Извлечение и интеграция информации из больших данных

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

Загрузка разнородных ресурсов данных (Домашнее задание 1)

Общие требования

- Срок выполнения задания – **23 марта**.
- Задание выполняется в виде программы на языках Jaql, AQL, Python (допускается использование и других языков программирования).
- 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>_HW1.rar (zip, gz, и.т.д.)
    - <Family_Name> - Ваша фамилия
    - Например: Briukhov_HW1.rar
- Выполнение практикума осуществляется на любой платформе Hadoop.
- 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

- Найти 2 источника данных, связанных по какому-либо атрибуту, в любом формате (html, txt, csv, ...).
- Загрузить эти данные в HDFS
- Написать программу преобразования этих данных в формат JSON
- В результате должно получиться 2 файла в формате JSON, содержащие данные из 2 разных источников

Примеры источников данных

- Один источник: данные о публикациях с нашего сайта (например, http://synthesis.ipi.ac.ru/synthesis/publications/17damdid-ilb.html)
  Второй источник: данные о сотрудниках с нашего сайта (например, http://synthesis.ipi.ac.ru/synthesis/staff/brd.html)
- Один источник: данные о продуктах с любого интернет магазина (например, https://www.oldi.ru/catalog/element/0463274/#harakteristiki)
  Второй источник: данные о продуктах с другого интернет магазина (например, https://www.dns-shop.ru/product/4e489f4faf9d3330/processor-intel-celeron-g3930-oem/characteristics/)
- Один источник: данные о продуктах с любого интернет магазина (например, https://www.dns-shop.ru/product/4e489f4faf9d3330/processor-intel-celeron-g3930-oem/characteristics/)
  Второй источник: другие данные о продуктах с того же интернет магазина (например, https://www.dns-shop.ru/product/4e489f4faf9d3330/processor-intel-celeron-g3930-oem/opinion/)
- Любые ресурсы из каталогов ресурсов: Data.gov.uk, Data.gov.us, data.mos.ru, hubofdata.ru, ...