

IS-601, BASE DE DATOS II

DATAWAREHOUSE PROYECTO



UNAH
UNIVERSIDAD NACIONAL
AUTÓNOMA DE HONDURAS

Estudiantes:

Gabriel Alexander Barrientos	20181000058
Andy Yankarlo Avelar	20191033226
Carlos Eduardo Alcerro	20181900319

Catedrático: Emilson Omar Acosta Giron

Sección: 1200

Fecha: 12 de Agosto del 2022

Índice

Índice	2
Introducción	3
Objetivo General	4
Objetivo Específico	5
Pentaho	6
Pentaho Data Integration	6
Creación de ETL con Pentaho Data Integration	7
Creación Del ETL	9
Configurando Las Transformaciones	10
Crear Conexión A Una Base De Datos	11
Obtener Y Procesar Registros Para Tablas De Dimensiones	13
Obtener Y Procesar Registros Para Tablas De Hechos	18
Tableau	21
Conexión Con MSSQL	21
Selección De Tablas	22
Reportes Generados En Tableau	25
Base De Datos OLTP	34
Base De Datos OLAP	35
Preguntas Del Negocio Utilizadas	36
Explicación De Métrica Utilizada	37
Conclusiones	38
Recomendaciones	39
Bibliografía	40

Introducción

Este informe mostrará cómo implementar la generación de ETL en Pentaho Spoon y también cómo establecer la conexión al servidor MSSQL utilizando diferentes gráficos de generación de informes con tableau, para realizar un análisis que nos brindara información detallada sobre el estado actual de la empresa para una mayor tomas de decisiones. Abordaremos esté tema desde la creación del modelo OLAP a partir de su base de datos OLTP, definiendo las preguntas del negocio junto a las métricas que requerimos analizar.

Objetivo General

La finalidad del trabajo se expresa en “desarrollar un proyecto de datawarehouse para una empresa real o ficticia”. Para esto, se debe tener una base de datos OLTP con toda la información almacenada y habilitar la generación de reportes mediante el uso de inteligencia de negocios.

Objetivo Específico

- Obtención de una base de datos ya sea real o ficticia.
- Crear un datamart y su respectivo modelo en estrella de la base de datos.
- Investigar e implementar ETLs usando pentaho.
- Investigar e implementar la reportería con tablou.
- Crear un reporte del total de productos vendidos y mostrar diferentes formatos del mismo.

Pentaho

Pentaho BI Suite es la suite de inteligencia de negocios más popular en el mundo que se utiliza para informes, análisis, tableros de control, minería de datos, flujo de trabajo y capacidades ETL o Export Transform Load. Existen numerosos proveedores de servicios de desarrollo de software que son expertos en análisis de negocios y desarrollo de integración de Pentaho y atienden todas las pilas que ofrecen soluciones de vanguardia a clientes estimados.

Pentaho Data Integration

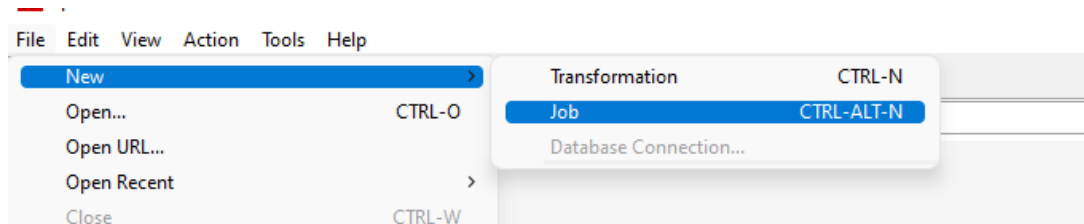
Es una herramienta creada en JAVA para crear procesos ETL de forma rápida y sencilla, en cualquier entorno.

Interfaz Drag & Drop, mediante la cual también se define el flujo de datos mediante conexión de "pasos" o "steps" con flechas.

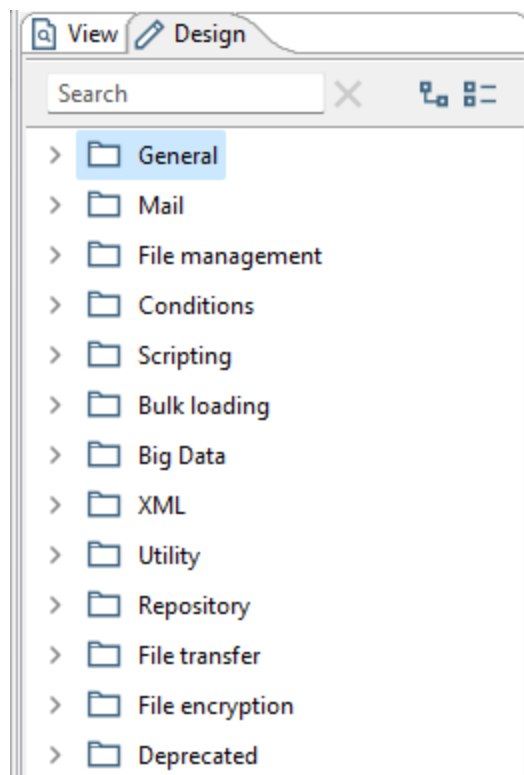
Es compatible con multitud de orígenes de datos y aplicaciones de terceros aplicadas al Big Data.

Creación de ETL con Pentaho Data Integration

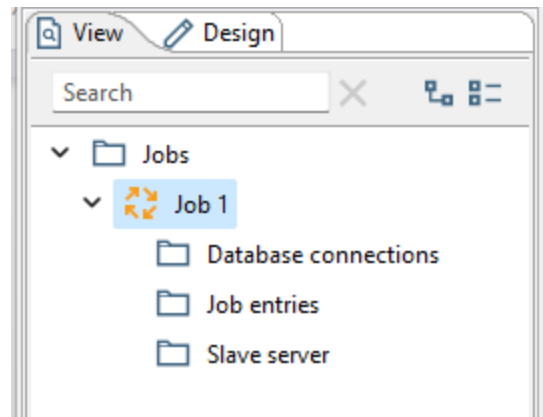
Dentro de Pentaho Data Integration, lo primero es crear un archivo de tipo trabajo (Job), el cual contendrá todas las entidades para que el ETL se ejecute correctamente.



Una vez creado en la parte izquierda de la interfaz nuestro trabajo (Job), las entidades en este caso el trabajo (Job) poseen dos pestañas una de vista (view) y diseño (design). En el diseño se presenta una lista de todos los componentes que se pueden utilizar en alguna entidad.



En la vista se puede ver los componentes que conforman una entidad. En el caso del trabajo creado anteriormente, dos son de los que se utilizarán, las conexiones a la o las bases de datos (Database connections) donde se establecen las conexiones a las bases de datos de origen y destino y las entradas del trabajo (Job entries) que serán los componentes que conformen el ETL.



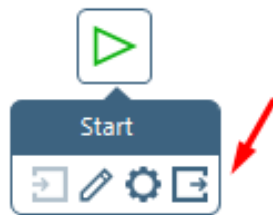
Para asignar un nombre a los trabajos, basta con dar doble clic en el elemento, esto nos despliega una ventana donde podemos asignar entre otras cosas, un nombre.

A screenshot of a configuration window titled 'Job'. It has tabs for 'Job', 'Parameters', 'Settings', and 'Log'. The 'Job' tab is active. It contains the following fields: 'Job name:' with the value 'BikeStores', 'Job filename:' with an empty text box, 'Description:' with an empty text box, and 'Extended description:' with a larger empty text area.

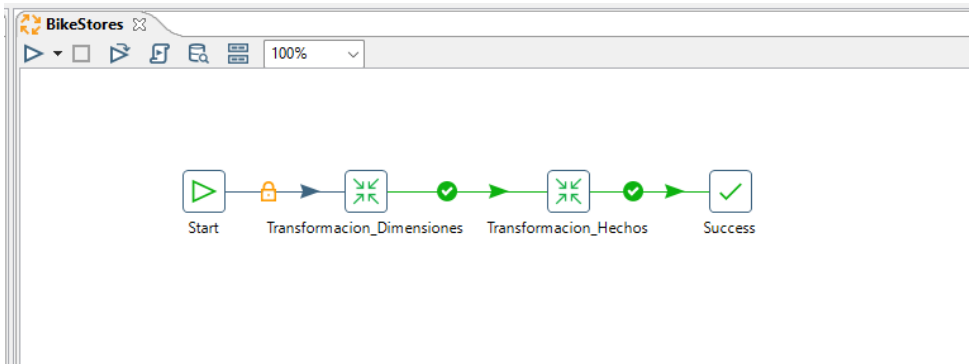
Creación Del ETL

Para crear un ETL se debe crear un flujo de datos en el trabajo, en la pestaña diseño, la categoría general contiene los elementos necesarios para realizar el ETL. los cuales son: el inicio (Start) que indica que nuestro flujo de datos ha iniciado, transformación (transformation) que permite realizar todas las operaciones necesarias con la información y éxito (success) que indica cuando el flujo de datos se ha completado con éxito. Para un modelo en estrella se deben hacer todas las operaciones necesarias primero a las tablas de dimensiones y finalmente a la tabla de hechos.

Para poder usar los elementos antes mencionados basta con arrastrarlos al documento del trabajo. Y para conectar cada elemento con el siguiente, para indicar hacia donde será el flujo de datos, se debe colocar el mouse sobre el elemento y dar clic en el botón que permite crear un conector.

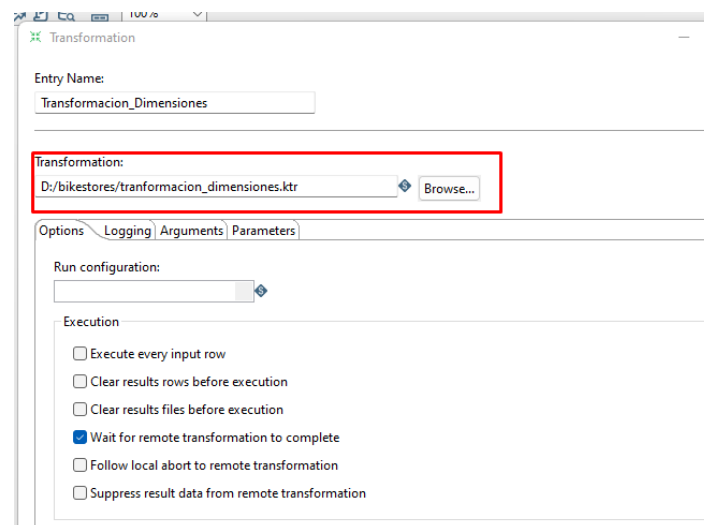


Para cambiar el nombre de algún elemento dentro del flujo de datos solo hay que dar doble clic sobre este y despliega una ventana donde se le puede asignar un nombre. El flujo de datos del ETL queda de la siguiente manera:



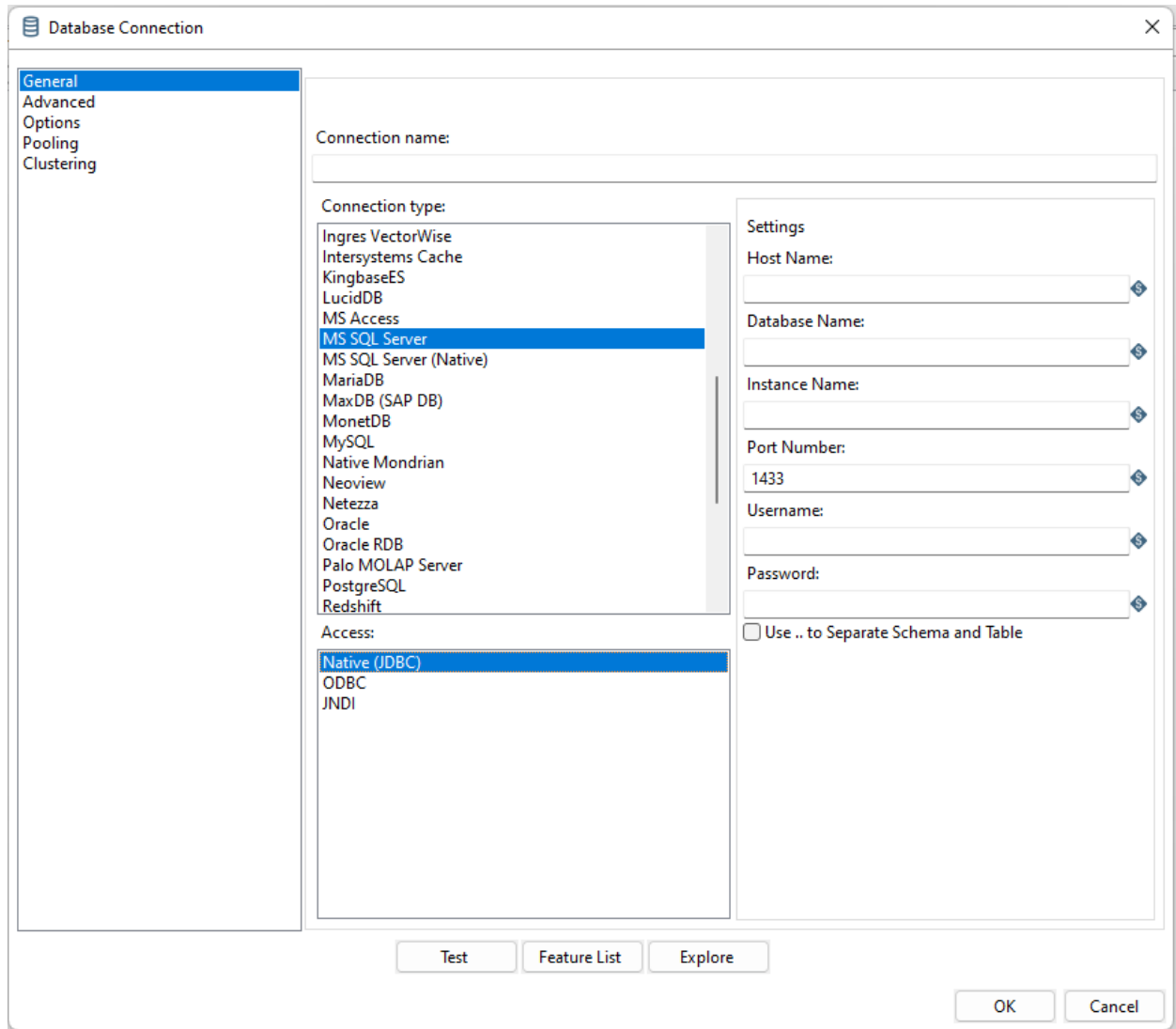
Configurando Las Transformaciones

Al dar doble clic en las transformaciones, despliega una ventana donde debemos asignar la ubicación del archivo de transformación que previamente tuvo que ser creado, para ser creado se sigue el mismo procedimiento para crear un trabajo.



Crear Conexión A Una Base De Datos

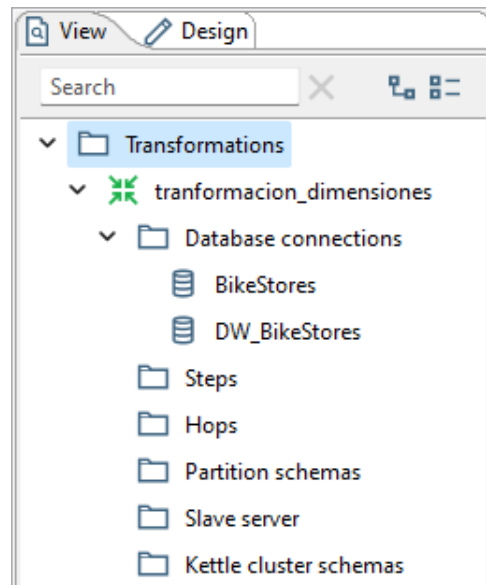
Una conexión a una base de datos ya sean en el trabajo o en la transformación, se crea dando clic derecho en las conexiones a base de datos (Database connections) y en una nueva conexión. Esto muestra en pantalla una ventana donde se configura la nueva conexión.



En el tipo de conexión se muestra una lista de gestores de base de datos a los que es posible conectarse, en este caso se hará una conexión a SQL Server por lo que el tipo de conexión será **MS SQL Server**, en el acceso se proporciona el driver de conexión en este caso se proporcionan tres: JDBC, ODBC y JINDI siendo JDBC el más utilizado y también el driver a implementar en las conexiones. En la configuración se tiene:

- **Host name:** Nombre del servidor donde está alojada la base de datos, en este caso la base de datos está en local, por lo que el parámetro a utilizar es **localhost**.
- **Database name:** Nombre de la base de datos a la que se desea hacer conexión.
- **Instance name:** Nombre de la instancia, por lo general se deja vacío.
- **Port Number:** Puerto expuesto para la conexión con la base de datos, en el caso SQL Server el puerto predeterminado es 1433.
- **Username:** Nombre de usuario.
- **Password:** Contraseña del usuario.

Luego de ingresar los datos necesarios, se puede probar si la conexión es correcta con el botón de Test, si todo está bien se da clic en OK crear la conexión. Por último es necesario asignar el nombre que tendrá la conexión dentro de Pentaho. Para el ETL se requieren dos conexiones. Una de la base de datos origen y la segunda que será la base de datos de destino. Al ser creadas las conexiones se podrán visualizar en el vista del trabajo o transformación. Para cada una de las transformaciones se debe hacer las conexiones.

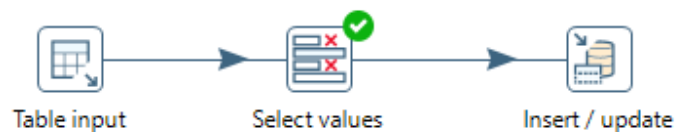


Obtener Y Procesar Registros Para Tablas De Dimensiones

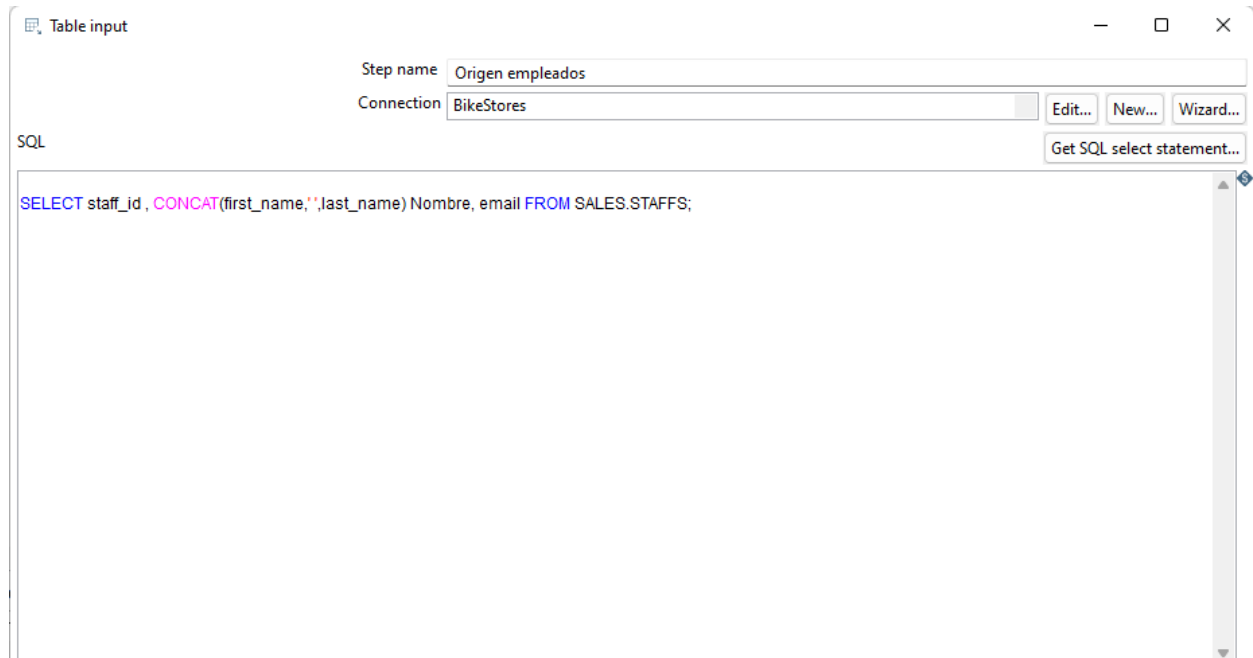
En la transformación de dimensiones se utilizarán tres elementos para poder procesar la información:

1. **Table input.** Permite obtener la información de la base de datos de origen a través de una consulta SQL.
2. **Select values.** En este se procesan los campos obtenidos en el paso anterior.
3. **Insert/ update.** Inserta o actualiza todos los campos en la tabla de la base de destino.

El flujo de datos para cada dimensión se verá exactamente igual. Por lo que para cada dimensión se verá así:



Es posible cambiar el nombre a cada uno de los elementos antes mencionados. Para configurar la tabla de entrada (Table input) es necesario dar doble clic sobre esta, despliega una ventana de configuración:



En step name(nombre del paso) se puede asignar un nombre al elemento, en la conexión se debe seleccionar una de las conexiones antes creadas, en este caso la base de datos de origen. En la caja de texto SQL se coloca la consulta SQL que se usará para obtener la información. Adicionalmente en la parte inferior se proporciona un botón para previsualizar si la información obtenida es la correcta, si es así solo resta dar en OK.

En el siguiente paso, Select values se debe convertir los datos al destino, seleccionando el tipo de dato apropiado que posee la tabla de destino y su correcto tamaño. Esto se hace en la sección meta-data del elemento, proporciona un botón donde podemos obtener los datos que fueron enviados del paso anterior:

Ya en el último paso, insert/ update se debe seleccionar la conexión, en este caso la base de datos de destino. Y buscar la tabla en la cual se insertarán los campos procesados anteriormente:

Insert / update

Step name: Insertar / actualizar empleados

Connection: DW_BikeStores [Edit...] [New...] [Wizard...]

Target schema: dbo [Browse...]

Target table: Empleados [Browse...]

Commit size: 100

Don't perform any updates: ☐

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1				

[Get fields]

Update fields:

#	Table field	Stream field	Update
1			

[Get update fields]
[Edit mapping]

[?] Help [OK] [Cancel] [SQL]

En el botón Get fields permite obtener los campos de origen y destino para compararlos, en este caso la inserción/ actualización se hará a través del ID por lo que solo se necesitan esos campos en ambas partes y compararlos:

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	Id_Empleado	=	staff_id	
2				
3				

Get fields

Si un registro no existe, entonces lo insertará como nuevo, pero en el caso de que si exista se debe actualizar, en la parte inferior se debe seleccionar que campos se actualizan los cuales serán todos a excepción del ID. Además de seleccionar a qué corresponde cada campo obtenido en los pasos anteriores en la tabla destino:

Update fields:

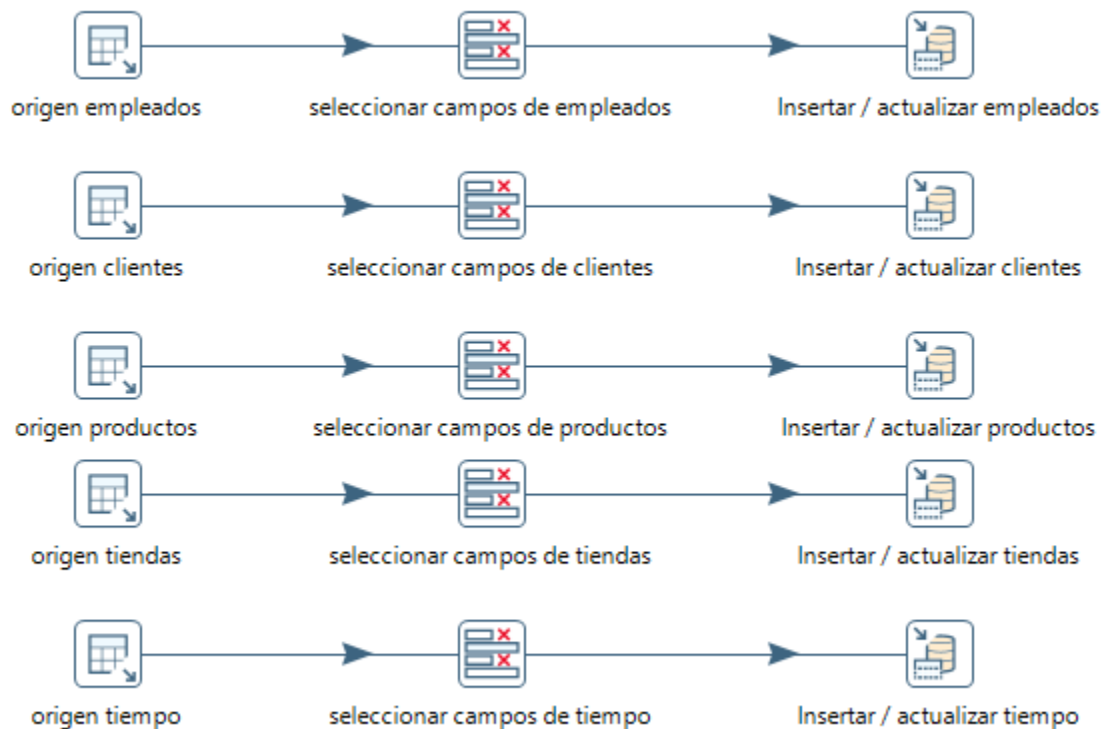
#	Table field	Stream field	Update
1	Id_Empleado	staff_id	N
2	Nombre_Empleado	Nombre	Y
3	Correo_Electronico	email	Y

Get update fields

Edit mapping

Help OK Cancel SQL

Basta con dar OK luego de realizar los pasos mencionados, para el resto de dimensiones se realizan exactamente los mismos pasos. Para este caso la configuración de todas las dimensiones se ve así:



Obtener Y Procesar Registros Para Tablas De Hechos

Para la tabla hechos, se utilizarán los primeros dos elementos del procedimiento anterior, a diferencia que el último paso se utiliza el elemento Table output, por lo que el flujo de datos para la tabla hechos será el siguiente:



En el caso de los primeros dos pasos es el mismo procedimiento. En Table output como primer paso se debe seleccionar la conexión a la base de datos de destino y la tabla donde se insertarán los datos. Adicionalmente se debe seleccionar la opción de Truncate Table para limpiar los registros de la tabla. Y

specify database fields para seleccionar los campos de la tabla destino.

The screenshot shows the 'Table output' configuration window. The 'Step name' is 'Table output'. The 'Connection' is 'DW_BikeStores' with buttons for 'Edit...', 'New...', and 'Wizard...'. The 'Target schema' is 'dbo' with a 'Browse...' button. The 'Target table' is 'Hechos_Ordenes' with a 'Browse...' button. The 'Commit size' is '1000'. The 'Truncate table' checkbox is checked. The 'Ignore insert errors' checkbox is unchecked. The 'Specify database fields' checkbox is checked.

Step name	Table output		
Connection	DW_BikeStores	Edit...	New...
		Wizard...	
Target schema	dbo	Browse...	
Target table	Hechos_Ordenes	Browse...	
Commit size	1000		
Truncate table	<input checked="" type="checkbox"/>		
Ignore insert errors	<input type="checkbox"/>		
Specify database fields	<input checked="" type="checkbox"/>		

Luego se deben seleccionar a que corresponde a cada campo de los campos anteriores en la tabla destino. Si todo está correcto, se da click en OK.

#	Table field	Stream field
1	Id_Cliente	customer_id
2	Id_Empleado	staff_id
3	Id_Tienda	store_id
4	Id_Producto	product_id
5	Id_Tiempo	Id_Tiemp
6	Total_Venta_Producto	Total_Venta_Producto
7		

Al ejecutar el ETL, si todo está correcto se debe ver de la siguiente manera:

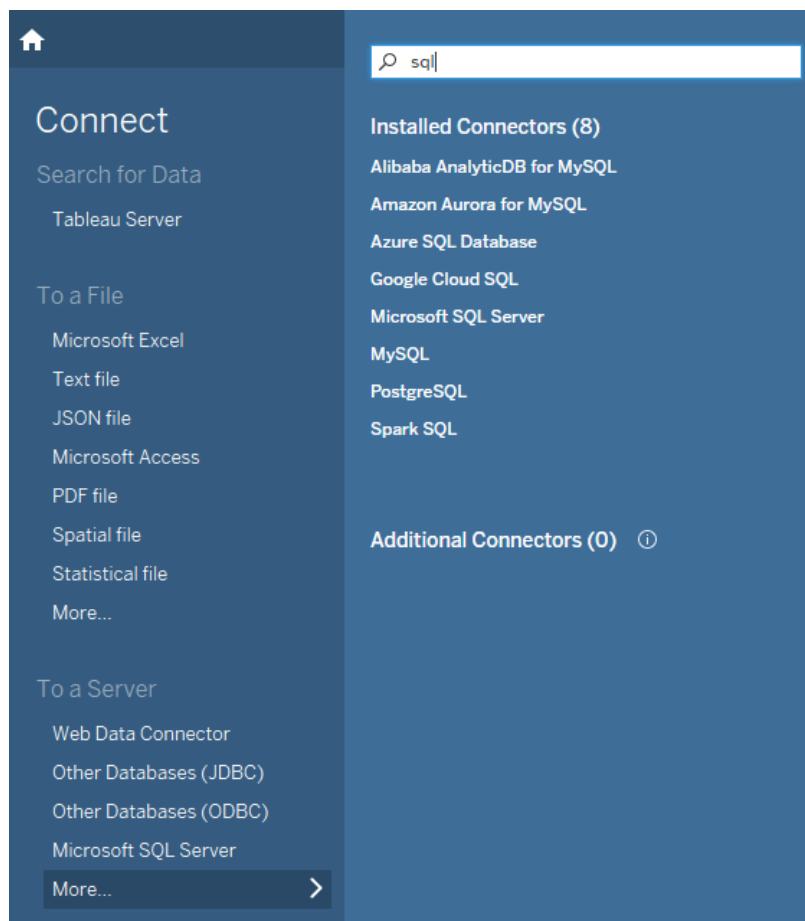


Tableau

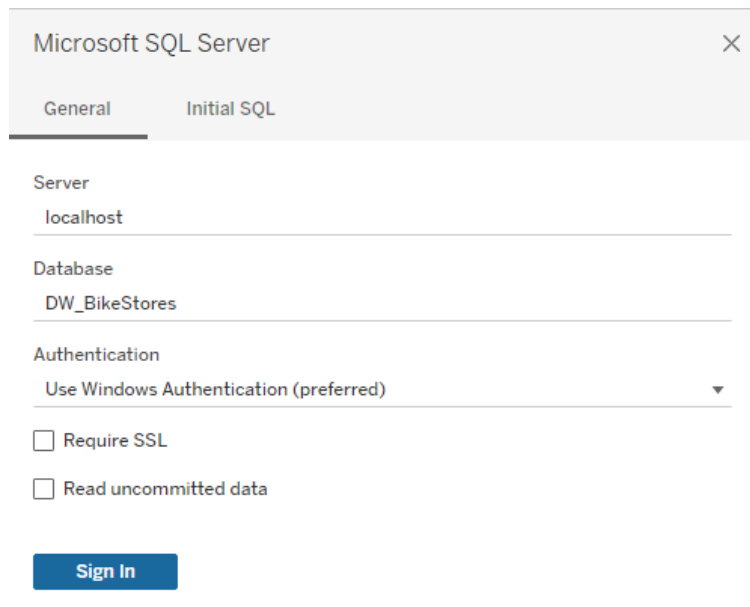
Tableau es visto como un programa de cambio empresarial, para cambiar la cultura y la forma de pensar sobre los datos y la toma de decisiones. A medida que trabajamos en nuestro viaje y mostramos a las personas lo fácil que era obtener información por sí mismos, surgieron los verdaderos momentos eureka. Ahora podemos dar un paso atrás y apoyarlos en sus propios viajes en lugar de tener que hacer todo de forma centralizada.

Conexión Con MSSQL

Para conectar a SQL Server dentro de Tableau, se debe ir a más conexiones y seleccionar SQL Server:



Esto abre una ventana donde se pide el nombre del servidor, en este caso **localhost** y los credenciales de autenticación, es posible autenticarse a través de windows authentication o con usuario y contraseña.



The screenshot shows the 'Microsoft SQL Server' connection window. It has two tabs: 'General' (selected) and 'Initial SQL'. Under 'General', there are three input fields: 'Server' with 'localhost', 'Database' with 'DW_BikeStores', and 'Authentication' with a dropdown menu set to 'Use Windows Authentication (preferred)'. Below these are two checkboxes: 'Require SSL' and 'Read uncommitted data', both of which are unchecked. At the bottom left is a blue 'Sign In' button. A close button (X) is in the top right corner.

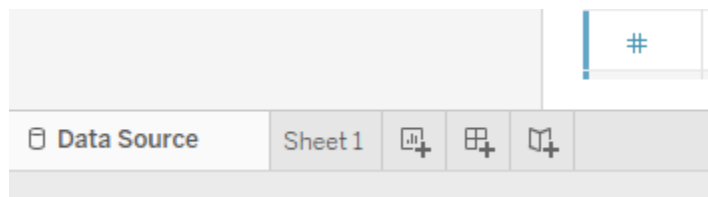
Selección De Tablas

El paso siguiente es seleccionar la base de datos a utilizar, en este caso el datawarehouse seleccionado todas las tablas relacionadas entre sí. Estas deben arrastrarse al espacio de trabajo correspondiente.

The screenshot shows a data modeling interface. On the left, a 'Table' list includes 'Clientes', 'Empleados', 'Hechos_Ordenes', 'products', 'Tabla_Tiempo', and 'Tienda'. In the center, a star schema diagram shows 'Hechos_Ordenes' as the central fact table, connected to five dimension tables: 'Clientes', 'Empleados', 'products', 'Tabla_Tiempo', and 'Tienda'. On the right, a data preview for 'Hechos_Ordenes' is shown with 7 fields and 4722 rows. The preview table has the following structure:

#	Hechos_Ordenes Id Codigo	#	Hechos_Ordenes Id Empleado	#	Hechos_Ordenes Id Tiempo	#	Hechos_Ordenes Id Producto	#	Hechos_Ordenes Id Cliente	#	Hechos_Ordenes Id Tienda	#	Hechos_Ordenes Total Venta Producto
1		2		21/8/2017		30		2		1		949.99	
2		2		21/8/2017		60		2		1		1,403.99	
3		2		21/8/2017		65		2		1		645.40	
4		2		21/8/2017		70		2		1		626.99	
5		2		21/8/2017		98		2		1		783.98	
6		2		27/3/2018		50		3		1		11,159.98	
7		2		27/3/2018		58		3		1		4,649.99	

Posteriormente, en la parte inferior izquierda están disponibles las pestañas para la generación de reportes. El data source es la selección de todas las tablas, y el resto serán los diferentes reportes creados:



Y finalmente, en la generación de reportes se deben arrastrar los diferentes campos con los que se desea crear un reporte.

Analytics

Hechos_Ordenes+ (DW_...

Search

Tables

Cientes

Correo Cliente

Id Cliente (Clientes)

Nombre completo

Clientes (Count)

Empleados

Correo Electronico

Id Empleado (Emple...

Nombre Empleado

Empleados (Count)

Hechos_Ordenes

Id Cliente

Id Codigo

Id Empleado

Id Producto

Id Tiempo

Id Tienda

Total Venta Producto

Hechos_Ordenes (C...

products

Anio Modelo

Id Producto (produc...

Nombre Categoria

Nombre Marca

Nombre Producto

products (Count)

Tabla_Tiempo

Dia Semana

Id Tiempo (Tabla Ti...

Mes

Trimestre

Anio

Semestre

Tabla_Tiempo (Cou...

Filters

Marks

Automatic

Color

Size

Text

Detail

Tooltip

Columns

Rows

Sheet 1

Drop field here

Reportes Generados En Tableau

Total de ventas por tienda



Figura 1

Nota. Reporte donde se puede visualizar la cantidad de ventas que cada sucursal de la tienda generó a lo largo del tiempo.

Cantidad de producto por tienda

		Nombre Producto																			
Direccion	Nombre Tienda	Electra A msterda..	Electra A msterda..	Electra A msterda..	Electra A msterda..	Electra A msterda..	Electra A msterda..	Electra Cruiser 1..	Electra Cruiser 1..	Electra Cruiser 1..	Electra Cruiser 1..	Electra Cruiser 7..	Electra Cruiser 7..	Electra Cruiser 7..	Electra Cruiser 7..	Electra Cruiser L..	Electra Cruiser L..	Electra Cruiser L..	Electra Cruiser L..	Electra Cruiser L..	Electra Cruiser L..
3700 Portol..	Santa Cruz ..	1	5	4	3		1	2	29	1		1				1	5	1		1	2
4200 Chest..	Baldwin Bik..	1	18	16	16	1	1	1	138	2			6	2	2	1	9	4	4	1	1
8000 Fairw..	Rowlett Bik..	1	1	2	4			1	26	1	1					1	5	1			

Figura 2

Nota. Reporte de la cantidad de cada producto por nombre que tiene cada tienda.

Cantidad de productos por cada cliente

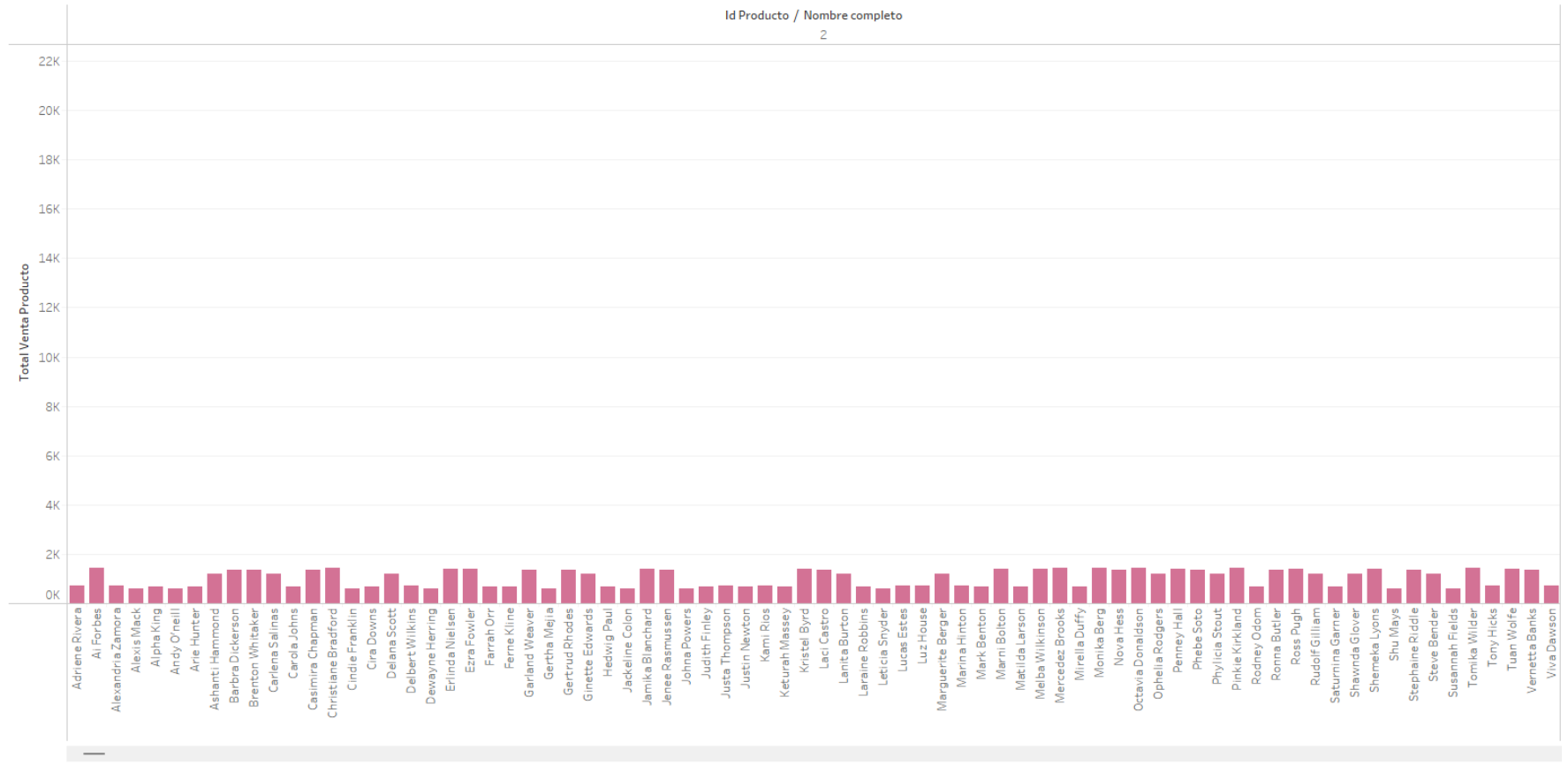


Figura 3

Nota. Reporte de la cantidad de producto que ha adquirido cada cliente en todas las tiendas.

Total ventas por mes

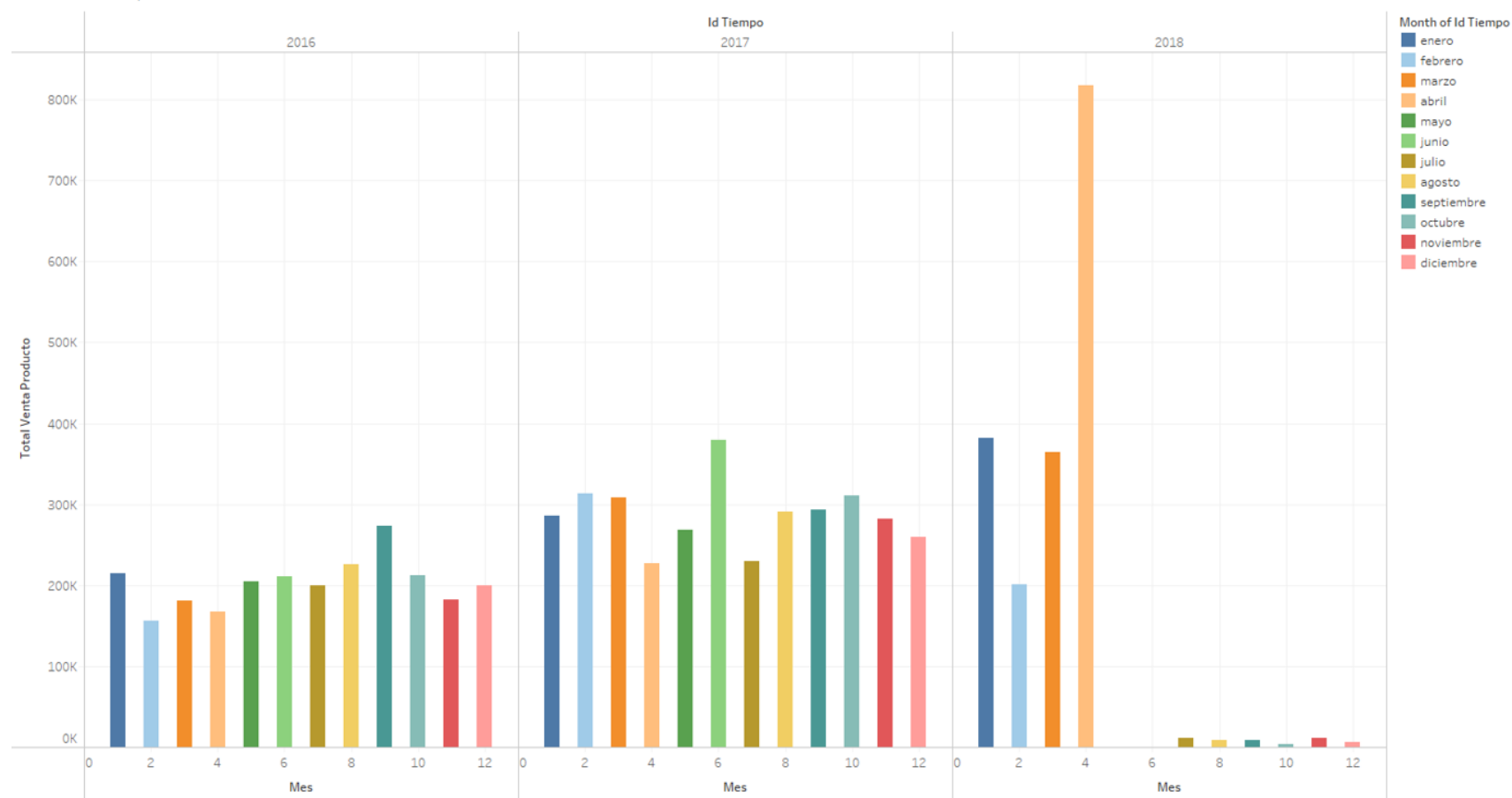


Figura 4

Nota. Reporte de las ventas hechas por mes y en cada año.

Cantidad de productos de cada categoría por marca

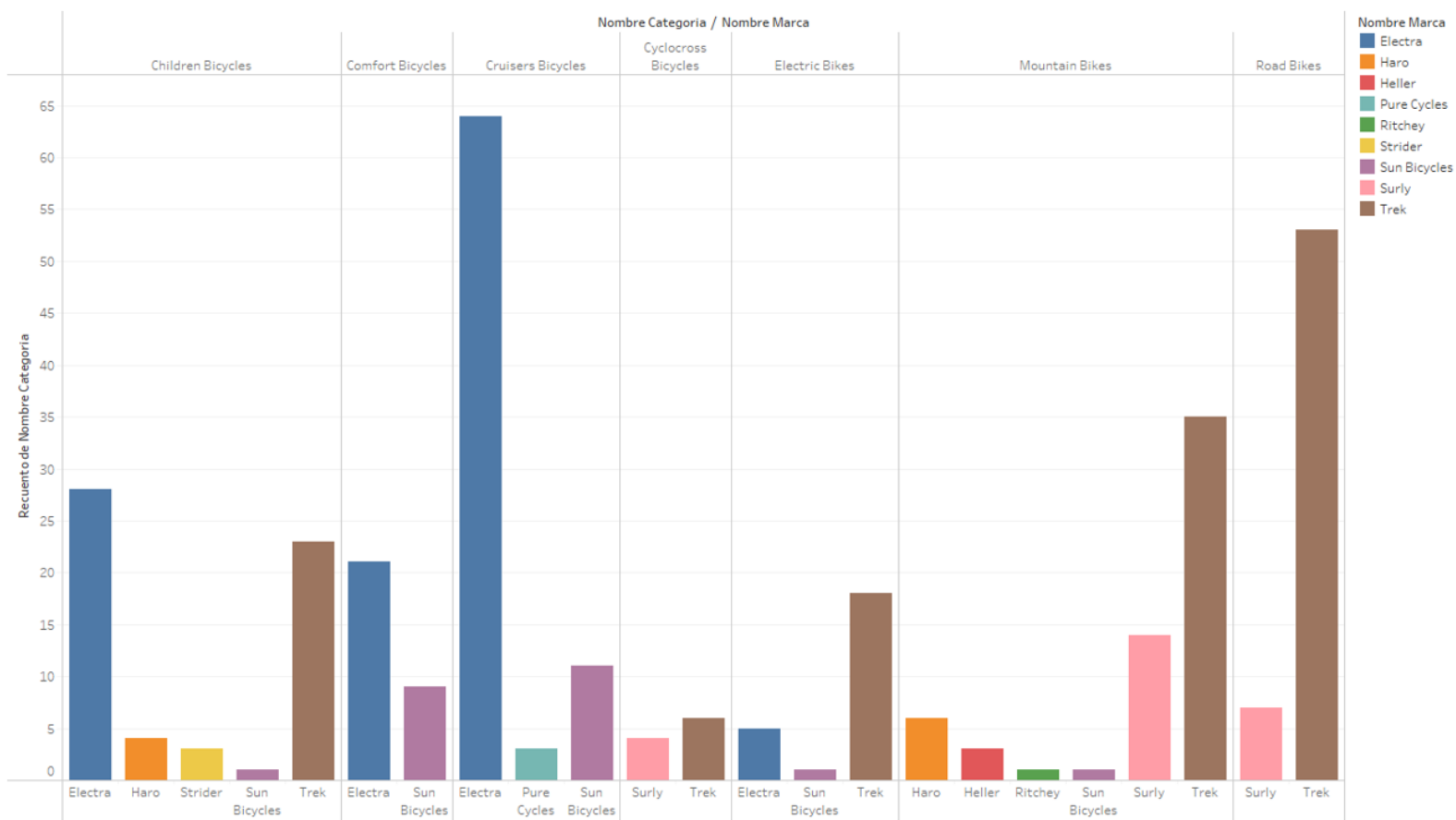


Figura 5

Nota. Reporte de la cantidad de productos de las diferentes categorías por marca.

Total ventas por trimestre

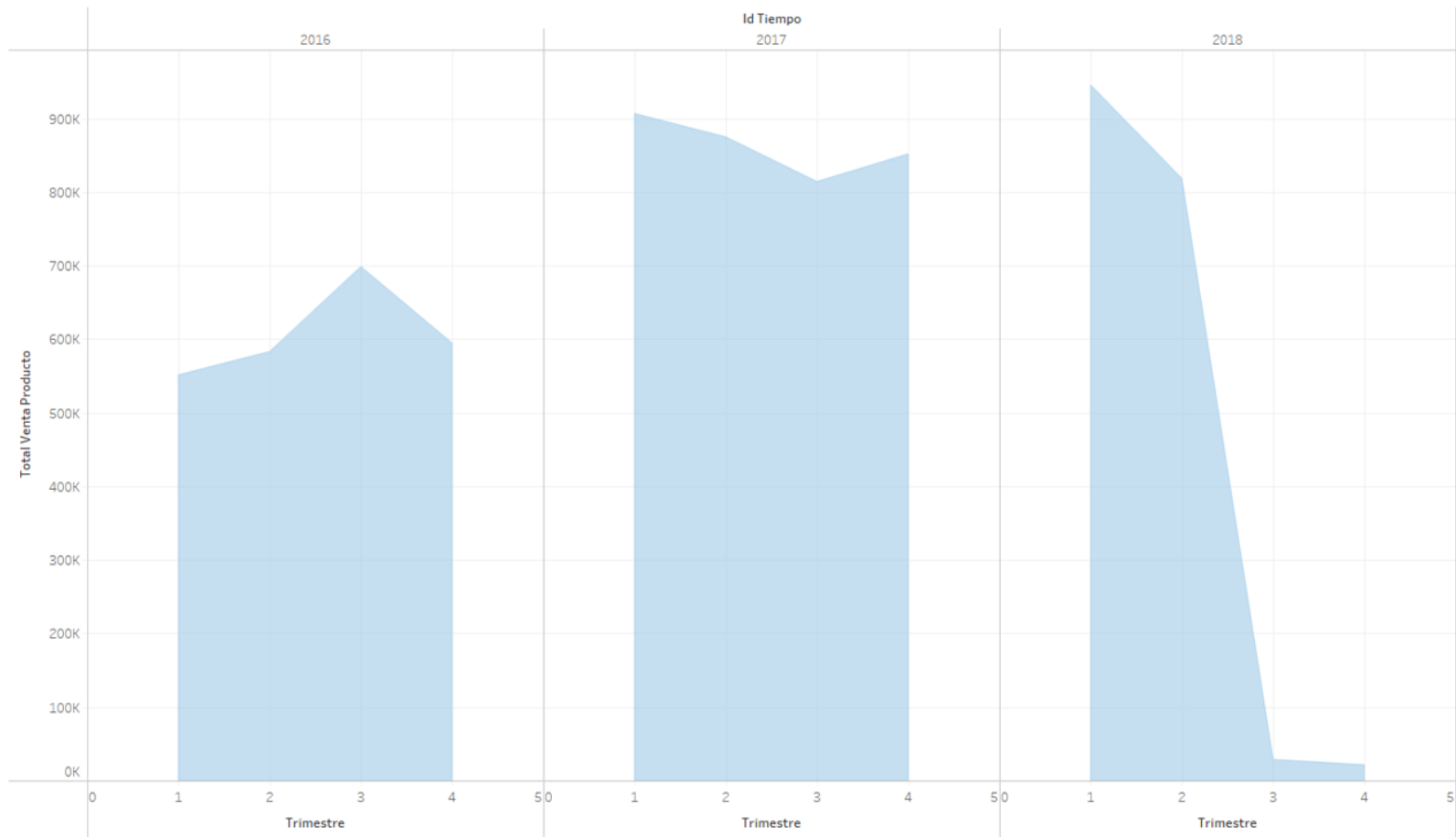


Figura 6

Nota. Reporte de la cantidad de ventas por trimestre.

Total de ventas por día



Figura 7

Nota. Reporte de la cantidad de ventas por día de la semana.

Total ventas por semestre

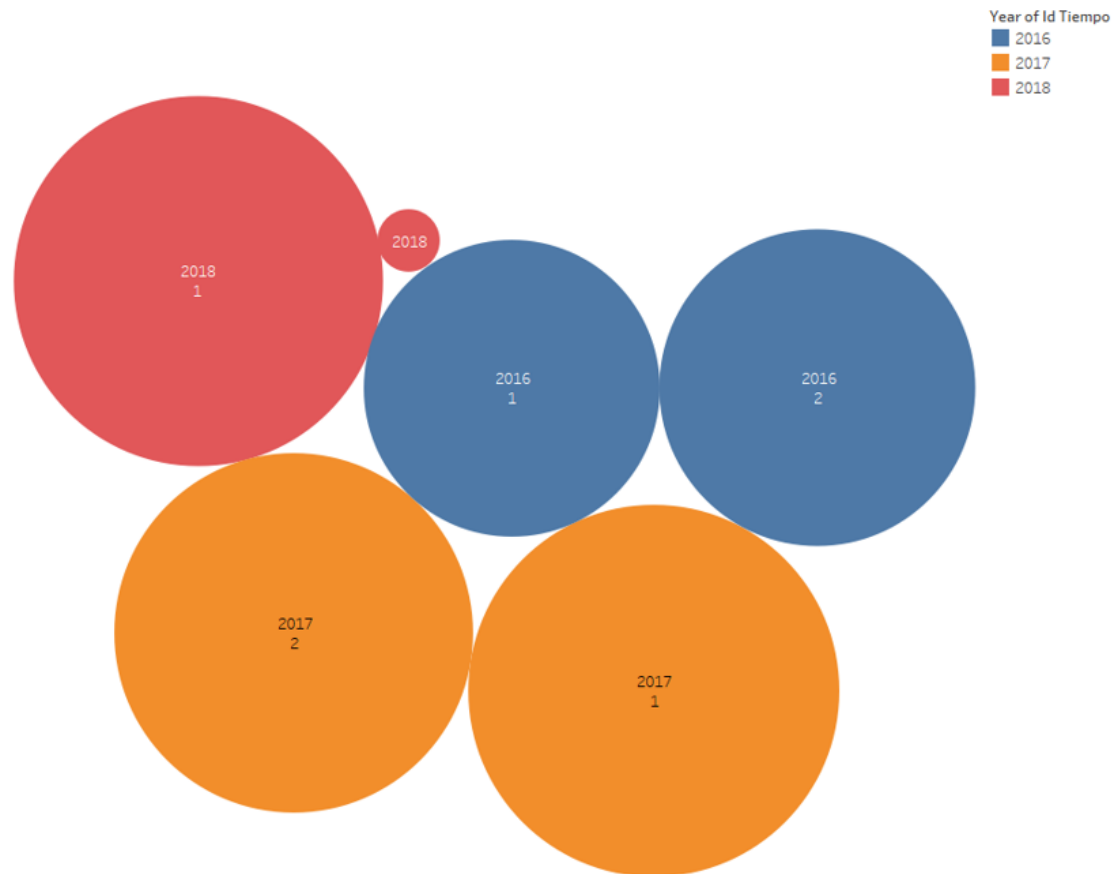


Figura 8

Nota. Reporte de la cantidad de ventas por semestre.

Total de ventas por empleado



Figura 9

Nota. Reporte de la cantidad de ventas de cada empleado.

Cantidad de clientes y productos por año

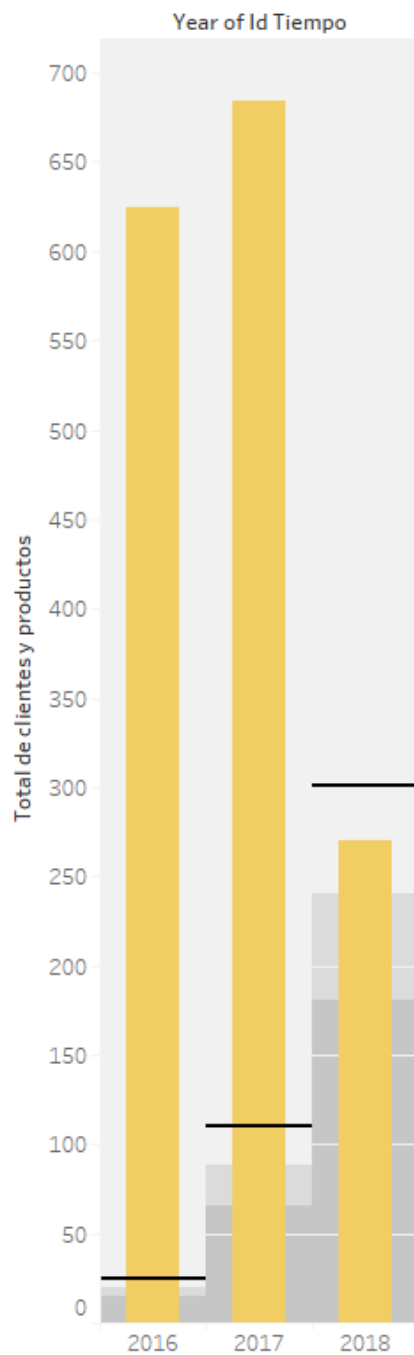
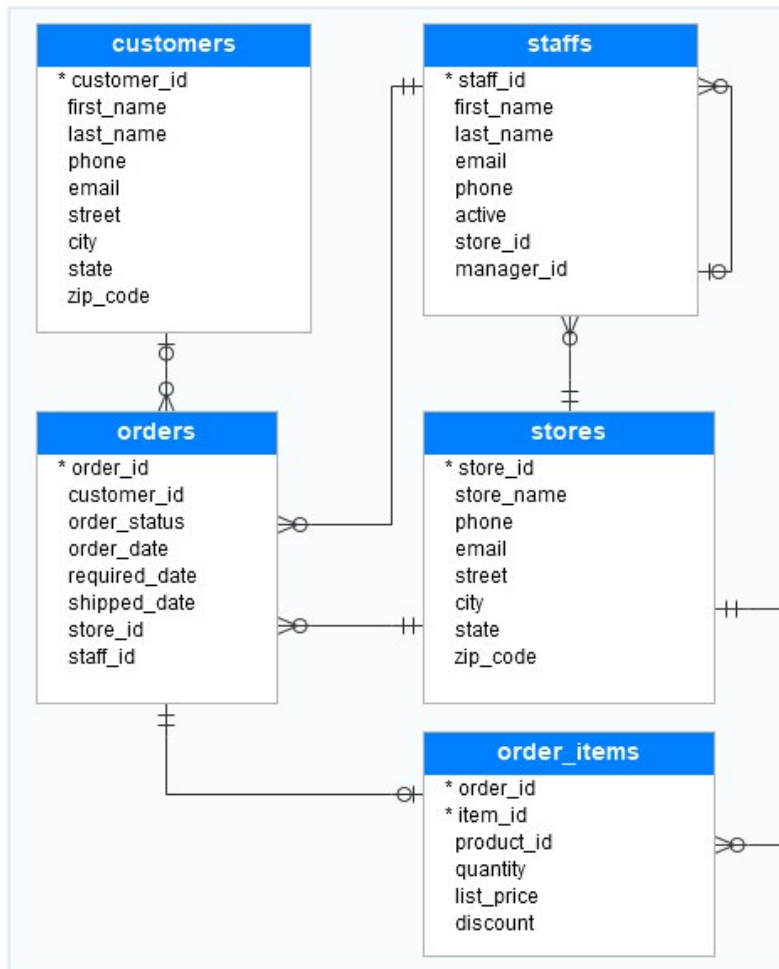


Figura 10

Nota. Reporte de la cantidad de productos y clientes registrados por año.

Base De Datos OLTP

Sales



Production

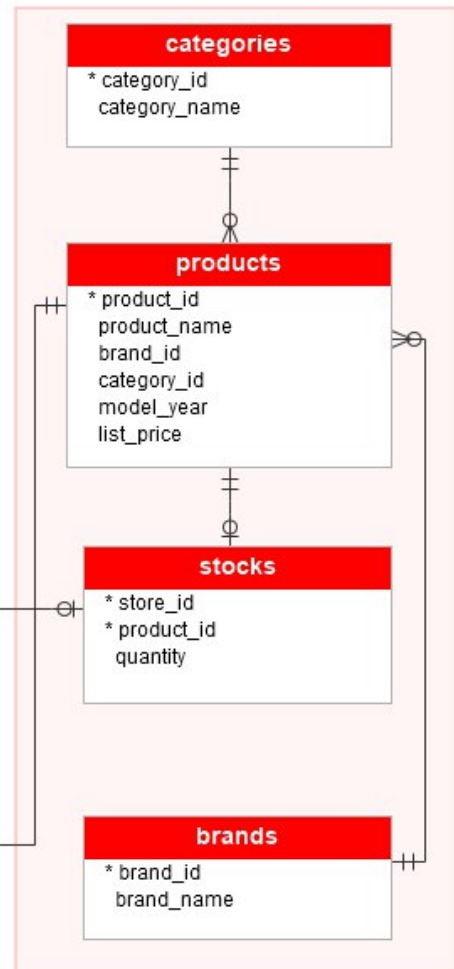


Figura 11

Base de datos OLTP.

Fuente. SQL Server Sample Database. (s. f.). [Ilustración]. SQL Server Sample Database.

<https://www.sqlservertutorial.net/wp-content/uploads/SQL-Server-Sample-Database.png>

Fuente de la base de datos: [sql server tutorial](https://www.sqlservertutorial.net)

Base De Datos OLAP

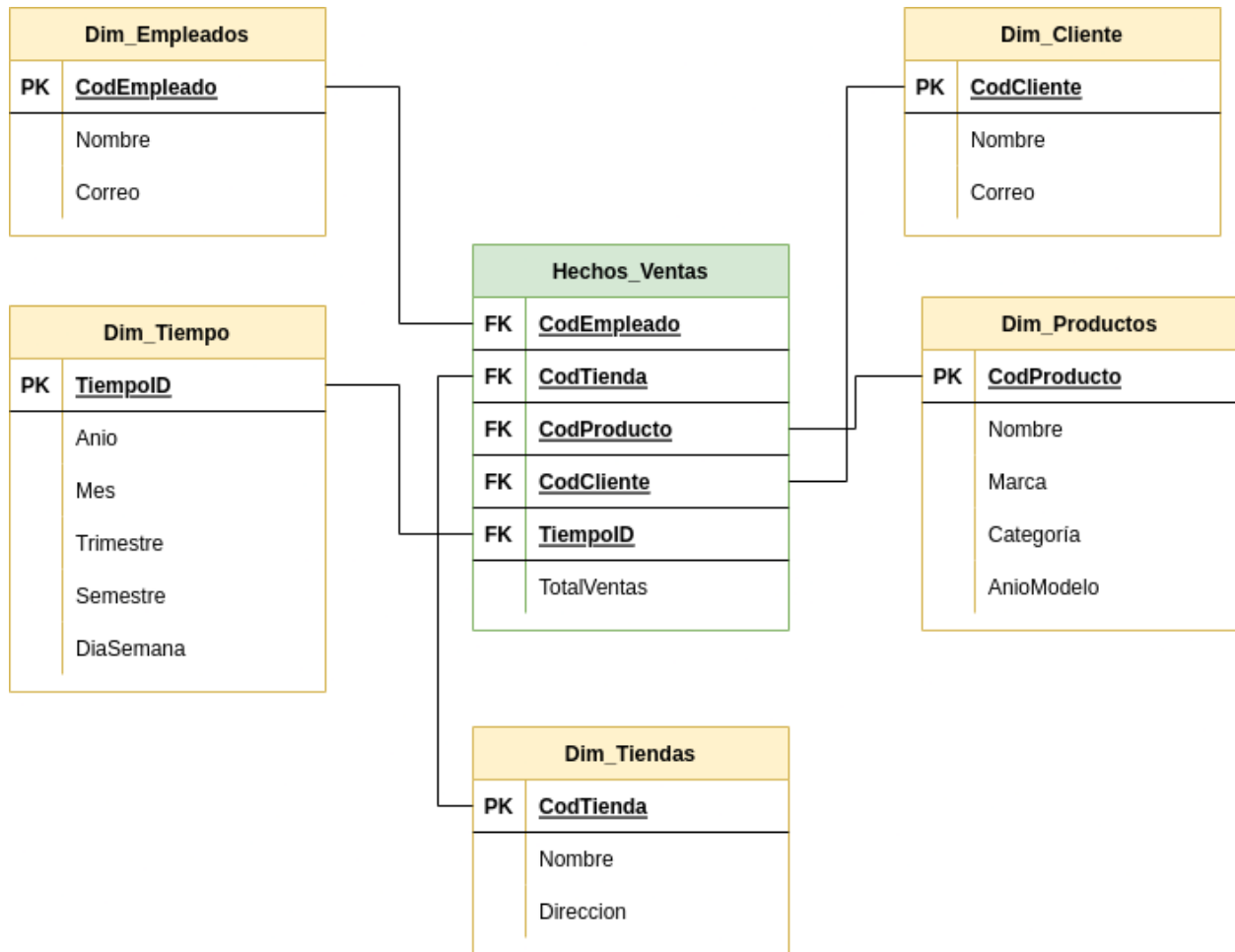


Figura 12

Modelo en estrella.

Preguntas Del Negocio Utilizadas

Preguntas del negocio:

- Se desea conocer el total de ventas de las órdenes que se han realizado.
- Se debe mostrar información de las tiendas, mostrando el código de la tienda, nombre de la tienda y dirección.
- Se debe mostrar información de los empleados, como el código, nombre, apellido y correo electrónico.
- Se debe mostrar información de los productos, el código del producto, nombre del producto, la marca, la categoría y el año del modelo.
- Se debe mostrar información de los clientes, código del cliente, nombre, apellido y correo electrónico.
- La información debe mostrarse en año, mes, trimestres, semestre, y día de la semana.

Explicación De Métrica Utilizada

1. Venta total de productos

Con dicha métrica se pretende medir y analizar, cuáles fueron las ventas totales realizadas en la compañía a través de la cantidad de órdenes realizadas en determinados tiempos ("día de la semana", "mes", "trimestre", "semestre" y "año").

Conclusiones

- De acuerdo a los reportes generados, el negocio arrancó por buen camino, aumentando la disponibilidad de varios productos en los dos primeros años y al mismo tiempo aumentando el número de clientes, pero en el tercer año se aumentó el volumen de inventario, comprando nuevos productos, sin embargo esto tiene un impacto negativo en la empresa debido a que el número de clientes ha disminuido significativamente. Esto se puede visualizar en la Figura 10.
- Tableau es muy fácil de usar y creemos que generar informes es simple, ya que se guía por la misma interfaz.
- En el reporte de la Figura 1, podemos ver que los ingresos totales se concentran principalmente en la tienda Baldwin Bikes, por lo que las otras dos tiendas no son rentables.
- En el informe contenido en la Figura 8, podemos ver que desde la segunda mitad de 2018, las ventas han disminuido significativamente.
- Crear un ETL con Pentaho Data Integration (Spoon) es similar a crear un ETL usando Visual Studio con el mismo sistema de arrastrar y soltar, por lo que es fácil de utilizar.

Recomendaciones

- En el caso de una conexión a SQL Server se debe tener instalado el driver correspondiente, en este caso el JDBC y este a su vez está dentro del driver JTDS el cual puede ser descargado de la siguiente página: [Driver JTDS](#).
- En caso de usar Pentaho Data Integration (Spoon) como alternativa en sistemas basados en Unix se puede hacer uso de docker con [webspoon](#).
- La herramienta Tableau si el tipo de dato es DATE automáticamente lo reconoce y es posible generar informes anuales, semestrales, trimestrales, mensuales, y diarios directamente desde Tableau, por lo que si una base de datos datawarehouse está destinada a ser visualizada en Tableau no es necesario crear los campos antes mencionados en la tabla de dimensiones de tiempo.
- En los datos de tipo DATE en Pentaho Data Integration al usar la sentencia CONVERT es recomendable convertirlo a VARCHAR debido a que si la conversión es a tipo DATE genera errores al ejecutar el ETL.
- Al utilizar pentaho Spoon, si se tiene instalada Pentaho Business Analytics 5.4 por defecto al instalarse este crea una variable de entorno llamada PENTAHO_JAVA_HOME el cual es una dirección de la versión de java incorporada en Pentaho Business Analytics, al querer ejecutarse pentaho Spoon intenta utilizar esta variable de entorno sin embargo esta versión de java es inferior a la requerida por lo que es recomendable cambiar la versión de java en la variable de entorno.

Bibliografía

Get Started. (s. f.). Tableau. Recuperado 11 de agosto de 2022, de
https://help.tableau.com/current/pro/desktop/en-us/gettingstarted_overview.htm

Install Drivers with the JDBC Distribution Tool. (2021, 4 agosto). Hitachi Vantara Lumada and Pentaho Documentation.

<https://help.hitachivantara.com/Documentation/Pentaho/5.3/0D0/160/030>

Analytics, P. (2018, 17 septiembre). *Tableau in Two Minutes - Tableau Basics for Beginners.* YouTube. <https://www.youtube.com/watch?v=jEgVto5QME8&feature=youtu.be>