Обогащение и агрегация

Цель: обогатить данные из источника расшифровкой кодов и сгенерировать агрегированное представление

Описание задания: Для выполнения задачи вам потребуется датасет классификатора ТН ВЭД, доступный на сайте налоговой службы: https://www.nalog.gov.ru/rn77/program/5961290/ (архив в ZIP-формате)

На данный момент у вас есть датасет в виде CSV-файла, полученный в предыдущем задании, где в колонке code указан идентификатор класса товаров из номенклатуры, которую вы скачали выше. В данном задании необходимо реализовать MapReduce-пайплайн (возможно, состоящий из нескольких MapReduce job'ов), выходным результатом которого должен быть один или несколько CSV-файлов (разделитель - символ табуляции) с колонками "code", "count" и "category", содержащих:

- В колонке "code" должны быть уникальные коды категорий из исходного CSV-файла
- В колонке "count" должно быть количество раз, сколько продукты данной категории ввозили за рассматриваемый период (игнорируйте количество единиц товара, считайте просто общее количество записей)
- В колонке "category" должно быть наименование категории по номенклатуре из справочника из файла TNVED3.txt (определяется только первыми двумя парами цифр, например, для кода, начинающегося на "3926", в колонке будет запись "ИЗДЕЛИЯ ПРОЧИЕ ИЗ ПЛАСТМАСС И ИЗДЕЛИЯ ИЗ ПРОЧИХ МАТЕРИАЛОВ ТОВАРНЫХ ПОЗИЦИЙ 3901 - 3914"). Брать нужно только актуальное название категории (последнее в списке, без даты окончания действия). В случае, если в исходных данных вам встретится категория, не упомянутая в номенклатуре, укажите здесь "ПРОЧЕЕ"

Выходные файлы должны быть отсортированы по колонке count от большего к меньшему и должны быть в кодировке UTF-8.

Файлы с кодом маппера/редьюсера для каждой стадии MR должны лежать в отдельном каталоге stage1, stage2, stage3 и т.д. README проекта должен содержать команды запуска Hadoop Streaming-пайплайнов каждой стадии по-очереди, использование дополнительных трюков с partitioner/combiner приветствуется.

Ответом на задание является ссылка на git-репозиторий с файлами кода для каждой стадии и простым описанием, как их запускать. Самих исходных файлов с датасетами или справочниками в репозитории быть не должно.

Инструменты, которые пригодятся для выполнения: Python, HDFS, Hadoop Streaming

Критерии, по которым будет оцениваться задание: После выполнения команд из README проекта (с потенциальными правками путей) в HDFS в выходном каталоге должен появляться один или несколько файлов CSV с результатом обработки.

Эти файлы должны содержать все категории товаров из исходных данных и корректное их количество, а также правильное наименование для каждой категории на основе первых двух цифр. Данные в файлах должны быть отсортированы по колонке count от большего к меньшему. Они должны быть в кодировке UTF-8 и использовать в качестве разделителя символ табуляции.