Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и
образовательным инновациям
_____ О.И. Чуприс
«12» 07 2019 г.

Регистрационный № УД-6906 /уч.

ТЕХНОЛОГИИ И КОМПЬЮТЕРНЫЕ СИСТЕМЫ ОБРАБОТКИ ДАННЫХ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 80 09 Прикладная математика и информатика

профилизации:
Алгоритмы и системы обработки больших данных
Аналитическая логистика
Компьютерный анализ данных
Математическая кибернетика

2019 г.

Учебная программа составлена на основе ОСВО 1-31 80 09-2019 и учебных планов G31-072/уч., G31-073/уч., G31-074/уч., G31-075/уч. от 11.04.2019 г.

СОСТАВИТЕЛЬ:

С.В. Баханович – доцент кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета, кандидат физ.-мат. наук.

РЕЦЕНЗЕНТЫ:

П.И. Соболевский, главный научный сотрудник Института математики НАН Беларуси, доктор физико-математических наук, профессор;

А.В. Жерело – заместитель начальника Центра информационных технологий Белорусского государственного университета, кандидат физико-математических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дискретной математики и алгоритмики
(протокол № 15 от 18 апреля 2019 года);

Научно-методическим Советом БГУ

(протокол № 5 от 28 июня 2019 года).

Заведующий кафедрой
дискретной математики и алгоритмики_____ В.М. Котов

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Современный уровень подготовки специалистов в области информационных технологий обязан обеспечиваться хорошим знанием как основных концепций в данной предметной области, так и свободным владением новейшими инструментальными методами и средствами разработки сложных информационных систем. Современный этап развития науки и производства, экономики характеризуется наличием огромных объемов информации и необходимости их обработки, существованием задач, связанных с выполнением больших объемов вычислений. Все это делает необходимым изучение принципов построения компьютерных систем распределенного хранения и обработки данных, в том числе, суперкомпьютерных систем, а также технологий параллельного проектирования и программирования.

Учебная дисциплина «Технологии и компьютерные системы обработки данных» знакомит студентов магистратуры с основными тенденциями развития современных технологий, прежде всего, в области параллельных вычислений. Дисциплина знакомит с принципами построения компьютерных систем, предназначенных для хранения и обработки больших объемов информации.

Цель учебной дисциплины – ознакомление студентов магистратуры с основными принципами построения параллельных вычислительных систем и основными технологиями программирования, используемыми для разработки параллельных приложений, формирование умения использовать технологии и компьютерные системы распределенных вычислений при решении прикладных задач, связанных с обработкой больших объемов данных.

Задачи учебной дисциплины:

- Изучить современные компьютерные системы, предназначенные для обработки больших объемов информации; сформировать представления об архитектуре суперкомпьютеров различного типа и принципах их функционирования.
- Изучить парадигмы и освоить основные технологий организации параллельных вычислений.
- Использовать высокопроизводительные компьютерные системы при решении прикладных задач.

Место учебной дисциплины: в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к циклу государственный компонент и входит в модуль «Программная инженерия».

Программа составлена с учетом **межпредметных связей** с учебными дисциплинами. Основой для изучения учебной дисциплины являются учебные дисциплины I ступени высшего образования «Программирование», «Операционные системы», «Теория алгоритмов», «Архитектура

компьютеров», «Компьютерные сети». Знания, полученные в учебной дисциплине, используются при изучении дисциплины «Системы хранения данных» модуля «Большие данные» компонента учреждения высшего образования.

Требования к компетенциям

Освоение учебной дисциплины «Технологии и компьютерные системы обработки данных» должно обеспечить формирование следующих универсальных, углубленных профессиональных компетенций.

универсальные компетенции:

УК-5. Обладать способностью в минимальные сроки изучать и профессионально эксплуатировать программные системы, модули и библиотеки.

углубленные профессиональные компетенции:

УПК-5. Владеть перспективными технологиями программирования.

В результате освоения учебной дисциплины студент должен:

знать:

- тенденции развития технологий и компьютерных систем обработки данных;
- современные технологии распределенных вычислений, хранения и обработки больших объемов данных, их достоинства и недостатки, а также области применения;
- базовые принципы построения и особенности параллельных вычислительных систем с общей и распределенной памятью;
- основные парадигмы и технологии разработки параллельных приложений;
- принципы организации и типовые модели взаимодействия параллельных процессов в рамках разных парадигм параллельных вычислений.

уметь:

- использовать технологии и компьютерные системы распределенных вычислений при решении задач, связанных с обработкой и анализом данных;
- принимать решения о выборе технологии с точки зрения ее оптимальности для решения поставленной задачи.
- производить анализ и выбор схемы распараллеливания под конкретные задачу и вычислительные ресурсы;
- производить оценку ускорения и эффективности параллельных приложений;

владеть:

- современными компьютерными технологиями хранения и обработки больших объемов данных;
- навыками выбора и обоснования методов и инструментов решения задач, требующих параллельного и распределенного программирования;

- навыками реализации и использования параллельных и распределенных программ.

Структура учебной дисциплины

Дисциплина изучается в 1-ом семестре. Всего на изучение учебной дисциплины «Технологии и компьютерные системы обработки данных» отведено:

– для очной формы получения высшего образования – 106 часов, в том числе 50 аудиторных часов, из них: лекции – 20 часов, лабораторные занятия – 20 часов, семинарские занятия – 10 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации по учебной дисциплине – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Параллельные вычисления

Тема 1.1. Параллельные вычислительные системы. Архитектура и технологии программирования

Классификация параллельных вычислительных систем. Параллельные компьютеры с общей памятью. Системы с распределенной памятью. Архитектура кластера. Коммуникационная среда. Парадигмы параллельного программирования. Парадигма программирования с разделяемой памятью. Парадигма программирования с передачей данных. Производительность суперкомпьютеров. Оценка эффективности параллельных приложений. Законы Амдала.

Тема 1.2. Параллельное программирование на системах с общей памятью. Технология OpenMP

Разработка параллельных приложений для систем с общей памятью, стандарт OpenMP. Основные конструкции, директивы, классы переменных, переменные среды, средства синхронизации.

Тема 1.3. Параллельное программирование на системах с распределенной памятью. Технология MPI

SPMD-модель программирования. Процесс и его атрибуты. Взаимодействие и синхронизация процессов. Общие функции MPI.

Тема 1.4. Передача и прием данных между процессами

Передача и прием данных между отдельными процессами. Передача и прием сообщений с блокировкой. Атрибуты сообщения. Режимы передачи сообщений. Передача и прием сообщений без блокировки. Функции контроля неблокирующих коммуникаций. Тупиковые ситуации. Типы данных. Создание пользовательских типов данных. Конструкторы типов. Использование пользовательских типов в коммуникациях.

Тема 1.5. Группы и коммуникаторы. Коллективные операции обмена данными

Группы процессов. Создание и уничтожение групп. Операции с группами процессов. Коммуникаторы. Создание и уничтожение коммуникаторов. Операции с коммуникаторами. Виртуальные топологии. Барьерная синхронизация. Широковещательный обмен. Сбор данных. Рассылка данных. Сборка и рассылка данных по схеме “каждый с каждым”. Операции редукции.

Раздел 2. Системы хранения данных.

Модель вычислений MapReduce и платформа Apache Hadoop

Тема 2.1. Введение в системы хранения данных

Различные типы компьютерных систем обработки и хранения больших объемов информации, их преимущества, недостатки и области применения. Данные и информация, типы данных, эволюция систем хранения данных, большие данные. Среда систем хранения данных, основные элементы. Виртуализация приложений и серверов. Архитектура систем хранения данных.

Тема 2.2. Модель вычислений MapReduce. Организация распределенных вычислений

Модель вычислений MapReduce. Принципы параллельной реализации вычислений. Область применения и примеры задач. Ограничения модели MapReduce, расширения и альтернативные подходы.

Тема 2.3. Платформа Apache Hadoop

Платформа Apache Hadoop. Интерфейсы прикладного программирования.

Тема 2.4. Инструментарий для работы с Hadoop

Высокоуровневые языки и инструментарий для работы с Hadoop.

Тема 2.5. Приемы и стратегии реализации MapReduce-программ

Приемы и стратегии реализации MapReduce-программ. Практические примеры использования MapReduce.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов						Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное	Количество часов УСР	
1	2	3	4	5	6	7	8	9
1.	**Параллельные вычисления**	**10**		**6**	**10**			
1.1	Параллельные вычислительные системы. Архитектура и технологии программирования	2						Устный опрос
1.2	Параллельное программирование на системах с общей памятью. Технология OpenMP	2		2	2			Собеседование
1.3	Параллельное программирование на системах с распределенной памятью. Технология MPI	2			2			Отчет по лабораторным работам с их устной защитой
1.4	Передача и прием данных между процессами	2		2	4			Отчет по лабораторным работам с их устной защитой. Выступление с докладом на семинаре.
1.5	Группы и коммуникаторы. Коллективные операции обмена данными	2		2	2			Отчет по лабораторным работам с их устной защитой. Контрольная работа № 1
2.	**Системы хранения данных. Модель вычислений MapReduce и платформа Apache Hadoop**	**10**		**4**	**10**			
2.1	Введение в системы хранения	2						Устный опрос

	данных.							
2.2	Модель вычислений MapReduce. Организация распределенных вычислений	2		2	2			Выступление с докладом на семинаре
2.3	Платформа Apache Hadoop	2			2			Отчет по лабораторным работам с их устной защитой
2.4	Инструментарий для работы с Hadoop	2		2	2			Коллоквиум
2.5	Приемы и стратегии реализации MapReduce-программ	2			4			Отчет по лабораторным работам с их устной защитой. Контрольная работа № 2

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Таненбаум Э., Остин Т. Архитектура компьютера. СПб.: Питер, 2015. – 816 с.
2. White T. Hadoop: The Definitive Guide, O'Reilly Media, Inc., 2015. – 728 p.
3. Антонов А.С. Технологии параллельного программирования MPI и OpenMP. М.: Изд-во МГУ, 2012. – 344 с.
4. Богачев К.Ю. Основы параллельного программирования. М.: БИНОМ. Лаборатория знаний, 2010. – 342 с.
5. Гергель В.П. Высокопроизводительные вычисления для многопроцессорных многоядерных систем. М.: Изд-во МГУ, 2010. – 544 с.

Перечень дополнительной литературы

1. Таненбаум Э., Уэзеролл Д. Компьютерные сети. СПб.: Питер, 2016. – 960 с.
2. Гергель В.П. Теория и практика параллельных вычислений. Москва: Бином, 2007. – 423 с.
3. Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. СПб.: БХВ-Петербург, 2002 г. – 608 с.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: устный опрос, собеседование, коллоквиум.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: отчеты по лабораторным работам с их устной защитой, оценивание на основе проектного метода.

Формой текущей аттестации по дисциплине «Технологии и компьютерные системы обработки данных» учебным планом предусмотрен экзамен.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку:

- отчёт по лабораторным работам с их устной защитой – 50 %;

- выступление с докладом на семинарских занятиях – 20 %;
- контрольные работы – 20 %;
- коллоквиум – 10 %.

Рейтинговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и экзаменационной оценки с учетом их весовых коэффициентов Вес оценка по текущей успеваемости составляет 30 %, экзаменационная оценка – 70 %.

Примерная тематика лабораторных занятий

Занятие № 1. Реализация эффективных механизмов сборки данных в MPI-приложениях.

Занятие № 2. Реализация “master-slave” модели взаимодействия процессов в рамках технологии обмена сообщениями MPI.

Занятие № 3. Конвейерный параллелизм и его эффективная реализация в модели обмена сообщениями.

Занятие № 4. Программная реализация алгоритма Фокса умножения матриц с использованием гибридных технологий (MPI+OpenMP).

Занятие № 5. Параллельные вычисления в модели с общей памятью. Разработка параллельного приложения для решения задачи Дирихле с использованием технологии OpenMP.

Занятие № 6. Реализация алгоритма подсчета количества уникальных слов в документе с использованием MapReduce/Spark/Pig/Hive.

Занятие № **7.** Реализация алгоритма BFS с использованием MapReduce/Spark/Pig/Hive.

Занятие № 8. Реализация алгоритма построения обратного индекса с использованием MapReduce/Spark/Pig/Hive.

Занятие № 9. Работа с HDFS. Загрузка данных, использование данных в заданиях MapReduce/Spark/Pig/Hive.

Занятие № 10. Работа с HBase. Загрузка данных, использование данных в заданиях MapReduce/Spark/Pig/Hive.

Примерная тематика семинарских занятий

Семинар № 1. Организация взаимодействия между потоками в OpenMP-приложениях. Балансировка вычислительной нагрузки.

Семинар № 2. Надежность и эффективность механизмов обмена сообщениями в MPI-приложениях. Балансировка вычислительной нагрузки.

Семинар № 3. Анализ эффективности параллельных приложений.

Семинар № 4. Решение типовых задач в MapReduce.

Семинар № 5. Реализация программ для Hadoop.

Описание инновационных подходов и методов к преподаванию учебной дисциплины (эвристический, проективный, практико-ориентированный)

При организации образовательного процесса большинства практических занятий используется практико-ориентированный подход, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация методов группового обучения, проектного обучения и учебной дискуссии. Комбинация методов предполагает: ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, кроме подготовки к экзамену, подготовка к зачету

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к зачёту, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к экзамену

1. Классификация параллельных вычислительных систем. Параллельные компьютеры с общей памятью.
2. Классификация параллельных вычислительных систем. Системы с распределенной памятью.
3. Оценка параллельных приложений. Эффективность и ускорение параллельных приложений. Законы Амдала.
4. Архитектура кластера. Вычислительные узлы. Коммуникационная среда. Память.

5. Вычислительные системы с распределенной памятью. Влияние архитектуры и характеристик коммуникационной среды на производительность параллельных приложений.

6. Технология MPI. SPMD-модель программирования. Процесс и его атрибуты. Общие функции MPI. Инициализация и завершение параллельной программы. Идентификация процесса.

7. Передача и прием данных между отдельными процессами. Особенности передачи данных с блокировкой и без блокировки.

8. Передача сообщений с блокировкой. Атрибуты сообщения. Базовые типы данных. Режимы передачи сообщений.

9. Прием сообщений с блокировкой. Атрибуты сообщения. Базовые типы данных. Статус сообщений.

10. Передача и прием сообщений. Функции анализа сообщения MPI_Probe, MPI_Get_count.

11. Передача и прием сообщений без блокировки. Атрибуты сообщения. Функции контроля не блокирующих коммуникаций.

12. Передача и прием данных между отдельными процессами. Правила корректной передачи данных. Тупиковые ситуации и их разрешение. Реализация взаимообмена данными между двумя процессами.

13. Типы данных MPI. Пользовательские типы данных. Конструктор MPI_Type_contiguous.

14. Пользовательские типы данных. Конструктор MPI_Type_vector.

15. Пользовательские типы данных. Конструктор типов MPI_Type_indexed.

16. Пользовательские типы данных. Конструктор типов MPI_Type_struct.

17. Коллективные операции обмена данными. Барьерная синхронизация. Широковещательная рассылка.

18. Коллективные операции обмена данными. Сбор данных.

19. Коллективные операции обмена данными. Рассылка данных.

20. Коллективные операции обмена данными. Сборка и рассылка данных по схеме “каждый с каждым”.

21. Коллективные операции обмена данными. Операции редукции.

22. Модель программирования с общей памятью. Стандарт OpenMP. Модель параллельной программы. Основные конструкции OpenMP. Директивы, функции.

23. Технология OpenMP. Параллельные и последовательные области. Классы переменных.

24. Технология OpenMP. Распределение работы между потоками. Параллельные циклы, секции.

25. Технология OpenMP. Синхронизация потоков. Барьерная синхронизация, критические секции, замки.

26. Распределенная файловая система HDFS. Особенности записи и чтения данных с кластера.

27. Распределенная файловая система HDFS. Ограничения и уязвимости.
28. Модель вычислений MapReduce.
29. Ограничения и возможности операций Map и Reduce.
30. Особенности операции Combine в MapReduce.
31. Решение типовых задач в MapReduce. Задача подсчета слов в документах.
32. Решение типовых задач в MapReduce. Реализация алгоритма BFS.
33. Распределенная вычислительная система Hadoop MapReduce. Улучшения в MapReduce 2.0.
34. Ограничения модели MapReduce, расширения и альтернативные подходы. Система Pig.
35. Ограничения модели MapReduce, расширения и альтернативные подходы. Система Hive.
36. Платформа планирования ресурсов YARN.
37. Табличное хранение информации на кластере. HBase.
38. Система Apache ZooKeeper. Примеры решаемых с использование ZK задач.
39. Система Apache Tez.
40. Система Apache Spark. RDD: Resilient Distributed Dataset.
41. Система Apache Spark. Отличия от Hadoop MapReduce.
42. Система Apache Spark. Операции обработки данных.
43. Система Apache Drill.
44. Система машинного обучения Apache Mahout.

Рекомендуемая тематика контрольных работ

1. *Контрольная работа № 1.* Организация параллельной обработки данных с использованием технологий OpenMP и MPI.
2. *Контрольная работа № 2.* Организация распределенной обработки данных на платформе Apache Hadoop.
3. *Коллоквиум*. «Технологии и компьютерные системы обработки данных».

Текущий контроль знаний проводится в соответствии с учебно-методической картой дисциплины.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Системы хранения данных	Дискретной математики и алгоритмики	Нет	Изменений в содержании учебной программы не требуется, протокол № 15 от 18 апреля 2019 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой
_____ _____ _____

УТВЕРЖДАЮ
Декан факультета
_____ _____ _____