Управление разно-структурированными большими данными

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

Извлечение информации из текстов (Домашнее задание 2)

Общие требования

- Задание выполняется на языке AQL
- Альтернатива: Задание выполняется в виде MapReduce приложения на Java или Python (допускается использование и других языков программирования).
- Входной файл: Facts.txt (http://synthesis.ipi.ac.ru/synthesis/student/BigData/seminar-hadoop/lectures2019/Facts.txt)
- Срок выполнения задания – **14 декабря**.
- Выполненное задание должно быть прислано на почту в виде архива (например, содержащего проект Eclipse).
- Имя должно быть названо: <Family_Name>_TextExtraction_Var<#>.rar (zip, gz, и.т.д.)
    - <Family_Name> - Ваша фамилия
    - <#> - номер варианта
    - Например: Briukhov_TextExtraction_Var4.rar

Платформа

Выполнение задания осуществляется в среде разработки Eclipse

- Eclipse версии 4.2
- JDK 7
- Инструменты разработки IBM BigInsights (https://yadi.sk/d/NZanVkEbpIxLQQ)
- Установка плагина описана в файле http://synthesis.ipi.ac.ru/synthesis/student/BigData/seminar-hadoop/lectures2019/06_MapReduce_Lab.pdf
-

Варианты заданий

- Написать представление для извлечения любой информации из файла Facts.txt (или подобного файла)
- Либо выполнить два из представленных ниже вариантов

1) Написать представление для извлечения информации (список стран и протяженность границы) о граничащих странах

Представление должно содержать 3 поля: country, bordered_country, distance.

Пример фрагмента текста с извлекаемой информацией:

Geography Afghanistan

…

Land boundaries: total: 5,529 km border countries: China 76 km, Iran 936 km, Pakistan 2,430 km, Tajikistan 1,206 km, Turkmenistan 744 km, Uzbekistan 137 km 2)

2) Написать представление для извлечения информации о продолжительности железных дорог по типу колеи

Представление должно содержать 2 поля: gauge_type, distance.

Пример фрагмента текста с извлекаемой информацией:

Railways: total: 24.6 km broad gauge: 9.6 km 1.524-m gauge from Railways: total: 965 km narrow gauge: 965 km 1.067-m gauge (2000 est.)

3) Написать представление для извлечения списка международных организаций, упорядоченных по числу стран, принимающих участие в этой организации

Представление должно содержать 2 поля: organization, number_of_countries. Пример фрагмента текста с извлекаемой информацией:

International organization participation: AsDB, CP, ECO, ESCAP, FAO, G-77, IAEA, IBRD, ICAO, ICRM, IDA, IDB, IFAD, IFC, IFRCS, ILO, IMF, IOC (suspended), IOM (observer), ITU, NAM, OIC, OPCW (signatory), UN, UNCTAD, UNESCO, UNIDO, UPU, WFTU, WHO, WMO, WToO

4) Написать представление для извлечения количества вхождений используемых природных ресурсов

Представление должно содержать 2 поля: natural_resource, number. Пример фрагмента текста с извлекаемой информацией:

Natural resources: coal, petroleum, natural gas, tin, limestone, iron ore, salt, clay, chalk, gypsum, lead, silica, arable land

5) Написать представление для извлечения количества импортируемых товаров для каждой страны

Представление должно содержать 2 поля: country, number_of_import_goods. Пример фрагмента текста с извлекаемой информацией:

Imports - commodities: capital goods, food and petroleum products;

…

Imports - commodities: machinery and transport equipment, construction materials, chemicals, food and live animals