

Retail Store Analytics

Proyecto de Inteligencia de Negocios



Autor:

Bryan Gustavo Guapulema Arellano

Plataforma:

Talend Cloud

Tabla de contenido

1. Introducción	3
1.1. Descripción del documento	3
1.2. Criterios técnicos	3
1.3. Criterios de evaluación	3
1.4. Definiciones	3
1.4.1. Retail Stores.....	3
1.4.2. Talend Cloud(TC)	4
2. Infraestructura del proyecto.....	5
2.1. Fuentes de datos	5
2.2. Herramientas y servicios	5
2.3. Arquitectura	6
3. Metodología y Desarrollo del Proyecto	7
3.1. Ingesta de datos: Talend Data Inventory (TDI).....	7
3.2. Limpieza de datos (ETL): Talend Data Preparation	7
3.3. Verificación de casos (ETL): Talend Data Stewardship	12
3.4. Volcado de datos (Single Store)	15
4. Visualización	18
4.1. Conexión a Power BI.....	18
4.2. Modelo de datos	18
4.3. Medidas DAX.....	19
5. Visualización	20
5.1. Dashboard.....	20
5.2. Historia construida a partir del dashboard	20
6. Recursos	22

1. Introducción

1.1. Descripción del documento

El presente documento tiene como objetivo describir de forma detallada el desarrollo técnico y analítico del proyecto de Inteligencia de Negocios “Retail Store Analytics”, implementado en la plataforma Talend Cloud.

Este informe recoge todos los procesos ejecutados para la creación de una solución end-to-end, que aborda el ciclo completo de vida de los datos desde la extracción y transformación de datos hasta la visualización final de indicadores comerciales.

El documento está estructurado para reflejar las fases del ciclo de vida del proyecto y los componentes utilizados en cada una de ellas. Además, se detallan las fuentes de información utilizadas, las herramientas de Talend empleadas, el modelo de datos creados, y el dashboard gerencial resultante.

1.2. Criterios técnicos

- **Fuente Datos:** Utilizar una fuente de datos que contenga estructuras de tipo descriptivo y que presente oportunidades de mejora en la calidad de la información, por ejemplo: clientes, productos o proveedores.
- **Reglas:** Aplicar al menos 10 reglas de limpieza y validación de datos, orientadas a garantizar la consistencia, integridad y precisión de la información.
- **Integración Plataforma:** Durante el desarrollo, se deberán emplear todas las siguientes herramientas de la suite Talend: Talend Data Inventory (TDI), Talend Data Preparation (TDP), Talend Data Stewardship (DTS), Talend Management Console (TMC)

1.3. Criterios de evaluación

- Extracción, transformación y carga correctamente ejecutadas en la plataforma elegida.
- Reglas de negocio a fuentes datos
- Integración de toda la suite Talend
- Claridad en la explicación de resultados final.
- Claridad en la explicación de resultados final, importancia de gobierno de datos dentro de las organizaciones.
- Soluciones novedosas, visualizaciones impactantes, enfoque original.

1.4. Definiciones

1.4.1. Retail Stores

Retail Stores es una empresa ficticia del sector comercial minorista (retail) dedicada a la venta de una amplia gama de productos, que incluye artículos de consumo masivo, tecnología, moda, hogar y entretenimiento. Opera a través de una red de tiendas físicas distribuidas en diferentes regiones, complementadas con un canal de ventas en línea que permite atender a clientes en todo el país.

La compañía basa su estrategia en la eficiencia operativa, la experiencia del cliente y la inteligencia de datos, utilizando la analítica avanzada para optimizar decisiones sobre inventario, precios, promociones y rentabilidad.

En el contexto del proyecto, Retail Stores representa una organización moderna que gestiona grandes volúmenes de datos transaccionales y de clientes

1.4.2. Talend Cloud(TC)

Talend Cloud es una plataforma integral de integración y gobierno de datos en la nube que permite a las organizaciones conectar, transformar, limpiar, gobernar y compartir datos de forma segura y escalable.

Su arquitectura combina herramientas visuales, conectores listos para usar y capacidades de automatización que facilitan la creación de canales de datos confiables, alineados con principios de calidad, trazabilidad y cumplimiento normativo.

Talend Cloud está compuesto por varios módulos especializados que, trabajando en conjunto, cubren todo el ciclo de vida de los datos: desde su descubrimiento y preparación hasta su validación, monitoreo y explotación analítica.

A. Talend Data Inventory (TDI)

- Es el catálogo central de datos dentro de Talend Cloud.
- Permite conectar fuentes de datos (bases de datos, archivos, APIs, etc.) y explorar su contenido.
- Evalúa la calidad de los datos mediante indicadores automáticos (completitud, unicidad, validez, etc.) y permite documentar, clasificar y certificar datasets para promover su reutilización.
- Actúa como punto de partida para proyectos de preparación, limpieza o gobierno de datos.

B. Talend Data Preparation (TDP)

- Herramienta orientada al perfilado, limpieza y transformación de datos de manera visual e interactiva.
- Permite aplicar operaciones como normalización, reemplazo de valores, fusiones, cálculos, formatos y filtros sin necesidad de programar.
- Los datasets preparados pueden exportarse a otros entornos (como bases de datos, archivos o Data Stewardship) para ser usados en procesos analíticos o de gobierno.
- Facilita el trabajo de analistas de datos o business users en la mejora continua de la calidad de la información.

C. Talend Data Stewardship (TDS)

- Plataforma de gestión colaborativa de la calidad y validación de datos.
- Permite crear campañas de revisión donde los *data stewards* (responsables de datos) corrigen, validan o completan registros según reglas de negocio.
- Se integra con Data Preparation o Data Inventory para recibir los datos “dudosos” o con errores detectados.
- Asegura que los datos finales sean precisos, coherentes y aprobados antes de ser reutilizados.

D. Talend Management Console (TMC)

- Es el panel de administración central de Talend Cloud.
- Permite gestionar usuarios, roles, entornos, ejecuciones, planes y permisos.
- Controla el despliegue y la ejecución de tareas (jobs o pipelines), ya sea en la nube o mediante Remote Engines locales.
- Incluye herramientas de monitoreo, auditoría y automatización, esenciales para la gobernanza operativa del ecosistema Talend.

E. Talend Open Studio (TOS)

- Es la versión de escritorio (on-premise) de Talend, orientada a la creación de Jobs ETL/ELT (extracción, transformación y carga).
- Ofrece una interfaz gráfica basada en componentes reutilizables para construir flujos de integración complejos.
- Aunque no es parte directa del entorno cloud, se integra con Talend Cloud para subir y ejecutar Jobs diseñados localmente.
- Ideal para desarrolladores técnicos que requieren control detallado sobre los procesos de integración.

2. Infraestructura del proyecto

2.1. Fuentes de datos

Fuente de datos	CSV
Descripción:	Archivos estructurados con información de clientes, productos y detalles de compra
Conexión	Disponible en: https://github.com/BryanGuapulema/Hackaton-3 Obtenido de: https://www.kaggle.com/datasets/ahmedmohamed2003/retail-store-sales-dirty-for-data-cleaning
Contenido	<ul style="list-style-type: none"> - mappings.csv - retail_store_sales_details.csv

2.2. Herramientas y servicios

A. Lenguajes:

Lenguaje	Uso principal
SQL	Consultas y transformaciones en los datasets publicados dentro de Talend Cloud y conexión con fuentes tabulares.
Expresiones Talend Preparations /	Aplicación de reglas, filtros y transformaciones dentro de Talend Data Preparation y Data Stewardship (condiciones, validaciones, sugerencias, etc.).
DAX (Power BI)	Cálculos analíticos y de control de calidad de datos en los dashboards (por ejemplo, % de registros válidos, completitud, cumplimiento de reglas).

B. Herramientas

Herramienta / Servicio	Función	Función principal
------------------------	---------	-------------------

Talend Data Inventory (TDI)	Descubrimiento / Evaluación	Punto de partida del proyecto. Permite conectar y explorar fuentes de datos, evaluar su calidad automática, clasificar columnas y documentar datasets. Desde aquí se identifican los registros o campos con baja calidad.
Talend Data Preparation (TDP)	Limpieza / Transformación visual	Espacio de trabajo interactivo y visual donde se preparan y limpian los datos detectados en TDI. Se aplican transformaciones, reemplazos, uniones y reglas de negocio. Los registros con valores sospechosos o sugeridos son enviados a Data Stewardship.
Talend Data Stewardship (TDS)	Validación colaborativa / Gobierno de datos	Herramienta de resolución colaborativa que permite crear campañas de validación para que los <i>data stewards</i> revisen, corrijan o aprueben registros. Permite controlar el estado de cada tarea (To Do, Resolved, Rejected).
Talend Management Console (TMC)	Administración / Ejecución	Panel de gestión centralizada donde se administran usuarios, permisos y pipelines. Permite ejecutar tareas y flujos de integración desde el Remote Engine Cloud y hacer seguimiento a la ejecución de procesos.
Talend Open Studio (TOS)	Integración / Desarrollo técnico	Entorno de desarrollo ETL local para diseñar y probar procesos de integración, exportación o carga de datos. Permite conectar con fuentes externas (por ejemplo, CSV o bases de datos) y preparar datasets para subir a Talend Cloud.
Power BI Desktop	Visualización	Conexión en modo Import a las vistas de Athena. Cálculos DAX para ventas acumuladas, brechas y porcentajes de cumplimiento de presupuesto.

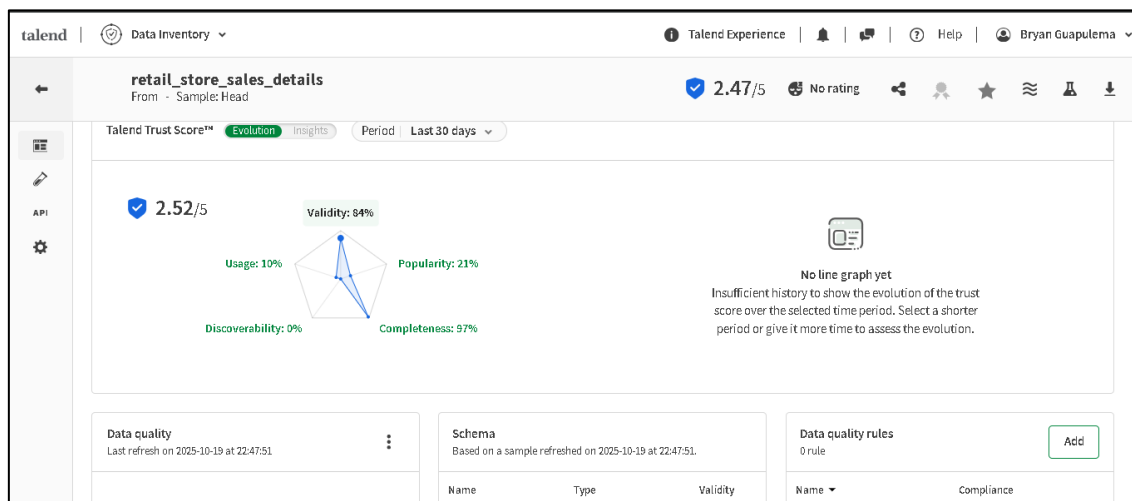
2.3. Arquitectura



3. Metodología y Desarrollo del Proyecto

3.1. Ingesta de datos: Talend Data Inventory (TDI)

Tras la búsqueda de la fuente de información con las características indicadas se hizo la carga del archivo mediante un pipeline creado en Talend Pipeline Designer. Así pues, con la fuente de datos disponible se hizo la carga de esta en Talend Data inventory

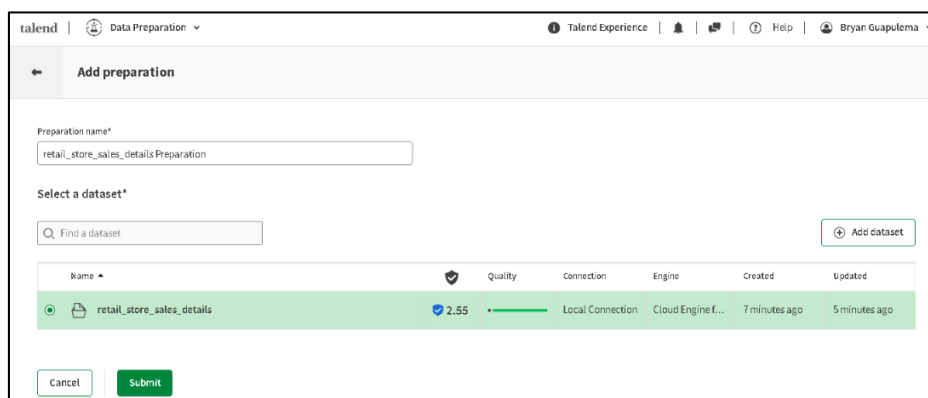


Como resultado de la carga se obtuvo un porcentaje de 2.47 puntos sobre 5 en lo referente a la calidad de los datos con notable presencia de valores inválidos, nulos y pobre documentación de semántica y metadata.

Adicionalmente se cargaron una fuente de mappings que contiene la información de los clientes y los productos por categoría, mismos que se unieron al dataset principal en Talend Data Preparation.

3.2. Limpieza de datos (ETL): Talend Data Preparation

Para el proceso de limpieza se crea una receta en Talend Data Preparation basada en la fuente de datos almacenada en Talend Data Inventory.



The screenshot shows the 'Add preparation' form in Talend Data Preparation. The 'Preparation name*' field contains 'retail_store_sales_details Preparation'. Below it is a 'Select a dataset*' section with a search bar and a table of datasets. The table has columns for Name, Quality, Connection, Engine, Created, and Updated. The dataset 'retail_store_sales_details' is highlighted with a quality score of 2.55. At the bottom are 'Cancel' and 'Submit' buttons.

Esta receta es la encargada de la limpieza de datos basada en reglas de calidad tanto de negocio como técnicas. Así pues, se definieron las siguientes reglas para la completitud, validación y limpieza del dataset.

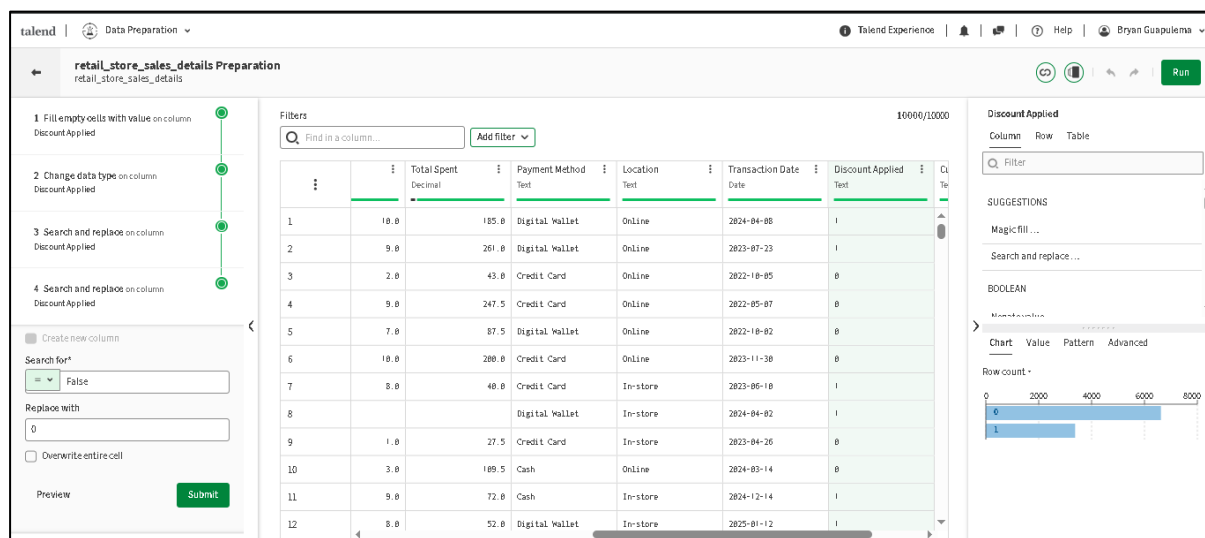
1. Regla 1: Normalizar Discount Applied a 0/1

Dado a que se encontraron múltiples valores para representar la presencia o ausencia de descuento en la compra como "True","False","No Discount","Yes","No" se normalizaron los valores que representan afirmaciones a 1 y las negaciones a 0 de modo que se tenga una bandera estandarizada

2. Regla 2: Corregir valores nulos en la columna Discount Applied

Para el tratamiento de valores nulos en la columna se interpretó el valor vacío como la ausencia de descuento, de modo que se imputaron los valores nulos con el valor de 0

Con estas dos reglas se unifica semántica y elimina 33% de nulos; hace medible el descuento en modelos y BI



The screenshot displays the Talend Data Preparation interface for the 'retail_store_sales_details' dataset. On the left, a list of rules is shown, including '1 Fill empty cells with value on column Discount Applied', '2 Change data type on column Discount Applied', '3 Search and replace on column Discount Applied', and '4 Search and replace on column Discount Applied'. The central table shows the following data:

	Total Spent (Decimal)	Payment Method (Text)	Location (Text)	Transaction Date (Date)	Discount Applied (Text)	
1	10.0	135.0	Digital Wallet	Online	2024-04-08	1
2	9.0	261.0	Digital Wallet	Online	2023-07-23	1
3	2.0	43.0	Credit Card	Online	2022-10-05	0
4	9.0	247.5	Credit Card	Online	2022-05-07	0
5	7.0	87.5	Digital Wallet	Online	2022-10-02	0
6	10.0	280.0	Credit Card	Online	2023-11-30	0
7	0.0	40.0	Credit Card	In-store	2023-06-10	1
8			Digital Wallet	In-store	2024-04-02	1
9	1.0	27.5	Credit Card	In-store	2023-04-26	0
10	3.0	109.5	Cash	Online	2024-03-14	0
11	9.0	72.0	Cash	In-store	2024-12-14	1
12	0.0	52.0	Digital Wallet	In-store	2025-01-12	1

The right-hand panel shows the 'Discount Applied' column configuration, including a 'Row count' bar chart with a value of 1 for the 'Discount Applied' column.

3. Regla 3: Completar Price Per Unit (PPU) con cálculos

Para la columna de precio unitario se noto la ausencia de 482 registros. Así pues para asegurar su completitud se calculo el precio unitario como la división entre el total gastado dividido para la cantidad para los casos en que esto sea posible

4. Regla 4: Completar Price Per Unit (PPU) con mediana por categoría

Para los casos de la columna de precio unitario que no se llenaron con la regla anterior se completaron mediante la imputación de precio por mediana de precios del producto basado en la categoría.

Estas reglas permitieron completar los valores faltantes y evitaron eliminar la porción de datos faltantes que podría representar un porcentaje significativo de la realidad de la empresa.

talend | Data Preparation

retail_store_sales_details Preparation

1 Fill empty cells with value on column Discount.Applied

2 Change data type on column Discount.Applied

3 Search and replace on column Discount.Applied

4 Search and replace on column Discount.Applied

5 Add, multiply, subtract or divide on column Total Spent

6 Fill empty cells with value on column Price Per Unit

7 Delete column on column Total Spent / Quantity

Filters

Find in a column... Add filter

	Price Per Unit Decimal	Quantity Decimal	Total Spent Decimal	Payment Method Text	Locat Text
1	18.5	10.0	185.0	Digital Wallet	On
2	29.0	9.0	261.0	Digital Wallet	On
3	21.5	2.0	43.0	Credit Card	On
4	27.5	9.0	247.5	Credit Card	On
5	12.5	7.0	87.5	Digital Wallet	On
6	20	10.0	200.0	Credit Card	On
7	5.0	8.0	40.0	Credit Card	In
8	33.5			Digital Wallet	In
9	27.5	1.0	27.5	Credit Card	In
10	36.5	3.0	109.5	Cash	On

Price Per Unit

Column Row Table

Q Filter

SUGGESTIONS

Add, multiply, subtract or divide ...

Compare numbers ...

Remove fractional part ...

Chart Value Pattern Advanced

Count:	10000	MIN:	5
Distinct:	38	MAX:	41
Duplicate:	9962	Mean:	23.4
Valid:	10000	Variance:	115.22
Empty:	0		
Invalid:	0		

5. Regla 5: Calcular Quantity si es posible

Con la columna de precio unitario completa se ejecutó el llenado de la cantidad basado en la división del total gastado sobre la cantidad. En caso de no ser posible el cálculo se podría optar con la cantidad promedio basado en la categoría del producto. El fin es eliminar los valores nulos por completo en esta columna

talend | Data Preparation

retail_store_sales_details Preparation

5 Add, multiply, subtract or divide on column Total Spent

6 Fill empty cells with value on column Price Per Unit

7 Delete column on column Total Spent / Quantity

8 Add, multiply, subtract or divide on column Total Spent

9 Fill empty cells with value on column Quantity

10 Delete column on column Total Spent / Price Per Unit

11 Fill empty cells from above on

Filters

Find in a column... Add filter

	Item Text	Price Per Unit Decimal	Quantity Decimal	Total Spent Decimal	P Text
919	Probiotic Drink 10L	32.0	8.0	256.0	
920	Instant Noodles Pa.	18.5	6.0	111.0	
921		6.5	1.0	6.5	
922	Chocolate Bar 100g	30.5	5.0	152.5	
923	Veal Outlets 250g	36.5	7.0	255.5	
924	lec... USB-C Cable 1m	6.5	10.0	65.0	
925	Jam Strawberry 300g	23.0	2.0	46.0	
926		8.0	2.0		
927	Laptop Stand Metal	30.5	8.0	244.0	

Quantity

Column Row Table

Q Filter

SUGGESTIONS

Add, multiply, subtract or divide ...

Compare numbers ...

Remove fractional part ...

Chart Value Pattern Advanced

Count:	10000	MIN:	1
Distinct:	10	MAX:	10
Duplicate:	9990	Mean:	5.54
Valid:	10000	Variance:	8.2
Empty:	0		
Invalid:	0		

6. Regla 6: Total Spent = Quantity * Price Per Unit

Con las columnas de cantidad y precio unitario se hizo una validación de la columna del gasto total multiplicando la cantidad y el precio unitario de modo que se detecten irregularidades que puedan ser mandadas a revisar al Data StewardShip.

talend | Data Preparation

retail_store_sales_details Preparation

retail_store_sales_details

Filters: Find in a column... Add filter

10000/10000

8 Add, multiply, subtract or divide on column Total Spent

9 Fill empty cells with value on column Quantity

10 Delete column on column Total Spent / Price Per Unit

11 Fill empty cells from above on column Quantity

12 Add, multiply, subtract or divide on column Price Per Unit

13 Fill empty cells with value on column Total Spent

14 Reorder columns on column

	Quantity Decimal	Price Per Unit x ... Decimal	Total Spent Decimal	Total Spent_eq... Boolean	Paymer Text
9991	3.0	33	33.0	true	Cred
9992	8.0	220	220.0	true	Cred
9993	5.0	100	100.0	true	Cash
9994	3.0	46.5	46.5	true	Cred
9995	10.0	335	335.0	true	Cred
9996	3.0	28.5	28.5	true	Digi
9997	6.0	237	237.0	true	Digi
9998	9.0	180	180.0	true	Digi
9999	9.0	288	288.0	true	Cash
10000	10.0	185	185.0	true	Digi

Total Spent_eq, Price Per Unit x Quantity?

Column Row Table

Filter

BOOLEAN

Negate value ...

COLUMNS

Chart Value Pattern Advanced

Row count

0 2500 5000 7500 10000

true

7. Regla 7: Completar nulos en Total Spent

Tras la validación se completaron los valores nulos con el resultado de la multiplicación.

talend | Data Preparation

retail_store_sales_details Preparation

retail_store_sales_details

Filters: Find in a column... Add filter

10000/10000

5 Add, multiply, subtract or divide on column Total Spent

6 Fill empty cells with value on column Price Per Unit

7 Delete column on column Total Spent / Quantity

8 Add, multiply, subtract or divide on column Total Spent

9 Fill empty cells with value on column Quantity

10 Delete column on column Total Spent / Price Per Unit

11 Fill empty cells from above on column Quantity

	Price Per Unit Decimal	Price Per Unit x ... Decimal	Quantity Decimal	Total Spent Decimal	Pa Text
919	10.0	32.0	256	8.0	256.0
920	18.5	111	6.0	111.0	
921	6.5	6.5	1.0	6.5	
922	30.5	152.5	5.0	152.5	
923	36.5	255.5	7.0	255.5	
924	6.5	65	10.0	65.0	
925	23.0	46	2.0	46.0	
926	8.0	16	2.0	16	
927	30.5	244	8.0	244.0	

Total Spent

Column Row Table

Filter

SUGGESTIONS

Add, multiply, subtract or divide ...

Compare numbers ...

Remove fractional part ...

Chart Value Pattern Advanced

Count: 10000 MIN: 5

Distinct: 365 MAX: 410

Duplicate: 9635 Mean: 130.03

Valid: 10000 Variance: 8995.22

Empty: 0

Invalid: 0

8. Regla 8: Normalización de Payment Method

Se establecieron todos los valores posibles dentro de 3 grupos "Cash", "Credit Card" y "Digital Wallet". Esto incluyó la conversión de todas las variantes de estos valores con mayúscula, minúsculas, entre otros.

retail_store_sales_details Preparation
retail_store_sales_details

10000/10000

1 Fill empty cells with value on column Discount Applied

2 Change data type on column Discount Applied

3 Search and replace on column Discount Applied

4 Search and replace on column Discount Applied

5 Add, multiply, subtract or divide on column Total Spent

6 Fill empty cells with value on column Price Per Unit

7 Delete column on column Total Spent / Quantity

Filters

Find in a column... Add filter

	Price Per Unit Decimal	Quantity Decimal	Total Spent Decimal	Payment Method Text	Location Text
1	18.5	10.0	185.0	Digital Wallet	Online
2	29.0	9.0	261.0	Digital Wallet	Online
3	21.5	2.0	43.0	Credit Card	Online
4	27.5	9.0	247.5	Credit Card	Online
5	12.5	7.0	87.5	Digital Wallet	Online
6	20	10.0	200.0	Credit Card	Online
7	5.0	8.0	40.0	Credit Card	In-store
8	33.5			Digital Wallet	In-store
9	27.5	1.0	27.5	Credit Card	In-store
10	36.5	3.0	109.5	Cash	Online

Price Per Unit

Column Row Table

Filter

SUGGESTIONS

Add, multiply, subtract or divide ...

Compare numbers ...

Remove fractional part ...

Chart Value Pattern Advanced

Count: 10000 MIN: 5
Distinct: 38 MAX: 41
Duplicate: 9962 Mean: 23.4
Valid: 10000 Variance: 115.22
Empty: 0
Invalid: 0

9. Regla 9: Estandarizar Transaction Date y validar no-futuro

Se convirtió la fecha al formato de fecha ISO con el formato “”yyyy-MM-dd”. Además se validó que la fecha sea coherente y sea menor que el día de ingesta. En caso de errores se dirige a Stewardship

retail_store_sales_details Preparation
retail_store_sales_details

10000/10000

14 Compare numbers on column Total Spent

15 Delete column on column Price Per Unit x Quantity

16 Rename column on column Total Spent_eq_Price Per Unit x Qua...

17 Change date format on column Transaction Date

18 Deduplicate rows with identical values on column Transaction ID

Matching criterion*
Exact value

Preview Submit

Filters

Find in a column... Add filter

	Location Text	Transaction Date Date	Discount Applied Text	Customer Text
1943	Online	2022-02-05	1	Noah Thompson
1944	In-store	2023-03-03	0	John Smith
1945	In-store	2022-03-27	0	Elijah White
1946	In-store	2022-03-09	1	James Wilson
1947	In-store	2024-01-22	0	Benjamin Kim
1948	Online	2023-10-12	1	Mia Taylor
1949	Online	2024-07-28	0	Daniel Garcia
1950	In-store	2024-01-13	0	John Smith
1951	Online	2022-01-19	1	John Smith

Transaction Date

Column Row Table

Filter

SUGGESTIONS

Magic fill ...

Extract date parts ...

Change date format ...

Chart Value Pattern Advanced

Row count

2021-01-01 2021-02-01 2021-03-01 2021-04-01 2021-05-01 2021-06-01 2021-07-01 2021-08-01 2021-09-01 2021-10-01 2021-11-01 2021-12-01

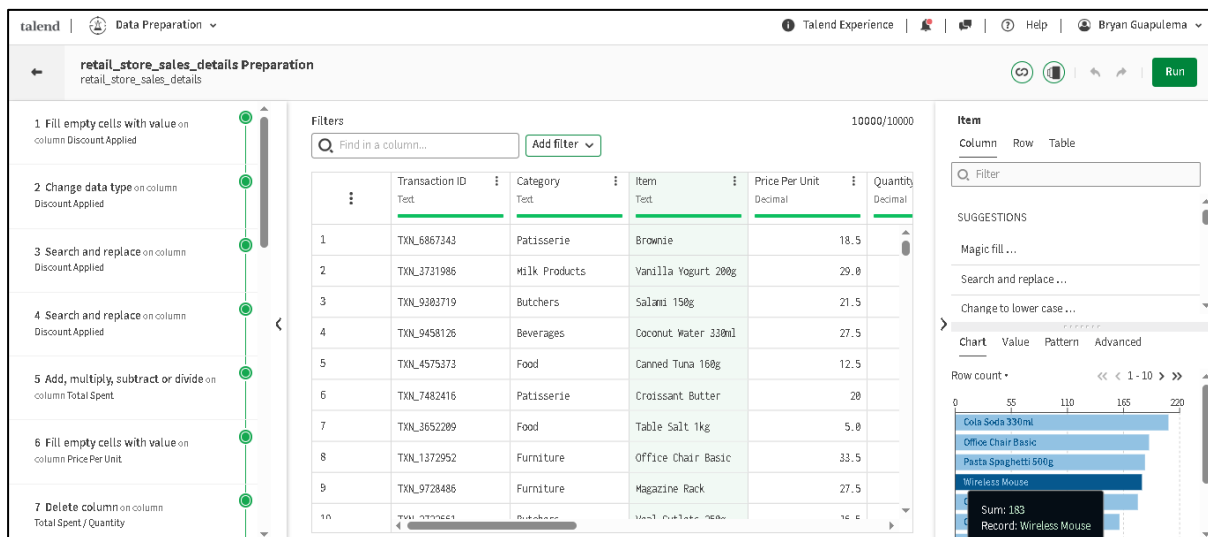
22-01-01 2023-01-01 2024-01-01 2025-01-01

2021-12-31 2025-03-31

Min 2021-12-31 Max 2025-03-31

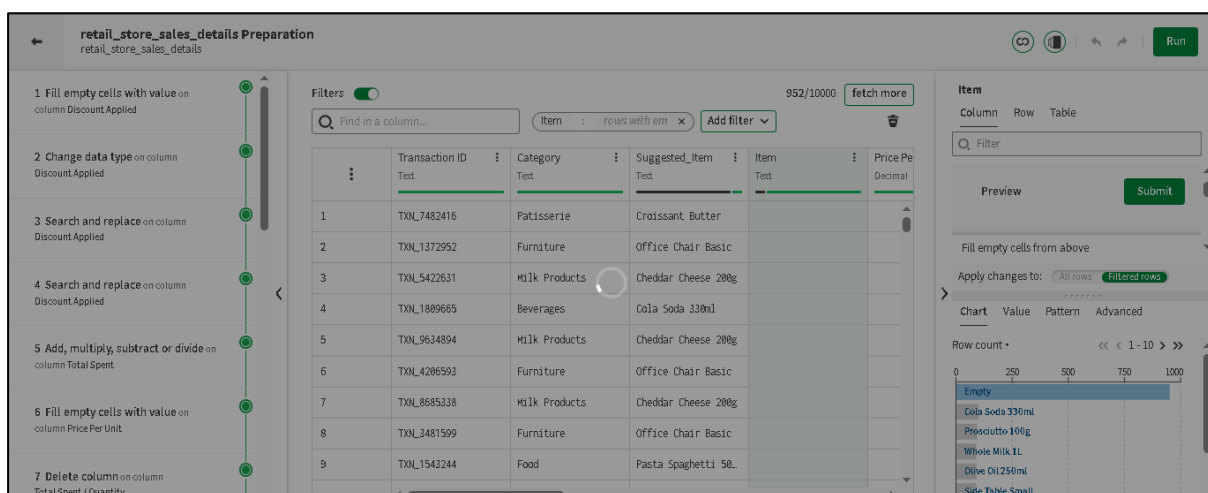
10. Regla 10. Transaction ID obligatorio y único

Adicionalmente se aplicó una deduplicación de los datos por Transaction ID a priori de eliminar duplicados exactos de fila.



11. Regla 11: Ítem no puede estar vacío

Finalmente se hizo el llenado de la columna Item. Pero esta no fue llenada de manera calculada pues corresponde a una decisión de negocio pues si se agrega al producto más vendido puede crearse outliers y si se agrega al de menos ventas puede crearse información falsa. Así pues, esta decisión el corresponde al custodio de la información: el Data Steward. Así pues, no se lleno directamente sino que se creó una columna Suggested Item con el producto más vendido para que sea validado e insertado o se cambie por el producto que se decida



12. Regla 12: Tipos coherentes

Finalmente se eliminaron columnas auxiliares y se verificó que cada una de las columnas tenga un tipo de dato coherente con su contenido

Nota: se creó una columna bandera que registró las columnas que necesitaban revisión por ser excepciones a la regla

3.3. Verificación de casos (ETL): Talend Data Stewardship

Para la revisión por parte del Data Steward fue necesario primero crear el modelo de datos con que la información que va a revisar cuenta:

talend | Data Stewardship | Bryan Guapulema

Retail_ItemReview

Description:
Modelo de datos para registros que necesitan revisión

Attributes Rules

+ Add attribute

Transaction_ID Text

Category Text

Suggested_Item Text

Cancel Save data model

Una vez creado el modelo se crea una campaña basada en el modelo que servirá para la conexión y asignación de tareas con los datos por validar el TDP.

talend | Data Stewardship | Bryan Guapulema

Store_Retail_Items_Review

DESCRIPTION: (optional)

TYPE: Resolution

ENABLE TASK RESOLUTION DELAY

Campaign owners

+ Add a campaign owner Bryan Guapulema

Cancel Edit campaign

talend | Data Stewardship | Bryan Guapulema

2/ Roles

Stewards ☒

+ Add a steward Bryan Guapulema

Add a role

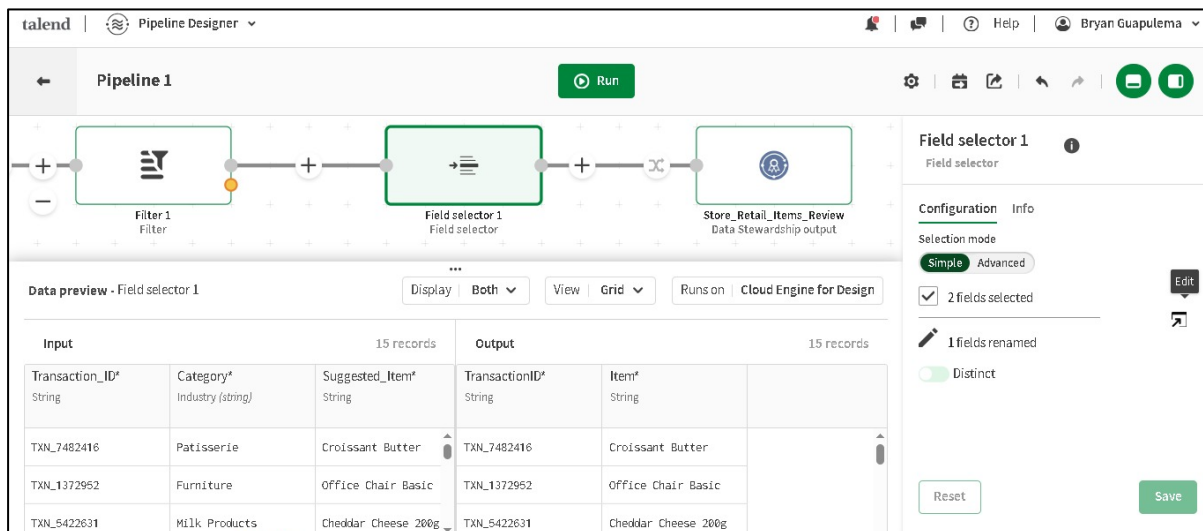
3/ Data model

Retail_ItemReview2

ALL ATTRIBUTES			
TRANSACTIONID	Text		
Stewards			

Cancel Edit campaign

Para poder enviar los registros de TDP a TDS se usó Talend Pipeline Designer. Una herramienta incluida en la suite de Talend y que se encuentra en el plan gratuito de esta. Así pues, se creó un pipeline que envió los registros que necesitan revisión a la campaña creada a priori



Pipeline 1

Run

Field selector 1

Store_Retail_Items_Review

Data preview - Field selector 1

Display Both View Grid Runs on Cloud Engine for Design

Input			Output		
Transaction_ID*	Category*	Suggested_Item*	TransactionID*	Item*	
String	Industry (string)	String	String	String	
TXN_7482416	Patisserie	Croissant Butter	TXN_7482416	Croissant Butter	
TXN_1372952	Furniture	Office Chair Basic	TXN_1372952	Office Chair Basic	
TXN_5422631	Milk Products	Cheddar Cheese 200g	TXN_5422631	Cheddar Cheese 200g	

Field selector 1

Configuration

Selection mode

Simple Advanced

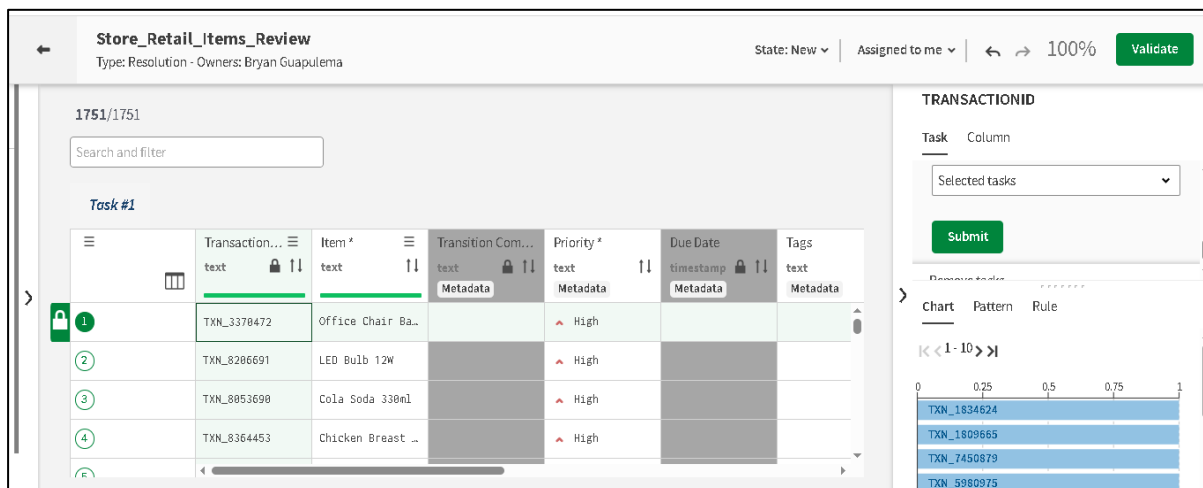
2 fields selected

1 fields renamed

Distinct

Reset Save

Tras ello se crean tareas con los registros que pueden ser revisados y validados por el usuario asignado. Así pues se revisaron y validaron dichos registros dejándolos con el estado resolved



Store_Retail_Items_Review

Type: Resolution - Owners: Bryan Guapulema

State: New Assigned to me 100% Validate

1751/1751

Search and filter

Task #1

	Transaction...	Item *	Transition Com...	Priority *	Due Date	Tags
	text	text	text	text	timestamp	text
	Metadata	Metadata	Metadata	Metadata	Metadata	Metadata
1	TXN_3370472	Office Chair Ba...		High		
2	TXN_8206691	LED Bulb 12W		High		
3	TXN_8053690	Cola Soda 330mL		High		
4	TXN_8364453	Chicken Breast ...		High		

TRANSACTIONID

Task Column

Selected tasks

Submit

Chart Pattern Rule

Chart

TXN_1834624

TXN_1809665

TXN_7450879

TXN_5980975

Store_Retail_Items_Review
Type: Resolution - Owners: Bryan Guapulema

State: Resolved

1751/1751

Search and filter

Task #1

	Transaction...	Item *	Transition Com...	Priority *	Due Date	Tags
	text	text	text	text	timestamp	text
	Metadata	Metadata	Metadata	Metadata	Metadata	Metadata
1	TXN_3370472	Office Chair Ba...		High		
2	TXN_8206691	LED Bulb 12W		High		
3	TXN_8053690	Cola Soda 330ml		High		
4	TXN_8364453	Chicken Breast ...		High		

TRANSACTIONID

Task Column

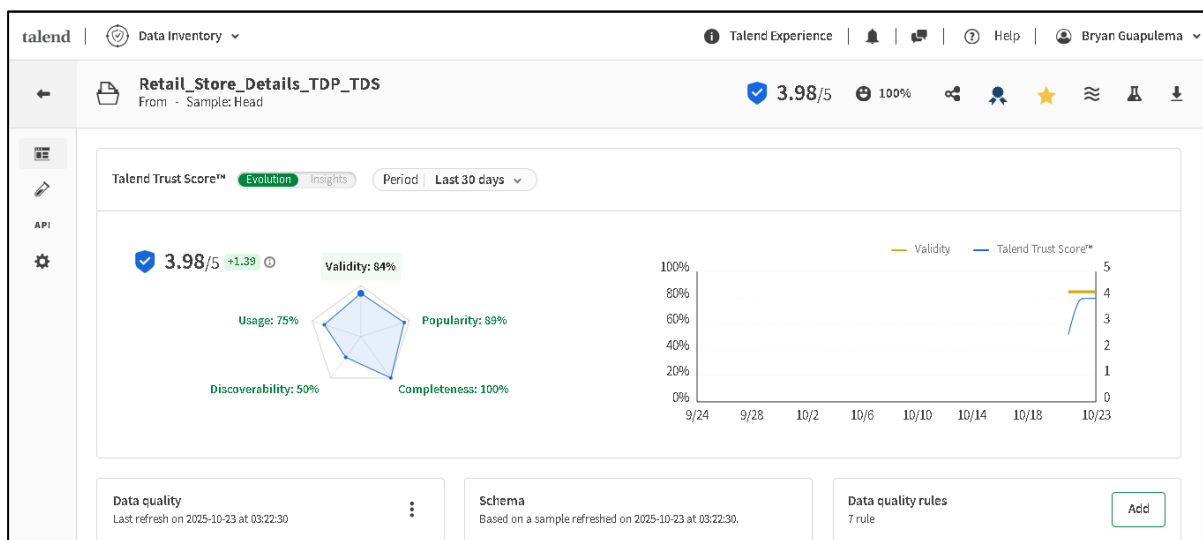
Find a function

No functions for the current selection

Chart Pattern Rule

TXN_1834624
TXN_1809665
TXN_7450879
TXN_5980975

Finalmente se unieron los datos validos al inicio con los validados con TDS en un solo dataset y se exportaron a un dataset almacenado en TDI cuyas métricas aumentaron en medida

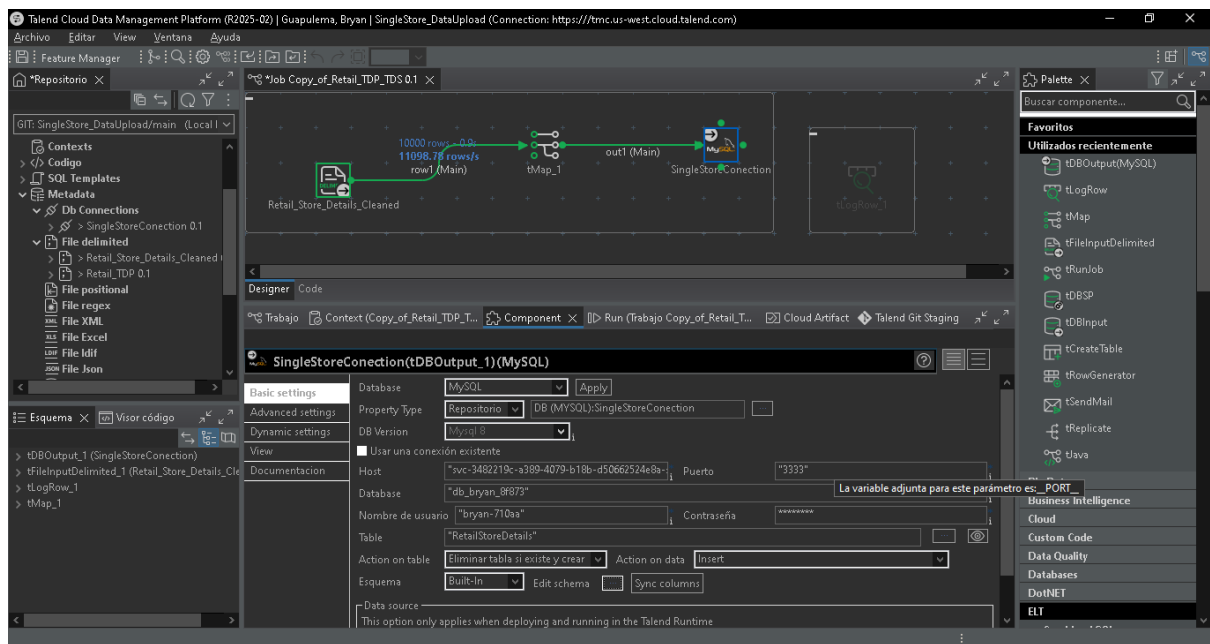


Nótese que el criterio de descubrimiento se ve afectado por el uso que se le da al dataset en la suite. Al ser el dataset final solo se usa para el volcado a single store mediante un job.

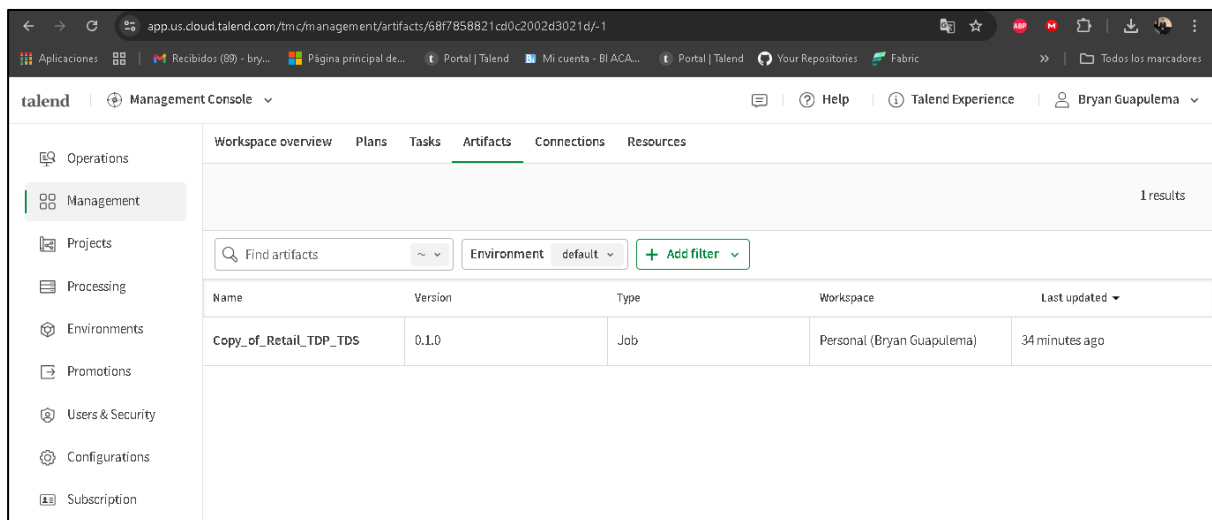
3.4. Volcado de datos (Single Store)

Para el volcado de datos se utilizó la consola de Talend: TMC junto con Talend Open Studio (la última de las herramientas de la suite de Talend. Así pues se utilizó la consola para conectar la nube de Talend con el estudio local mediante la creación de un token en TMC.

Así pues, se creó un job en Talend que consuma el dataset final generado, lo mapee asegurando el orden, semántica y tipo de datos y se publicó la tabla resultante a una base de datos en SinlgeStore previamente creada.



Finalmente se publicó el job a la nube de cloud como un artefacto que se puede ejecutar de manera automatizada



Finalmente se ejecutó el job desde Talend Management Console y se verificó el volcado de datos en SingleStore desde DBeaver y en la propia plataforma de SingleStore

SingleStore Customer Portal

portal.singlestore.com/organizations/ef7d6c50-67b5-4a0b-996a-2defea739149/clusters/680819b8-756c-48ba-a148-d94035c8230b/databases...

SingleStore.. Search Ctrl + k Free Trial Credits Used 0 / 167 CR Upgrade bryan guapulema BRYAN GUAPULEMA'S ORGA...

starter-workspace > db_bryan_8f873

Tables 2 Views 0 Procedures 1 Functions 0 Aggregates 0 Pipelines 1

Name	Storage	Row Count	Memory Usage	Index Disk Usage...	Data Disk Usage	Compre
RetailStoreDetails	Columnstore	12.58K	7 MB	0 B	0 B	
uk_price_paid	Columnstore	114.46K	0 B	448 KB	8 MB	

19:26 22/10/2025

SingleStore Customer Portal

portal.singlestore.com/organizations/ef7d6c50-67b5-4a0b-996a-2defea739149/clusters/680819b8-756c-48ba-a148-d94035c8230b/databases...

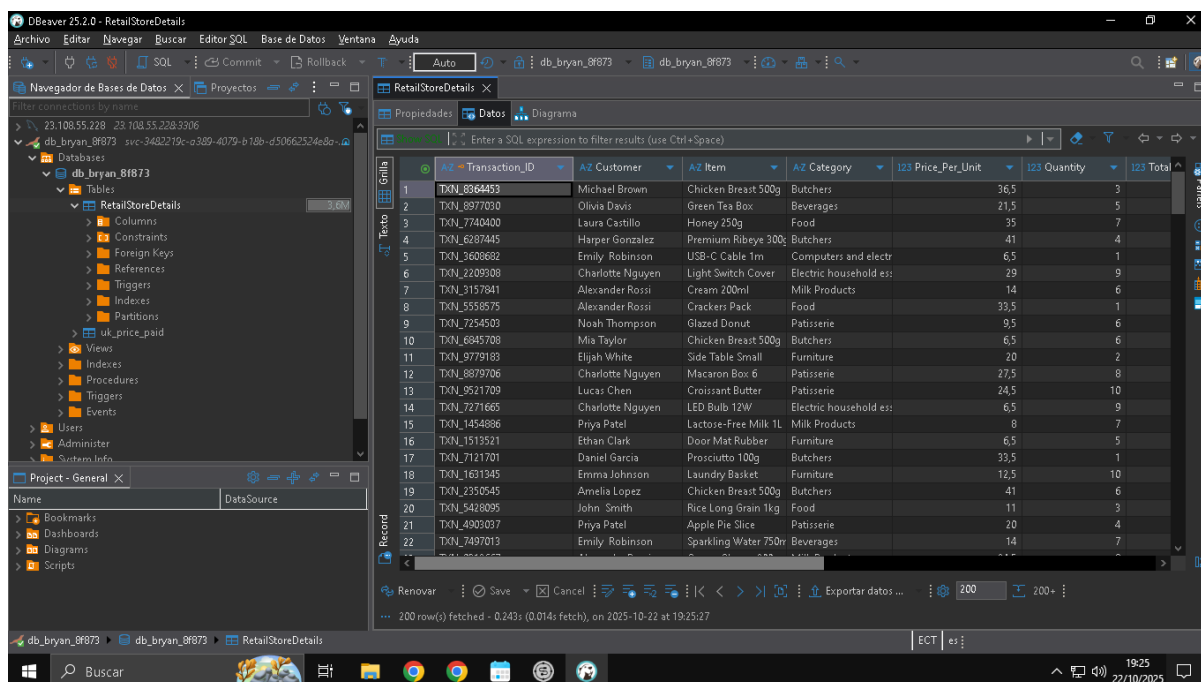
SingleStore.. Search Ctrl + k Free Trial Credits Used 0 / 167 CR Upgrade bryan guapulema BRYAN GUAPULEMA'S ORGA...

starter-workspace > db_bryan_8f873 > |||| RetailStoreDetails

Columns 11 Indexes 2 Sample Data SQL

Name	Data Type	Computed	Nullable	Default	Memory Usage	Disk Usage	Compression Ratio
A Transaction_ID	varchar(11)	No	No	—	0 B	0 B	—
A Customer	varchar(30)	No	Yes	—	0 B	0 B	—
A Item	varchar(100)	No	Yes	—	0 B	0 B	—
A Category	varchar(50)	No	Yes	—	0 B	0 B	—
# Price_Per_Unit	float(4,2)	No	Yes	—	0 B	0 B	—
# Quantity	float(4,3)	No	Yes	—	0 B	0 B	—
# Total_Spent	float(10,3)	No	Yes	—	0 B	0 B	—

19:26 22/10/2025



	AZ Transaction_ID	AZ Customer	AZ Item	AZ Category	123 Price_Per_Unit	123 Quantity	123 Total
1	TXN_8364453	Michael Brown	Chicken Breast 500g	Butchers	36,5	3	
2	TXN_8977030	Olivia Davis	Green Tea Box	Beverages	21,5	5	
3	TXN_7740400	Laura Castillo	Honey 250g	Food	35	7	
4	TXN_6287445	Harper Gonzalez	Premium Ribeye 300g	Butchers	41	4	
5	TXN_3608682	Emily Robinson	USB-C Cable 1m	Computers and electr	6,5	1	
6	TXN_2209308	Charlotte Nguyen	Light Switch Cover	Electric household es	29	9	
7	TXN_3157841	Alexander Rossi	Cream 200ml	Milk Products	14	6	
8	TXN_5558575	Noah Thompson	Crackers Pack	Food	33,5	1	
9	TXN_7254503	Mia Taylor	Glazed Donut	Patisserie	9,5	6	
10	TXN_6945708	Elijah White	Chicken Breast 500g	Butchers	6,5	6	
11	TXN_9779183	Charlotte Nguyen	Side Table Small	Furniture	20	2	
12	TXN_8879706	Lucas Chen	Macaron Box 6	Patisserie	27,5	8	
13	TXN_9521709	Charlotte Nguyen	Croissant Butter	Patisserie	24,5	10	
14	TXN_7271665	Priya Patel	LED Bulb 12W	Electric household es	6,5	9	
15	TXN_1454886	Ethan Clark	Lactose-Free Milk 1L	Milk Products	8	7	
16	TXN_1519521	Daniel Garcia	Door Mat Rubber	Furniture	6,5	5	
17	TXN_7121701	Emma Johnson	Prosciutto 100g	Butchers	33,5	1	
18	TXN_1631345	Amelia Lopez	Laundry Basket	Furniture	12,5	10	
19	TXN_2350545	John Smith	Chicken Breast 500g	Butchers	41	6	
20	TXN_5428099	Priya Patel	Rice Long Grain 1kg	Food	11	3	
21	TXN_4905037	Emily Robinson	Apple Pie Slice	Patisserie	20	4	
22	TXN_7497013		Sparkling Water 750ml	Beverages	14	7	

4. Visualización

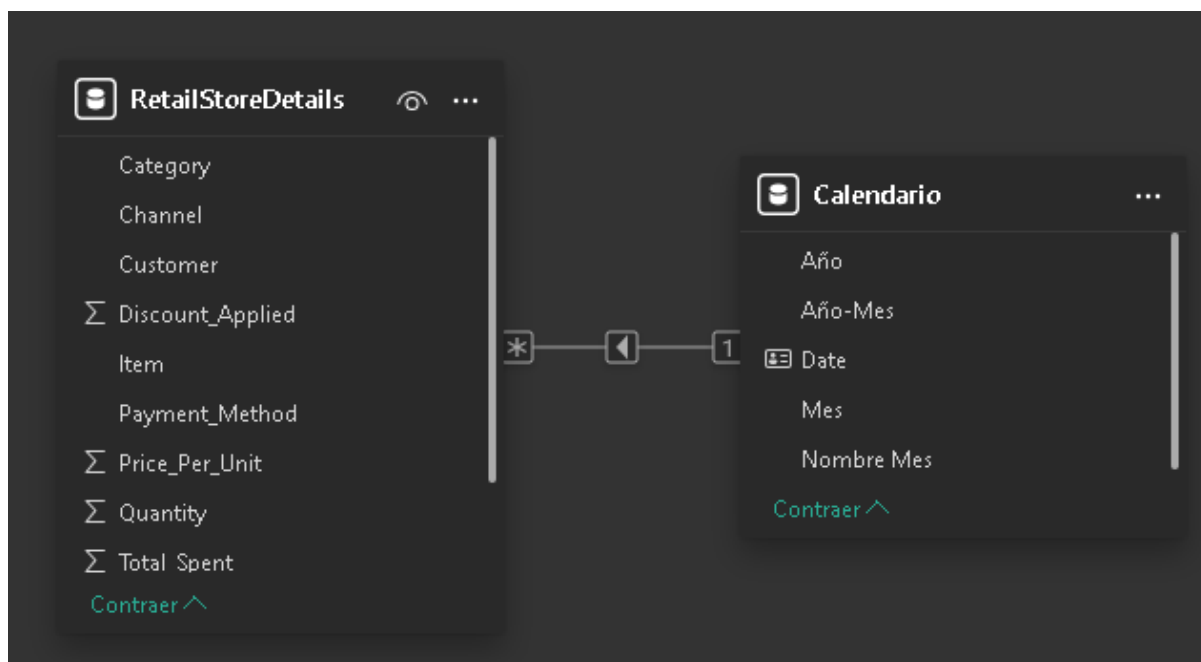
4.1. Conexión a Power BI

Para conectar Single Store con Power BI se usó el driver de la base de datos incluida en el software de visualización. Para iniciar sesión se utilizaron las credenciales proporcionadas en la plataforma y se configuró el certificado SSL.

Se eligió el modo Import, que permite mayor rendimiento y refrescos del modelo cuando se actualiza con los nuevos meses cargados por la lambda incremental.

4.2. Modelo de datos

Tras conectar la tabla de SingleStore Power BI Desktop, se construyó el esquema presentado enseguida (nótese que se está usando una big table con toda la información de la empresa):



4.3. Medidas DAX

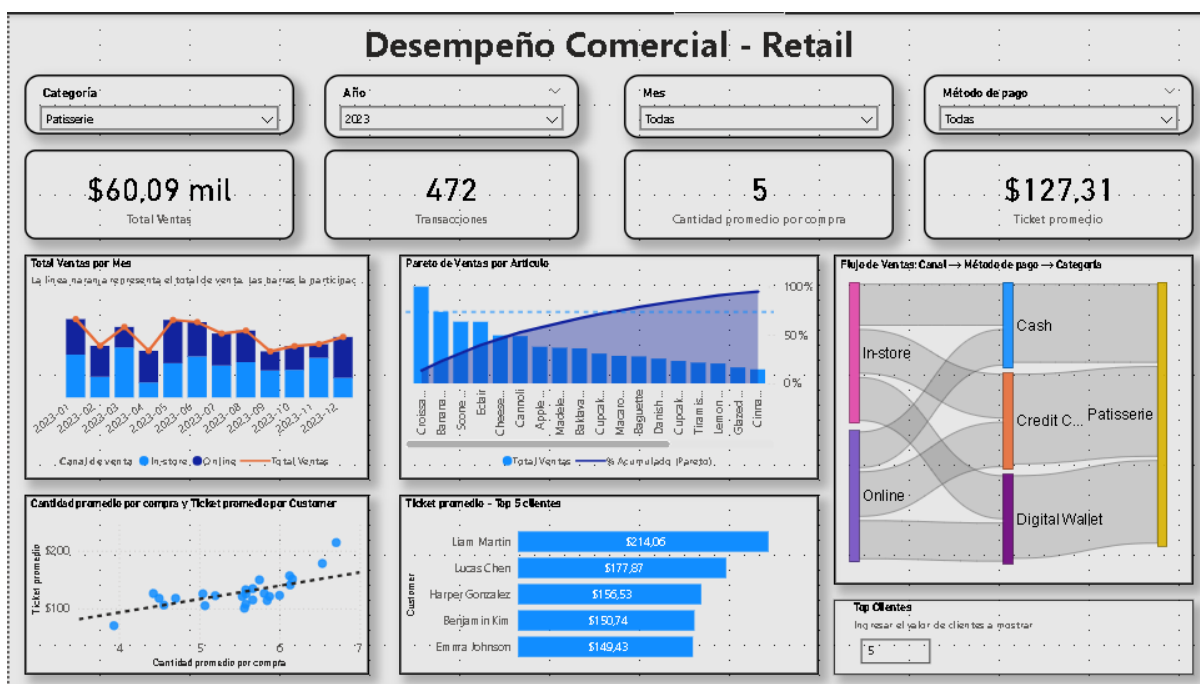
Se definió la siguiente lista de medidas dentro de una tabla contenedora de medidas:

Medida	Fórmula
% Acumulado (Pareto)	% Acumulado (Pareto) = DIVIDE([Ventas acumuladas (Pareto)], [Ventas total visible (Pareto)])
% Participación canal	% Participación canal = DIVIDE([Ventas (contexto mes)], [Ventas total mes (todos los canales)])
Cantidad	Cantidad = SUM(RetailStoreDetails[Quantity])
Cantidad promedio por compra	Cantidad promedio por compra = DIVIDE([Cantidad], [Transacciones])
Cantidad promedio por compra	Cientes únicos = DISTINCTCOUNT(RetailStoreDetails[Customer])
Mostrar cliente (Top N)	Mostrar cliente (Top N) = IF([Ranking cliente (Ticket Promedio)] <= SELECTEDVALUE('TOP'[TOP]), 1, 0)
Ticket promedio	Ticket promedio = DIVIDE([Total Ventas], [Transacciones])
Título gráfico Top N	Título gráfico Top N = "Ticket promedio - Top " & SELECTEDVALUE('TOP'[TOP]) & " clientes"
Total Ventas	Total Ventas = (SUM(RetailStoreDetails[Total_Spent]))
Transacciones	Transacciones = DISTINCTCOUNT(RetailStoreDetails[Transaction_ID])
Valor Sankey =	Valor Sankey = VAR origen = SELECTEDVALUE('Enlaces Sankey'[Origen]) VAR destino = SELECTEDVALUE('Enlaces Sankey'[Destino]) VAR isCanal_Metodo = origen IN VALUES(RetailStoreDetails[Channel]) && destino IN VALUES(RetailStoreDetails[Payment_Method]) VAR isMetodo_Categoria = origen IN VALUES(RetailStoreDetails[Payment_Method]) && destino IN VALUES(RetailStoreDetails[Category]) RETURN SWITCH(TRUE(), isCanal_Metodo, CALCULATE([Total Ventas], RetailStoreDetails[Channel] = origen, RetailStoreDetails[Payment_Method] = destino))

	<pre>), isMetodo_Categoria, CALCULATE([Total Ventas], RetailStoreDetails[Payment_Method] = origen, RetailStoreDetails[Category] = destino), BLANK()) </pre>
Ventas (contexto mes)	Ventas (contexto mes) = [Total Ventas]
Ventas acumuladas (Pareto)	<pre> Ventas acumuladas (Pareto) = VAR r = [Ranking Item] VAR Tabla = ADDCOLUMNS(ALLSELECTED(RetailStoreDetails[Item]), "VentasItem", CALCULATE([Total Ventas])) VAR TopR = TOPN(r, Tabla, [VentasItem], DESC) RETURN SUMX(TopR, [VentasItem]) </pre>

5. Visualización

5.1. Dashboard



5.2. Historia construida a partir del dashboard

Este proyecto contempla el análisis de una tienda de retail especializada en productos de consumo, con información transaccional comprendida entre los años 2022 y 2025. El objetivo principal fue evaluar el desempeño comercial durante este periodo, identificando los factores que impulsan las ventas, los productos más rentables y los clientes de mayor valor para el negocio.

Una visión general del desempeño muestra que en este periodo, la tienda ha registrado ventas totales por 60 mil dólares, distribuidas en 472 transacciones, con un ticket promedio

de 127 dólares y una cantidad promedio de cinco unidades por compra, lo que refleja una estructura de ventas saludable y sostenida.

En la parte superior del dashboard se pueden seleccionar los filtros por categoría, año, mes y método de pago, lo que permite analizar de manera dinámica el comportamiento comercial bajo distintos escenarios.

Ahora bien, al observar la evolución de ventas por mes, notamos un comportamiento estacional pero con una tendencia general de crecimiento, especialmente en el canal Online, que ha ganado terreno frente a la tienda física. Este aumento del canal digital demuestra que los clientes están adoptando cada vez más las compras en línea, lo cual abre oportunidades para fortalecer las estrategias de marketing digital, descuentos web y experiencia omnicanal.

Sin embargo, se observa que ciertos meses —como marzo y octubre— presentan leves caídas en el volumen total de ventas, lo que podría relacionarse con ciclos de consumo y con la falta de campañas promocionales específicas durante esos periodos.

Pasando al análisis de productos, el gráfico de Pareto muestra que un grupo reducido de artículos concentra la mayoría de los ingresos. En concreto, se confirma la regla 80/20: el 20% de los productos genera aproximadamente el 80% de las ventas. Esto significa que la rentabilidad del negocio está altamente concentrada en pocos artículos, por lo que las decisiones de reposición, marketing y descuentos deben centrarse en ese núcleo de productos estrella. Al mismo tiempo, los productos del tramo inferior del Pareto deberían ser revisados para determinar si conviene mantenerlos, liquidarlos o reemplazarlos por variantes más rentables.

Si analizamos el flujo de ventas desde el canal hasta la categoría, podemos ver cómo se comporta el cliente en su proceso de compra. En el Sankey diagram se observa que el canal Online representa el flujo más grande, con predominio de pagos con tarjeta de crédito, especialmente en la categoría Patisserie. En cambio, las ventas In-store muestran una mayor proporción de pagos en efectivo. Este análisis es clave porque revela las preferencias de los consumidores según el canal: mientras los clientes online valoran comodidad y medios de pago digitales, los clientes presenciales prefieren compras inmediatas y pagos tradicionales.

En la parte inferior izquierda, el gráfico de dispersión relaciona la cantidad promedio comprada con el ticket promedio por cliente, lo cual permite identificar los distintos tipos de comportamiento.

En la zona superior derecha del gráfico se ubican los clientes VIP, aquellos que compran con alta frecuencia y alto ticket promedio; en la zona inferior derecha, los premium ocasionales, que gastan mucho pero compran con menor frecuencia; y en la parte superior izquierda se ubican los clientes frecuentes de bajo ticket, que podrían ser fidelizados con promociones o descuentos.

Este análisis es fundamental para segmentar a los clientes y establecer estrategias diferenciadas: por ejemplo, programas de fidelización para los VIP, combos para los frecuentes de bajo ticket y campañas de recompra para los premium ocasionales.

En la parte inferior derecha se observa el Top de clientes con mayor ticket promedio, el cual es totalmente dinámico gracias al parámetro Top N. En este caso, los clientes Liam Martin y Lucas Chen encabezan el ranking con tickets promedio superiores a los 170 dólares. Este

tipo de visualización permite concentrar los esfuerzos de fidelización y recompra en los clientes que más contribuyen a la rentabilidad general del negocio.

Finalmente, si observamos el conjunto de KPIs y tendencias, podemos concluir que el negocio mantiene una estructura comercial saludable, con crecimiento en el canal online, altas oportunidades de segmentación de clientes, y una cartera de productos concentrada pero altamente rentable.

En términos estratégicos, las recomendaciones principales son: primero, fortalecer el canal online mediante campañas dirigidas y beneficios de pago digital; segundo, optimizar el catálogo de productos, priorizando los del tramo superior del Pareto y revaluando los de baja contribución; y tercero, desarrollar programas de fidelización segmentados por tipo de cliente para aumentar el valor promedio de cada compra.

6. Recursos

Documentación completa: <https://github.com/BryanGuapulema/Hackaton-3>