

# Управление разно-структурированными большими данными

к.т.н. Брюхов Д.О. ([dbriukhov@ipiran.ru](mailto:dbriukhov@ipiran.ru))

## Hadoop Map Reduce (Домашнее задание 1)

### Общие требования

- Задание выполняется в виде MapReduce приложения на Java или Python (допускается использование и других языков программирования).
- Срок выполнения задания – **22 ноября**.
- Выполненное задание должно быть прислано на почту в виде архива (например, содержащего проект Eclipse).
- Имя должно быть названо: <Family\_Name>\_MapReduce\_Var<#>.rar (zip, gz, и.т.д.)
  - <Family\_Name> - Ваша фамилия
  - <#> - номер варианта
  - Например: Briukhov\_MapReduce\_Var4.rar

### Платформа Hadoop

Выполнение задания осуществляется на любой платформе Hadoop (например, IBM BigInsights, HortonWorks).

Если домашний компьютер позволяет запускать виртуальные машины дома, можно скачать виртуальную машину и работать с ней.

Если такой возможности нет, то можно использовать удаленный сервер

### Использование виртуальной машины

Скачать виртуальную машину

Например, <https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html>

Установить Python

```
yum install python-pip
```

Установить MRJob

```
pip install mrjob
```

### Соединение с сервером

Установить SSH клиент, например, Putty (<https://www.putty.org/>)

Установить WinSCP (<https://winscp.net/eng/download.php>)

Адрес: 83.149.227.84

Порт: 38222

Логин/пароль: student/msucmc19

### Варианты заданий

1. Реализовать алгоритм TF\*IDF

Входными данными являются: директория, содержащая несколько текстовых файлов, и поисковая строка, состоящая из нескольких слов.

Формат представления входных и выходных данных выбирается Вами.

2. Реализовать алгоритм поиска кратчайшего пути в графе

Формат представления входных и выходных данных выбирается Вами. Ребра графа должны содержать веса.

3. Реализовать метод кластеризации K-means

Формат представления входных и выходных данных выбирается Вами. Не использовать Евклидову метрику в качестве метрики расстояния между точками.