Задание на лабораторную работу № 5

1. В качестве исходных данных для работы использовать записи, сформированные в файловой системе DDFS в рамках выполнения лабораторной работы № 4.

2. Разработать MapReduce-алгоритм решения выбранного варианта задания с учетом формата входных данных формируемого сценарием, реализованным в рамках лабораторной работы № 4.

3. Реализовать разработанный в пункте 2 алгоритм в виде сценария на языке программирования Python с использованием API-интерфейса доступа к вычислительным ресурсам MapReduce-кластера. Разработанный сценарий должен за один запуск обрабатывать несколько групп входных данных и измерять время обработки для каждой группы. Интерфейс созданного сценария должен быть реализован следующим образом:

```
python computeInDisco.py --input tagName --metrics metrics.csv
```

где `computeInDisco.py` – имя файла разработанного сценария; `--input` – ключ, определяющий следующий аргумент командной строки как общую часть имени ссылок (tag), под которыми в файловой системе DDFS размещены исходные данные; `tagName` – общая часть для имени группы входных файлов; `--tag-name` – ключ, определяющий следующий аргумент командной строки как имя ссылки (tag), под которой будут размещены загружаемые данные; `tagName` – общая часть имён ссылок (tag), под которыми в файловой системе DDFS размещены исходные 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 обработки исходных данных в зависимости от их объёма.

7. Полученные результаты эксперимента записать в таблицу (см. таблицу 3.3). 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.