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 олимпийский дисциплины на английском (полностью маленькими буквами);
- season - сезон дисциплины (summer / winter);

*Укажите не менее 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.

2.1.1.1. 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.

2.1.1.2. Требования к демонстрации работы:

- Все скрипты и решения необходимо опубликовать в github и предоставить ссылку на репозиторий.
- 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.п.) и предоставьте доступ по ссылке;
- Приложите в репозиторий github текстовый файл с ссылкой на ваше видео