

Documentación Mage ai

Tubería ELT

Lista de bloques:

- 1. **load_from_mysql**: carga los datos desde una base de datos MySQL local
- 2. **export_2_snowflake**: exporta los datos procesados al schema RAW en la base de datos INSTACART_DB en Snowflake.

Modelamiento Star-Schema

Tablas actuales:

- AISLES (AISLE_ID , AISLE)
- DEPARTMENTS (DEPARTMENT_ID , DEPARTMENT)
- ORDERS (ORDER_ID , USER_ID , ORDER_NUMBER, ORDER_DOW, ORDER_HOUR_OF_DAY, DAYS_SINCE_PRIOR_ORDER)
- ORDER_PRODUCTS (ORDER_ID , PRODUCT_ID , ADD_TO_CART_ORDER, REORDERED)
- PRODUCTS (PRODUCT_ID, PRODUCT_NAME, AISLE_ID , DEPARTMENT_ID)

Tablas de hechos:

FACT_ORDERS
ORDER_ID (PK)
USER_ID
DAY_ID (FK)
ORDER_NUMBER
DAYS_SINCE_PRIOR_ORDER

FACT_ORDER_PRODUCTS
ORDER_ID (FK)
PRODUCT_ID (FK)
ADD_TO_CART_ORDER
REORDERED

Tablas de dimensiones:

DIM_PRODUCTS
PRODUCT_ID (PK)
PRODUCT_NAME
DEPARTMENT
AISLE

DIM_DAY
DAY_ID (PK)
ORDER_DOW
ORDER_HOUR_OF_DAY
IS_WEEKEND
TIME_OF_DAY

ETL Pipeline

- 1. **raw_from_snowflake**: Carga datos de RAW desde Snowflake.
- 2. **manage_nan**: Maneja valores nulos (NaN) en los datos de acuerdo al plan de acción del EDA.
 - a. print:

Tabla: 'PRODUCTS', Columna: 'PRODUCT_NAME' → Nans a rellenar: 1258

Tabla: 'ORDERS', Columna: 'DAYS_SINCE_PRIOR_ORDER' → Nans a rellenar: 28819

Tabla: 'ORDER_PRODUCTS', Columna: 'ADD_TO_CART_ORDER' → Nans a rellenar: 836

- 3. **reformat_strings**: Aplica transformaciones a las cadenas de texto, como poner todo en minúsculas y quitar espacios incesarios.
- 4. **manage_duplicates**: Identifica y maneja registros duplicados en los datos de acuerdo al plan de acción del EDA.

a. print:

```
PRODUCTS - registros antes: 49694 después: 49590 eliminados: 104

ORDERS - registros antes: 478967 después: 478952 eliminados: 15
```

5. **stats_after_clean**: Bloque para verificar la limpieza en los datos.

a. print:

```
=====

Análisis de la tabla: AISLES

Dimensiones: (134, 2)

=====

Estadísticas por columna:

Columna: AISLE_ID

valores_unicos: 134

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

Columna: AISLE

valores_unicos: 134

tipos_datos: ["<class 'str'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

=====

=====

Análisis de la tabla: DEPARTMENTS

Dimensiones: (21, 2)

=====

Estadísticas por columna:

Columna: DEPARTMENT_ID
```

valores_unicos: 21

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

Columna: DEPARTMENT

valores_unicos: 21

tipos_datos: ["<class 'str'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

=====

=====

Análisis de la tabla: PRODUCTS

Dimensiones: (49590, 4)

=====

Estadísticas por columna:

Columna: PRODUCT_ID

valores_unicos: 49590

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

Columna: PRODUCT_NAME

valores_unicos: 48333

tipos_datos: ["<class 'str'>"]

nulos: 0

duplicados_unicos: 1257

duplicados_totales: 1258

Columna: AISLE_ID

valores_unicos: 134

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 49456

duplicados_totales: 49590

Columna: DEPARTMENT_ID

valores_unicos: 21

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 49569

duplicados_totales: 49590

=====

=====

Análisis de la tabla: ORDERS

Dimensiones: (478952, 6)

=====

Estadísticas por columna:

Columna: ORDER_ID

valores_unicos: 478952

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 0

duplicados_totales: 0

Columna: USER_ID

valores_unicos: 157437

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 321515

duplicados_totales: 423595

Columna: ORDER_NUMBER

valores_unicos: 100

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 478852

duplicados_totales: 478952

Columna: ORDER_DOW

valores_unicos: 7

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 478945

duplicados_totales: 478952

Columna: ORDER_HOUR_OF_DAY

valores_unicos: 24

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 478928

duplicados_totales: 478952

Columna: DAYS_SINCE_PRIOR_ORDER

valores_unicos: 31

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 478921

duplicados_totales: 478952

=====

=====

Análisis de la tabla: ORDER_PRODUCTS

Dimensiones: (4545007, 4)

```

=====

Estadísticas por columna:

Columna: ORDER_ID

valores_unicos: 450046

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 4094961

duplicados_totales: 4523160

Columna: PRODUCT_ID

valores_unicos: 45477

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 4499530

duplicados_totales: 4539980

Columna: ADD_TO_CART_ORDER

valores_unicos: 65

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 4544942

duplicados_totales: 4545007

Columna: REORDERED

valores_unicos: 2

tipos_datos: ["<class 'int'>"]

nulos: 0

duplicados_unicos: 4545005

duplicados_totales: 4545007

=====

```

6. **modeling**: Crea las nuevas tablas de hechos y dimensiones a partir de las tablas limpias.
7. **export_2_clean_snowflake**: Exporta las nuevas tablas al schema CLEAN de la base de datos INSTACART_DB en Snowflake.

a. print:

Cargando datos de la tabla: FACT_ORDERS con dimensiones: (478952, 5)
Cargando datos de la tabla: FACT_ORDER_PRODUCTS con dimensiones: (4545007, 4)
Cargando datos de la tabla: DIM_PRODUCTS con dimensiones: (49590, 4)
Cargando datos de la tabla: DIM_DAY con dimensiones: (168, 5)

Relaciones tablas creadas

INSTACART_DB / CLEAN / FACT_ORDERS

Table PLOMERO 4 minutes ago 479.0K 3.4MB

Table Details Columns Data Preview Copy History

Table definition

```
1 create or replace TABLE INSTACART_DB.CLEAN.FACT_ORDERS (  
2   ORDER_ID NUMBER(38,0),  
3   USER_ID NUMBER(38,0),  
4   DAY_ID NUMBER(38,0),  
5   ORDER_NUMBER NUMBER(38,0),  
6   DAYS_SINCE_PRIOR_ORDER NUMBER(38,0),  
7   constraint PK_FACT_ORDERS primary key (ORDER_ID),  
8   constraint FK_FACT_ORDERS_DAY foreign key (DAY_ID)  
   references INSTACART_DB.CLEAN.DIM_DAY(DAY_ID)  
9 );  
  
Show less ^
```

INSTACART_DB / CLEAN / FACT_ORDER_PROD...

Table PLOMERO 2 minutes ago 4.5M 24.8MB

Table Details Columns Data Preview Copy History

Table definition

```
1 create or replace TABLE  
INSTACART_DB.CLEAN.FACT_ORDER_PRODUCTS (  
2   ORDER_ID NUMBER(38,0),  
3   PRODUCT_ID NUMBER(38,0),  
4   ADD_TO_CART_ORDER NUMBER(38,0),  
5   REORDERED BOOLEAN,  
6   constraint FK_FACT_ORDER_PRODUCTS_ORDER foreign key  
   (ORDER_ID) references  
INSTACART_DB.CLEAN.FACT_ORDERS(ORDER_ID),  
7   constraint FK_FACT_ORDER_PRODUCTS_PRODUCT foreign key  
   (PRODUCT_ID) references  
INSTACART_DB.CLEAN.DIM_PRODUCTS(PRODUCT_ID)  
8 );  
  
Show less ^
```

INSTACART_DB / CLEAN / DIM_DAY

...

Load Data

Table

PLOMERO

3 minutes ago

168

2.5KB

Table Details

Columns

Data Preview

Copy History

Table definition

1 create or replace TABLE INSTACART_DB.CLEAN.DIM_DAY (
2 DAY_ID NUMBER(38,0),
3 ORDER_DOW NUMBER(38,0),
4 ORDER_HOUR_OF_DAY NUMBER(38,0),
5 IS_WEEKEND BOOLEAN,
6 TIME_OF_DAY VARCHAR(16777216),
7 constraint PK_DIM_DAY primary key (DAY_ID)
8);

Show less ^

INSTACART_DB / CLEAN / DIM_PRODUCTS

...

Load Data

Table

PLOMERO

1 minute ago

49.6K

1.3MB

Table Details

Columns

Data Preview

Copy History

Table definition

1 create or replace TABLE INSTACART_DB.CLEAN.DIM_PRODUCTS (
2 PRODUCT_ID NUMBER(38,0),
3 PRODUCT_NAME VARCHAR(16777216),
4 DEPARTMENT VARCHAR(16777216),
5 AISLE VARCHAR(16777216),
6 constraint PK_DIM_PRODUCTS primary key (PRODUCT_ID)
7);

Show less ^


Automatización


A modo de ejemplo se implemento que el pipeline ETL se corriera de manera diaria pero se podría establecer cualquier frecuencia de tiempo.


instacart_project > Pipelines > etl > Triggers > ETL Instacart diaria


v0.9.7514:09 UTCLive help


</>

















ETL Instacart diaria

Settings

<

Trigger type

schedule

•

Status

active

⌚

Frequency

daily

📅

Next run date

2025-02-18 00:00:00

📅

Start date

2025-02-17 14:08:00

📅

Last enabled at

2025-02-17 14:09:14

Runtime variables

tables

["AISLES", "DEPARTMENTS", "PRODUCTS", "C

execution date

<run_datetime>

⛔ Disable trigger

🔄 Run@once

✎ Edit trigger

All statuses ▾

Runs for this trigger

No runs available