Оленчикова Т.Ю.

# Лабораторная работа 3.

# Hadoop. Разработка распределенных программ

**Цель:** Ознакомиться с технологией распоеденных вычислений и архитектурой фреймворка Hadoop/

### Порядок выполнения работы:

Напишите программу на Python, используя фреймворк Hadoop, согласно вашему варианту для датасета “brooklyn\_sales\_map.csv”. Для помощи в выполнении задания можно воспользоваться книгами Чак, Л. «Hadoop в действии» – 2012, Лебедев, А. С. «Методы Big Data» – 2021 или интернетом.

### Индивидуальные задания

**Вариант 1.**

* • Найдите среднюю стоимость жилья (sale\_price) и выведите новую таблицу, содержащую две колонки – стоимость жилья и отклонение стоимости от среднего значения.
* • Найдите среднюю стоимость жилья (sale\_price) для каждого района.
* • Выведите среднюю полную площадь жилья (gross\_sqft) для всех сочетаний налоговых категорий (tax\_class) и лет продажи (year\_of\_sale).
* • Выведите таблицу, содержащую количество пустых (null) значений для каждой колонки.

**Вариант 2.**

* • Найдите среднюю площадь жилья (gross\_sqft) и выведите новую таблицу, содержащую стоимость жилья и отклонение стоимости от среднего значения.
* • Найдите среднюю площадь жилья (gross\_sqft) для каждого года, в котором оно было построено (year\_built).
* • Найдите среднюю стоимость жилья (sale\_price) для всех сочетаний соседств (neighborhood) и категорий класса здания (building\_class\_category).
* • В исходном датафрейме удалите все строчки с записями домов, которые были построены позже 2000 года (year\_built), а также те, которые содержат только нулевые значения.

**Вариант 3.**

* • Найдите среднюю стоимость жилья (sale\_price) и выведите новую таблицу, содержащую две колонки – стоимость жилья и процент отклонения стоимости от среднего значения.
* • Выведите таблицу, содержащую все категории класса зданий (building\_class\_category) и количество записей, которые к ним относятся.
* • Выведите таблицу, содержащую средние значения по каждому столбцу в датафрейме.
* • В исходном датафрейме заполните все нулевые значения средними по столбцу.

**Вариант 4.**

* • Добавьте в датасет новую колонку, содержащую «возраст» жилья.
* • Выведите таблицу, содержащую среднюю дату продажи для всех сочетаний индексов (zip\_code) и налоговых категорий (tax\_class) жилья.
* • Выведите таблицу, содержащую суммарную стоимость жилья (sale\_price) по всем сочетаниям налоговых категорий (tax\_class) и индексов (zip\_code).
* • Создайте новую таблицу, в которой есть 10 колонок исходного датафрейма, в которых нулевые значения есть, но не преобладают. После этого удалите все строки, в которых содержатся исключительно нули.

**Вариант 5.**

* • Найдите средний год постройки жилья (year\_built) и выведите новую таблицу, содержащую год постройки жилья и отклонение года постройки от среднего значения.
* • Отсортировать датасет по возрастанию цены продажи (sale\_price) и убыванию индексов (zip\_code) одновременно.
* • Выведите таблицу с наибольшими ценами продажи (sale\_price) и количеством зданий по каждому сочетанию соседства (neighborhood) и категории класса здания (building\_class\_category

### Требования к оформлению отчета.

Отчет по ЛР состоит из

**а)** Титульный лист: название ЛР, ФИО студента и номер группы, № варианта.

б)Текст задания; в) UML-диаграмму классов (создаете в Umlet):

г) Исходный текст программы;

д) Скрины результатов выполнения;

е) Выводы.

### Критерии оценивания.

Оценка суммируется из следующих оценок:

1. Приложение отлажено и работоспособно -2 балла; имеются ошибки или реализованы не все функции – 1 балл; приложение не работоспособно – 0 баллов

2. Отчет оформлен в соответствии с требованиями – 1 балл, иначе – 0 баллов

3. Полные и верные ответы на контрольные вопросы – 1 балл, иначе – 0 баллов

Итого – максимум 4 балла.

Внимание! Полученная оценка автоматически снижается на 2% за каждую полную неделю задержки сдачи отчета по работе, но не более, чем на 40%.

### Контрольные вопросы

1. Что такое большие данные?
2. Объясните пять V больших данных
3. Сколько форматов ввода существует в Hadoop?
4. Что такое Thread?
5. Кто использует Hadoop?
6. Каковы основные функции Hadoop?
7. В каких трех режимах может работать Hadoop?
8. В каких операционных системах работает Hadoop?
9. Что делает JobTracker?
10. Расскажите о различных методах редуктора
11. Что делает команда jps?
12. Как вы проверяете NameNode при использовании команды jps?
13. В чем разница между левым полусоединением и внутренним соединением?
14. Как вы развертываете решение для работы с большими данными?
15. Каковы компоненты HDFS?
16. Какие минусы или опасные места HDFS?
17. Что входит в экосистему Hadoop?
18. Что является основным ядром всех продуктов Hadoop?
19. Что такое контрольная точка?
20. Как работает Hadoop MapReduce?
21. Что происходит при сбое узла данных?
22. Объясните, что такое спекулятивное исполнение
23. Укажите, в чем разница между СУБД и Hadoop
24. Объясните, как можно отлаживать код Hadoop
25. Объясните, что такое узлы хранения и вычисления
26. Укажите принципиальные различия между реляционной базой данных и HDFS?

### Литература

1. Чак, Л. Hadoop в действии / Л. Чак. — Москва : ДМК Пресс, 2012. — 424 с. – URL: URL: https://e.lanbook.com/book/39997 (дата обращения: 13.09.2023)
2. Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. — 91 с. — URL: https://e.lanbook.com/book/182452 (дата обращения: 13.09.2023)