Государственное образовательное учреждение высшего образования
«КОМИ РЕСПУБЛИКАНСКАЯ АКАДЕМИЯ ГОСУДАРСТВЕННОЙ СЛУЖБЫ И УПРАВЛЕНИЯ»
(ГОУ ВО КРАГСиУ)

«КАНМУ СЛУЖБАӦ ДА ВЕСЬКӦДЛЫНЫ ВЕЛӦДАН КОМИ РЕСПУБЛИКАСА АКАДЕМИЯ»
вылыс тшупӧда велӧдан канму учреждение
(КСдаВВКРА ВТШВ КУ)

Утверждена в структуре
ОПОП 46.03.02 Документоведение и архивоведение
(решение Ученого совета
от 17.06.2022 № 12)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«ОСНОВЫ РАБОТЫ С Big Data»

Направление подготовки – *46.03.02 Документоведение и архивоведение;* направленность (профиль) – *«Организационное и документационное обеспечение управления»*

Уровень высшего образования – *бакалавриат*

Форма обучения – *очная, заочная*

Год начала подготовки – 2022

Сыктывкар
2022

Рабочая программа дисциплины «Основы работы с Big Data» составлена в соответствии с требованиями:

- Федерального государственного образовательного стандарта высшего образования по направлению подготовки 46.03.02 Документоведение и архивоведение (*уровень бакалавриата*), утвержденного приказом Минобрнауки России от 29.10.2020 г. № 1343;
- Приказа Минобрнауки России «Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры» от 06.04.2021 № 245;
- учебного плана ГОУ ВО «Коми республиканская академия государственной службы и управления» по направлению 46.03.02 Документоведение и архивоведение (бакалавриата) направленность (профиль) «Организационное и документационное обеспечение управления».

РАЗДЕЛ 1. ОСНОВНЫЕ СВЕДЕНИЯ

1. Цель и задачи учебной дисциплины

1.1. Цель изучения учебной дисциплины

Целью освоения дисциплины «Основы работы с Big Data» формирование у обучающихся компетенций в области использования систем обработки и анализа больших массивов данных.

1.2. Задачи учебной дисциплины

Задачами освоения дисциплины «Основы работы с Big Data» являются:

– изучение теории, базовых понятий Big Data;

– изучение технологий подготовки, хранения, обработки и анализа больших данных;

– освоение статистических методов для анализа больших объемов информации;

– приобретение базовых умений работы с большими данными.

1.3. Виды компетенций, формируемые в результате освоения дисциплины

Изучение дисциплины «Основы работы с Big Data» направлено на формирование следующих:

1) универсальные:

− УК-1 – способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач.

2) общепрофессиональные:

− ОПК-4. Способен использовать базовые знания в области информационно-коммуникационных технологий в сфере своей профессиональной деятельности

ОПК-5. Способен самостоятельно работать с различными источниками информации и применять основы информационно-аналитической деятельности при решении профессиональных задач

1.4. Место дисциплины в структуре образовательной программы:

Дисциплина «Основы работы с Big Data» относится к *части, формируемой участниками образовательных отношений* Блока 1 «Дисциплины (модули)» образовательной программы.

2. Требования к результатам освоения учебной дисциплины

2.1. Изучение дисциплины «Основы работы с Big Data» направлено на формирование следующих компетенций и индикаторов их достижений, заявленных в образовательной программе:

1) универсальные:

Наименование категории (группы) компетенций	**Формируемые компетенции** (код, наименование компетенции)	**Код и наименование индикатора достижений компетенций**	**Содержание индикатора достижений компетенций**
Системное и критическое мышление	УК-1. Способен осуществлять поиск, критический анализ и синтез информации, применять системный	УК-1.И-1. Осуществляет поиск необходимой информации,	УК-1.И-1.У-2. Умеет осуществлять поиск необходимой для решения поставленной задачи информации,

	подход для решения поставленных задач	опираясь на результаты анализа поставленной задачи	критически оценивая надежность различных источников информации

2) общепрофессиональные:

Формируемые компетенции (код, наименование компетенции)	**Код и наименование индикатора достижений компетенций**	**Содержание индикатора достижений компетенций**
ОПК-4. Способен использовать базовые знания в области информационно-коммуникационных технологий в сфере своей профессиональной деятельности	ОПК-4. И-1 использует базовые знания в области информационно-коммуникационных технологий в сфере своей профессиональной деятельности	ОПК-4.И-1.З-1. Имеет базовые знания в области информационных технологий (программные продукты, используемые в управлении документами) ОПК-4.И-1.У-1. Умеет пользоваться базовыми знаниями в области информационных технологий (программные продукты, используемые в управлении документами)
ОПК-5. Способен самостоятельно работать с различными источниками информации и применять основы информационно-аналитической деятельности при решении профессиональных задач	ОПК-5. И-1 самостоятельно работает с различными источниками информации	ОПК-5.И-1.З-1. Знает методы самостоятельной работы с различными источниками информации ОПК-5.И-1.У-1. Умеет самостоятельно работать с различными источниками информации
	ОПК-5. И-2 Применяет основы информационно-аналитической деятельности при решении профессиональных задач	ОПК-5.И-2.З-1 Знает методы самостоятельной работы с различными источниками информации ОПК-5.И-2.У-1.Умеет пользоваться основами информационно-аналитической деятельности и способностью применять их в профессиональной сфере

2.2. Запланированные результаты обучения по дисциплине «Основы работы с Big Data»:

Должен знать:

– принципы работы современных информационных технологий, соответствующих содержанию профессиональных задач на основе методов и принципов хранения, выборки и обработки больших данных.

Должен уметь:

– осуществлять поиск необходимой информации, хранящейся в структуре больших данных, опираясь на результаты анализа поставленной задачи;

− разрабатывать варианты решения проблемной ситуации на основе критического анализа доступных источников информации в структуре больших данных;

− использовать основные методы, средства получения, представления, хранения и обработки статистических данных с использованием методов и технологий больших данных;

− применять статистические методы обработки собранных данных, использовать анализ данных, необходимых для решения поставленных задач на основе больших данных;

3. Объём учебной дисциплины

Очная форма обучения

Виды учебной работы	Распределение учебного времени
Контактная работа	***56,35***
Аудиторные занятия (всего):	54
Лекции	18
Практические занятия	36
Лабораторные занятия	-
Промежуточная аттестация	0,35
Консультация перед экзаменом	2
Экзамен	0,35
Зачет	-
Контрольная работа	-
Руководство курсовой работой	-
Самостоятельная работа	***51,65***
Самостоятельная работа в течение семестра	15,65
Подготовка контрольной работы	-
Написание курсовой работы	-
Подготовка к промежуточной аттестации	36
Вид текущей аттестации	контрольная работа
Общая трудоёмкость дисциплины:	
часы	***108***
зачётные единицы	***3***

Заочная форма обучения

Виды учебной работы	Распределение учебного времени
Контактная работа	***12,35***
Аудиторные занятия (всего):	10
Лекции	4
Практические занятия	6
Лабораторные занятия	
Промежуточная аттестация	
Консультация перед экзаменом	2
Экзамен	0,35
Зачет	

Контрольная работа	
Руководство курсовой работой	
Самостоятельная работа	***95,65***
Самостоятельная работа в течение семестра	86,65
Подготовка контрольной работы	-
Написание курсовой работы	-
Подготовка к промежуточной аттестации	9
Вид текущей аттестации	контрольная работа
Общая трудоёмкость дисциплины:	
часы	***108***
зачётные единицы	***3***

4. Содержание разделов и тем учебной дисциплины

Наименование темы учебной дисциплины	Содержание темы
Тема 1. Данные. Подходы и определения (ОПК-4; ОПК-5)	Определение данных: философский, юридический подходы. Классификации объемов данных. Жизненный цикл данных. Создание данных (Data Generation/Data Capture). Обслуживание данных (Data Maintenance). Синтез данных (Data Synthesis). Использование данных (Data Usage). Публикация данных (Data Publication). Архивация данных (Data Archival). Уничтожение данных (Data Purging). Метаданные: понятие и жизненный цикл. Мировой потенциал для хранения, передачи и вычисления информации (динамика роста).
Тема 2. Big Data. Понятие и характеристики Big Data (ОПК-4; ОПК-5)	Понятие больших данных (Big Data). История и причины появления термина Big Data. Глобальный феномен Big Data. Источники больших данных. Основные сферы использования больших данных. Характеристики и формы больших данных. Структурированная форма. Неструктурированная форма. Полуструктурированная форма. Структуризация данных и катологизация. Принципы работы с данными.
Тема 3. Системы управления большими данными (ОПК-4; ОПК-5)	Распределенные файловые системы. Распределенные фреймворки. Бенчмаркинг. Серверное программирование. Планирование. Системы развертывания. Интеграция данных. Информационная безопасность. Машинное обучение. Базы данных NoSQL и новые SQL базы данных.
Тема 4. Визуализация данных и результатов анализа (УК-1; ОПК-4; ОПК-5)	Типы, задачи и виды визуализации. Графики, диаграммы, инфографика. Интерактивный сторителлинг, дашборды. Язык R и его возможности. AmazonS3. Дедупликация данных.
Тема 5. Архитектура системы обработки Big Data (ОПК-4; ОПК-5)	Прием данных (Data Ingestion). Сбор данных (Data Staging). Анализ данных (Analysis Layer). Представление результатов (Consumption Layer).
Тема 6. Технологии анализа и принципы обработки больших данных	Существующие технологии обработки данных. Big Data-ориентированные информационные системы. Оптимизация структуры объектов данных в ходе вычислений. Пакет RHadoop. Система Hadoop и R. Операторы Map и Reduce.

(УК-1; ОПК-4; ОПК-5)	Оператор Reduce (свертка). Оператор Map. Распределенная файловая система HDFS. Моделирование в условиях ограниченного объема памяти.
Тема 7. Классификация задач анализа данных. Проведение анализа с применением технологии больших данных, их интерпретация и формирование аналитических отчетов (УК-1; ОПК-4; ОПК-5)	Data Mining. Интеллектуальный анализ данных, его, отличия и задачи. Text Mining. Web Mining. Web Content Mining. Web Usage Mining. Social media mining. Rapid Miner. Работа в системе Hadoop. Подходы к формированию отчета по результатам анализа. Виды представления результатов анализа.
Тема 8. Статистические методы анализа данных (УК-1; ОПК-4; ОПК-5)	Предсказание и прогнозирование социально-экономических прогнозов. Статистические гипотезы и критерии. Методы многомерного статистического анализа и анализа нечисловой информации. Метрический и линейных классификаторы. ROC–кривая. Кластерный анализ.
Тема 9. Программные платформы и системы для Big Data (ОПК-4; ОПК-5)	Системы управления потоками данных. Системы хранения больших данных. Платформы больших данных. Обработка данных в реальном времени. Системы управления большими данными. Аналитические платформы. Специализированные статистические пакеты.

5. Учебно-методическое и информационное обеспечение учебной дисциплины

5.1. Основная литература (в том числе из ЭБС):

1. Бродовская, Е.В. Большие данные в исследовании политических процессов: учебное пособие: / Е.В. Бродовская, А.Ю. Домбровская; Министерство науки и высшего образования Российской Федерации, Московский педагогический государственный университет. – Москва: Московский педагогический государственный университет (МПГУ), 2018. – 88 с.: схем., табл., ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=563578

2. Радченко, И.А, Технологии и инфраструктура Big Data / И.А. Радченко, И.Н. Николаев. – СПб: Университет ИТМО, 2018. – 52 с. – Режим доступа: https://books.ifmo.ru/file/pdf/2326.pdf

5.2. Дополнительная литература (в том числе из ЭБС):

1. Кобзаренко, Д.Н. Анализ больших данных: учебное пособие / Д.Н. Кобзаренко, А.Г. Мустафаев. – Махачкала: ДГУНХ, 2019 г. – 107 с. – Режим доступа: https://dgunh.ru/content/glavnay/ucheb_deyatel/uposob/up-it_ib-fgos-80.pdf

2. Сенько А.С. Работа с BigData в облаках. Обработка и хранение данных с примерами из Microsoft Azure / А.С. Сенько. – СПб.: Питер, 2019. — 448 с.: ил. — (Серия «Для профессионалов»). – Режим доступа: https://biconsult.ru/files/books/DG/Rabota_s_BigData_v_oblakakh_Obrabotka_i_khranenie_dannykh_s_primerami_iz_Microsoft_Azure_2019_Senko.pdf

3. Управление данными в госсекторе. Навигатор для начинающих / под ред. О. М. Гиацинтова, В. А. Сазонова, М. С. Шклярук. 2-е изд., доп. и перераб. – М.; Берлин: Директ-Медиа, 2014. – 112 с.: ил., табл. – Режим доступа: https://cdto.ranepa.ru/media/reports/pdfs/Data_1_CDTO_RANEPA_.pdf

5.3. Профессиональные базы данных, информационно-справочные и поисковые системы:

Справочно-правовая система «КонсультантПлюс»;

ЭБС «Университетская библиотека онлайн».

5.4. Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. Интернет – университет информационных технологий (ИНТУИТ). (http://www.intuit.ru/).

2. Научно-технический и научно-производственный журнал «Информационные технологии» - http://novtex.ru/IT/index.htm

3. Центр подготовки руководителей и команд цифровой трансформации ВШГУ РАНХиГС - https://cdto.ranepa.ru/

5.5. Нормативно-правовые акты:

1. Об информации, информационных технологиях и о защите информации: федеральный закон от 27.07.2006 г. № 149-ФЗ // Собр. закон-ва РФ. – 2006. – № 31 (1 ч.). – Ст. 3448.

2. ГОСТ Р ИСО/МЭК 15288-2005. Информационная технология. Системная инженерия. Процессы жизненного цикла систем. 2006 г. https://standartgost.ru

3. ГОСТ Р ИСО/МЭК 17799-2005. Информационная технология. https://standartgost.ru

4. ГОСТ Р ИСО 11442-2014. Техническая документация на продукцию. Управление документацией. 2015 г. https://standartgost.ru

5. ГОСТ 34.320-96. Информационные технологии. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной базы. 2001г. https://standartgost.ru

6. Средства обеспечения освоения учебной дисциплины

В учебном процессе при реализации учебной дисциплины «Основы работы с Big Data» используются следующие программные средства:

Информационные технологии	***Перечень программного обеспечения и информационных справочных систем***
Офисный пакет для работы с документами	Microsoft Office Professional Свободно распространяемое программное обеспечение Only Office. https://www.onlyoffice.com
Информационно-справочные системы	Справочно-правовая система «Консультант Плюс»
	Справочно-правовая система «Гарант»
Электронно-библиотечные системы	ЭБС «Университетская библиотека онлайн»
	Национальная электронная библиотека (https://нэб.рф) (в здании ГОУ ВО КРАГСиУ)
	Научная электронная библиотека «КиберЛенинка» https://cyberleninka.ru
	Российская научная электронная библиотека https://www.elibrary.ru
Электронная почта	Электронная почта в домене krags.ru
Средства для организации вебинаров, телемостов и конференций	Сервисы веб- и видеоконференцсвязи, в том числе BigBlueButton

Сопровождение освоения дисциплины обучающимся возможно с использованием электронной информационно-образовательной среды ГОУ ВО КРАГСиУ, в том числе

образовательного портала на основе Moodle (https://moodle.krags.ru)

7. Материально-техническое обеспечение освоения учебной дисциплины

При проведении учебных занятий по дисциплине «Основы работы с Big Data» задействована материально-техническая база академии, в состав которой входят следующие средства и ресурсы для организации самостоятельной и совместной работы обучающихся с преподавателем:

– специальные помещения для реализации данной дисциплины представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Специальные помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории, наборами демонстрационного оборудования и учебно-наглядных пособий, обеспечивающие тематические иллюстрации;

– лаборатории, оснащенные лабораторным оборудованием;

– помещение для самостоятельной работы обучающихся, которое оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду организации;

– компьютерные классы, оснащенные современными персональными компьютерами, работающими под управлением операционных систем Microsoft Windows, объединенными в локальную сеть и имеющими выход в Интернет;

– библиотека Академии, книжный фонд которой содержит научно-исследовательскую литературу, научные журналы и труды научных конференций, а также читальный зал;

– серверное оборудование, включающее, в том числе, несколько серверов серии IBM System X, а также виртуальные сервера, работающие под управлением операционных систем Calculate Linux, включенной в Реестр Российского ПО;

– сетевое коммутационное оборудование, обеспечивающее работу локальной сети, предоставление доступа к сети Интернет с общей скоростью подключения 100 Мбит/сек, а также работу беспроводного сегмента сети Wi-Fi в помещениях Академии;

– интерактивные информационные киоски «Инфо»;

– программные и аппаратные средства для проведения видеоконференцсвязи.

Кроме того, в образовательном процессе обучающимися широко используются следующие электронные ресурсы:

– сеть Internet (скорость подключения – 100 Мбит/сек);

– сайт https://www.krags.ru/;

– беспроводная сеть Wi-Fi.

Конкретные помещения для организации обучения по дисциплине «Основы работы с Big Data» определяются расписанием учебных занятий и промежуточной аттестации. Оборудование и техническое оснащение аудитории, представлено в паспорте соответствующих кабинетов ГОУ ВО КРАГСиУ.

РАЗДЕЛ II. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ

Важнейшим условием успешного освоения материала является планомерная работа обучающегося в течение всего периода изучения дисциплины. Обучающемуся необходимо ознакомиться со следующей учебно-методической документацией:

программой дисциплины; учебником и/или учебными пособиями по дисциплине; электронными ресурсами по дисциплине; методическими и оценочными материалами по дисциплине.

Учебный процесс при реализации дисциплины основывается на использовании *традиционных, инновационных и информационных образовательных технологий*.

Традиционные образовательные технологии представлены *лекциями и занятиями семинарского типа (практические занятия)*.

Инновационные образовательные технологии используются в виде широкого применения активных и интерактивных форм проведения занятий. Аудиторная работа обучающихся может предусматривать интерактивную форму проведения лекционных и практических занятий: *лекции-презентации, анализ практических ситуаций и др*.

Информационные образовательные технологии реализуются путем активизации самостоятельной работы обучающихся в информационной образовательной среде.

Все аудиторные занятия преследуют цель обеспечения высокого теоретического уровня и практической направленности обучения.

Подготовка к лекционным занятиям

В ходе лекций преподаватель излагает и разъясняет основные и наиболее сложные понятия темы, а также связанные с ней теоретические и практические проблемы, дает рекомендации по подготовке к занятиям семинарского типа и самостоятельной работе. В ходе лекционных занятий обучающемуся следует вести конспектирование учебного материала.

С целью обеспечения успешного освоения дисциплины обучающийся должен готовиться к лекции. При этом необходимо:

- внимательно прочитать материал предыдущей лекции;
- ознакомиться с учебным материалом лекции по рекомендованному учебнику и/или учебному пособию;
- уяснить место изучаемой темы в своей профессиональной подготовке;
- записать возможные вопросы, которые обучающийся предполагает задать преподавателю.

Подготовка к занятиям семинарского типа

Этот вид самостоятельной работы состоит из нескольких этапов:

1) повторение изученного материала. Для этого используются конспекты лекций, рекомендованная основная и дополнительная литература;

2) углубление знаний по теме. Для этого рекомендуется выписать возникшие вопросы, используемые термины;

При подготовке к занятиям семинарского типа рекомендуется с целью повышения их эффективности:

– уделять внимание разбору теоретических задач, обсуждаемых на лекциях;

– уделять внимание краткому повторению теоретического материала, который используется при выполнении практических заданий;

– выполнять внеаудиторную самостоятельную работу;

– ставить проблемные вопросы, по возможности использовать примеры и задачи с практическим содержанием;

– включаться в используемые при проведении практических занятий активные и интерактивные методы обучения.

При разборе примеров в аудитории или дома целесообразно каждый их них обосновывать теми или иными теоретическими положениями.

Активность на занятиях семинарского типа оценивается по следующим критериям:

- ответы на вопросы, предлагаемые преподавателем;
- участие в дискуссиях;
- выполнение проектных и иных заданий;
- ассистирование преподавателю в проведении занятий.

Организация самостоятельной работы

Самостоятельная работа обучающихся представляет собой процесс активного, целенаправленного приобретения ими новых знаний, умений без непосредственного участия преподавателя, характеризующийся предметной направленностью, эффективным контролем и оценкой результатов деятельности обучающегося.

Задачами самостоятельной работы являются:

- систематизация и закрепление полученных теоретических знаний и практических умений обучающихся;
- углубление и расширение теоретических знаний;
- формирование умений использовать нормативную и справочную документацию, специальную литературу;
- развитие познавательных способностей, активности обучающихся, ответственности и организованности;
- формирование самостоятельности мышления, творческой инициативы, способностей к саморазвитию, самосовершенствованию и самореализации;
- развитие исследовательских умений.

При изучении дисциплины организация самостоятельной работы обучающихся представляет собой единство трех взаимосвязанных форм:

1) внеаудиторная самостоятельная работа;

2) аудиторная самостоятельная работа, которая осуществляется под непосредственным руководством преподавателя при проведении практических занятий и во время чтения лекций;

3) творческая, в том числе научно-исследовательская работа.

Перед выполнением обучающимися внеаудиторной самостоятельной работы преподаватель может давать разъяснения по выполнению задания, которые включают:

- цель и содержание задания;
- сроки выполнения;
- ориентировочный объем работы;
- основные требования к результатам работы и критерии оценки;
- возможные типичные ошибки при выполнении.

Контроль результатов внеаудиторной самостоятельной работы обучающихся может проходить в письменной, устной или смешанной форме.

Подготовка к промежуточной аттестации

Видами промежуточной аттестации по данной дисциплине являются сдача *экзамена*. При проведении промежуточной аттестации выясняется усвоение основных теоретических и прикладных вопросов программы и умение применять полученные знания к решению практических задач. При подготовке к *экзамену* учебный материал рекомендуется повторять по учебному изданию, рекомендованному в качестве основной литературы, и конспекту. *Экзамен* проводится в назначенный день, по окончании изучения дисциплины. После контрольного мероприятия преподаватель учитывает активность работы обучающегося на аудиторных занятиях, качество самостоятельной работы, результаты текущей аттестации, посещаемость и выставляет итоговую оценку.

Изучение дисциплины с использованием дистанционных образовательных технологий

При изучении дисциплины с использованием дистанционных образовательных технологий необходимо дополнительно руководствоваться локальными нормативными актами ГОУ ВО КРАГСиУ, регламентирующими организацию образовательного процесса с использованием дистанционных образовательных технологий.

РАЗДЕЛ III. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

8. Контрольно-измерительные материалы, необходимые для проверки сформированности индикаторов достижения компетенций (знаний и умений)

8.1. Задания для проведения текущего контроля (контрольная работа)

Задание 1.

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах для экосистемы Hadoop. Механизм ее работы предлагается изучить путем пошаговой реализации ее алгоритмов на условном примере.

Задача обучающегося – выполнить подсчет количества разнородных элементов в наборе данных, выданных преподавателем: в следующей последовательности, согласно алгоритма MapReduce:

Шаг 1: Сохраните набор данных в 4 разделах (как в HDFS). Цель научиться балансировать нагрузку на ресурсы.

Шаг 2: Сопоставьте данные. Цель понять суть кластеризации и присвоения ключей.

Шаг 3: Сортировка и перемешивание. Цель научиться балансировать нагрузку на ресурсы.

Шаг 4: Сгенерировать пары "ключ-значение".

Задание 2.

Hadoop Distributed File System (HDFS) – распределённая файловая система, позволяющая хранить информацию практически неограниченного объёма.

Cloudera Virtual Machine (VM) Image – преднастроенный виртуальный образ/машина создаваемый на персональном компьютере.

Задача обучающегося – посчитать, сколько раз встречается слово в тексте.

Шаг 1: Загрузите и установите VirtualBox.

Шаг 2: Загрузите и установите образ виртуальной машины (ВМ) Cloudera.

Шаг 3: Запустите Cloudera VM.

Шаг 4: Загрузите файл в HDFS с текстом, выданный преподавателем;

Шаг 5: Запустите приложение WordCount.

Шаг 6: Скопируйте результаты WordCount из Hadoop Distributed File System (HDFS).

8.2. Вопросы для подготовки к экзамену

1. Понятие и жизненный цикл данных.
2. Сущность больших данных и перспективы их использования в различных сферах.
3. Условия и возможности использования больших данных в различных сферах.
4. Примеры использования больших данных в зарубежной практике.
5. Примеры использования больших данных на национальном уровне.
6. Понятие и цели Big Data. Роль цифровой информации в современных условиях.
7. Базовые принципы обработки больших данных.
8. Главные характеристики Big Data.
9. Причины появления технологий больших данных.
10. Источники получения больших данных.
11. Виды и формы данных.
12. Системы управления большими данными.
13. Распределенные фреймворки
14. Бенчмаркинг в больших данных
15. Технологии обработки больших данных: NoSQL
16. Примеры и инструменты для визуализации.
17. Data Mining. Постановка основных задач.
18. Машинное обучение.

19. Бизнес-решения с помощью алгоритмов Data Mining.
20. Распараллеливание данных: сущность и алгоритмы.
21. Экосистема Hadoop.
22. Возможности и ограничения использования ресурсов среды программирования R при анализе больших данных.
23. Вычислительная модель MapReduce.
24. В чем суть HDFS?
25. Цели кластеризации.
26. Платформы больших данных

8.3. Вариант заданий для проведения промежуточного контроля

1. Big Data – это …

(1) представление фактов, понятий или инструкций в форме, приемлемой для интерпретации, или обработки;
(2) комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов;
(3) колоссальный объем данных, собранных человечеством;
(4) класс в Java, предназначенный для хранения данных от 100 Гб.

2. Объём накопленных человечеством цифровых данных на 2012 год измеряется:

(1) петабайтами;
(2) зеттабайтами;
(3) экзабайтами;
(4) йоттабайтами.

3. Укажите фактор, способствовавший появлению тренда больших данных

(1) маркетинговые кампании крупных корпораций;
(2) снижение издержек на хранение данных;
(3) появление новых технологий обработки потоковых данных;
(4) выпуск баз данных с обработкой данных в памяти.

4. Какие вероятные разочарования тренда больших данных?

(1) из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных;
(2) из-за угрозы безопасности личной жизни (privacy) граждан будут упрощены процедуры сбора данных, что приведёт к падению ценности больших данных;
(3) нет.

5. Отметьте значимые события, повлиявшие на формирование тренда больших данных:

(1) разработка Hadoop;
(2) изобретение принципа MapReduce;
(3) разработка языка Pyhton.

6. Какие данные занимают больше мировой памяти относительно остальных?

(1) Structured Data;
(2) Unstructured Data;
(3) Semi-Structured Data;
(4) Quasi-Structured Data.

7. Выберите верный ответ
(1) большие данные – это обработка или хранение более 1 Тб информации;
(2) проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна;
(3) большие данные – это огромная PR-акция крупных вендоров и не более того;
(4) большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.

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

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

10. Перечислите четыре основных характеристики Big Data:
(1) Virtualization, Volume, Variability, Vehicle;
(2) Variety, Velocity, Volume, Value;
(3) Verification, Volume, Velocity, Visualization;
(4) Video, Value, Variety, Volume.

11. Разбиение системы на более мелкие структурные компоненты и разнесение их по отдельным физическим машинам (или их группам), и (или) увеличение количества серверов, параллельно выполняющих одну и ту же функцию, это:
(1) Горизонтальное масштабирование;
(2) Вертикальное масштабирование;
(3) Master- slave репликация;
(4) Peer-to-peer репликация;

12. Принцип MapReduce состоит в том, чтобы
(1) производить вычисления на узлах, где информация изначально была сохранена;
(2) использовать вычислительные мощности систем хранения;
(3) использовать функциональное программирование для решения задач массивно-параллельной обработки.

13. Что из этого является недостатком MapReduce?
(1) Фиксированный алгоритм обработки данных;

(2) Масштабируемость;
(3) Отказоустойчивость;
(4) Возможность автоматического распараллеливания.

14. Данные имеющие определенный тип, формат и структуру (например, транзакции) являются:
(1) Структурированными;
(2) Полуструктурированными;
(3) Квазиструктурированными;
(4) Неструктурированными;

15. Какая компания создала технологию MapReduce?
(1) Google;
(2) Yahoo;
(3) EMC;
(4) Oracle.

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

17. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
(1) 100Гб;
(2) 1Тб;
(3) 100Тб;
(4) 1Пб.

18. Человек покупает товары через интернет. Государство хочет знать насколько могут возрасти такие продажи в ближайшем будущем и когда. К какому типу относится эта задача анализа данных?
(1) прогнозирование;
(2) кластеризация;
(3) классификация;
(4) цензурирование.

19. Инвестиционный фонд интересуется тем, почему часть финансируемых им проектов успешно переходят на второй год, а часть - нет. К какому типу относится эта задача анализа данных?
(1) поиск информативных признаков;
(2) построение решающего правила;
(3) классификация;
(4) цензурирование.

20. Инвестиционный фонд имеет ряд проектов, который успешно переходят на второй год финансирования и тех, кто не переходит. Как бы в данном случае формулировалась задача поиска информативных признаков?
(1) определить, почему ряд проектов успешно переходят на второй год, а ряд – нет;
(2) определить для нового проекта, перейдёт ли он через год на второй этап финансирования или нет;
(3) восстановить некоторые характеристики проектов, которые изначально не заполнялись;
(4) определить критерий успешности.

21. Инвестиционный фонд имеет ряд проектов, который успешно переходят на второй год финансирования и тех, кто не переходит. Фонд поставил задачу определить критерий успешности проекта. К какому типу задач анализа данных наиболее близка эта задача?
(1) прогнозирование;
(2) построение решающего правила;
(3) поиск информативных признаков;
(4) цензурирование.

22. Поликлиникой ставится цель определения структуры своих клиентов с точки зрения числа обращений. К какому типу относится эта задача анализа данных?
(1) прогнозирование;
(2) кластеризация;
(3) классификация;
(4) цензурирование.

23. Поликлиника обладает некоторыми данными о клиентах и о их возрасте. Как бы в данном случае формулировалась задача кластеризации?
(1) определить основные группы клиентов;
(2) определить, сколько раз придет тот или иной клиент в следующем периоде;
(3) определить, когда вернётся тот или иной клиент.

24. Компания, проводящая социологические опросы, испытывает сложности с верификацией данных, поступающих от волонтеров непосредственно опрашивающих респондентов: многие анкеты заполнены не полностью; волонтеры фальсифицируют результаты опроса, самостоятельно заполняя часть анкет. К какому типу наиболее близка эта задача анализа данных?
(1) прогнозирование;
(2) кластеризация;
(3) классификация;
(4) цензурирование.

25. С некоторой периодичностью на госпредприятии списываются группы расходных материалов на различных участках учета. Для выявления ошибок, акты списания выборочно проверяются аудитором. Руководство заинтересовано в сокращении количества проверок, при сохранении точности выявления ошибочного списания на уровне 97%. Требуется выявлять сомнительные акты списания,

подлежащие обязательной проверке аудитором. К какому типу относится эта задача анализа данных?
(1) прогнозирование;
(2) кластеризация;
(3) классификация;
(4) цензурирование.

26. С некоторой периодичностью на госпредприятии списываются группы расходных материалов на различных участках учета. Для выявления ошибок, акты списания выборочно проверяются аудитором. Как бы в данном случае формулировалась задача классификации?
(1) определить характерные признаки ошибочных списаний;
(2) научиться автоматически выявлять ошибочные списания с ожидаемой ошибкой не ниже 97%;
(3) классифицировать типичные ошибки и составить их список;
(4) определить три категории: "ошибочные", "под сомнением", "безошибочные" и найти правило отнесения к этим категориям.

27. С некоторой периодичностью на госпредприятии списываются группы расходных материалов на различных участках учета. Для выявления ошибок, акты списания выборочно проверяются аудитором. Определены три категории: "ошибочные", "под сомнением", "безошибочные". К какому типу задач анализа данных относится задача о построении правила автоматического отнесения списаний к этим категориям.
(1) поиск информативных признаков;
(2) кластеризация;
(3) классификация;
(4) цензурирование.

28. К какому типу шкал относится шкала "очень плохо"-"плохо"-"средне"-"хорошо"-"очень хорошо"?
(1) порядковая;
(2) абсолютная;
(3) бинарная;
(4) номинальная.

9. Критерии выставления оценок по результатам изучения дисциплины

Освоение обучающимся каждой учебной дисциплины в семестре, независимо от её общей трудоёмкости, оценивается по 100-балльной шкале, которая затем при промежуточном контроле в форме экзамена и дифференцированного зачета переводится в традиционную 4-балльную оценку («отлично», «хорошо», «удовлетворительно», «неудовлетворительно»), а при контроле в форме зачёта – в 2-балльную («зачтено» или «незачтено»). Данная 100-балльная шкала при необходимости соотносится с Европейской системой перевода и накопления кредитов (ECTS).

Соотношение 2-, 4- и 100-балльной шкал оценивания освоения обучающимися учебной дисциплины со шкалой ECTS

Оценка по 4-бальной шкале	Зачёт	Сумма баллов по дисциплине	Оценка ECTS	Градация

5 (отлично)	Зачтено	90 – 100	A	Отлично
4 (хорошо)		85 – 89	B	Очень хорошо
		75 – 84	C	Хорошо
3 (удовлетворительно)		70 – 74	D	Удовлетворительно
		65 –69		
		60 – 64	E	Посредственно
2 (неудовлетворительно)	Не зачтено	Ниже 60	F	Неудовлетворительно

Критерии оценок ECTS

5	A	«**Отлично**» – теоретическое содержание дисциплины освоено полностью, без пробелов, необходимые практические умения работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному
4	B	«**Очень хорошо**» – теоретическое содержание дисциплины освоено полностью, без пробелов, необходимые практические умения работы с освоенным материалом в основном сформированы, все предусмотренные программой обучения учебные задания выполнены, качество выполнения большинства из них оценено числом баллов, близким к максимальному, однако есть несколько незначительных ошибок
	C	***«*Хорошо*»*** – теоретическое содержание дисциплины освоено полностью, без пробелов, некоторые практические умения работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками
3	D	***«*Удовлетворительно*»*** – теоретическое содержание дисциплины освоено частично, но пробелы не носят существенного характера, необходимые практические умения работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки
	E	***«*Посредственно*»*** – теоретическое содержание дисциплины освоено частично, некоторые практические умения работы не сформированы, многие предусмотренные программой обучения учебные задания не выполнены, либо качество выполнения некоторых из них оценено числом баллов, близким к минимальному
2	F	**«Неудовлетворительно»** – теоретическое содержание дисциплины не освоено, необходимые практические умения работы не сформированы, все выполненные учебные задания содержат грубые ошибки, дополнительная самостоятельная работа над материалом дисциплины не приведет к какому-либо значимому повышению качества выполнения учебных заданий

Оценивание результатов обучения по дисциплине осуществляется в форме текущего и промежуточной контроля. Текущий контроль в семестре проводится с целью обеспечения своевременной обратной связи, с целью активизации самостоятельной работы обучающихся. Объектом промежуточного контроля являются конкретизированные результаты обучения (учебные достижения) по дисциплине.

Структура итоговой оценки обучающихся
Критерии и показатели оценивания результатов обучения

№	Критерии оценивания	Показатели (оценка в баллах)
1	Работа на аудиторных занятиях	20
2	Посещаемость	5
3	Самостоятельная работа	15
4	**Текущая аттестация**	**20**
	Итого	**60**
5	**Промежуточная аттестация**	**40**
	Всего	**100**

Критерии и показатели оценивания результатов обучения в рамках аудиторных занятий

№	Критерии оценивания	Показатели (оценка в баллах)
1	Подготовка и выступление с докладом	до 5 баллов
2	Активное участие в обсуждении доклада	до 5 баллов
3	Выполнение практического задания (анализ практических ситуаций, составление документов, сравнительных таблиц)	до 5 баллов
4	Другое	до 5 баллов
	Всего	**20**

Критерии и показатели оценивания результатов обучения в рамках посещаемости обучающимся аудиторных занятий

Критерии оценивания	Показатели (оценка в баллах)
100% посещение аудиторных занятий	5
100% посещение аудиторных занятий. Небольшое количество пропусков по уважительной причине	4
До 30% пропущенных занятий	3
До 50% пропущенных занятий	2
До 70% пропущенных занятий	1
70% и более пропущенных занятий	0

Критерии и показатели оценивания результатов обучения в рамках самостоятельной работы обучающихся

Критерии оценивания	Показатель (оценка в баллах)
Раскрыты основные положения вопроса или задания через систему аргументов, подкрепленных фактами, примерами, обоснованы предлагаемые в самостоятельной работе решения, присутствуют полные с детальными пояснениями выкладки, оригинальные предложения, обладающие элементами практической значимости, самостоятельная работа качественно и чётко оформлена	15–12
В работе присутствуют отдельные неточности и замечания непринципиального характера	11–9
В работе имеются серьёзные ошибки и пробелы в знаниях	8–5
Задание не выполнено или выполнено с грубыми ошибками	0

Критерии и показатели оценивания результатов обучения в рамках текущей аттестации

Критерии оценивания	Показатели (оценка в баллах)
Задание полностью выполнено, правильно применены теоретические положения дисциплины. Отмечается чёткость и структурированность изложения, оригинальность мышления	20–17
Задание полностью выполнено, при подготовке применены теоретические положения дисциплины, потребовавшие уточнения или незначительного исправления	16–13
Задание выполнено, но теоретическая составляющая нуждается в доработке. На вопросы по заданию были даны нечёткие или частично ошибочные ответы	12–5
Задание не выполнено или при ответе сделаны грубые ошибки, демонстрирующие отсутствие теоретической базы знаний обучающегося	0

Критерии и показатели оценивания результатов обучения в рамках промежуточного контроля

Промежуточный контроль в форме *экзамена* имеет целью проверку и оценку знаний обучающихся по теории и применению полученных знаний и умений.

Критерии и показатели оценки результатов экзамена в устной/письменной форме

Критерии оценивания	Показатели (оценка в баллах)
продемонстрировано глубокое и прочное усвоение знаний материала; исчерпывающе, последовательно, грамотно и логически стройно изложен теоретический материал; правильно формулированы определения; продемонстрировано умение делать выводы по излагаемому материалу;	40–35
продемонстрировано достаточно полное знание материала, основных теоретических понятий; достаточно последовательно, грамотно и логически стройно изложен материал; продемонстрировано умение делать достаточно обоснованные выводы по излагаемому материалу; с некоторыми неточностями	34–25
продемонстрировано общее знание изучаемого материала, основной рекомендуемой программой дисциплины учебной литературы, умение строить ответ в соответствии со структурой излагаемого вопроса; показано общее владение понятийным аппаратом дисциплины;	24–15
продемонстрировано незнание значительной части программного материала; невладение понятийным аппаратом дисциплины; сделаны существенные ошибки при изложении учебного материала; продемонстрировано неумение строить ответ в соответствии со структурой излагаемого вопроса, делать выводы по излагаемому материалу,	14–0