Sebastián Martínez
martinez.sebastian314159@gmail.com

# Documento de Diseño de Modelo Dimensional para Innova.

Prueba técnica para Data Engineer en Alegra.

Sebastián Martínez
martinez.sebastian314159@gmail.com

# Índice

Sebastián Martínez
martinez.sebastian314159@gmail.com

# Introducción

## Contexto y proceso de negocio

Innova es una empresa de retail líder en el mercado mexicano, la cual pretende realizar una transformación tecnológica para obtener mejor insight de sus datos. Para esto desea implementar una solución Cloud a modo de Data Warehouse que beba de sus sistemas OLTP. Esta solución debe responder y atender tanto a las necesidades de almacenamiento on demand y streaming de data como a la de reportería en tiempo real.

## Descripción de la solución

Se espera que la data recibida del cliente sea en tiempo real, dado a el gran volumen y frecuencia de las compras en retail, la granularidad del sistema debe ser tolerante hasta el el intérvalo del segundo. Por otro lado se entiende que el cliente pueda otorgar data on demand con frecuencia diara, semanal o mensual. Debido a esto, los primeros pasos de las tablas albergarán la data en tablas particionadas por día de ingesta.
Se propone una arquitectura de medallas o salto múltiple, que permita mantener la misma data repartida en diferentes capas según calidad de data.

**Data Lake - Landing (Capa 1):**

- Almacenamiento de la data del cliente en tablas y buckets particionados por fecha de ingesta.
- Poca a nula validación y limpieza de datos.

**Data Lake - Stagging (Capa 2):**

- Tablas con validaciones de datos y enriquecida en granularidad y especificidad.

**DataWarehouse - Production (Capa 3):**

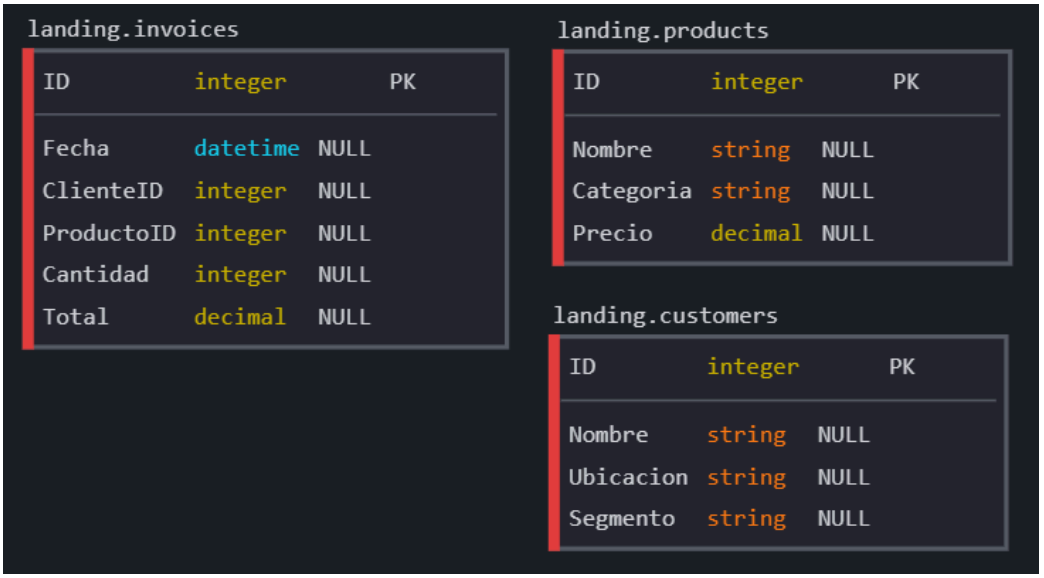
- Modelo relacional establecido y definido.
- Data llega 100% validada.

Sebastián Martínez
martinez.sebastian314159@gmail.com

# Diagramas

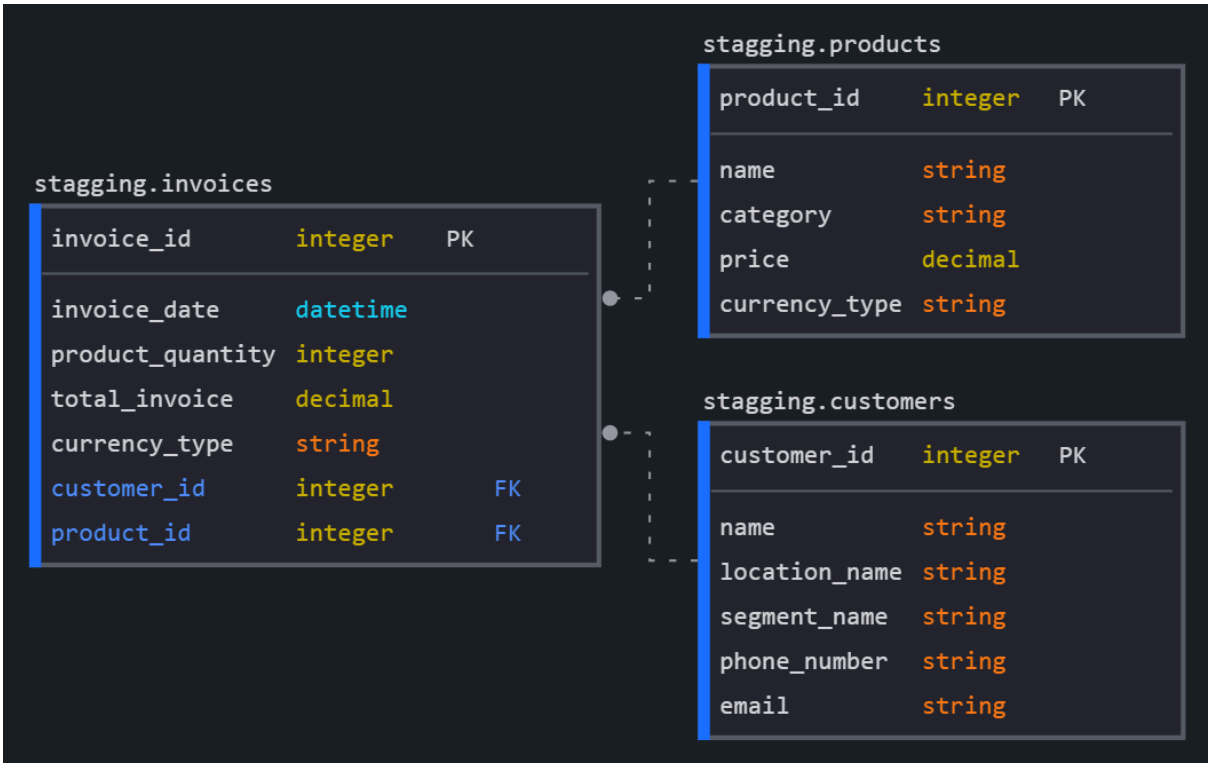
## Etapa Landing

Las tablas corresponden a los archivos entregados en csv, todas particionadas por fecha de ingesta, no hay validaciones ni trabajo de datos.




## Etapa Stagging

Tablas en esta etapa mantienen la estructura anterior, pero los campos están estandarizados en inglés, usando minúsculas. Campos agregados que añaden granularidad considerada necesaria.



Sebastián Martínez
martinez.sebastian314159@gmail.com

# Etapa Production

Modelo dimensional completo, dimensiones añadidas para especificidad.

Sebastián Martínez
martinez.sebastian314159@gmail.com

# Diccionario de datos: Modelo dimensional

## 1 - Dimensión Productos: `dw_prod.products_dim`

| **Columna** | **Descripción** |
|---|---|
| id | Clave subrogada autoincrementable. |
| product_id | Id heredado desde el archivo original. |
| name | Nombre del producto, proveniente del archivo original. |
| category | Categoría del producto, proveniente del archivo original. |
| price | Precio de producto, en decimal. |
| currency_type | Tipo de moneda, por default ‘MXN’ (divisa mexicana) |
| ingestion_date | Fecha de recepción del dato de manos del cliente, si es que no se especifica por parte del cliente. |
| last_modified_date | Última fecha de modificación de la fila, proveniente de los pipelines ETL. |

## 2 - Dimensión Tiempo: `dw_prod.time_dim`

| **Columna** | **Descripción** |
|---|---|
| id | Clave subrogada autoincrementable. |
| date | Fecha en formato ‘yyyy-mm-dd’. |
| year | Año de la fecha, como entero. |
| quarter | Trimestre asociado a la fecha, como entero, partiendo del 0. |
| semester | Semestre asociado a la fecha, como entero, partiendo del 0. |
| month | Mes asociado a la fecha, como entero, partiendo del 1. |
| month_string | Mes escrito en letras minúsculas y en inglés. |
| day | día del mes, como entero. |
| day_of_week_string | día de la semana escrito letras minúsuclas y en inglés. |
| hour_24 | Hora asociada a la fecha, como entero, en formato de 24 horas. |
| hour_12 | Hora asociada a la fecha, como entero, en formato de 12 horas. |
| minutes | Minutos asociados a la fecha, como entero, partiendo del 0. |
| seconds | Segundos asociados a la fecha, como entero, partiendo del 0. |
| max_date_ingested | Última fecha ingestada en la tabla. |
| min_date_ingested | Primer fecha ingestada en la tabla. |
| ingestion_date | Fecha de recepción del dato de manos del cliente, si es que no se especifica por parte del cliente. |

Sebastián Martínez
martinez.sebastian314159@gmail.com

| | |
|---|---|
| last_modified_date | Última fecha de modificación de la fila, proveniente de los pipelines ETL. |

### 3 - Dimensión Clientes: `dw_prod.customers_dim`

| **Columna** | **Descripción** |
|---|---|
| id | Clave subrogada autoincrementable. |
| customer_id | Id heredado desde el archivo original. |
| name | Nombre del cliente, proveniente del archivo original. |
| location_name | Ubicación del cliente, proveniente del archivo original. |
| segment_name | Segmento del cliente, proveniente del archivo original. |
| phone_number | Número de teléfono del cliente, sugerido como dato adicional. |
| email | Dirección de correo electrónico del cliente, sugerido como dato adicional. |
| ingestion_date | Fecha de recepción del dato de manos del cliente, si es que no se especifica por parte del cliente. |
| last_modified_date | Última fecha de modificación de la fila, proveniente de los pipelines ETL. |

### 4 - Dimensión Tipos de Pago: `dw_prod.payment_methods_dim`

| **Columna** | **Descripción** |
|---|---|
| id | Clave subrogada autoincrementable. |
| payment_method_id | Id del tipo de pago, sugerido como nuevo dato. |
| method | Nombre del tipo de pago. |
| description | Breve descripción del tipo de pago. |
| ingestion_date | Fecha de recepción del dato de manos del cliente, si es que no se especifica por parte del cliente. |
| last_modified_date | Última fecha de modificación de la fila, proveniente de los pipelines ETL. |

Sebastián Martínez
martinez.sebastian314159@gmail.com

## 5 - Hechos Facturas: `dw_prod.fact_invoices`

| Columna | Descripción |
|---|---|
| id | Clave subrogada autoincrementable. |
| invoice_id | Id de la factura, repetible como tantos elementos tenga el detalle de la factura. |
| time_id | Id de la dimensión de tiempo según la fecha de la factura. |
| customer_id | Id de la dimensión del cliente asociado a la factura. |
| product_id | Id de la dimensión de productos asociado a la factura. |
| payment_method_id | Id de la dimensión de tipos de pago asociado a la factura. |
| product_quantity | Cantidad del producto asociado al ítem del detalle de la factura. |
| total_per_product | Total del precio asociado al producto del detalle de la factura. |
| total_invoice | Total del precio de la factura completa, campo desnormalizado por eficiencia. |
| currency_type | Tipo de divisa empleado en la factura. |
| invoice_state | Estado asociado a la factura. (emitida, cancelada, enviada, pagada, en revisión.) |
| ingestion_date | Fecha de recepción del dato de manos del cliente, si es que no se especifica por parte del cliente. |
| last_modified_date | Última fecha de modificación de la fila, proveniente de los pipelines ETL. |

Sebastián Martínez
martinez.sebastian314159@gmail.com

# Aclaciones y Justificación

El cliente hizo entrega de archivos útiles, pero carentes de granularidad para el detalle de la factura, se asume que el detalle quedará disperso en la tabla de hechos, uniendo el detalle por el id compuesto de la factura.

Cada dimensión posee `ingestion_date` y `last_modified_date`, metadata necesaria de cada registro, que lo ubica temporalmente.

Se agregó la dimensión de tiempo time_dim para facilitar el cálculo por diferentes periodos. Esta dimensión posee el campo `max_date_ingested` y `min_date_ingested`, estos cumplen el fin de indicar qué intérvalo de años existen dentro de la tabla, sin necesidad de hacer una query extensa antes de la ingesta de nuevos años.

Se añadió una dimensión acerca el tipo de pago de la factura `payment_methods_dim`, esto debido a la posibilidad de internacionalizar el modelo, o de la futura implementación de esta dimensión.

Se sugiere la inclusión del valor del dolar estadounidense (USD) asociado a la fecha de cada factura, de esta forma se puede realizar el cálculo necesario para la conversión en reportería internacional.

Las capas sugeridas están sometidas a evaluación, por lo que puede que se añadan o se retiren ciertas capas, en pos de asegurar la rapidez y disponibilidad de la data.

El modelo multicapas otorga un entorno de trabajo seguro e iterativo, al no modificar directamente las tablas operativas, permitiendo que las modificaciones posteriores sean realizables con facilidad.