МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭНЕРГЕТИЧЕСКИЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «КГЭУ»)

Подписан: ФГБОУ ВО «КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭНЕРГЕТИЧЕСКИЙ УНИВЕРСИТЕТ»
Владелец: Беляев Эдуард Ирекович
Директор Института цифровых технологий и экономики
Сертификат: 01E4B4530091B044A4464D6B8C379A8EBA
Действителен с 04.10.2023 по 04.10.2024

УТВЕРЖДАЮ
Директор Института цифровых технологий и экономики
_____ Э.И. Беляев
«28» ноября 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.В.ДЭ.01.01.01 Пакетная и потоковая обработка данных

Направление подготовки 09.03.03 Прикладная информатика
(*Код и наименование направления подготовки*)

Направленность(и) (профиль(и)) Прикладной искусственный интеллект
(*Наименование направленности (профиля) образовательной программы*)

Квалификация Бакалавр

г. Казань, 2023

Программу разработал(и):

Наименование кафедры	Должность, уч.степень, уч.звание	ФИО разработчика
УрФУ ИТСУ	Ассистент	Аксёнов А.С.
УрФУ ИТСУ	Доцент, к.т.н.	Созыкин А.В.
УрФУ ИТСУ	Доцент, к.т.н.	Ронкин М.В.
КГЭУ ИТИС	Доцент, к.п.н., доцент	Еремина И.И.

Согласование	Наименование подразделения	Дата	№ протокола	Подпись
Одобрена	кафедра ИТИС	27.11.2023	11	_____ И.о. зав.каф.,к.ф.-м.н., доц. Соловьев С. А.
Согласована	Учебно-методический совет института ИЦТЭ	27.11.2023	3	_____ Директор, к.т.н., доц. Беляев Э.И.
Одобрена	Ученый совет института ИЦТЭ	28.11.2023	3	_____ Директор, к.т.н., доц. Беляев Э.И.

1. Цель, задачи и планируемые результаты обучения по дисциплине

Целью освоения дисциплины «Пакетная и потоковая обработка данных» является формирование у студентов базовых знаний и навыков в области обработки данных, а также в развитии их компетенций для решения практических задач в сфере анализа и обработки больших объемов данных.

Задачи обучения:

- Изучение основных понятий и методов пакетной и потоковой обработки данных.
- Развитие навыков работы с инструментами и технологиями для пакетной и потоковой обработки данных.
- Обучение студентов применению полученных знаний и умений в решении прикладных задач.
- Формирование у студентов навыков самостоятельной работы с источниками информации и анализа данных.
- Развитие критического мышления и умения применять полученные знания на практике.

Планируемые результаты обучения:

- Знание основных методов и инструментов пакетной и потоковой обработки данных и умение применять их на практике.
- Умение работать с большими объемами данных, анализировать и обрабатывать их с помощью специализированных программных средств.
- Навыки самостоятельной работы с литературой, базами данных и другими источниками информации, умение извлекать необходимую информацию для решения поставленных задач.
- Способность применять полученные знания и навыки для решения прикладных задач, связанных с обработкой и анализом данных.

Компетенции и индикаторы, формируемые у обучающихся:

Код и наименование компетенции	Код и наименование индикатора
ПК-6 Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-6.1 Осуществляет поиск данных в открытых источниках, специализированных библиотеках и репозиториях
	ПК-6.2 Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения
ПК-7 Способен разрабатывать системы анализа больших данных	ПК-7.1 Разрабатывает программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных
	ПК-7.2 Разрабатывает программные компоненты обработки, удаленной, распределенной и объединенной аналитики,

Код и наименование компетенции	Код и наименование индикатора
	использования результатов анализа, описания и управления качеством и достоверностью больших данных

2. Место дисциплины в структуре ОП

Предшествующие дисциплины :
Управление ИТ-проектами
Аналитика и визуализация данных
Машинное обучение
Сбор и верификация данных
Анализ данных и искусственный интеллект
Этика искусственного интеллекта
Программирование глубоких нейронных сетей
Производственная практика (технологическая (проектно-технологическая))

Последующие дисциплины
Рекомендательные системы
Платформы бизнес-анализа
Приложения искусственного интеллекта

3. Структура и содержание дисциплины

3.1. Структура дисциплины

Для очной формы обучения

Вид учебной работы	Всего ЗЕ	Всего часов	Семестр 7
			7
ОБЩАЯ ТРУДОЕМКОСТЬ ДИСЦИПЛИНЫ	3	108	108
КОНТАКТНАЯ РАБОТА*	-	61	61
АУДИТОРНАЯ РАБОТА	1,38	50	50
Лекции	0,44	18	18
Практические (семинарские) занятия	-	-	-
Лабораторные работы	0,94	32	32
САМОСТОЯТЕЛЬНАЯ РАБОТА ОБУЧАЮЩЕГОСЯ	1,62	22	22
Проработка учебного материала	0,62	22	22
Курсовой проект	-	-	-
Курсовая работа	-	-	-
Подготовка к промежуточной аттестации	1	36	36
Промежуточная аттестация:			Э

3.2. Содержание дисциплины, структурированное по разделам и видам занятий

Разделы дисциплины	Всего часов	Распределение трудоемкости по видам учебной работы				Формы и вид контроля	Индексы индикаторов формируемых компетенций
		лекции	лаб. раб.	пр. зан.	сам. раб.		
Распределенные данные, распределенная обработка и современные требования к скорости и надежности вычислений	12	4	4		4	ТК1	ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Spark. Основные парадигмы. Стадии обработки. Типы операций. API взаимодействия	13	4	6		3		ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Spark. Работа с dataframe	9	2	4		3	ТК2	ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Работа с данными типа "ключ — значение" (rdd)	9	2	4		3		ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Spark. Обработка данных и ML	9	2	4		3		ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Очереди и брокеры	11	2	6		3	ТК3	ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Потоковая обработка с учетом состояний и основы потоковой обработки	9	2	4		3		ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2
Экзамен	36				36	**ОМ 3**	**ПК-6.1, ПК-6.2, ПК-7.1, ПК-7.2**
Итого за 7 семестр	**108**	**18**	**32**	**0**	**58**		
ИТОГО	**108**	**18**	**32**	**0**	**58**		

3.3. Содержание дисциплины

Раздел 1. Распределенные данные, распределенная обработка и современные требования к скорости и надежности вычислений.

Понятие Big Data и его составляющие. Требования к высоконагруженным сервисам. Модели данных и языки запросов. Распределенные хранилища и распределенные данные. Производные данные. Введение в понятие пакетной и потоковой обработки.

Раздел 2. Spark. Основные парадигмы. Стадии обработки. Типы операций.

API взаимодействия.

Что такое spark? Место в экосистеме больших данных. Модель параллельных вычислений: отложенные вычисления, хранение данных в памяти и управление памятью, неизменяемость и интерфейс RDD, широкие и узкие зависимости, планирование заданий.

Раздел 3. Spark. Работа с dataframe.

Сессии и контексты. Схемы. API DataFrame. Оптимизация запросов. Эффективные преобразования.

Раздел 4. Работа с данными типа "ключ — значение" (rdd)

Как использовать классы PairRDDFunctions и OrderedRDDFunctions. Действия над парами "ключ — значение". Выбор операции агрегирования. Предотвращение ошибок нехватки памяти при операциях агрегирования.

Раздел 5. Spark. Обработка данных и ML

Выбор между библиотеками Spark MLlib и Spark ML. Кодирование признаков в библиотеке MLlib и подготовка данных. Обучение моделей библиотеки MLlib. Оценка модели MLlib. Этапы конвейера ML и Mllib.

Раздел 6. Очереди и брокеры.

Суть понятий. Обзор брокеров. Kafka. Основные понятия, архитектура, пакеты исходного кода. Разработка проекта, производители и потребители. Брокеры и темы. Kafka, как хранилище. Kafka Streams.

Раздел 7. Потоковая обработка с учетом состояний и основы потоковой обработки.

Flink: архитектура, DataStream, операторы на основе времени и оконные операторы. Обработка потоков с учетом состояния. Архитектура Apache Flink Настройка Flink для потоковых приложений.

3.4. Тематический план практических занятий

Данный вид работы не предусмотрен учебным планом.

3.5. Тематический план лабораторных работ

Лабораторная работа 1. Установка и развертывание Apache Spark

Лабораторная работа 2. Применение Apache Spark для считывания, обработки и записи данных

Лабораторная работа 3. Применение data frame API Apache Spark

Лабораторная работа 4. Применение rdd API Apache Spark

Лабораторная работа 5. Применение Apache Spark для решения задачи преобразования данных

Лабораторная работа 6. Применение Apache Spark для решения задачи

анализа данных

Лабораторная работа 7. Применение Apache Spark для решения задачи машинного обучения

Лабораторная работа 8. Установка и развертывание Apache Kafka

Лабораторная работа 9. Создания простейшего сервиса Apache Kafka, который «слушает» источник и передает данные на Apache Spark job

Лабораторная работа 10. Установка и развертывание Apache Flink

Лабораторная работа 11. Создание простейшего сервиса Apache Flink поставляющего данные на основе состояний

3.6. Курсовой проект /курсовая работа

Данный вид работы не предусмотрен учебным планом.

4. Оценивание результатов обучения

Оценивание результатов обучения по дисциплине осуществляется в рамках текущего контроля и промежуточной аттестации, проводимых по балльно-рейтинговой системе (БРС).

Шкала оценки результатов обучения по дисциплине:

Код компе-тенции	Код индикатора компетенции	Запланированные результаты обучения по дисциплине	Уровень сформированности индикатора компетенции			
			Высокий	Средний	Ниже среднего	Низкий
			от 85 до 100	от 70 до 84	от 55 до 69	от 0 до 54
			Шкала оценивания			
			отлично	хорошо	удовлет-ворительно	неудов-летвори-тельно
			зачтено			не зачтено
ПК-6	ПК-6.1	знать:				
		- технологии поиска данных в открытых источниках, специализированных библиотеках и репозиториях	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок	Уровень знаний в объеме, соответствующем программе, имеет место несколько негрубых ошибок	Минимально допустимый уровень знаний, имеет место много негрубых ошибок	Уровень знаний ниже минимальных требований, имеют место грубые ошибки
		уметь:				
		ориентироваться в теориях, концепция	Продемонстрированы все основные умения,	Продемонстрированы все основные	Продемонстрированы основные	При решении стандартных задач не

		х и направлениях дисциплины и давать им критическую оценку, используя научные достижения других дисциплин	решены все основные задачи, выполнены все задания в полном объеме	умения, решены все основ-ные задачи с негрубыми ошибка-ми, вы-полнены все зада-ния в полном объеме, но неко-торые с недочетами	умения, решены типовые задачи с негрубыми ошиб-ками, выполнены все задания, но не в полном объеме	продемонстрированы основные умения, имеют место грубые ошибки
		владеть:				
		- навыками высокого уровеня сформированности заявленных в рабочей программе компетенций; - владеть навыками самостоятельно и творчески решать сложные проблемы и нестандартные ситуации; - навыками применения теоретических знаний для выбора методики выполнения заданий; - навыками грамотного обоснования ход решения задач; - безупречно	Продемонстрированы навыки при решении нестандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов без ошибок и недочетов	Продемонстрированы базовые навыки при решении стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов с некоторыми недочетами.	Имеется минимальный набор навыков для решения стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов с некоторыми недочетами	Не продемонстрированы базовые навыки при решении стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов, имеют место грубые ошибки.

		владеть инструментарием учебной дисциплины, навыками его эффективного использования в постановке научных и практических задач; - навыкамитворческой самостоятельной работы на практических/семинарских/лабораторных занятиях, активно участвовать в групповых обсуждениях, высоким уровнем культуры исполнения заданий				
	ПК-6.2	знать:				
		- систематизированно, глубоко и полно концептуальные основы подготовки и разметки структурированных и неструктурированных данных для машинного	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок	Уровень знаний в объеме, соответствующем программе, имеет место несколько негрубых ошибок	Минимально допустимый уровень знаний, имеет место много негрубых ошибок	Уровень знаний ниже минимальных требований, имеют место грубые ошибки

		обучения; - все основные понятия и термины, используемые подготовке и разметке структурированных и неструктурированных данных для машинного обучения; - типовые модели подготовки и разметки структурированных и неструктурированных данных для машинного обучения.				
		уметь:				
		- проводить анализ основных методических приемов различных моделей подготовки и разметки структурированных и неструктурированных данных для машинного обучения; - использовать современные методики, разрабатывать регламент	Продемонстрированы все основные умения, решены все основные задачи, выполнены все задания в полном объеме	Продемонстрированы все основные умения, решены все основ-ные задачи с негрубыми ошибками, выполнены все задания в полном объеме, но некоторые с недочетами	Продемонстрированы основные умения, решены типовые задачи с негрубыми ошибками, выполнены все задания, но не в полном объеме	При решении стандартных задач не продемонстрированы основные умения, имеют место грубые ошибки

		ы подготовки и разметки структурированных и неструктурированных данных для машинного обучения.				
		владеть:				
		- навыками проведения анализа подготовки и разметки структурированных и неструктурированных данных для машинного обучения.	Продемонстрированы навыки при решении нестандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов без ошибок и недочетов	Продемонстрированы базовые навыки при решении стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов с некоторыми недочетами.	Имеется минимальный набор навыков для решения стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов с некоторыми недочетами	Не продемонстрированы базовые навыки при решении стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов, имеют место грубые ошибки.
ПК-7	ПК-7.1	знать:				
		современные программные компоненты извлечения, хранения, подготовки больших	-глубокие, всесторонние и аргументированные знания программного материала; -полное понимание	-знание и понимание основных вопросов контролируемого объема программного материала;	-знания теоретического материала; - неполные ответы на основные вопросы, ошибки в	- существенные пробелы в знаниях учебного материала; - допускаются принципиа

		данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	сущности и взаимосвязи рассматриваемых процессов и явлений, точное знание основных понятий, в рамках обсуждаемых заданий; - способность устанавливать и объяснять связь практики и теории, - логически последовательные, содержательные, конкретные и исчерпывающие ответы на все задания билета, а также дополнительные вопросы экзаменатора.	- знания теоретического материала - способность устанавливать и объяснять связь практики и теории, выявлять противоречия, проблемы и тенденции развития; - правильные и конкретные, без грубых ошибок, ответы на поставленные вопросы	ответе, недостаточное понимание сущности излагаемых вопросов; - неуверенные и неточные ответы на дополнительные вопросы.	льные ошибки при ответе на основные вопросы билета, отсутствует знание и понимание основных понятий и категорий; - непонимание сущности дополнительных вопросов в рамках заданий билета.
		уметь:				
		использовать современные программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных,	Правильно выполнил практическое задание билета. Показал отличные умения в рамках освоенного учебного материала. Решает предложенные практические задания без ошибок Ответил на	Выполнил практическое задание билета с небольшими неточностями. Показал хорошие умения в рамках освоенного учебного материала. Предложенные практические задания	Выполнил практическое задание билета с существенными неточностями. Допускаются ошибки в содержании ответа и решении практических	При выполнении практического задания билета обучающийся продемонстрировал недостаточный уровень умений. Практические задания не выполнены Обучающи

		определений, словарей и эталонной архитектуры больших данных	все дополнительные вопросы	решены с небольшими неточностями. Ответил на большинство дополнительных вопросов.	заданий. При ответах на дополнительные вопросы было допущено много неточностей.	йся не отвечает на вопросы билета при дополнительных наводящих вопросах преподавателя.
		владеть:				
		современными программными компонентами извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	Применяет теоретические знания для выбора методики выполнения заданий. Не допускает ошибок при выполнении заданий. Самостоятельно анализирует результаты выполнения заданий. Грамотно обосновывает ход решения задач.	Без затруднений выбирает стандартную методику выполнения заданий. Допускает ошибки при выполнении заданий, не нарушающие логику решения задач Делает корректные выводы по результатам решения задачи. Обосновывает ход решения задач без затруднений.	Испытывает затруднения по выбору методики выполнения заданий. Допускает ошибки при выполнении заданий, нарушения логики решения задач. Испытывает затруднения с формулированием корректных выводов. Испытывает затруднения при обосновании алгоритма выполнения заданий.	Не может выбрать методику выполнения заданий. Допускает грубые ошибки при выполнении заданий, нарушающие логику решения задач. Делает некорректные выводы. Не может обосновать алгоритм выполнения заданий.
	ПК-7.2	знать:				
	ПК-7.2	современные программн	-глубокие, всесторонние и	-знание и понимание основных	-знания теоретического	-существенные пробелы

		ые компоненты обработки, удаленной, распределенной и объединенной аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных	аргументированные знания программного материала; -полное понимание сущности и взаимосвязи рассматриваемых процессов и явлений, точное знание основных понятий, в рамках обсуждаемых заданий; - способность устанавливать и объяснять связь практики и теории, - логически последовательные, содержательные, конкретные и исчерпывающие ответы на все задания билета, а также дополнительные вопросы экзаменатора.	вопросов контролируемого объема программного материала; - знания теоретического материала - способность устанавливать и объяснять связь практики и теории, выявлять противоречия, проблемы и тенденции развития; - правильные и конкретные, без грубых ошибок, ответы на поставленные вопросы	материала; - неполные ответы на основные вопросы, ошибки в ответе, недостаточное понимание сущности излагаемых вопросов; - неуверенные и неточные ответы на дополнительные вопросы.	в знаниях учебного материала; - допускаются принципиальные ошибки при ответе на основные вопросы билета, отсутствует знание и понимание основных понятий и категорий; - непонимание сущности дополнительных вопросов в рамках заданий билета.
		уметь:				
		использовать современные программные компоненты обработки, удаленной, распределенной и	Правильно выполнил практическое задание билета. Показал отличные умения в рамках освоенного учебного материала.	Выполнил практическое задание билета с небольшими неточностями. Показал хорошие умения в рамках	Выполнил практическое задание билета с существенными неточностями. Допускаются	При выполнении практического задания билета обучающийся продемонстрировал недостаточный

		объединенной аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных	Решает предложенные практические задания без ошибок Ответил на все дополнительные вопросы	освоенного учебного материала. Предложенные практические задания решены с небольшими неточностями. Ответил на большинство дополнительных вопросов.	ошибки в содержании ответа и решении практических заданий. При ответах на дополнительные вопросы было допущено много неточностей.	уровень умений. Практические задания не выполнены Обучающийся не отвечает на вопросы билета при дополнительных наводящих вопросах преподавателя.
		владеть:				
		современными программными компонентами извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	Применяет теоретические знания для выбора методики выполнения заданий. Не допускает ошибок при выполнении заданий. Самостоятельно анализирует результаты выполнения заданий. Грамотно обосновывает ход решения задач.	Без затруднений выбирает стандартную методику выполнения заданий. Допускает ошибки при выполнении заданий, не нарушающие логику решения задач Делает корректные выводы по результатам решения задачи. Обосновывает ход решения задач без затруднений.	Испытывает затруднения по выбору методики выполнения заданий. Допускает ошибки при выполнении заданий, нарушения логики решения задач. Испытывает затруднения с формулированием корректных выводов. Испытывает затруднения при обосновании алгоритма	Не может выбрать методику выполнения заданий. Допускает грубые ошибки при выполнении заданий, нарушающие логику решения задач. Делает некорректные выводы. Не может обосновать алгоритм выполнения заданий.

					выполнения заданий.	

Оценочные материалы для проведения текущего контроля и промежуточной аттестации приведены в Приложении к рабочей программе дисциплины.

Полный комплект заданий и материалов, необходимых для оценивания результатов обучения по дисциплине, хранится на кафедре разработчика.

5. Учебно-методическое и информационное обеспечение дисциплины

5.1. Учебно-методическое обеспечение

5.1.1. Основная литература

1. Гулямов, С. С., Искусственный интеллект и когнитивные технологии в экономике : учебное пособие / С. С. Гулямов, А. Т. Шермухамедов, Б. М. Холбоев. — Москва : Русайнс, 2024. — 285 с. — ISBN 978-5-466-04173-6. — URL: https://book.ru/book/951458. — Текст : электронный.
2. Криволапов, С. Я., Математика на Python : учебник / С. Я. Криволапов, М. Б. Хрипунова. — Москва : КноРус, 2024. — 455 с. — ISBN 978-5-406-12069-9. — URL: https://book.ru/book/950432. — Текст : электронный.

5.1.2.Дополнительная литература

1. Волгина, О. А., Математическое моделирование экономических процессов и систем : учебное пособие / О. А. Волгина, Г. И. Шуман. — Москва : КноРус, 2022. — 256 с. — ISBN 978-5-406-08869-2. — URL: https://book.ru/book/941747. — Текст : электронный.
2. Лесковец, Юре. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман, Дж. Ульман ; пер. с анг. А. А. Слинкина. - Москва : ДМК Пресс, 2016. - 500 с. - URL: http://new.ibooks.ru/bookshelf/364297 . - ISBN 978-5-97060-190-7. - Текст : электронный.
3. Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 4-е изд., стер. — Санкт-Петербург : Лань, 2024. — 188 с. — ISBN 978-5-507-47346-5. — Текст : электронный // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/362318. - Текст : электронный.
4. Анализ больших данных : учебное пособие / И. Б. Тесленко, В. Е. Крылов, А. М. Губернаторов [и др.]. — Москва : КноРус, 2023. — 295 с. — ISBN 978-5-406-10550-4. — URL: https://book.ru/book/950469. — Текст : электронный.

5.2. Информационное обеспечение

5.2.1. Электронные и интернет-ресурсы

№ п/п	Наименование электронных и интернет-ресурсов	Ссылка
1	Электронно-библиотечная система «Лань»	https://e.lanbook.com/
2	Электронно-библиотечная система «ibooks.ru»	https://ibooks.ru/
3	Электронно-библиотечная система «book.ru»	https://www.book.ru/
4	Портал «Открытое образование»	https://npoed.ru
5	Российская национальная библиотека	https://nlr.ru/
6	КиберЛенинка	https://cyberleninka.ru
7	Техническая библиотека	https://techlibrary.ru
8	Национальная электронная библиотека (НЭБ)	https://rusneb.ru/

5.2.2. Профессиональные базы данных / Информационно-справочные системы

№ п/п	Наименование профессиональных баз данных	Адрес	Режим доступа
1	Официальный интернет-портал правовой информации	http://pravo.gov.ru	http://pravo.gov.ru
2	Справочная правовая система «Консультант Плюс»	http://consultant.ru	http://consultant.ru
3	Справочно-правовая система по законодательству РФ	http://garant.ru	http://garant.ru

5.2.3. Лицензионное и свободно распространяемое программное обеспечение дисциплины

№ п/п	Наименование программного обеспечения	Описание	Реквизиты подтверждающих документов
1	Операционная система Microsoft Windows 10	Пользовательская операционная система	Договор №133/2021 от 12.10.2021, лицензиар - ЗАО «Софт Лайн Трейд», тип (вид) лицензии - неискл. право, срок действия лицензии - бессрочно
2	Microsoft Office 2019	Пакет офисных приложений	Договор №133/2021 от 12.10.2021, лицензиар - ЗАО «Софт Лайн Трейд», тип (вид) лицензии - неискл. право, срок действия лицензии - бессрочно
3	LMS Moodle	ПО для эффективного онлайн- взаимодействия преподавателя и студента	Свободная лицензия, тип (вид) лицензии - неискл. право, срок действия лицензии - бессрочно.

№ п/п	Наименование программного обеспечения	Описание	Реквизиты подтверждающих документов
4	Браузер Chrome	Система поиска информации в сети интернет	Свободная лицензия, тип (вид) лицензии - неискл. право, срок действия лицензии - бессрочно.

6. Материально-техническое обеспечение дисциплины

Наименование вида учебной работы	Наименование учебной аудитории, специализированной лаборатории	Перечень необходимого оборудования и технических средств обучения
Лекции	Учебная аудитория для проведения занятий лекционного типа	Специализированная учебная мебель, технические средства обучения, служащие для представления учебной информации большой аудитории (мультимедийный проектор, компьютер (ноутбук), экран), демонстрационное оборудование, учебно-наглядные пособия
Лабораторные работы	Учебная лаборатория программной инженерии, ауд. В-608	Специализированное лабораторное оборудование по профилю лаборатории программной инженерии, учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Компьютерный класс с выходом в Интернет, ауд. В-610	Специализированная учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Учебная лаборатория информационной безопасности, ауд. В-615	Специализированное лабораторное оборудование по профилю лаборатории информационной безопасности, учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Компьютерный класс с выходом в Интернет, ауд. В-617	Специализированная учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное

Наименование вида учебной работы	Наименование учебной аудитории, специализированной лаборатории	Перечень необходимого оборудования и технических средств обучения
		обеспечение
	Компьютерный класс с выходом в Интернет, ауд. В-619	Специализированная учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Компьютерный класс с выходом в Интернет, ауд. В-621	Специализированная учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Учебная лаборатория реинжиниринга и управления бизнес-процессами, ауд. В-623	Специализированное лабораторное оборудование по профилю лаборатории реинжиниринга и управления бизнес-процессами, учебная мебель, технические средства обучения (мультимедийный проектор, мультимедийная доска, моноблоки), необходимое лицензионное программное обеспечение
	Компьютерный класс с выходом в Интернет В-600	Специализированная учебная мебель на 30 посадочных мест, 30 компьютеров, технические средства обучения (мультимедийный проектор, компьютер (ноутбук), экран), видеокамеры, программное обеспечение
Самостоятельная работа	Компьютерный класс с выходом в Интернет В-600	Специализированная учебная мебель на 30 посадочных мест, 30 компьютеров, технические средства обучения (мультимедийный проектор, компьютер (ноутбук), экран), видеокамеры, программное обеспечение
	Читальный зал библиотеки	Специализированная мебель, компьютерная техника с возможностью выхода в Интернет и обеспечением доступа в ЭИОС, экран, мультимедийный проектор, программное обеспечение

7. Особенности организации образовательной деятельности для лиц с ограниченными возможностями здоровья и инвалидов

Лица с ограниченными возможностями здоровья (ОВЗ) и инвалиды имеют возможность беспрепятственно перемещаться из одного учебно-лабораторного

корпуса в другой, подняться на все этажи учебно-лабораторных корпусов, заниматься в учебных и иных помещениях с учетом особенностей психофизического развития и состояния здоровья.

Для обучения лиц с ОВЗ и инвалидов, имеющих нарушения опорно-двигательного аппарата, обеспечены условия беспрепятственного доступа во все учебные помещения. Информация о специальных условиях, созданных для обучающихся с ОВЗ и инвалидов, размещена на сайте университета www//kgeu.ru. Имеется возможность оказания технической помощи ассистентом, а также услуг сурдопереводчиков и тифлосурдопереводчиков.

Для адаптации к восприятию лицами с ОВЗ и инвалидами с нарушенным слухом справочного, учебного материала по дисциплине обеспечиваются следующие условия:

- для лучшей ориентации в аудитории, применяются сигналы оповещения о начале и конце занятия (слово «звонок» пишется на доске);
- внимание слабослышащего обучающегося привлекается педагогом жестом (на плечо кладется рука, осуществляется нерезкое похлопывание);
- разговаривая с обучающимся, педагогический работник смотрит на него, говорит ясно, короткими предложениями, обеспечивая возможность чтения по губам.

Компенсация затруднений речевого и интеллектуального развития слабослышащих обучающихся проводится путем:

- использования схем, диаграмм, рисунков, компьютерных презентаций с гиперссылками, комментирующими отдельные компоненты изображения;
- регулярного применения упражнений на графическое выделение существенных признаков предметов и явлений;
- обеспечения возможности для обучающегося получить адресную консультацию по электронной почте по мере необходимости.

Для адаптации к восприятию лицами с ОВЗ и инвалидами с нарушениями зрения справочного, учебного, просветительского материала, предусмотренного образовательной программой по выбранному направлению подготовки, обеспечиваются следующие условия:

- ведется адаптация официального сайта в сети Интернет с учетом особых потребностей инвалидов по зрению, обеспечивается наличие крупношрифтовой справочной информации о расписании учебных занятий;
- педагогический работник, его собеседник (при необходимости), присутствующие на занятии, представляются обучающимся, при этом каждый раз называется тот, к кому педагогический работник обращается;
- действия, жесты, перемещения педагогического работника коротко и ясно комментируются;
- печатная информация предоставляется крупным шрифтом (от 18 пунктов), тотально озвучивается;
- обеспечивается необходимый уровень освещенности помещений;
- предоставляется возможность использовать компьютеры во время занятий и право записи объяснений на диктофон (по желанию обучающихся).

Форма проведения текущей и промежуточной аттестации для

обучающихся с ОВЗ и инвалидов определяется педагогическим работником в соответствии с учебным планом. При необходимости обучающемуся с ОВЗ, инвалиду с учетом их индивидуальных психофизических особенностей дается возможность пройти промежуточную аттестацию устно, письменно на бумаге, письменно на компьютере, в форме тестирования и т.п., либо предоставляется дополнительное время для подготовки ответа.

8. Методические рекомендации для преподавателей по организации воспитательной работы с обучающимися.

Методическое обеспечение процесса воспитания обучающихся выступает одним из определяющих факторов высокого качества образования. Преподаватель вуза, демонстрируя высокий профессионализм, эрудицию, четкую гражданскую позицию, самодисциплину, творческий подход в решении профессиональных задач, в ходе образовательного процесса способствует формированию гармоничной личности.

При реализации дисциплины преподаватель может использовать следующие методы воспитательной работы:

- методы формирования сознания личности (беседа, диспут, внушение, инструктаж, контроль, объяснение, пример, самоконтроль, рассказ, совет, убеждение и др.);
- методы организации деятельности и формирования опыта поведения (задание, общественное мнение, педагогическое требование, поручение, приучение, создание воспитывающих ситуаций, тренинг, упражнение, и др.);
- методы мотивации деятельности и поведения (одобрение, поощрение социальной активности, порицание, создание ситуаций успеха, создание ситуаций для эмоционально-нравственных переживаний, соревнование и др.)

При реализации дисциплины преподаватель должен учитывать следующие направления воспитательной деятельности:

Гражданское и патриотическое воспитание:

- формирование у обучающихся целостного мировоззрения, российской идентичности, уважения к своей семье, обществу, государству, принятым в семье и обществе духовно-нравственным и социокультурным ценностям, к национальному, культурному и историческому наследию, формирование стремления к его сохранению и развитию;
- формирование у обучающихся активной гражданской позиции, основанной на традиционных культурных, духовных и нравственных ценностях российского общества, для повышения способности ответственно реализовывать свои конституционные права и обязанности;
- развитие правовой и политической культуры обучающихся, расширение конструктивного участия в принятии решений, затрагивающих их права и интересы, в том числе в различных формах самоорганизации, самоуправления, общественно-значимой деятельности;
- формирование мотивов, нравственных и смысловых установок личности, позволяющих противостоять экстремизму, ксенофобии, дискриминации по социальным, религиозным, расовым, национальным

признакам, межэтнической и межконфессиональной нетерпимости, другим негативным социальным явлениям.

Духовно-нравственное воспитание:

- воспитание чувства достоинства, чести и честности, совестливости, уважения к родителям, учителям, людям старшего поколения;
- формирование принципов коллективизма и солидарности, духа милосердия и сострадания, привычки заботиться о людях, находящихся в трудной жизненной ситуации;
- формирование солидарности и чувства социальной ответственности по отношению к людям с ограниченными возможностями здоровья, преодоление психологических барьеров по отношению к людям с ограниченными возможностями;
- формирование эмоционально насыщенного и духовно возвышенного отношения к миру, способности и умения передавать другим свой эстетический опыт.

Культурно-просветительское воспитание:

- формирование эстетической картины мира;
- формирование уважения к культурным ценностям родного города, края, страны;
- повышение познавательной активности обучающихся.

Научно-образовательное воспитание:

- формирование у обучающихся научного мировоззрения;
- формирование умения получать знания;
- формирование навыков анализа и синтеза информации, в том числе в профессиональной области.

Вносимые изменения и утверждения на новый учебный год

№ п/п	№ раздела внесения изменений	Дата внесения изменений	Содержание изменений	«Согласовано» Зав. каф. реализующей дисциплину	«Согласовано» председатель УМК института (факультета), в состав которого входит выпускающая
1	2	3	4	5	6
1					
2					
3					

Приложение к рабочей программе дисциплины

К Г Э У

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭНЕРГЕТИЧЕСКИЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «КГЭУ»)

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ
по дисциплине

Б1.В.ДЭ.01.01.01 Пакетная и потоковая обработка данных

г. Казань, 2023

Оценочные материалы по дисциплине *Пакетная и потоковая обработка данных*, предназначенны для оценивания результатов обучения на соответствие индикаторам достижения компетенций.

Оценивание результатов обучения по дисциплине осуществляется в рамках текущего контроля (ТК) и промежуточной аттестации, проводимых по балльно-рейтинговой системе (БРС).

1.Технологическая карта

Семестр 7.

Наименование раздела	Формы и вид контроля	Рейтинговые показатели							
		I текущий контроль	Дополнительные баллы к ТК1	II текущий контроль	Дополнительные баллы к ТК2	III текущий контроль	Дополнительные баллы к ТК3	Итого	Промежуточная аттестация
Раздел 1. «Распределенные данные, распределенная обработка и современные требования к скорости и надежности вычислений»	**ТК1**	**20**	**0-15**					**20-35**	**20-35**
Тест или письменный опрос		3							
Защита лабораторной работы		7							
Выполнение индивидуальных заданий (рефератов)			0-7						
Раздел 2. «Spark. Основные парадигмы. Стадии обработки. Типы операций. API взаимодействия»									
Тест или письменный опрос		3							
Защита лабораторной работы		7							
Выполнение индивидуальных заданий (рефератов)			0-8						
Раздел 3. «Spark. Работа с dataframe»	**ТК2**			**20**	**0-15**			**20-35**	**20-35**
Тест или письменный опрос				3					
Защита лабораторной работы				4					
Выполнение индивидуальных заданий (рефератов)					0-5				
Раздел 4. «Работа с данными типа "ключ — значение" (rdd)»									
Тест или письменный опрос				3					
Защита лабораторной работы				4					
Выполнение индивидуальных заданий (рефератов)					0-5				

Раздел 5. «Spark. Обработка данных и ML»									
Тест или письменный опрос				3					
Защита лабораторной работы				3					
Выполнение индивидуальных заданий (рефератов)					0-5				
Раздел 6. «Очереди и брокеры»	**ТКЗ**					**15**	**0-15**	**15-30**	**15-30**
Тест или письменный опрос						3			
Защита лабораторной работы						7			
Выполнение индивидуальных заданий (рефератов)							0-7		
Раздел 7. «Потоковая обработка с учетом состояний и основы потоковой обработки»									
Тест или письменный опрос						3			
Защита лабораторной работы						7			
Выполнение индивидуальных заданий (рефератов)							0-8		
Промежуточная аттестация (экзамен)	**ОМ**								**0-45**
В форме теста									0-45

2. Оценочные материалы текущего контроля и промежуточной аттестации

Шкала оценки результатов обучения по дисциплине:

Код компе-тенции	Код индикатора компетенции	Запланированные результаты обучения по дисциплине	Уровень сформированности индикатора компетенции			
			Высокий	Средний	Ниже среднего	Низкий
			от 85 до 100	от 70 до 84	от 55 до 69	от 0 до 54
			Шкала оценивания			
			отлично	хорошо	удовлет-ворительно	неудов-летвори-тельно
			зачтено			не зачтено
ПК-6	ПК-6.1	знать:				
		- технологии поиска данных в открытых источниках, специализированных библиотеках и репозиториях	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок	Уровень знаний в объеме, соответствующем программе, имеет место несколько негрубых ошибок	Минимально допустимый уровень знаний, имеет место много негрубых ошибок	Уровень знаний ниже минимальных требований, имеют место грубые ошибки
		уметь:				

		ориентироваться в теориях, концепциях и направлениях дисциплины и давать им критическую оценку, используя научные достижения других дисциплин	Продемонстрированы все основные умения, решены все основные задачи, выполнены все задания в полном объеме	Продемонстрированы все основные умения, решены все основ-ные задачи с негрубыми ошибка-ми, вы-полнены все зада-ния в полном объеме, но неко-торые с недочетами	Продемонстрированы основные умения, решены типовые задачи с негрубыми ошиб-ками, выполнены все задания, но не в полном объеме	При решении стандартных задач не продемонстрированы основные умения, имеют место грубые ошибки
		владеть:				
		- навыками высокого уровеня сформированности заявленных в рабочей программе компетенций; - владеть навыками самостоятельно и творчески решать сложные проблемы и нестандартные ситуации; - навыками применения теоретических знаний для выбора методики выполнения заданий; - навыками грамотного обоснования ход решения задач;	Продемонстрированы навыки при решении нестандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов без ошибок и недочетов	Продемонстрированы базовые навыки при решении стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов с некоторыми недочетами.	Имеется минимальный набор навыков для решения стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов с некоторыми недочетами	Не продемонстрированы базовые навыки при решении стандартных задач, поиск данных в открытых источниках, специализированных библиотеках и репозиториях, обработки и интерпретации их результатов, имеют место грубые ошибки.

		- безупречно владеть инструментарием учебной дисциплины, навыками его эффективного использования в постановке научных и практических задач; - навыкамитворческой самостоятельной работы на практических/семинарских/лабораторных занятиях, активно участвовать в групповых обсуждениях, высоким уровнем культуры исполнения заданий				
	ПК-6.2	знать:				
		- систематизированно, глубоко и полно концептуальные основы подготовки и разметки структурированных и неструктурированных данных для машинного обучения; - все основные	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок	Уровень знаний в объеме, соответствующем программе, имеет место несколько негрубых ошибок	Минимально допустимый уровень знаний, имеет место много негрубых ошибок	Уровень знаний ниже минимальных требований, имеют место грубые ошибки

		понятия и термины, используемые подготовке и разметке структурированных и неструктурированных данных для машинного обучения; - типовые модели подготовки и разметки структурированных и неструктурированных данных для машинного обучения.				
		уметь:				
		- проводить анализ основных методических приемов различных моделей подготовки и разметки структурированных и неструктурированных данных для машинного обучения; - использовать современные методики, разрабатывать регламенты подготовки и разметки структурированных и неструктурированных	Продемонстрированы все основные умения, решены все основные задачи, выполнены все задания в полном объеме	Продемонстрированы все основные умения, решены все основ-ные задачи с негрубыми ошибками, выполнены все задания в полном объеме, но некоторые с недочетами	Продемонстрированы основные умения, решены типовые задачи с негрубыми ошибками, выполнены все задания, но не в полном объеме	При решении стандартных задач не продемонстрированы основные умения, имеют место грубые ошибки

		данных для машинного обучения.				
		владеть:				
		- навыками проведения анализа подготовки и разметки структурированных и неструктурированных данных для машинного обучения.	Продемонстрированы навыки при решении нестандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов без ошибок и недочетов	Продемонстрированы базовые навыки при решении стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов с некоторыми недочетами.	Имеется минимальный набор навыков для решения стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов с некоторыми недочетами	Не продемонстрированы базовые навыки при решении стандартных задач подготовки и разметки структурированных и неструктурированных данных для машинного обучения, обработки и интерпретации их результатов, имеют место грубые ошибки.
ПК-7	ПК-7.1	знать:				
		современные программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей	-глубокие, всесторонние и аргументированные знания программного материала; -полное понимание сущности и взаимосвязи рассматриваемых процессов и явлений, точное	-знание и понимание основных вопросов контролируемого объема программного материала; - знания теоретического материала - способность устанавлив	-знания теоретического материала; - неполные ответы на основные вопросы, ошибки в ответе, недостаточное понимание сущности излагаем	- существенные пробелы в знаниях учебного материала; - допускаются принципиальные ошибки при ответе на основные вопросы билета, отсутствует знание и понимание основных

		и эталонной архитектуры больших данных	знание основных понятий, в рамках обсуждаемых заданий; - способность устанавливать и объяснять связь практики и теории, - логически последовательные, содержательные, конкретные и исчерпывающие ответы на все задания билета, а также дополнительные вопросы экзаменатора.	ать и объяснять связь практики и теории, выявлять противоречия, проблемы и тенденции развития; - правильные и конкретные, без грубых ошибок, ответы на поставленные вопросы	ых вопросов; - неуверенные и неточные ответы на дополнительные вопросы.	понятий и категорий; - непонимание сущности дополнительных вопросов в рамках заданий билета.
		уметь:				
		использовать современные программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	Правильно выполнил практическое задание билета. Показал отличные умения в рамках освоенного учебного материала. Решает предложенные практические задания без ошибок Ответил на все дополнительные вопросы	Выполнил практическое задание билета с небольшими неточностями. Показал хорошие умения в рамках освоенного учебного материала. Предложенные практические задания решены с небольшими неточностями. Ответил на большинст	Выполнил практическое задание билета с существенными неточностями. Допускаются ошибки в содержании ответа и решении практических заданий. При ответах на дополнительные вопросы	При выполнении практического задания билета обучающийся продемонстрировал недостаточный уровень умений. Практические задания не выполнены Обучающийся не отвечает на вопросы билета при дополнительных наводящих вопросах преподавателя.

				во дополнительных вопросов.	было допущено много неточностей.	
		владеть:				
		современными программными компонентами извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	Применяет теоретические знания для выбора методики выполнения заданий. Не допускает ошибок при выполнении заданий. Самостоятельно анализирует результаты выполнения заданий. Грамотно обосновывает ход решения задач.	Без затруднений выбирает стандартную методику выполнения заданий. Допускает ошибки при выполнении заданий, не нарушающие логику решения задач Делает корректные выводы по результатам решения задачи. Обосновывает ход решения задач без затруднений.	Испытывает затруднения по выбору методики выполнения заданий. Допускает ошибки при выполнении заданий, нарушения логики решения задач. Испытывает затруднения с формулированием корректных выводов. Испытывает затруднения при обосновании алгоритма выполнения заданий.	Не может выбрать методику выполнения заданий. Допускает грубые ошибки при выполнении заданий, нарушающие логику решения задач. Делает некорректные выводы. Не может обосновать алгоритм выполнения заданий.
	ПК-7.2	знать:				
		современные программные компоненты обработки, удаленной, распределенной и объединенн	-глубокие, всесторонние и аргументированные знания программного материала; -полное понимание	-знание и понимание основных вопросов контролируемого объема программного материала;	-знания теоретического материала; - неполные ответы на основные вопросы, ошибки в	- существенные пробелы в знаниях учебного материала; - допускаются принципиальные ошибки при ответе

		ой аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных	сущности и взаимосвязи рассматриваемых процессов и явлений, точное знание основных понятий, в рамках обсуждаемых заданий; - способность устанавливать и объяснять связь практики и теории, - логически последовательные, содержательные, конкретные и исчерпывающие ответы на все задания билета, а также дополнительные вопросы экзаменатора.	- знания теоретического материала - способность устанавливать и объяснять связь практики и теории, выявлять противоречия, проблемы и тенденции развития; - правильные и конкретные, без грубых ошибок, ответы на поставленные вопросы	ответе, недостаточное понимание сущности излагаемых вопросов; - неуверенные и неточные ответы на дополнительные вопросы.	на основные вопросы билета, отсутствует знание и понимание основных понятий и категорий; - непонимание сущности дополнительных вопросов в рамках заданий билета.
		уметь:				
		использовать современные программные компоненты обработки, удаленной, распределенной и объединенной аналитики, использования результатов анализа, описания и	Правильно выполнил практическое задание билета. Показал отличные умения в рамках освоенного учебного материала. Решает предложенные практические задания без ошибок Ответил на	Выполнил практическое задание билета с небольшими неточностями. Показал хорошие умения в рамках освоенного учебного материала. Предложенные практические задания	Выполнил практическое задание билета с существенными неточностями. Допускаются ошибки в содержании ответа и решении практических	При выполнении практического задания билета обучающийся продемонстрировал недостаточный уровень умений. Практические задания не выполнены Обучающийся не отвечает на вопросы

		управления качеством и достоверностью больших данных	все дополнительные вопросы	решены с небольшими неточностями. Ответил на большинство дополнительных вопросов.	заданий. При ответах на дополнительные вопросы было допущено много неточностей.	билета при дополнительных наводящих вопросах преподавателя.
		владеть:				
		современными программными компонентами извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных	Применяет теоретические знания для выбора методики выполнения заданий. Не допускает ошибок при выполнении заданий. Самостоятельно анализирует результаты выполнения заданий. Грамотно обосновывает ход решения задач.	Без затруднений выбирает стандартную методику выполнения заданий. Допускает ошибки при выполнении заданий, не нарушающие логику решения задач Делает корректные выводы по результатам решения задачи. Обосновывает ход решения задач без затруднений.	Испытывает затруднения по выбору методики выполнения заданий. Допускает ошибки при выполнении заданий, нарушения логики решения задач. Испытывает затруднения с формулированием корректных выводов. Испытывает затруднения при обосновании алгоритма выполнения заданий.	Не может выбрать методику выполнения заданий. Допускает грубые ошибки при выполнении заданий, нарушающие логику решения задач. Делает некорректные выводы. Не может обосновать алгоритм выполнения заданий.

Оценка **«отлично»** выставляется за выполнение *лабораторных работ в семестре; тестовых заданий; глубокое понимание современных программных компонентов обработки, удаленной, распределенной и объединенной*

аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных, полные и содержательные ответы на вопросы билета (теоретическое и практическое задание)/качественные ответы на тест и умение аргументировать выбранный вариант ответа;

Оценка **«хорошо»** выставляется за выполнение *лабораторных работ в семестре; тестовых заданий; понимание современных программных компонентов обработки, удаленной, распределенной и объединенной аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных, ответы на вопросы билета (теоретическое или практическое задание)/ ответы на тест и умение аргументировать выбранный вариант ответа;*

Оценка **«удовлетворительно»** выставляется за выполнение *лабораторных работ в семестре и тестовых заданий;*

Оценка **«неудовлетворительно»** выставляется за слабое и неполное выполнение *лабораторных работ в семестре и тестовых заданий*.

3. Перечень оценочных средств

Краткая характеристика оценочных средств, используемых при текущем контроле успеваемости и промежуточной аттестации обучающегося по дисциплине:

Наименование оценочного средства	Краткая характеристика оценочного средства	Описание оценочного средства
Тест (Тест)	Система стандартизированных заданий, позволяющая автоматизировать процедуру измерения уровня знаний и умений обучающегося	Комплект тестовых заданий
Отчет по лабораторной работе (ОЛР)	Выполнение лабораторной работы, обработка результатов эксперимента. Оформление отчета, защита результатов лабораторной работы по отчету	Перечень заданий и вопросов для защиты лабораторной работы, перечень требований к отчету
Реферат (Рфр)	Продукт самостоятельной работы студента, представляющий собой краткое изложение в письменном виде полученных результатов теоретического анализа определенной научной (учебно-исследовательской) темы	Темы рефератов
Экзамен	Оценочные материалы, вынесенные на экзамен, состоят из системы стандартизированных заданий, позволяющих автоматизировать процедуру оценки уровня знаний и умений обучающегося	Комплект тестовых заданий База вопросов и задании не менее 500

4. Перечень контрольных заданий или иные материалы, необходимые для оценки знаний, умений и навыков, характеризующих этапы формирования компетенций в процессе освоения дисциплины

Пример задания

Для текущего контроля ТК1:

Проверяемая компетенция: ПК-6

Тест

1. Большинство данных в мире в 2011 году содержалось:
– в цифровом виде
– в аналоговом виде

2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?
– 20

3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
– петабайтами
– зеттабайтами
– экзабайтами
– йоттабайтами

4. Сколько Петабайт в Зеттабайте? Укажите число.
– 1024

5. укажите фактор, способствовавший появлению тренда больших данных
– маркетинговые кампании крупных корпораций
– снижение издержек на хранение данных
– появление новых технологий обработки потоковых данных
– выпуск баз данных с обработкой данных в памяти

6. Какие вероятные разочарования тренда больших данных?
– из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных
– из-за угрозы безопасности личной жизни (privacy) граждан будут упрощены процедуры сбора данных, что приведёт к падению ценности больших данных
– нет

7. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
– разработка Hadoop
– изобретение принципа MapReduce
– разработка языка Pyhton
– победа Deepblue в матче с Г.Каспаровым

8. Выберите верный ответ
– большие данные – это обработка или хранение более 1 Тб информации
– проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна
– большие данные – это огромная PR-акция крупных вендоров и не более того

– большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект

9. Выберите неверный ответ:
– большие данные – это данные объёма свыше 1 Тб
– проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна
– большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров
– большие данные как правило не структурированы

10. Отметьте те из вариантов, в которых данные структурированы:
– данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word
– таблица с ежедневными показаниями температуры помещения за год в файле формата csv
– текст педагогической поэмы А.С. Макаренко, представленный в формате PDF
– библиотека фильмов, представленных в формате mpeg4 на одном жестком диске

11. Перечислите четыре основных характеристики Big Data:
– Virtualization, Volume, Variability, Vehicle
– Variety, Velocity, Volume, Value
– Verification, Volume, Velocity, Visualization
– Video, Value, Variety, Volume

12. Выберите неверное высказывание:
– большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных
– увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации
– удешевление систем хранения на единицу информации привело к росту рынка больших данных

13. Отметьте неверное понимание Variety в контексте характеристик Big Data:
– высокая скорость генерирования данных
– разные типы данных в колонках таблиц реляционных СУБД
– разнообразие отраслей, являющихся источниками данных
– разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные

14. Принцип MapReduce состоит в том, чтобы
– производить вычисления на узлах, где информация изначально была сохранена
– использовать вычислительные мощности систем хранения
– использовать функциональное программирование для решения задач массивно-параллельной обработки

Проверяемая компетенция: ПК-7

1. Выберите одно неверное высказывание про MapReduce:
– интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
– MapReduce – это две операции: распределения и сборки данных
– MapReduce был придуман разработчиками Hadoop
– MapReduce был анонсирован разработчиками Google

2. Во сколько раз теоретически вырастет производительность при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум? (Введите число.)

– 2

3. Какие из следующих технологий СУБД не используют принцип MapReduce

– Hadoop
– Cassandra
– HDInsight
– Redis

4. Какие СУБД полностью полагаются на оперативную память при хранении информации:

– Oracle Exalytics
– SAP HANA
– BigTable
– HBase

5. В чём преимущество колоночно-ориентированных СУБД?

– они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
– они позволяют динамически дополнять содержание записей новыми полями
– они имеют более гибкие возможности аналитики
– они позволяют эффективно делать межколоночные сравнения

6. Для чего аналитику необходима "песочница"?

– для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций
– для хранения всех полученных от заказчика данных
– для построения отчётов о результатах анализа
– для снижения затрат, связанных с репликацией данных

7. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:

– Hadoop
– Data Warehouse
– "Песочница"
– Python

8. Выберите верное утверждение:

– Data Warehouse создаются для проверки гипотез при анализе больших данных
– "Песочница" используется для снижения нагрузки на основной Data Warehouse
– каждый Data Warehouse должен содержать "песочницу"
– "Песочница" необходима для любого процесса аналитики

9. Ниже приведена последовательность этапов проекта аналитики в соответствии с CRISP-DM, укажите первый этап.

– моделирование (Modeling)
– внедрение (Deployment)
– подготовка данных (Data Preparation)
– понимание бизнеса (Business understanding)
– оценка (Evaluation)
– понимание данных (Data Understanding)

10. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
– понимание бизнеса (Business understanding)
– понимание данных (Data Understanding)
– моделирование (Modeling)
– оценка (Evaluation)

11. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
– понимание бизнеса (Business understanding)
– подготовка данных (Data Preparation)
– моделирование (Modeling)
– оценка (Evaluation)

12. Пример благоразумного использования Hadoop
– анализ 10 Гб данных
– ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт)
– посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт)
– построение графика пульса пациента в реальном времени

13. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
– 100Гб
– 1Тб
– 100Тб
– 1Пб

14. Hadoop – это:
– набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах
– распределённая СУБД, позволяющая обрабатывать большие данные
– язык выполнения заданий в парадигме MapReduce
– распределённая файловая система, предназначенная для хранения файлов большого объёма

Отчет по лабораторной работе

Пример лабораторной работы 1. Базовый и расширенный Spark. Установка и тестирование Spark

Две искровые установки

1 Развертывание машины Linux

Здесь мы используем кластерный режим для установки, подготовки более двух серверов Linux и установки JDK1.8.

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
http://maven.apache.org/xsd/maven-4.0.0.xsd">
   <modelVersion>4.0.0</modelVersion>
```

```
    <groupId>com.atguigu</groupId>
    <artifactId>spark</artifactId>
    <version>1.0-SNAPSHOT</version>
    <modules>
        <module>sparkcore</module>
    </modules>

    <! - Указывает, что текущий проект является родительским проектом, нет конкретного кода, только объявленная общая информация->
    <packaging>pom</packaging>

        <! - Объявить открытые атрибуты->
    <properties>
        <spark.version>2.1.1</spark.version>
        <scala.version>2.11.8</scala.version>
        <log4j.version>1.2.17</log4j.version>
        <slf4j.version>1.7.22</slf4j.version>
    </properties>

        <! - Объявить и ввести публичные зависимости->
    <dependencies>
        <!-- Logging -->
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>jcl-over-slf4j</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>slf4j-api</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>slf4j-log4j12</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>log4j</groupId>
            <artifactId>log4j</artifactId>
            <version>${log4j.version}</version>
        </dependency>
        <!-- Logging End -->
        <dependency>
            <groupId>org.scala-lang</groupId>
            <artifactId>scala-library</artifactId>
            <version>${scala.version}</version>
            <!--<scope>provided</scope>-->
        </dependency>
    </dependencies>
```

```
    <! - Объявить только публичные зависимости->
<dependencyManagement>
   <dependencies>
      <!-- https://mvnrepository.com/artifact/org.apache.spark/spark-core -->
      <dependency>
         <groupId>org.apache.spark</groupId>
         <artifactId>spark-core_2.11</artifactId>
         <version>${spark.version}</version>
                  <! - Среда компиляции доступна, среда выполнения недоступна->
         <!--<scope>provided</scope>-->
      </dependency>

   </dependencies>
</dependencyManagement>

    <! - Информация о сборке конфигурации->
<build>

         <! - Объявить и представить встроенный плагин->
   <plugins>
            <! - Установить скомпилированную версию проекта->
      <plugin>
         <groupId>org.apache.maven.plugins</groupId>
         <artifactId>maven-compiler-plugin</artifactId>
         <version>3.6.1</version>
         <configuration>
            <source>1.8</source>
            <target>1.8</target>
         </configuration>
      </plugin>

            <! - Используется для компиляции кода Scala в класс->
      <plugin>
         <groupId>net.alchim31.maven</groupId>
         <artifactId>scala-maven-plugin</artifactId>
         <version>3.2.2</version>
         <executions>
            <execution>
               <goals>
                  <goal>compile</goal>
                  <goal>testCompile</goal>
               </goals>
            </execution>
         </executions>
      </plugin>

   </plugins>

         <! - Объявить только встроенные плагины->
   <pluginManagement>

      <plugins>
```

```
            <plugin>
              <groupId>org.apache.maven.plugins</groupId>
              <artifactId>maven-assembly-plugin</artifactId>
              <version>3.0.0</version>
              <executions>
                <execution>
                  <id>make-assembly</id>
                  <phase>package</phase>
                  <goals>
                    <goal>single</goal>
                  </goals>
                </execution>
              </executions>
            </plugin>
          </plugins>

      </pluginManagement>

    </build>

</project>
```

2 Установите JDK

1) Удалить существующий JDK

(1) Запросить, установлено ли программное обеспечение Java:

[root@hadoop101 opt]# rpm -qa|grep java

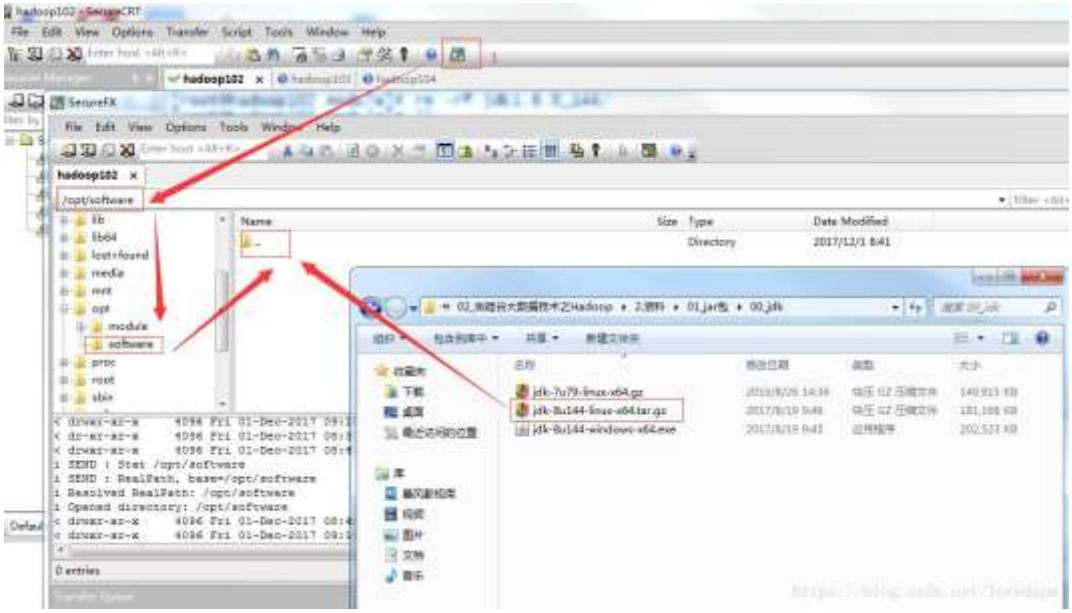
(2) Если установленная версия ниже 1.7, удалите jdk:

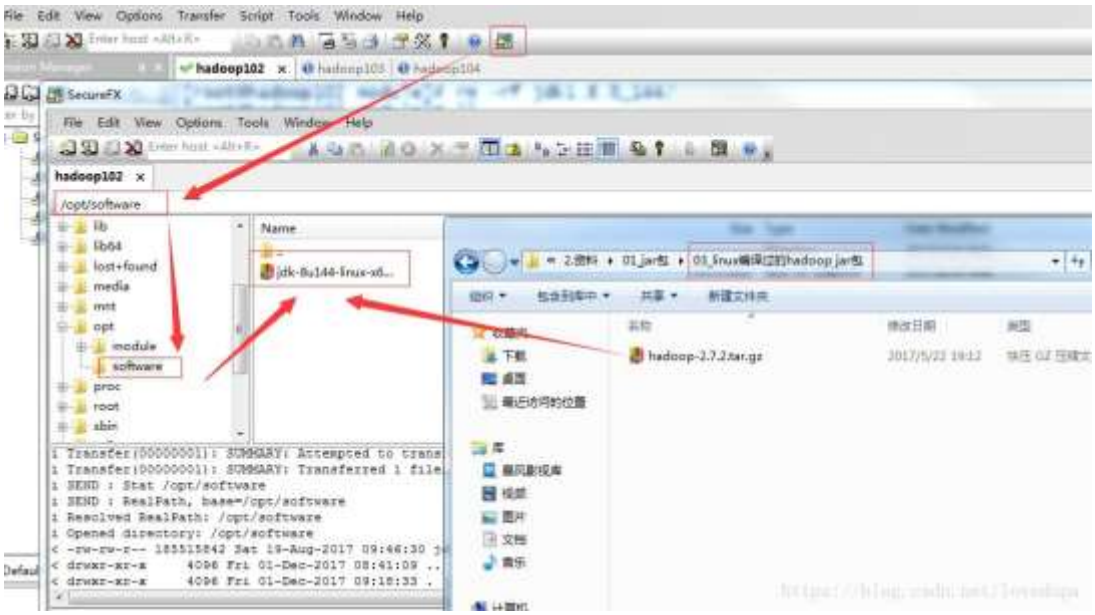
[root @ hadoop101 opt] # rpm -e package

(1) Скачать

Войдите на официальный сайт Oracle http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html, примите условия соглашения и выберите 64-битный пакет tar для Linux версии Ссылка, вы можете щелкнуть ссылку для прямой загрузки, рекомендуется использовать многопоточные инструменты загрузки (например, Thunder) для ускорения загрузки.

2) Используйте инструмент SecureCRT для импорта jdk и Hadoop-2.7.2.tar.gz в папку программного обеспечения в каталоге opt





3) Проверьте, успешно ли импортирован пакет программного обеспечения в каталоге opt в системе Linux.

[root@hadoop101opt]# cd software/

[root@hadoop101software]# ls

hadoop-2.7.2.tar.gz jdk-8u144-linux-x64.tar.gz

4) Извлеките jdk в каталог / opt / module

[root@hadoop101software]# tar -zxvf jdk-8u144-linux-x64.tar.gz -C /opt/module/

5) Настройте переменные среды jdk

(1) Сначала получите путь JDK:

[root@hadoop101 jdk1.8.0_144]# pwd

/opt/module/jdk1.8.0_144

(2) Откройте файл / etc / profile:

[root@hadoop101 jdk1.8.0_144]# vi /etc/profile

Добавьте путь jdk в конец файла profie:

##JAVA_HOME

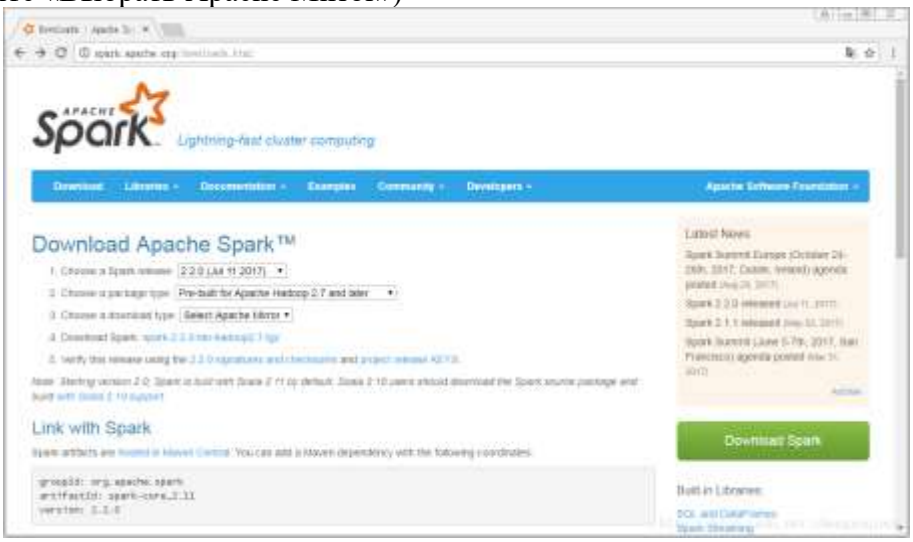
export JAVA_HOME=/opt/module/jdk1.8.0_144

export PATH=$PATH:$JAVA_HOME/bin

(3) Выход после сохранения:

:wq !
(4) Сделайте так, чтобы измененный файл вступил в силу:
[root@hadoop101 jdk1.8.0_144]# source /etc/profile
(5) Перезапустите (если доступна версия Java, не перезапускайте):
[root@hadoop101 jdk1.8.0_144]# sync
[root@hadoop101 jdk1.8.0_144]# reboot
6) Успешно протестировать установку jdk
[root@hadoop101 jdk1.8.0_144]# java -version
java version "1.8.0_144"

3 Загрузить установку Spark
(1) Войти на официальный сайт Spark
http://spark.apache.org/downloads.html
(2) Первый выбирает дистрибутив Spark (выберите 2.2.0), второй выбирает тип пакета (выбирает Hadoop 2.7), а третий выбирает тип загрузки (прямая загрузка) Медленнее выберите «Выбрать Apache Mirror»)



(3) Нажмите ссылку spark-2.2.0-bin-hadoop2.7.tgz и выберите домашнее зеркало

Другие способы тоже
Загрузите загруженный файл в каталог opt и распакуйте его в
(4) Затем распакуйте каталог / opt. Мы согласны с тем, что сторонние программные пакеты на платформе Linux размещаются в каталоге / opt.

```
[root@master ~]# tar -zxvf spark-2.2.0-bin-hadoop2.7.tgz -C /opt
```

5. Настройка Spark

Здесь мы используем автономный режим

Скопируйте slaves.template как slaves

Скопируйте spark-env.sh.template в spark-env.sh

Настройка Spark 【Автономный one

Режимами развертывания Spark являются Local, Standalone, Yarn, Mesos. Мы выбираем наиболее представительный автономный режим.

Измените файл spark-env.sh и добавьте следующую конфигурацию:

Переименуйте и измените файл slaves.template

mv slaves.template slaves

vi slaves

Добавьте местоположение дочернего узла в файл (рабочий узел)

linux102

linux103

linux104

Сохранить и выйти

Распространение сконфигурированного файла Spark на другие узлы

Кластер Spark настроен. В настоящее время он является главным, двумя рабочими, и искровой кластер запущен на master01.

/home/bigdata/hadoop/spark-2.1.1-bin-hadoop2.7/sbin/start-all.sh

После запуска выполните команду jps. На главном узле есть главный процесс, а на других дочерних узлах - работа и работа. Войдите в интерфейс управления Spark, чтобы просмотреть состояние кластера (главный узел): http: // master01: 8080 /

Примечание. Если вы столкнулись с исключением «JAVA_HOME not set», вы можете добавить следующую конфигурацию в файл spark-config.sh в каталоге sbin:

export JAVA_HOME=XXXX

6. Тестирование на ОК

После запуска выполните команду jps. Главный процесс находится на главном узле, а работа выполняется на других дочерних узлах. Войдите в интерфейс управления Spark, чтобы проверить состояние кластера (главный узел):http://linux102:8080/

Пример по лабораторной работе 2. Применение Apache Spark для считывания, обработки и записи данных

1 Выполнить первое искровое приложение

```
/opt/module/sparktest/bin/spark-submit  \
--class org.apache.spark.examples.SparkPi \
--master spark://linux102:7077 \
--executor-memory 1G \
--total-executor-cores 2 \
/opt/module/sparktest/examples/jars/spark-examples_2.11-2.1.1.jar \
```

100
Этот алгоритм использует алгоритм Монте-Карло, чтобы найти PI

2 Spark-shell
/opt/module/sparktest/bin/spark-submit \
--class org.apache.spark.examples.SparkPi \
--master spark://node01:7077 \
--executor-memory 1G \
--total-executor-cores 2 \
/usr/local/spark-1.6.1-bin-hadoop2.6/lib/spark-examples-1.6.1-hadoop2.6.0.jar \
100
Этот алгоритм использует алгоритм Монте-Карло, чтобы найти PI
Описание параметра:
--master spark: // master01: 7077 указывает адрес мастера
--executor-memory 1G Укажите 1G доступной памяти для каждого исполнителя
--total-executor-cores 2 Указать количество чашечных ядер, используемых каждым исполнителем, равным 2

Написание WordCount в оболочке Spark

1 запуск sparkshell

Spark-shell - это интерактивная программа Shell, поставляемая со Spark, которая удобна для пользователей, чтобы выполнять интерактивное программирование.Пользователи могут писать программы spark, используя scala под этой командной строкой.
/home/bigdata/hadoop/spark-2.1.1-bin-hadoop2.7/bin/spark-shell \
--master spark://master01:7077 \
--executor-memory 2g \
--total-executor-cores 2

Примечание:

Если вы не укажете главный адрес при запуске оболочки spark, вы также можете запустить оболочку spark и нормально запускать программы в оболочке spark. Фактически запускается локальный режим spark. Этот режим только запускает процесс на локальной машине без установления соединения с кластером.

Класс SparkContext был инициализирован для объекта sc по умолчанию в Spark Shell. Если нужен код пользователя, напрямую примените sc

2 Написание wordCount в спарк-оболочке

1. Сначала запустите hdfs
2. Загрузите файл в hdfs на hdfs: // node01: 9000 / words.txt
3. Напишите программу spark на языке scala в оболочке spark
sc.textFile("hdfs://node01:9000/words.txt").flatMap(_.split(" "))
.map((_,1)).reduceByKey(_+_).saveAsTextFile("hdfs://node01:9000/out")
4. Используйте команду hdfs для просмотра результатов
hdfs dfs -ls hdfs://node01:9000/out/p*
Описание:
sc - это объект SparkContext, который является точкой входа для отправки программы spark
textFile (hdfs: // node01: 9000 / words.txt) считывает данные из hdfs
сначала отображается flatMap (_. split (""), а затем выравнивается
Карта ((_, 1)) образует кортеж слов и 1
lowerByKey (_ + _) уменьшает на ключ и накапливает значения
saveAsTextFile ("hdfs: // node01: 9000 / out") записывает результат в hdfs

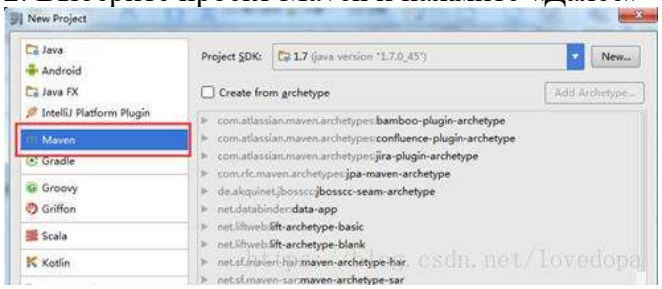
3 Написание программ Spark в идее

Оболочка spark используется только при тестировании и проверке наших программ. В производственной среде программы обычно компилируются в среде IDE, затем упаковываются в файлы jar и затем передаются в кластер. Наиболее часто используемым является создание проекта Maven и его использование. Maven управляет зависимостями пакета jar.

Оболочка spark используется только при тестировании и проверке наших программ. В производственной среде программы обычно компилируются в среде IDE, затем упаковываются в файлы jar и затем передаются в кластер. Наиболее часто используемым является создание проекта Maven и его использование. Maven управляет зависимостями пакета jar.

1. создать проект

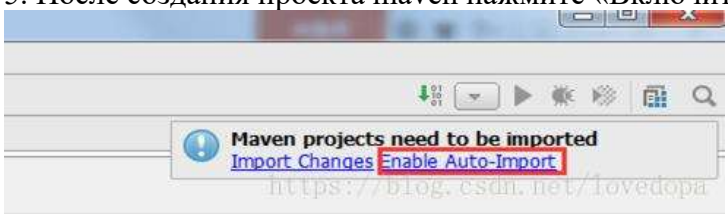
2. Выберите проект Maven и нажмите «Далее»



3. Заполните GAV Maven и нажмите кнопку Далее

4. Введите название проекта и нажмите «Готово»

5. После создания проекта maven нажмите «Включить автоматический импорт».



6. Настройте Maven's pom.xml

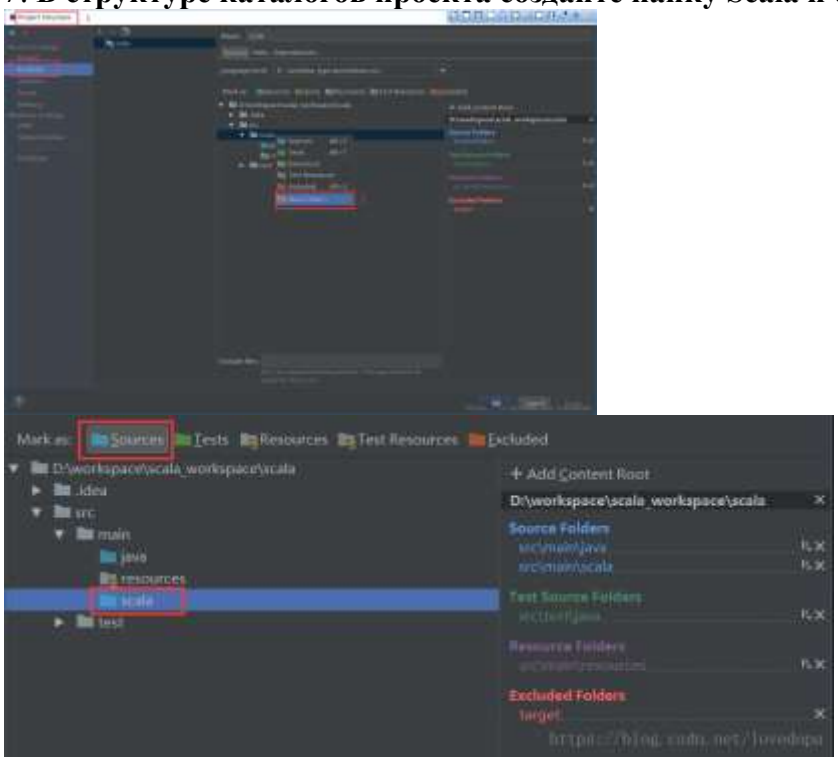
```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<project xmlns="http://maven.apache.org/POM/4.0.0"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
http://maven.apache.org/xsd/maven-4.0.0.xsd">
    <modelVersion>4.0.0</modelVersion>

    <groupId>com.atguigu</groupId>
    <artifactId>spark</artifactId>
    <version>1.0-SNAPSHOT</version>
    <modules>
        <module>sparkcore</module>
    </modules>

        <! - Указывает, что текущий проект является родительским проектом, нет конкретного
кода, только объявленная общая информация->
    <packaging>pom</packaging>

        <! - Объявить открытые атрибуты->
    <properties>
        <spark.version>2.1.1</spark.version>
        <scala.version>2.11.8</scala.version>
        <log4j.version>1.2.17</log4j.version>
        <slf4j.version>1.7.22</slf4j.version>
    </properties>

        <! - Объявить и ввести публичные зависимости->
    <dependencies>
        <!-- Logging -->
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>jcl-over-slf4j</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>slf4j-api</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>org.slf4j</groupId>
            <artifactId>slf4j-log4j12</artifactId>
            <version>${slf4j.version}</version>
        </dependency>
        <dependency>
            <groupId>log4j</groupId>
            <artifactId>log4j</artifactId>
            <version>${log4j.version}</version>
        </dependency>
        <!-- Logging End -->
        <dependency>
            <groupId>org.scala-lang</groupId>
            <artifactId>scala-library</artifactId>
```

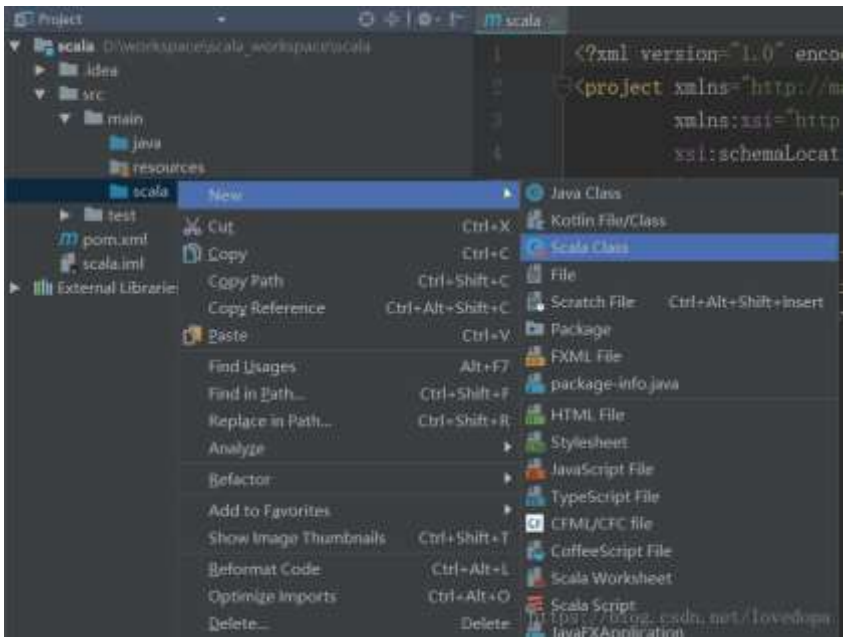
```
      <version>${scala.version}</version>
      <!--<scope>provided</scope>-->
    </dependency>
</dependencies>

    <! - Объявить только публичные зависимости->
<dependencyManagement>
    <dependencies>
      <!-- https://mvnrepository.com/artifact/org.apache.spark/spark-core -->
      <dependency>
        <groupId>org.apache.spark</groupId>
        <artifactId>spark-core_2.11</artifactId>
        <version>${spark.version}</version>
                    <! - Среда компиляции доступна, среда выполнения недоступна->
        <!--<scope>provided</scope>-->
      </dependency>

    </dependencies>
</dependencyManagement>

    <! - Информация о сборке конфигурации->
<build>

          <! - Объявить и представить встроенный плагин->
    <plugins>
                <! - Установить скомпилированную версию проекта->
      <plugin>
        <groupId>org.apache.maven.plugins</groupId>
        <artifactId>maven-compiler-plugin</artifactId>
        <version>3.6.1</version>
        <configuration>
          <source>1.8</source>
          <target>1.8</target>
        </configuration>
      </plugin>

                <! - Используется для компиляции кода Scala в класс->
      <plugin>
        <groupId>net.alchim31.maven</groupId>
        <artifactId>scala-maven-plugin</artifactId>
        <version>3.2.2</version>
        <executions>
          <execution>
            <goals>
              <goal>compile</goal>
              <goal>testCompile</goal>
            </goals>
          </execution>
        </executions>
      </plugin>

    </plugins>
```

```
            <! - Объявить только встроенные плагины->
        <pluginManagement>

            <plugins>
                <plugin>
                    <groupId>org.apache.maven.plugins</groupId>
                    <artifactId>maven-assembly-plugin</artifactId>
                    <version>3.0.0</version>
                    <executions>
                        <execution>
                            <id>make-assembly</id>
                            <phase>package</phase>
                            <goals>
                                <goal>single</goal>
                            </goals>
                        </execution>
                    </executions>
                </plugin>
            </plugins>

        </pluginManagement>

    </build>

</project>
```

7. **В структуре каталогов проекта создайте папку Scala и отметьте ее как исходную.**



8.**После завершения вышеуказанной конфигурации вы можете Щелкните правой кнопкой мыши, чтобы создать класс Scala**



9. Напиши искровые программы

```
package com.qf.spark
import org.apache.spark.{SparkContext, SparkConf}
object WordCount {
  def main(args: Array[String]) {
// Создаем SparkConf () и устанавливаем имя приложения
    val conf = new SparkConf().setAppName("WC")
// Создать SparkContext, этот объект является входом для отправки Spark App
    val sc = new SparkContext(conf)
// Создать RDD с помощью sc и выполнить соответствующее преобразование и действие
        sc.textFile(args(0)).flatMap(_.split(" ")).map((_, 1)).reduceByKey(_+_, 1).sortBy(_._2, false).saveAsTextFile(args(1))
// Остановить sc и завершить задачу
    sc.stop()
  }
}
```

10. Упаковка с Maven: сначала измените основной класс в pom.xml

Нажмите на опцию Maven Project справа от идеи

Нажмите «Жизненный цикл», выберите «Очистить и упаковать», затем нажмите «Запустить Maven Build».

```
<transformers>
    <transformer implementation="org.apache.maven.plugins.shade.resource.AppendingTransforme
        <resource>reference.conf</resource>
    </transformer>
    <transformer implementation="org.apache.maven.plugins.shade.resource.ManifestResourceTra
        <mainClass>com.qf.spark.WordCount</mainClass>
    </transformer>
</transformers>
```


11. Выберите успешно скомпилированный пакет jar и загрузите jar на узел в кластере Spark.

12. Сначала запустите кластер HDFS и Spark.
запустите hdfs
/usr/local/hadoop-2.6.1/sbin/start-dfs.sh
начать искру
/usr/local/spark-1.6.1-bin-hadoop2.6/sbin/start-all.sh

13. Отправьте приложение Spark с помощью команды spark-submit (обратите внимание на порядок параметров)

```
 /opt/module/sparktest/bin/spark-submit  \
--class com.itstar.spark.WordCount\
--master spark://master01:7077\
--executor-memory 1G \
--total-executor-cores 2 \
wordcount-jar-with-dependencies.jar\
hdfs://master01:9000/RELEASE\
hdfs://master01:9000/out
```

Посмотреть результаты выполнения программы
Используйте команду hdfs для просмотра результатов
hdfs dfs -cat hdfs://linux102:9001/hbase/*

Лабораторная работа 3. Применение data frame API Apache Spark

Лабораторная работа 4. Применение rdd API Apache Spark

Лабораторная работа 5. Применение Apache Spark для решения задачи преобразования данных

Критерии оценки выполнения задания

Оценка	Критерии оценивания
Неудовлетворительно	Работа выполнена не полностью и объем выполненной части работы не позволяет сделать правильных выводов
Удовлетворительно	Работа выполнена не полностью, но не менее 50% объема, что позволяет получить правильные результаты и выводы; в ходе проведения работы были допущены ошибки
Хорошо	Работа выполнена в полном объеме с соблюдением необходимой последовательности действий, но допущена одна ошибка или не более двух недочетов и обучающийся может их исправить самостоятельно или с небольшой помощью преподавателя
Отлично	Работа выполнена в полном объеме без ошибок с соблюдением необходимой последовательности действий

Отчет оформляется каждым студентом индивидуально и должен содержать:

номер и название работы, цель работы, дату выполнения, краткое описание теории изучаемого вопроса, основные характеристики измерительных приборов, записи результатов прямых измерений и расчетов косвенных измерений, оформленные в виде таблицы, графики полученных зависимостей (если требуются), запись вычислений, приводящих к окончательному результату, расчет ошибок измерений и окончательный результат с указанием ошибки измерения, скриншоты выполнения в ПП, исходные файлы выпаленного задания в ПП, выводы.

К каждой лабораторной работе содержится перечень вопросов для защиты лабораторной работы.

Вопросы для проведения промежуточной аттестации по итогам освоения дисциплины

Раздел 1. Распределенные данные, распределенная обработка и современные требования к скорости и надежности вычислений

1. Что такое распределенные данные и распределенная обработка?
2. Какие существуют типы распределенных систем обработки данных и в чем их основные отличия?
3. Каковы основные преимущества и недостатки распределенной обработки данных?
4. Как распределенная обработка помогает улучшить скорость и надежность вычислений?
5. Какие технологии и протоколы используются для реализации распределенной обработки?
6. Что такое “горизонтальная масштабируемость” и как она связана с распределенными данными?
7. Какие требования предъявляются к современным системам распределенной обработки с точки зрения производительности и надежности?
8. Как решаются проблемы безопасности, связанные с обработкой распределенных данных?
9. Какие алгоритмы и структуры данных лучше всего подходят для распределенной обработки и почему?
10. Какие примеры успешных проектов или компаний, использующих распределенные системы обработки данных, вы можете привести?

Раздел 2. Spark. Основные парадигмы. Стадии обработки. Типы операций. API взаимодействия

1. Что из себя представляет платформа Spark?
2. Каковы основные парадигмы обработки данных в Spark?
3. Какие стадии обработки данных включает в себя Spark?
4. Какие типы операций доступны в Spark для обработки данных?
5. Опишите API взаимодействия с Spark на примере языка программирования Python.
6. В чем преимущество использования Spark перед другими системами обработки данных?
7. Как в Spark реализована поддержка потоковой обработки данных?
8. Что такое RDD в Spark и как они используются для обработки данных?
9. Как в Spark реализовано параллельное выполнение задач?
10. В чем разница между DataFrames и RDDs в Spark с точки зрения обработки данных?

Темы рефератов:

1. Spark как платформа для обработки больших данных.
2. Парадигмы обработки данных в Spark.
3. Стадии обработки данных в Spark.
4. Типы операций в Spark для обработки данных.
5. API взаимодействия с Spark на примере Python.
6. Преимущества использования Spark перед другими системами.
7. Потоковая обработка данных в Spark.
8. RDD в Spark и их использование.
9. Параллельное выполнение задач в Spark.
10. Сравнение DataFrames и RDDs в Spark.
11. Распределённые данные и их обработка.
12. Современные требования к распределённой обработке данных.
13. Технологии распределённой обработки данных.
14. Горизонтальная масштабируемость в распределённой обработке.
15. Проблемы безопасности в распределённой обработке данных.
16. Алгоритмы и структуры данных для распределённой обработки.
17. Проекты и компании, использующие распределённую обработку данных.

Для текущего контроля ТК2:
Проверяемая компетенция: ПК-6

Тест

1. Михаил получает на электронную почту в среднем 1000 писем в месяц, из них 2,44% - это спам. Известно, что среди спама слово "знакомство" встречается в 0,01% писем, а среди обычных писем в 10 раз реже. Какова вероятность того, что письмо, попавшее на почтовый ящик Михаила, в тексте которого встречается указанное слово, не является спамом? (Ответ укажите в целых процентах без знака процента.)

80

2. Выберите оптимальный параметр для следующей модели согласно принципу ML (Maximum Likelihood / Максимальное правдоподобие): "Вероятность того что идет дождь если есть тучи сильнее, чем вероятность того что идет дождь, если туч нет":
   - Падают капли
   - Наличие туч
   - Не видно небо
   - Мокрая земля

3. Дома на четной стороне улицы имеют номера 2, 4, 6, … . Номер дома – это признак:
   - Бинарный
   - Номинальный
   - Порядковый
   - Количественный
   - Нет правильного ответа

4. Недостаток алгоритма Expectation Maximization (EM) заключается в следующем:
   - На каждом из шагов возможно, как возрастание, так и убывание likelihood (вероятности)

- Невозможно оптимизировать аналитически
- Не гарантируется глобальная оптимизация
- В ряде случаев достигнуть экстремум невозможно

5. Какому этапу CRISP-DM соответствует Exploratory data analysis:
   - Business Understanding
   - Data Understanding
   - Data Preparation
   - Modeling
   - Evaluation
   - Deployment

6. На практике, более какого числа раз имеет смысл проводить запуск алгоритма K-means ?

1000

Проверяемая компетенция: ПК-7

Тест

1. Выберите лишний этап методологии CRISP-DM:
   - Понимание бизнес-целей
   - Понимание данных
   - Подготовка данных
   - Обучение модели
   - Моделирование
   - Оценка
   - Внедрение

2. Пример задачи эффекта "проклятие размерности". Даны два случайных вектора x и y в пространстве размерности D. Как зависит математическое ожидание косинус-расстояния между x и y от размерности D, при наблюдениях, что числитель стремится к нулю, а знаменатель положительный ? Ответ укажите с точность до 2-го знака после запятой.

1,57

3. Какие проблемы решают задачи кластеризации, отыскивая "скрытую структуру" исследуемых данных и не имея опорной целевой переменной?
   - разметка данных "в ручную" очень дорого и трудозатратно
   - построение признаков из очень большего количества данных
   - возможность отслеживать эволюционные изменения
   - поиск выбросов и шумов в исследуемых данных
   - исследование и визуализация больших данных

4. Что из перечисленного является средством EDA?
   - Histogram
   - Scatter plot
   - Visual estimation
   - Piechart

5. Подходы к построению моделей Data Mining
   - статистический и на основании машинного обучения
   - на основании машинного обучения и вычислительный
   - вычислительный и статистический
   - все перечисленное

6. Для преобразования многомерного пространства в пространство низшей размерности и формирования малого количества признаков из большого количества признаков следует использовать следующий алгоритм:
   - T-SNE
   - DBSCAN
   - PAM
   - CWM

7. На диаграмме процесса CRISP-DM после этапа Моделирование (Modeling) следует этап
   - Понимание бизнес-целей (Business Understanding
   - Подготовка данных (Data Preparation)
   - Оценка (Evaluation)
   - Внедрение (Deployment)

8. Какие характерные активности этапа подготовки данных (Data Preparation) для процесса CRISP-DM
   - Удаление шума
   - Заполнение отсутсвующих значений
   - Понять чем характеризуется задача
   - Какого результата нужно достичь

Отчет по лабораторным работам

Лабораторная работа 6. Применение Apache Spark для решения задачи анализа данных

Лабораторная работа 7. Применение Apache Spark для решения задачи машинного обучения

Лабораторная работа 8. Установка и развертывание Apache Kafka

Вопросы для проведения промежуточной аттестации по итогам освоения дисциплины

Раздел 3. Spark. Работа с dataframe

1. Что такое Spark и для чего он используется?
2. Каковы основные компоненты Spark?
3. Как работает механизм обработки данных в Spark?
4. Что такое DataFrame в Spark? Как он отличается от других типов данных?
5. Какие операции можно выполнять с DataFrames?
6. Что означает распределение данных по разным узлам кластера? Как это влияет на производительность вычислений?
7. В чем разница между локальным и распределенным режимами работы с данными в Spark?

8. Какие функции используются для работы с датафреймами в Spark? Приведите примеры.
9. Как можно преобразовать один тип данных в другой в Spark?
10. Как осуществляется чтение и запись данных из различных источников в Spark?
11. Как в Spark работают с разными типами данных, такими как строки, числа, даты и т.д.?
12. Какие параметры конфигурации доступны для настройки работы с данными в Spark и как они влияют на производительность?
13. Какие методы используются для обработки ошибок и исключений в Spark?
14. Как происходит параллельная обработка данных в Spark с использованием нескольких ядер и процессоров?

Раздел 4. Работа с данными типа "ключ — значение" (rdd)

1. Что такое RDD в Spark? Каковы его основные характеристики и отличия от других структур данных?
2. Как происходит работа с RDD? Опишите основные этапы работы с ними.
3. Какие операции можно проводить над RDD? Приведите несколько примеров.
4. Какие функции используются для создания и работы с RDD в Spark? Опишите их назначение и использование.
5. Как производится чтение и запись RDD из разных источников данных?
6. Как в Spark происходит работа с различными типами данных в RDD (строки, числа, дата и т. д.)?
7. Какие параметры конфигурации используются для настройки работы RDD в Spark и какое влияние они оказывают на производительность системы?

   8 .Как происходит обработка ошибок и исключительных ситуаций в RDD?
8. Как осуществляется параллельная обработка RDD с использованием нескольких процессоров и ядер?
9. Как Spark обеспечивает масштабируемость и отказоустойчивость при работе с RDD ?
10. Какие существуют способы оптимизации производительности при работе с RDD и какие методы используются для их реализации?
11. Как осуществляется взаимодействие между различными RDD в рамках одного задания или разных заданий в Spark?

Раздел 5. Spark. Обработка данных и ML

1. Какие этапы обработки данных включает в себя Spark ML?
2. Какие типы данных могут быть обработаны с помощью Spark ML?
3. Какие алгоритмы машинного обучения поддерживаются Spark ML?
4. Как выбрать подходящий алгоритм машинного обучения для конкретной задачи?
5. Какие метрики используются для оценки качества модели машинного

обучения?
6. Какие этапы включает в себя процесс построения модели машинного обучения с помощью Spark?
7. Как выбирать параметры для алгоритмов машинного обучения в Spark ML?
8. Какие виды нормализации данных поддерживаются в Spark ML?
9. Какие методы оценки качества моделей машинного обучения реализованы в Spark MLlib?
10. Какие подходы к регуляризации моделей машинного обучения поддерживаются в Spark ML?
11. Как осуществлять выбор модели машинного обучения на основе кросс-валидации?
12. Какие методы для оценки производительности моделей машинного обучения существуют в Spark ML и как их использовать?
13. Как оценивать качество модели машинного обучения и интерпретировать результаты?
14. Какие возможности предоставляет Spark для анализа данных и визуализации результатов?
15. Как оптимизировать модели машинного обучения в Spark?

Темы рефератов:

1. Spark как платформа для обработки и анализа данных: архитектура, компоненты и возможности.
2. Machine Learning в Spark: основные алгоритмы, их особенности и примеры использования.
3. Применение Spark для предварительной обработки данных: основные этапы и инструменты.
4. Spark Streaming: технология обработки данных в реальном времени, ее особенности и применение.
5. Использование Spark для работы с большими данными: масштабируемость, производительность и надежность.
6. Применение Spark в области анализа данных: примеры проектов и кейсы.
7. Spark и Apache Hadoop: сравнение архитектур и возможностей.
8. Spark для работы с данными типа ключ-значение: особенности и преимущества.
9. Spark GraphX: обработка графовых данных, алгоритмы и примеры применения.
10. Spark SQL: возможности и преимущества для работы с SQL запросами.
11. Использование Spark в задачах машинного обучения: подходы, инструменты и метрики качества.
12. Spark MLib: библиотека алгоритмов машинного обучения, ее возможности и примеры использования.
13. Spark для обработки естественных языков (NLP): возможности и применение.
14. Spark в биоинформатике: анализ геномных данных, их обработка и визуализация.

Для текущего контроля ТК3:

Тест

Вопрос 1: Что такое очередь в контексте систем обмена сообщениями?

Ответ 1: Очередь - это механизм, который позволяет сообщениям ожидать обработки, когда обработчик занят или недоступен.

Вопрос 2: Что такое брокер сообщений?

Ответ 2: Брокер сообщений - это программное обеспечение, которое обеспечивает обмен сообщениями между отправителем и получателем.

Вопрос 3: В чем разница между очередями и брокерами сообщений? Ответ 3: Очереди используются для временного хранения сообщений, пока обработчик не станет доступным, в то время как брокеры сообщений обеспечивают взаимодействие между отправителями и получателями.

Вопрос 4: Какие типы очередей вы знаете?

Ответ 4: Очереди могут быть разных типов, например, очереди сообщений, очереди задач, очереди обработки и т.д.

Вопрос 5: Что такое “очередь сообщений”?

Ответ 5: Очередь сообщений - это система, которая хранит сообщения до тех пор, пока они не будут обработаны.

Правильные ответы:

1 2 3 4 5

Отчеты по лабораторным работам

Лабораторная работа 9. Создания простейшего сервиса Apache Kafka, который «слушает» источник и передает данные на Apache Spark job

Лабораторная работа 10. Установка и развертывание Apache Flink

Лабораторная работа 11. Создание простейшего сервиса Apache Flink поставляющего данные на основе состояний

Вопросы для проведения промежуточной аттестации по итогам освоения дисциплины

Раздел 6. Очереди и брокеры

1. Что такое очереди и брокеры в контексте программирования?
2. В каких случаях используются очереди и брокеры?
3. Чем отличаются очереди от брокеров?
4. Какие основные типы очередей вы знаете?
5. Как работает система с очередями и брокерами?
6. Какие проблемы могут возникнуть при использовании очередей и как их избежать?
7. Что такое брокер сообщений и каковы его основные функции?
8. Как выбрать подходящий брокер сообщений для вашего проекта?
9. Какие существуют типы брокеров сообщений?

10. Что такое “очередь сообщений” и зачем она нужна?
11. Каковы основные преимущества использования очередей и брокеров в программном обеспечении?
12. Какие виды очередей используют наибольшее количество разработчиков?
13. Какие наиболее популярные брокеры сообщений вы знаете?
14. В чем разница между очередями в памяти и дисковыми очередями?
15. Что такое многопоточность и как она связана с очередями?
16. Что значит “очередь ограничена” и “очередь без ограничений”?
17. Какие функции обычно предоставляются библиотеками для работы с очередями сообщений?

Раздел 7. Потоковая обработка с учетом состояний и основы потоковой обработки

1. Потоковая обработка данных: определение и основные характеристики.
2. Основы потоковой обработки: понятие состояний, событий и переходов.
3. Задачи, решаемые с помощью потоковой обработки данных.
4. Примеры приложений, использующих потоковую обработку данных.
5. Отличие потоковой обработки от традиционной обработки данных.
6. Основные архитектуры систем потоковой обработки.
7. Понятие состояния в потоковой обработке данных: определение, виды и применение.
8. Состояния в потоковой обработке: конечные и бесконечные, дискретные и непрерывные состояния.
9. События в потоковой обработке данных: определения, виды и использование.
10. Переходы между состояниями в потоковой обработке данных.
11. Моделирование систем с помощью конечных автоматов.
12. Задачи, связанные с обработкой потоков данных: классификация и примеры.
13. Алгоритмы и методы обработки потоков данных.
14. Основы теории графов и сетей для потоковой обработки данных.
15. Методы оптимизации и улучшения производительности систем потоковой обработки данных.

Темы рефератов:

1. Введение в потоковую обработку с учетом состояний.
2. Основы и принципы потоковой обработки с учетом состояний.
3. Применение потоковой обработки с учетом состояний в реальных приложениях.
4. Архитектура систем потоковой обработки с учетом состояний.
5. Конечные автоматы и их использование в потоковой обработке с учетом состояний.
6. Обработка потоков событий в системах с учетом состояний.
7. Методы моделирования систем с использованием конечных автоматов и потоковой обработки с учетом состояний.

8. Алгоритмы обработки потоков данных в системах с учетом состояний.
9. Применение теории графов в задачах потоковой обработки с учетом состояний.
10. Методы оптимизации систем потоковой обработки с учетом состояний.

Для промежуточной аттестации ОМ:
Вопросы к экзамену

1. Что Такое Spark?
2. Каковы ключевые особенности Spark?
3. Что такое SCC?
4. Что такое RDD?
5. Что такое неизменность (Immutability)?
6. Что такое YARN?
7. Какие самые распространённые языки программирования в Spark?
8. Сколько менеджеров кластера доступны в Spark?
9. Каковы обязанности движка Spark?
10. Что такое ленивые вычисления’?
11. Что такое раздел (Partition)?
12. Для чего нужен Spark Streaming?
13. Нормально ли запускать все ваши процессы на локализованном ноде?
14. Для чего используется SparkCore?
15. Имеет ли применение File System API в Spark?
16. Чем MapReduce отличается от Spark?
17. Что вы понимаете под трансформациями в Spark?
18. Что такое Apache Kafka?
19. Как запустить сервер в Kafka?
20. Что такое традиционные методы передачи данных и чем Kafka лучше?
21. Что такое zookeeper в Kafka и можем ли мы использовать эту программу без него?
22. Почему Kafka является такой важной частью технологии?
23. Объясните, что такое последователь и лидер в Kafka.
24. Что такое потребители и пользователи в Kafka?
25. Как вы используете Kafka в качестве системы хранения данных?
26. Объясните максимальный размер сообщения, которое может принять Kafka.
27. Как разбалансировать кластер в Kafka?