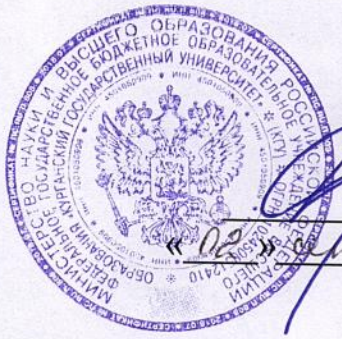
Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Курганский государственный университет»
(КГУ)

Кафедра «Программное обеспечение автоматизированных систем»



УТВЕРЖДАЮ:
Врио ректора
_____ / Н.В. Дубив/
«02» сентября 2019 г.

Рабочая программа учебной дисциплины

СОВРЕМЕННЫЕ СИСТЕМЫ И МЕТОДЫ ВЫСОКО-ПРОИЗВОДИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

образовательной программы высшего образования –
программы магистратуры

09.04.04 – Программная инженерия
направленность:
Методы и алгоритмы интеллектуальной обработки данных
в информационно-вычислительных системах
Форма обучения: заочная

Курган 2019

Рабочая программа дисциплины «Современные системы и методы высокопроизводительной обработки данных» составлена в соответствии с учебными планами по программе магистратуры «Программная инженерия» (Методы и алгоритмы интеллектуальной обработки данных в информационно-вычислительных системах), утвержденными для заочной формы обучения «29» августа 2019 года.

Рабочая программа дисциплины одобрена на заседании кафедры «Программное обеспечение автоматизированных систем» «30» августа 2019 года, протокол № 1.

Рабочую программу составили:

Доцент кафедры
«Программное обеспечение
автоматизированных систем», к.т.н. Н.В. Агапова

Согласовано:

Заведующий кафедрой
«Программное обеспечение
автоматизированных систем»
к.т.н., доцент Т.Р. Змызгова

Специалист по учебно-методической
работе Учебно-методического отдела Г.В. Казанкова

Начальник Управления образовательной
деятельности С.Н. Синицын

1. ОБЪЕМ ДИСЦИПЛИНЫ

Всего: 4 зачетных единиц трудоемкости (144 академических часа)

Вид учебной работы	На всю дисциплину	Семестр
		2
Аудиторные занятия (контактная работа с преподавателем), всего часов в том числе:	**14**	**14**
Лекции	10	10
Практические занятия	4	4
Аудиторные занятия в интерактивной форме, часов	-	-
Самостоятельная работа, всего часов в том числе:	**130**	**130**
Подготовка к зачету	18	18
Другие виды самостоятельной работы	112	112
Контрольная работа	-	-
Вид промежуточной аттестации	**зачет**	**зачет**
Общая трудоемкость дисциплины и трудоемкость по семестрам, часов	**144**	**144**

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина «Современные системы и методы высокопроизводительной обработки данных» относится к вариативной части блока 1 «Высокопроизводительные системы обработки данных».

Для успешного освоения данной дисциплины необходимо и достаточно знаний и умений, приобретенных студентами при изучении на предыдущем уровне образования таких дисциплин, как «Объектно-ориентированный анализ и программирование», «Вычислительные системы, сети и телекоммуникации», «Операционные сети и сети».

Дисциплина является базовой при проведении научно-исследовательской работы магистра, прохождении научно-исследовательской практики, подготовке магистерской диссертации.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Цель изучения дисциплины состоит в формировании знаний умений и навыков в области разработки и эксплуатации программного обеспечения современных высокопроизводительных распределенных систем. В данном курсе рассматриваются программные технологии построения масштабируемых многомашинных информационно-вычислительных систем, обеспечивающих параллельную обработку сверхбольших массивов данных. За рубежом совокупность таких технологий обозначается термином Big Data (англ. - большие данные).

Рассматриваются также типовые методы и алгоритмы параллельной обработки сверхбольших массивов данных с использование стека технологий Big Data.

Задачи изучения дисциплины:

1) ознакомление с теоретическими основами организации параллельной распределенной обработки данных на программном уровне;

2) получение опыта практической работы с современными программными инструментами для параллельной распределенной обработки данных.

Компетенции, формируемые в результате освоения дисциплины:

- Способность разрабатывать и использовать программное обеспечение для моделирования, анализа, распознавания и обработки информации, в том числе - в системах искусственного интеллекта (ПК-3);
- Способность проектировать архитектуры высокопроизводительных программных систем и проводить оценку их производительности (ПК-5).

В результате изучения дисциплины **обучающийся должен**

знать:

- теоретические основы организации распределенных вычислений (для ПК-3);

- состав и принципы построения ПО параллельных распределенных вычислений (для ПК-3);
- методы измерения производительности вычислительных систем (для ПК-5);

уметь:

- реализовывать параллельные алгоритмы обработки данных на высокоуровневых языках программирования с использованием библиотек (для ПК-3);
- устанавливать и настраивать окружение распределенных вычислений с использованием современных программных продуктов (для ПК-3);
- **владеть:**
- средствами выполнения и отладки прикладного ПО для распределенных систем (для ПК-3);
- средствами профилирования и измерения производительности при решении задач на распределенных вычислительных системах (для ПК-5).

4. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1. Учебно-тематический план

Заочная форма обучения

№	Наименование раздела	Количество часов контактной работы с преподавателем	
		Лекции	Практические занятия
1	Обзор технологий высокопроизводительных систем	2	-
2	Технологии Hadoop и Map/Reduce	4	4
3	Системы разработки, сборки и доставки кода	2	-
4	Системы поточной обработки данных	2	-
	Всего:	**10**	**4**

4.2. Содержание лекционных занятий

Наименование и содержание лекции	Часов контактной работы с преподавателем
Раздел №1. **Обзор технологий высокопроизводительных систем**	**2**
Понятие высокопроизводительных вычислений. История развития. Классические подходы: конвейерные, массово-параллельные, кластерные подходы. Новые подходы, ориентированные на данные.	2
Раздел №2. **Технологии Hadoop и Map/Reduce**	**4**
Распределенная файловая система Hadoop. Принцип доставки вычислений к данным. Метод Map/Reduce. Примеры реализации на языках программирования	4
Раздел №3. **Системы разработки, сборки и доставки кода**	**2**
Базовые компоненты классической системы очередей. Типичный жизненный цикл сообщений в системах очередей. Kafka и классические сервисы очередей. Структура данных. Consumer Groups. Consumer offsets. Apache ZooKeeper	2
Раздел №4. **Системы поточной обработки данных**	**2**
Фреймворк Apache Spark. Основные характеристики. Приложение Spark - состав. Обзор архитектуры Spark. Операции с RDD. Ленивые вычисления. Направленный ациклический граф. Циклы выполнения. Apache Flume. Apache Flume — передача данных в Hadoop. Apache Sqoop	2
Итого:	**10**

4.3 Практические занятия

Наименование и содержание практического занятия	Часов контактной работы с преподавателем
Раздел №2. **Технологии Hadoop и Map/Reduce**	
Практическое занятие № 1. **Технология Hadoop**	**2**
Практическое занятие №2. **Технология Map/Reduce**	**2**
Всего часов практических занятий	**4**

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Во время лекций по дисциплине студентам рекомендуется конспектировать теоретический материал, отмечая важные моменты, на которые заострил внимание преподаватель, участвовать в опросах и дискуссиях. Перед лекцией необходимо повторить выданный материал, зафиксировать непонятные места, чтобы обсудить их на занятии. Конспект лекций представлен в виде мультимедийных презентаций и включен в состав методического комплекса дисциплины.

Для текущего контроля успеваемости преподавателем используется система контроля и оценки академической активности. Поэтому настоятельно рекомендуется тщательно прорабатывать материал дисциплины при самостоятельной работе, участвовать во всех формах обсуждения и взаимодействия, как на лекциях, так и на практических занятиях в целях лучшего освоения материала и получения высокой оценки по результатам освоения дисциплины.

Выполнение самостоятельной работы подразумевает самостоятельное изучение разделов дисциплины, подготовку к практическим занятиям, подготовку к зачету.

Рекомендуемая трудоемкость самостоятельной работы представлена в таблице:

Таблица 5.1 – Рекомендуемая трудоемкость самостоятельной работы

Виды самостоятельной работы	Рекомендуемая трудоемкость, акад. часов
Самостоятельное изучение тем дисциплины:	**94**
Поисковые задачи на Map/Reduce	6
Современные BigData-решения и архитектуры	7
Система Hive	9
Система Shark	9
Система Impala	9
Система Phoenix	9
Система Mahout	9
Система YARN	9
Система MLBase	9
Система Mesos	69
Способы повышения производительности BigData-систем	9
Написание реферата (по одной из выбранных тем)	**10**
Подготовка и выполнение практических заданий (по 4 часа на каждое занятие)	**8**

Подготовка к зачету	**18**
Всего:	**130**

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

6.1. Перечень оценочных средств

1. Отчеты обучающихся по практическим работам
2. Тестовые задания
3. Реферат
4. Вопросы к зачету

6.2. Процедура оценивания результатов освоения дисциплины

Защита практических занятий, реферата проводятся в форме устного опроса.

Преподаватель прорабатывает со студентами основной материал соответствующих разделов дисциплины в форме краткой лекции-дискуссии.

Зачет проводится в форме беседы по билетам, которые состоят из 2 вопросов. Время, отводимое студенту на подготовку к ответу, составляет 1 астрономический час.

Результаты зачета заносятся преподавателем в зачетную ведомость, которая сдается в организационный отдел института в день зачета, а также выставляются в зачетную книжку студента.

6.4. Примеры оценочных средств

Ниже приведены примерные темы опросов на занятиях, дающие представление об их направленности и уровне сложности:

– Понятие высокопроизводительных вычислений. История развития. Классические подходы: конвейерные, массово-параллельные, кластерные подходы. Новые подходы, ориентированные на данные.

– Распределенная файловая система Hadoop. Принцип доставки вычислений к данным.

– Метод Map/Reduce. Примеры реализации на языках программирования

– Недостатки Map/Reduce. Надстройки над Hadoop.

– Обзор технологий стека Apache Big Data.

– Рассмотрение элементов стека Apache с примерами на языках высокого уровня

– Проблемы применения и узкие места Big Data. Точки роста технологий Big Data. Взаимосвязь Big Data и классических технологий высокопроизводительных вычислений.

Примеры оценочных средств для зачета

1. Конфигурирование и системные утилиты Hadoop, взаимодействие с файловой системой HDFS
2. Конфигурирование и системные утилиты Map/Reduce, запуск примеров программ обработки данных
3. Разработка собственной программы на языке Java для Map/Reduce Hadoop
4. Конфигурирование и системные утилиты Apache Spark, взаимодействие с классическими и распределенными файловыми системами
5. Запуск примеров программ в системе Spark на языке Java. Измерение производительности
6. Разработка собственной программы на языке Java для Apache Spark
7. Интерактивная среда на языке R в системе Apache Spark
8. Запуск примеров программ в системе Spark на языке R. Измерение производительности
9. Разработка собственной программы на языке R для Apache Spark
10. Запуск примеров программ для Apache HBase. Измерение производительности
11. Запуск примеров программ для Apache Hive. Измерение производительности
12. Разработка собственных программ на языке Java для Apache Hive и HBase
13. Система Hive
14. Система Impala
15. Система Shark
16. Система Phoenix
17. Сравнение технологий MapReduce: Hadoop и Twister
18. Среда R
19. Система Mahout
20. Система MLBase
21. Файловая система MapR
22. Системы планирования задач
23. Система YARN
24. Система Mesos

6.5. Фонд оценочных средств

Полный банк заданий для текущего контроля и промежуточной аттестации по дисциплине, показатели, критерии, шкалы оценивания компетенций, методические материалы, определяющие процедуры оценивания образовательных результатов, приведены в учебно-методическом комплексе дисциплины.

7. ОСНОВНАЯ И ДОПОЛНИТЕЛЬНАЯ УЧЕБНАЯ ЛИТЕРАТУРА

7.1 Основная литература

1. Лесковец Ю., Раджараман А., Ульман Д. - Анализ больших наборов данных - 2016

2. Ын А., Су К. - Теоретический минимум по Big Data (Библиотека программиста) – 2019

3. Гергель В.П. Современные языки и технологии параллельного программирования: учебник для вузов. – М.: Изд-во Московского ун-та, 2012. – 408 с.

2. Макшанов А.В. Технологии интеллектуального анализа данных [электронный ресурс]: учеб. Пособие/А.В. Макшанов, А.Е.Журавлев.-Санкт-Петербург: Лань, 2019.-212 с. – Доступ из ЭБС «Лань»

7.2 Дополнительная литература

1. Hadoop and Big Data [Электронный ресурс] – Режим доступа : http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html.

2. Бойченко И.В. Высокопроизводительные распределенные системы: метод. указания к практическим занятиям и по организации самостоятельной работы магистрантов, обучающихся по направлению «Программная инженерия». – 2015. – 7 с. [Электронный ресурс]: сайт каф. АОИ ТУСУРа. – URL:

http://aoi.tusur.ru/upload/methodical_materials/High_performance_distributed_systems_file__646_2666.pdf

8 МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

8.1 Техническое обеспечение

№	Наименование	Использование
1	Комплект: ноутбук, медиа-проектор, экран	Для демонстрации иллюстративного материала при чтении лекций
2	Персональный компьютер стандартной комплектации	Используется в качестве инструмента и объекта исследования при выполнении лабораторных и контрольных работ

8.2 Программное обеспечение

№	Наименование	Использование
1	Oracle Java, Python, Apache Hadoop, Apache Spark	Выполнение практических заданий

Аннотация

рабочей программы учебной дисциплины

СОВРЕМЕННЫЕ СИСТЕМЫ И МЕТОДЫ ВЫСОКОПРОИЗВОДИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

образовательной программы высшего образования –

программы магистратуры

09.04.04 – Программная инженерия

направленность:

Методы и алгоритмы интеллектуальной обработки данных в информационно-вычислительных системах

Форма обучения - заочная

Трудоемкость освоения дисциплины – 4 зач. ед. (144 акад. часов)

Семестры: 2-й

Промежуточная аттестация: зачет (2-й семестр)

Содержание дисциплины

Цель изучения дисциплины состоит в формировании знаний умений и навыков в области разработки и эксплуатации программного обеспечения современных высокопроизводительных распределенных систем. В данном курсе рассматриваются программные технологии построения масштабируемых многомашинных информационно-вычислительных систем, обеспечивающих параллельную обработку сверхбольших массивов данных. За рубежом совокупность таких технологий обозначается термином Big Data (англ. - большие данные). Рассматриваются также типовые методы и алгоритмы параллельной обработки сверхбольших массивов данных с использование стека технологий Big Data.

Задачи изучения дисциплины:

1) ознакомление с теоретическими основами организации параллельной распределенной обработки данных на программном уровне;

2) получение опыта практической работы с современными программными инструментами для параллельной распределенной обработки данных.