# Table of content:

[**Document Introduction**](#_zcuau2ot5kfa) **1**

[**Introduction**](#_sinarej6j672) **2**

[**Source System Integration**](#_lkwhdinwlm02) **3**

[**Data Format**](#_lkekslijhu0) **3**

[**Integration Risks and Recovery**](#_hulu37t2cd1o) **4**

[**ETL Mapping Rules**](#_irjopu1zr654) **4**

[**Discussions and Concerns**](#_qf2a4erf0330) **5**

# 

# **Document** **Introduction**

*This is an example of some source system data for different businesses and it doesn’t relate or correlate to any actual telecom company data. We simulate some data for general demo purposes as part of educational free content.*

*The dataset is a sample generated using a script and it does not describe the full telecom dataset for signalling but it simulates the general idea.*

# **Introduction**

This specification aims to document the source system data analysis and its properties. It provides the information related to the process of integration, data formats, and constraints of the data.

# **Source System Integration**

1. Source system will push a new dump file every 5 mins using SFTP protocol.

| **Interface Metric** | **Interface Detail** |
| --- | --- |
| Source Name | Signal\_4G |
| Source System Team | Engineering Team |
| Source System Owner | Ali Ali Ali |
| Protocol | SFTP |
| Latency | 5 min |
| Data Format | Pip separated '|' |
| Data Volume Per Dump (every 5 min) | 3.5GB |
| Daily Volume | ~ 1TB |
| File Naming Convention | ercsn\_4g\_(start-data YYYYmmddHHMMSS)\_(end-data YYYYmmddHHMMSS).CSV, Example -->  ercsv\_4g\_20201215164500\_20201215165000.csv |

# 

# 

# 

# 

# 

# **Data Format**

This section describes the input data and its format.

**Landing Zone File Format**

| **Column Name** | **Column Type** | **Is nullable** | **Column Sample Data** | **Description** |
| --- | --- | --- | --- | --- |
| IMSI | String | FALSE | 500300100 |  |
| IMEI | String | TRUE | 4000123498764356 |  |
| CELL | Int | FALSE | 123 |  |
| LAC | Int | FALSE | 12 |  |
| EVENT\_TYPE | String | TRUE | 1 |  |
| EVENT\_TS | Timestamp | FALSE | 2020-06-10 12:45:56 |  |

# **Integration Risks and Recovery**

Some risks could affect this project delivery as following:

1. **Data volume:** could be doubled/tripled in the peak hours, weekends, or public holidays.
2. **Event Latency Delay:** events could be delayed due to some network lags or high traffic in some areas. This latency affects some real-time use cases.
3. **Data format:** some rows could have null imsi due to network issues or some individual cases. This formatting issue could result in not being able to detect the member detail.

Due to the data volume and some performance improvement, the source files are available in the transient server for 24 hours from the file generation (number of files, server storage).

# **ETL Mapping Rules**

This section describes the ETL mapping from the landing zone to the modeled zone.

**Model Zone: cust\_loc\_dtl**

| **Column Name** | **Column Type** | **Mapping Rules** | **Target Model** |
| --- | --- | --- | --- |
| IMSI | String | As-Is (reject the record if null) | IMSI |
| IMSI | String | Join with IMSI references and get subscriber Id. Add -99999 if null. | subscriber Id |
| IMEI | String | substring first 8 chars, substring (0,7). If null or size is less than 14 Add -99999 | TAC |
| IMEI | String | substring first 8 chars, substring (8,13). If null or size is less than 14 Add -99999 | SNR |
| IMEI | String | As-Is | IMEI |
| CELL | Int | As-Is (reject the record if null) | CELL |
| LAC | Int | As-Is (reject the record if null) | LAC |
| EVENT\_TYPE | String | As-Is | EVENT\_TYPE |
| EVENT\_TS | Timestamp | validate the timestamp format to be YYYY-MM-dd HH:MM:SS (reject the record if null) | EVENT\_TS |

References:

* IMEI:<https://en.wikipedia.org/wiki/International_Mobile_Equipment_Identity>
* IMEI Numbers<https://youtu.be/BMGdigu4VI0>

# **Discussions and Concerns**

| **Issue** | **Details** | **Engineering Reply** | **Comments** |
| --- | --- | --- | --- |
| Data Latency | The business and development teams have critical concerns regarding the data latency every 10 min | * Latency has decreased for 5 min. * Reduce the latency could affect network performance, and it needs more time for assessment and testing. | Agreed based on further enhancement in the future to max 2 min. |
| Data Transfer mechanism | The development team needs to steam pull the data from the source. | * There is a security aspect which avoids any system to read directly from the network. * Pulling from the source could affect network functionality. However, we are assessing the solution to create a replica from the source systems for any systems that need to read directly from the server. This point depends on the first point to be approved. | Agreed |
| Data Format issues | There are some records that come with an unknown number and can’t match with imsi reference. | * This issue is acceptable as it has many reasons one of them is roaming subscribers. * In case of other problems with the data, we need to have more samples to analyze and try to avoid such behavior in the future. | Agreed |
| IMEI Format | We need more info to understand the IMEI format structure | * Wiki page link added to provide more info. | Agreed |