



ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Informe de: Inteligencia de negocios

Tema: Proceso ETL con Pentaho

Nombre: Danny Sebastián Díaz Padilla

Fecha: 06/11/2019

1. Objetivos:

1.1. Objetivo General

- Crear un proceso de extracción, transformación y carga.

1.2. Objetivos Específicos

- Entender la plataforma pentaho para el análisis de datos.
- Crear visualizaciones con qlik.
- Conectar distintas fuentes de datos para extraer y transformar su data.

2. Marco teórico:

Pentaho

Pentaho es una plataforma de Business Intelligence (BI) orientada a la solución y centrada en procesos que incluye los componentes requeridos para implementar soluciones basadas en procesos como minería de datos, ETL y generación de informes.

Pentaho nos ofrece una serie de útiles productos como son los siguientes:

- Pentaho Reporting: Es un motor de presentación capaz de generar informes programáticos sobre la base de un archivo de definición XML.
- Pentaho Dashboard: Es una plataforma integrada para proporcionar información sobre sus datos, donde se pueden ver informes, gráficos, etc.
- Pentaho Data Mining: Es una suite de software que usa estrategias de aprendizaje de máquina, automático y minería de datos.
- Pentaho para Apache Hadoop: Es un conector de bajo nivel para facilitar el acceso a grandes volúmenes manejados en el proyecto Apache Hadoop.
- Pentaho Analysis Services: Es compatible con el MDX, y el lenguaje de conducta XML para el análisis y especificaciones de la interfaz.

Qlik sense

Qlik Sense es una aplicación de visualización y descubrimiento de datos gobernada, basada en servidor, ideal para las necesidades analíticas de grupos, departamentos o toda una organización. Los usuarios de negocio obtienen un análisis de datos potente, flexible y personalizado y colaboración en cualquier dispositivo, a la vez que se adhieren a unas políticas de gobierno y seguridad centralizada de datos.

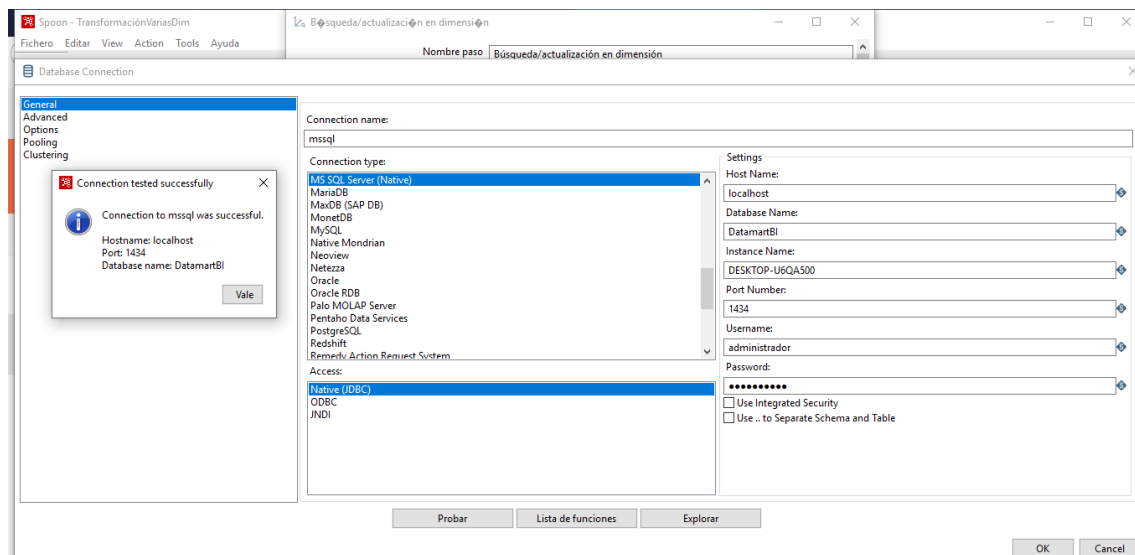
3. Desarrollo de la práctica:

Para trabajar en la plataforma pentaho se debe descargar el archivo del siguiente enlace <https://sourceforge.net/projects/pentaho/files/Data%20Integration/>

Una vez descargado y descomprimido se procede a ejecutar el archivo Spoon.bat. Como prerequisite el jdk de Java debe haber sido instalado.

Para almacenar las dimensiones se crea una conexión con la base de datos, en este caso MS SQL. Solo se puede usar username registrados en el login de MS SQL.

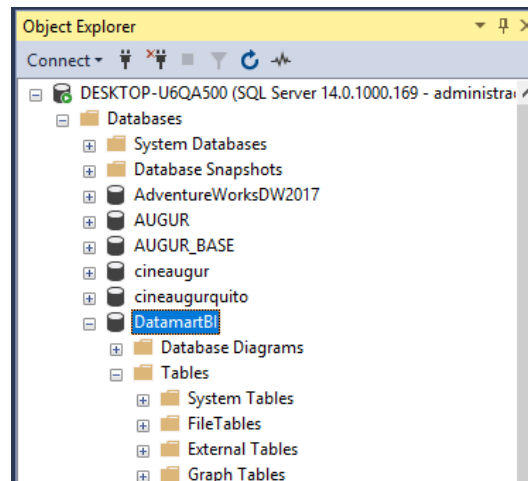
Es importante guardar en la carpeta "lib" el conector JDBC de la página oficial de Microsoft. Por lo general MS SQL suele escuchar en el puerto 1433, sin embargo la instalación que se tiene escucha en el 1434.





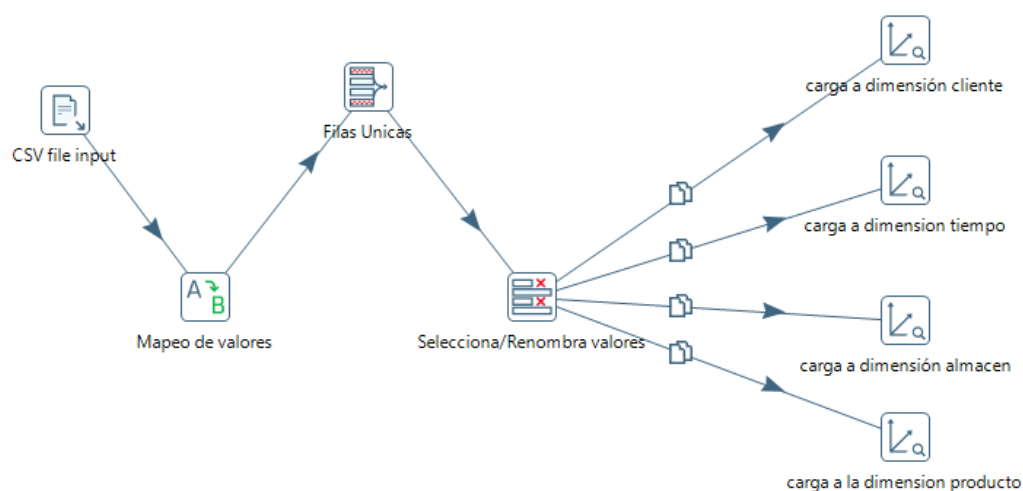
ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Para el datamart es necesario tener una base de datos y en lo posible un esquema de las dimensiones, en este caso se crearán las dimensiones desde el mismo Spoon.

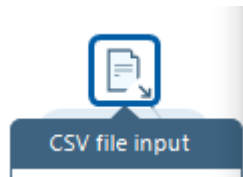


Siguiendo el siguiente tutorial <https://www.youtube.com/watch?v=0qGOBFhykek> se aplica una extracción desde un archivo CSV, se realiza 3 transformaciones y se carga a una dimensión.

Esquema ETL



Extracción



En base a la factura del negocio elegido se crea un CSV con el registro de varias facturas.

Herrajes e Herrerías Sur
DOCUMENTO COTIZADO-VAPORADO
RUC: 1281224224001
Fecha: 05/11/2019
Cliente: JORDY ALEXIS JACOME GAROFALO
Dirección: CATÓLICA AV MACHACHI 1111, LAMAYO, DISTRITO CANTÓN CATÓLICA, PROV. CAHAZUZA
Forma de pago: Efectivo
Monto: 237.15
Factura de Autorización: 258847203819
001-001-NE 0021675

Cantidad	Descripción	Unidad	Precio Unitario	Total
1	1	2829917	E	27.45
2	2	2829237	E	56.25
3	3	2829917	E	87.45
4	4	2129917	TD	127.45
5	5	2229917	TC	327.99
6	6	2119917	E	87.45
7	7	2833917	E	10.00
8	8	2119917	E	0.45
9	9	2877917	E	3.45
10	10	2222917	TD	237.15

1	Numero, RUC, Dirección, Fecha, Nombre, Apellido, VendedorID, Teléfono, Forma de pago, Total
2	1, 1720254224001, Católica av Machachi, 05/11/2019, Cliente1, ApCliente1, 1, 2829917, E, 27.45
3	2, 1776254224001, Católica av Machachi, 01/11/2019, Cliente2, ApCliente2, 1, 2829237, E, 56.25
4	3, 1720904224001, Católica av Machachi, 04/11/2019, Cliente3, ApCliente3, 2, 2829917, E, 87.45
5	4, 1744254224001, Católica av Machachi, 03/11/2019, Cliente4, ApCliente4, 2, 2129917, TD, 127.45
6	5, 1711254224001, Católica av Machachi, 05/11/2019, Cliente1, ApCliente1, 1, 2229917, TC, 327.99
7	6, 1660254224001, Católica av Machachi, 04/11/2019, Cliente5, ApCliente5, 2, 2119917, E, 87.45
8	7, 1798254224001, Católica av Machachi, 03/11/2019, Cliente6, ApCliente6, 1, 2833917, E, 10.00
9	8, 1789754224001, Católica av Machachi, 05/11/2019, Cliente7, ApCliente7, 2, 2119917, E, 0.45
10	9, 1520254224001, Católica av Machachi, 03/11/2019, Cliente6, ApCliente6, 1, 2877917, E, 3.45
11	10, 0820254224001, Católica av Machachi, 04/11/2019, Cliente8, ApCliente8, 1, 2222917, TD, 237.15



ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

En el paso de extracción se selecciona el directorio y automáticamente traerá los campos.

CSV file input

Step name: CSV file input

Filename: C:\Users\COMPANY\Downloads\BI\Facturas.csv Examinar...

Delimiter: , Insert TAB

Enclosure: ""

NIO buffer size: 50000

Lazy conversion? ☒

Header row present? ☒

Add filename to result ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

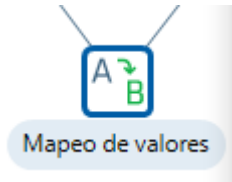
Format: mixed

File encoding: utf-8

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim
1	Numero	Integer	#	15	0	€	,	.	ningi
2	RUC	Integer	#	15	0	€	,	.	ningi
3	Direccion	String		21		€	,	.	ningi
4	Fecha	String		11		€	,	.	ningi
5	Nombre	String		9		€	,	.	ningi
6	Apellido	String		11		€	,	.	ningi

Transformación

La primera transformación consiste en buscar ciertos valores de la columna “Forma de pago” y cambiarlos por otros (TD por Tarjeta de débito, TC por Tarjeta de crédito, E por Efectivo). De modo que acoplemos a las reglas del negocio de la tabla hechos



Mapeo de valores

Nombre de paso : Mapeo de valores

Nombre de campo origen : Forma de pago

Nombre de campo destino (vacío=sobreescribir) :

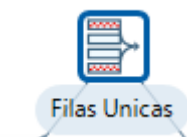
Default upon non-matching :

Valores de campo:

#	Valor origen	Valor destino
1	TD	Tarjeta de débito
2	TC	Tarjeta de crédito
3	E	Efectivo

Help Vale Cancelar

La segunda transformación se encarga de eliminar duplicados.



Filas Unicas

Nombre de paso : Filas Unicas

Settings

☒ Agregar contador a la salida? ☐ Campo Contador

☐ Redirect duplicate row ☐ Error description

Campos de comparación (dejar vacío para utilizar la fila entera)

#	Nombre de campo	Ignorar Mayúsculas/Minúsculas
1		

Help Vale Cancelar Obtener



ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Y por último la última transformación renombra un campo y selecciona las filas que se pasarán a las distintas dimensiones de la base de datos.

Selecciona/Renombra valores

Nombre paso

Selecciona & Modifica Eliminar Meta-información

Campos :

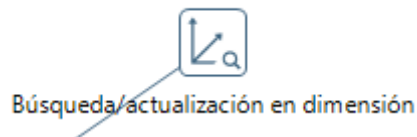
#	Nombre campo	Renombrar a	Longitud	Precisión
1	Numero	FacturalID		
2	RUC			
3	Direccion			
4	Fecha			
5	Nombre			
6	Apellido			
7	VendedorID			
8	Teléfono			
9	Forma de pago			
10	Total			

Obtener campos a seleccionar Edit Mapping

Include unspecified fields, ordered ☐

Help Vale Cancelar

Carga



DIMENSION_CLIENTE

Esta tabla será creado con la llave primaria igual al RUC almacenado en la factura.

Búsqueda/actualización en dimensión

Tabla destino: DIMENSION_CLIENTE [Examinar...]

Tamaño transacción: 100

Enable the cache? ☒

Pre-load the cache? ☐

Cache size in rows (0 = cache all): 5000

Claves Campos

Campos clave (para buscar fila en dimensión):

#	Campo de Dimensión	Campo en flujo
1	RUC	RUC

Campo de clave técnica: ClientelD [Nuevo nombre]

Creación de clave técnica

☒ Utilizar máximo tabla + 1

☐ Utilizar secuencia []

☐ Utilizar campo auto-incrementativo

Campo de Versión: version

Campo Fecha flujo: []

Campo inicio rango fecha: date_from Año m. 1900

Use an alternative start date? ☐ <Select Option> []

Campo final rango fecha: date_to Año m. 2199

Vale Cancelar Obtener Campos SQL

? Help



ESCUELA POLITÉCNICA NACIONAL FACULTAD DE INGENIERÍA DE SISTEMAS INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Y los campos solo se tomarán el nombre, apellido y teléfono, los demás se los dejará a lado para las otras dimensiones.

Búsqueda/actualización en dimensión

Tabla destino: DIMENSION_CLIENTE [Examinar...](#)

Tamaño transacción: 100

Enable the cache? ☒

Pre-load the cache? ☐

Cache size in rows (0 = cache all): 5000

Claves Campos

Campos Búsqueda/Actualización

#	Campo de Dimensión	Campo de flujo con el que comparar	Tipo de actualización de dimensión
1	Nombre	Nombre	Insertar
2	Apellido	Apellido	Insertar
3	Teléfono	Teléfono	Insertar

Campo de clave técnica: ClientelD Nuevo nombre

Creación de clave técnica

☒ Utilizar máximo tabla + 1

☐ Utilizar secuencia

☐ Utilizar campo auto-incrementativo

Campo de Versión: version

Campo Fecha flujo:

Campo inicio rango fecha: date_from Año mín.: 1900

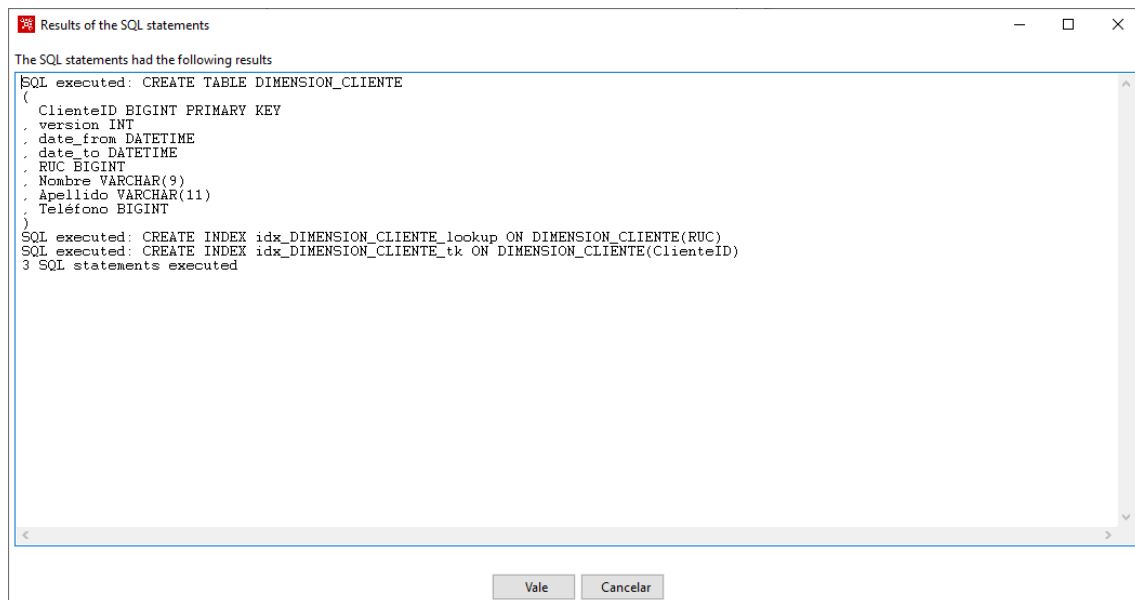
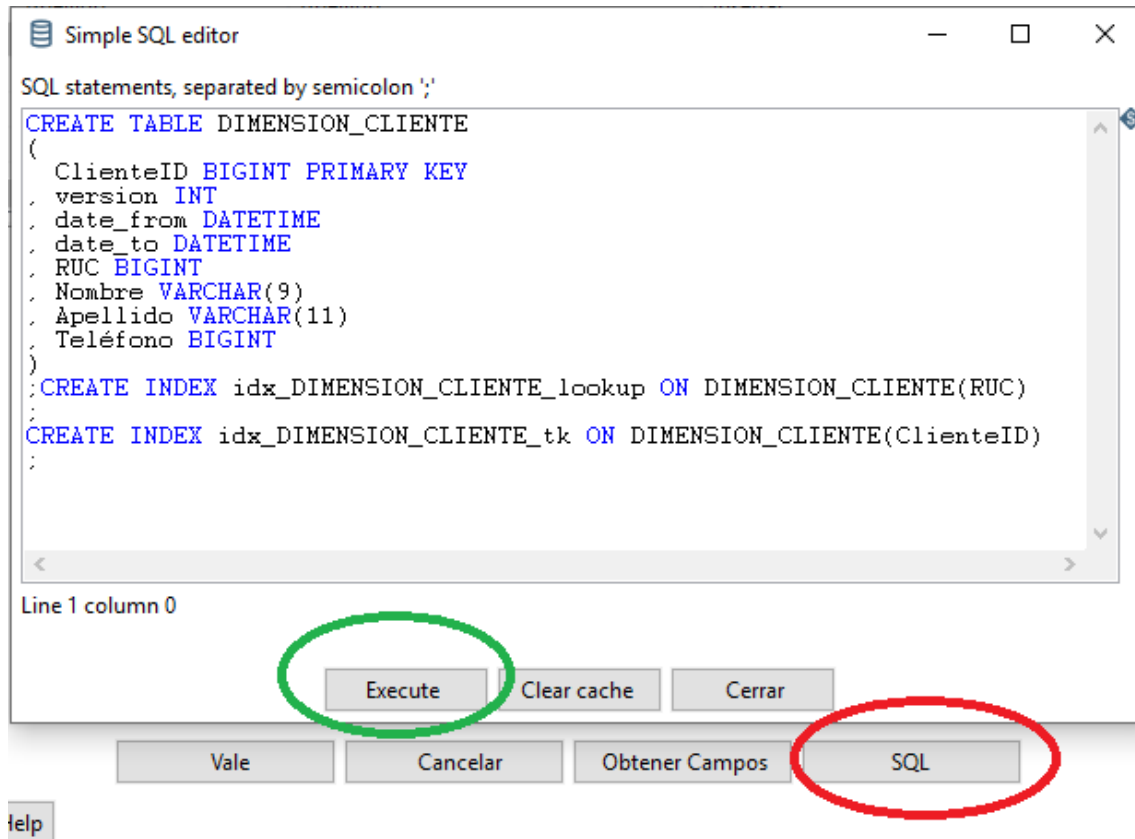
Use an alternative start date? ☐

Campo final rango fecha: date_to Año máx.: 2199

Vale Cancelar Obtener Campos SQL

[Help](#)

Se pulsa en SQL y se ejecuta el código auto generado.





ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Finalmente se creará un esquema en blanco como el siguiente:

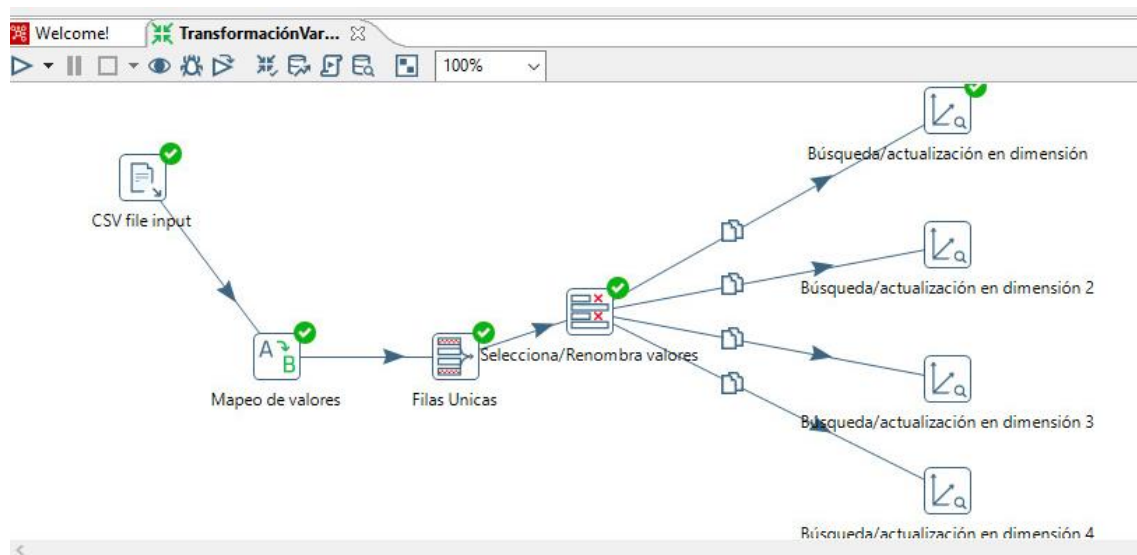
```
/****** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [ClienteID]
, [version]
, [date_from]
, [date_to]
, [RUC]
, [Nombre]
, [Apellido]
, [Teléfono]
FROM [DatamartBI].[dbo].[DIMENSION_CLIENTE]
```

10 %

Results Messages

ClienteID	version	date_from	date_to	RUC	Nombre	Apellido	Teléfono
-----------	---------	-----------	---------	-----	--------	----------	----------

Al ejecutar veremos un resultado como este.



Y en la base de datos se habrá cargado datos especificados para el cliente, la versión indica la actualización realizada sobre esa tupla.

SQLQuery3.sql - DE...administrador (56) SQLQuery2.sql - DE...administrador (55) SQLQuery1.sql - DE...administrador

```

/***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [ClienteID]
, [version]
, [date_from]
, [date_to]
, [RUC]
, [Nombre]
, [Apellido]
, [Teléfono]
FROM [DatamartBI].[dbo].[DIMENSION_CLIENTE]

```

100 %

Results Messages

	ClienteID	version	date_from	date_to	RUC	Nombre	Apellido	Teléfono
1	0	1	NULL	NULL	NULL	NULL	NULL	NULL
2	1	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1720254224001	Cliente1	ApCliente1	2829917
3	2	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1776254224001	Cliente2	ApCliente2	2829237
4	3	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1720904224001	Cliente3	ApCliente3	2829917
5	4	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1744254224001	Cliente4	ApCliente4	2129917
6	5	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1711254224001	Cliente1	ApCliente1	2229917
7	6	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1660254224001	Cliente5	ApCliente5	2119917
8	7	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1798254224001	Cliente6	ApCliente6	2833917
9	8	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1789754224001	Cliente7	ApCliente7	2119917
10	9	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1520254224001	Cliente6	ApCliente6	2877917
11	10	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	820254224001	Cliente8	ApCliente8	2222917

Visualizaciones

Qlik

First Name* Danny

Last Name* Diaz Padilla

Username* magody
Username should be between 8 to 20 characters long

Password* *****

Confirm password* *****

Company* Ninguna

Job Title* Student

Country* Ecuador

State Pichincha

Phone* +593 978654041

Work Email* melodastfery@gmail.com

☒ Email me my account activation link
☐ Send my activation code to: +593978654041

Terms and Conditions*
☒ I agree to abide by the Qlik [Terms and Conditions](#).

Privacy
☐ Yes, I would like to receive email communications related to Qlik including product information and invitation-only social events. I understand that any information I provide will be treated in accordance with the Qlik [Privacy Policy](#).

Para utilizar Qlik debe crearse una cuenta y descargarse la aplicación de escritorio.



ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Se crea entonces una conexión a cualquier base de datos en este caso se elige PostgreSQL.

Create a connection - PostgreSQL

Database properties

Host name
localhost

Port
5432

Name
PostgreSQL_localhost

Cancelar Crear

Entonces se marcará la tabla de hechos y automáticamente traerá todos sus datos incluyendo los indicadores.

Qlik Sense Desktop

Centro de control de Qlik Sense Desktop

Añadir datos a VisualizaciónHechos

Nuevo

EN APP

Entrada manual

UBICACIONES DE ARCHIVO

Mi ordenador

CONEXIONES DE DATOS

PostgreSQL PostgreSQL_localhost

CONTENIDO DE DATOS

Qlik DataMarket

PostgreSQL PostgreSQL_localhost

Owner: public

Filter data

Campos

Vista previa de datos Metadatos

Filtrar campos

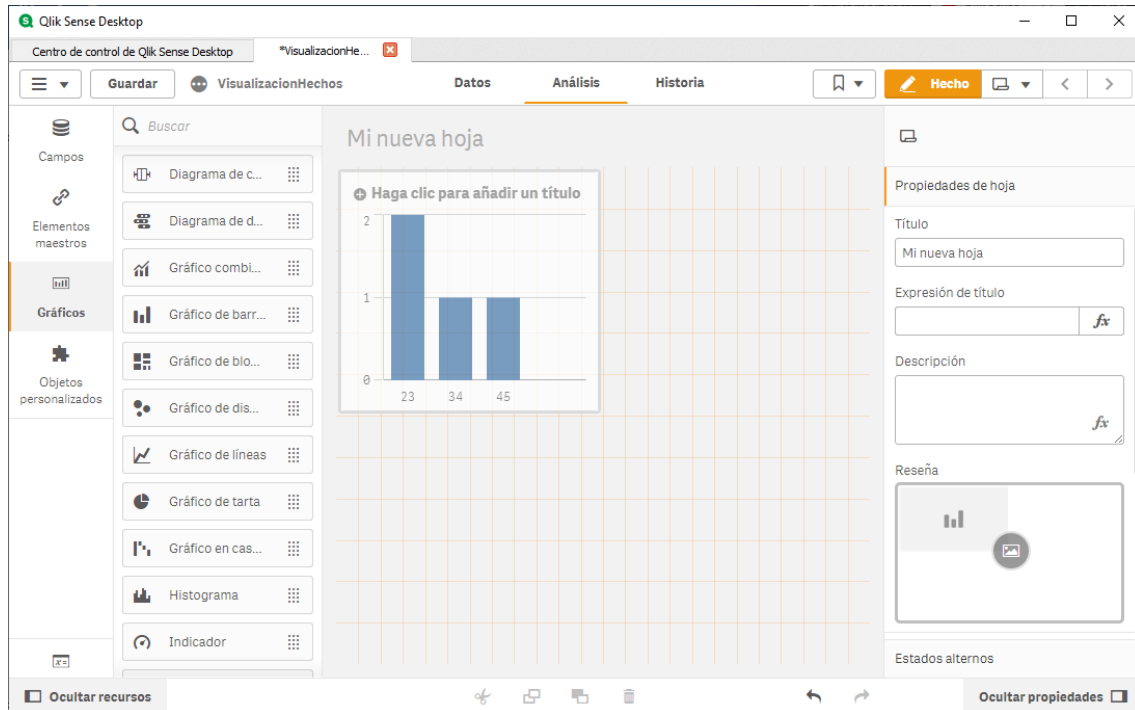
indicador id

23	1
34	2
45	3
23	4

Añadir datos

Para crear visualizaciones se debe seleccionar “Editar la hoja de cálculo” dirigirse a gráficos y arrastrar al cuadro del medio.

Seleccionamos los ejes y graficará los indicadores existentes en la tabla hechos.



4. Análisis de resultados:

La limpieza de datos realizada con pentaho es bastante útil, pero requiere de una planificación sobre lo que se necesita cambiar en las tablas, para obtener un resultado de calidad al final de todo el proceso.

El siguiente error era bastante frecuente

```
org.pentaho.di.core.exception.KettleDatabaseException:
Error occurred while trying to connect to the database

Driver class 'com.microsoft.sqlserver.jdbc.SQLServerDriver' could not be found, make sure the 'MS SQL Server (Native)' driver (jar file) is installed.
com.microsoft.sqlserver.jdbc.SQLServerDriver

at org.pentaho.di.core.database.Database.normalConnect(Database.java:472)
at org.pentaho.di.core.database.Database.connect(Database.java:370)
at org.pentaho.di.core.database.Database.connect(Database.java:341)
at org.pentaho.di.core.database.Database.connect(Database.java:331)
at org.pentaho.di.core.database.DatabaseFactory.getConnectionTestReport(DatabaseFactory.java:80)
at org.pentaho.di.core.database.DatabaseMeta.testConnection(DatabaseMeta.java:370)
```

Y esto ocurre por la incompatibilidad del Java jdk y la versión de los conectores.



ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

5. Conclusiones y recomendaciones:

- Trabajar con dimensiones y una tabla de hechos facilita el trabajo de analítica de datos.
- Pentaho es un software open source.
- Los cuadros de mando permiten tomar decisiones de forma más clara.
- Se recomienda usar una base de datos Postgres para mayor compatibilidad.
- Es recomendable utilizar la infraestructura Apache Hadoop con bases de datos NoSQL. Esto por la eficiencia al momento de realizar consultas.

6. Bibliografía:

- <https://www.itop.es/blog/item/que-es-pentaho-y-cuales-son-sus-beneficios.html>
- <https://forums.pentaho.com/threads/218553-Error-connecting-to-SQL-server-could-not-find-driver/>
- <https://wiki.pentaho.com/pages/viewpage.action?pageId=14844841#id-.01Introducci%C3%B3naSpoon-Instalaci%C3%B3ndeSpoon>
- <https://www.qlik.com/es-es/products/qlik-sense#/es/videos/how-to/sense-product-tour>
- <https://www.youtube.com/watch?v=pmCTSaCntC4>
- <https://www.youtube.com/watch?v=a6nMj6M7IUU>
- <https://www.youtube.com/watch?v=0qGOBFhykek>