

Извлечение и интеграция информации из больших данных

к.т.н. Брюхов Д.О. (dbriukhov@ipiran.ru)

Слияние данных (Домашнее задание 2)

Общие требования

- Срок выполнения задания – **27 апреля**.
- Задание выполняется в виде программы на языках Pig (допускается использование и других языков программирования).
- Входными данными являются JSON файлы, являющиеся результатом домашнего задания 1.
- Результатом являются данные в формате JSON.
- Выполненное задание должно быть прислано на почту в виде архива, содержащего текст программы, исходные данные и результат.
- Имя должно быть названо: <Family_Name>_HW2.rar (zip, gz, и.т.д.)
 - <Family_Name> - Ваша фамилия
 - Например: Briukhov_HW2.rar
- Выполнение практикума желательно осуществлять на платформе IBM BigInsights.
- Если домашний компьютер позволяет запускать виртуальные машины дома, можно скачать виртуальную машину и работать с ней.
- Если такой возможности нет, то можно использовать наш сервер Hadoop (подробности доступа к серверу в файле Лабораторная на Python/MRJob)

Задание

- Написать функцию реализующую одну из четырех функций слияния данных: Minimum Union, Complement Union, Merge или Fuse By;
- Применить эту функцию к данным, полученным в результате выполнения домашнего задания 1;
- В результате должно получиться множество объектов, включающих данные (атрибуты) из обоих входных файлов;
- Результат представить в формате JSON