Задание на лабораторную работу № 4

1. В качестве исходных данных для работы использовать файлы, формируемые генератором исходных данных, разработанным в рамках лабораторной работы № 1.

2. Разработать формат данных для размещения исходных данных в файловой системе DDFS MapReduce-кластера Disco.

3. Разработать алгоритм размещения исходных данных в файловой системе DDFS MapReduce-кластера Disco.

4. Реализовать разработанный в пункте 3 алгоритм в виде сценария на языке программирования Python с использованием API-интерфейса доступа к файловой системе DDFS. Разработанный сценарий должен за один запуск загружать несколько групп входных данных и измерять время загрузки для каждой группы. Интерфейс созданного сценария должен быть реализован следующим образом:

```
python load2DDFS.py --input inputFilesName
--tag-name tagName --metrics metrics.csv
```

где `load2DDFS.py` – имя файла разработанного сценария; `--input` – ключ, определяющий следующий аргумент командной строки как общую часть для имени группы входных файлов, обрабатываемых данных; `inputFilesName` – общая часть для имени группы входных файлов; `--tag-name` – ключ, определяющий следующий аргумент командной строки как общую часть имени ссылок (tag), под которыми будут размещены загружаемые данные; `tagName` – общая часть имён ссылок (tag), под которыми будут размещены загружаемые данные; `--metrics` – ключ, определяющий следующий аргумент командной строки как имя файла, содержащего результаты измерения времени; `metrics.csv` – имя файла, содержащего результаты измерений в формате CSV. Кроме перечисленных допускается наличие других аргументов командной строки, но в таком случае они должны указываться (следовать) после обязательных аргументов, перечисленных выше.

5. Проверить корректность работы разработанных алгоритма и сценария.

6. Провести эксперимент по оценке времени загрузки данных в файловую систему DDFS в зависимости от объёма исходных данных.

7. Полученные результаты эксперимента записать в таблицу следующего вида.

Таблица 3.3 – Форма представлния результатов лабораторной работы № 4

№ п/п	Объём входных данных (ед. изм.)	Среднее время работы однопоточного алгоритма с определённой погрешностью измерений (ед. изм.)

Подготовленные данные о производительности реализованного алгоритма представить в виде диаграммы зависимости объема обрабатываемых данных от времени их загрузки в файловую систему.

8. Обосновать полученные результаты эксперимента, проведя анализ разработанного алгоритма и условий работы разработанного сценария.