

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

Факультет инновационных технологий

УТВЕРЖДЕНО:
Декан
С. В. Шидловский

Оценочные материалы по дисциплине

Анализ больших данных

по направлению подготовки / специальности

27.04.05 Инноватика

Направленность (профиль) подготовки/ специализация:
Отраслевой инжиниринг

Форма обучения
Очная

Квалификация
инженер-исследователь

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
Э.А. Соснин

Председатель УМК
О.В. Вусович

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

БК 2 – Способен использовать научные методы для решения профессиональных задач

ОПК 1 – Способен решать профессиональные задачи и использовать в практической деятельности новые знания и умения, полученные в том числе в областях, непосредственно не связанных с инновационной деятельностью

ОПК 2 – Способен аргументировано выбирать и обосновывать структурные, алгоритмические, технологические и программные решения для управления инновационными процессами и проектами, реализовывать их на практике применительно к инновационным системам предприятия, отраслевым и региональным инновационным системам

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

РОБК 2.1 Знает основные методы научных исследований

РООПК 1.3 Рассматривает и предлагает возможные варианты решения профессиональной задачи с привлечением математических методов и моделей, опираясь на стратегические задачи развития экономики

РООПК 2.3 Способен применять на практике алгоритмические, технологические и программные решения для управления инновационными процессами и проектами

2. Оценочные материалы текущего контроля и критерии 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 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 | 1 |
| | <input type="text" value="возможность одновременно обрабатывать различные типы данных."/> | |
| Velocity | Ответ | 2 |
| | <input type="text" value="скорость прироста и необходимости быстрой обработки данных для получения результатов."/> | |
| Volume | Ответ | 3 |
| | <input type="text" value="величина физического объёма"/> | |

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

б) 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 связанных 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 память для хранения данных

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

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.

1. 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 для 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 изложении 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.

Информация о разработчиках

Погуда Алексей Андреевич, доцент кафедры информационного обеспечения инновационной деятельности факультета инновационных технологий, кандидат технических наук.