**Паспорт Образовательной программы**

**«BIG DATA в медицине»**

|  |  |
| --- | --- |
| **Версия программы** | **1** |
| **Дата Версии** | **15.10.2020** |

1. **Сведения о Провайдере**

|  |  |  |
| --- | --- | --- |
| 1.1 | Провайдер | **АНО ДПО "СофтЛайн Эдюкейшн"** |
| 1.2 | Логотип образовательной организации |  |
| 1.3 | Провайдер ИНН | **7736228783** |
| 1.4 | Ответственный за программу ФИО | **Максименко Денис Владиславович** |
| 1.5 | Ответственный должность | Руководитель проектного офиса |
| 1.6 | Ответственный Телефон | +7-495-232-0023 доб. 0889 |
| 1.7 | Ответственный Е-mail | Edu2035@softline.com |

1. **Основные Данные**

|  |  |  |
| --- | --- | --- |
| **№** | **Название** | **Описание** |
| 2.1 | Название программы | **BIG DATA в медицине** |
| 2.2 | Ссылка на страницу программы | <https://go.teachbase.ru/course_sessions/vhodnoe-testirovanie-kurs-big-data-v-meditsine-11-01/apply> - входное тестирование  <https://go.teachbase.ru/course_sessions/big-data-v-meditsine-11-01/apply> - страница программы |
| 2.3 | Формат обучения | Заочный с применение дистанционных образовательных технологий |
|  | Подтверждение от ОО наличия возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа | Подтверждаем возможность реализации образовательной программы с применением дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа |
| 2.4 | Уровень сложности | Базовый |
| 2.5 | Количество академических часов | 72 академических часа |
|  | Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы | 57 ак. часов отведено практическим занятиям и выполнению практических заданий в режиме самостоятельной работы, что составляет около 79,2 % трудоёмкости учебной деятельности |
| 2.6 | Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг | 29 900 руб.   1. <https://netology.ru/programs/web-developer?utm_source=advcake&utm_medium=cpa&utm_campaign=affiliate&utm_content=shapolovalexey&utm_term=8568b5dc891a4da085cd42cd1e42e274&stop=1#/>   54 000 руб.   1. <https://loftschool.com/course/web-beginner/>   90 000 руб.   1. <https://teachline.ru/courses/kak-sozdat-sayt-samostoyatelno/?partner=m48x3591zj0>   110 000 руб.  Аналогичные русскоязычные программы повышения квалификации отсутствуют, имеются программы/курсы на зарубежных образовательных площадках (Coursera), курсы от облачных вендоров и университетские программы для бакалавриата или магистратуры. Однако и эти программы на английском языке и не ориентированы прицельно на повышение квалификации специалистов по Большим данным в медицине |
| 2.7 | Минимальное количество человек на курсе | 1 |
| 2.8 | Максимальное количество человек на курсе | 2000 |
| 2.9 | Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе | отсутствует |
| 2.10 | Формы аттестации | **тестирование** |
|  | Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей | Большие данные |

1. **Аннотация программы**

Большие Данные (Big Data) – одно из направлений федерального проекта «Цифровые технологии», запущенного в рамках национальной программы «Цифровая экономика Российской Федерации». Системы класса Больших Данных позволяют организовывать сбор, хранение и обработку как структурированных, так и неструктурированных данных, практически в неограниченных объемах. Это позволяет создавать эффективные реестры медицинских данных, включая хранилища медицинских изображений, с целью построения сервисов, облегчающих и повышающих качество труда медицинского персонала, а также повышающих эффективность проведения исследовательских работ. Программа повышения квалификации направлена на формирование профессиональной компетенции «Способен ставить и решать прикладные задачи по подготовке медицинских данных для проведения аналитических работ по исследованию больших данных с использованием инструментов Hadoop Common». Для успешного освоения программы обучающимся необходимо владеть базовыми компетенциями цифровой грамотности, в том числе навыками владения ПК и ОС Linux на уровне среднего или продвинутого пользователя; компетенции в области алгоритмизации, базовыми знаниями в области языков программирования Java, Python и SQL. В процессе обучения по программе обучающимся рекомендуется просмотреть и прослушать все видеолекции и видеоролики по практическим занятиям, выполнить по каждой теме обязательные практические задания, проработать самостоятельно дополнительные материалы и выполнить задания для самостоятельной работы, после этого необходимо выполнить тесты по каждой теме и для успешного завершения обучения выполнить итоговый тест (70% верных ответов). Обучающиеся в результате освоения программы: будут знать сущность понятия «большие данные» и примеры применимости больших данных в медицине, концепцию распределенных вычислений MapReduce и особые техники использования MapReduce, компонент программно-технических архитектур Hadoop и компоненты распределенной файловой системы HDFS, особенности реляционных и нереляционных баз данных и примеры этих баз данных в медицине; существующие приложения Hadoop и интерфейсы взаимодействия с ними; подходы к извлечению знаний из данных по медицине; подходы и технологии обработки и анализа текстовых данных в медицине; будут уметь использовать возможности имеющейся в медицинской организации технологической инфраструктуры больших данных, выбранные инструментальные средства для хранения, преобразования и обработки больших данных в медицине, а также облачные варианты технологической инфраструктуры больших данных применительно к медицине; будут иметь практический опыт работы с распределённой файловой системой HDFS; написания программного кода на Java и на Python для решения MapReduce задач применительно к медицине; запуска MapReduce задач с помощью YARN; написания SQL-подобных запросов для создания выборок в Hive; написания скриптов Pig Latin для подготовки медицинских данных и проведения аналитических работ. В качестве бонуса в курсе предлагается развёртывание виртуальной машины с предустановленной инфраструктурой обработки больших данных.

Результаты обучения по программе будут востребованы при решении профессиональных задач по подготовке медицинских данных для проведения аналитических работ по исследованию Больших данных. Данные исследования являются актуальными в области доказательной медицины, в фармацевтике и фармакологии, могут составлять один, либо нескольких этапов комплексных проектов медицинской сферы. Полученная компетенция также может быть востребована в проектах цифровизации отрасли медицины при участии медицинских информационно-аналитических и специализированных информационно-вычислительных центров, как в целом, так и конкретных служб и/или подразделений.

**Автономная некоммерческая организация**

**дополнительного профессионального образования**

**«СофтЛайн Эдюкейшн»**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
|  | УТВЕРЖДАЮ:  Генеральный директор  АНО ДПО «СофтЛайн Эдюкейшн»  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.Э. Разуваев  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г. |

**Дополнительная профессиональная программа повышения квалификации**

**«BIG DATA в медицине»**

72 часа

Москва, 2020

**ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ**

**1.Цель программы**

Формирование профессиональных компетенций, необходимых для постановки и решения прикладных задач по подготовке медицинских данных для проведения аналитических работ по исследованию Больших данных с использованием инструментов Hadoop Common.

**2.Планируемые результаты обучения:**

2.1.Знание (осведомленность в областях)

2.1.1 сущности понятия «Большие данные» и примеров применимости Больших данных в медицине;

2.1.2 концепции распределенных вычислений MapReduce;

2.1.3 особых техник использования MapReduce;

2.1.4 компонент программно-технических архитектур Hadoop;

2.1.5 компонент распределенной файловой системы HDFS;

2.1.5 особенностей реляционных и нереляционных базах данных и примеров применения этих баз данных в медицине;

2.1.6 существующих приложений Hadoop и интерфейсах взаимодействия с ними;

2.1.7 существующих походов извлечения знаний из данных по медицине;

2.1.8 подходов и технологий обработки и анализа текстовых данных в медицине.

2.2. Умение (способность к деятельности)

2.2.1 использовать возможности имеющейся в медицинской организации технологической инфраструктуры Больших данных;

2.2.2 использовать выбранные инструментальные средства для хранения, преобразования и обработки Больших данных в медицине;

2.2.3 использовать облачные варианты технологической инфраструктуры Больших данных применительно к медицине;

2.3.Навыки (использование конкретных инструментов)

2.3.1 работы с распределённой файловой системой HDFS;

2.3.2. написания программного кода на Java для решения MapReduce задач (применительно к медицине);

2.3.3. написания программного кода на Python для решения MapReduce задач (применительно к медицине);

2.3.4. запуска MapReduce задач с помощью YARN;

2.3.5. написания SQL-подобных запросов для создания выборок в Hive;

2.3.6. написания скриптов Pig Latin для подготовки медицинских данных и проведения аналитических работ.

**3.Категория слушателей** (возможно заполнение не всех полей)

* 1. Образование: среднее профессиональное, высшее
  2. Квалификация: уровень квалификации не ниже 6
  3. Наличие опыта профессиональной деятельности: обработка и анализ данных
  4. Предварительное освоение иных дисциплин/курсов/модулей:

**4.Учебный план программы «BIG DATA в медицине»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль** | **Всего, час** | **Виды учебных занятий** | | |
| **лекции** | **практические занятия** | **самостоятельная работа** |
|  | Модуль 1. Введение в Большие данные в медицине | **5** | 1 | 3 | 1 |
|  | Модуль 2. Введение в Hadoop | **5** | 1 | 2 | 2 |
|  | Модуль 3. Развертывание Hadoop | **6** | 1 | 4 | 1 |
|  | Модуль 4. Ядро Hadoop | **19** | 4 | 11 | 4 |
|  | Модуль 5. Написание скриптов Hadoop | **7** | 2 | 4 | 1 |
|  | Модуль 6. Базы данных Hadoop | **11** | 3 | 6 | 2 |
|  | Модуль 7. Введение в Data Lake | **4** | 1 | 2 | 1 |
|  | Модуль 8. Интеллектуальный анализ Больших данных в медицине | **14** | 2 | 8 | 4 |
| **Итоговая аттестация** | | **1** | **Тест, зачет** | | **1** |
|  | | **72** | **15** | **40** | **17** |

**5.Календарный план-график реализации образовательной программы «BIG DATA в медицине»**

(дата начала обучения – дата завершения обучения) в текущем календарном году, указания на периодичность набора групп (не менее 1 группы в месяц)

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Наименование учебных модулей** | **Трудоёмкость (час)** | **Сроки обучения** |
| **1** | Модуль 1. Введение в Большие данные в медицине | **5** | **01.11** |
| **2** | Модуль 2. Введение в Hadoop | **5** | **02.11** |
| **3** | Модуль 3. Развертывание Hadoop | **6** | **03.11-04.11** |
| **4** | Модуль 4. Ядро Hadoop | **19** | **04.11-07.11** |
| **5** | Модуль 5. Написание скриптов Hadoop | **7** | **08.11-09.11** |
| **6** | Модуль 6. Базы данных Hadoop | **11** | **09.11-10.11** |
| **7** | Модуль 7. Введение в Data Lake | **4** | **11..** |
| **8** | Модуль 8. Интеллектуальный анализ Больших данных в медицине | **14** | **12.11-14.11** |
| **11** | Итоговая аттестация | **1** | **15.11** |
| **Всего:** | | **72** | **15 календарных дней** |

**6.Учебно-тематический план программы «BIG DATA в медицине»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль / Тема** | **Всего, час** | **Виды учебных занятий** | | | **Формы контроля** |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| 1 | **Модуль 1. Введение в Большие данные** **в медицине** | **5** | 1 | 3 | 1 |  |
| 1.1 | Тема 1.1.  Большие данные в медицине | 5 | 1 | 3 | 1 | тест |
| 2 | **Модуль 2. Введение в Hadoop** | **5** | 1 | 2 | 2 |  |
| 2.1 | Тема 2.1. Технология и инструменты Hadoop | 5 | 1 | 2 | 2 | тест |
| 3 | **Модуль 3.**  **Развертывание Hadoop** | **6** | 1 | 4 | 1 |  |
| 3.1 | Тема 3.1. Развертывание инфраструктуры Hadoop | 6 | 1 | 4 | 1 | тест |
| 4 | **Модуль 4. Ядро Hadoop** | **19** | 4 | 11 | 4 |  |
| 4.1 | Тема 4.1.  Файловая система HDFS | 7 | 2 | 4 | 1 | тест |
| 4.2 | Тема 4.2. Вычислительное ядро Hadoop | 6 | 1 | 3 | 2 | тест |
| 4.3 | Тема 4.3.  Запуск MapReduce задач | 6 | 1 | 4 | 1 | тест |
| 5 | **Модуль 5. Написание скриптов Hadoop** | **7** | 2 | 4 | 1 |  |
| 5.1 | Тема 5.1.  Скрипты Pig для данных из медицины | 7 | 2 | 4 | 1 | тест |
| **6** | **Модуль 6. Базы данных Hadoop** | **11** | 3 | 6 | 2 |  |
| 6.1 | Тема 6.1.  SQL и NoSQL базы данных | 5 | 1 | 3 | 1 | тест |
| 6.2 | Тема 6.2.  Работа с инструментом Hive для данных из медицины | 6 | 2 | 3 | 1 | тест |
| **7** | **Модуль 7. Введение в Data Lake** | **4** | 1 | 2 | 1 |  |
| 7.1 | Тема 7.1.  Data Lake (Озеро Данных) в медицине | 4 | 1 | 2 | 1 | тест |
| **8** | **Модуль 8. Интеллектуальный анализ Больших данных** **в медицине** | **14** | 2 | 8 | 4 |  |
| 8.1 | Тема 8.1. Извлечение знаний из данных в медицине | 7 | 1 | 4 | 2 | тест |
| 8.2 | Тема 8.2.  Обработка текстовых данных в медицине | 7 | 1 | 4 | 2 | тест |
|  | **Итоговая**  **аттестация** | 1 |  |  | 1 | **Итоговый тест** |

**7. Учебная (рабочая) программа повышения квалификации «BIG DATA в медицине»**

**Модуль 1. Введение в Большие данные в медицине (5 час.)**

Тема 1.1. Большие данные в медицине(5 час.)

Понятия данных, информации и знаний;

- Понятие Больших данных;

- Признаки Больших данных;

- Сравнение с классическими инструментами;

- Примеры использования Больших данных в задачах медицины.

**Модуль 2** **Введение в Hadoop (5 час.)**

Тема 2.1. Технология и инструменты Hadoop (5 час.)

- Что такое Hadoop;

- История возникновения и развития;

- Поставщики Hadoop;

- Экосистема Hadoop.

**Модуль 3. Развертывание Hadoop (7 час.)**

Тема 3.1. Развертывание инфраструктуры Hadoop (7 час.)

- Принцип развёртывания кластера Hadoop;

- Управляемые услуги «облачных» поставщиков инфраструктуры;

- Развертывание виртуальной машины Cloudera QuickStart VM

**Модуль 4. Ядро Hadoop (19 час.)**

Тема 4.1. Файловая система HDFS (7 час.)

- Принцип устройства файловых систем;

- Распределенная файловая система HDFS;

- Форматы хранения данных в HDFS;

- Работа в HDFS с помощью Java API;

- Работа в HDFS с помощью shell команд

Тема 4.2. Вычислительное ядро Hadoop (6 час.)

- Концепция вычислений MapReduce;

- Архитектура MapReduce ver1;

- Архитектура YARN;

- Особенные случаи использования MapReduce.

Тема 4.3. Запуск MapReduce задач (6 час.)

- Создание и запуск MapReduce задач на языке Java;

- Создание и запуск MapReduce задач на языке Python.

**Модуль 5. Написание скриптов Hadoop (7 час)**

Тема 5.1. Скрипты Pig для данных из медицины (7 час.)

- Инструмент скриптов Pig;

- Режимы работы Pig;

- Синтаксис Pig Latin;

- Обработка данных с помощью Pig.

**Модуль 6. Базы данных Hadoop (11 час.)**

Тема 6.1. SQL и NoSQL базы данных (5 час.)

- Понятие реляционных баз данных;

- Понятие NoSQL баз данных;

- Типы NoSQL баз данных;

- SQL и NoSQL инструменты в Hadoop.

Тема 6.2. Работа с инструментом Hive для данных из медицины (5 час.)

- Hive - SQL инструмент в Hadoop;

- Архитектура Hive;

- Модель данных Hive;

- Диалект HiveQL;

- Расширение функций Hive с помощью Python скриптов;

- Выборка и обработка данных в Hive (на примерах медицинских данных).

**Модуль 7. Введение в Data Lake (4 час.)**

Тема 7.1. Data Lake (Озеро Данных) в медицине (4 час.)

- Понятие озера данных (на примерах из медицины);

- Построение корпоративного озера данных (применительно к медицине);

- Принципы управления данными (применительно к медицине).

**Модуль 8. Интеллектуальный анализ Больших данных** **в медицине (14 час.)**

Тема 8.1. Извлечение знаний из данных в медицине (7 час.)

- История возникновения и развития машинного обучения;

- Примеры задач машинного обучения в медицине;

- Формальная постановка задач и типы задач;

- Оценка качества, метрики, переобучение.

Тема 8.2. Обработка текстовых данных в медицине (7 час)

- Основы предобработки текстовых данных;

- Примеры задач обработки естественного языка в медицине;

- Введение в задачи обработки естественного языка.

**Итоговая аттестация** (1 час.)

**7.1. Описание практико-ориентированных заданий и кейсов**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Номер темы/модуля** | **Наименование практического занятия** | **Описание** |
| **1.1** | **Модуль 1. Тема 1.** | Большие данные в медицине | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практическое задание на отработку терминологии проблематики Больших данных  *Пример задания.*  Составьте глоссарий терминов, касающихся технологии Больших данных. *(по видеолекции)*  Практическое задание на отработку постановки задач класса Больших данных.  *Пример задания.*  Опишите пример задачи из области медицины, относящейся к проблематике Больших данных. *(по видеолекции)*  *Пример задания.*  Опишите пример задачи из области медицины, в которых объем данных относительно велик, но её нельзя отнести к проблематике Больших данных. |
| **2.1** | **Модуль 2. Тема 2.1** | Технология, инструменты, архитектура Hadoop | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практическое задание на отработку выбора архитектуры решения класса Больших данных.  *Пример задания.*  Составьте список инструментов и утилит, входящих в экосистему Hadoop. *(по видеолекции)*  *Пример задания.*  Предложите архитектуру (состав компонент) кластера Hadoop для реализации следующего функционала:  a. Сбор данных из реляционных баз медицинских данных  b. Распределенное хранение медицинских данных  c. Предоставление SQL интерфейса для систем потребителей медицинских данных |
| **3.1** | **Модуль 3. Тема 3.1** | Развертывание инфраструктуры Hadoop | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практическое задание на отработку процесса выбора инфраструктуры для развертывания Hadoop с учетом различных факторов  *Пример задания*  Выделите и опишите основные преимущества развёртывания кластера Hadoop в «облаке». *(по видеолекции)*  *Пример задания*  Установите и запустите виртуальную машину Сloudera QuickStart. *(по видеолекции)*  *Пример задания.*  Проведите сравнительный анализ инструментов Hadoop, доступных в облаке Yandex и Mail.ru |
| **4.1** | **Модуль 4. Тема 4.1** | Файловая система HDFS | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения работать с файловой системой HDFS.  *Пример задания*.  Создайте в HDFS рабочую папку. *(по видеолекции)*  *Пример задания*.  Произведите загрузку в HDFS всех файлов из архива, доступного по ссылке https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files в созданную ранее директорию. Выведите на экран первые 10 строчек файла. *(по видеолекции)*  *Пример задания.*  Изучите код mkdir.java  Используя скомпилированный jar-пакет hdfs\_client.jar с помощью команды «hadoop jar hdfs\_client.jar mkdir [Directory\_Path]» создайте рабочую директорию Drugs\_files. Опишите вывод работы jar-пакета при его корректном и некорректном использовании, а также в случаях, когда директория уже существует. |
| **4.2** | **Модуль 4. Тема 4.2** | Вычислительное ядро Hadoop | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения формулировать задачи, решаемые с помощью распределенных вычислений.  *Пример задания*  Опишите концепцию MapReduce на примере задачи поиска самых часто встречаемых слов в нескольких научных публикациях медицинской тематики *(по видеолекции)*  *Пример задания*  Сформулируйте задачу обработки медицинских данных, которую можно решить, используя только Map функцию. |
| **4.3** | **Модуль 4. Тема 4.3** | Запуск MapReduce задач | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения создания и запуска MapReduce задач.  *Пример задания*  Запустите скомпилированный WordCount.jar пакет используя YARN *(по видеолекции)*  *Пример задания*  Запустите python скрипты mapper.py и reducer.py в виде hadoop-streaming задачи. *(по видеолекции)*  *Пример задания*  Опишите каким образом необходимо изменить код WordCount.java, чтобы скомпилированный пакет можно было запускать с аргументами входная и выходная директория.  *Пример задания*  Опишите каким образом необходимо изменить код WordCount.java, чтобы результат подсчета частот ошибочно показывал удвоенные значения. Предложите 2 варианта правок: для этапа Map и для этапа Reduce |
| **5.1** | **Модуль 5. Тема 5.1** | Скрипты Pig для данных из медицины | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умений создавать и запускать скрипты Pig  *Пример задания*  Произведите обработку файла Products.txt из архива, доступного по ссылке <https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files> с помощью скрипта Pig latin:   * Произведите загрузку * Извлеките первые 10 строк файла * Выведите их на экран * Произведите группировку по признаку ActiveIngredient * Произведите частотный анализ по выделенным группам * Произведите сортировку результатов * Выведите на экран 10 строк результата   *(по видеолекции)*  *Пример задания*  Повторите операции для других файлов архива. Например, для файла Application проведите расчёт количества ApplNo в группировке по ApplType. |
| **6.1** | **Модуль 6. Тема 6.1** | SQL и NoSQL базы данных | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения осуществлять выбор типа базы данных, оптимальный для задачи  *Пример задания*  Опишите архитектуру решения, какие SQL и NoSQL инструменты Вы бы использовали для реализации компонент:   * хранение журнала посещений медицинского учреждения * карточек пациентов   *(по видеолекции)*  *Пример задания*  Опишите какой тип(ы) SQL/NoSQL баз данных Вы бы использовали для создания хранилища публикаций и их метаданных (авторы, абстракт, ссылки, рецензии и т.д.) |
| **6.2** | **Модуль 6. Тема 6.2** | Работа с инструментом Hive для данных из медицины | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения создавать и запускать SQL-подобные запросы в Hive  *Пример задания*  Произведите обработку файла Products.txt из архива, доступного по ссылке <https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files> с помощью скрипта Pig latin:   * Создайте новую базу данных * Создайте схему реляционной таблицы * Произведите загрузку данных * Извлеките первые 10 строк файла * Создайте view c группировкой по признаку ActiveIngredient и частотным анализом по выделенным группам * Сделайте запрос к view и произведите сортировку результатов * Сохраните результат в таблице   *(по видеолекции)*  *Пример задания*  Повторите операции для других файлов архива. Например, для файла AppDoc проведите расчёт количества AppDocID в группировке по ActionType. |
| **7.1** | **Модуль 7. Тема 7.1** | Data Lake (Озеро Данных) в медицине | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения разрабатывать архитектуру озера данных  *Пример задания*  Опишите архитектуру озера данных. Какие инструменты могут использоваться для реализации озера данных.  *(по видеолекции)*  *Пример задания*  Опишите пример реализации озера данных в сфере медицины. Какие данные, из каких источников и в каком формате будут храниться. |
| **8.1** | **Модуль 8. Тема 8.1** | Извлечение знаний из данных в медицине | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения постановки задач машинного обучения  *Пример задания*  Опишите виды задач машинного обучения. Приведите примеры.  *(по видеолекции)*  *Пример задания*  Опишите пример задачи машинного обучения с учителем из области медицины  *Пример задания*  Опишите пример задачи машинного обучения с учителем из области медицины. |
| **8.2** | **Модуль 8. Тема 8.2** | Обработка текстовых данных в медицине | *Видеоролик с демонстрацией практических вопросов курса и примеров выполнения заданий. Самостоятельное выполнение обязательных практических заданий.*  Практические задания на отработку умения постановки задач обработки текстовых данных и анализа естественного языка  *Пример задания*  Опишите этапы и инструменты обработки текстовых данных. Опишите задачи анализа естественного языка. Приведите примеры.  *(по видеолекции)*  *Пример задания*  Опишите пример задачи обработки естественного языка на примере анализа научных публикаций медицинской тематики.  *Пример задания*  Опишите пример задачи обработки естественного языка в совокупности с задачей машинного обучения из области медицины. |

**8.Оценочные материалы по образовательной программе «BIG DATA в медицине»**

**8.1. Вопросы тестирования по модулям**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ модуля** | **Вопросы входного тестирования** | **Вопросы промежуточного тестирования** | **Вопросы итогового тестирования** |
| **0 (Входное тестирование)** | 1. С помощью какого инструмента MS Office можно собирать и анализировать табличные данные   1. MS Word 2. MS Excel 3. MS Access 4. Варианты 2 и 3 5. Варианты 1-2-3   2. Можно ли хранить и обрабатывать в одном файле MS Excel данные объемом более 10Гб   1. Да 2. Нет   3. Совокупность нескольких вычислительных систем, работающих совместно для выполнения общих приложений, называется   1. Локальная вычислительная сеть 2. Кластер 3. Облако   4. «База данных» и «Хранилище данных» – это одно и то же   1. Да 2. Нет   5. Для работы c реляционной базой данных чаще всего используется язык   1. С++ 2. Java 3. Python 4. SQL   5. К свойствам данных можно отнести   1. Репрезентативность 2. Достоверность 3. Полнота 4. Все варианты   6. Для анализа данных используется   1. Арифметика 2. Статистика 3. Вычислительная математика 4. Варианты 1 и 2 5. Варианты 1-2-3   7. Среднее и медианное значение всегда совпадают   1. Да 2. Нет   8. К объектно-ориентированным языкам программирования относятся   1. С 2. С++ 3. Java 4. Варианты 2 и 3 5. Варианты 1-2-3   9. Основным элементом ООП программы является   1. Переменная 2. Объект 3. Класс 4. Варианты 2 и 3 5. Варианты 1-2-3   10. К реляционным системам управления базами данных относятся   1. MS SQL Server 2. ORACLE Database 3. MongoDB 4. Варианты 1 и 2 5. Варианты 1-2-3 | **-** | **-** |
| **1.1** | **-** | 1. Вычислительная машина обрабатывает   1. Данные 2. Информацию 3. Знания   2. Научная публикация представляет собой следующий тип данных   1. Структурированные 2. Полу-структурированные 3. Неструктурированные   3. Большие данные — это данные, которые   1. Имеют большой объем 2. Разнообразные по типу 3. Быстро накапливаются 4. Все варианты   4. Платформа Больших данных, должна поддерживать функции   1. Сбор 2. Хранение 3. Анализ 4. Все варианты   5. Платформа Больших данных, должна поддерживать следующий вид масштабируемости   1. Вертикальная 2. Горизонтальная 3. Оба варианта   6. Отказоустойчивость платформ Больших данных предотвращает от потери данных при выходе из строя   1. Вычислительного узла кластера 2. Инфраструктуры Центра Обработки Данных 3. Внешнего канала связи   7. Принцип локальности данных подразумевает   1. Хранение всех данных на одной вычислительной машине 2. Хранение данных в контуре одного Центра Обработки Данных 3. Возможность обработки данных на месте их хранения   8. MPP базы данных можно отнести к технологиям Больших данных   1. Да 2. Нет   9. Какой подход дает больше гибкости и универсальности при работе с данными   1. MPP база данных 2. Большие данные   10. Платформы базы данных полностью заменяют классические инструменты работы с данными и при этом работают не менее эффективно   1. Да 2. Нет | **-** |
| **2.1** | **-** | 1. Hadoop — проект   1. Фонда Apache 2. Google 3. Yahoo!   2. Разработку Hadoop инициировали сотрудники   1. Фонда Apache 2. Google 3. Yahoo!   3. Импортозамещенный дистрибутив Hadoop предоставляет компания   1. Cloudera 2. Arenadata 3. Yandex   4. Кластер Hadoop размещается на   1. Одном сервере 2. Нескольких узлах 3. Нескольких стойках 4. Вариант 2 или 3   5. Файловая система Hadoop называется   1. FAT 2. NTFS 3. Ext3 4. HDFS   6. Какой инструмент Hadoop используется для распределения вычислительных ресурсов кластера   1. HDFS 2. YARN 3. ZooKeeper   7. Какой инструмент Hadoop используется для написания скриптов?   1. MapReduce 2. Hive 3. Hbase 4. Pig   8. Какой инструмент Hadoop используется для выполнения SQL запросов?   1. MapReduce 2. Hive 3. Hbase 4. Pig   9. Какой инструмент Hadoop используется для обработки данных в оперативной памяти   1. Spark 2. Solr 3. Storm   10. Какой инструмент Hadoop используется для импорта из реляционных баз данных   1. Sqoop 2. Flume 3. Kafka | **-** |
| **3.1** | **-** | Кластер Hadoop включает узлы   1. Master 2. Slave 3. Оба варианта   2. Для развертывания кластера Hadoop необходимо высококлассное оборудование с обязательной корпоративной поддержкой   1. Да 2. Нет   3. Продуктивный кластер Hadoop рекомендуется составлять минимум из данного количества узлов   1. 1 2. 3 3. 4 4. 5   4. Сбои и отказы каких типов узлов не являются критичными для Hadoop   1. Master 2. Slave 3. Оба варианта   5. Масштабировать ресурсы кластера Hadoop возможно путем   1. Добавления новых узлов 2. Расширения ресурсов внутри имеющихся узлов 3. Оба варианта   6. Ресурс для Master узла зависит от объема кластера   1. Да 2. Нет   7. Какие российские компании предлагают «облачные» сервисы Hadoop   1. Yandex 2. Mail.ru 3. Rambler   8. Возможно ли развернуть Hadoop на одной рабочей станции   1. Да 2. Да, но только для ознакомительного использования 3. Нет | **-** |
| **4.1** | **-** | 1. Файловая система определяет   1. Максиммальный размер имён файлов 2. Максимальный возможный размер файла 3. Оба варианта   2. Прикладные программы имеют информацию о физическом расположении файле на носителе   1. Да 2. Нет   3. Классическая архитектура файловой системы называется   1. FAT 2. SLIM 3. TALL   4. Пространство файловой системы логически содержит   1. Область указателей 2. Область данных 3. Оба варианта   5. Укажите основные отличительные особенности HDFS   1. Распределенное хранение файлов 2. Возможность хранить огромные объемы данных 3. Встроенная отказоустойчивость 4. Все варианты   6. Сервер имён HDFS называется   1. DataNode 2. NameNode 3. ZooKepeer   7. Необходимо ли использование аппаратных средств отказоустойчивости накопителей (RAID) для серверов данных HDFS   1. Да 2. Нет   8. Какое количество копий блоков файлов HDFS хранит по умолчанию   1. 1 2. 2 3. 3   9. Какая модель записи-чтения используется в HDFS   1. write-once / read-many 2. write- many / read- once   10. Механизм репликации данных в HDFS производит операции   1. выбор новых DataNode-серверов для новых реплик 2. балансировка размещения данных по DataNode-серверам 3. Оба варианта | **-** |
| **4.2** | **-** | 1. Алгоритмы Map-reduce обрабатывают данные   1. На одной машине с аналитическим ядром 2. Распределённо, по возможности там, где они хранятся   2. Алгоритмы Map-reduce включают этапы   1. Map 2. Reduce 3. Shuffle 4. Map и Reduce 5. Map, Reduce и Shuffle   3. Master узел для реализации Map-reduce задач запускает процесс   1. JobTracker 2. TaskTracker   4. Архитектура MRv1 способна эффективно работать на больших кластерах (более 4000 узлов)   1. Да 2. Нет   5. Служба, которая в Hadoop ver 2 используется для распределения вычислительных ресурсов кластера называется   1. MRv2 2. YARN 3. ZooKepeer   6. Служба YARN, управляющая выполнением конкретного экземпляра приложения называется   1. ResourceManager 2. ApplicationMaster 3. NodeManager   7. YARN способна запускать только Map-reduce задачи   1. Да 2. Нет   8. Возможен ли запуск только Map задач   1. Да 2. Нет   9. Операция Combine чаще всего совпадает с функцией Reduce   1. Да 2. Нет   10. Фреймворк Tez (надстройка YARN) позволяет   1. Оптимизировать выполнение цепочек и графов Map-reduce задач 2. Балансировать данные в HDFS 3. Оба варианта | **-** |
| **4.3** | **-** | 1. Код языка Java является   1. Интерпретируемым 2. Исполняемым после компиляции   2. Код языка Python является   1. Интерпретируемым 2. Исполняемым после компиляции   3. Для реализаций функций Map и Reduce на языке Java их необходимо разделить в два отдельных файла/пакета   1. Да 2. Нет   4. Для реализаций функций Map и Reduce на языке Python их необходимо разделить в два отдельных файла/пакета   1. Да 2. Нет   5. Для реализаций функций Map и Reduce на языке Java в коде необходимо явно произвести импорт библиотек Hadoop   1. Да 2. Нет   6. Для реализаций функций Map и Reduce на языке Python в коде необходимо явно произвести импорт библиотек Hadoop   1. Да 2. Нет   7. Для запуска Java пакета с MapReduce задачей необходимо использование Streaming-интерфейса   1. Да 2. Нет   8. Для запуска Python реализации MapReduce задач необходимо использование Streaming-интерфейса   1. Да 2. Нет | **-** |
| **5.1** | **-** | Код синтаксиса Pig называется   1. Latin 2. Cyrillic 3. Alfabet   2. Код синтаксиса Pig по сравнению с аналогичным MapReduce кодом выполняется   1. Быстрее 2. Медленнее   3. Pig используется для   1. Предобработки данных 2. Простых вычислений 3. Реляционных вычислений 4. Все варианты   4. Pig можно использовать для вычислений над локальной файловой системой   1. Да 2. Нет   5. Pig можно использовать для вычислений над HDFS   1. Да 2. Нет   6. Pig может работать в режимах   1. Интерактивного ввода команд 2. Пакетного выполнения скрипта 3. Оба варианта   7. Pig поддерживает циклы   1. Да 2. Нет   8. Pig поддерживает операцию группировки   1. Да 2. Нет   9. Pig поддерживает пользовательские функции на языке   1. Java 2. Python 3. Оба варианта   10. Pig поддерживает режимы оптимизированных по скорости вычислений с помощью   1. Tez 2. Spark 3. Оба варианта | **-** |
| **6.1** | **-** | 1. Какой объект наиболее близок к понятию «реляция» в рамках модели данных   1. Матрица 2. Таблица   2. Возможные (допустимые) значения данных конкретного атрибута задаются   1. Доменом 2. Кортежем   3. SQL это   1. Язык структурированных запросов 2. Язык неструктурированных запросов 3. Язык программирования   4. Любые данные можно эффективно обрабатывать в рамках реляционной модели   1. Да 2. Нет   5. К NoSQL относят   1. Хранилище «ключ-значение» 2. Документо-ориентированные базы данных 3. Графовые базы данных 4. Все варианты   6. К решениям класса SQL над Hadoop относят   1. Hive 2. Impala 3. Оба варианта   7. Решение Impala использует парадигму Map-Reduce   1. Да 2. Нет   8. Решения класса SQL над Hadoop используют для   1. «Пакетной» обработки больших массивов данных 2. Случайных операций «чтения/записи» в больших таблицах 3. Оба варианта   9. К решениям класса NoSQL над Hadoop относят   1. Hbase 2. Cassandra 3. Оба варианта   10. Решение HBase обеспечивает   1. Согласованность 2. Доступность 3. Оба варианта | **-** |
| **6.2** | **-** | 1. В Hive реализован полноценный диалект SQL   1. Да 2. Нет   2. К базе данных Hive можно подключится, используя ODBC-драйвер (например, из Excel), для создания выборки   1. Да 2. Нет   3. Для хранения данных Hive использует   1. HDFS 2. HBase 3. Оба варианта   4. Для вычислений Hive использует   1. MapReduce (YARN) 2. Tez 3. Оба варианта   5. Данные Hive могут хранится в форматах   1. CSV/text files 2. Apache Parquet 3. Apache ORC 4. Все варианты   6. Для дробления данных в Hive используется   1. Партиции 2. «Бакеты» 3. Оба варианта   7. Результаты операции выброки (select) могут быть   1. Помещены в таблицу базы данных 2. Сохранены в файл 3. Выведены на экран консоли 4. Все варианты   8. Hive поддерживает расширение функционала за счет создания пользовательских функций   1. Да 2. Нет   9. Hive поддерживает циклы   1. Да 2. Нет   10. Hive поддерживает операцию группировки   1. Да 2. Нет | **-** |
| **7.1** | **-** | 1. Озеро данных – это корпоративное хранилище данных реализованное на Hadoop   1. Да 2. Нет   2. Озеро данных может хранить   1. Структурированные данные 2. Текстовые документы 3. Изображения 4. Аудиофайлы 5. Все варианты   3. Озеро данных хранит только данные, актуальные для текущих задач   1. Да 2. Нет   4. При появлении новых источников или новых форматов данных, процесс адаптации будет сложнее для   1. Корпоративного хранилища данных 2. Озера данных   5. Данные из Озера данных легко доступны для анализа   1. Да 2. Нет   6. Стоимость владения в расчёте на 1 ГБ данных минимальна для   1. Корпоративного хранилища данных 2. Озера данных   7. Озеро данных может содержать пред обработанные данные   1. Да 2. Нет   8. Концепция озера данных реализуется только с помощью инструментов Hadoop   1. Да 2. Нет | **-** |
| **8.1** | **-** | 1. Какой из указанных ниже терминов не имеет никакого отношения к искусственному интеллекту   1. AlphaGo 2. Siri 3. DeepBlue 4. SQL   2. Какая из этих задач не относится к машинному обучению   1. Распознавание лиц по фото-видео 2. Автоматический перевод 3. Прогнозирование исхода заболевания 4. Дизайн сайтов   3. Выберите неверные утверждения   1. В задаче машинного обучения необходимо восстановить зависимость 2. В задаче машинного обучения дано конечное количество пар “объект - ответ” 3. В задаче машинного обучения объекты обладают признаками 4. В задаче машинного обучения наличие метрики качества необязательно   4. Какой из типов данных не может использоваться в качестве входных данных для задач обучения с учителем   1. Признаковое описание 2. Матрица расстояний 3. Изображение 4. Блок-схема   5. Какие из этих задач относятся к регрессии   1. Прогноз четвертной оценки ученика 2. Прогноз продаж магазина 3. Оценка времени в пути из точки А в точку Б 4. Визуализация кривой заболеваемости   6. Какая из задач не относится к классификации   1. Прогноз возврата кредита (вернет или нет) 2. Прогноз положительного теста на коронавирус (положительный/отрицательный) 3. Определение пола по истории посещения веб-сайтов (м/ж) 4. Определение необходимого количества младшего медицинского персонала   7. Какой из типов задач не относится к обучению без учителя   1. Понижение размерности 2. Кластеризация 3. Визуализация данных 4. Классификация текстов   8. Для чего нужна кластеризация   1. Для поиска скрытой структуры в данных 2. Для очистки данных 3. Для оценки точности алгоритмов 4. Для определения аномалий в данных   9. Что можно использовать в качестве метрики алгоритма   1. Доля верных ответов в задаче классификации 2. Среднеквадратичное относительное отклонение в задаче регрессии 3. Среднеквадратичное абсолютное отклонение в задаче регрессии 4. Количество верных ответов в задаче классификации   10. Какие из нижеприведенных утверждений характерны для переобучения   1. Низкая точность алгоритма на отложенной выборке 2. Высокая точность алгоритма на отложенной выборке 3. Чрезмерный подгон данных под обучающую выборку 4. Сопоставимый уровень точности алгоритма на обучающей и на тестовой выборках | **-** |
| **8.2** | **-** | 1. Какие из приведенных ниже задач относят к обработке естественного языка   1. Определение эмоционального окраса 2. Суммаризация текстов 3. Автоматический перевод 4. Распознавание текстов на фото   2. Какие данные можно использовать для задач обработки естественного языка   1. Последовательности слов 2. Последовательности символов 3. Последовательности картинок 4. Точки на координатной плоскости   3. Что из этого не имеет никакого отношения к лемматизации   1. Приведение существительных к форме именительного падежа единственного числа 2. Приведение прилагательных к форме именительного падежа единственного числа мужского рода 3. Приведение глаголов к форме инфинитива несовершенного вида 4. Приведение существительного к основе слова   4. Что из этого не имеет никакого отношения к стеммингу   1. Приведение существительного к его корню 2. Может выполняться с помощью поиска по словарю 3. Может выполняться с помощью лексических правил, но могут возникать ошибки 4. Приведение слова к основе   5. Какое из слов не входит в мешок слов, построенных на тексте «Граф Патто играл в лото. Графиня Патто знала про то, что граф Патто играл в лото. А граф Патто не знал про то, что графиня Патто знала про то, что граф Патто играл в лото.»   1. Граф 2. не 3. в 4. узнала   6. Посчитайте TF-IDF для слова “лингвист” которое встречается 5 раз в документе Х, содержащем 100 слов, а всего 1000 документов, из которых только 10 содержат данное слово (для удобства возьмите десятичный логарифм)   1. 0.05 2. 0.1 3. 5 4. 0.01   7. Выберите самые удаленные друг от друга слова в пространстве векторных представлений слов   1. man (мужчина) – king (король) 2. man (мужчина) – woman (женщина) 3. queen (королева) – king (король) 4. man (мужчина) – queen (королева)   8. Какие задачи можно решать с помощью алгоритмов информационного поиска   1. Поиск статьи 2. Поиск документов в базе знаний 3. Поиск товаров на сайте 4. Выделение основных тезисов из документа (суммаризация)   9. Какие задачи можно решать с помощью вопросно-ответных систем   1. Консультирование посетителей сайта медицинского учреждения 2. Техническая поддержка банка 3. Создание баз знаний 4. Создание баз данных   10. Выберите существующие подходы для решения задачи машинного перевода   1. Статистический машинный перевод 2. Машинный перевод на правилах 3. Адаптивный машинный перевод 4. Глубокий машинный перевод | **-** |
| **Итоговое тестирование** |  |  | Примеры вопросов (тест формируется из базы тестовых вопросов):  1. Большие данные — это данные, которые   1. Имеют большой объем 2. Разнообразные по типу 3. Быстро накапливаются 4. Все варианты   2. Кластер Hadoop размещается на   1. Одном сервере 2. Нескольких узлах 3. Нескольких стойках 4. Вариант 2 или 3   3. Файловая система Hadoop называется   1. FAT 2. NTFS 3. Ext3 4. HDFS   4. Кластер Hadoop включает узлы   1. Master 2. Slave 3. Оба варианта   5. Укажите основные отличительные особенности HDFS   1. Распределенное хранение файлов 2. Возможность хранить огромные объемы данных 3. Встроенная отказоустойчивость 4. Все варианты   6. Алгоритмы Map-reduce обрабатывают данные   1. На одной машине с аналитическим ядром 2. Распределённо, по возможности там, где они хранятся   7. Для запуска Python реализации MapReduce задач необходимо использование Streaming-интерфейса   1. Да 2. Нет   8. Pig используется для   1. Предобработки данных 2. Простых вычислений 3. Реляционных вычислений 4. Все варианты   9. SQL это   1. Язык структурированных запросов 2. Язык неструктурированных запросов 3. Язык программирования   10. Решения класса SQL над Hadoop используют для   1. «Пакетной» обработки больших массивов данных 2. Случайных операций «чтения/записи» в больших таблицах 3. Оба варианта   11. К базе данных Hive можно подключится, используя ODBC-драйвер (например, из Excel), для создания выборки   1. Да 2. Нет   12. Hive поддерживает расширение функционала за счет создания пользовательских функций   1. Да 2. Нет   13. Озеро данных может хранить   1. Структурированные данные 2. Текстовые документы 3. Изображения 4. Аудиофайлы 5. Все варианты   12. Данные из Озера данных легко доступны для анализа   1. Да 2. Нет   13. Какая из этих задач не относится к машинному обучению   1. Распознавание лиц по фото-видео 2. Автоматический перевод 3. Прогнозирование исхода заболевания 4. Дизайн сайтов   14. Какие из этих задач относятся к регрессии   1. Прогноз четвертной оценки ученика 2. Прогноз продаж магазина 3. Оценка времени в пути из точки А в точку Б 4. Визуализация кривой заболеваемости   15. Для чего нужна кластеризация   1. Для поиска скрытой структуры в данных 2. Для очистки данных 3. Для оценки точности алгоритмов 4. Для определения аномалий в данных   16. Какие из приведенных ниже задач относят к обработке естественного языка   1. Определение эмоционального окраса 2. Суммаризация текстов 3. Автоматический перевод 4. Распознавание текстов на фото   17. Что из этого не имеет никакого отношения к стеммингу   1. Приведение существительного к его корню 2. Может выполняться с помощью поиска по словарю 3. Может выполняться с помощью лексических правил, но могут возникать ошибки 4. Приведение слова к основе   18. Какие задачи можно решать с помощью вопросно-ответных систем   1. Консультирование посетителей сайта медицинского учреждения 2. Техническая поддержка банка 3. Создание баз знаний 4. Создание баз данных   19. Какие российские компании предлагают «облачные» сервисы Hadoop   1. Yandex 2. Mail.ru 3. Rambler   20. Вычислительная машина обрабатывает   1. Данные 2. Информацию 3. Знания |

* 1. **описание показателей и критериев оценивания, шкалы оценивания.**

Шкалы оценивания

* 1. Входное тестирование

**Шкала критериев оценивания по входному тестированию**

|  |  |  |
| --- | --- | --- |
| **Шкала оценивания** | **Критерии** | **Описание** |
| 70%-100% верных ответов на вопросы входного теста | **зачтено** | Потенциальный обучающийся успешно выполнил входной тест и может быть допущен к обучению по программе |
| 69% и менее верных ответов на вопросы входного теста | **не зачтено** | Потенциальный обучающийся не выполнил успешно входной тест и не может быть допущен к обучению по программе |

* 1. Промежуточное тестирование по темам/модулям

**Шкала критериев оценивания по промежуточному тестированию**

|  |  |  |  |
| --- | --- | --- | --- |
| **Шкала** | **Критерии** | **Описание** | **Уровень сформированности компетенции** |
| 80%-100% верных ответов на вопросы промежуточного теста | зачтено | Обучающийся успешно выполнил промежуточный тест по теме/модулю | Продвинутый уровень |
| 70%-79% | зачтено | Обучающийся успешно выполнил промежуточный тест по теме/модулю | Базовый уровень |
| 50%-69% верных ответов на вопросы промежуточного теста | не зачтено | Обучающийся не выполнил успешно промежуточный тест, необходимо повторное назначение теста и консультации преподавателя | Начальный уровень |
| Менее 50% верных ответов на вопросы промежуточного теста | не зачтено | Обучающийся не выполнил успешно промежуточный тест, необходимо повторное назначение теста и консультации преподавателя | Не владеет, компетенция не сформирована |

* 1. Итоговое тестирование по программе

**Шкала критериев оценивания по итоговому тестированию**

|  |  |  |  |
| --- | --- | --- | --- |
| **Шкала** | **Критерии** | **Описание** | **Уровень сформированности компетенции** |
| 80%-100% верных ответов на вопросы промежуточного теста | зачтено | Обучающийся успешно выполнил итоговый тест по программе | Продвинутый уровень |
| 70%-79% | зачтено | Обучающийся успешно выполнил итоговый тест по программе | Базовый уровень |
| 50%-69% верных ответов на вопросы промежуточного теста | не зачтено | Обучающийся не выполнил успешно итоговый тест по программе | Начальный уровень |
| Менее 50% верных ответов на вопросы промежуточного теста | не зачтено | Обучающийся не выполнил успешно итоговый тест по программе | Не владеет, компетенция не сформирована |

* 1. **примеры контрольных заданий по модулям или всей образовательной программе.**

**Пример контрольного задания по программе**

1. В этом задании вам нужно продемонстрировать умение использования компонент Hadoop:

* HDFS
* MapReduce
* Pig
* Hive

2. Выберите набор табличных данных и сохраните его (например, с помощью MS Excel) в текстовый формат (CSV).

3. Произведите исследование по плану:

(а) Выберите вариант инфраструктуры Hadoop

(б) Произведите загрузку данных

(в) Проведите обработку данных средствами Hadoop

(г) Поставьте задачу машинного обучения для имеющихся данных

(д) Предложите варианты по обогащению (расширению набора признаков) имеющихся данных

* 1. **тесты и обучающие задачи (кейсы), иные практико-ориентированные формы заданий.**

Примеры практических заданий представлены в таблице 7.1, примеры тестов представлены в таблице 8.1.

В программе используются задания, которые являются обязательными для выполнения по каждой теме. Для организации самостоятельной работы используются задания со свободным ответом (в том числе проблемные), которые слушатели выполняют самостоятельно с последующей проверкой преподавателем.

**Примеры заданий для самостоятельной работы**

**Модуль 1. Тема 1.1**

Исследовательская компания Gartner регулярно публикует цикл зрелости технологий (Hype Cycle) <https://habr.com/ru/post/198506/>. Проследите траекторию и динамику продвижения технологии Больших данных (Big Data) на данном цикле Gartner.

**Модуль 2. Тема 2.1**

Задание 1

Важной задачей в управлении кластером Hadoop с большим количеством компонент и ресурсов является проблема координации и синхронизации. Для этого зачастую используется инструмент Zookeeper. Проанализируйте информационные ресурсы и составьте краткий отчет: в каких сценариях чаще всего используется данный компонент.

Задание 2

Для упрощения работы с Hadoop, также часто используют инструмент Ambari. Произведите анализ функционала Ambari и составьте краткий отчет.

**Модуль 3. Тема 3.1**

При выборе варианта размещения кластера Hadoop (в собственном ЦОД, облаке от российского или иностранного поставщика) необходимо учитывать множество факторов. Помимо общей стоимости владения (которая включает как капитальные, так и операционные расходы), доступных функций и наличия квалифицированных кадров, важно учитывать местное законодательство.

Изучите информационные ресурсы и опишите случаи, когда развертывание кластера в облачной инфраструктуре (как от российского, так и от иностранного поставщика) недопустимо.

**Модуль 4. Тема 4.1**

Файловая система HDFS очень часто используется с целью хранения больших объемов информации. Однако существуют другие подходы и инструменты, альтернативные Hadoop HDFS. Проанализируйте информационные ресурсы и на примере облачных поставщиков управляемых услуг (таких как Mail.ru MCS, Amazon Web Services и других) опишите доступные инструменты, альтернативные Hadoop HDFS, и основные сценарии их использования.

**Модуль 4. Тема 4.2**

Задание 1

В 2017 вышло большое обновление в виде Hadoop 3. Проанализируйте информационные ресурсы и опишите основные новые возможности YARN в Hadoop 3.

Задание 2

В видеолекции мы подробно разобрали подход MapReduce задач, концепция которой дала толчок к созданию Hadoop. Однако, начиная с Hadoop версии 2, компонента YARN позволяет запускать и другие типы задач. Составьте таблицу – запуск и управление какими стандартными типами задач поддерживается в актуальной версии YARN.

Задание 3

Ядро Hadoop часто используют для пакетной обработки больших массивов данных. Однако большое количество задач требуют итерационного и иногда даже случайного доступа к данным с малыми задержками. Для этих целей используют фреймворк Spark. Проанализируйте информационные ресурсы и составьте краткий отчет – в каких сценариях и для решения каких задач используется фреймворк Spark.

**Модуль 4. Тема 4.3**

В видеолекции мы разобрали наиболее популярные языки для реализации MapReduce задач в Hadoop: Java и Python. Проанализируйте информационные ресурсы и составьте таблицу – какие языки программирования и скриптов используются в практике реализации MapReduce задач в Hadoop.

**Модуль 5. Тема 5.1**

Задание 1

Составьте аннотированный каталог базовых функций pig latin.

Задание 2

Для расширения функциональности в pig latin имеется поддержка UDF (User Defined Functions) – пользовательских функций. Составьте таблицу, какие языки программирования и скриптов поддерживаются для создания UDF.

**Модуль 6. Тема 6.1**

Задание 1

Проанализируйте информационные ресурсы и составьте таблицу реляционных СУБД, которые чаще всего используются в промышленной эксплуатации. Укажите тип лицензии (проприетарная, свободная).

Задание 2

Проанализируйте информационные ресурсы и составьте таблицу НЕреляционных СУБД, которые чаще всего используются в промышленной эксплуатации. Укажите тип СУБД (документо-ориентированная, «ключ-значение» и т.п.) и тип лицензии (проприетарная, свободная).

**Модуль 6. Тема 6.2**

Часто в альтернативу Hive ставят инструмент Impala. Проанализируйте информационные ресурсы и составьте краткий отчет с описанием сходства и различия данных SQL-инструментов для Apache Hadoop.

**Модуль 7. Тема 7.1**

Проанализируйте информационные ресурсы и составьте таблицу какие облачные поставщики управляемых услуг (такие как Mail.ru MCS, Amazon Web Services и другие) предлагают услуги создания и управления Озерами данных. Укажите какие технологии использует облачный поставщик для организации Озера данных.

**Модуль 8. Тема 8.1**

Задание 1

Составьте список статей из области медицины (не менее 10), в которых ставятся и описываются задачи, решаемые, в том числе, методами машинного обучения. Укажите тип задач машинного обучения.

Задание 2

Проанализируйте информационные ресурсы и составьте таблицу какие инструменты используются для решения задач машинного обучения над большими данными. Приведите краткое описание данных инструментов.

**Модуль 8. Тема 8.2**

Задание 1

Одной из задач обработки естественного языка является информационный поиск. Проанализируйте информационные ресурсы и составьте краткое описание функциональных возможностей инструмента

Apache Solr. Также составьте таблицу инструментов, которые являются популярными альтернативами для Apache Solr.

Задание 2

При обработке больших массивов (корпусов) текстов важной задачей является тематическое моделирование. Проанализируйте информационные ресурсы и составьте краткое описание данной задачи. Составьте таблицу алгоритмов, используемых для решения данной задачи.

**8.5.**  **описание процедуры оценивания результатов обучения.**

В процессе обучения слушатели должны выполнить обязательные практические задания с последующей проверкой преподавателем (в рамках практических занятий), а также задания для самостоятельной работы также с последующей проверкой преподавателем; пройти промежуточную аттестацию по каждой теме/модулю – электронное тестирование. Промежуточное тестирование включает 8-10 вопросов по теме/модулю. Промежуточное тестирование считается успешно завершённым, если слушателем даны верные ответы на не менее чем 70% вопросов теста. Каждый слушатель имеет три попытки на выполнение промежуточного теста, время тестирования – 0,5 часа.

К итоговой аттестации допускаются слушатели, **выполнившие все обязательные практические задания (в рамках практических занятий) и промежуточные тесты по темам (не менее 70% верных ответов по каждому тесту по теме/модулю)**, предусмотренные учебной программной. В ходе итогового электронного тестирования слушателю предлагается ответить на 20 вопросов. Тестирование считается успешно завершённым, если слушателем даны верные ответы на не менее чем 70% вопросов теста. Каждый слушатель имеет одну попытку на выполнение итогового теста, время тестирования – 1 час.

**9.Организационно-педагогические условия реализации программы**

**9.1. Кадровое обеспечение программы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Фамилия, имя, отчество (при наличии)** | **Место основной работы и должность, ученая степень и ученое звание (при наличии)** | **Ссылки на веб-страницы с портфолио (при наличии)** | **Фото в формате jpeg** | **Отметка о полученном согласии на обработку персональных данных** |
| **1** | **Милков Максим Леонидович** | Softline, технический директор департамента анализа данных | https://www.linkedin.com/in/maxim-milkov-195730185/ |  | **+** |

**9.2.Учебно-методическое обеспечение и информационное сопровождение**

|  |  |
| --- | --- |
| **Учебно-методические материалы** | |
| Методы, формы и технологии | Методические разработки,  материалы курса, учебная литература |
| Методы: проблемно-развивающие методы; исследовательский, объяснительно-иллюстративный, метод проблемного изложения, частично-поисковый, методы контроля и самоконтроля  Формы: самостоятельная работа по освоению теоретического материала; самоконтроль через ответы на вопросы по видеолекциям и выполнение практических заданий для самоконтроля; промежуточный контроль (тесты по темам для промежуточной аттестации), итоговый контроль (итоговое тестирование по курсу). | 1. Марц Н., Большие данные: принципы и практика построения масштабируемых систем обработки данных в реальном времени. — Вильямс, 2019 2. Judith Hurwitz, Big Data For Dummies. John Wiley & Sons Limited, 2013 3. Барсегян Арменак, Анализ данных и процессов. BHV, 2018 |
|  |  |

|  |  |
| --- | --- |
| **Информационное сопровождение** | |
| Электронные  образовательные ресурсы | Электронные  информационные ресурсы |
|  | 1. Блог на коллективном ресурсе разработчиков <https://habr.com/ru/hub/bigdata/> |

**9.3.Материально-технические условия реализации программы**

|  |  |
| --- | --- |
| Вид занятий | Наименование оборудования,  программного обеспечения |
| Лекционные (видеолекции) | Операционная система семейств Linux Mac Os, Windows.  Браузер Google Chrome  Рекомендуется скорость канала от 2 Мбит в секунду |
| Практические | Операционная система семейств Linux Mac Os, Windows.  Открытая система управления виртуальными машинами VirtualBox  Рекомендуется скорость канала от 2 Мбит в секунду  Вычислительное оборудование:   * Не менее 4х ядер CPU * Не менее 10 Гб RAM * Не менее 25Гб HDD |
| Самостоятельная работа | Операционная система семейств Linux Mac Os, Windows.  Открытая система управления виртуальными машинами VirtualBox  Рекомендуется скорость канала от 2 Мбит в секунду  Вычислительное оборудование:   * Не менее 4х ядер CPU * Не менее 10 Гб RAM * Не менее 25Гб HDD |

**III.Паспорт компетенций (Приложение 2)**

Автономная некоммерческая организация дополнительного профессионального образования

«СофтЛайн Эдюкейшн»

.

ПАСПОРТ КОМПЕТЕНЦИИ

**Дополнительная профессиональная программа повышения квалификации**

**«BIG DATA в медицине»**

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Наименование компетенции | Способен ставить и решать прикладные задачи по подготовке медицинских данных для проведения аналитических работ по исследованию Больших данных с использованием инструментов Hadoop Common | |
| 2. | Указание типа компетенции | профессиональная | |
| 3. | Определение, содержание и основные сущностные характеристики компетенции | Под компетенцией понимается способность постановки и нахождения путей решения прикладных задач по подготовке медицинских данных для проведения аналитических работ по исследованию Больших данных с использованием инструментов Hadoop Common.  Слушатель должен:  *знать*:  - сущность понятия «большие данных» и примеры применимости Больших данных в медицине;  - концепцию распределенных вычислений MapReduce;  - особые техники использования MapReduce;  - компоненты программно-технических архитектур Hadoop;  - компоненты распределенной файловой системы HDFS;  - особенности реляционных и нереляционных баз данных, примеры этих баз данных в медицине;  - существующие приложения Hadoop и интерфейсах взаимодействия с ними;  - существующие походы к извлечению знаний из данных по медицине;  - подходы и технологии обработки и анализа текстовых данных в медицине.  *уметь***:**  - использовать возможности имеющейся в медицинской организации технологической инфраструктуры Больших данных;  - использовать выбранные инструментальные средства для хранения, преобразования и обработки Больших данных в медицине;  - использовать облачные варианты технологической инфраструктуры Больших данных применительно к медицине;  *владеть*:  - приемами работы с распределённой файловой системой HDFS;  - написанием программного кода на Java для решения MapReduce задач (применительно к медицине);  - написанием программного кода на Python для решения MapReduce задач (применительно к медицине);  - методами запуска MapReduce задач с помощью YARN;  - написанием SQL-подобных запросов для создания выборок в Hive;  - написанием скриптов Pig Latin для подготовки медицинских данных и проведения аналитических работ. | |
| 4. | Дескриптор знаний, умений и навыков по уровням | Уровни сформированности компетенции  обучающегося | Индикаторы |
|  | Начальный уровень | *Знает*: сущность понятия «большие данных» и примеры применимости Больших данных в медицине; концепцию распределенных вычислений MapReduce; особенности реляционных и нереляционных баз данных, примеры этих баз данных в медицине;  *Умеет*: - использовать выбранные инструментальные средства для хранения Больших данных в медицине.  *Владеет*: отдельными приемами работы с распределённой файловой системой HDFS |
|  | Базовый уровень | *Знает*: отдельные техники использования MapReduce; - компоненты программно-технических архитектур Hadoop; компоненты распределенной файловой системы HDFS; существующие приложения Hadoop и интерфейсы взаимодействия с ними; отдельные походы к извлечению знаний из данных по медицине; отдельные подходы и технологии обработки и анализа текстовых данных в медицине.  *Умеет*: использовать отдельные инструментальные средства для хранения, преобразования и обработки Больших данных в медицине; использовать облачные варианты технологической инфраструктуры Больших данных применительно к медицине.  *Владеет*: написанием программного кода на Java для решения MapReduce задач (применительно к медицине); написанием программного кода на Python для решения MapReduce задач (применительно к медицине); отдельными методами запуска MapReduce задач с помощью YARN; написанием SQL-подобных запросов для создания выборок в Hive |
|  | Продвинутый | *Знает* особые техники использования MapReduce; существующие походы к извлечению знаний из данных по медицине; современные подходы и технологии обработки и анализа текстовых данных в медицине.  *Умеет*: использовать возможности имеющейся в медицинской организации технологической инфраструктуры Больших данных; использовать разнообразные инструментальные средства для хранения, преобразования и обработки Больших данных в медицине.  *Владеет*: написанием скриптов Pig Latin для подготовки медицинских данных и проведения аналитических работ |
| 5. | Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции | Для формирования данной компетенции слушатель должен владеть:   * базовыми компетенциями цифровой грамотности, в том числе навыками владения ПК и ОС Linux на уровне среднего или продвинутого пользователя; * компетенциями в области алгоритмизации; * базовыми знаниями в области языков программирования Java, Python и SQL | |
| 6. | Средства и технологии оценки | Тесты, электронное промежуточное (по темам/модулям) и итоговое тестирование | |

**IV.Иная информация о качестве и востребованности образовательной программы** (результаты профессионально-общественной аккредитации образовательной программы, включение в системы рейтингования, призовые места по результатам проведения конкурсов образовательных программ и др.) (при наличии)

**V.Рекомендаций к программе от работодателей**: наличие не менее двух писем и/или подтверждения на цифровой платформе Государственной системы предоставления ПЦС от работодателей о рекомендации образовательной программы для реализации в рамках Государственной системы предоставления ПЦС на формирование у трудоспособного населения компетенций цифровой экономики с указанием востребованности результатов освоения программы в сфере деятельности соответствующих компаний и готовности к рассмотрению заявок наиболее успешно освоивших образовательную программу граждан на прохождение стажировки и (или) собеседования на предмет трудоустройства путем проставления отметки в профиле программы

**Загружено на платформу**

**VI.Указание на возможные сценарии профессиональной траектории граждан** по итогам освоения образовательной программы (в соответствии с приложением)

**СЦЕНАРИИ ПРОФЕССИОНАЛЬНОЙ ТРАЕКТОРИИ ГРАЖДАН**

|  |  |
| --- | --- |
| **Цели получения персонального цифрового сертификата** | |
| **текущий статус** | **цель** |
| **Развитие компетенций в текущей сфере занятости** | |
| работающий по найму в организации, на предприятии | сохранение текущего рабочего места |
| работающий по найму в организации, на предприятии | развитие профессиональных качеств |
| работающий по найму в организации, на предприятии | повышение заработной платы |
| работающий по найму в организации, на предприятии | смена работы без изменения сферы профессиональной деятельности |

**VII.Дополнительная информация**

**VIII.Приложенные Скан-копии**

Утвержденной рабочей программа (подпись, печать, в формате pdf)

Загружено на платформу:

1. Два рекомендательных письма

2. Скан утвержденной образовательной программы

3. Скан утвержденного паспорта образовательной программы

4. Паспорт образовательной программы в формате word

**Генеральный директор**

**АНО ДПО «СофтЛайн Эдюкейшн» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.Э. Разуваев**

М.П.