МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Факультет информационных технологий

УТВЕРЖДЕНО
Декан факультета
Информационных технологий
_____ / Д.Г. Демидов /
«16» ____02_____ 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Большие данные»

Направление подготовки/специальность
09.03.02 Информационные системы и технологии

Профиль/специализация
Цифровая трансформация

Квалификация
Бакалавр

Формы обучения
Очная

Москва, 2023 г.

Разработчик(и):

преподаватель кафедры
«Информатика и информационные технологии» /К.М. Кононенко/

Согласовано:
Заведующий кафедрой «Информатика
и информационные технологии»,
к.т.н., доцент  /Е.В. Булатников/

Содержание

1 Цели, задачи и планируемые результаты обучения по дисциплине

Целью освоения дисциплины «Большие данные» является формирование у обучающихся практических навыков, связанных с применением в своей профессиональной деятельности современных технологий в области обработки больших данных, решать задачи сбора, организации, хранения, анализа больших данных. На практике получить навыки разработки алгоритмов, программных модулей и моделей для обработки больших данных. Полученные знания могут быть использованы в профессиональной деятельности для обработки больших данных.

К основным **задачам** освоения дисциплины следует отнести:

- ознакомление с основными видами теоретических и практических основ и навыков работы с большими данными;
- ознакомление с современными способами обработки больших данных;
- разъяснение назначения и роли об основных научно-технических проблемах и перспективах развития больших данных и их связи со смежными отраслями;
- освоение программных средств, предназначенных для обработки больших данных.

Обучение по дисциплине «Большие данные» направлено на формирование у обучающихся следующих компетенций:

Код и наименование компетенций	Индикаторы достижения компетенции
ОПК-7. Способен осуществлять выбор платформ и инструментальных программноаппаратных средств для реализации информационных систем	ИОПК-7.1 знает основные платформы, технологии и инструментальные программноаппаратные средства для реализации информационных систем ИОПК-7.2 умеет применять современные технологии для реализации информационных систем ИОПК-7.3 имеет навыки владения технологиями, применения инструментальных программно-аппаратных средств реализации информационных систем

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к базовой обязательной части Блока 1. Дисциплины (модули) учебного плана программы бакалавриата.

Дисциплина взаимосвязана логически и содержательно-методически со следующими дисциплинами и практиками ОПОП:

- Технологии программирования;
- Архитектура информационных систем;
- Базы данных
- Анализ данных;
- Производственная практика.

3 Структура и содержание дисциплины

Разделы дисциплины изучаются на 3 курсе 5 семестре.

Общая трудоемкость дисциплины составляет 3 зачетные единицы (108 часов), форма промежуточной аттестации – экзамен.

3.1 Виды учебной работы и трудоемкость (по формам обучения)

1.1.1. Очная форма обучения

№ п/п	Вид учебной работы	Количество часов	Семестры
			5
1	**Аудиторные занятия**	**54**	**54**
	В том числе:		
1.1	Лекции	18	18
1.2	Семинарские/практические занятия		
1.3	Лабораторные занятия	36	36
2	**Самостоятельная работа**	**54**	**54**
	В том числе:		
2.1	Подготовка и выполнение лабораторных работ	54	54
3	**Промежуточная аттестация**		
	Экзамен		✔
7	**Итого:**	**108**	**108**

3.2 Тематический план изучения дисциплины (по формам обучения)

1.1.2. Очная форма обучения

№ п/п	Разделы/темы дисциплины	Трудоемкость, час					
		Всего	Аудиторная работа				Самостоятельная работа
			Лекции	Семинарские/практические занятия	Лабораторные занятия	Практическая подготовка	
1.	Вводное занятие. Введение в науку о данных и большие данные.	**8**	**2**				6
2.	Характеристики больших данных и их источники.	**14**	**2**		6		6
3.	Основные задачи и методы анализа больших данных.	**14**	**2**		6		6
4.	Области работы с большими данными на одном компьютере.	**14**	**2**		6		6
5.	Распределенные и облачные хранилища больших данных.	**8**	**2**				6
6.	Архитектура экосистемы анализа и обработки больших данных Hadoop.	**14**	**2**		6		6
7.	Базы данных NoSQL.	**14**	**2**		6		6
8.	Технологии MapReduce и Spark.	**8**	**2**				6
9.	Графовые базы данных.	**14**	**2**		6		6
Итого:		**108**	**18**		**36**		**54**

3.3 Содержание дисциплины

Введение

Вводное занятие. Введение в науку от данных и большие данные. Наука о данных (Data Science). Большие данные (Big Data). «Большая триада». Пять «V» больших данных. Объем. Скорость. Разнородность. Достоверность. Ценность. Проблемы обработки и анализа данных. Эволюция данных от библии до искусственного интеллекта. Иерархия DIKW. Пирамида DIKW. Пирамида Data Science. Процесс CRISP-DM. CRISP-DM. Жизненный цикл CRISP-DM. Понимание бизнес-целей и начальное изучение данных. Подготовка данных. Моделирование. Оценка, тестирование, проверка. Внедрение CRISP-DM. Распределение времени между основными задачами обработки больших данных.

Тема 1. Характеристики больших данных и их источники.

Основные характеристики больших данных**.** Структурирование данных. Неструктурированные данные. Данные на естественном языке. Машинные данные. Графовые или сетевые данные. Аудио, видео и графика. Потоковые данные. Основные источники больших данных. Использование телематических данных. Использование данных о времени и местоположении. Данные радиочастотной идентификации (RFID). Данные, генерируемые интеллектуальными сетями. Использование больших данных с датчиков на промышленном оборудовании. Использование больших данных, полученных из социальных сетей. Открытые ист. очники больших данных для исследователей и аналитиков. World Bank Open Data data.worldbank.org. www.imf.org/en/Data Data. xVIEW. MIMIC-III. Berkeley DeepDrive BDD 100k. CREMA-D.

Тема 2. Основные задачи и методы анализа больших данных.

Три основные задачи обработки больших данных. Хранение и управление Big Data. Обработка неструктурированной информации. Основные типы задач и методы их решения при анализе больших данных. Классификация. Методы классификации. Кластеризация. Основные методы кластеризации. Регрессия. Основные методы решения задач регрессии. Анализ социальных сетей. Анализ связей на графе. Задачи анализа сетей (графов). Поиск ассоциаций. Примеры использования поиска ассоциаций. Предсказание и прогноз. Методы и модели прогнозирования. Анализ адекватности и точности построения прогноза.

Тема 3. Области работы с большими данными на одном компьютере.

Проблемы при работе с большими объемами данных. Нехватка памяти. Долгое время выполнения или «зависание». Узкие места и простаивание. Способы решения проблем. Адаптация алгоритмов. Онлайновые алгоритмы. Особенности онлайновых алгоритмов. Режимы работы онлайновых алгоритмов. Примеры онлайновых алгоритмов. Линейный метод наименьших квадратов. Метод стохастического градиентного спуска. Инкрементальный стохастический градиентный спуск. Блочные алгоритмы. Некоторые библиотеки реализации блочных алгоритмов. Алгоритмы MapReduce. Алгоритмы MapReduce на псевдокоде. Правильный выбор структуры данных. Разреженные данные. Древовидные структуры. Хеш-таблицы. Правильный выбор инструментов.

Тема 4. Распределенные и облачные хранилища больших данных.

Распределенные хранилища. Репликация данных. Облачные хранилища. Прозрачный доступ к данным. Избыточность/устойчивость системы. Разнородные вычислительные среды. Несложное перемещение данных. Администрирование базы данных/системы. Примеры распределенных и облачных хранилищ данных. Услуги, предоставляемые облачными системами. Облачные сервисы, относящиеся к большим данным. Модели работы с облаками для разных групп пользователей. SAAS. PaaS. IaaS. Как выбрать распределенное или облачное хранилище данных? Способы создания

ресурсов и управления ими в облаке. Использование WEB-портала. Использование программных библиотек (SDK). Специализированные расширения для языков командной строки. Использование шаблонов. Преимущества распределенных и облачных баз данных. Преимущества распределенных и облачных хранилищ больших данных. Недостатки распределенных и облачных хранилищ больших данных. Развитие распределенных и облачных хранилищ больших данных.

Тема 5. Архитектура экосистемы анализа и обработки больших данных Hadoop.

Hadoop. Для чего нужен Hadoop. Архитектура Hadoop. Компоненты Hadoop. Компоненты, которые вместе образуют экосистему Hadoop. Hadoop Distributed File System (HDFS). Особенности HDFS. YARN. MapReduce. Apache Spark. PIG. HIVE. Apache HBase. Mahout. Другие компоненты. Hadoop и озёра данных. Как работают Data Lake. Озёра данных VS хранилища данных. Гибридные хранилища. Гиганты использующие Hadoop. Организации обеспечивающие Hadoop решения.

Тема 6. Базы данных NoSQL.

Базовые принципы реляционных БД: acid. Теорема CAP. Треугольник cap для SQL и NoSQL субд. Пример рассогласования данных. Принципы base баз данных nosql. Базовая доступность (Basically Available). Неустойчивое состояние (Soft state). Согласованность в конечном счете (Eventual consistency). Базы данных NoSQL («Not Only SQL»). Задачи которые решает NoSQL. Базы данных Newsql. Сравнение баз данных SQL (реляционных) и NoSQL (не реляционных). Классификация баз данных nosql и newsql. Основные типы данных NoSQL. Документирование базы данных. Key-value хранилища. Семейство столбцовых баз данных. Графовые базы данных. Примеры использования баз данных NoSQL. Интернет вещей (IoT). Альтернатива больших данных (Big Data). Социальные сети и медиа. Использование NoSQL с реляционными БД. Примеры популярных систем управления базами данных NoSQL. Преимущества баз данных NoSQL. Недостатки баз данных NoSQL.

Тема 7. Технологии MapReduce и Spark.

Что такое MapReduce и Spark? Возможности MapReduce. Особенности MapReduce. Основополагающие функции вычислительной модели. Модель mapreduce. Принцип работы MapReduce. Архитектура MapReduce в HADOOP. Пример. Решение 2. Размещаем лог-флаг в hadoop. Пишем функции map и reduce. Примеры использования MapReduce. Продукты, использующие mapreduce. Особенности Spark. Как Spark решает проблемы MapReduce? Архитектура Spark. Компоненты экосистемы spark. Примеры использования Spark. Системы/компании, где применяется Spark. Запуск Spark в google colab. Сравнение MapReduce и Spark.

Тема 8. Графовые базы данных.

Что такое графовая база данных? Что такое граф? Примеры графов. Терминология графовых БД. Как использовать графовые БД. Графовые БД, сложность и размер данных. Место графовых БД среди других типов БД. Примеры использования графовых БД. Различные реализации графовых БД. Графовая субд Neo4j. Особенности Neo4j. Принципы REST API. Место Neo4j в системах управления БД. Архитектура Neo4j. Объекты в Neo4j. Пользовательский интерфейс в Neo4j. Пример модели сети от Neo4j. Интерфейсы и алгоритмы Neo4j. Язык запросов Cypher. Примеры создания графов. Преимущества графовых баз данных. Недостатки графовых баз данных.

3.4 Тематика семинарских/практических и лабораторных занятий

3.4.1 Лабораторные занятия

Лабораторная работа № 1 «Интерпретируемый язык программирования R для анализа данных». 1. Загрузите и установите консольное приложение для интерпретации команд R 2. Запуск R. 3. В этом курсе будет использовать программный пакет Swirl для R, чтобы проиллюстрировать некоторые ключевые концепции. Пакет Swirl превращает консоль R в интерактивную среду обучения. Использование Swirl также даст вам возможность исследовать данные, как это сделал бы специалист по данным. В этой лабораторной работе у вас будет возможность попрактиковаться в некоторых ключевых концепциях этого курса. 4. Для выполнения лабораторной работы установите Swirl. 5. Swirl предлагает множество интерактивных курсов, но для наших целей вам нужен курс под названием "***Exploratory Data Analysis***" (Разведочный анализ данных). 6. Для выполнения этой лабораторной работы выполните первые десять уроков курса: Principles of Analytic Graphs, Exploratory Graphs, Graphics Devices in R, Plotting Systems, Base Plotting System, Lattice Plotting System, Working with Colors, GGPlot2 Part1, GGPlot2 Part2, GGPlot2 Extras . 7. Общий лог выполнения работы и скриншоты используемых/выводимых графиков(рисунков) загрузите в качестве отчета на данную лабораторную работу, оформив в виде файла doc или pdf с титульным листом.

Лабораторная работа № 2 «Кластерный анализ данных с использованием языка R». В рамках этой лабораторной работы пройти следующие уроки в учебном курсе пакета swirl: 1. Hierarchical Clustering, 2. K Means Clustering, 3. Dimension Reduction, 4. Clustering Example. В отчет необходимо включить логи выполнения работы со скриншотами графиков.

Лабораторная работа № 3 «Анализ больших данных на локальном компьютере с помощью библиотек языка Python». 1. Реализуйте указанный код в любой среде программирования Python или используйте блокноты Google Colab или Jupyter. 2. В отчет включите листинг вашей программы и все диаграммы, представленные выше в **руководстве**. 3. Для каждой диаграммы включите в отчет Ваш личный словестный анализ/интерпретацию полученного результата: почему получен такой результат, в чем основная суть этого результата и т.п. 4. Включите в отчет дополнительный анализ на свое усмотрение - не менее 3х **диаграмм**, графиков. Например, анализ по городам, анализ по хронологии (по годам и десятилетиям (или пятилеткам), когда появлялись НЛО), а также анализ по какой-то отдельно взятой стране, например по России.

Лабораторная работа № 4 «Работа с программной платформой HADOOP» 1. Установите и настройте HADOOP по шагам, как описано в данном руководстве. 2. Протоколируйте в отчет скриншотами и комментариями каждый шаг вашей установки. 3.Выполните программу Word Count для нескольких текстовых книг. Результаты включите в отчет. 4. В папке примеров найдите еще типовые задачи и включите их выполнение в отчет.

Лабораторная работа № 5 «Основы MongoDB». 1. Установите MongoDB. 2. Произведите подключение к тестовой базе данных. Выполните тестовые примеры 1 и 2. 3. Сойздайте собственную БД и добавьте произвольные данные в БД. 4. Извлеките добавленные на предыдущем шаге данные. 5. Все результаты по пп.1-4 вностите в отчет со скриншотами, кодом и комментариями.

Лабораторная работа № 6 «Изучение основных возможностей Neo4j» 1. Установите сервер Neo4j, запустите клиента и выберите пункт: «Начать обучение» (Start Learning). Далее следуйте инструкциям (для следующей страницы нажимайте справа вверху стрелку в виде знака «>») и в результате создайте следующие графы, включите в отчет результаты. Для выполнения команд нужно щелкнуть на ячейку с командой, и она автоматически появится в редакторе команд. Далее следует просто запустить команду на выполнение (play). 2. Наберите команду **:play movie graph.** Запустится руководство по созданию графовой БД о кинофильмах. Далее по шагам следуйте инструкциям и в результате создайте БД, выполните поисковые запросы к БД, аналитические запросы по поиску закономерностей и рекомендаций. Все результаты, листинги кода и скриншоты включите в отчет. 3. Наберите команду **:play northwind graph.** Запустится руководство по конвертации данных из реляционной модели (из файлов таблиц CSV) в графовую модель. Далее по шагам следуйте инструкциям и в результате загрузите таблицы в БД, создайте индексы и связи, выполните запросы к БД. Все результаты, листинги кода и скриншоты включите в отчет. 4. Наберите команду **:play query-template.** Далее по шагам следуйте инструкциям по созданию шаблонов запросов. Все результаты, листинги кода и скриншоты включите в отчет.
5. Наберите команду **:play cypher.** Далее по шагам следуйте инструкциям по изучению основных возможностей языка Cypher. Все результаты, листинги кода и скриншоты включите в отчет.

3.5 Тематика курсовых проектов (курсовых работ)

Курсовые проекты не предусмотрены.

4 Учебно-методическое и информационное обеспечение

4.1 Нормативные документы и ГОСТы

1. Федеральный закон от 29 декабря 2012 года No 273-ФЗ «Об образовании в Российской Федерации» (с изменениями и дополнениями);
2. Федеральный государственный образовательный стандарт высшего образования - бакалавриат по направлению подготовки 09.03.02 Информационные системы и технологии, утвержденный Приказом Министерства образования и науки РФ от 19 сентября 2017 г. N 929 "Об утверждении федерального... Редакция с изменениями N 1456 от 26.11.2020;
3. Приказ Министерства образования и науки РФ от 05 апреля 2017 г. No 301 «Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры.

4.2 Основная литература

1. Рындина, С. В. Цифровая трансформация бизнеса: использование аналитики на основе больших данных : учебное пособие / С. В. Рындина. — Пенза : ПГУ, 2019. — 182 с. — ISBN 978-5-907262-04-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/162301— Режим доступа: для авториз. пользователей.
2. Кобзаренко, Д. Н. Учебное пособие дисциплины «Анализ больших данных» для направления подготовки 38.03.05 «Бизнес-информатика», профиль «Электронный бизнес» : учебное пособие / Д. Н. Кобзаренко, А. Г. Мустафаев ; составитель Д. Н. Кобзаренко. — Махачкала : ДГУНХ, 2019. — 107 с. — Текст :

электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/246542— Режим доступа: для авториз. пользователей.
3. Документация по языку R [Электронный ресурс]. – URL: https://cran.r-project.org/manuals.html (дата обращения: 09.05.2023).
4. Документация Apache Spark [Электронный ресурс]. – URL: https://spark.apache.org/docs/latest/ (дата обращения: 09.05.2023).
5. Документация Neo4j [Электронный ресурс]. – URL: https://neo4j.com/docs/(дата обращения: 09.05.2023).
6. Документация MongoDB [Электронный ресурс]. – URL: https://www.mongodb.com/docs/launch-manage/ (дата обращения: 09.05.2023).
7. Документация Hadoop, HDFS, MapReduce, YARN [Электронный ресурс]. – URL: https://hadoop.apache.org/docs/stable/ (дата обращения: 09.05.2023).
8. Большие данные в образовании [Электронный ресурс]. – URL: http://www.unkniga.ru/vishee/9614-bolshie-dannye-vobrazovanii.html (дата обращения: 09.05.2023).
9. Большие данные (Big Data) [Электронный ресурс]. – URL: https://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data) (дата обращения: 09.05.2023).

4.3 Дополнительная литература

1. **Силен Дэви, Мейсман Арно, Али Мохамед.** Основы Data Science и Big Data. Python и наука о данных. - СПб.: Питер, 2017. - 336 с.
2. **Ын Анналин, Су Кеннет.** Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. — СПб.: Питер, 2019. — 208 с.
3. **Дейтел Пол, Дейтел Харви.** Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.
4. **Брюс П.** Практическая статистика для специалистов Data Science: Пер. с англ. /П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.
5. **Дж. Д. Лонг и Пол Титор.** R. Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации данных / пер. с анг. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 510 с.

4.4 Электронные образовательные ресурсы

Курс «Большие данные» в СДО Мосполитеха:
https://online.mospolytech.ru/course/view.php?id=12507

4. 5 Лицензионное и свободно распространяемое программное обеспечение

1. Google Colab;
2. Visual Studio Code (свободная лицензия);
3. R (свободная программная среда);
4. Hadoop (свободная лицензия);
5. MongoDB (свободная лицензия);
6. Neo4j Desktop (свободная лицензия).

4.6. Современные профессиональные базы данных и информационные справочные системы

1. ОП "Юрайт" https://urait.ru/
2. IPR Smart https://www.iprbookshop.ru/
3. ЭБС "Лань" https://e.lanbook.com/

5 Материально-техническое обеспечение

Аудитории общего фонда для лекционных и лабораторных занятий, Москва, ул. Прянишникова, д. 2а со следующей оснащенностью: столы, стулья, аудиторная доска, использование переносного мультимедийного комплекса (переносной проектор, персональный ноутбук). Персональные компьютеры, мониторы, манипулятор «мышь», клавиатуры. Рабочее место преподавателя: стол, стул.

Лицензионное программное обеспечение: Microsoft Windows 10/11, Microsoft Office.

6 Методические рекомендации

6.1 Методические рекомендации для преподавателя по организации обучения

Методика преподавания дисциплины «Большие данные» предусматривает использование онлайн-курса в системе дистанционного обучения Университета, изучение теоретической части в процессе посещения лекционных онлайн занятий, групповых и индивидуальных консультаций обучающихся, аудиторных занятий в сочетании с внеаудиторной работой с целью формирования и развития профессиональных навыков использования современных информационных технологий для решения задач профессиональной деятельности.

Лабораторные работы по дисциплине «Большие данные» осуществляются в форме самостоятельной проработки теоретического материала обучающимися; выполнения практического задания; защиты преподавателю лабораторной работы (знание теоретического материала и выполнение практического задания по теме лабораторной работы).

6.2 Методические указания для обучающихся по освоению дисциплины

Изучение дисциплины осуществляется в соответствии с учебным планом.

На занятиях осуществляется закрепление полученных, в том числе и в процессе самостоятельной работы, знаний. Особое внимание обращается на умение применять полученные знания на практике, в том числе при решении реальных задач, отличающихся от проработанных.

В процессе самостоятельной работы студенты закрепляют и углубляют знания, полученные во время аудиторных занятий, дополнительно изучают лекционный теоретический материал, выполняют индивидуальные задания на лабораторных занятиях, оформляют отчеты по выполненным работам, готовятся к текущему контролю и промежуточной аттестации.

Текущий контроль осуществляется на аудиторных занятиях в форме тестирования в системе дистанционного обучения Университета.. Критериями оценки результатов являются:

- уровень освоения теоретического материала;
- уровень владения практическими навыками (в виде вопросов по процессу выполнения практических заданий);
- умения обучающегося использовать теоретические знания при выполнении практических задач (в виде дополнительных заданий);
- сформированность компетенций;

– оформление материала в соответствии с требованиями.

Промежуточный контроль осуществляется на аудиторных занятиях в форме тестирования в системе дистанционного обучения Университета..

7 Фонд оценочных средств

7.1 Методы контроля и оценивания результатов обучения

В процессе обучения используются следующие оценочные формы самостоятельной работы студентов, оценочные средства текущего контроля успеваемости и промежуточных аттестаций: **лабораторные работы, экзамен**.

7.2 Шкала и критерии оценивания результатов обучения

К промежуточной аттестации допускаются только студенты, выполнившие все виды учебной работы, предусмотренные рабочей программой по дисциплине «Большие данные».

7.2.1 Критерии оценки ответа на экзамене

(формирование компетенций — ОПК-7)

«Отлично»:

Выполнены все виды учебной работы, предусмотренные учебным планом. Обучающийся выполнил и защитил лабораторные работы со средним баллом от 4,5 до 5. Итоговое тестирование выполнено на 85 — 100%. Обучающийся демонстрирует прочные теоретические знания, практические навыки, владеет терминами, делает аргументированные выводы и обобщения, приводит примеры, оперирует приобретенными знаниями, умениями, навыками, применяет их в ситуациях повышенной сложности. При этом могут быть допущены незначительные ошибки, неточности, которые обучающийся может исправить самостоятельно.

«Хорошо»:

Выполнены все виды учебной работы, предусмотренные учебным планом. Обучающийся выполнил и защитил лабораторные работы со средним баллом от 4 до 4,5. Итоговое тестирование выполнено на 70 — 84%. Обучающийся демонстрирует достаточные теоретические знания, практические навыки, владеет терминами, делает аргументированные выводы и обобщения, приводит примеры, оперирует приобретенными знаниями, умениями, навыками. При этом могут быть допущены незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации, которые обучающийся может исправить при незначительной коррекции преподавателем.

«Удовлетворительно»:

Выполнены все виды учебной работы, предусмотренные учебным планом. Обучающийся выполнил и защитил лабораторные работы со средним баллом ниже 4. Итоговое тестирование выполнено на 55 — 69%. Обучающийся демонстрирует неполное соответствие теоретических знаний, практических навыков, владеет терминами, делает аргументированные выводы и обобщения, приводит примеры, оперирует приобретенными знаниями, умениями, навыками. При этом могут быть допущены ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации, которые обучающийся может исправить при коррекции преподавателем.

«Неудовлетворительно»:

Не выполнен один или более видов учебной работы, предусмотренных учебным планом. Обучающийся не выполнил одно или более заданий текущего и промежуточного

контроля. Итоговое тестирование выполнено на 0 — 54%. Обучающийся демонстрирует незнание теоретических основ предмета, отсутствие практических навыков, не умеет делать аргументированные выводы и приводить примеры, не владеет терминами, проявляет отсутствие логичности и последовательности изложения, делает ошибки, которые не может исправить даже при коррекции преподавателем, отказывается отвечать на дополнительные вопросы, допускает значительные ошибки, испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.

7.2.2 Критерии оценки работы обучающегося на лабораторных и семинарских (практических) занятиях:

(формирование компетенций — ОПК-7)

«5» (отлично): выполнены все практические задания, предусмотренные лабораторными работами, обучающийся четко и без ошибок ответил на все контрольные вопросы, проявил творческий подход при выполнении заданий, смог выполнить дополнительные задания.

«4» (хорошо): выполнены все практические задания, предусмотренные лабораторными работами, обучающийся с корректирующими замечаниями преподавателя ответил на все контрольные вопросы, проявил творческий подход при выполнении заданий, смог частично выполнить дополнительные задания.

«3» (удовлетворительно): выполнены все практические задания, предусмотренные лабораторными работами, с замечаниями преподавателя; обучающийся ответил на все контрольные вопросы с замечаниями, дополнительные задания выполнены с замечаниями.

«2» (неудовлетворительно): обучающийся не выполнил или выполнил неправильно практические задания, предусмотренные лабораторными работами, обучающийся ответил на контрольные вопросы с ошибками или не ответил на контрольные вопросы, дополнительные задания выполнены неверно или не выполнены.

В процессе обучения используются следующие оценочные формы самостоятельной работы студентов, оценочные средства текущего контроля успеваемости и промежуточных аттестаций: **тестирование, экзамен.**

7.3 Оценочные средства

7.3.1 Текущий контроль

Текущий контроль осуществляется на аудиторных занятиях в форме тестирования в системе дистанционного обучения Университета. Лабораторные занятия – средство контроля усвоения учебного материала темы, раздела или разделов дисциплины, организованное как учебное занятие в виде демонстрации полученных навыков при решении поставленных практических задач.

Примеры тестовых заданий к защите лабораторных работ (оцениваемые компетенции — ОПК-7).

Тест по теме 1:

1. К проблемам обработки и анализа данных относятся: …

- формат, отличный от традиционных баз данных

- четкая структурированность данных, не позволяющая обработать разноформатные данные
- отсутствие нужных инструментов, чтобы связать данные между собой
- хранение больших вычислительных мощностей

2. Тест Тьюринга умышленно исключает прямое когнитивное взаимодействие между тестирующим и компьютером …

- Верно
- Неверно

3. Ключевые характеристики Big Data с условным наименованием «5V»: …

- Value
- Voice
- Volume
- Veracity
- Variety

4. На этапе внедрения изучается: ...

- Каким образом можно развернуть выбранные модели в бизнес-среде
- Соответствует ли модель целям процесса?
- Как интегрировать модели в бизнес-процессы организации
- Как интегрировать модели в техническую инфраструктуру

5. Соответствие распределения времени между основными задачами обработки больших данных

- построение обучающих моделей
- очистка и организация данных
- анализ данных для выявления закономерностей
- сбор данных
- уточнение алгоритмов
- другие задачи

Тест по теме 2:

1. Особенности Электронной библиотеки Всемирного Банка: …

- содержит Обширный набор экономических, социальных и экологических показателей
- включает показатели по странам за период с 1930 года
- предоставляет свободный и открытый доступ к данным о развитии в странах по всему миру
- является полнотекстовой базой статистических, аналитических анализов Всемирного Банка

2. Основные источники больших данных: …

- статистика

- реляционные базы данных
- показания считывающих устройств
- интернет

3. Верные заключения о больших данных: …

- интеллектуальные сети обеспечивают возможность коммунальным предприятиям гораздо лучше управлять их энергосистемами
- текстовые данные не являются типами данных, применимыми в больших данных
- со многими источниками больших данных связаны проблемы соблюдения конфиденциальности
- одни и те же базовые технологии могут быть использованы в различных отраслях для решения различных задач

4. Особенности данных, генерируемых интеллектуальными сетями: …

- предполагает наличие сложных систем мониторинга, связи и генерации энергии
- уступают по надежности традиционным линиям электропередач
- обеспечивают более надежное обслуживание и восстановление после отключения питания
- домовладелец может проверить, какую мощность потребляют приборы, включив их по очереди

Тест по теме 3:

1. Особенности хранения и управления Big Data: …

- %-100 Big Data всегда хранятся и организуются в централизованных файловых системах
- информация хранится на нескольких (иногда тысячах) жестких дисках, на стандартных компьютерах
- для обеспечения отказоустойчивости и надежности, каждую часть информации сохраняют несколько раз
- Big Data обычно хранятся и организуются в распределенных файловых системах

2. Три основные задачи, связанные с большими данными: …

- Хранение и управление
- Совершенствование методов анализа Big Data
- Обеспечение атомарности RDBMS
- Обработка неструктурированной информации

3. Недостаток неструктурированных данных в Big Data: …

- сохранение “всех данных”, независимо от того, какая часть данных актуальна для последующего принятия решения
- сохранение “всех данных”, независимо от того, какая часть данных актуальна для последующего принятия решения
- для извлечения полезной информации требуется последующая обработка этих огромных массивов данных

4. Метод классификации больших данных, основанный на таблице частот, при котором алгоритм разбивает датасет на все меньшие куски данных, формируя тем самым дерево, называется: …

- ZeroR
- Decision Tree
- OneR
- Naive Bayesian

Тест по теме 4:

1. Библиотека, используемая при работе с большим данными для повышения скорости выполнения кода, компилирующая код непосредственно перед выполнением (JIT-компиляция), называется …

- Numexpr
- Hadoopy
- NumPy
- Numba

2. Особенности баз данных на древовидных структурах: …

- Индексы строятся на базе деревьев и хеш-таблицах
- 100просматривают содержимое таблиц от первой строки до последней
- используют индекс для ускорения поиска

3. Типы алгоритмов, при работе с большими данными, оптимизирующие процесс обработки и анализ больших данных: …

- блочные алгоритмы
- алгоритмы MapReduce
- онлайновые алгоритмы
- структурные алгоритмы

4. Особенности «Узких мест» и простаивания при работе с большими данными: …

- SSD работают медленнее, чем HDD
- Некоторые системы простаивают, т.к. компоненты компьютера создают «узкие места»,
- SSD работают быстрее центрального процессора
- Некоторые программы не успевают быстро поставлять данные процессору, т.к. читают данные с HDD

5. Примеры онлайновых алгоритмов: …

- Линейный метод наименьших квадратов
- Онлайновая выпуклая оптимизация
- Регрессивное обучение
- Метод стохастического градиентного спуска
- Инкрементальный стохастический градиентный спуск

Тест по теме 5:

1. Примеры распределенных и облачных хранилищ данных: …

- APACHE HADOOP
- MICROSOFT ACCESS
- GOOGLE CLOUD STORAGE
- AMAZON S3

2. Особенности распределенных хранилищ данных: …

- предусмотрено централизованное хранение данных
- данные расположены на нескольких компьютерах в разных местах
- компьютерная сеть, в которой информация реплицируемым образом
- информация хранится в распределенной базе данных

3. Для подключения специализированных расширений для языков командной строки к облачным ресурсам: …

- не требуется дополнительных действий
- необходимо импортировать ключи
- необходимо выполнить вход в аккаунт через форму ввода логина/пароля

4. Системы, применяющие шаблон для конфигурирования различных инфраструктур: …

- Ansible
- Chef
- GitHub
- Puppet

Тест по теме 6:

1. Компоненты Hadoop: …

- YARN
- MapReduce
- Dask
- HDFS

2. Особенности HDFS: …

- Отказоустойчивая
- Файловая система на основе JavaScript
- Надежная и экономичная
- Масштабируемая

3. Hadoop состоит из двух основных компонентов: …

- узла каталога
- узла данных
- формата узла

– имени узла

4. Особенности HIVE: …

– 30%имеет два компонента - драйверы и командную строку
– не позволяет обрабатывать данные реальном времени
– 40%осуществляет пакетную обработку
– 30%поддерживает все типы данных SQL

Тест по теме 7:

1. Основные принципы BASE баз данных NOSQL: …

– Soft state
– Basically Available
– Eventual consistency
– Atomicity

2. Принцип при котором: если блок данных включается в базу, то он либо включается полностью, либо не включается вообще, называется: …

– Atomicity
– Durability
– Consistency
– Isolation

3. Задачи, которые решает NoSQL: …

– приводит все БД таблицы к нормальной форме
– анализ и хранение больших объемов данных в распределенной среде
– Задачи, связанные с доступом к большим объемам разнородных данных
– Задачи, связанные с обработкой больших объемов разнородных данных

4. Характерные свойства ACID для реляционной базы данных: …

– атомарность
– горизонтальное масштабирование
– изолированность
– непротиворечивость

5. Способность системы, сети или процесса справляться с увеличением рабочей нагрузки при добавлении ресурсов, называется …

– Масштабируемость
– Согласованность данных
– Изолированность
– Атомарная операция

Тест по теме 8:

1. Технология MapReduce может использоваться для: …

- счётчиков частоты обращений к заданному адресу
- индексации веб-контента
- подсчета слов в большом файле
- вычисления объема памяти

2. Функция Reduce: …

- каждой итерации заданной функции передаются новый элемент списка
- принимает на вход список значений
- применяет к каждому элементу списка некую функцию и возвращает новый список
- преобразует список к единственному атомарному значению

3. При возможности использования MapReduce и Spark наилучшим считается: …

- отдать предпочтение MapReduce
- отдать предпочтение одному из них в зависимости от задачи
- совместное использование
- отдать предпочтение Spark

4. Особенности Spark

- разработан компанией Apache
- занимается хранением файлов и управлением ресурсами
- является взаимодополняющей системой к Hadoop
- инфраструктура анализа больших данных и кластерных вычислений

Тест по теме 9:

1. Принципы написания кода интерфейсов: …

- Отсутствие записи состояния клиента
- Отделение клиента от сервера
- Кэшируемость
- Одноуровневость системы

2. Графовая база данных использует модель графа, чтобы представить данные в виде: …

- ребер
- атомов
- вершин
- узлов
- связей

3. Принцип REST API, в котором отражено, что сервера могут располагаться на разных

уровнях, при этом каждый сервер взаимодействует только с ближайшими уровнями и не связан запросами с другими, называется …

- Stateless
- Layered System
- Uniform Interface
- Client-Server

7.2.3 Промежуточная аттестация

Промежуточная аттестация обучающихся в форме экзамена осуществляется по результатам выполнения всех видов учебной работы, предусмотренных учебным планом по данной дисциплине, при этом учитываются результаты текущего контроля успеваемости в течение семестра. Экзамен проводится в форме тестирования в системе дистанционного обучения Университета или в форме ответа на вопросы экзаменационного билета. По итогам промежуточной аттестации по дисциплине выставляется оценка «отлично», «хорошо», «удовлетворительно» или «неудовлетворительно».

Примеры тестовых заданий промежуточного контроля (оцениваемые компетенции — ОПК-7).

Итоговый тест:

1. Базовый принцип реляционных БД, при котором если блок данных включается в базу, то он либо включается полностью, либо не включается вообще, называется: …
   - Долгосрочность
   - Атомарность
   - Изолированность
   - Согласованность

2. Принципы BASE баз данных NOSQL: …
   - Согласованность в конечном счете
   - Атомарность
   - Неустойчивое состояние
   - Базовая доступность

3. Задачи хранения и управления больших данных связаны с тем, что: …
   - работа с Big Data не позволяет использовать NoSQL
   - Big Data работает только со структурированной информацией
   - Big Data работает не может работать со структурированной информацией
   - большой объем данных не позволяет легко хранить и управлять ими с помощью реляционных БД

4. При отсутствии функции GPS сотовые телефоны: …
   - не смогут определить местоположение
   - достаточно точно определяют местоположение
   - определяют местоположение с большой погрешностью
   - используют сигналы базовых станций операторов мобильной связи для определения местоположение

5. Особенности Hadoop: …
- популярный фреймворк
- содержит платный набор утилит
- разработан на Java
- предназначен для разработки и выполнения централизованных программ

6. Сравнение MapReduce и Spark: …
- Spark обладает более широким спектром возможностей для работы с данными
- MapReduce имеет более статическую архитектуру
- Преимущество Spark – линейная обработка огромных наборов данных
- Spark имеет более динамическую архитектуру

7. Целостность и внутренняя непротиворечивость данных, называется …
- Атомарная операция
- Масштабируемость
- Изолированность
- Согласованность данных

8. Процесс изучения наборов данных с целью получения выводов о содержащейся в них информации, все чаще с помощью специализированных систем и программного обеспечения, называется: …
- Data Science
- Expert system
- Data Analytics
- Анализ данных

9. Для хранения и построения запросов графовых данных используются: …
- SQL
- SPARQL
- графовые базы данных
- реляционные базы данных

10. К принципам REST API относятся: …
- Casheable
- Starting with the Null Style
- Stateless
- Interface Laurence Null

11. Фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов …
- Hadoop
- Reduce
- YARN
- HDFS

12. Примеры неструктурированных данных: …
- посты в соц. сетях
- электронный справочник
- сообщения электронной почты

- реляционная база данных

13. Особенности графовых баз данных: …
    - эффективно работают с данными, имеющими сложную структуру
    - полезны в областях, связанных с социальными сетями и логистикой
    - предоставляют эффективный способ хранения и обработки связанных данных
    - обычно применяются для работы с данными, имеющими простую структуру

14. Компоненты, которые образуют экосистему Hadoop: …
    - HDFS
    - YARN, MapReduce
    - SSD
    - HBase

15. Решение задачи прогнозирования сводится к решению подзадач: …
    - анализ точности прогноза
    - анализ гомогенности
    - выбор модели прогнозирования
    - анализ адекватности прогноза

16. Типы данных для больших данных в: …
    - структурированные
    - неструктурированные
    - квантовые
    - на естественном языке

17. Базовый принцип Acid, при котором: когда в базе данных что-то изменяется, ничего не может происходить точно с одними и теми же данными точно в один момент …
    - Согласованность
    - Долгосрочность
    - Изолированность
    - Атомарность

18. xVIEW – это: …
    - обезличенные медицинские данные пациентов
    - датасеты о финансах и ценах на товары
    - содержание изображения сложных сцен со всего мира, аннотированные с помощью ограничительных рамок
    - набор воздушных снимков Земли с аннотациями

19. В REST API: …
    - код запросов остается на стороне клиента
    - код запросов и код для доступа к данным всегда на стороне сервера
    - код запросов и код для доступа к данным всегда на стороне сервера
    - код для доступа к данным на стороне сервера

20. К основным методам кластеризации относятся: …
    - Полиномиальная кластеризация
    - Послойная кластеризация
    - Минимальное покрывающее дерево

– Выделение связных компонент

21. Основные методы кластеризации: …
– Минимальное покрывающее дерево
– Гребневая
– Иерархический
– k-средних

22. Ценность информации, как ключевой параметр для оценки эффективности вложений в ее обработку, включает ответы на вопросы: …
– Дают ли собираемые данные ответы на поставленные вопросы?
– Какой объем данных нужно собрать за единицу времени?
– Оправдываются ли затраты на внедрение аналитических механизмов и систем?
– Способна ли компания извлекать пользу из собираемых данных?

23. Информационная иерархия, в которой каждый следующий уровень характеризуется большим уровнем зрелости (пригодностью к продолжению жизни) и кратно меньшим объёмом сведений, называется ...
– CRISP-DM
– JSON
– Data Science
– DIKW

24. При работе с большими данными нехватка оперативной памяти выражается в том, что ОС начинает выгружать блоки памяти на диск и: …
– скорость обработки данных не меняется
– скорость работы с данными резко растет
– ОС резко виснет
– скорость работы с данными резко падает

25. Динамично развивающееся направление облачных вычислений, ориентированное прежде всего на веб-разработчиков, называется …
– SaaS
– IaaS
– ZaaS
– PaaS

26. Режимы работы онлайновых алгоритмов: …
– мини-пакетное обучение
– полнопакетное обучение
– полупакетное обучение
– онлайновое обучение

27. Специализированные расширения для языков командной строки: …
– shell
– Pyton
– CMD
– AWS CLI

28. Не относится к основным методам анализа больших данных …
- прогнозирование
- лемматизация
- классификация
- кластеризация

29. Примеры использования поиска ассоциаций: …
- анализ посещений веб-страниц
- регрессивный анализ
- анализ ДНК живых организмов
- анализ рыночной корзины

30. Особенности использования программных библиотек …
- программное манипулирование ресурсами облака выполняется принципалом
- SDK должна содержать ключи учетной записи, которая будет иметь доступ к облаку
- SDK представляет собой набор классов и методов для работы с ресурсами облака
- невозможность создания программ, которые сами себе создают облачные ресурс

31. Распределенные и облачные хранилища данных: …
- Яндекс.Диск
- Dropbox
- MySQL
- OneDrive

32. Библиотека, используемая при работе с большим данными для повышения скорости выполнения кода, компилирующая код непосредственно перед выполнением (JIT-компиляция), называется …
- Numba
- Numexpr
- NumPy
- Hadoopy

33. Учитывая “Разнородность” данных самым важным считается: …
- объединить данные в структуру, поддающуюся анализу
- добиться однообразия данных, переводом их в буквенные знаки или цифры
- добиться взаимосвязанной структуры данных
- объединить разнородные данные в общую структуру

34. Программы, использующие MapReduce: …
- исполняются на распределенных узлах кластера
- исполняются на центральном узле последовательно
- автоматически централизуются
- автоматически распараллеливаются

35. Библиотека, используемая при работе с большим данными, помогающая решить проблемы нехватки памяти, которые могут возникнуть при использовании NumPy, позволяющая сохранять массивы и работать с ними в оптимальной сжатой форме, называется …
- Blaze

- Numba
- Bcolz
- Numexpr

36. Онлайновые алгоритмы: …
- алгоритмы, не работающие с внешней памятью
- в которых данные образуются как функция от времени
- являются общей техникой, когда невозможна тренировка по всему набору данных
- противоположны пакетной технике обучения

37. На первых двух этапах – понимания бизнес-целей и начального изучения данных – специалист: ...
- создает наборы данных, которые можно использовать для анализа
- знакомится с данными
- осуществляет интеграцию источников из нескольких баз данных
- пытается сформулировать цели проекта с точки зрения бизнеса

38. Теорема CAP. Если БД распределена по разным серверам, то она может обеспечить не более двух из трёх следующих свойств (найдите все 3 свойства): …
- изолированность
- согласованность
- доступность
- долгосрочность
- устойчивость к разделению

39. Функция MapReduce, выполняющая сортировку и фильтрацию данных, организуя их в виде группы, генерирующая результат на основе пары ключ-значение, называется …
- Oozie
- Reduce
- HBase
- Map

40. Область, которая относится к коллективным процессам, теориям, концепциям, инструментам и технологиям, которые позволяют просматривать, анализировать и извлекать ценные знания и информацию из необработанных данных, называется: …
- Expert system
- Data Analytics
- Data Science
- Наука о данных

41. Архитектура Spark состоит из следующих основных компонентов: …
- Spark RDD
- Spark Duo
- Spark Core
- Spark SQL

42. К основным методам решения задачи регрессии относятся: …
- Ридж
- Дискриминационная
- Линейная

– Полиномиальная

43. Большая часть собранной информации Big Data в распределенной файловой системе состоит из: …

– неструктурированных данных
– текста, изображений, фотографий или видео
– реляционных таблиц
– структурированных данных

44. Лучшими, считаются модели: ...

– которые плавно вписываются в существующую практику
– ориентированные на широкий круг пользователей, столкнувшихся разными проблемами
– ориентированные на конкретных пользователей, столкнувшихся с четко обозначенной проблемой