Описание датасета telecomX.

Д.В.Яценко *

5 мая 2024 г.
v1.0

Содержание

1 Введение **2**

2 Описание данных **2**

3 Постановка задачи **6**

*Southern Federal University: ydv@arenadata.io

1 Введение

Представляемый синтетический набор данных создан на основании реального кейса, встречающегося в работе телеком-компаний. Этот кейс встречается как часть реального производственного процесса, решаемого аналитиками данных компании – оператора связи.

Общее описание кейса. Компания, предоставляющая услуги связи, имеет различные тарифы на предоставляемые каналы связи интернет. Однако, несмотря на то, что каналы ограничены рамками тарифов, компания отслеживает динамику потребления интернет-трафика потребителями, и в случае фиксации нетипичного всплеска потребления, предполагает, что возможно компьютер абонента был взломан, и при помощи вредоносного ПО стал выполнять роль сервера спам-рассылки, элемента DDOS сети, сайта с запрещенным контентом и т. д. В этом случае, предоставление услуг интернет-связи временно блокируется, а с абонентом связываются для выяснения обстоятельств.

2 Описание данных

При анализе, дата аналитики получают выгрузку из различных систем:

- MDM (см. раздел Аббревиатуры) - мастер данные о бизнес-сущностях, которые обеспечивают контекст для бизнес-транзакций,
- OLTP - операционные данные по текущим бизнес-транзакциям,
- OBS - оперативные данные из систем онлайн-биллинга (ПАК обеспечения систем связи) о подключениях абонентов, выгружаемые через небольшие интервалы времени. Обычно такие данные собираются по протоколам управления с сетвых коммутаторов, например NetFlow, аггрегируются в микробатчи и выгружаются как пребиллинговый уровень данных.

Все эти данные выгружаются в виде файлов различного формата на хранилище аналитики (например NFS, HDFS). В некоторых случаях, оперативные данные биллинга могут поставляться не в виде файлов, а через потоковые системы, например Kafka. На (рис 1) приведена общая логическая модель данных, необходимаых для анализа. Рассмотрим данные более детально.

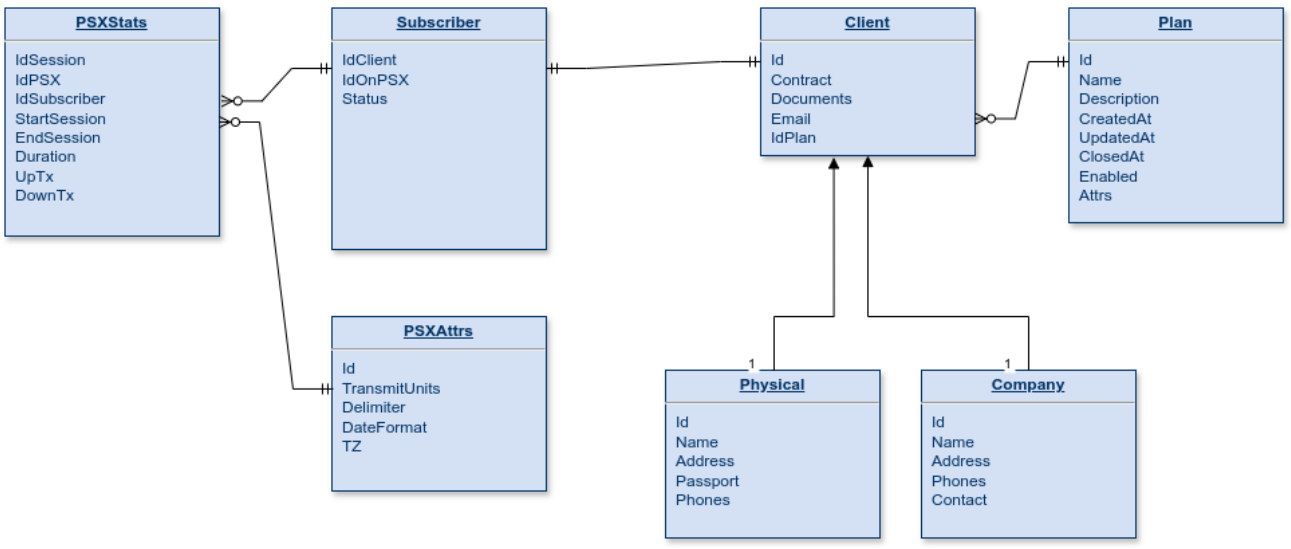

Рис. 1: Логическая модель данных

Информация о бизнес-сущностях содержится в таблицах Client, Physical, Company и Plan.

Client - это базовая сущность для Physical и Company. Содержит общую часть информации из наследованных сущностей Physical и Company. Данные поставляются как сжатый файл Parquet со встроеной схемой данных. Пример данных приведен в (таб 1) Описание аттрибутов сущности:

- Id - идентификатор клиента,

- Contract - номер контракта,
- Documents - хранилище копий документов клиентов,
- Email - электорнный адрес клиента,
- IdPlan - идентификатор тарифного плана (см сущность Plan).

Таблица 1: Пример данных Client.

Id	**Contract**	**Documents**	**Email**	**IdPlan**
7971b0c1-6000-41ea-ac8c-f443e61 f20c5	GB29J LOH44	internal.store.com/ clients/ documents/GB29J LOH44	wheelerjames@ williams.com	4

Physical - это сущность описывает клиента компании физическое лицо. Эта сущность наследуется от Clients. Файл с данными содержит соединенные данные базовой и наследованной сущности. Данные поставляются как сжатый файл Parquet со встроеной схемой данных. Пример данных приведен в (таб 2).

- Id - идентификатор клиента,
- Name - ФИО клиента,
- Address - адрес проживания,
- Passport - паспортные данные,
- Phones - телефоны.

Таблица 2: Пример данных Physical.

Id	**Contract**	**Documents**	**Email**	**Id Plan**	**Name**	**Address**	**Passport**	**Phones**
7971b0c1-6000-41ea-ac8c-f443e61 f20c5	GB29J LOH44	internal.store.com/ clients/ documents /GB29J LOH44	wheeler jame@ williams.com	4	Billy Briggs PhD	3047 Marissa Lights Apt. 603 South Ryanside	Billy Briggs PhD M 1989-05-09, 2022-01-29, 2026-05-16 G859 49583	b'["+1 747 922 9784"+1 752-237-5341"]'

Company - это сущность описывает клиента компании юридическое лицо. Эта сущность наследуется от Clients. Файл с данными содержит соединенные данные базовой и наследованной сущности. Данные поставляются как сжатый файл Parquet со встроеной схемой данных. Пример данных приведен в (таб 3).

- Id - идентификатор клиента,
- Name - название компании,
- Address - юридический адрес,
- Phones - телефоны компании,
- Contact - контактные данные сотрудников компании.

Таблица 3: Пример данных Physical.

Id	**Contract**	**Documents**	**Email**	**Id Plan**	**Name**	**Address**	**Phones**	**Contatct**
3f5f7714-b633-4929-bc5b-0e98a3790a13	GB42SHNI52	internal.store.com/clients/documents/GB42SHNI52	michaelsullivan@ball.com	11	Brock-Nunez and Sons	4755 Foster Locks Apt. 715 Scottshire, TN 37405	b'["+1 517-644-9726"+1 468 973 6401"]'	b'["Stephanie Duncan"Thomas Robinson"]'

Plan - эта сущность описывает тарифный план. Файл с данными предоставляетя в JSON формате. Пример данных приведен в листинге (лист 1).

- Id - int,идентификатор тарифного плана,
- Name - text, название тарифного плана,
- Description - text, краткое описание тарифного плана,
- CreatedAt - long, дата открытия (эпохи в милисекундах),
- UpdatedAt - long, дата изменения (эпохи в милисекундах),
- ClosedAt - long, дата закрытия (эпохи в милисекундах),
- Enabled - bool, активный/отключенный,
- Attrs - text, внутренние аттрибуты плана.

Листинг 1: Пример данных Plan

```
[{
    "Id":0,
    "Name":"Start",
    "Description":"Minimal plan for beginer internet users",
    "CreatedAt":1606089600000,
    "UpdatedAt":null,
    "ClosedAt":null,
    "Enabled":true,
    "Attrs":"I,1000,24"
}]
```

Subscriber - эта сущность описывает текущий статус абонента, номер абонента в оборудовании связи и его идентификатор в справочнике первичных данных. Файл с данными предоставляетя в CSV формате. Пример данных приведен в листинге (лист 2).

- IdClient - uuid,идентификатор клиента в сущности Client,
- IdOnPSX - int, номер абонента в линейном оборудовании (PSX),
- Status - text, абонент разрешен/запрещен (ON/OFF) ,

Листинг 2: Пример данных Subscriber

```
IdClient,IdOnPSX,Status
7971b0c1-6000-41ea-ac8c-f443e61f20c5,3792,ON
```

Следующие сущности уже не справочные и содержат оперативную информацию.

PSXStats - эта сущность описывает текущее состояние подключений абонента. Данные выгружаются из линейного оборудования каждые 10 мин. При этом каждая выгрузка, с каждого коммутатора создает новый, дополнительный к уже имеющимся ранее, файл. Файлы с данными предоставляютя в CSV или в TXT формате, в зависимотси от типа оборудования-источника. Отличия в форме выгружаемых данных касаются символа разделителя, формата даты, единиц измерения объема трафика (см сущнсть PSXAttrs). Пример данных приведен в листинге (лист 3, 4).

- IdSession - long, номер сессия подключения абонента, этот номер не меянется, до тех пор, пока клиент не отключится и не подключится заново, длительность сессии обычно составляет от часа до суток.
- IdPSX - int, идентификатор коммутатора (см сущность PSXAttrs),
- IdSubscriber - int, номер абонента в линейном оборудовании (см сущность Subscriber),
- StartSession - text, дата начала сессии связи абонента, формат даты см. PSXAttrs,
- EndSession - text, дата окончания сессии связи абонента, формат даты см. PSXAttrs, если сессия не завершена в текущую десятиминутку, то поле пустое,
- Duartion - int, длительность сеанса связи абонентав секундах,
- UpTx - long, объем переданного от абонента трафика, единицы измерения см. PSXAttrs,
- DownTx - long, объем скачанного абонентом трафика, единицы измерения см. PSXAttrs.

Листинг 3: Пример данных PSXStats (CSV)

```
IdSession,IdPSX,IdSubscriber,StartSession,EndSession,Duartion,UpTx,DownTx
9934,3,98502,31-12-2023 15:50:39,,53961,2498801833,2619386264
```

Листинг 4: Пример данных PSXStats (TXT)

```
IdSession|IdPSX|IdSubscriber|StartSession|EndSession|Duartion|UpTx|DownTx
5938|1|63981|01/01/2024 04:32:56||7624|766867096|809999496
```

PSXAttrs - эта сущность описывает параметры (профили) коммутаторов связи. Файл с данными предоставляютя в CSV

- Id - int, идентификатор коммутатора,
- PSX - text, название коммутатора,
- TransmitUnits - next, единица измерения трафика биты/байты,
- Delimiter - символ разделителя в файлах выгрузки,
- DateFormat - формат представления данных файлах выгрузки,
- TZ - таймзона даты коммутатора.

Листинг 5: Пример данных PSXAttrs

```
Id,PSX,TransmitUnits,Delimiter,DateFormat,TZ
0,66.1,bits,|,%d/%m/%Y %H:%M:%S,GMT-5
```

3 Постановка задачи

Итак, данные из описанной предметной области представлены в виде файлов датасета. При этом для сущностей Client, Physical, Company, Plan, Subscriber, PSXAttrs данные предоставляются как один текущий снимок - т.е. одна сущность - один файл с актуальными данными.

Данные с коммутаторов выгружаются каждые 10 минут, т.е. при наличии например 10и коммутаторов за сутки будет выгружено 24*6*10 = 1440 файлов. Имена файлов содержат название коммутатора и время выгрузки. В качестве альтернативы можно организовать выгрузку с оборудования не в ввиде файлов, а в виде сообщений в потоковую систему (например Kafka).

В данной работе представленны три варианта датасета – выгрузки за 7 дней с 6-и коммутаторов для различных по размеру операторов связи:

- telecom10k - оператор связи с 10000 абонентов (51M),
- telecom100k - оператор связи с 100000 абонентов (696M),
- telecom1000k - оператор связи с 1000000 абонентов (7,2G).

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).

Для решения этой задачи необходимо использовать методы: очистки данных, загрузки данных, расчет витрин данных и т. д. Помимо анализа данных, важно применять практики управления данными (data governance) для контроля качества данных на всех этапах анализа (data quality), определения происхождения данных (data lineage), описания глоссария предметной области.

Ожидаемый результат. Необходимо на основе имеющихся данных построить набор витрин с интервалом расчета в 1 час входных данных и содержащих информацию о расходе трафика потребителями и признака подозрения на взлом.

Формат решения. Результаты, которые демонстрируются аналитики должны содержать:

- список имен взломанных абонентов,
- набор выходных таблиц (витрин), содержащих расчетные данные, построенные по данным выгрузки с телеком-оборудования,
- набор отчетов о контроле качества данных (data quality) на всех этапах обработки данных,
- диаграмму происхождения данных (data lineage) для предоставляемых витрин,
- глоссарий предметной области,
- документ — презентация с демонстрацией хода решения.

Аббревиатуры

MDM Master Data Management, системы управления мастер-данными. 2

OBS Online Billing System, систем онлайн - биллинга(ПАК обеспечения систем связи). 2

OLTP Online Transaction Processing, система обработки транзакций в реальном времени.. 2

ПАК Программно аппаратный комплекс. 7