ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНЫХ КИБЕРНЕТИЧЕСКИХ СИСТЕМ

КАФЕДРА КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕХНОЛОГИЙ

ОДОБРЕНО

**РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ**

НАУКА О ДАННЫХ И АНАЛИЗ БОЛЬШИХ ДАННЫХ (DATA SCIENCE AND BIG DATA ANALYTICS)

|  |  |
| --- | --- |
| Направление подготовки (специальность) | 09.04.01 Информатика и вычислительная техника |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Семестр** | **Трудоемкость, кред.** | **Общий объем курса, час.** | **Лекции, час.** | **Практич. занятия, час.** | **Лаборат. работы, час.** | **СРС, час.** | **КСР, час.** | **Форма(ы) контроля, экз./зач./КР/КП** |
| 2 | 3 | 108 | 30 | 0 | 15 | 63 | 0 | З |
| Итого | 3 | 108 | 30 | 0 | 15 | 63 | 0 |  |

АННОТАЦИЯ

В настоящее время процесс генерации новых данных приобрел «лавинный» характер. В результате всеобщей информатизации и активного совершенствования современных вычислительных мощностей темпы роста объема хранимых данных можно охарактеризовать как крайне высокие. Помимо, собственно, хранения данные нуждаются в обработке. При этом анализ данных в общем случае не является разовым. В следствии развития методов DataMining и концепции BigData, старые данные обычно сохраняются в первозданном виде для последующего анализа с учетом новых тенденций и подходов, которые отсутствовали на момент первоначального анализа.

В качестве примера таких данных можно привести: экспериментальные данные, статистику обращений к веб-сервисам, метеоданные, поток информации из социальных сетей, микроблогов и др. В каждом из представленных случаев возможно осуществить параллельную обработку данных на распределенной вычислительной системе. При этом, необходимы системы хранения, способны обрабатывать миллионы довольно простых по форме запросов.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ УЧЕБНОЙ ДИСЦИПЛИНЫ

Целями освоения учебной дисциплины «Data Science and Big Data Analytic» являются формирование у студентов целостного представления о принципах разработки, анализа и реализации параллельных алгоритмов обработки структур данных; освоение студентами технологий разработки программных продуктов для суперкомпьютерных систем обработки и хранения больших объемов данных данных.

2. МЕСТО УЧЕБНОЙ ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

Учебная дисциплина входит в профессиональный цикл, раздел M2. Для освоения дисциплины необходимы компетенции, сформированные у обучающихся в результате освоения дисциплин общенаучного цикла М1 (в частности, дисциплины разделов М1.Б.2 «Методы оптимизации», М1.В.1 «Архитектура информационных систем» и М1.В.2. «Параллельные вычисления»).

Сформированные при изучении данной дисциплины компетенции необходимы для выполнения научно-исследовательской работы и подготовки магистерских диссертаций по проблематике организации высокопроизводительных систем анализа данных.

3. ФОРМИРУЕМЫЕ КОМПЕТЕНЦИИ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Универсальные и(или) общепрофессиональные компетенции:

|  |  |
| --- | --- |
| Код и наименование компетенции | Код и наименование индикатора достижения компетенции |

4. СТРУКТУРА И СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Разделы учебной дисциплины, их объем, сроки изучения и формы контроля:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **№ п.п** | **Наименование раздела учебной дисциплины** | **Недели** | **Лекции/ Практ. (семинары )/ Лабораторные работы, час.** | **Обязат. текущий контроль (форма\*, неделя)** | **Максимальный балл за раздел\*\*** | **Аттестация раздела (форма\*, неделя)** | **Индикаторы освоения компетенции** |
|  | *2 Семестр* |  |  |  |  |  |  |
| 1 | Организация Big Data - систем | 1-8 | 16/0/8 |  | КИ-8 | 25 |  |
| 2 | Методы и средства организации распределенного анализа данных | 9-15 | 14/0/7 |  | КИ-15 | 25 |  |
|  | *Итого за 2 Семестр* |  | 30/0/15 |  |  | 50 |  |
|  | **Контрольные мероприятия за 2 Семестр** |  |  |  | З | 50 |  |

\* – сокращенное наименование формы контроля

\*\* – сумма максимальных баллов должна быть равна 100 за семестр, включая зачет и (или) экзамен

Сокращение наименований форм текущего контроля и аттестации разделов:

|  |  |
| --- | --- |
| **Обозначение** | **Полное наименование** |
| КИ | Контроль по итогам |
| З | Зачет |

КАЛЕНДАРНЫЙ ПЛАН

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Недели** | **Темы занятий / Содержание** | **Лек., час.** | **Пр./сем., час.** | **Лаб., час.** |
|  | *2 Семестр* | 30 | 0 | 15 |
| **1-8** | **Организация Big Data - систем** | 16 | 0 | 8 |
| 1 | **Вводное занятие** Общие вопросы организации BigData-систем. Проблемы организации BigData-систем. Классификация характера задач для выбора подходящей реализации. Методы декомпозиции BigData-задач. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 2 |  | 1 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 2 | **Архитектура BigData систем. Часть 1** Метрики производительности BigData-систем, состав и назначение подсистем внутренних/внешних обменов, подсистем online/offline аналитики. Особенности использования внешних и внутренних каналов связи при передиче больших объемов данных. Основные метриками для оценки производительности BigData-систем. Способы организации внешних сетевых интерфейсов BigData-систем и построения модулей для online/offline обработки. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 2 |  | 1 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 3 | **Архитектура BigData систем. Часть 2** Способы интеграции подсистем online/offline обработки, подсистем индексирования и хранения данных. Введение в проблематику построения систем мониторинга вычислительных ресурсов. Состав и назначение (на примере системы мониторинга вычислительных ресурсов) модулей online/offline обработки, подсистем индексирования и хранения данных.  Стандартные средства разработки для организации процессов поточной обработки больших объемов данных.  Принципы построения интеграционного кластерного интерфейса для организации взаимодействия распределенных систем. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 2 |  | 1 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 4 - 5 | **Инфраструктура BigData систем** Системы построения виртуализированных сетевых инфраструктур. Системы легковесной контейнеризации. Вопросы виртуализации сетевых функций и построения программно определяемых сетей.  Назначение технологий SDN/NFV. Механизмы работы систем управления виртуализированными контейнерами.  Технология контейнеризации OpenShift для развертывания программных решений в облаке.  Принципы использования технологий OpenStack и OpenShift для организации сетей виртуальных машин и систем управления контейнерами. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 4 |  | 2 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 6 - 7 | **Пакетная распределенная обработка больших объемов данных** Технология организации пакетной обработки больших объемов данных MapReduce.  Фазы и действия, выполняемые программным каркасом Hadoop при исполнении MapReduce-программы.  Разработка Java-приложения, использующие Hadoop для выполнения пакетной обработки данных по таймеру.  Управления жизненным циклом Hadoop-кластера. Способами развертывания MapReduce-программ в Hadoop-окружении. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 4 |  | 2 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 8 | **Оценка производительности BigData-систем** Элементы теории массового обслуживания для оценки производительности распределенных вычислительных BigData-систем  Основы теории массового обслуживания для расчет интенсивности поступления запросов на каждый узел сети.  Оценка необходимой производительности узлов сети исходя из прогнозируемой нагрузки.  Математический аппарат теории вероятностей и теории массового обслуживания для построения моделей потоков данных в BigData-системах. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 2 |  | 1 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| **9-15** | **Методы и средства организации распределенного анализа данных** | 14 | 0 | 7 |
| 9 - 10 | **Асинхронная обработка больших объемов данных** Принципы работы с технологией Apache Spark для выполнения асинхронных вычислительных операций и системы очередей для управления асинхронными процессами в BigData-системах. Состав кластера Apache Spark и Apache Kafka. Организация загрузки/выгрузки информации в системах Apache Spark и Apache Kafka. Администрирование систем Apache Spark и Apache Kafka, способами организации асинхронного взаимодействия нескольких вычислительных задач. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 4 |  | 2 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 11 - 12 | **Хранение больших объемов данных** Высокопроизводительные NoSQL-системы. Предпосылки, типы и характеристика. Состав и характеристики высокопроизводительных файловых систем на примере GFS, HDFS и NFS v4.1. Состав кластера Apache Cassandra. Средства обеспечения согласованности в высокопроизводительных системах хранения данных.  Определение необходимого типа системы хранения и схему упаковки данных в зависимости от задачи.  Принципы трансформации потока входящей информации в поток объектов хранения BigData-системы. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 4 |  | 2 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 13 - 14 | **Построение индексов** Модули индексации данных для BigData-систем. Назначение модулей индексации данных и требования к ним.  Проектирование отказоустойчивых высокопроизводительных модулей индексации данных, предназначенные для поддержки конкретных алгоритмов BigData-аналитики.  Принципы построения и оценки производительности подсистем индексации данных. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 4 |  | 2 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |
| 15 - 16 | **Библиотеки машинного обучения для BigData-систем** Алгоритмы машинного обучения, предназначенные для пакетной и поточной обработки.  Виды алгоритмов машинного обучения, пригодные для использования в offline-модулях, Виды алгоритмов машинного обучения, пригодные для использования в onfline-модулях. Применение гибридных вычислительных технологий в задачах машинного обучения. Способы повышения производительности алгоритмов машинного обучения с применением гибридных вычислительных технологий. | Всего аудиторных часов | Всего аудиторных часов | Всего аудиторных часов |
| 2 |  | 1 |
| Онлайн | Онлайн | Онлайн |
|  |  |  |

Сокращенные наименования онлайн опций:

|  |  |
| --- | --- |
| **Обозначение** | **Полное наименование** |
| ЭК | Электронный курс |
| ПМ | Полнотекстовый материал |
| ПЛ | Полнотекстовые лекции |
| ВМ | Видео-материалы |
| АМ | Аудио-материалы |
| Прз | Презентации |
| Т | Тесты |
| ЭСМ | Электронные справочные материалы |
| ИС | Интерактивный сайт |

5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Проведение лекций предусматривает использование технических средств обучения (ТСО) для показа презентаций, иллюстрации процесса разработки, отладки и профилирования программ.

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств по дисциплине обеспечивает проверку освоения планируемых результатов обучения (компетенций и их индикаторов) посредством мероприятий текущего, рубежного и промежуточного контроля по дисциплине.

Связь между формируемыми компетенциями и формами контроля их освоения представлена в следующей таблице:

|  |  |
| --- | --- |
| **Компетенция** | **Индикаторы освоения** |

Оценочные средства приведены в Приложении.

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

ОСНОВНАЯ ЛИТЕРАТУРА:

1. ЭИ A22 Advances in Data Mining. Applications and Theoretical Aspects : 16th Industrial Conference, ICDM 2016, New York, NY, USA, July 13-17, 2016. Proceedings, Cham: Springer International Publishing, 2016

2. ЭИ A22 Advances in Knowledge Discovery and Data Mining : 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I, Cham: Springer International Publishing, 2016

3. ЭИ Б 82 Основы работы с технологией CUDA : , Москва: ДМК Пресс, 2010

4. ЭИ С 18 Технология CUDA в примерах: введение в программирование графических процессоров : , Москва: ДМК Пресс, 2011

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА:

1. 510 А45 Алгоритмы : построение и анализ, Москва [и др.]: Вильямс, 2011

2. 004 П18 Параллельные вычисления на GPU. Архитектура и программная модель CUDA : учебное пособие, Москва: Издательство Московского университета, 2012

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ:

Специальное программное обеспечение не требуется

LMS И ИНТЕРНЕТ-РЕСУРСЫ:

https://online.mephi.ru/

http://library.mephi.ru/

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Специальное материально-техническое обеспечение не требуется

9. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ СТУДЕНТОВ

Приведёны в приложении

Автор(ы):

|  |  |  |
| --- | --- | --- |
|  | Ровнягин Михаил Михайлович |  |

Рецензент(ы):

|  |  |  |
| --- | --- | --- |
|  | Васильев Н.П. |  |