МИНОБРНАУКИ РФ
Федеральное бюджетное государственное образовательное учреждение
высшего профессионального образования
«Восточно–Сибирский государственный университет технологий и управления»

Кафедра «Программная инженерия и искусственный интеллект»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к лабораторным работам по дисциплине

"Технология распределенной обработки данных"

Составитель: к.т.н., доц. Л.П. Бильгаева

Улан-Удэ
2023

Лабораторная работа №1. Установка платформы Hadoop.

Цель: получить навыки установки и настройки программного обеспечения Hadoop.

Задачи:

Установка платформы Hadoop:

- установите программу для виртуализации операционных систем Oracle VM;
- создайте виртуальную машину в VirtualBox, установите операционную систему Ubuntu (дистрибутив GNU/Linux);
- установите и настройте в виртуальной машине Ubuntu платформу Java и сетевой протокол SSH;
- установите Hadoop.

В качестве ответа на задание приложите один скриншот вашего полного экрана с запущенным терминалом, в котором выполнена команда jps, и запущенным браузером, в котором открыта страница localhost:8042.

Контрольные вопросы

1. Какие функции выполняет NameNode в Hadoop?
2. Что такое фактор репликации?
3. На сколько блоков разбиваются данные при загрузке в HDFS?
4. Какими командами можно считать данные из HDFS?

Лабораторная работа №2. Разработка приложения MapReduce.

Цель: получить навык создания MapReduce-приложения.

Задачи:

- Установите в виртуальной машине Ubuntu среду разработки «Eclipse».
- Напишите программу «WordCount» на языке программирования Java, которая считает количество слов в файле, хранящимся в HDFS.

В качестве ответа на задание приложите исходный код программы.

Контрольные вопросы

1. Какие функции выполняет ResourceManager в Hadoop?
2. Из каких частей состоит приложение MapReduce?
3. Что такое shuffle?

Лабораторная работа №3. Разработка статистических отчетов с использованием Apache Hive

Цель: получить навыки работы в СУБД Apache Hive и создания статистических отчетов.

Задачи:

1. Установить в виртуальной машине Ubuntu СУБД Apache Hive..

2. Создать в СУБД Apache Hive базу данных, согласно описанию, приведенному в приложенном файле.

В качестве ответа на задание приложите текст запросов для создания описанных в задании таблиц базы данных и скриншоты вашего полного экрана с запущенным в терминале Hive и выполненными запросами.

Контрольные вопросы

1. Какой язык запросов используется в Hive?
2. Что хранится в Metastore?
3. Может ли Hive хранить данные в HDFS?

Лабораторная работа №4. Анализ данных в Hadoop.

Цель: получить навыки разработки приложения для анализа и визуализации данных в Hadoop.

Задачи:

– Установить в виртуальной машине Ubuntu Apache Spark.

– Импортировать свободнораспространяемый датасет.

– Создать программу с помощью Apache Spark MLlib для анализа данных.

– Визуализировать результат в виде диаграммы.

В качестве ответа на задание приложите текст программы и скриншоты вашего полного экрана с визуализацией результата анализа в виде диаграммы.

Контрольные вопросы

1. Укажите иерархию между объектами рисунка Figure, Axes, Axis в matplotlib.
2. Как отобразить легенду к графику в matplotlib?
3. Какая команда в matplotlib рисует круговую диаграмму?