Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Проектный практикум по разработке ETL-решений

Лабораторная работа 4.1

Работа с XML и JSON файлами

Выполнила: st\_98, группа: АДЭУ-211

Преподаватель:

Москва

2025

**Цель:** анализ данных с помощью DASK и визуализация ориентированных ациклических графов (DAG)

**Задачи:**

4.1.1. Настроить среду и рабочий каталог.

4.1.2. Загрузить данные.

4.1.3. Проверить качество данных (например, отсутствующие значения и выбросы).

4.1.4. Удалить столбцы (множество пропусков в значениях, бесполезные столбцы для анализа).

4.2.1. Визуализировать DAG с одним узлом и зависимостями.

4.2.2. Визуализировать DAG с более чем одним узлом и зависимостями.

**Вариант 11.** UK Property Price official data 1995-202304.zip.

Для начала необходимо настроить рабочий каталог, проверить файлы и установить dask[complete] (рисунок 1).

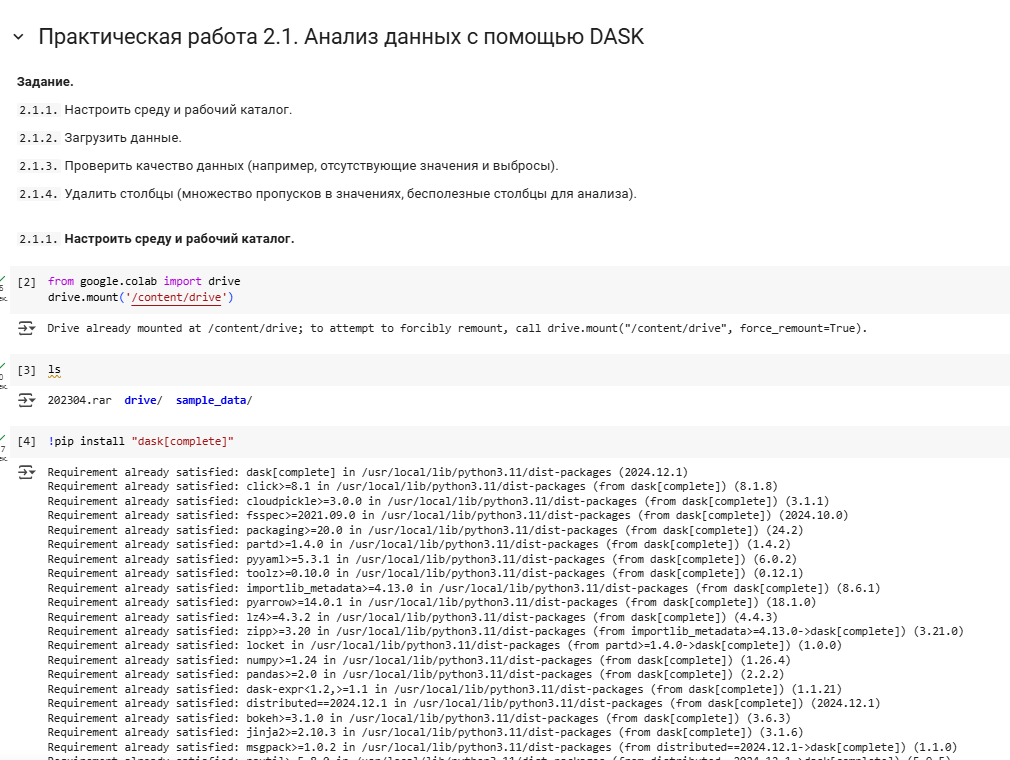


Рисунок 1 – Настройка среды и рабочего каталога

Далее нужно импортировать все библиотеки, загрузить данные. Так как загружался изначально архив, то надо будет его сначала распаковать (рисунок 2).



Рисунок 2 – Импорт библиотек и загрузка данных

Как можно заметить на рисунке 3, архив успешно распаковался



Рисунок 3 – Архив распаковался

На рисунке 4 показано изменение типов данных, а также чтение файла с помощью библиотеки Dask. Была разбивка на 6 партиций, так как файл в общей сложности весит 393/64 (64 мб блок) = 6.

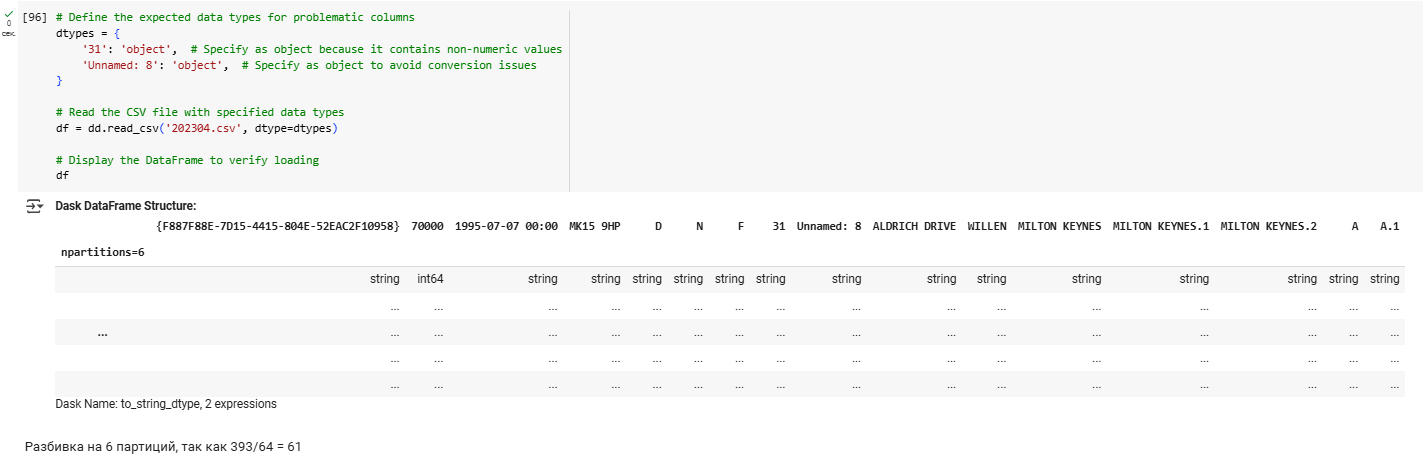


Рисунок 4 – Изменение типов данных

На рисунке 5 показана проверка качества данных

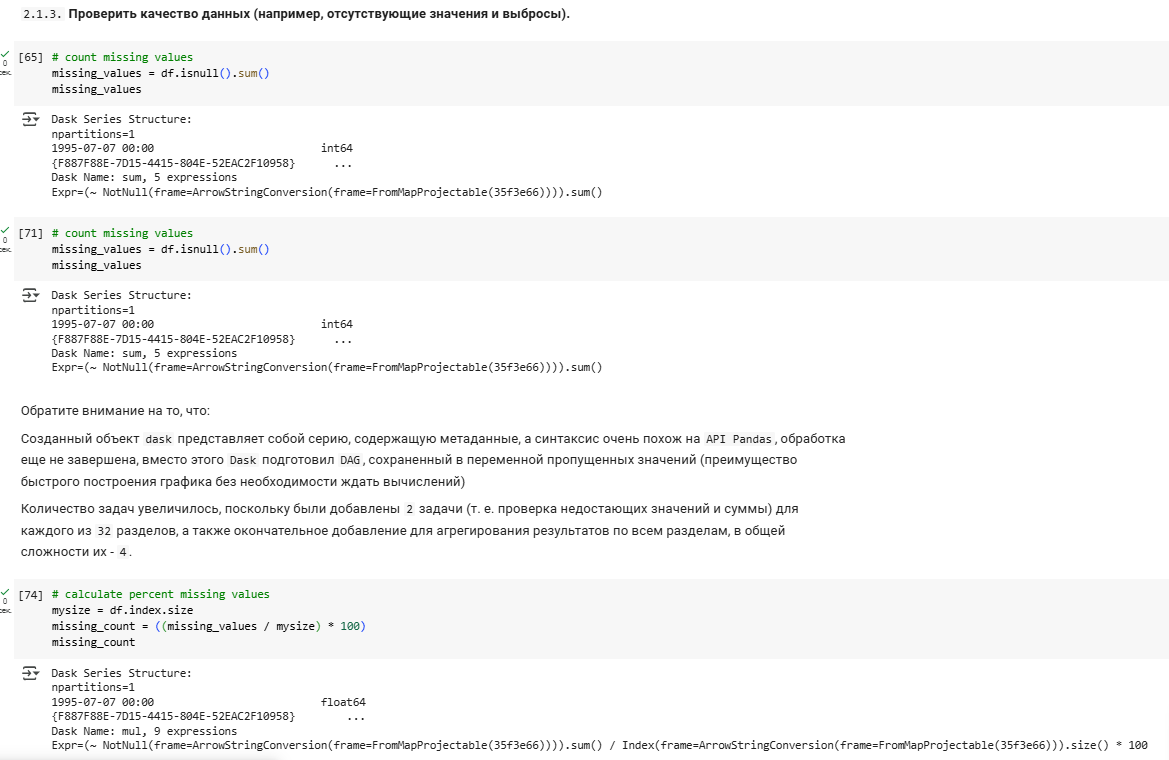


Рисунок 5 – Проверка качества данных

На рисунке 6 показан вывод всех столбцов.

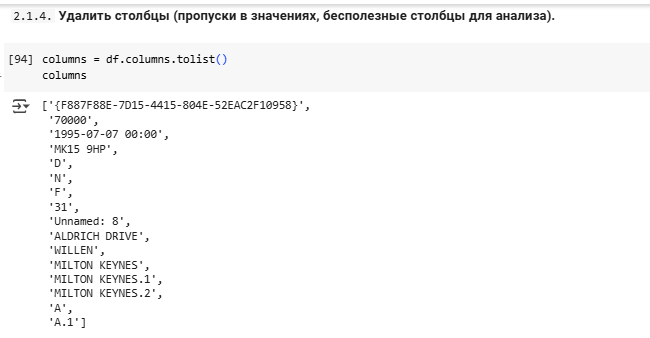


Рисунок 6 – Вывод всех столбцов

На рисунке 7 показан вывод 20 первых значений датасета

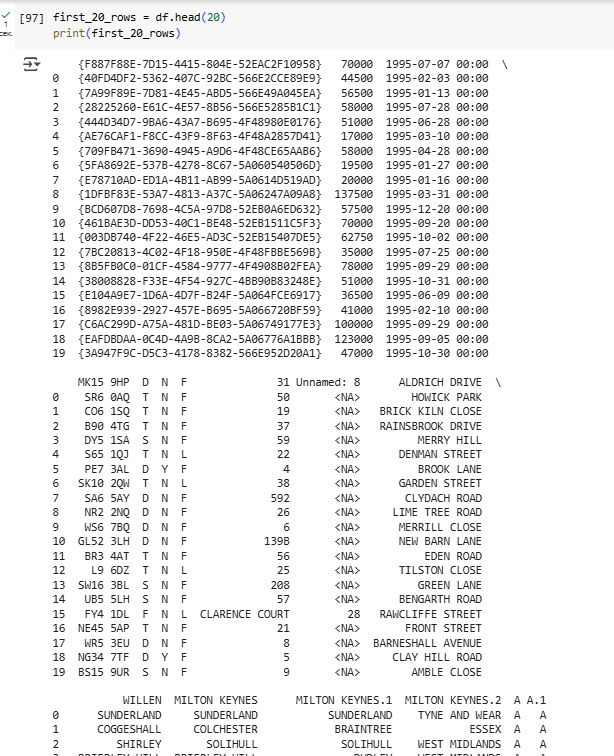


Рисунок 7 – Вывод датасета

Далее необходимо удалить столбец с именем Unnamed: 8, так как там пропуски (рисунок 8).

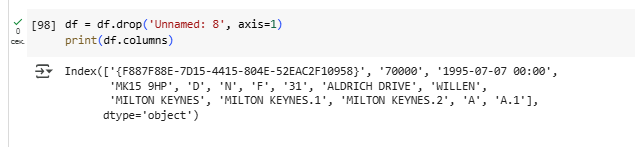


Рисунок 8 – Удаление столбца с именем Unnamed: 8

На рисунке 9 показан Dag с одним узлом и зависимостями.

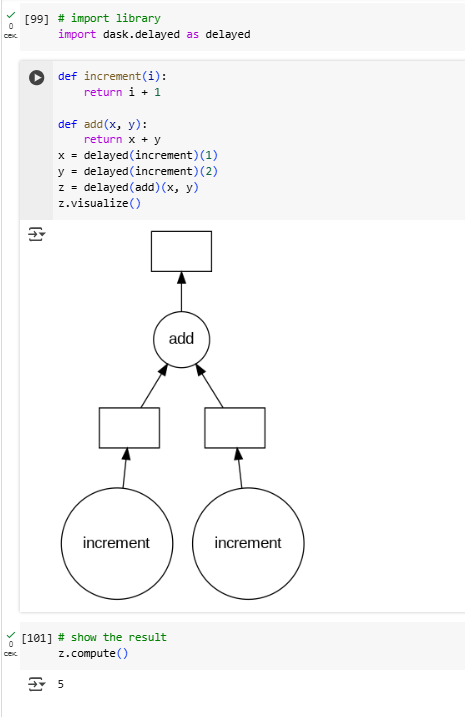


Рисунок 9 – Dag с одним узлом

На рисунке 10 показан первый слой Dag с несколькими зависимостями.

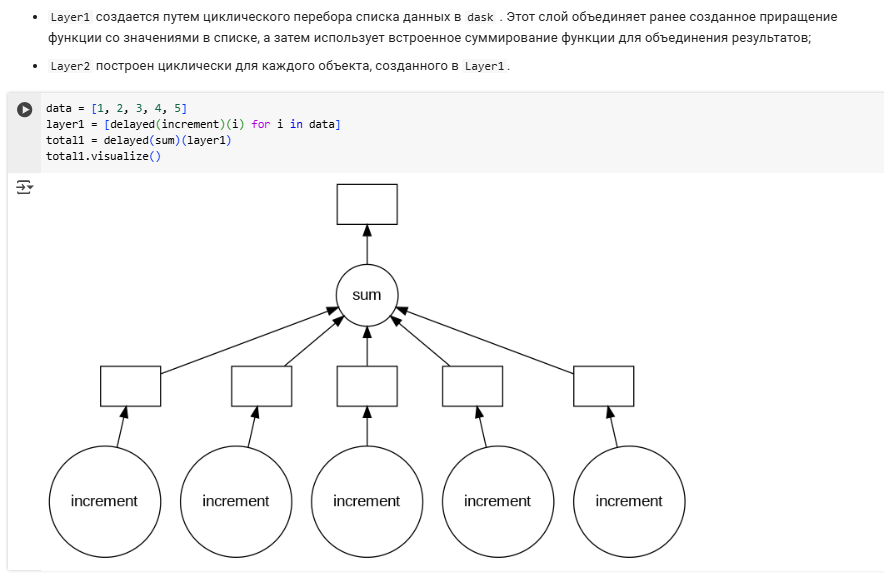


Рисунок 10 – Первый слой Dag с несколькими зависимостями

На рисунке 11 показан второй слой Dag с несколькими зависимостями

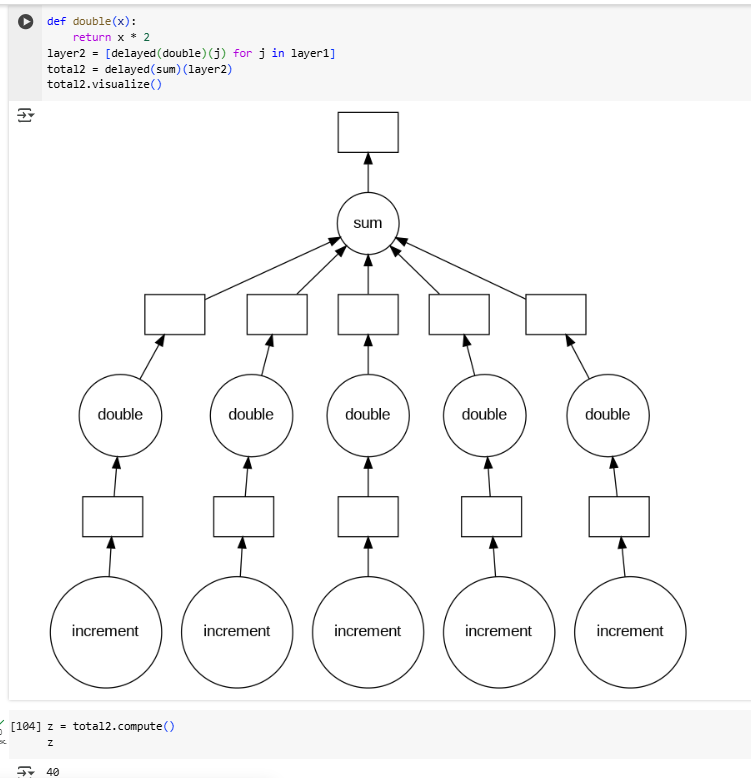


Рисунок 11 – Второй слой Dag с несколькими зависимостями

На рисунке 12 показана визуализация Dag этого датасета.

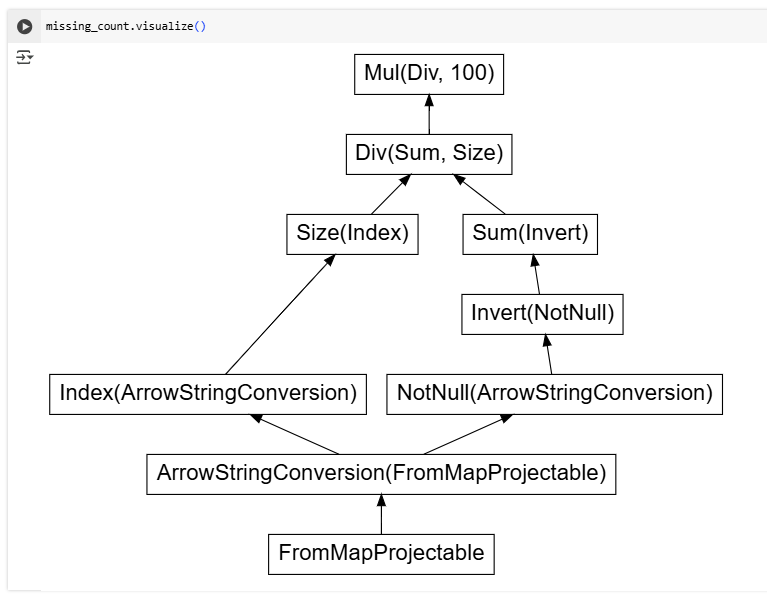


Рисунок 12 – Визуализация Dag этого датасета

**Выводы:**

1. Настроена среда и рабочий каталог.

2. Загружены данные.

3. Проверено качество данных

4. Удалены столбцы (множество пропусков в значениях, бесполезные столбцы для анализа).

5. Визуализирован DAG с одним узлом и зависимостями.

6. Визуализирован DAG с более чем одним узлом и зависимостями.