**ПРИЛОЖЕНИЕ 1**

**ШКАЛА ОЦЕНИВАНИЯ ТЕСТА (рубежный контроль)**

1. В одном тестовом задании некоторое количество закрытых вопросов.
2. К заданиям даются готовые ответы на выбор, один правильный и остальные неправильные.
3. Обучающемуся необходимо помнить: в каждом задании с выбором одного правильного ответа правильный ответ должен быть.
4. За каждый правильно ответ – 5 баллов
5. Определяется сумма набранных баллов.
6. Вычисляется отношение набранных баллов к максимально возможным за тест. Отметка выражается в %.

**ШКАЛА ОЦЕНИВАНИЯ ПРАКТИЧЕСКИХ РАБОТ (текущий контроль)**

* 85-100 % - Демонстрирует полное понимание проблемы. Все требования, предъявляемые к заданию выполнены.
* 70-84 % - Демонстрирует значительное понимание проблемы. Все требования, предъявляемые к заданию выполнены.
* 60-69 % - Демонстрирует частичное понимание проблемы. Большинство требований, предъявляемых к заданию выполнены.
* 31-60 % - Демонстрирует небольшое понимание проблемы. Многие требования, предъявляемые к заданию не выполнены.
* 0-30 % - Демонстрирует непонимание проблемы и даже не было попытки решить задачу.

**ПРИЛОЖЕНИЕ 2**

**Примеры тестовых заданий**

**(выбрать правильный вариант или указать число)**

1. О соотношении аналоговой и цифровой информации:

* 1. Большинство данных в мире в 2011 году содержалось:
     1. В цифровом виде
     2. В аналоговом виде
  2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?
  3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
     1. Петабайтами
     2. Зеттабайтами
     3. Экзабайтами
     4. Йоттабайтами
  4. Сколько Петабайт в Зеттабайте?

1. История больших данных
   1. Укажите фактор, способствовавший появлению тренда больших данных
      1. Маркетинговые кампании крупных корпораций
      2. Снижение издержек на хранение данных
      3. Появление новых технологий обработки потоковых данных
      4. Выпуск баз данных с обработкой данных в памяти
   2. Какие вероятные разочарования тренда больших данных?
      1. Из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных.
   3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
      1. Разработка Hadoop
      2. Изобретение принципа MapReduce
      3. Разработка языка Pyhton
      4. Победа Deepblue в матче с Г.Каспаровым.
2. Определение больших данных:
   1. Выберите верный ответ
      1. Большие данные – это обработка или хранение более 1 Тб информации.
      2. Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущност-ная обработка данных затруднена или невозможна.
      3. Большие данные – это огромная PR-акция крупных вендоров и не более того.
      4. Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.
   2. Выберите неверный ответ:
      1. Большие данные – это данные объёма свыше 1 Тб
      2. Проблема больших данных – это проблема, когда при суще-ствующих технологиях хранения и обработки сущностная об-работка данных затруднена или невозможна.
      3. Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.
      4. Большие данные как правило не структурированы.
   3. Отметьте те из вариантов, в которых данные структурированы:
      1. Данные о продажах компании, представленные в виде поме-сячных отчётов в формате MS Word.
      2. Таблица с ежедневными показаниями температуры помеще-ния за год в файле формата csv.
      3. Текст педагогической поэмы А.С. Макаренко, представлен-ный в формате PDF.
      4. Библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.
3. Характеристики Big Data:
   1. Перечислите четыре основных характеристики Big Data:
      1. Virtualization, Volume, Variability, Vehicle
      2. Variety, Velocity, Volume, Value
      3. Verification, Volume, Velocity, Visualization
      4. Video, Value, Variety, Volume
   2. Выберите неверное высказывание:
      1. Большие объёмы данных приводят к слабой их структуриза-ции, поэтому появляется такое разнообразие данных.
      2. Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
      3. Удешевление систем хранения на единицу информации при-вело к росту рынка больших данных.
      4. Большое разнообразие источников данных
   3. Отметьте неверное понимание Variety в контексте характеристик

Big Data:

* + 1. Высокая скорость генерирования данных.
    2. Разные типы данных в колонках таблиц реляционных СУБД.
    3. Разнообразие отраслей, являющихся источниками данных.
    4. Разнообразие типов данных, включающих в себя структури-рованные, полуструктурированные и неструктурированные.

1. Принцип MapReduce
   1. Принцип MapReduce состоит в том, чтобы
      1. Производить вычисления на узлах, где информация изначаль-но была сохранена
      2. Использовать вычислительные мощности систем хранения
      3. Использовать функциональное программирование для реше-ния задач массивно-параллельной обработки
   2. Выберите одно неверное высказывание про MapReduce:
      1. Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изна-чально была сохранена
      2. MapReduce – это две операции: распределения и сборки дан-ных
      3. MapReduce был придуман разработчиками Hadoop
      4. MapReduce был анонсирован разработчиками Google
   3. Каков теоретический прирост производительности при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум?
2. Технологии хранения
   1. Какие из следующих технологий СУБД не используют принцип

MapReduce

* + 1. Hadoop
    2. Cassandra
    3. HDInsight
    4. Redis
  1. Какие СУБД полностью полагаются на оперативную память при хранении информации:
     1. Oracle Exalytics
     2. SAP HANA
     3. BigTable
     4. HBase
  2. В чём преимущество колоночно-ориентированных СУБД?
     1. Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
     2. Они позволяют динамически дополнять содержание записей новыми полями
     3. Они имеют более гибкие возможности аналитики.
     4. Они позволяют эффективно делать межколоночные сравне-ния.

1. «Песочница» в аналитическом процессе
   1. Для чего аналитику необходима «песочница»?
      1. Для высокопроизводительной аналитики за счёт использова-ния оперативной памяти и inDB операций.
      2. Для хранения всех полученных от заказчика данных.
      3. Для построения отчётов о результатах анализа
      4. Для снижения затрат, связанных с репликацией данных
   2. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более

100Гб:

* + 1. Hadoop
    2. Data Warehouse
    3. «Песочница»
    4. Python
  1. Выберите верное утверждение:
     1. Data Warehouse создаются для проверки гипотез при анализе больших данных.
     2. «Песочница» используется для снижения нагрузки на основ-

ной Data Warehouse.

* + 1. Каждый Data Warehouse должен содержать «песочницу».
    2. «Песочница» необходима для любого процесса аналитики.

1. CRISP-DM
   1. Расставьте последовательность этапов проекта аналитики в соот-ветствии с CRISP-DM.
      1. Понимание бизнеса (Business understanding)
      2. Понимание данных (Data Understanding)
      3. Подготовка данных (Data Preparation)
      4. Моделирование (Modeling)
      5. Оценка (Evaluation)
      6. Внедрение (Deployment)
   2. На каком из этапов процесса CRISP-DM происходит проверка гипо-тез?
      1. Понимание бизнеса (Business understanding)
      2. Понимание данных (Data Understanding)
      3. Моделирование (Modeling)
      4. Оценка (Evaluation)
   3. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. про-даж). Какие из этапов CRISP-DM можно опустить:
      1. Понимание бизнеса (Business understanding)
      2. Подготовка данных (Data Preparation)
      3. Моделирование (Modeling)
      4. Оценка (Evaluation)
2. Hadoop
   1. Пример благоразумного использования Hadoop
      1. Анализ 10 Гб данных.
      2. Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
      3. Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
      4. Построение графика пульса пациента в реальном времени.
   2. Начиная с каких размеров данных обоснованно применение класте-ра Hadoop для хранения данных?
      1. 100Гб
      2. 1Тб
      3. 100Тб
      4. 1Пб
   3. Hadoop – это:
      1. Набор утилит, и программный каркас для выполнения рас-пределённых программ, работающих на кластерах.
      2. Распределённая СУБД, позволяющая обрабатывать большие данные.
      3. Язык выполнения заданий в парадигме MapReduce.
      4. Распределённая файловая система, предназначенная для хранения файлов большого объёма.

# ПРИЛОЖЕНИЕ 3

# Практическая работа 1. Библиотека NumPy

1. **Задания:**
2. Загрузите массив из файла
3. Измените форму массива
4. Упорядочите элементы массива
5. Разбейте и объедените массив

# Практическая работа 2. Анализ набора данных usa.gov data from Bitly

1. 2011 году служба сокращения URL-адресов bit.ly заключила партнерское соглашение с сайтом правительства США USA.gov (<https://www.usa.gov/>) о син-хронном предоставлении анонимных данных о пользователях, которые со-кращают ссылки, заканчивающиеся на .gov или .mil. В 2011 году, помимо синхронной ленты, формировались ежечасные мгновенные снимки, доступ-ные в виде текстовых файлов. В 2017 году эта служба уже закрылась, но мы сохранили файлы данных и приводим их в качестве примеров.
2. **Задания**
3. Загрузите данные из файла
4. Подсчитайте количество часовых поясов

# Практиктическая работа 3. Анализ набора данных MovieLens 1M

Исследовательская группа GroupLens Research ([https://grouplens.org/datasets/](https://grouplens.org/datasets/movielens/) [movielens/](https://grouplens.org/datasets/movielens/)) предлагает несколько наборов данных о рейтингах фильмов, проставленных пользователями сайта MovieLens в конце 1990-х – начале 2000-х годов. Наборы содержат рейтинги фильмов, метаданные о фильмах (жанр и год выхода) и демографические данные о пользователях (возраст, почтовый индекс, пол и род занятий). Такие данные часто представляют интерес для разработки систем рекомендования, основанных на алгоритмах машинного обучения. И хотя в этой книге методы машинного обучения не рассматриваются, я все же покажу, как формировать продольные и попереч-ные разрезы таких наборов данных с целью привести их к нужному виду.

**Задания**

1. Загрузите данные из zip-файла
2. Выведите начало и конец данных
3. Выполните статистический анализ данных
4. Измерьте несогласия в оценках

**Практиктическая работа 4. Анализ имен, которые давали детям в США за период с 1880 по 2010 год**

Управление социального обеспечения США выложило в Сеть данные о часто-те встречаемости детских имен за период с 1880 года по настоящее время. Хэдли Уикхэм (Hadley Wickham), автор нескольких популярных пакетов для R, часто использует этот пример для иллюстрации манипуляций с данными в R.

**Задания**

1. Создайте набор данных
2. Постройте график общего количество родившихся в зависимости от пола и года рождения
3. Проанализируйте тенденции в выборе имен

# Практиктическая работа 5. База данных о продуктах питания министерства сельского хозяйства США

Министерство сельского хозяйства США публикует данные о пищевой цен-ности продуктов питания. Программист Эшли Уильямс (Ashley Williams) пре-образовал эту базу данных в формат JSON. Записи выглядят следующим об-разом:

{

"id": 21441,

"description": "KENTUCKY FRIED CHICKEN, Fried Chicken, EXTRA CRISPY, Wing, meat and skin with breading",

"tags": ["KFC"],

"manufacturer": "Kentucky Fried Chicken",

"group": "Fast Foods",

"portions": [

{

"amount": 1,

"unit": "wing, with skin",

"grams": 68.0

},

...

],

"nutrients": [

{

"value": 20.8,

"units": "g",

"description": "Protein",

"group": "Composition"

},

...

]

}

Для каждого продукта питания есть ряд идентифицирующих атрибутов и два списка: питательные элементы и размеры порций. Для анализа данные в та-кой форме подходят плохо, поэтому необходимо их переформатировать.

**Задания**

1. Загрузите архив с сайта
2. Покажите распределение продуктов питания по группам
3. Постройте медианные значения цинка по группе питательных элементов

# Практическая работа 6. Анализ базы данных федеральной избирательной комиссии

Федеральная избирательная комиссия США публикует данные о пожертво-ваниях участникам политических кампаний. Указывается имя жертвователя, род занятий, место работы и сумма пожертвования. Интерес представляет на-бор данных, относящийся к президентским выборам 2012 года.

**Задания**

1. Загрузите набор данных из файла
2. Покажите статистику пожертвований по роду занятий и месту работы
3. Постройте график Общая сумма пожертвований по партиям для родов занятий с максимальной суммой пожертвований

**Приложение 4**

**Технологическая карта дисциплины Аналитика больших данных**

**2 семестр**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Название модулей  дисциплины согласно РПД** | **Контроль** | **Форма контроля** | **зачетный  минимум** | **зачетный  максимум** | **График контроля** |
| 1. Введение в большие данные | текущий | Сдача практических работ №1 | 5 | 8 | 26 |
| рубежный | Тест №1 | 5 | 8 | 26 |
| 1. Жизненный цикл проекта по аналитике больших данных | текущий | Сдача практических работ №2 | 5 | 9 | 29 |
| рубежный | Тест №2 | 5 | 9 | 29 |
| 3. Основные техники работы с большими данными | текущий | Сдача практических работ №3, №4, №5 | 5 | 9 | 33 |
| рубежный | Тест №3 | 5 | 9 | 33 |
| 4. Приложение больших данных | текущий | Сдача практических работ №6 | 5 | 9 | 35 |
| рубежный | Тест №4 | 5 | 9 | 35 |
| ВСЕГО за семестр | | | 40 | 70 |  |
| Промежуточный контроль (Экзамен) | | | 20 | 30 |  |
| Семестровый рейтинг по дисциплине | | | 60 | 100 |  |

85 – 100 баллов– «отлично»

70 – 84 баллов– «хорошо»

60-69 баллов– «удовлетворительно»

менее 60 баллов– «неудовлетворительно»