**Задание**

Двигаемся дальше, и теперь на очереди — Hive!

* Первым делом нужно достать данные — заходим и грузим все, что есть из репозитория: [GitHub - sultanmurad/csv\_files](https://github.com/sultanmurad/csv_files).
* Эти файлы будут подвержены анализу, но в них кое-чего недостает — добавьте в каждый файл столбец с номером группы таким образом, чтобы файл был разделен на 10 групп. В файл customers.csv добавьте столбец с номером года, в который была совершена подписка (**Subscription Date**). Используйте средства python + pandas.
* Загрузите полученные файлы на hdfs.
* Теперь ваша задача следующая: аналитики хотят сводную статистику на уровне каждой компании и на уровне каждого года получить целевую возрастную группу подписчиков — то есть, возрастную группу, представители которой чаще всего совершали подписку именно в текущий год на текущую компанию.

Например:

|  |  |  |
| --- | --- | --- |
| **Company** | **Year** | **Age\_group** |
| Apple | 2023 | 0-18 |
| Xiaomi | 2023 | 25-35 |

Все операции необходимо выполнить в Hive. Работать с Hive можно через интерфейс HUE, если перейти в раздел «Query», или выбрав нужный пункт в разделе «Editor».

Таким образом вам нужно создать под каждый csv-файл отдельную таблицу. Для оптимизации используйте свои знания партиционирования и бакетирования. А затем на основе 3-х таблиц собрать витрину, которая решает поставленную задачу. В качестве результата предоставьте код SQL-запросов для создания исходных таблиц и создания итоговой витрины.

Результат выполнения задания необходимо выложить в github/gitlab и указать ссылку на Ваш репозиторий (не забудьте: репозиторий должен быть публичным).