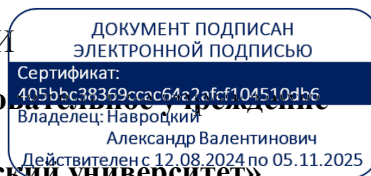




МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образование
высшего образования
«Волгоградский государственный технический университет»



Факультет электроники и вычислительной техники

УТВЕРЖДЕНО

Факультет электроники и вычислительной
техники

Декан Авдеюк О.А.
Г.

Системы обработки больших данных

рабочая программа дисциплины (модуля, практики)

Закреплена за кафедрой **Электронно-вычислительные машины и системы**

Учебный план Направление 09.04.01 Информатика и вычислительная техника
Программа "Анализ данных и интеллектуальные технологии"

Профиль

Квалификация **Магистр**

Срок обучения **2 года**

Форма обучения **очная** Общая трудоемкость **4 ЗЕТ**

Виды контроля в экзамены 1
семестрах: курсовые работы 1

Семестр(Курс.Номер семестра на курсе)	1(1.1)		Итого	
	УП	ПП	УП	ПП
Лекции	16	16	16	16
Лабораторные	16	16	16	16
Итого ауд.	32	32	32	32
Контактная работа	32.35	32.35	32.35	32.35
Сам. работа	76	76	76	76
Часы на контроль	35.65	35.65	35.65	35.65
Практическая подготовка	0	0	0	0
Итого трудоемкость в часах	144	144	0	0

ЛИСТ ОДОБРЕНИЯ, СОГЛАСОВАНИЯ И АКТУАЛИЗАЦИИ РАБОЧЕЙ ПРОГРАММЫ

Разработчик(и) программы:

ст. преподаватель Кравченя Павел Дмитриевич кфмн

Рецензент(ы):

(при наличии)

Рабочая программа дисциплины (модуля, практики)

Системы обработки больших данных

разработана в соответствии с ФГОС ВО:

Федеральный государственный образовательный стандарт высшего образования - магистратура по направлению подготовки 09.04.01 Информатика и вычислительная техника (приказ Минобрнауки России от 19.09.2017 г. № 918)

составлена на основании учебного плана:

Направление 09.04.01 Информатика и вычислительная техника

Программа "Анализ данных и интеллектуальные технологии"

Профиль:

утвержденного учёным советом вуза от 05.06.2019 протокол № 12.

Рабочая программа одобрена на заседании кафедры

Электронно-вычислительные машины и системы

номер протокола 2019 г.

Зав. кафедрой Андреев Андрей Евгеньевич

Рабочая программа дисциплины (модуля, практики) актуализирована 30.08.2024

СОГЛАСОВАНО:

Факультет электроники и вычислительной техники

Председатель НМС факультета: Авдеюк О.А.

Протокол заседания НМС от

г. №

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ). ВИД, ТИП ПРАКТИКИ, СПОСОБ И ФОРМА (ФОРМЫ) ЕЕ ПРОВЕДЕНИЯ.
Цель изучения дисциплины:
изучение свойств, особенностей больших данных и современных инструментов работы с ними, знакомство с парадигмой MapReduce и ее применением для обработки больших данных в рамках систем Apache Hadoop и Apache Spark, изучение технологии NoSQL и ее применение к хранению и обработке больших данных.
Основными задачами изучения дисциплины являются:
ознакомление с понятием больших данных и их свойств;
получение базовых знаний по принципам хранения и обработки больших данных;
выработка базовых теоретических и практических навыков использования инструментов экосистемы Apache Hadoop для работы с большими данными;
овладение навыками проектирования и разработки программного обеспечения на основе технологии MapReduce на базе фреймворков Apache Hadoop и Apache Spark для работы в распределенных вычислительных системах;
выработка навыков использования СУБД NoSQL Apache HBase для хранения и обработки больших массивов данных.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ) В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ	
Цикл (раздел) ОП:	Б1.О
2.1	Требования к предварительной подготовке обучающегося:
2.1.1	Технологии программирования
2.1.2	Технологии анализа данных
2.1.3	Параллельные и распределенные вычисления
2.2	Дисциплины (модули) и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:
2.2.1	Машинное обучение
2.2.2	Учебная практика: Технологическая (проектно-технологическая) практика
2.2.3	Системы искусственного интеллекта
2.2.4	Выполнение и защита выпускной квалификационной работы
3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)	
ОПК-1: Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;	
<i>ОПК-1.1: Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности.</i>	
Результаты обучения: Знает математические, естественнонаучные и социально-экономические методы, используемые при проектировании и использовании систем хранения и обработки больших данных.	
<i>ОПК-1.2: Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний</i>	
Результаты обучения: Умеет решать нестандартные профессиональные задачи, возникающие при проектировании и использовании систем хранения и обработки больших данных, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний.	
<i>ОПК-1.3: Иметь навыки: теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте.</i>	
Результаты обучения: Имеет навыки теоретического и экспериментального исследования систем хранения, обработки и визуализации больших данных, в том числе, в новой или незнакомой среде и в междисциплинарном контексте.	
ОПК-6: Способен разрабатывать компоненты программно-аппаратных комплексов обработки информации и автоматизированного проектирования;	
<i>ОПК-6.1: Знать: аппаратные средства и платформы инфраструктуры информационных технологий, виды, назначение, архитектуру, методы разработки и администрирования программно-аппаратных комплексов объекта профессиональной деятельности.</i>	
Результаты обучения: Знает аппаратные средства и платформы систем хранения, обработки и визуализации больших данных, их назначение, архитектуру и методы администрирования.	
<i>ОПК-6.2: Уметь: анализировать техническое задание, разрабатывать и оптимизировать программный код для решения задач обработки информации и автоматизированного проектирования.</i>	
Результаты обучения: Умеет анализировать техническое задание, разрабатывать и оптимизировать программный код для решения задач машинного обучения с применением технологий больших данных.	

ОПК-6.3: Владеть: навыками составления технической документации по использованию и настройке компонентов программно-аппаратного комплекса.

Результаты обучения: Владеет навыками составления технической документации по использованию и настройке компонентов систем хранения, обработки и визуализации больших данных.

ПК-1: Управление развитием БД

ПК-1.1: Знает: основные направления развития способов сбора и хранения данных.

Результаты обучения: Знает основные современные направления развития способов сбора и хранения больших данных.

ПК-1.2: Умеет: управлять изменениями при организации баз данных.

Результаты обучения: Умеет управлять изменениями при организации баз и хранилищ, предназначенных для хранения больших данных.

ПК-1.3: Владеет навыками: применения современных инструментов управления базами данных и механизмами изменений.

Результаты обучения: Владеет навыками применения современных программных инструментов для управления хранилищами и базами больших данных, а также механизмами изменений в них.

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Форма контроля
1	Раздел 1. Обучение			
1.1	Понятие и свойства больших данных. Подходы к анализу больших данных /Тема/	1	0	
1.1.1	Большие данные: понятие и свойства. Модель MapReduce /Лек/	1	2	Эк, Ко
1.1.2	Подготовка к текущему контролю успеваемости /Ср/	1	4	Эк, Ко
1.2	Фреймворк Apache Hadoop /Тема/	1	0	
1.2.1	Фреймворк обработки больших данных Hadoop /Лек/	1	2	Эк, Ко, К
1.2.2	Поток данных в Hadoop MapReduce /Лек/	1	2	Эк, Ко, К
1.2.3	Подготовка к лабораторной работе /Ср/	1	4	Ко
1.2.4	Изучение технологий Hadoop и MapReduce /Лаб/	1	5	Эк, Ко
1.2.5	Выполнение курсовой работы /Ср/	1	16	К
1.2.6	Подготовка к текущему контролю успеваемости /Ср/	1	8	Эк, Ко, К
1.3	Фреймворк Apache Spark /Тема/	1	0	
1.3.1	Фреймворк обработки больших данных Apache Spark (часть 1) /Лек/	1	2	Эк, Ко, К
1.3.2	Фреймворк обработки больших данных Apache Spark (часть 2) /Лек/	1	2	Эк, Ко, К
1.3.3	Подготовка к лабораторной работе /Ср/	1	4	Ко
1.3.4	Изучение технологии Apache Spark /Лаб/	1	5	Эк, Ко
1.3.5	Выполнение курсовой работы /Ср/	1	16	К
1.3.6	Подготовка к текущему контролю успеваемости /Ср/	1	8	Эк, Ко, К
1.4	Распределенная NoSQL-база данных Apache HBase /Тема/	1	0	
1.4.1	Организация хранения и обработки больших данных. NoSQL. HBase /Лек/	1	2	Эк, Ко
1.4.2	Архитектура и принципы работы СУБД Apache HBase /Лек/	1	2	Эк, Ко
1.4.3	Подготовка к лабораторной работе /Ср/	1	4	Ко
1.4.4	Изучение технологии NoSQL на основе нереляционной базы данных Apache HBase /Лаб/	1	6	Эк, Ко
1.4.5	Подготовка к текущему контролю успеваемости /Ср/	1	8	Эк, Ко
1.5	Обработка потоковых данных /Тема/	1	0	
1.5.1	Обработка потоковых больших данных. Apache Kafka /Лек/	1	2	Эк, Ко
1.5.2	Подготовка к текущему контролю успеваемости /Ср/	1	4	Эк, Ко
2	Раздел 2. Промежуточная аттестация			
2.1	Экзамен /Тема/	1	0	
2.1.1	Подготовка к экзамену /Экзамен/	1	35.65	Эк
2.1.2	Контактная работа /КоПа/	1	0.35	

Примечание. Формы контроля: Эк – экзамен, К- контрольная работа, Ко- контрольный опрос, Сз- семестровое задание, З-зачет, ОП- отчет по практике, Зд-задание, Р-реферат.

5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства планируемых результатов обучения представлены в виде фондов оценочных средств (ФОС), разработанных в соответствии с локальным нормативным актом университета. ФОС может быть представлен в Приложении к рабочей программе.

5.1 Контрольные вопросы и задания

Оценочные средства планируемых результатов обучения представлены в виде фондов оценочных средств (ФОС), разработанных в соответствии с локальным нормативным актом университета. В целях освоения компетенций, указанных в рабочей программе дисциплины, предусмотрены следующие вопросы, задания текущего контроля:

ПК-1: Управление развитием БД;

ПК-1.1: Знает: основные направления развития способов сбора и хранения данных.

Результаты обучения: Знает основные современные направления развития способов сбора и хранения больших данных.

Вопросы, задания:

1. Какие направления развития способов сбора больших данных Вы знаете?
2. Какие направления развития способов хранения больших данных Вы знаете?
3. Какие технологии хранения больших данных получили развитие в последнее время?

ПК-1.2: Умеет: управлять изменениями при организации баз данных.

Результаты обучения: Умеет управлять изменениями при организации баз и хранилищ, предназначенных для хранения больших данных.

1. Приведите последовательность действий, которые нужно совершить пользователю при добавлении информации в HDFS.
2. Приведите последовательность действий, которые нужно совершить пользователю при добавлении информации в HBase.
3. Приведите последовательность действий, которые нужно совершить пользователю при извлечении из HBase предыдущей версии данных.

ПК-1.3: Владеет навыками: применения современных инструментов управления базами данных и механизмами изменений.

Результаты обучения: Владеет навыками применения современных программных инструментов для управления хранилищами и базами больших данных, а также механизмами изменений в них.

1. Напишите команду CLI, позволяющую увеличить фактор репликации данных в HDFS до четырех.
2. Напишите команду CLI, позволяющую изменить количество версий ячейки в HBase до десяти.
3. Напишите команду CLI, позволяющую извлечь из HBase все данные требуемой версии.

ОПК-1: Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;

ОПК-1.1: Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности.

Результаты обучения: Знает математические, естественнонаучные и социально-экономические методы, используемые при проектировании и использовании систем хранения и обработки больших данных.

1. Назовите математические методы, используемые при проектировании систем хранения больших данных.
2. Назовите математические методы, положенные в основу алгоритма MapReduce.
3. Назовите математические методы, используемые при решении задач классификации на больших данных.

ОПК-1.2: Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний

Результаты обучения: Умеет решать нестандартные профессиональные задачи, возникающие при проектировании и использовании систем хранения и обработки больших данных, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний.

1. Примените линейные модели машинного обучения для решения задачи банковского скоринга на больших данных.
2. Примените ансамблевые методы машинного обучения для решения задачи банковского скоринга на больших данных.
3. Примените метод TF-IDF для решения задачи сентимент-анализа на больших данных.

ОПК-1.3: Иметь навыки: теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте.

Результаты обучения: Имеет навыки теоретического и экспериментального исследования систем хранения, обработки и визуализации больших данных, в том числе, в новой или незнакомой среде и в междисциплинарном контексте.

1. Оцените масштабируемость конкретной программы для обработки больших данных с использованием Apache Spark.
2. Оцените, во сколько раз уменьшится объем данных, передаваемый по сети, при выполнении условия локальности данных и вычислений.
3. Оцените минимальное количество экземпляров службы DataNode в HDFS, достаточной для хранения 150 ТБ данных.

ОПК-6: Способен разрабатывать компоненты программно-аппаратных комплексов обработки информации и

автоматизированного проектирования;

ОПК-6.1: Знать: аппаратные средства и платформы инфраструктуры информационных технологий, виды, назначение, архитектуру, методы разработки и администрирования программно-аппаратных комплексов объекта профессиональной деятельности.

Результаты обучения: Знает аппаратные средства и платформы систем хранения, обработки и визуализации больших данных, их назначение, архитектуру и методы администрирования.

1. Какую архитектуру имеет система HDFS?
2. Какие методы администрирования кластера Hadoop Вы знаете?
3. Поясните назначение системы YARN в Hadoop.

Результаты обучения: ОПК-6.2: Уметь: анализировать техническое задание, разрабатывать и оптимизировать программный код для решения задач обработки информации и автоматизированного проектирования.

Умеет анализировать техническое задание, разрабатывать и оптимизировать программный код для решения задач машинного обучения с применением технологий больших данных.

1. Разработайте программный код, выполняющий анализ заданного датасета большого объема методом линейной регрессии.
2. Разработайте программный код, выполняющий анализ заданного датасета большого объема методом градиентного бустинга на деревьях решений.
3. Предложите способы оптимизации представленного программного кода, реализующего алгоритм MapReduce.

Результаты обучения: ОПК-6.3: Владеть: навыками составления технической документации по использованию и настройке компонентов программно-аппаратного комплекса.

Владеет навыками составления технической документации по использованию и настройке компонентов систем хранения, обработки и визуализации больших данных.

1. Составьте документацию по настройке системы Apache Spark для работы с системой хранения данных HDFS.
2. Составьте документацию по использованию системы Apache Spark с языком программирования Python.
3. Составьте документацию по использованию системы HDFS с внешними клиентами.

5.2 Темы письменных работ (курсовые работы)

1. Реализация программы по выявлению негативных оценок среди трендовых видео платформы YouTube.
2. Разработка программы для анализа аннотации статей на ArXiv с помощью MapReduce.
3. Определение наиболее популярного кандидата в президенты в США в 2020 году с помощью модели распределенных вычислений MapReduce.
4. Определение количества твитов с отрицательной эмоциональной окраской по нишевым тематикам с помощью MapReduce.
5. Анализ влияния принадлежности к определённой категории популярных видео с YouTube на количество просмотров, комментариев, лайков и дизлайков пользователей.
6. Реализация программы для анализа метеорологических данных с помощью модели распределённых вычислений MapReduce.
7. Создание алгоритма для нахождения авторов с наибольшим количеством опубликованных статей в определённой категории.
8. Разработка программы для анализа аннотации статей на ArXiv с помощью MapReduce.
9. Исследование направления движения финансового рынка с помощью алгоритма распределенных вычислений MapReduce.
10. Анализ зависимости зарплаты баскетболистов от гражданства и возрастной группы.

5.3 Показатели и критерии оценивания компетенций, шкалы оценивания

В рамках изучаемой дисциплины студент может демонстрировать следующие уровни овладения компетенциями.

Повышенный уровень: обучающийся демонстрирует глубокое знание учебного материала; способен использовать сведения из различных источников для успешного исследования и поиска решения в нестандартных ситуациях; способен анализировать, проводить сравнение и обоснование выбора методов решения практико-ориентированных заданий. Оценка промежуточной аттестации (экзамен): 5 (отлично) – 90 баллов и более.

Базовый уровень: обучающийся способен понимать и интерпретировать освоенную информацию; демонстрирует осознанное владение учебным материалом и учебными умениями, навыками и способами деятельности, необходимыми для решения практико-ориентированных заданий. Оценка промежуточной аттестации (экзамен): 4 (хорошо) – 76-89 баллов.

Пороговый уровень: обучающийся обладает необходимой системой знаний и владеет некоторыми умениями; демонстрирует самостоятельность в применении знаний, умений и навыков к решению учебных заданий на

репродуктивном уровне. Оценка промежуточной аттестации (экзамен): 3 (удовлетворительно) – 61-75 баллов.

Уровень ниже порогового: система знаний, необходимая для решения учебных и практико-ориентированных заданий, не сформирована; обучающийся не владеет основными умениями, навыками и способами деятельности. Оценка промежуточной аттестации (экзамен): 2 (неудовлетворительно) – ниже 61 балла.

В рамках данной дисциплины используются следующие критерии оценки знаний студентов.

Отлично

Обучающийся демонстрирует:

- систематизированные, глубокие и полные знания по всем разделам учебной дисциплины, а также по основным вопросам, выходящим за ее пределы;
- точное использование научной терминологии, грамотное, логически правильное изложение ответа на вопросы;
- безупречное владение инструментарием учебной дисциплины, умение его эффективно использовать в постановке и решении научных и профессиональных задач;
- выраженную способность самостоятельно и творчески решать сложные проблемы в нестандартной ситуации;
- полное и глубокое усвоение основной, и дополнительной литературы, по изучаемой учебной дисциплине;
- умение свободно ориентироваться в теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку, использовать научные достижения других дисциплин;
- творческую самостоятельную работу на учебных занятиях, активное творческое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

Хорошо

Обучающийся демонстрирует:

- систематизированные, глубокие и полные знания по всем разделам учебной дисциплины;
- использование научной терминологии, грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы и обобщения;
- владение инструментарием учебной дисциплины (методами комплексного анализа, техникой информационных технологий), умение его использовать в постановке и решении научных и профессиональных задач;
- способность решать сложные проблемы в рамках учебной дисциплины;
- свободное владение типовыми решениями;
- усвоение основной и дополнительной литературы, рекомендованной рабочей программой по учебной дисциплине;
- умение ориентироваться в теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку;
- активную самостоятельную работу на учебных занятиях, систематическое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

Удовлетворительно

Обучающийся демонстрирует:

- достаточные знания в объеме рабочей программы по учебной дисциплине;
- использование научной терминологии, грамотное, логически правильное изложение ответа на вопросы, умение делать выводы без существенных ошибок;
- владение инструментарием учебной дисциплины, умение его использовать в решении учебных и профессиональных задач;
- способность самостоятельно применять типовые решения в рамках изучаемой дисциплины;
- усвоение основной литературы, рекомендованной рабочей программой по дисциплине;
- умение ориентироваться в базовых теориях, концепциях и направлениях по дисциплине;
- работу на учебных занятиях под руководством преподавателя, фрагментарное участие в групповых обсуждениях, достаточный уровень культуры исполнения заданий.

Неудовлетворительно

Обучающийся демонстрирует:

- фрагментарные знания в рамках изучаемой дисциплины; знания отдельных литературных источников, рекомендованных рабочей программой по учебной дисциплине;
- неумение использовать научную терминологию учебной дисциплины, наличие в ответе грубых, логических ошибок;
- пассивность на занятиях или отказ от ответа, низкий уровень культуры исполнения заданий.

5.4. Вопросы промежуточной аттестации

1. Четвертая промышленная революция. Концепция «Индустрии 4.0». Значение и роль данных в современном мире.
2. Понятие больших данных (Big Data). Модель mV для больших данных. Источники больших данных. Применение больших данных.
3. Методы анализа больших данных. Подходы к анализу больших данных. Проблемы, связанные с обработкой больших данных традиционными инструментами. Требования к системам BigData.
4. Понятие технологии MapReduce, основные принципы, лежащие в ее основе. Характеристики этапов map и reduce. Примеры задач MapReduce.
5. Фреймворк Hadoop как распределенная система обработки больших данных. Экосистема Hadoop. Масштабируемость и производительность Hadoop. Принцип определения доступности узлов. Понятие Heartbeat.

6. Понятие HDFS. Организация HDFS. Архитектура HDFS. NameNode, SecondaryNameNode и DataNode: понятие и функции. Обеспечение надежности хранения данных. Понятие репликации. Характеристики и ключевые особенности HDFS как файловой системы.
7. Процесс чтения и записи данных в HDFS. Консольный клиент.
8. Понятие и назначение YARN. Архитектура YARN. ResourceManager и NodeManager: понятие и функции. Понятие контейнеров (container) YARN. Отличия MapReduce v1 от MapReduce v2.
9. Процесс запуска и исполнения приложения под управлением YARN. Понятие ApplicationMaster. Взаимодействие компонентов YARN. Понятие локальности данных. Организация отказоустойчивости YARN.
10. Основные планировщики YARN: FIFO, Capacity, Fair. Их назначение и ключевые особенности.
11. Процесс выполнения приложения в Hadoop MapReduce. Понятие задания (job), задачи (task), контекста (context). Конфигурация параметров Hadoop, примеры. Поток данных в Hadoop.
12. Подготовка данных для этапа Map в Hadoop. Класс InputFormat, его наследники. Split и RecordReader, их взаимосвязь. Поток данных до этапа Map.
13. Этап Map в Hadoop. Класс Mapper. Методы map(), setup(), cleanup(). Пример реализации Map. Класс Partitioner.
14. Обработка промежуточных записей. Кольцевой буфер и операция spill. Этапы sort и combine. Перемешивание данных (shuffle) и его влияние на производительность.
15. Операция Reduce в Hadoop. Класс Reducer. Методы reduce(), setup(), cleanup(). Пример реализации Reduce. Сохранение полученных результатов. Класс OutputFormat. Класс RecordWriter.
16. Обмен данными между основными этапами в MapReduce Hadoop. Понятие сериализации. Интерфейсы Writable и WritableComparable. Основные Writable-классы в Hadoop.
17. Фреймворк обработки больших данных Apache Spark, его назначение, функции и отличия от Hadoop MapReduce. Экосистема Spark. API фреймворка.
18. Архитектура приложения Apache Spark. Модели запуска Spark-приложений (YARN, Mesos, Standalone, Kubernetes). Понятие драйвера (driver) и исполнителей (executors). Процесс запуска Spark-приложения под управлением YARN.
19. Понятие устойчивого распределенного набора данных (RDD). Понятие раздела RDD (partition). Способы создания RDD. Трансформации (transformations) и действия (actions). Примеры. Внутренние свойства RDD.
20. RDD и PairRDD: понятие, назначение. Основные трансформации (transformations) и действия (actions) над ними. Понятие кортежа (tuple2), связь между JavaPairRDD и JavaRDD посредством кортежа.
21. Реализация концепции MapReduce в фреймворке Spark. Методы map, flatMap, mapValues, flatMapValues, mapPartitions, mapToPair, filter.
22. Реализация концепции MapReduce в Spark. Особенности реализации Reduce, требования к функции редукции. Методы reduce и reduceByKey.
23. Клиентский (client) и кластерный (cluster) режимы работы Apache Spark.
24. Модель ленивых вычислений (lazy) и ее применение в Spark. Понятие ориентированного ациклического графа (Directed Acyclic Graph, DAG) и плана исполнения, их построение. Понятие задания (job), этапа (stage) и задачи (task). Кэширование (cache) данных в Spark.
25. Понятие узких и широких трансформаций в Apache Spark. Перемешивание данных (shuffle): причины возникновения и влияние на производительность. Алгоритмы перемешивания.
26. Управление распределением данных. Класс Partitioner и его наследники. Способы распределения по разделам. Сохранение раздела при трансформациях. Влияние распределения по разделам на производительность. Примеры.
27. Работа с переменными в Spark, понятие замыкания (closure). Примеры замыканий. Аккумуляторы (accumulators) и широкоовещательные (broadcast) переменные, их назначение и использование.
28. Понятие базы данных и хранилища данных (warehouse). ELT и ETL. Архитектура хранилищ данных. Примеры.
29. Понятие озера данных. Основные элементы озера данных. Архитектура озера данных. Примеры технологий для построения озера. Отличия озера данных от хранилища данных. Перспективы развития озер данных.
30. Понятие распределенной системы, проблемы построения распределенных систем. Частичные сбои. CAP-теорема и BASE-принцип. Строгая (strict) согласованность и согласованность в конечном счете (eventual).
31. Понятие NoSQL баз данных. Особенности NoSQL БД и отличия их от реляционных БД. Типы NoSQL баз данных. Их особенности, преимущества и недостатки. Примеры.
32. База данных Apache HBase, ее особенности. Сравнение HDFS и HBase как систем хранения данных. Паттерны использования HBase.
33. Модель данных HBase. Понятия строк, колонок, семейства колонок. Свойства семейств колонок. Временные метки.
34. Архитектура HBase. MasterServer и RegionServer, их функции. Понятие BlockCache, MemStore и Write Ahead Log, их применение. Понятие региона, правила их образования. Принципы образования новых регионов при добавлении данных в БД. Сервис ZooKeeper. Шардинг, виды шардинга.
35. Процесс записи данных в HBase и чтения из нее.
36. Хранение данных в HBase. Файлы HFile, их организация. Процедура удаления записей из БД. Tombstones. Minor и major compactions.
37. Способы доступа к HBase. Консольный интерфейс HBase (HBase Shell) и Java API, их основные функции. Основные операции над данными в HBase. Понятие фильтров в HBase, способы их использования.
38. Организация обмена данными между Spark и HBase. Метод saveAsNewAPIHadoopDataset. Понятие Bulk load и механизм его работы.
39. Понятие потоковых данных. Архитектуры обработки потоковых данных (λ и μ), их особенности. Основные технологии, применяемые для обработки потоковых данных.
40. Потоковая передача событий. Apache Kafka: понятие и предназначение. Архитектура Kafka. Понятия события (event), темы (topic), поставщиков (producers) и потребителей (consumers), разделов (partitions). API Apache Kafka.

41. Apache Kafka Streams. Понятия темы (topic), потока (stream) и таблицы (table) в Kafka Streams, их взаимосвязь. Логика обработки данных в Kafka Streams. Модель параллелизма Kafka Streams.
42. Системы обработки данных в инженерии знаний. Применимость СОБД при построении систем, основанных на знаниях.

5.5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Промежуточная аттестация обучающихся ведется непрерывно и включает в себя текущую аттестацию (контроль текущей работы в семестре, включая оценивание промежуточных результатов обучения по дисциплине, – как правило, по трем модулям) и семестровую аттестацию (экзамен) – оценивание окончательных результатов обучения по дисциплине.

По данной дисциплине, завершающейся экзаменом, по обязательным формам текущего контроля студенту предоставляется возможность набрать в сумме не менее 60 баллов. Оценивание окончательных результатов обучения по дисциплине ведется по 100-балльной шкале, оценка формируется автоматически как сумма количества баллов, набранных обучающимися за выполнение заданий обязательных форм текущего контроля и количества баллов, набранных на семестровой аттестации (экзамене).

Система оценивания

Текущий контроль представляет собой проверку усвоения учебного материала теоретического и практического характера, регулярно осуществляемую на протяжении семестра. К основным формам текущего контроля можно отнести устный опрос, письменные задания, лабораторные работы, курсовую работу.

Курсовая работа

Курсовая работа по настоящей дисциплине представляет собой законченную работу, включающую в себя разработку программы, реализующей алгоритм MapReduce для решения поставленной прикладной задачи (в соответствии с заданием), описания процессов компиляции программы, запуска ее на вычислительном кластере, получения и анализ результатов.

Данная работа позволяет оценить умения учащихся решать практические задачи анализа больших данных, оценить приобретенные навыки реализации приложений для обработки данных. Полностью выполненная курсовая работа оценивается в 29 баллов.

Лабораторная работа.

Лабораторная работа является формой контроля и средством применения и реализации полученных обучающимися знаний, умений и навыков в ходе выполнения учебно-практической задачи, связанной с получением значимого результата с помощью реальных средств деятельности. За первое полностью выполненное лабораторное задание начисляется 11 баллов, за остальные – 10 баллов. В рамках данной дисциплины планируется 3 лабораторные работы. Темы лабораторных работ указаны в разделе “4. Структура и содержание дисциплины (модуля, практики)”.

Устный опрос, собеседование.

Устный опрос, собеседование являются формой оценки знаний и предполагают специальную беседу преподавателя с обучающимися на темы, связанные с изучаемой дисциплиной. Процедуры направлены на выяснение объема знаний обучающегося по определенному разделу, теме, проблеме и т.п. Устный ответ или собеседование может практиковаться преподавателем для уточнения знаний на практических и лабораторных занятиях.

Устный опрос включает 1 вопрос из группы вопросов “5.1 Контрольные вопросы и задания”, собеседование может включать более 1-го вопроса того же списка. Ответ оценивается от 0 до 3 баллов следующим образом:

3 балла - полный, логически безупречный ответ;

2 балла - ответ в целом полный, но могут иметь место несущественные пробелы в знаниях; логика ответа правильная, но некоторые моменты в своих рассуждениях студент обосновать затрудняется;

1 балл - ответ частичный, содержит значительные изъяны; нарушений логики ответа нет, но имеется ряд логических переходов в рассуждениях, которые студент обосновать затрудняется.

Промежуточная аттестация. Экзамен.

Промежуточная аттестация осуществляется в конце семестра и завершает изучение дисциплины. Промежуточная аттестация помогает оценить более крупные совокупности знаний, умений и навыков, в некоторых случаях – даже формирование определенных компетенций. В рамках данного предмета к форме промежуточного контроля относится экзамен.

Экзамен по дисциплине имеет цель оценить сформированность компетенций, теоретическую подготовку студента, его способность к творческому мышлению, приобретенные им навыки самостоятельной работы, умение синтезировать полученные знания и применять их при решении практических задач. Экзамен проводится в устной форме. В ходе экзамена студент отвечает на вопросы билета. Билет включает два вопроса из списка “5.4. Вопросы промежуточной аттестации”, оцениваемых на 40 баллов. Каждый вопрос оценивается в 20 баллов. Дополнительные баллы, помимо баллов, полученных за курсовую и лабораторные работы, могут быть заработаны за правильные ответы в ходе опросов и собеседований.

Если суммарное число баллов набранных в семестре по результатам модулей и полученных на экзамене:

- от 61 до 75 , то ставится итоговая оценка "Удовлетворительно",
- от 76 до 90, то ставится итоговая оценка "Хорошо",
- от 91 до 100, то ставится итоговая оценка "Отлично".

Если суммарное число баллов, набранных студентом не менее 60 баллов, то студент может согласиться с соответствующей итоговой оценкой без экзамена.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

6.1. Рекомендуемая литература

	Авторы, составители	Заглавие	Издательство, год.	Электронный адрес
Л1.1	Бонцанини М., Логунова А. В.	Анализ социальных медиа на Python. Извлекайте и анализируйте данные из всех уголков социальной паутины на Python	Москва: ДМК Пресс, 2018	https://e.lanbook.com/reader/book/108129/#6
Л1.2	Маккинни У.	Python и анализ данных	Москва: ДМК Пресс, 2020	https://e.lanbook.com/book/131721
	Авторы, составители	Заглавие	Издательство, год.	Электронный адрес
Л2.1	Москвитин А. А.	Данные, информация, знания: методология, теория, технологии: монография	Санкт-Петербург: Лань, 2019	https://e.lanbook.com/reader/book/113937/#233
	Авторы, составители	Заглавие	Издательство, год.	Электронный адрес
Л3.1	Кузнецов С. Ю., Костикова А. В., Сайкина Ю. А.	Интеллектуальный анализ данных: учеб. пособие	Волгоград: ВолгГТУ, 2019	
Л3.2	Тюрин Ю. Н., Макаров А. А.	Статистический анализ данных на компьютере: [учеб. пособие]	М.: ИНФРА-М, 1998	
Л3.3	Орешков В. И.	Интеллектуальный анализ данных: учебное пособие	Рязань: РГРТУ, 2017	https://e.lanbook.com/reader/book/168028/#3

6.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	Онлайн-курс Академии BIG DATA: Введение в аналитику больших массивов данных // Национальный открытый университет Intuit. 2019. URL: https://intuit.ru/studies/courses/12385/1181/info (дата обращения: 01.04.2019).
----	---

6.3 Перечень программного обеспечения

6.3.1.1	Apache Hadoop — система хранения и обработки больших данных
6.3.1.2	PuTTY — SSH-клиент для удаленного доступа к вычислительному кластеру
6.3.1.3	WinSCP — SCP-клиент для обмена файлами с вычислительным кластером
6.3.1.4	СДО "Moodle" — система дистанционного обучения
6.3.1.5	Операционные системы Windows (на компьютерах пользователей) и CentOS (на вычислительном кластере)
6.3.1.6	Adobe Acrobat Reader DC — бесплатное решение для просмотра файлов PDF
6.3.1.7	LibreOffice — офисный пакет

6.4 Перечень информационных справочных систем и электронных библиотечных систем (ЭБС)

6.3.2.1	Библиотека (НТБ), http://library.vstu.ru/
6.3.2.2	Электронная информационно-образовательная среда университета, https://eos2.vstu.ru
6.3.2.3	ЭБС "Лань", https://e.lanbook.com/
6.3.2.4	ЭБС "Book.ru", https://www.book.ru/
6.3.2.5	Электронная библиотека "Grebennikon", https://grebennikon.ru/

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ) /ОБОРУДОВАНИЕ

7.1	Мультимедийная учебная аудитория для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации. /Учебная доска, учебная мебель, интерактивная трибуна, видеопроектор.
7.2	Лаборатория информационных технологий. /Учебная мебель, компьютерная техника, оснащенная программным обеспечением, доступом в Интернет и в электронную информационно-образовательную среду университета.
7.3	Аудитория для самостоятельной работы обучающихся. /Учебная мебель, компьютерная техника с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду университета (читальный зал информационно-библиотечного центра).

8. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

Организация образовательного процесса по данной дисциплине регламентируется учебным планом и расписанием учебных занятий. При формировании своей индивидуальной образовательной траектории обучающийся имеет право на перезачет дисциплины (переаттестации ее части), если она была освоена в процессе предшествующего обучения.

Перезачёт (переаттестации ее части) освобождает обучающегося от необходимости повторного освоения дисциплины (полностью или частично).

Учебный процесс при преподавании курса основывается на использовании традиционных, инновационных и информационных образовательных технологий. Традиционные образовательные технологии представлены лекциями и лабораторными работами. Инновационные образовательные технологии используются в виде широкого применения активных и интерактивных форм проведения занятий. Информационные образовательные технологии реализуются путем активизации самостоятельной работы студентов в электронной информационной образовательной среде.

Лекционный курс предполагает систематизированное изложение основных вопросов учебного плана. На первой лекции лектор информирует студентов о рекомендуемой литературе и электронных источниках информации по дисциплине, с указанием, какой учебник (учебное пособие) является базовым.

Лабораторные работы предполагают выполнение и отчет заданий по темам, рассмотренным на лекционных занятиях. Каждому лабораторному занятию предшествует самостоятельная подготовка студента, включающая: ознакомление с содержанием лабораторной работы по методическим указаниям; проработку теоретической части по лекционному материалу и учебникам, рекомендованным в методических указаниях;

Самостоятельная работа студентов включает изучение законспектированного на лекционных занятиях материала, дополнение его с учетом рекомендованной по данной теме литературы, самостоятельную подготовку к лабораторным работам, самостоятельное выполнение и оформление курсовой работы.

Перечень методических указаний для освоения дисциплины:

1. Методические указания к лабораторным работам по дисциплине «Вычислительные системы и сетевые технологии».

Часть 2 : метод. указания к лабораторным работам по дисциплине «Вычислительные системы и сетевые технологии» / сост.

А.Е. Андреев, П.Д. Кравченя, С.В. Гаевой – Волгоград : ИУНЛ ВолгГТУ, 2017. – 32 с.

В течении семестра для студентов проводятся групповые текущие консультации по учебной дисциплине, а также консультация перед экзаменом.

Методические рекомендации по обучению лиц с ограниченными возможностями здоровья и инвалидов

Профессорско-педагогический состав знакомится с психолого-физиологическими особенностями обучающихся инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ), индивидуальными программами реабилитации инвалидов (при наличии). При необходимости осуществляется дополнительная поддержка преподавания тьюторами, психологами, социальными работниками, прошедшими подготовку ассистентами.

В соответствии с методическими рекомендациями Минобрнауки РФ (утв. 8 апреля 2014 г. N АК-44/05вн), в курсе предполагается использовать социально-активные и рефлексивные методы обучения, технологии социокультурной реабилитации с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создании комфортного психологического климата в студенческой группе. Подбор и разработка учебных материалов производится с учетом предоставления материала в различных формах: аудиальной, визуальной, с использованием специальных технических средств и информационных систем.

Освоение дисциплины лицами с ОВЗ осуществляется с использованием средств обучения общего и специального назначения (персонального и коллективного использования). Материально-техническое обеспечение предусматривает приспособление аудиторий к нуждам лиц с ОВЗ (при необходимости).

Форма проведения аттестации для студентов-инвалидов устанавливается с учетом индивидуальных психофизических особенностей. Для студентов с ОВЗ предусматривается доступная форма предоставления заданий оценочных средств. Студентам с инвалидностью увеличивается время на подготовку ответов на контрольные вопросы. Для таких студентов предусматривается доступная форма предоставления ответов на задания.

При необходимости для обучающихся с инвалидностью процедура оценивания результатов обучения может проводиться в несколько этапов.