TelecomX Dataset Description

By D.V. Iatsenko *

May 5, 2024
v1.0

Contents

*Southern Federal University: iatcenko@sfedu.ru, d.yacenko@gmail.com

1 Introduction

We present a synthetic dataset based on a real case commonly encountered in the telecommunications industry. This case is part of the real production process solved by the data analytics team of a telecom operator.

General Case Description. The company providing communication services offers various tariffs for Internet communication channels. However, despite the channels being limited by tariffs, the company monitors the dynamics of Internet traffic consumption by users. If an atypical surge in consumption is detected, it is assumed that the subscriber's computer may have been hacked and used for malicious activities such as serving as a spam distribution server, a DDOS network element, or hosting prohibited content. In such cases, the provision of Internet services is temporarily blocked, and the subscriber is contacted to clarify the situation.

2 Data Description

During analysis, data analysts receive exports from various systems:

- MDM (see Abbreviations section) - master data about business entities that provide context for business transactions,
- OLTP - operational data on current business transactions,
- OBS -operational data from online billing systems about subscriber connections, exported at short intervals. Typically, such data is collected from network switches, e.g., NetFlow, aggregated in micro-batches, and exported as a pre-billing level of data.

All these data are exported in various file formats to the analytics storage (e.g., NFS, HDFS). In some cases, operational billing data may be delivered not as files but through streaming systems, e.g., Kafka. See Figure 1 for the general logical data model. Let's examine the data in more detail.

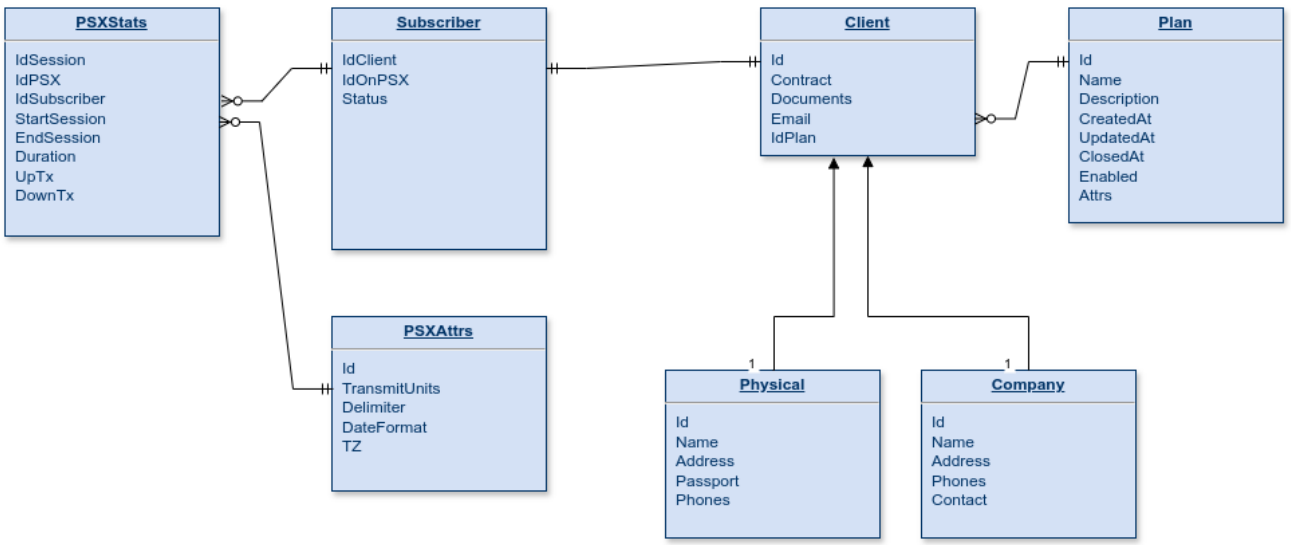

Figure 1: Logical Data Model

Information about business entities is contained in tables for Client, Physical, Company, and Plan.

Client - - the base entity for Physical and Company. Contains a general part of information from inherited entities. Data is delivered as a compressed Parquet file with an embedded data schema. An example of data is given in Table 1. Description of entity attributes:

- Id - client identifier,
- Contract - contract number,
- Documents - storage of copies of client documents,
- Email - client's email address,

- IdPlan - tariff plan identifier (see Plan entity).

Table 1: Example of Client Data.

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 - this entity describes the company's individual clients. Inherits from Clients. Data is delivered as a compressed Parquet file with schema combining data from the base and inherited entities. An example of data is given in Table 2.

- Id - client identifier,
- Name - client's full name,
- Address - residential address,
- Passport - passport data,
- Phones - phones.

Table 2: Example of Physical Data.

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 - this entity describes the company's corporate clients. Inherits from Clients. Data is presented as combine of parent and inherited entities. Data is delivered as a compressed Parquet file with schema combining data from the base and inherited entities. An example of data is given in Table 3.

- Id - client identifier,
- Name - company name,
- Address - legal address,
- Phones - company phones,
- Contact - contact details of company employees.

Table 3: Example of Physical Data.

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 - this entity describes the tariff plan. Data is provided in JSON format. An example of data is given in the Listing 1.

- Id - *int*, tariff plan identifier,
- Name - *text*, name of the tariff plan,
- Description - *text*, brief description of the tariff plan,
- CreatedAt - *long*, opening date (epochs in milliseconds),
- UpdatedAt - *long*, date change (epoch in milliseconds),
- ClosedAt - *long*, closing date (epochs in milliseconds),
- Enabled - *bool*, active/disabled,
- Attrs — *text*, internal attributes of the plan.

Listing 1: Example of Plan Data

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

Subscriber - this entity describes the current status of the subscriber, the subscriber number in the communication equipment, and their identifier in the primary data directory. Data is provided in CSV format. An example of data is given in the Listing 2.

- IdClient - *uuid*, client identifier in the Client entity,
- IdOnPSX - *int*, subscriber number in line communication equipment (PSX),
- Status - *text*, subscriber allowed/denied (ON/OFF).

Listing 2: Example of Subscriber Data

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

Additional entities contain operational information.

PSXStats - This entity describes the current state of subscriber connections. Data are extracted from line equipment every 10 minutes. Each extraction from each switch creates a new file, additional to those already existing. Data files are provided in either CSV or TXT format, depending on the type of source equipment. Differences in the format of the extracted data pertain to the delimiter symbol, date format, and units of traffic volume measurement (see PSXAttrs entity for details). An example of the data is provided in the listing (list 3,4).

- IdSession - *long*, session ID for the subscriber connection. This ID remains unchanged unless the client disconnects and reconnects. The duration of a session typically ranges from one hour to a full day.
- IdPSX - *int*, identifier of the switch (refer to the PSXAttrs entity).
- IdSubscriber - *int*, subscriber number in the line equipment (refer to the Subscriber entity).
- StartSession - *text*, start date of the subscriber's connection session, date format as per PSXAttrs.
- EndSession - *text*, end date of the subscriber's connection session, formatted as per PSXAttrs; if the session does not end in the current ten-minute interval, this field is left blank.
- Duration - *int*, duration of the connection session in seconds.
- UpTx - *long*, volume of traffic transmitted from the subscriber, units of measurement as per PSXAttrs.
- DownTx - *long*, volume of traffic downloaded by the subscriber, units as per PSXAttrs.

Listing 3: Example of PSXStats Data (CSV)

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

Listing 4: Example of PSXStats Data (TXT)

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

PSXAttrs —this entity describes the parameters (profiles) of communication switches. The data is provided in a CSV format:

- Id - *int*, identifier of the switch equipment,
- PSX - *text*, name of the switch,
- TransmitUnits - *text*, traffic measurement unit in bits/bytes,
- Delimiter - *text*, character used as a delimiter in export files,
- DateFormat - *text*, format for presenting data in export files,
- TZ - *text*, timezone of the switch equipment date.

Listing 5: Example of PSXAttrs Data

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

3 Problem Statement

The described subject area data is presented as dataset files. For entities such as Client, Physical, Company, Plan, Subscriber, and PSXAttrs, data are provided as a current snapshot — one entity, one file with current data.

Data from switches are exported every 10 minutes. For example, with 10 switches, 24*6*10 = 1440 files are exported per day. File names contain the switch name and export time. Alternatively, data could be streamed from equipment via systems like Kafka.

This work presents three dataset variants — exports from 6 switches over 7 days for operators of varying sizes:

- telecom10k - operator with 10,000 subscribers (51MB),
- telecom100k - operator with 100,000 subscribers (696MB),
- telecom1000k - operator with 1,000,000 subscribers (7.2GB).

The task is to analyze the presented data on Internet traffic consumption by subscribers, by examining the data on the volume of transmitted and received traffic from the communication equipment (switches) exports. During the analysis, it is necessary to compare the retrospective consumption of a subscriber's traffic with the current, and upon detecting atypical consumption, to conclude hacking. A data showcase table (data mart) should be constructed for each hour of data from the switches, i.e., the number of showcases should equal the period (in hours) for which operational data were exported, e.g., 24*7 = 168. The showcase should present the following data:

- Time,
- Client name,
- Client contract number,
- Contact data for communication with the client,
- Presumed hacking status (hacked/clear),
- Justification of the presumed hacking status (brief history of traffic consumption).

The methods required for this task include data cleaning, data loading, data mart calculation, etc. In addition to data analysis, it is important to apply data governance practices to control data quality at all stages of the analysis (data quality), determine data origins (data lineage), and describe the glossary of the subject area.

Expected Result. Based on the available data, a set of data marts with a calculation interval of 1 hour of input data should be constructed, containing information on consumer traffic consumption and signs of suspected hacking.

Solution Format. The results demonstrated by the analysts should include:

- A list of names of hacked subscribers,
- A set of output tables (data marts) containing calculated data, constructed from the telecom equipment data,
- A set of reports on data quality control at all data processing stages,
- A data lineage diagram for the provided showcases,
- A glossary of the subject area,
- A document—presentation demonstrating the course of the solution.

Acronyms

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