Strix

Data Ingestion Contract

Client: Strix

Project: DataLake for Growth

Date: 2025-05-26

Version: 1.1

---

1. Overview

This document outlines the agreed-upon standards and requirements for delivering datasets to the data for growth project. It includes
storage locations, data format requirements, and schema definitions to ensure consistency, compatibility, and quality throughout the
data pipeline.

---

2. Data Delivery Location

All datasets must be uploaded to the following Amazon S3 bucket:

- **Bucket Name**: `s3://strix-production-datalake-for-growth/`
- **Folder Structure**:
  - `/bronze/{source_name}/{dataset_name}/execution_date=YYYY-MM-DD/`
- **Permissions**:
  - Ensure the appropriate IAM roles are granted write access to the specified paths.
  - Use server-side encryption (SSE-S3) for all files.

---

3. File Format Requirements

All delivered files must conform to the following format:

- **File Format**: `Parquet`
- **Compression**: `Snappy`
- **Encoding**: `UTF-8`
- **Partitioning**: By execution_date (`YYYY-MM-DD`)

---

4. Dataset Specifications

Below is a list of datasets and their expected schemas.

4.1 Source: `calipso`

4.1.1 Dataset: `WINCLAP_CALIPSO_CLIENTES`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_CLIENTES/`
- **File Name Format**: `WINCLAP_CALIPSO_CLIENTES_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id_cliente	STRING
tipo_cliente	STRING
tipo_cartera	STRING
cliente	STRING
fecha_alta_cliente	DATETIME
categoria_iva_cliente	STRING
id_tipo_documento_cliente	STRING
numero_documento_cliente	STRING

4.1.2 Dataset: `WINCLAP_CALIPSO_DATOS_CONTACTO_CLIENTES`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_DATOS_CONTACTO_CLIENTES/`
- **File Name Format**: `WINCLAP_CALIPSO_DATOS_CONTACTO_CLIENTES_YYYYMMDD_HHMMSS.parquet`

- **Schema**:

Column Name	Data Type
id_cliente	STRING
tipo_cliente	STRING
tipo_cartera	STRING
ciudad_acuerdo	STRING
email_acuerdo	STRING
provincia_acuerdo	STRING
telefono_acuerdo	STRING

4.1.3 Dataset: `WINCLAP_CALIPSO_INVENTARIO_X_VEHICULOS`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_INVENTARIO_X_VEHICULOS/`
- **File Name Format**: `WINCLAP_CALIPSO_INVENTARIO_X_VEHICULOS_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id_place	STRING
tipo_cliente	STRING
tipo_cartera	STRING
tipo_vehiculo	STRING
nro_serie_equipo	STRING
estado_equipo	STRING
modelo_equipo	STRING
marca_equipo	STRING
modelo_equipo_normalizado	STRING

4.1.4 Dataset: `WINCLAP_CALIPSO_SERVICIOS_CLIENTES`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_SERVICIOS_CLIENTES/`
- **File Name Format**: `WINCLAP_CALIPSO_SERVICIOS_CLIENTES_YYYYMMDD_HHMMSS.parquet`

- **Schema**:

Column Name	Data Type
id_cliente_serv	STRING
id_cliente	STRING
id_place	STRING
id_plan_cuenta	STRING
plan_cuenta	STRING
estado_plan_cuenta	STRING
id_cuenta	STRING
cuenta	STRING
servicio_cuenta	STRING
cliente_serv	STRING

4.1.5 Dataset: `WINCLAP_CALIPSO_VEHICULOS`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_VEHICULOS/`
- **File Name Format**: `WINCLAP_CALIPSO_VEHICULOS_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id_place	STRING
id_cliente	STRING
tipo_cliente	STRING
tipo_cartera	STRING
id_cuenta	STRING
tipo_vehiculo	STRING
dominio	STRING
marca_vehiculo	STRING
valor_mercado	FLOAT
marca_modelo	STRING

4.1.6 Dataset: `WINCLAP_CALIPSO_ESTADO_CREDITICIO_CLIENTE`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/calipso/WINCLAP_CALIPSO_ESTADO_CREDITICIO_CLIENTE/`
- **File Name Format**: `WINCLAP_CALIPSO_ESTADO_CREDITICIO_CLIENTE_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id_cliente	STRING
id_acuerdo	STRING
estado_saldo	STRING
empresa	STRING

4.2 Source: `turnero`

4.2.1 Dataset: `WINCLAP_TURNOS_GESTIONES`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/turnero/WINCLAP_TURNOS_GESTIONES/`
- **File Name Format**: `WINCLAP_TURNOS_GESTIONES_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
solicitud_id	STRING
turno_id	STRING
fecha_turno	DATETIME
fecha_asignacion_turno	DATETIME
tipo_solicitud	STRING
servicio	STRING
es_autoturnado	BOOLEAN
fecha_alta	DATETIME
no_gestiona	BOOLEAN
producto	STRING

4.2.2 Dataset: `WINCLAP_TURNOS_PERSONAS`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/turnero/WINCLAP_TURNOS_PERSONAS/`
- **File Name Format**: `WINCLAP_TURNOS_PERSONAS_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
solicitud_id	STRING
apellido	STRING
nombre	STRING
tipo_documento	STRING
nro_documento	STRING
telefono	STRING
email	STRING
direccion	STRING
localidad	STRING
provincia	STRING

4.2.3 Dataset: `WINCLAP_TURNOS_SERVICIOS`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/turnero/WINCLAP_TURNOS_SERVICIOS/`
- **File Name Format**: `WINCLAP_TURNOS_SERVICIOS_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
solicitud_id	STRING
turno_id	STRING
calipso_plan_id	STRING
calipso_plan_name	STRING
calipso_service_id	STRING
calipso_service_name	STRING
sponsor_id	STRING
sponsor	STRING

4.2.4 Dataset: `WINCLAP_TURNOS_VEHICULOS`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/turnero/WINCLAP_TURNOS_VEHICULOS/`
- **File Name Format**: `WINCLAP_TURNOS_VEHICULOS_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
solicitud_id	STRING
dominio	STRING
marca	STRING
modelo	STRING
valor_vehiculo	FLOAT

4.3 Source: `magenta`

4.3.1 Dataset: `accounts`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/accounts/`
- **File Name Format**: `accounts_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
identification_type	STRING
identification_number	STRING
name	STRING
active	BOOLEAN
country_id	STRING
created_datetime	DATETIME
last_update_datetime	DATETIME
services	STRING

4.3.2 Dataset: `devices`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/devices/`
- **File Name Format**: `devices_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
user_id	STRING
app_installation_id	STRING
app_version_id	STRING
battery_level	FLOAT
created_by	STRING
created_timestamp	DATETIME
identifier	STRING
last_modified_by	STRING
last_modified_timestamp	DATETIME
location_accuracy	FLOAT
location_coordinates	STRING
location_type	STRING
location_timestamp	DATETIME
make	STRING
model	STRING
name	STRING
push_notifications_enabled	BOOLEAN
system_name	STRING
system_version	STRING
token	STRING
tracking_enabled	BOOLEAN
last_update_datetime	DATETIME

4.3.5 Dataset: `flexes`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/flexes/`

- **File Name Format**: `flexes_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
label	STRING
latitude	DECIMAL
longitude	DECIMAL
battery_level	FLOAT
things	STRING
location	STRING
location_recorded_at	DATETIME
created_datetime	DATETIME

4.3.4 Dataset: `gpses`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/gpses/`
- **File Name Format**: `gpses_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
make	STRING
model	STRING
serial_number	STRING
parent_id	STRING
template_id	STRING
created_datetime	DATETIME

4.3.6 Dataset: `homes`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/homes/`

- **File Name Format**: `homes_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
label	STRING
address_line1	STRING
city	STRING
state	STRING
latitude	DECIMAL
longitude	DECIMAL
things	STRING
status_datetime	DATETIME
created_datetime	DATETIME

4.3.3 Dataset: `users`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/users/`
- **File Name Format**: `users_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
username	STRING
first_name	STRING
last_name	STRING
signup_completed	BOOLEAN
has_ios	BOOLEAN
has_android	BOOLEAN
has_device	BOOLEAN
last_device_login	DATETIME

4.3.6 Dataset: `vehicles`

- **Path**: `s3://strix-production-datalake-for-growth/bronze/magenta/vehicles/`
- **File Name Format**: `vehicles_YYYYMMDD_HHMMSS.parquet`
- **Schema**:

Column Name	Data Type
id	STRING
account_id	STRING
make	STRING
year	INTEGER
color	STRING
label	STRING
model	STRING
domain	STRING
subtype	STRING
engine_number	STRING
chassis_number	STRING
mileage	FLOAT
latitude	FLOAT
longitude	FLOAT
things	STRING
location_datetime	DATETIME
created_datetime	DATETIME

---

5. Delivery Frequency

All dataset must be updated each first day of the month except the datasets from `turnero` that must be daily.

---

6. Contact

For any questions or issues related to this data contract:

- **Technical Contact**: Tomas Recalt, Solution Architect
  Email: tomas.recalt@winclap.com
- **Technical Contact**: Federico Jaureguialzo, Data Engineer
  Email: federico@winclap.com