**ОТЧЕТ по практической работе №3.1**

**Цель и задачи**

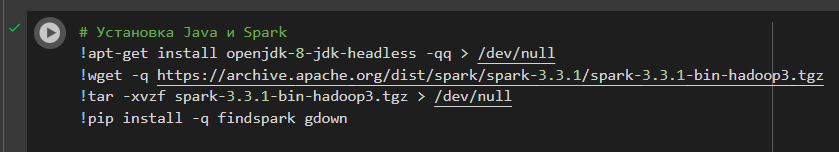
**Цель:** Изучить основы работы с Apache Spark и PySpark в среде Google Colab, применить методы обработки и анализа большого объема данных на примере датасета endomondoHR.json, получить практические навыки выполнения расчетов и визуализации с использованием распределенных вычислений.

**Задачи:**

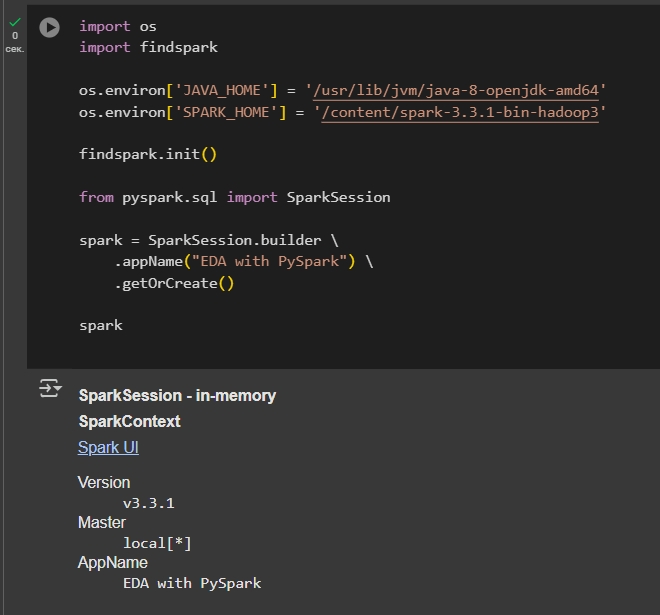
* Установить и настроить Apache Spark в Google Colab;
* Загрузить и обработать датасет endomondoHR.json;
* Изучить структуру данных и определить ключевые признаки;
* Построить визуализации признаков для анализа распределения данных;
* Построить визуализации и вычислить корреляции;
* Выполнить агрегацию по тренировкам (например, средняя высота);
* Сделать выводы на основе анализа.

**1. Подготовка среды и установка зависимостей**

Работа выполнялась в Google Colab. Был установлен Apache Spark версии 3.3.1 и Java 8.

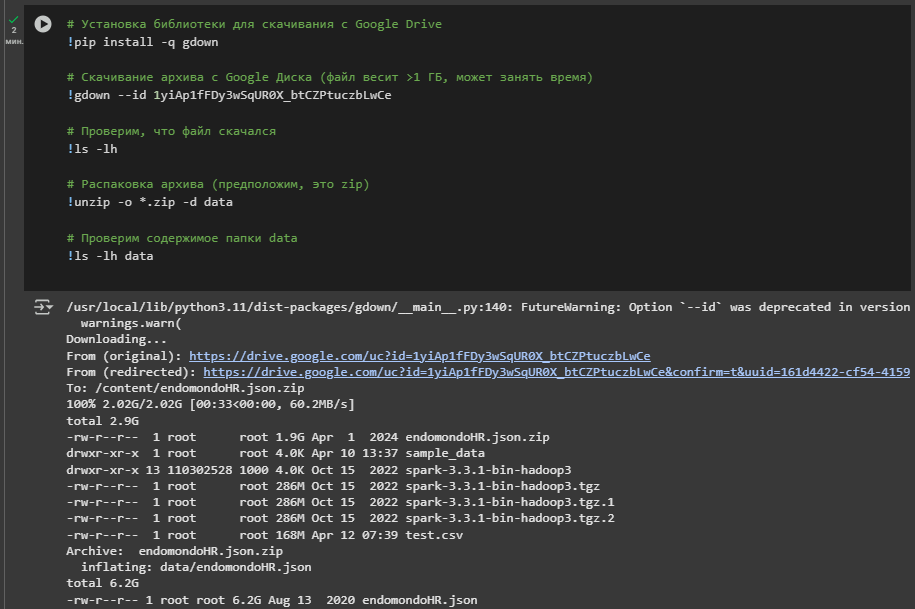


Установка Java и Spark через apt-get и wget

  
Настройка переменных среды и создание SparkSession

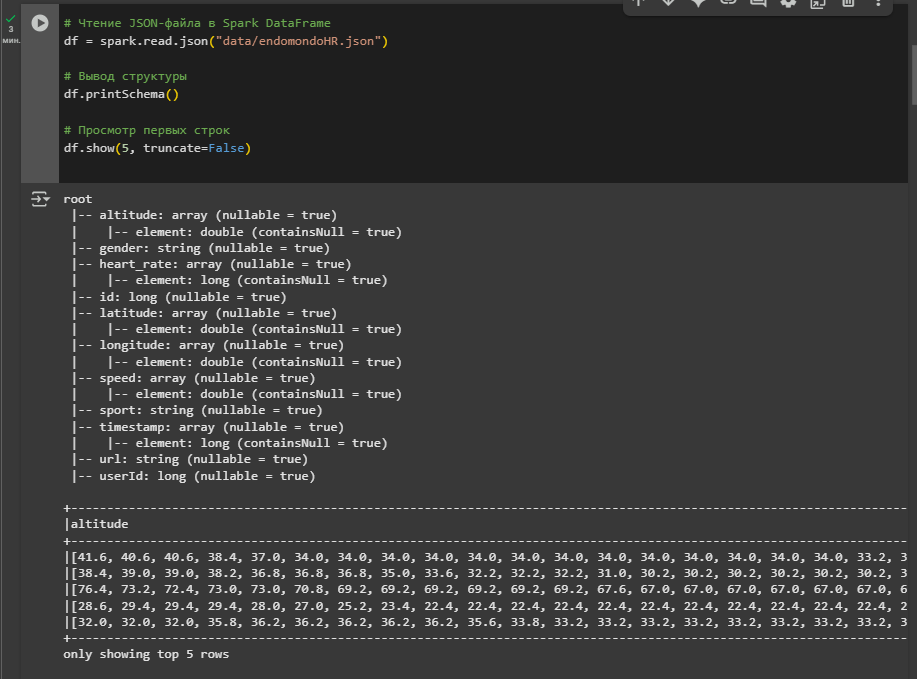
**2. Скачивание и загрузка данных**

Файл endomondoHR.json был скачан с Google Drive и распакован. Затем загружен в Spark DataFrame:

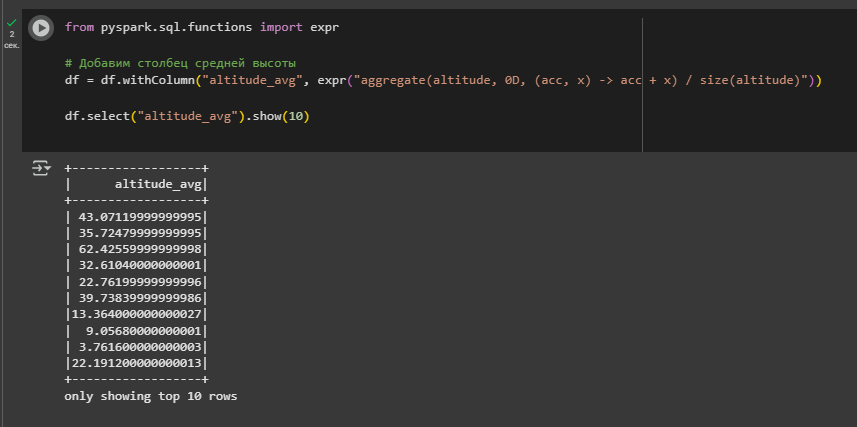
Скачивания файла с Гугл диска

**3. Анализ структуры данных**

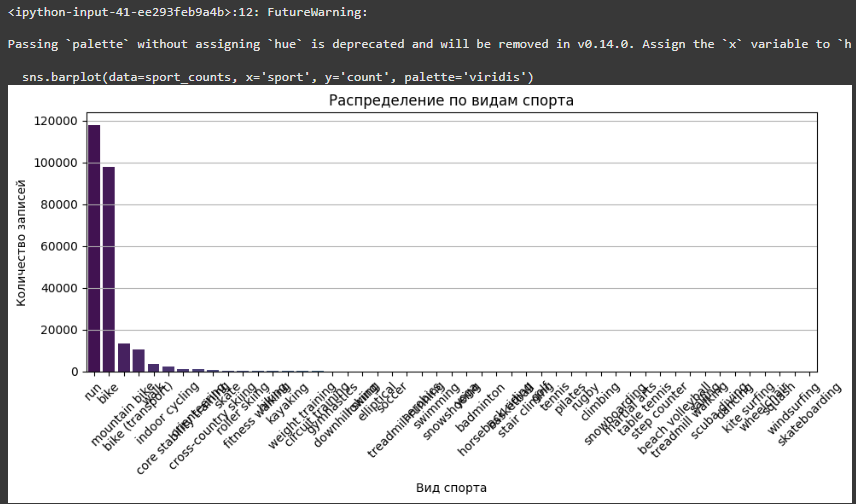
Были определены основные поля: altitude, distance, heart\_rate, longitude, latitude, speed, timestamp.

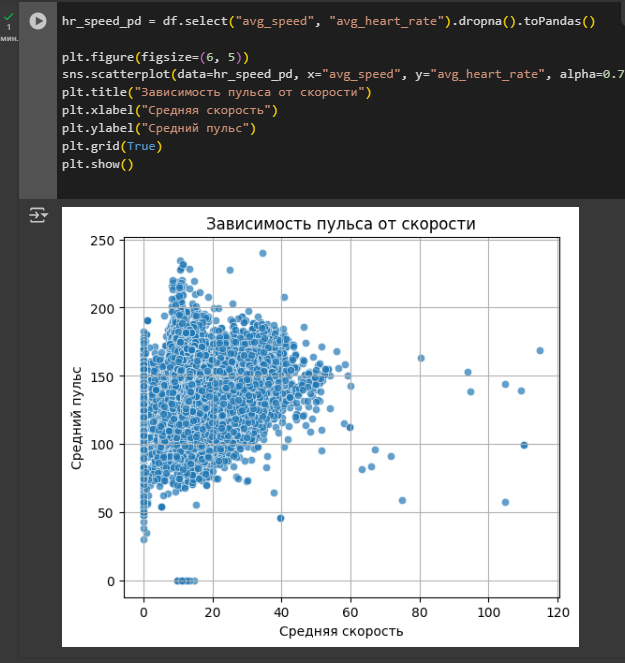
Структура данных (printSchema) и просмотр первых 5 строк

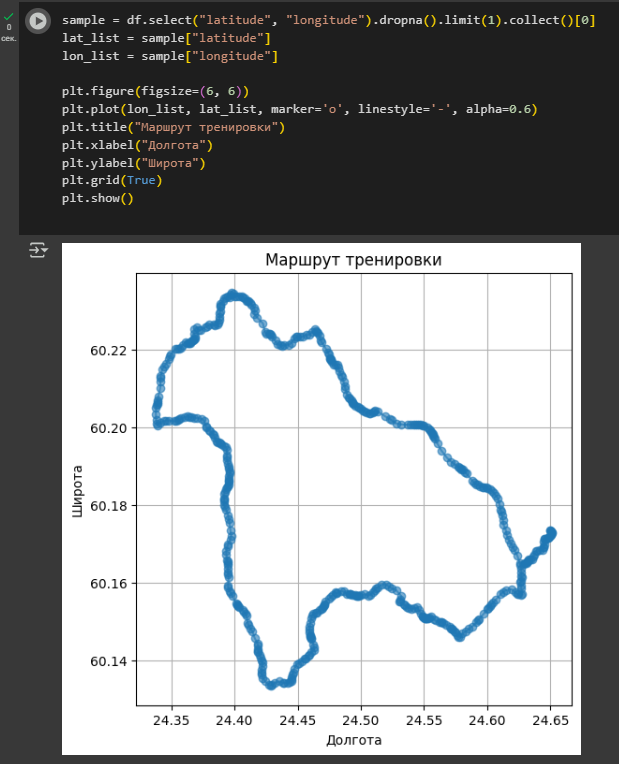
**4. Подсчет средней высоты (altitude) по каждой тренировке**

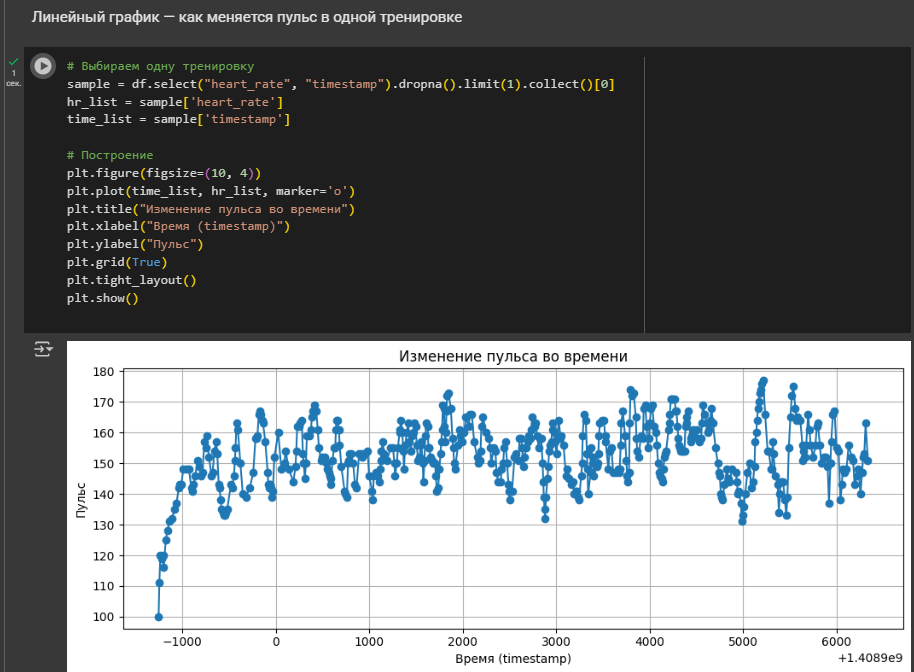


**5. Визуализация данных**

****

****

****

****

**Вывод**

В ходе выполнения практической работы была освоена технология обработки больших данных с использованием Apache Spark и библиотеки PySpark в среде Google Colab. Установлены и настроены необходимые компоненты (Spark, Java), что позволило выполнить анализ большого JSON-файла endomondoHR.json, содержащего записи о физических тренировках.

В рамках разведочного анализа были:

* определены основные признаки датасета и их типы;
* обработаны пропущенные значения;
* рассчитаны агрегаты, включая среднюю высоту по каждой тренировке;
* построены визуализации для анализа распределений данных.

Также были реализованы пользовательские функции для преобразования данных. Работа с PySpark продемонстрировала высокую эффективность при работе с большим объемом информации и позволила получить ценные аналитические сведения о структуре и взаимосвязях признаков.