Агрегация для витрины

Цель: построить витрину торговых партнеров как по импорту, так и по экспорту по каждому из представленных в данных временных диапазонов.

Описание задания: В исходных данных присутствуют колонки direction, month и country, содержащие, соответственно, направление (импорт/экспорт), месяц/год записи и аббревиатуру страны, с которой идет товарообмен. Также у вас есть датасет с названиями категорий, полученный на предыдущем этапе проекта.

Необходимо реализовать программу на Python с использованием PySpark для составления витрины (набора Parquet-файлов в HDFS), где в результате даты преобразуются в структуру каталогов с годом и месяцем вот в таком виде:

```
/partners/2016/01/<parquet_files>
/partners/2016/02/<parquet_files>
…
/partners/2021/12/<parquet_files>
```

В каждом таком каталоге должны лежать parquet-файлы, содержащие колонки “month”, “country”, “direction”, “code”, “category”, "measure", "value", "netto”, "quantity", "region", "district" с соответствующими данными из исходных датасетов.

Ответом на задание является ссылка на git-репозиторий с кодом на Python+Pyspark и описанием команды запуска в README. Самих исходных файлов с датасетами или справочниками в репозитории быть не должно.

Инструменты, которые пригодятся для выполнения: Python, HDFS, Spark

Критерии, по которым будет оцениваться задание: выходные файлы должны быть в формате Parquet и содержать перечисленные в задании колонки, а также они должны быть расположены в иерархии каталогов с указанием года и месяца

Ваше решение

Отправить на проверку

Сохранить как черновик

Далее