Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

УТВЕРЖДАЮ
Директор по образовательной деятельности

_____ С.Т. Князев
«__» _____ 2022

РАБОЧАЯ ПРОГРАММА МОДУЛЯ

Код модуля	Модуль
1161154	Обработка больших объемов данных

Екатеринбург
2022

Перечень сведений о рабочей программе модуля	Учетные данные
Образовательная программа Алгоритмы искусственного интеллекта	**Код ОП** 09.03.01
Направление подготовки Информатика и вычислительная техника	**Код направления и уровня подготовки** 09.03.01

Области образования, в рамках которых реализуется модуль образовательной программы по СУОС УрФУ :

№ п/п	Перечень областей образования, для которых разработан СУОС УрФУ	Уровень подготовки
1.	Инженерное дело, технологии и технические науки	бакалавриат

Программа модуля составлена авторами:

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	к.т.н.	Доцент	Кафедра информационных технологий и систем управления

1. ОБЩАЯ ХАРАКТЕРИСТИКА МОДУЛЯ Обработка больших объемов данных

1.1. Аннотация содержания модуля

Модуль "Обработка больших объемов данных" состоит из одноименной дисциплины и способствует формированию у студентов представления об основах технологий обработки больших объемов данных и жизненного цикла разработки приложений обработки больших объемов данных, изучение особенностей использования командной строки Linux в системах обработки больших объемов данных, а также получения навыков использования современного инструмента анализа больших данных Apache Spark.

1.2. Структура и объем модуля

Таблица 1

№ п/п	Перечень дисциплин модуля в последовательности их освоения	Объем дисциплин модуля и всего модуля в зачетных единицах
1	Обработка больших объемов данных	3
	ИТОГО по модулю:	3

1.3.Последовательность освоения модуля в образовательной программе

Пререквизиты модуля	Не предусмотрены
Постреквизиты и кореквизиты модуля	Не предусмотрены

1.4. Распределение компетенций по дисциплинам модуля, планируемые результаты обучения (индикаторы) по модулю

Таблица 2

Перечень дисциплин модуля	Код и наименование компетенции	Индикаторы достижения компетенции	Планируемые результаты обучения
1	2	3	4
Обработка больших объемов данных	ПК-6. Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-6.1. Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения	ПК-6.1. З-1. Знает методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных ПК-6.1. З-2. Знает методы планирования вычислительного эксперимента, формирования

			обучающей и контрольной выборок ПК-6.1. У-1. Умеет выявлять и исключать из массива данных ошибочные данные и выбросы ПК-6.1. У-2. Умеет выделять входные и выходные переменные с целью использования предиктивных моделей ПК-6.1. У-3. Умеет осуществлять разметку структурированных и неструктурированных данных ПК-6.1. У-4. Умеет использовать инструменты, библиотеки и технологии Data Science для подготовки и разметки структурированных и неструктурированных данных для машинного обучения ПК-6.1. У-5. Умеет использовать методы и технологии массово параллельной обработки и анализа данных

1.5. Форма обучения

Обучение по дисциплинам модуля может осуществляться в очной форме.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Обработка больших объемов данных

Рабочая программа дисциплины составлена авторами:

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	к.т.н.	Доцент	Кафедра информационных технологий и систем управления

1. СОДЕРЖАНИЕ И ОСОБЕННОСТИ РЕАЛИЗАЦИИ ДИСЦИПЛИНЫ Обработка больших объемов данных

1.1. Технологии реализации, используемые при изучении дисциплины модуля

- Традиционная (репродуктивная) технология
- Разноуровневое (дифференцированное) обучение
    - Базовый уровень

**Базовый I уровень – сохраняет логику самой науки и позволяет получить упрощенное, но верное и полное представление о предмете дисциплины, требует знание системы понятий, умение решать проблемные ситуации. Освоение данного уровня результатов обучения должно обеспечить формирование запланированных компетенций и позволит обучающемуся на минимальном уровне самостоятельности и ответственности выполнять задания;*
Продвинутый II уровень – углубляет и обогащает базовый уровень как по содержанию, так и по глубине проработки материала дисциплины. Это происходит за счет включения дополнительной информации. Данный уровень требует умения решать проблемы в рамках курса и смежных курсов посредством самостоятельной постановки цели и выбора программы действий. Освоение данного уровня результатов обучения позволит обучающемуся повысить уровень самостоятельности и ответственности до творческого применения знаний и умений.

1.2. Содержание дисциплины

Таблица 1.1

Код раздела, темы	**Раздел, тема дисциплины***	**Содержание**
1.	Анализ больших данных	Большие данные и методы их обработки. Жизненный цикл приложений обработки больших объемов данных. Распределенные системы хранения больших данных. Инструменты распределенной обработки больших объемов данных.
2.	ОС Linux и работа в командной строке	Основы работы в командной строке Linux. Редактирование файлов в Linux. Сетевое взаимодействие в ОС Linux. Работа с кластером обработки больших данных в Linux.
3.	Анализ больших данных с помощью Apache Spark	Алгоритмы обработки больших данных с помощью Apache Spark. Разработка программ Spark на Python с помощью PySpark. Использование DataFrame API в Apache Spark.

1.3. Направление, виды воспитательной деятельности и используемые технологии

Таблица 1.2

Направление воспитательной деятельности	Вид воспитательной деятельности	Технология воспитательной деятельности	Компетенция	Результаты обучения
Профессиональное воспитание	профориентационная деятельность	Технология формирования уверенности и готовности к самостоятельной успешной	ПК-6. Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-6.1. З-1. Знает методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки

		профессиональной деятельности Технология самостоятельной работы		структурированных и неструктурированных данных

1.4. Программа дисциплины реализуется на государственном языке Российской Федерации .

2. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ Обработка больших объемов данных

Электронные ресурсы (издания)

1. Мейер, Б. Инструменты, алгоритмы и структуры данных / Б. Мейер. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 543 с. : схем., ил. - Библиогр. в кн. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=429033
2. Крутиков, В.Н. Анализ данных : учебное пособие / В.Н. Крутиков, В.В. Мешечкин ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Кемеровский государственный университет». - Кемерово : Кемеровский государственный университет, 2014. - 138 с. : ил. - Библиогр. в кн. - ISBN 978-5-8353-1770-7 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=278426

Печатные издания

1. Spark для профессионалов. Современные паттерны обработки больших данных / С. Риза [и др.] ; [пер. с англ. И. Пальти] .— Санкт-Петербург ; Москва ; Екатеринбург [и др.] : Питер, 2017 .— 272 с. : ил. — (Бестселлеры O'Reilly) .— Пер. изд.: Advanced Analytics with Spark / S. Ryza et al, Beijing etc. 2016 .— ISBN 978-5-496-02401-3.
2. Парфенов, Юрий Павлович. Постреляционные хранилища данных : [учебное пособие для студентов вузов, по программе магистратуры по направлению подготовки "Информатика и вычислительная техника"] / Ю. П. Парфенов ; [под науч. ред. Н. В. Папуловской] ; Урал. федер. ун-т имени первого Президента России Б. Н. Ельцина .— Москва ; Екатеринбург : Юрайт : Издательство Уральского университета, 2017 .— 121 с. : ил. — (Университеты России) .— Рек. метод. советом УрФУ .— Библиогр.: с. 119-121 .— ISBN 978-5-534-03408-0 (Юрайт) .— ISBN 978-5-7996-1827-8 (Изд-во Урал. ун-та).
3. Бурнаева, Эльфия Гарифовна. Обработка и представление данных в MS Excel : учебное пособие / Э. Г. Бурнаева, С. Н. Леора .— Санкт-Петербург ; Москва ; Краснодар : Лань, 2016 .— 156 с. : ил. — (Учебники для вузов. Специальная литература) .— Библиогр.: с. 153 (11 назв.) .— ISBN 978-5-8114-1923-4.
4. Мэтиз, Эрик. Изучаем Python. Программирование игр, визуализация данных, веб-приложения / Э. Мэтиз ; [пер. с англ. Е. Матвеева] .— Санкт-Петербург ; Москва ; Нижний Новгород [и др.] : Питер, 2017 .— 496 с. : ил. — (Библиотека программиста) .— Пер. изд.: Python crash course. / E. Matthes, San Francisco. 2016 .— ISBN 978-5-496-02305-4.

Профессиональные базы данных, информационно-справочные системы

http://e.lanbook.com/

http://www.tandfonline.com

http://onlinelibrary.wiley.com/

http://www.biblioclub.ru/

Материалы для лиц с ОВЗ

Весь контент ЭБС представлен в виде файлов специального формата для воспроизведения синтезатором речи, а также в тестовом виде, пригодном для прочтения с использованием экранной лупы и настройкой контрастности.

Базы данных, информационно-справочные и поисковые системы

1. Государственная публичная научно-техническая библиотека. Режим доступа: http://www.gpntb.ru, свободный.

2. Список библиотек, доступных в Интернет и входящих в проект «Либне». Режим доступа: http://www.valley.ru/-nicr/listrum.htm, свободный.

3. Российская национальная библиотека. Режим доступа: http://www.rsl.ru, свободный.

4. Библиотека нормативно-технической литературы. Режим доступа: http://www.tehlit.ru, свободный.

5. Электронная библиотека нормативно-технической документации. Режим доступа: http://www.technormativ.ru, свободный.

6. Библиотека В. Г. Белинского. Режим доступа: http://book.uraic.ru, свободный.

7. Электронный каталог Зональной научной библиотеки УрФУ. Режим доступа http://opac.urfu.ru/, свободный.

8. Электронно-библиотечная система «Лань». Режим доступа https://e.lanbook.com/

9. CONSENSUS: корпоративная сеть библиотек Урала. Режим доступа: http://consensus.urfu.ru.

10. Научная электронная библиотека Elibrary. Режим доступа: http://elibrary.ru

11. Информационные технологии и сервисы. Онлайн-курс. Режим доступа: https://openedu.ru/course/urfu/ITS/

12.http://eor.edu.ru/

13. https://www.computerra.ru/

3. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Обработка больших объемов данных

Сведения об оснащенности дисциплины специализированным и лабораторным оборудованием и программным обеспечением

Таблица 3.1

№ п/п	Виды занятий	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения

1	Лекции	Мебель аудиторная с количеством рабочих мест в соответствии с количеством студентов Рабочее место преподавателя Доска аудиторная Подключение к сети Интернет	Microsoft Windows 8.1 Pro 64-bit RUS OLP NL Acdmc Office 365 EDUA3 ShrdSvr ALNG SubsVL MVL PerUsr B Faculty EES
2	Лабораторные занятия	Мебель аудиторная с количеством рабочих мест в соответствии с количеством студентов Рабочее место преподавателя Доска аудиторная Подключение к сети Интернет	Microsoft Windows 8.1 Pro 64-bit RUS OLP NL Acdmc Office 365 EDUA3 ShrdSvr ALNG SubsVL MVL PerUsr B Faculty EES Интегрированная среда разработки Microsoft Visual Studio Apache Hadoop

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ
ПО ДИСЦИПЛИНЕ
Обработка больших объемов данных

Екатеринбург

Оценочные материалы составлены автором(ами):

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	к.т.н.	Доцент	Кафедра информационных технологий и систем управления

1. СТРУКТУРА И ОБЪЕМ ДИСЦИПЛИНЫ Обработка больших объемов данных

1.	● **Объем дисциплины в зачетных единицах**	● 3	
2.	● **Виды аудиторных занятий**	Лекции Лабораторные занятия	
3.	● **Промежуточная аттестация**	Экзамен	
4.	● **Текущая аттестация**	Контрольная работа Домашняя работа	1 1

2. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ (ИНДИКАТОРЫ) ПО ДИСЦИПЛИНЕ МОДУЛЯ Обработка больших объемов данных

Индикатор – это признак / сигнал/ маркер, который показывает, на каком уровне обучающийся должен освоить результаты обучения и их предъявление должно подтвердить факт освоения предметного содержания данной дисциплины, указанного в табл. 1.3 РПМ-РПД.

Таблица 1.

Код и наименование компетенции	**Индикаторы достижения компетенции**	**Планируемые результаты обучения**	**Контрольно-оценочные средства для оценивания достижения результата обучения по дисциплине**
1	**2**		**3**
ПК-6. Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-6.1. Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения	ПК-6.1. З-1. Знает методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных ПК-6.1. З-2. Знает методы планирования вычислительного эксперимента, формирования обучающей и контрольной выборок ПК-6.1. У-1. Умеет выявлять и исключать из массива данных ошибочные данные и выбросы	Лекции Лабораторные занятия Контрольная работа Домашняя работа Экзамен

		ПК-6.1. У-2. Умеет выделять входные и выходные переменные с целью использования предиктивных моделей ПК-6.1. У-3. Умеет осуществлять разметку структурированных и неструктурированных данных ПК-6.1. У-4. Умеет использовать инструменты, библиотеки и технологии Data Science для подготовки и разметки структурированных и неструктурированных данных для машинного обучения ПК-6.1. У-5. Умеет использовать методы и технологии массово параллельной обработки и анализа данных	

3. ПРОЦЕДУРЫ КОНТРОЛЯ И ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ В РАМКАХ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ МОДУЛЯ В БАЛЬНО-РЕЙТИНГОВОЙ СИСТЕМЕ (ТЕХНОЛОГИЧЕСКАЯ КАРТА БРС)

3.1. Процедуры текущей и промежуточной аттестации по дисциплине

1. Лекции: коэффициент значимости совокупных результатов лекционных занятий – 0.6		
Текущая аттестация на лекциях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
Контрольная работа	6, 5	40
Домашняя работа	6, 9	60
Весовой коэффициент значимости результатов текущей аттестации по лекциям – 0.4		
Промежуточная аттестация по лекциям – экзамен **Весовой коэффициент значимости результатов промежуточной аттестации по лекциям – 0.6**		
2. Практические/семинарские занятия: коэффициент значимости совокупных результатов практических/семинарских занятий – не предусмотрено		
Текущая аттестация на практических/семинарских занятиях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**

Весовой коэффициент значимости результатов текущей аттестации по практическим/семинарским занятиям– не предусмотрено		
Промежуточная аттестация по практическим/семинарским занятиям–нет **Весовой коэффициент значимости результатов промежуточной аттестации по практическим/семинарским занятиям– не предусмотрено**		
3. Лабораторные занятия: коэффициент значимости совокупных результатов лабораторных занятий – 0.4		
Текущая аттестация на лабораторных занятиях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
Защита лабораторных работ	6, 1-16	100
Весовой коэффициент значимости результатов текущей аттестации по лабораторным занятиям - 1		
Промежуточная аттестация по лабораторным занятиям – не предусмотрено **Весовой коэффициент значимости результатов промежуточной аттестации по лабораторным занятиям– 0**		

3.2. Процедуры текущей и промежуточной аттестации курсовой работы/проекта

Текущая аттестация выполнения курсовой работы/проекта	**Сроки − семестр, учебная неделя**	**Максимальная оценка в баллах**
Весовой коэффициент текущей аттестации выполнения курсовой работы/проекта– не предусмотрено		
Весовой коэффициент промежуточной аттестации выполнения курсовой работы/проекта– защиты – не предусмотрено		

4. КРИТЕРИИ И УРОВНИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ МОДУЛЯ

4.1. В рамках БРС применяются утвержденные на кафедре/институте критерии (признаки) оценивания достижений студентов по дисциплине модуля (табл. 4) в рамках контрольно-оценочных мероприятий на соответствие указанным в табл.1 результатам обучения (индикаторам).

Таблица 4

Критерии оценивания учебных достижений обучающихся

Результаты обучения	**Критерии оценивания учебных достижений, обучающихся на соответствие результатам обучения/индикаторам**
Знания	Студент демонстрирует знания и понимание в области изучения на уровне указанных индикаторов и необходимые для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Умения	Студент может применять свои знания и понимание в контекстах, представленных в оценочных заданиях, демонстрирует освоение умений на уровне указанных индикаторов и необходимых для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Опыт /владение	Студент демонстрирует опыт в области изучения на уровне указанных индикаторов.

Другие результаты	Студент демонстрирует ответственность в освоении результатов обучения на уровне запланированных индикаторов. Студент способен выносить суждения, делать оценки и формулировать выводы в области изучения. Студент может сообщать преподавателю и коллегам своего уровня собственное понимание и умения в области изучения.

4.2 Для оценивания уровня выполнения критериев (уровня достижений обучающихся при проведении контрольно-оценочных мероприятий по дисциплине модуля) используется универсальная шкала (табл. 5).

Таблица 5

Шкала оценивания достижения результатов обучения (индикаторов) по уровням

Характеристика уровней достижения результатов обучения (индикаторов)				
№ п/п	**Содержание уровня выполнения критерия оценивания результатов обучения (выполненное оценочное задание)**	**Шкала оценивания**		
		Традиционная характеристика уровня		**Качественная характеристика уровня**
1.	Результаты обучения (индикаторы) достигнуты в полном объеме, замечаний нет	Отлично (80-100 баллов)	Зачтено	Высокий (В)
2.	Результаты обучения (индикаторы) в целом достигнуты, имеются замечания, которые не требуют обязательного устранения	Хорошо (60-79 баллов)		Средний (С)
3.	Результаты обучения (индикаторы) достигнуты не в полной мере, есть замечания	Удовлетворительно (40-59 баллов)		Пороговый (П)
4.	Освоение результатов обучения не соответствует индикаторам, имеются существенные ошибки и замечания, требуется доработка	Неудовлетворительно (менее 40 баллов)	Не зачтено	Недостаточный (Н)
5.	Результат обучения не достигнут, задание не выполнено	Недостаточно свидетельств для оценивания		Нет результата

5. СОДЕРЖАНИЕ КОНТРОЛЬНО-ОЦЕНОЧНЫХ МЕРОПРИЯТИЙ ПО ДИСЦИПЛИНЕ МОДУЛЯ

5.1. Описание аудиторных контрольно-оценочных мероприятий по дисциплине модуля

5.1.1. Лекции

Самостоятельное изучение теоретического материала по темам/разделам лекций в соответствии с содержанием дисциплины (п. 1.2. РПД)

5.1.2. Лабораторные занятия

Примерный перечень тем

1. Выбор задачи анализа больших данных, поиск источников данных для анализа.
2. Планирование шагов по реализации задачи анализа больших данных на основе жизненного цикла.
3. Работа с файловой системой HDFS, загрузка данных в HDFS, выгрузка данных в локальную файловую систему.
4. Запуск задач MapReduce в кластере Hadoop. Просмотр и анализ результатов работы. Выгрузка результатов анализа больших данных в локальную файловую систему.
5. Работа в командной строке Linux.
6. Редактирование файлов в Linux с помощью редактора vi.
7. Настройка доступа на Linux машину без пароля с использованием открытого ключа.
8. Команды для работы с HDFS и YARN в командной строке Linux.
9. Алгоритмы параллельной обработки данных в Apache Spark. Граф обработки данных.
10. Работы с PySpark. Загрузка данных в программу PySpark. Запуск трансформаций и действий Spark. Сохранение результатов обработки данных в файловую систему HDFS.
11. Манипуляции с данными с помощью DataFrame API: загрузка, фильтрация, объединение, очистка.

5.2. Описание внеаудиторных контрольно-оценочных мероприятий и средств текущего контроля по дисциплине модуля

Разноуровневое (дифференцированное) обучение.

Базовый

5.2.1. Контрольная работа

Тестовые задания для контрольной работы по теме «Анализ больших данных»

1. Кто ввел термин Большие данные?
А) **Клиффорд Линч**
Б) Алан Тьюринг
В) Бьерн Страуструп
Г) Дональд Кнут

2. Какие данные занимают больше мировой памяти относительно остальных?
А) Structured Data
Б) **Unstructured Data**
В) Semi-Structured Data
Г) Quasi-Structured Data

3. BigData – это …
А) Представление фактов, понятий или инструкций в форме, приемлемой для интерпретации, или обработки.
Б) **Комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов.**
В) Колоссальный объем данных, собранных человечеством.
Г) Класс в Java, предназначенный для хранения данных от 100 Гб

4. Какая компания создала технологию MapReduce?
А) **Google**

Б) Yahoo
В) EMC
Г) Oracle

5. Данные текстовых файлов с определенными паттернами для их обработки (например, XML) являются:
А) Структурированными
Б) **Полуструктурированными**
В) Квазиструктурированными
Г) Неструктурированными

6. Данные, имеющие определенный тип, формат и структуру (например, транзакционные данные) являются:
А) **Структурированными**
Б) Полуструктурированными
В) Квазиструктурированными
Г) Неструктурированными

7. Какой язык программирования из перечисленных является наиболее важным для аналитика?
А) C++
Б) PHP
В) F#
Г) **Python**

8. Языком, на котором был разработан RabbitMQ, является:
А) Java
Б) Python
В) C++
Г) **Erlang**

9. Что из перечисленного не является средством анализа?
А) Продвинутая визуализация
Б) **Reporting**
В) Predictive Modelling
Г) Data Mining

10. Процессом создания и выбора модели для предсказания вероятности наступления некоторого события является:
А) OLAP
Б) Data Mining
В) **Predictive Modelling**
Г) Data Science

11. Что из этого не является реализацией Hadoop?
А) Google MapReduce
Б) Phoenix
В) **GreenMint**

Г) Qizmt

12. Какие из перечисленных пунктов являются достоинствами MapReduce?
А) Оптимальная производительность
Б) Эффективное применение в маленьких кластерах с небольшим объемом данных
В) **Масштабируемость**
Г) **Отказоустойчивость**

13. Что такое Oozie?
А) Распределенный координационный сервис
Б) Нереляционная распределенная база данных
В) Язык управления потоком данных и исполнительная среда для анализа больших объемов данных
Г) **Сервис для записи и планировки заданий Hadoop**

14. Сколько уровней имеет лямбда-архитектура?
А) 2
Б) **3**
В) 4
Г) 5

15. Какие компоненты являются частями MapReduce?
А) Task Tracker
Б) Name Node и Data Node
В) **Job Tracker и Task Tracker**
Г) Job Tracker, Task Tracker, Name Node и Data Node

16. Что такое Spark?
А) **Инструмент для кластерных вычислений**
Б) Графический движок
В) Библиотека для работы с графами
Г) Технология распределенных вычислений

17. Дайте определение Map Reduce…
А) **Модель распределенных вычислений, предназначенная для параллельных вычислений над очень большими (до нескольких петабайт) объемами данных**
Б) Набор компонентов и интерфейсов для распределенных файловых систем и общего ввода-вывода
В) Распределенная файловая система, работающая на больших кластерах типовых машин
Г) Распределенный сервис для коллекционирования, сбора, и перемещения больших массивов данных

18. Что из этого является недостатком MapReduce?
А) **Фиксированный алгоритм обработки данных**
Б) Масштабируемость
В) Отказоустойчивость
Г) Возможность автоматического распараллеливания

19. Какое API было добавлено в Hadoop v2.0?

А) YAWN
Б) **YARN**
В) SARN
Г) DARN

20. Какая цель у NameNode в HDFS?
А) **Хранить индекс того, какая часть данных находится в каком узле**
Б) Хранить имя файла, хранящегося в конкретном узле
В) Хранить индекс узла, в котором хранится имя файла
Г) Хранить имена узлов

Ключ к тесту:

Вопрос	Ответ
1	А)
2	Б)
3	Б)
4	А)
5	Б)
6	А)
7	Г)
8	Г)
9	Б)
10	В)
11	В)
12	В), Г)
13	Г)
14	Б)
15	В)
16	А)
17	А)
28	А)
19	Б)
20	А)

5.2.2. Домашняя работа

Работа с DataFrame в Apache Spark

Цель: рассмотреть, как использовать DataFrame API в Apache Spark на примере анализа информации о героях комиксов.

Задание:

1. Определите, какой цвет глаз наиболее популярен у героев вселенных Marvel и DC.
2. Найдите героев, у которых один глаз, а также героев, у которых нет глаз.
3. Определите, сколько персонажей комиксов живо, а сколько умерло (столбец ALIVE) в целом по всем данным, а также отдельно по вселенным Marvel и DC.
4. Определите, сколько персонажей комиксов появлялось в каждом году. Выведите ТОП 10 лет с наибольшим количеством появившихся героев комиксов.
5. Прочитайте статью Comic Books Are Still Made By Men, For Men And About Men. Согласны ли вы с выводами? Напишите запросы для получения данных для диаграмм из статьи.

Для выполнения задания необходимо установить Apache Spark и загрузить данные о супергероях Marvel и DC (ссылки на данные выдаются преподавателем). После выполнения работы студенты оформляют домашнюю работу в виде отчета.

5.3. Описание контрольно-оценочных мероприятий промежуточного контроля по дисциплине модуля

5.3.1. Экзамен

1. Определение больших данных.
2. Источники больших данных.
3. Задачи, для решения которых требуются большие данные.
4. Особенности технологий обработки больших данных.
5. Постановка задачи анализа данных.
6. Подготовка набора данных.
7. Очистка данных.
8. Выбор алгоритма анализа данных.
9. Оценка качества работы алгоритма анализа данных.
10. Принятие решений на основе данных.
11. Распределенная файловая система HDFS.
12. Архитектура файловой системы HDFS.
13. Обеспечение надежности хранения данных и производительности.
14. Экосистема Apache Hadoop.
15. Технология MapReduce.
16. Кластер Apache Hadoop.
17. Командная строка Linux.
18. Копирование/перемещение/удаление файлов. Создание/удаление/перемещение каталогов.
19. Создание файлов в Linux. Редактирование текстовых файлов.
20. Файловый менеджер.
21. Доступ к Linux по SSH. Аутентификация в SSH.
22. Использование открытых/закрытых ключей для аутентификации. Подключение сетевых дисков.
23. Кластеры Linux для обработки больших данных.
24. Утилиты Linux для работы с кластерами.
25. Команды Linux для работы с экосистемой Apache Hadoop,
26. Система обработки больших данных Apache Spark.
27. Интеграция Apache Spark с экосистемой Hadoop.
28. Отличия Spark от Map/Reduce.
29. Командная строка Spark.
30. Анализ выходные данных результатов работы программ PySpark.
31. Оптимизация производительности работы программ PySpark с DataFrame