

Министерство науки и высшего образования Российской Федерации  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Факультет инновационных технологий

УТВЕРЖДАЮ:  
Руководитель ОПОП

  
В. И. Сырямкин  
« 13 » мая 2023 г.

Оценочные материалы  
текущего контроля и промежуточной аттестации по дисциплине

**Анализ больших данных**

по направлению подготовки

**27.03.02 Управление качеством**

Направленность (профиль) подготовки:  
**Управление качеством в производственно-технологических системах**

Форма обучения

**Очная**

Квалификация

**Бакалавр**

## 1. Планируемые результаты освоения дисциплины

Результаты освоения дисциплины (индикатор достижения компетенции)	Планируемые образовательные результаты (ОР) обучения по дисциплине
ИОПК-7.1 Понимает принцип работы современных информационных технологий.	ОР 7.1.1 – Применяет современные среды разработки в профессиональной деятельности ОР 7.1.2 – Понимает принципы работы в современных информационных системах для решения прикладных задач в профессиональной деятельности
ИОПК-7.2 Знает и способен применять современные программные платформы в области профессиональной деятельности.	ОР 7.2.1 – Применяет современные программные платформы и алгоритмы анализа информации для решения прикладных задач
ИОПК-8.1 Владеет методами сбора и анализа информации в области управления качеством продукции, процессов, услуг.	ОР 8.1.1 – Осуществляет отбор и анализ материала, а также решает задачи в области управления качеством продукции, процессов, услуг
ИОПК-8.2 Владеет методами оценки профессиональной информации.	ОР 8.2.1 – Использует современные алгоритмы и методы оценки для анализа исходной информации в профессиональной деятельности
ИПК-1.1 Умеет собирать, систематизировать и анализировать данные по показателям качества, характеризующим разрабатываемую и выпускаемую продукцию (работы, услуги), в том числе, с использованием средств и технологий цифровизации.	ОР 1.1.1 – Собирает и систематизирует информацию в области контроля качества ОР 1.1.2 – Использует современные средства анализа данных в профессиональной деятельности

## 2. Этапы достижения образовательных результатов в процессе освоения дисциплины

№	Разделы и(или) темы дисциплин	Образовательные результаты	Формы текущего контроля и промежуточной аттестации
	Тема 1. Большие данные (введение)	ОР 7.1.1	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 2. Методики анализа больших данных	ОР 7.1.1 ОР 7.1.2	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 3. Инструменты Больших данных	ОР 7.1.1 ОР 7.1.2	Текущий контроль: Тест  Промежуточная аттестация:

			зачет
	Тема 4. Технологии хранения и обработки больших данных	ОР 7.1.2	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 5. Вычислительное ядро Hadoop	ОР 1.1.1	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 6. Скрипты Pig	ОР 7.1.2	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 7. Базы данных	ОР 1.1.1	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Тема 8. Озеро данных	ОР 7.1.1	Текущий контроль: Тест  Промежуточная аттестация: зачет
	Практическая работа №1	ОР 7.1.1 ОР 7.1.2 ОР 7.2.1 ОР 8.2.1 ОР 1.1.2	Текущий контроль: Отчет по практической работе  Промежуточная аттестация: зачет
	Практическая работа №2	ОР 7.1.2 ОР 7.2.1 ОР 8.1.1 ОР 1.1.1 ОР 1.1.2	Текущий контроль: Отчет по практической работе  Промежуточная аттестация: зачет
	Практическая работа №3	ОР 7.1.2 ОР 7.2.1 ОР 8.1.1 ОР 8.2.1 ОР 1.1.1 ОР 1.1.2	Текущий контроль: Отчет по практической работе  Промежуточная аттестация: зачет

### 3. Оценочные средства для проведения текущего контроля и методические материалы, определяющие процедуру их оценивания

Текущий контроль проводится в течение семестра с целью определения уровня усвоения обучающимися знаний, формирования умений и навыков, своевременного выявления преподавателем недостатков в подготовке обучающихся и принятия необходимых мер по ее корректировке, а также для совершенствования методики обучения, организации учебной работы, и фиксируется в форме контрольной точки не менее одного раза в семестр.

Текущий контроль включает в себя: тестовые задания, посещаемость, самостоятельную работу.

Для проведения текущего контроля используется:

Типовые задания для проведения текущего контроля успеваемости по дисциплине (тестовые задания).

#### Тест №1

1. Впервые термин «большие данные» появился в прессе в ..... году, когда редактор журнала Nature Клиффорд Линч выпустил статью на тему развития будущего науки с помощью технологий работы с большим количеством данных.\
  - A. 2008
  - B. 1998
  - B. 2000
  - Г. 2014
2. Основные источники информации для Big Data. Выберите один или несколько ответов:
  - A. интернет-коммерция
  - Б. телекоммуникации
  - В. финансовая сфера
  - Г. ритейл
3. Структурированные и неструктурированные данные огромных объёмов и значительного многообразия это...
4. Соотнесите важнейшие направления Big Data и их определения
  - A. Выберите... возможность одновременно обрабатывать различные типы данных. скорость прироста и необходимости быстрой обработки данных для получения результатов. величина физического объёма
  - Б. Выберите... возможность одновременно обрабатывать различные типы данных. скорость прироста и необходимости быстрой обработки данных для получения результатов. величина физического объёма
  - В. Выберите... возможность одновременно обрабатывать различные типы данных. скорость прироста и необходимости быстрой обработки данных для получения результатов. величина физического объёма
5. Соотнесите методики анализа больших данных и их определения

1. Genetic algorithms
2. Machine learning
3. Visualization
4. Crowdsourcing

А. В этой методике возможные решения представляют в виде «хромосом», которые могут комбинироваться и мутировать. Как и в процессе естественной эволюции, выживает наиболее приспособленная особь.

Б. Направление, которое преследует цель создания алгоритмов самообучения на основе анализа эмпирических данных

В. Методы графического представления результатов анализа больших данных в виде диаграмм или анимированных изображений

Г. Методика сбора данных из большого количества источников.

Д. Набор методик, который позволяет определить наиболее восприимчивые для продвигаемого продукта или услуги категории потребителей, выявить особенности наиболее успешных работников, предсказать поведенческую модель потребителей

6. Технология выявления скрытых взаимосвязей внутри больших баз данных это...

## Тест №2

- 1) Модель распределённых вычислений, представленная компанией Google, используется компанией в компьютерных кластерах для параллельных вычислений над очень большими, даже несколько петабайт, наборами данных это...
- 2) Соотнесите шаги MapReduce и их действие

А. Происходит свёртка предварительно обработанных данных.

Б. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map.

В. Происходит предварительная обработка входных данных.

- 3) Проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов это...
- 4) Соотнесите основные компоненты Hadoop и их предназначение

1. Hadoop Distributed File System

2. Hadoop YARN

А. фреймворк для управления ресурсами кластера и менеджмента задач, в том числе включает фреймворк MapReduce.

Б. распределённая файловая система, позволяющая хранить информацию практически

В. библиотеки управления файловыми системами, поддерживаемыми Hadoop, и сценарии создания необходимой инфраструктуры и управления распределённой обработкой.

5) База данных, в которой в отличие от большинства традиционных систем баз данных не используется табличная схема строк и столбцов. В этих базах данных применяется модель хранения, оптимизированная под конкретные требования типа хранимых данных.

- А. Реляционная база данных
- Б. Нереляционная база данных
- В. База данных в памяти

б) Соотнесите типы баз данных NoSQL с их предназначением

Документно-ориентированные  
Графовые  
Столбцовые (колоночные)  
Хранилище «Ключ-значение»

- А. Предназначены для хранения иерархических структур данных.
- Б. Предназначены для обеспечения удобства создания и запуска приложений с наборами сложносвязных данных.
- В. Данные хранятся в виде разреженной матрицы, строки и столбцы которой используются как ключи.
- Г. Тип баз данных, в котором для хранения данных используется простой метод «ключ-значение».

7. In-memory database - это тип нереляционной базы данных, которая опирается главным образом на (кэш **оперативную** постоянную) память для хранения данных

#### 4. Оценочные средства для проведения промежуточной аттестации

Темы и содержание практических работ

Практическая работа №1 Terminal

1. Выделите и опишите основные преимущества развёртывания кластера Hadoop в «облаке». Составьте краткий отчет.
2. Скачайте образ виртуальной машины Cloudera QuickStart (скачать) предоставленный спонсорами для образовательных целей.
3. Установите и запустите виртуальную машину Cloudera QuickStart. Составьте краткий отчет.
4. Создайте в HDFS рабочую папку "lab1".
5. Произведите загрузку в HDFS всех файлов из архива data\_lab1.zip в созданную ранее директорию. Выведите на экран первые 15 строчек файла.
6. Изучите код mkdir.java из вложения hdfs\_mkdir.zip. Используя скомпилированный создайте рабочую директорию lab1\_files. Опишите вывод работы jar-пакета при его корректном и некорректном использовании, а также в случаях, когда директория уже существует.

Практическая работа №2 MapReduce

1. Запустите скомпилированный WordCount.jar пакет используя YARN.
2. Запустите python скрипты mapper.py и reducer.py в виде hadoop-streaming задачи для данных приложенных в архиве.
3. Опишите каким образом необходимо изменить код WordCount.java, чтобы скомпилированный пакет можно было запускать с аргументами входная и выходная директория?
4. 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.

### Практическая работа №3 Pig Latin

1. Произведите обработку файла 2018.txt или 2019.txt из архива, data\_lab3.zip с помощью скрипта Pig latin:
2. Произведите загрузку.
3. Извлеките первые 30 строк файла.
4. Выведите их на экран.
5. Произведите группировку по признаку DATE.
6. Произведите анализ усреднения по выделенным группам.
7. Произведите сортировку результатов.
8. Выведите на экран 10 строк результата.
9. Повторите операции для файлов из архива lab3\_variant.zip согласно вашему варианту. Совместно с усреднением используйте также агрегирующие функции минимума и максимума.

### Критерии оценивания

Оценка	Характеристика ответа
Зачтено	Работа выполнена полностью. Студент владеет теоретическим материалом, отсутствуют ошибки при описании теории, формулирует собственные, самостоятельные, обоснованные, аргументированные суждения, допуская незначительные ошибки на дополнительные вопросы
Не зачтено	Работа выполнена полностью. Студент не владеет теоретическим материалом, допуская грубые ошибки, испытывает затруднения в формулировке собственных суждений, не способен ответить на дополнительные вопросы