Курсовой проект для студентов учебного центра Neoflex по направлению «Data-Engineering (BDS)»

Задание №2

В предыдущем задании вы занимались восстановлением потерянных данных и сломавшегося функционала расчёта витрины 101-й формы в реляционной БД (Oracle / PostgreSQL).

Во время общения с архитектором вы узнаёте, что банк решил мигрировать свой DWH в Hadoop, а ETL-процессы перевести на Spark. В виду этого необходимо настроить среду для исследования возможностей новых инструментов и того как в них можно переложить существующий ETL-функционал.

**Задача 2.1**

В помощь, архитектор поделился с вами скриптами по настройке среды разработки со Spark. По мимо этого он отправил вам файл «Athletes.csv» и попросил выполнить несколько запросов для проверки работоспособности Spark-приложений.

Настройте виртуальную машину с Ubuntu, установите на неё Spark (PySpark по желанию). Так же вы можете установить среду разработки - например Jupyter. Команды по установке и первичной настройке находятся в файле "ubuntu\_commands.txt", рядом с ним ещё должен быть прикреплён файл «PySpark\_Simple\_example.txt».

Проверочные задачи:

1. Сгенерировать DataFrame из трёх колонок (row\_id, discipline, season) - олимпийские дисциплины по сезонам.

* row\_id - число порядкового номера строки;
* discipline - наименование олимпиский дисциплины на английском (полностью маленькими буквами);
* season - сезон дисциплины (summer / winter);

\*Укажите не мнее чем по 5 дисциплин для каждого сезона.

Сохраните DataFrame в csv-файл, разделитель колонок табуляция, первая строка должна содержать название колонок.

Данные должны быть сохранены в виде 1 csv-файла а не множества маленьких.

---------------------------------------------------------------------------------

2. Прочитайте исходный файл "Athletes.csv".

Посчитайте в разрезе дисциплин сколько всего спортсменов в каждой из дисциплин принимало участие.

Результат сохраните в формате parquet.

---------------------------------------------------------------------------------

3. Прочитайте исходный файл "Athletes.csv".

Посчитайте в разрезе дисциплин сколько всего спортсменов в каждой из дисциплин принимало участие.

Получившийся результат нужно объединить с сгенерированным вами DataFrame из 1-го задания и в итоге вывести количество участников, только по тем дисциплинам, что есть в вашем сгенерированном DataFrame.

Результат сохраните в формате parquet.

Примечания:

* Виртуальную среду можете развернуть в VirtualBox – самый популярный инструмент для подобных учебных целей;
* Если вы уже работаете в Linux- среде, то виртуальную среду с Ubuntu разворачивать не обязательно – можете настраивать Spark в своей среде.

Требования к демонстрации работы:

* Записать видео с экрана компьютера, в котором вы демонстрируйте и комментируете в слух, то что вы делаете / уже разработали;
* Обязательно продемонстрируйте процесс вызова функции и сохранения её результата;
* Это видео загрузите к себе на облако (гугл-диск, яндекс-диск и т.д.) и предоставьте доступ по ссылке;
* Приложите в домашнее задание внутри архива текстовый файл с ссылкой на ваше видео и все скрипты решения.