Distributed Dataset (RDD)
Самостоятельная работа. 16 ч. Spark SQL
Лабораторная работа. 6 ч. SparkSQL и Hive
Самостоятельная работа. 2(0) ч. Spark MLlib
Самостоятельная работа. 2(0) ч. Spark GraphX
Самостоятельная работа. 2(0) ч. Spark Streaming
Самостоятельная работа. 8 ч. Операции над RDD
Самостоятельная работа. 8 ч. Hive для DataFrame
Самостоятельная работа. 16 ч. Платформа Kaggle
Лабораторная работа. 4(0) ч. Cloudera’s Distribution including Apache Hadoop (CDH) и подобные
Лабораторная работа. 6 ч. Абстракция DataFrame
Лабораторная работа. 6 ч. Работа с Spark MLlib
Самостоятельная работа. 16 ч. Основные возможности Spark MLlib
Самостоятельная работа. 8 ч. Необходимые сведения из Scala. Основные структуры данных: Списки, Наборы, Кортеж, Карты
Самостоятельная работа. 16 ч. Необходимые сведения из Scala. Функциональные Комбинаторы:, map, foreach, filter, zip, partition, find, drop и dropWhile, foldRight и foldLeft, flatten, flatMap, Обобщенные функциональные комбинаторы
Самостоятельная работа. 2(0) ч. Платформа Kaggle
Самостоятельная работа. 16 ч. Работа с Spark GraphX
Самостоятельная работа. 16 ч. Работа с Spark Streaming
Самостоятельная работа. 4(0) ч. Spark MLlib
Самостоятельная работа. 14 ч. Hadoop Cloudera
Самостоятельная работа. 10 ч. Решение задач Kaggle

БРС

Семестр	Контрольные точки	Баллы
2	**Текущий контроль** в разделе «Стек технологий Hadoop и методы хранения больших данных-2»	
	Разработка проекта	30
	Составление структурно-логической схемы	10
2	**Текущий контроль** в разделе «Стек технологий Hadoop и методы хранения больших данных-1»	
	Разработка проекта	30
	Составление структурно-логической схемы	10
2	**Зачет**	
	Разработка проекта	20

Семестр	Контрольные точки	Баллы
	Итого за семестр 2: 100	
3	**Текущий контроль** в разделе «Фреймворк Apache Spark и иной инструментарий обработки данных»	
	Разработка проекта	40
	Составление структурно-логической схемы	20
3	**Экзамен**	
	Разработка проекта	40
	Итого за семестр 3: 100	

Учебно-методическое и информационное обеспечение учебного процесса

Образовательные технологии (в том числе на занятиях, проводимых в интерактивных формах).
При изучении данного курса применяются как традиционные (лекции, практические занятия, экзамен), так и инновационные образовательные технологии, которые реализуются в учебном процессе в активных и интерактивных формах проведения занятий, из которых можно выделить следующие:
1) лекция-дискуссия при рассмотрении тем "NoSQL базы данных","Платформа Kaggle".
2) метод группового решения задач при изучении тем "Работа с Caffe","Работа с Spark MLlib".
3) перекрестная самопроверка при изучении тем "Настройка и запуск Hadoop и Pig","Hadoop Cloudera".

Учебно-методические материалы, в том числе методические указания для обучающихся по освоению дисциплины
К современному специалисту общество предъявляет достаточно широкий перечень требований, среди которых немаловажное значение имеет наличие у выпускников определенных способностей и умения самостоятельно добывать знания из различных источников, систематизировать полученную информацию, давать оценку конкретной ситуации. Формирование такого умения происходит в течение всего периода обучения через участие студентов в лабораторных, практических занятиях, выполнение заданий и тестов, написание курсовых и выпускных квалификационных работ. При этом самостоятельная работа студентов играет решающую роль в ходе всего учебного процесса.
Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

По данной дисциплине разработано учебно-методическое обеспечение для самостоятельной работы обучающихся и размещено в электронной информационно-образовательной среде университета (личном кабинете студента).

- СРС_маг_010401_Математика_2015_Технологии сбора и обработки больших данных_05151м.doc
  Режим доступа: https://my.bsu.ru/content/file/3/36/363/103424_srs_mag_010401_matematika_2015_tehnologii-sbora-i-obrabotki-bolshih-dannih_05151m.doc

Учебно-методические материалы, в том числе методические указания для обучающихся по освоению дисциплины
— На лекционных занятиях студент слушает доклад преподавателя, составляет конспект лекции. Во время лекции студенту рекомендуется делать отметки на полях тетради, касающиеся того теоретического материала, который вызвал затруднения в понимании. После лекции трудности необходимо устранить путем консультации у преподавателя или самостоятельной работы с рекомендованной учебной литературой.
— На лабораторных занятиях студенту предлагается ряд заданий по теме, прослушанной на лекции. По заданию необходимо решить поставленную задачу, пользуясь персональным компьютером и необходимым программным обеспечением. Контроль знаний, умений и навыков студентов осуществляется путём проведения на лабораторных занятиях компьютерных тестов, сдачи проектов. Студенту, выполнившему то или иное задание на низкое количество баллов, по согласованию с преподавателем, необходимо выполнить работу над ошибками для улучшения результатов.
— Предусмотрена самостоятельная работа студентов, в рамках которой студентом должны выполняться работа над проектами, анализ требований, ошибок, перспектив и т.п.. Теоретический материал студентом должен быть проработан с использованием конспектов лекций (если в плане предусмотрены лекции) или рекомендуемой учебной литературы.

Советы по планированию и организации времени, необходимого для изучения дисциплины.
Рекомендуется следующим образом организовать время, необходимое для изучения дисциплины:
В день после лекционного занятия в течение 1 астрономического часа изучить лекцию, выявить моменты, где имеются вопросы и недостаточное понимание. По этим вопросам изучить рекомендованную литературу, устранить непонимание. В день перед проведением практического занятия в течение 30 минут повторить

пройденный материал, предшествующий теме занятия.
При изучении дисциплины очень полезно самостоятельно изучать материал, который еще не прочитан на лекции, не применялся на лабораторном или практическом занятии. Тогда предстоящее занятие будет гораздо понятнее.
Рекомендации по работе с литературой. Теоретический материал курса становится более понятным, когда дополнительно к прослушиванию лекций, изучению конспектов, изучаются и книги по изучаемой дисциплине. Литературу по дисциплине рекомендуется изучать в библиотеке. Полезно использовать несколько учебников по дисциплине. Рекомендуется, кроме «заучивания» материала, добиться состояния понимания изучаемой темы дисциплины. С этой целью рекомендуется после изучения очередного параграфа выполнить несколько простых упражнений на данную тему.

Советы по подготовке к экзамену:
1) Ознакомиться с процедурой проведения экзамена, ознакомиться со списком вопросов для подготовки к экзамену. Повторить весь материал пройденных лекций (если в плане были лекции), проработать конспекты, рекомендованную учебную литературу. Составить конспекты ответов на экзаменационные вопросы.
2) Собрать итоговые выполненные лабораторные и проекты (если в плане были лабораторные и проекты), если произведено их улучшение согласно замечаниям преподавателя. В рамках экзамена/зачета возможен пересмотр преподавателем баллов (по усмотрениюпреподавателя, с учетом общей работы студента во время курса), выставленных ранее за выполненные задания, при условии получения улучшения результатов

Оценочные средства

По данной дисциплине разработаны оценочные средства, критерии их оценивания, а также методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций.

- СРС_маг_010401_Математика_2015_Технологии сбора и обработки больших данных_05151м.doc
- ФОС_маг_010401_Математика_2015_Технологии сбора и обработки больших данных_05151м.docx

Список литературы

Перечень основной и дополнительной литературы, необходимой для освоения дисциплины.

Основная

1. ОСНОВЫ ИСПОЛЬЗОВАНИЯ И ПРОЕКТИРОВАНИЯ БАЗ ДАННЫХ: Учебник/Илюшечкин В.М.. —М.: Издательство Юрайт, 2016. —213 с.
   Режим доступа: http://www.biblio-online.ru/book/1C650A7F-DC7D-4834-998E-42D06FC8EF33
2. ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ: Учебник и практикум/Миркин Б.Г.. —М.: Издательство Юрайт, 2016. —174 с.
   Режим доступа: http://www.biblio-online.ru/book/E486A3B0-1AE9-4179-8D48-FA24B626C7C9
3. АНАЛИЗ ДАННЫХ: Учебник/Мхитарян В.С. - Отв. ред.. —М.: Издательство Юрайт, 2016. —490 с.
   Режим доступа: http://www.biblio-online.ru/book/AF1D197F-1759-422E-9593-8B43E2D1093B

Дополнительная

1. Применение искусственных нейронных сетей и системы остаточных классов в криптографии/[Н.И. Червяков и др.]. —Москва: Физматлит, 2012. —279 с.
   Режим доступа: http://e.lanbook.com/books/element.php?pl1_cid=25&pl1_id=5300
2. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных: научно-популярная литература/П. Флах ; пер. с англ. А. А. Слинкин. —Москва: ДМК Пресс, 2015. —400 с.
   Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=69955
3. Scala для нетерпеливых/К. Хостманн. —Москва: ДМК Пресс, 2013. —408 с.
   Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=66473
4. Python и анализ данных: научно-популярная литература/У. Маккинли ; [пер. с англ. А. А. Слинкин]. —Москва: ДМК Пресс, 2015. —482 с.
   Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=73074
5. Python на практике/Саммерфилд М.. —Москва: ДМК Пресс, 2014
   Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=66480

Перечень ресурсов информационно-коммуникационной сети «Интернет», необходимых для освоения дисциплины
https://www.ibm.com/developerworks/ru/library/os-log-process-hadoop/
https://www.kaggle.com/c/digit-recognizer

Перечень информационных технологий, используемых при осуществлении образовательного

процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем (при необходимости)
http://hadoop.apache.org/

Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине
Учебная аудитория для проведения лекционных занятий (1205, 1209, 1208, 1211); Учебная аудитория для проведения занятий лабораторного типа с доступом в Интернет (1313, 1312, 1316); Помещение для самостоятельной работы с доступом в Интернет (1312, 1313, 1316);
Учебная аудитория для проведения индивидуальных и групповых консультаций (1313, 1312, 1316); Учебная аудитория для проведения текущей и промежуточной аттестации (1312, 1316, 1313); Требуемый перечень программного обеспечения: ОС Windows/Ubuntu; Интегрированная среда разработки Code::Blocks 13.12 и выше.

Автор: Цыбиков Анатолий Сергеевич

Рабочая программа обсуждена на заседании кафедры _____ от «__» _____ 20__г. Протокол №__.