Краткое описание набора данных telecomX.

Д.В.Яценко,Учебный центр Аренадата, Россия ydv@arenadta.io

Датасет telecomX представляет синтетический набор данных, созданный на основе реального кейса, встречающегося в работе телеком-компаний, связанный с мониторингом потребления интернет-трафика абонентами. Общая схема приведена на рис 1.

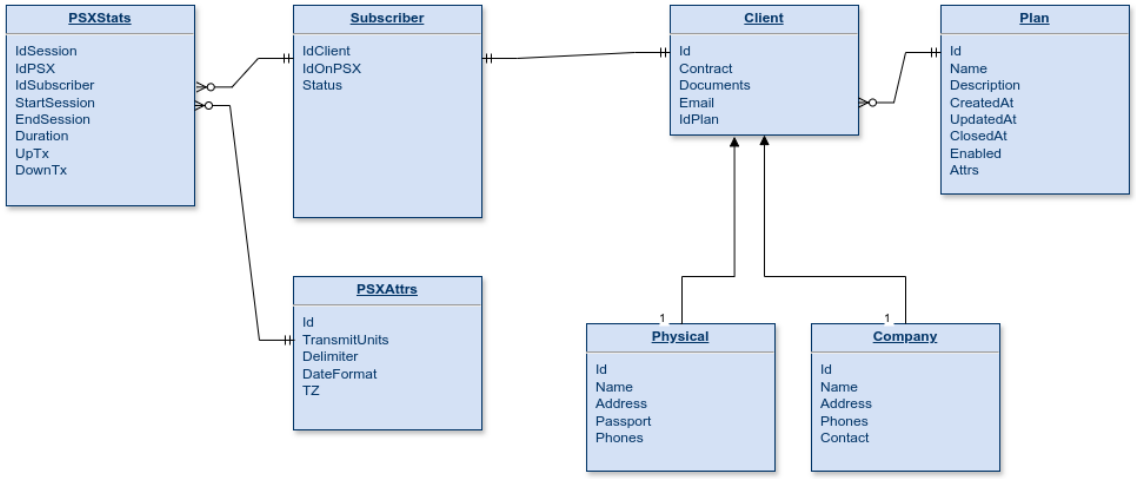

Рис. 1: Логическая модель данных

Данные представлены в следующих форматах файлов и содержат следующие типы данных:

Client (Клиент):
Формат: сжатый файл Parquet
Атрибуты: идентификатор клиента, номер контракта, хранилище копий документов клиентов, электронный адрес клиента, идентификатор тарифного плана.

Physical (Физическое лицо):
Формат: сжатый файл Parquet
Атрибуты: идентификатор клиента, ФИО клиента, адрес проживания, паспортные данные, телефоны.

Company (Компания):
Формат: сжатый файл Parquet
Атрибуты: идентификатор клиента, название компании, юридический адрес, телефоны компании, контактные данные сотрудников компании.

Plan (Тарифный план):
Формат: JSON
Атрибуты: идентификатор тарифного плана, название тарифного плана, краткое описание, даты открытия/изменения/закрытия, активность, внутренние атрибуты плана.

Subscriber (Абонент):
Формат: CSV
Атрибуты: идентификатор клиента в сущности Client, номер абонента в линейном оборудовании, статус абонента (разрешен/запрещен).

PSXStats (Статистика по подключениям):
Формат: CSV или TXT
Атрибуты: номер сессии подключения, идентификатор коммутатора, номер абонента, начало и конец сессии, длительность сеанса, объем переданного и скачанного трафика.

PSXAttrs (Атрибуты коммутатора):
Формат: CSV
Атрибуты: идентификатор коммутатора, название, единицы измерения трафика, символ разделителя, формат даты, таймзона.

Задание. Требуется провести анализа представленных данных о потреблении интернет-трафика абонентами, путем исследования данных о объеме переданного и полученного трафика из файлов выгрузок с коммутаторов. В процессе исследования нужно сравнивать ретроспетивное потребление трафика абонентом с текущим, и при обнаружении нетипичного потребления сделать вывод о взломе. Необходимо строить по одной таблице-отчету (витрине данных) на каждый час данных с коммутаторов. Т.е. количество витрин должно быть равно периоду (в часах) за который выгружались оперативные данные, т.е. 24*7 = 168.
В витрине должны быть представлены следуюшие данные:

- время,
- название клиента,
- номер договора клиента,
- контактные данные для связи с клиентом,
- предполагаемый статус взлома (hacked/clear),
- обоснование предполагаемый статус взлома (краткая историрия потребления).