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

Факультет инновационных технологий

УТВЕРЖДАЮ:
Руководитель ОПОП
_____ С. В. Шидловский
«_29_»____08_____ 2022 г.

Оценочные материалы
текущего контроля и промежуточной аттестации по дисциплине

Анализ больших данных

по направлению подготовки

09.04.02 Информационные системы и технологии

Направленность (профиль) подготовки :
Компьютерная инженерия: искусственный интеллект и робототехника

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Томск – 2022

1. Планируемые результаты освоения дисциплины

Результаты освоения дисциплины *(индикатор достижения компетенции)*	**Планируемые образовательные результаты (ОР) обучения по дисциплине**
ИОПК 2.1 Владеет методами алгоритмизации и программирования	ОР 2.1.1 – Использует современные алгоритмы анализа информации для решения прикладных задач ОР 2.1.2 – Применяет современные языки программирования для решения профессиональных задач.
ИОПК 2.2 Знает современные подходы, методы и технологии в области интеллектуального анализа данных	ОР 2.2.1 – Осуществляет отбор и анализ материала для профессиональной деятельности. ОР 2.2.2 – Применяет современные технологии анализа данных в профессиональной деятельности
ИОПК 2.3 Использует методы современных интеллектуальных технологий для решения профессиональных задач	ОР 2.3.1 – Применяет современные методы анализа и профессиональные средства анализа данных в профессиональной деятельности

2. Этапы достижения образовательных результатов в процессе освоения дисциплины

№	Разделы и(или) темы дисциплин		Формы текущего контроля и промежуточной аттестации
1.	Тема 1. Большие данные (введение)	ОР 2.1.1	Текущий контроль: Тест
2.	Тема 2. Методики анализа больших данных	ОР 2.1.1	Текущий контроль: Тест
3.	Тема 3. Инструменты Больших данных	ОР 2.1.1	Текущий контроль: Тест
4.	Тема 4. Технологии хранения и обработки больших данных	ОР 2.1.1	Текущий контроль: Тест
5.	Тема 5. Вычислительное ядро Hadoop	ОР 2.1.1, ОР 2.2.2	Текущий контроль: Тест
6.	Тема 6. Скрипты Pig	ОР 2.1.1, ОР 2.2.2, ОР 2.3.1	Текущий контроль: Тест
7.	Тема 7. Базы данных Hadoop.	ОР 2.1.1, ОР 2.2.2 , ОР 2.3.1	Текущий контроль: Тест
8.	Тема 8. Озеро данных	ОР 2.1.1	Текущий контроль: Тест
9.	Практическая работа №1 Terminal	ОР 2.1.1, ОР 2.1.2, ОР 2.2.2	Текущий контроль: Отчет по практической работе
10.	Практическая работа №2 MapReduce	ОР 2.1.1, ОР 2.1.2, ОР 2.2.1, ОР 2.2.2, ОР 2.3.1	Текущий контроль: Отчет по практической работе
11.	Практическая работа №3 Pig Latin	ОР 2.1.1, ОР 2.1.2, ОР 2.2.1, ОР 2.2.2, ОР 2.3.1	Текущий контроль: Отчет по практической работе
12.	Практическая работа №4 SQL Hive	ОР 2.1.1, ОР 2.1.2, ОР 2.2.1, ОР 2.2.2, ОР 2.3.1	Текущий контроль: Отчет по практической работе

13.	Практическая работа №5. Итоговое задание	ОР 2.1.1, ОР 2.1.2, ОР 2.2.1, ОР 2.2.2, ОР 2.3.1	Текущий контроль: Отчет по практической работе

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 (тесты и выполнение практических заданий).

3.1. Тест №1

1) Впервые термин «большие данные» появился в прессе в [2008] году, когда редактор) журнала Nature Клиффорд Линч выпустил статью на тему развития будущего науки с помощью технологий работы с большим количеством данных.

2) Основные источники информации для Big Data. Выберите один или несколько ответов:

a.интернет-коммерция

b.телекоммуникации

c.финансовая сфера

d. ритейл

3) Структурированные и неструктурированные данные огромных объёмов и значительного многообразия это...

4) Соотнесите важнейшие направления Big Data и их определения

Variety	Ответ: возможность одновременно обрабатывать различные типы данных.	1
Velocity	Ответ: скорость прироста и необходимости быстрой обработки данных для получения результатов.	2
Volume	Ответ: величина физического объёма	3

5) Соотнесите методики анализа больших данных и их определения

Genetic algorithms

Ответ 1

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

Machine learning

Ответ 2

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

Visualization

Ответ 3

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

Crowdsourcing

Ответ 4

Методика сбора данных из большого количества источников.

Data mining

Ответ 5

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

Reduce

Ответ 1

Происходит свёртка предварительно обработанных данных.

Shuffle

Ответ 2

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

Map

Ответ 3

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...

Соотнесите основные компоненты Hadoop и их предназначение

Hadoop YARN

Ответ 1

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

Hadoop Distributed File System

Ответ 2

распределённая файловая система, позволяющая хранить информацию практически неограниченного объёма.

Hadoop Common

Ответ 3

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

4) 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.

Выберите один ответ:

a. Реляционная база данных

b. Нереляционная база данных

c. База данных в памяти

Соотнесите типы баз данных NoSQL с их предназначением

Документо-ориентированные

Ответ 1

Предназначены для хранения иерархических структур данных.

Графовые

Ответ 2

Предназначены для обеспечения удобства создания и запуска приложений с наборами сложно связных данных.

Столбцовые (колоночные)

Ответ 3

Данные хранятся в виде разреженной матрицы, строки и столбцы которой используются как ключи.

Хранилище «Ключ-значение»

Ответ 4

Тип баз данных, в котором для хранения данных используется простой метод «ключ- значение».

5) In-memory database - это тип нереляционной базы данных, которая опирается главным образом на оперативную 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 jar-пакет hdfs_client.jar с помощью команды «hadoop jar hdfs_client.jar mkdir [Directory_Path]» 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:
    1. Произведите загрузку.
    2. Извлеките первые 30 строк файла.
    3. Выведите их на экран.
    4. Произведите группировку по признаку DATE.
    5. Произведите анализ усреднения по выделенным группам.
    6. Произведите сортировку результатов.
    7. Выведите на экран 10 строк результата.
2. Повторите операции для файлов из архива lab3_variant.zip согласно вашему варианту. Совместно с усреднением используйте также агрегирующие функции минимума и максимума.

Практическая работа №4 SQL Hive

1. Произведите обработку файла 2018.txt, из архива приложенного к заданию, с помощью скрипта Pig latin:

- Создайте новую базу данных
- Создайте схему реляционной таблицы
- Произведите загрузку данных
- Извлеките первые 10 строк файла
- Создайте view с группировкой по признаку DATE и анализом усреднения по выделенным группам
- Сделайте запрос к view и произведите сортировку результатов
- Сохраните результат в таблице

2. Повторите операции для других файлов архива, согласно своему варианту. Совместно с усреднением используйте также агрегирующие функции минимума и максимума.

Практическая работа №5. Итоговое задание

В этом задании вам нужно продемонстрировать умение использования компонент Hadoop:

- HDFS
- MapReduce
- Pig
- Hive

1. Выберите набор табличных данных и сохраните его (например, с помощью MS Excel) в текстовый формат (CSV). Это могут быть данные, связанные с Вашей деятельностью, открытые данные, модельные данные.
2. Произведите исследование по плану:

- Выберите вариант инфраструктуры Hadoop

- Произведите загрузку данных
- Проведите обработку данных средствами Hadoop
- Предложите варианты по обогащению (расширению набора признаков) имеющихся данных.

Приложите краткий отчет, содержащий описание данных, проблематику их накопления и обработки, шаги исследования (по плану выше), вывод.

Критерии оценивания зачета с оценкой:

Оценка «отлично» выставляется, при условии глубокого и прочного знания материала курса, исчерпывающего, последовательного, четкого и логически выстроенного ответа. При ответе на вопрос студент не только излагает материал, но умеет увязывать теорию с практикой, приводить примеры иллюстрирующие ответ. Студент свободно справляется с вычислительными задачами, не затрудняется с ответом при видоизменении заданий, использует в ответе материал из различных источников литературы, правильно обосновывает свои решения, владеет разносторонними навыками и приемами выполнения заданий по формированию профессиональных компетенций.

Оценка «хорошо» выставляется студенту, при условии твердого знания материала. Отвечая, студент грамотно и по существу, излагает материал курса, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические знания при решении практических задач, решает типовые задачи без ошибок, может затрудняться с ответом при видоизменении заданий, испытывает трудности в приведения практических примеров.

Оценка «удовлетворительно» выставляется студенту, когда он имеет знания только основного материала, использует в ответах неточные формулировки, при ответе есть нарушения логической последовательности в изложения вопроса, студент испытывает сложности при выполнении практических заданий, затрудняется связать теорию с практическими примерами.

Оценка «неудовлетворительно» выставляется студенту, который не знает большей части программного материала, неуверенно отвечает на вопрос, допускает грубые ошибки, не может решить типовые задачи.