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, departmentos 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.

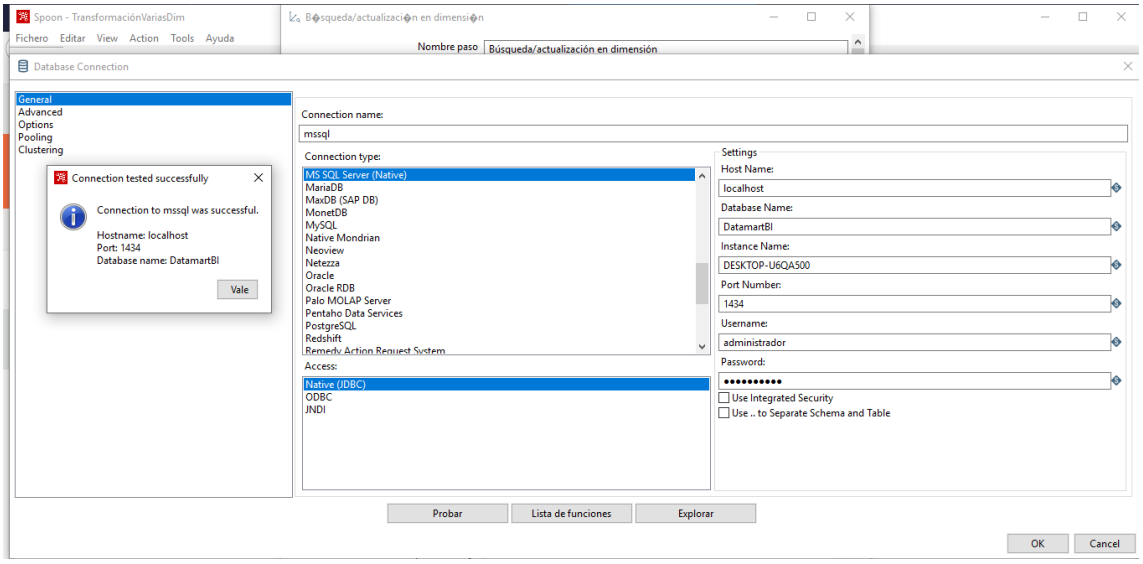
3. Desarrollo de la práctica:

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

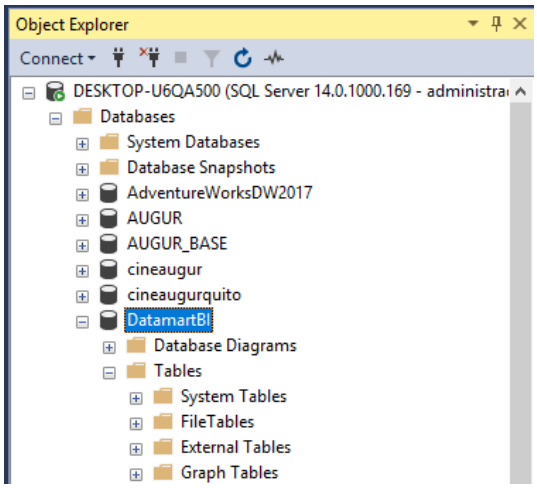
Una vez descargado y descomprimido se procede a ejecutar el archivo Spoon.bat. Como prerequisito el jdk de Java debe haber sido instalado.

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.

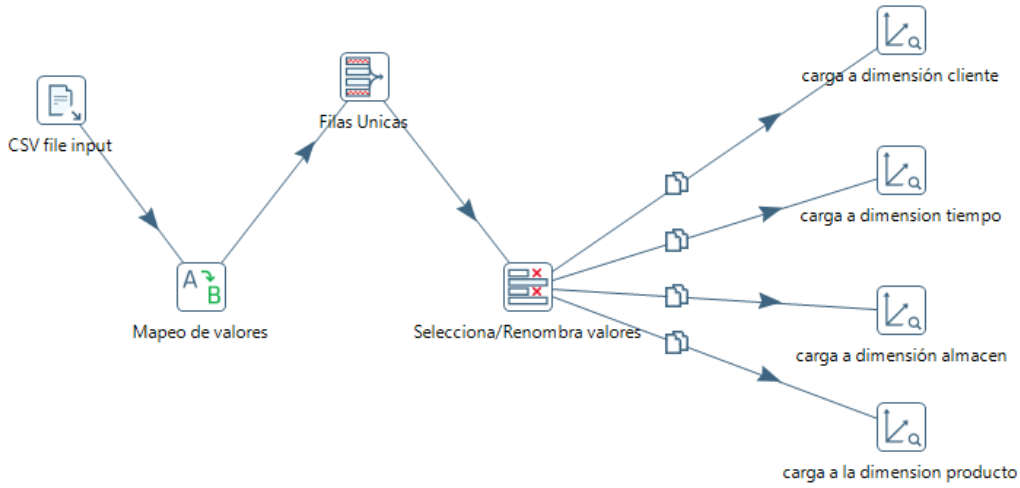


Para el datamart es necesario tener una base de datos y en lo posible un esquema de las dimensiones, en este caso se crearán las dimensiones desde el mismo Spoon.

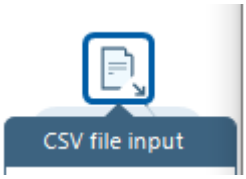


Siguiendo el siguiente tutorial https://www.youtube.com/watch?v=0qGObFhykek se aplica una extracción desde un archivo CSV, se realiza 3 transformaciones y se carga a una dimensión.

Esquema ETL



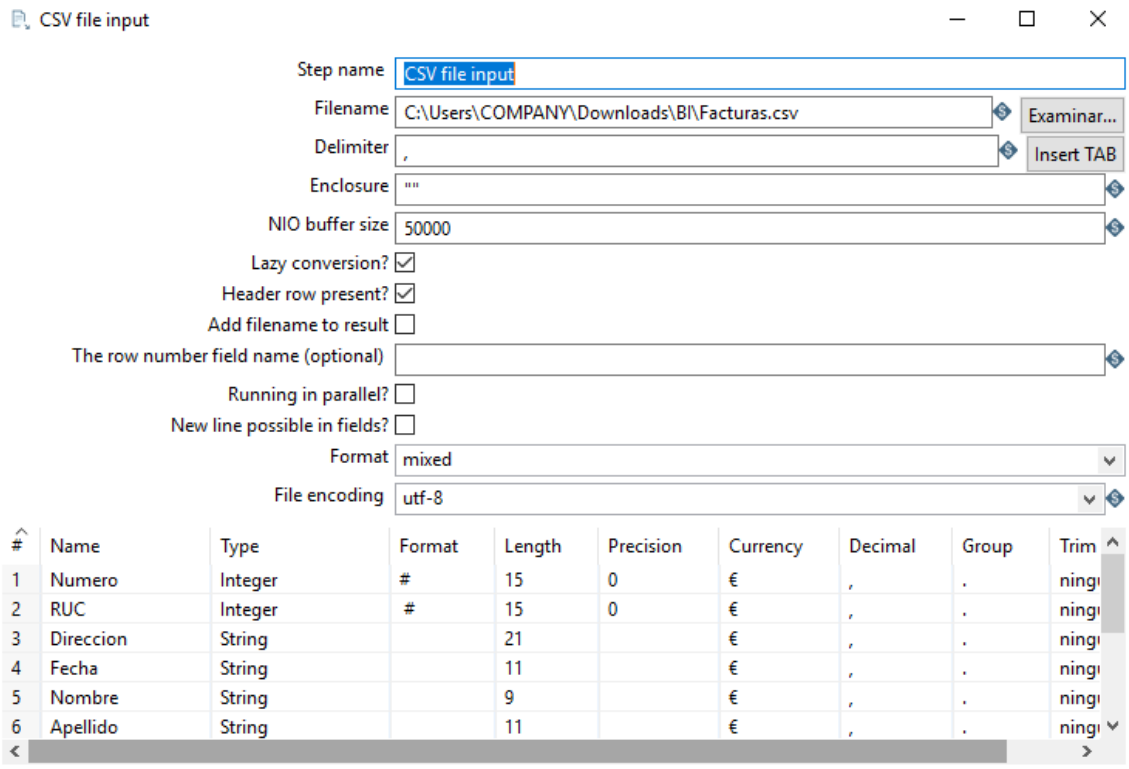
Extracción



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

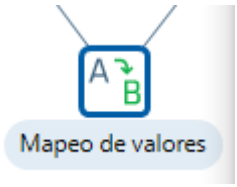
1	Numero, RUC, Direccion, Fecha, Nombre, Apellido, VendedorID, Teléfono, Forma de pago, Total
2	1, 1720254224001, Católica av Machachi, 05/11/2019, Cliente1, ApCliente1, 1, 2829917, E, 27.45
3	2, 1776254224001, Católica av Machachi, 01/11/2019,Cliente2, ApCliente2, 1, 2829237, E, 56.25
4	3, 1720904224001, 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

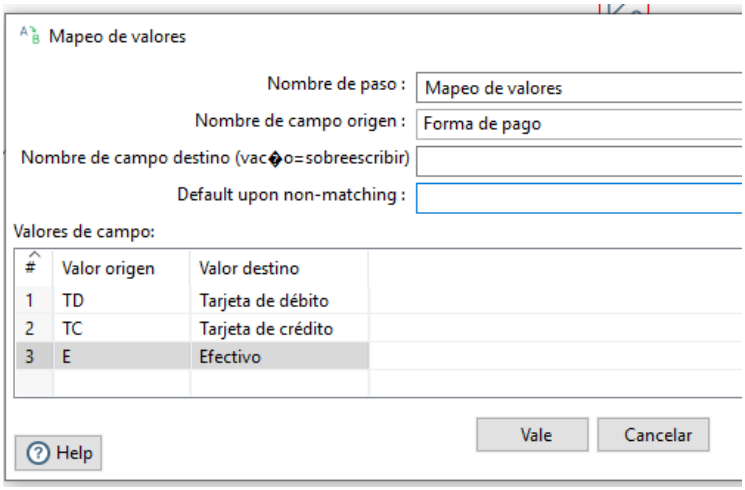
En el paso de extracción se selecciona el directorio y actomáticamente traerá los campos.



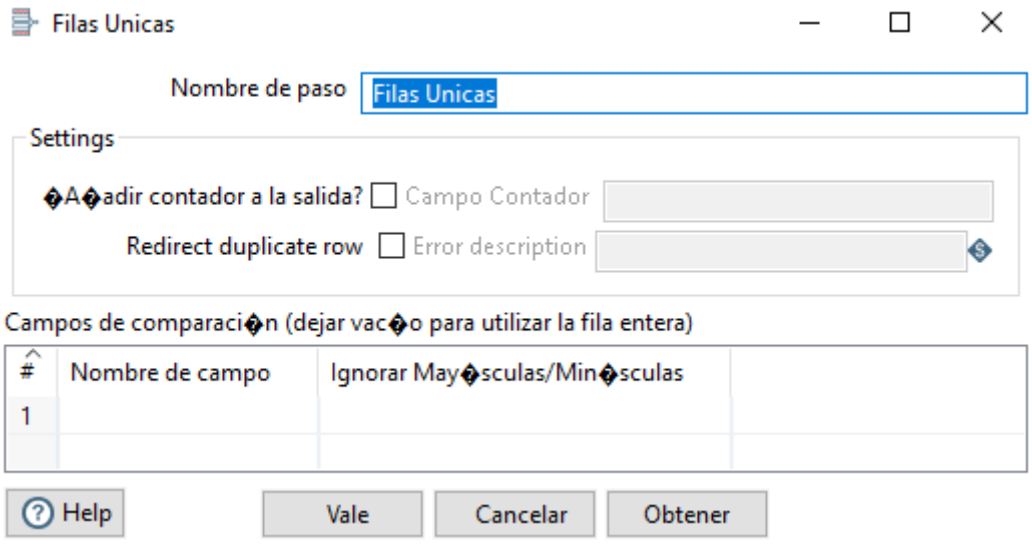
Transformación

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

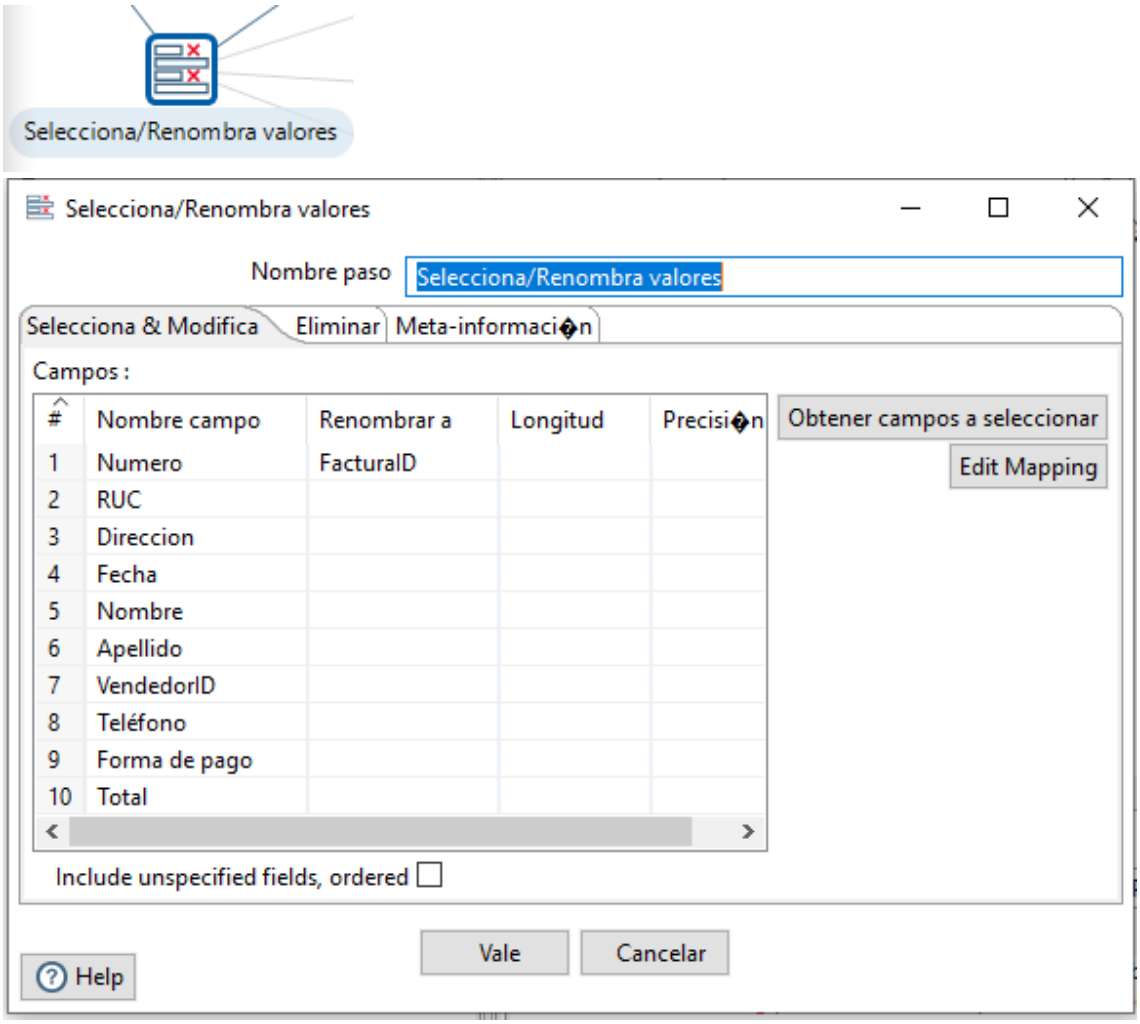




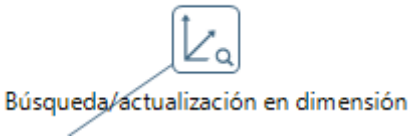
La segunda transformación se encarga de eliminar duplicados.



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.

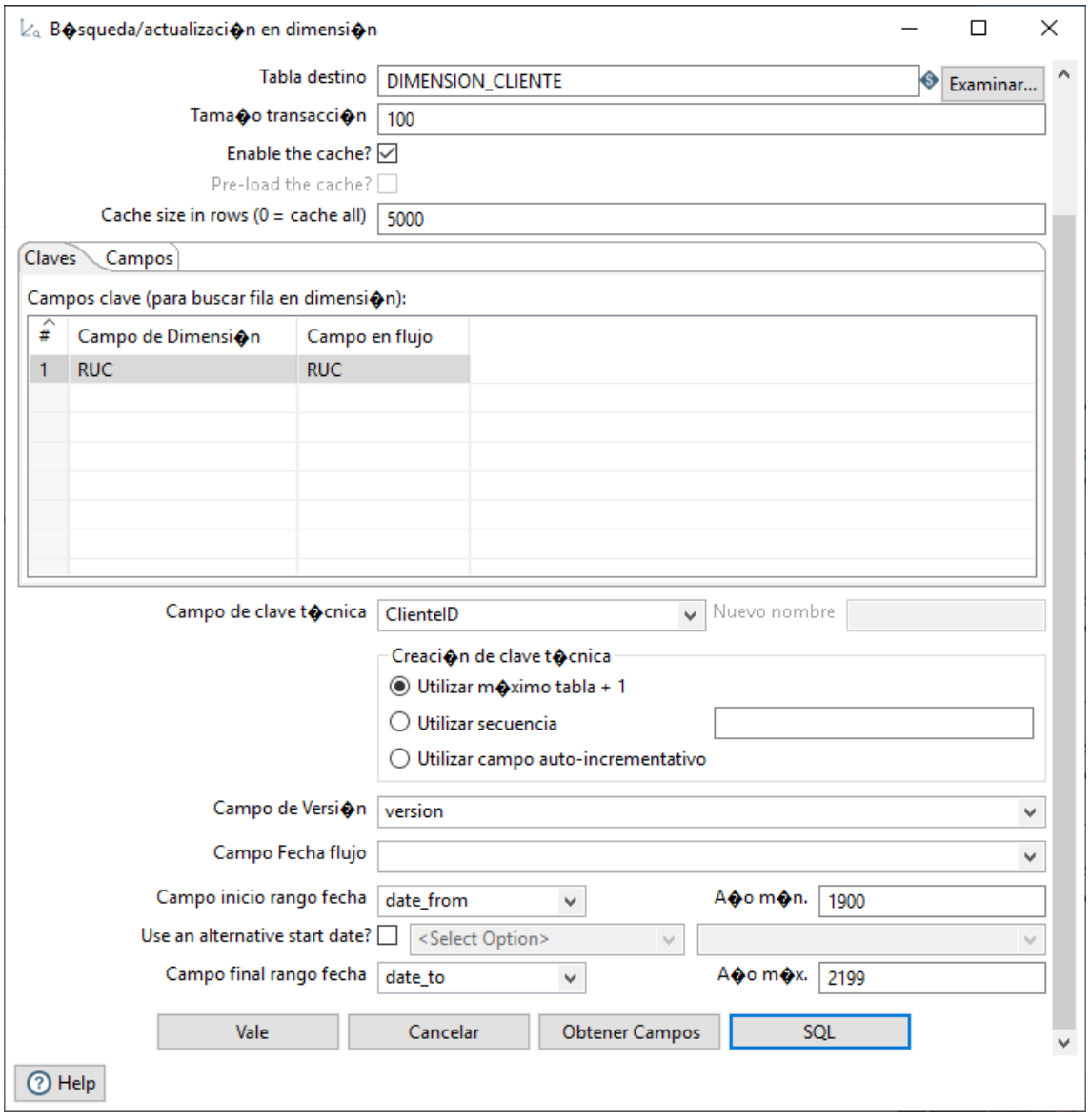


Carga

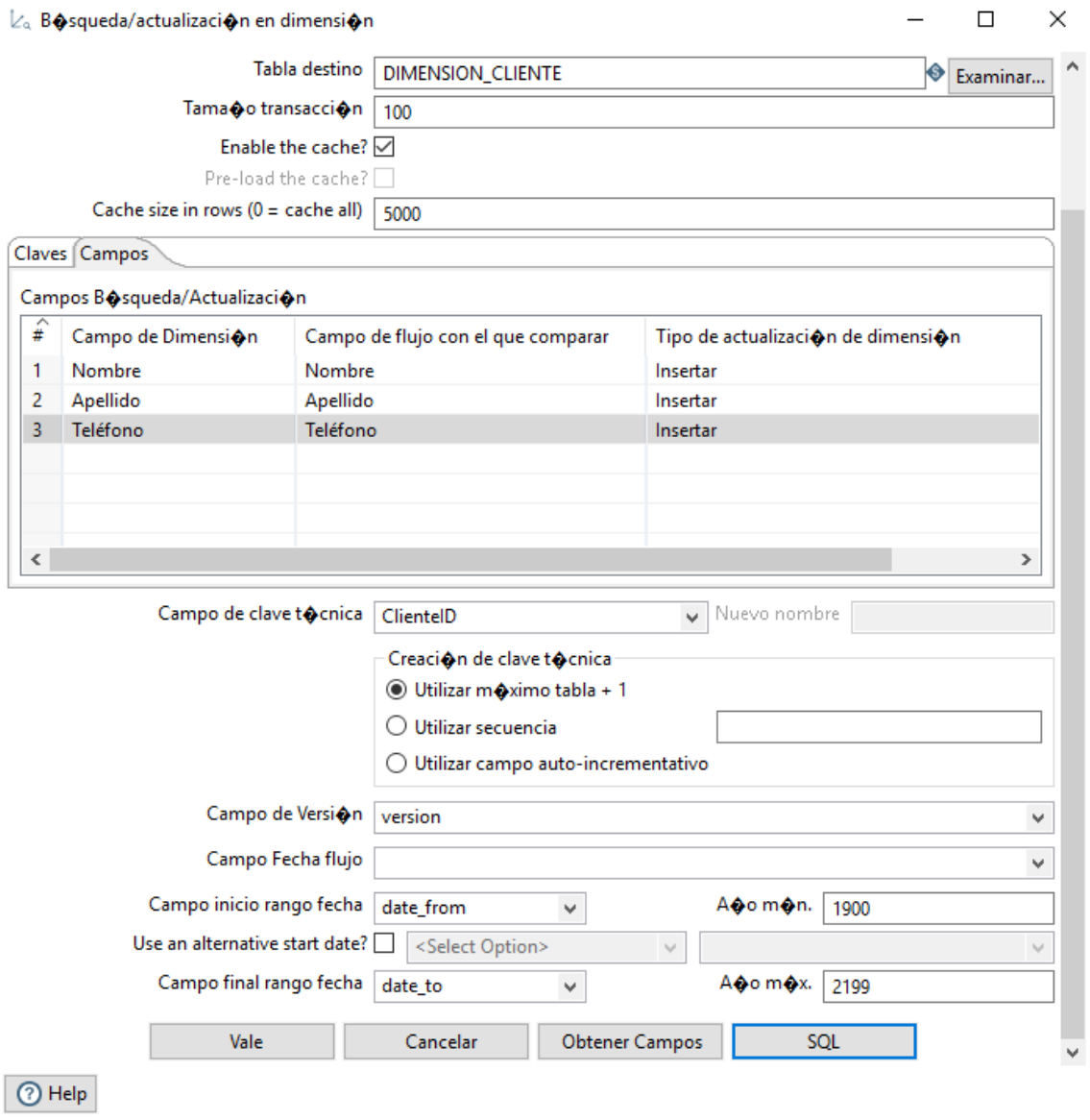


DIMENSION_CLIENTE

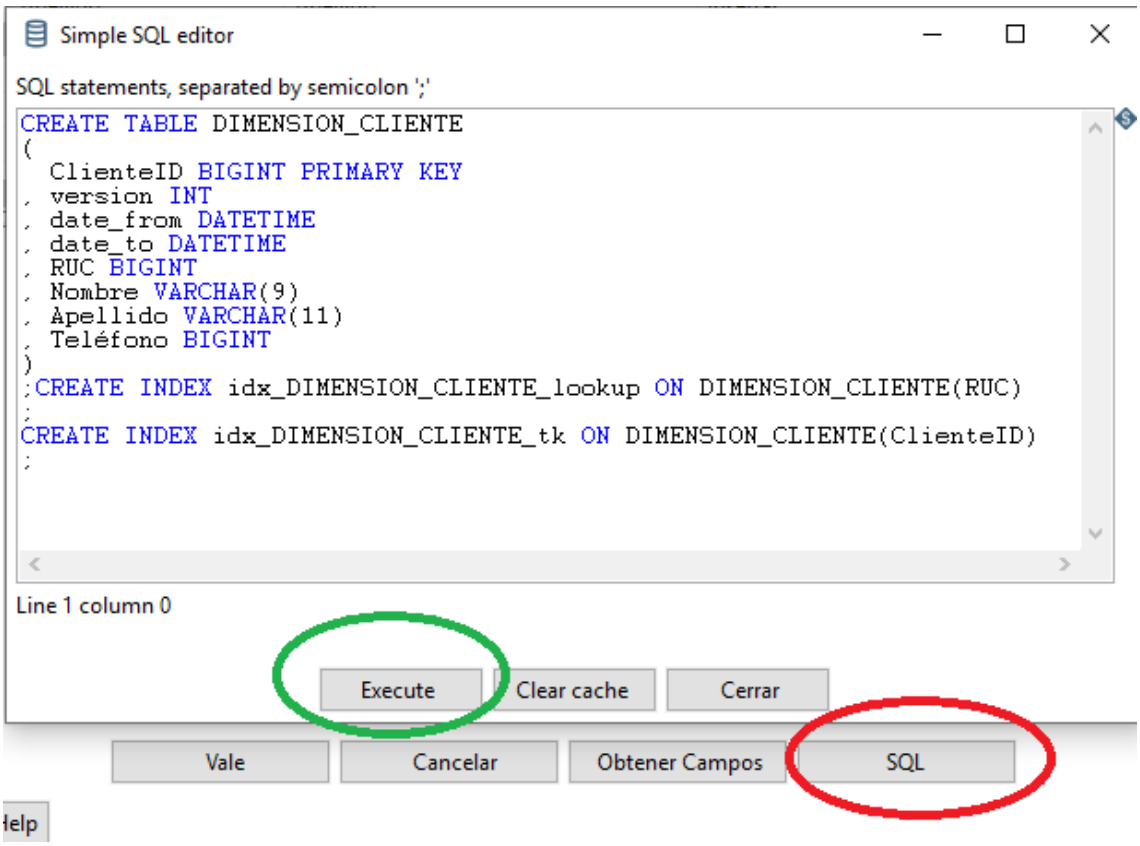
Esta tabla será creado con la llave primaria igual al RUC almacenado en la factura.

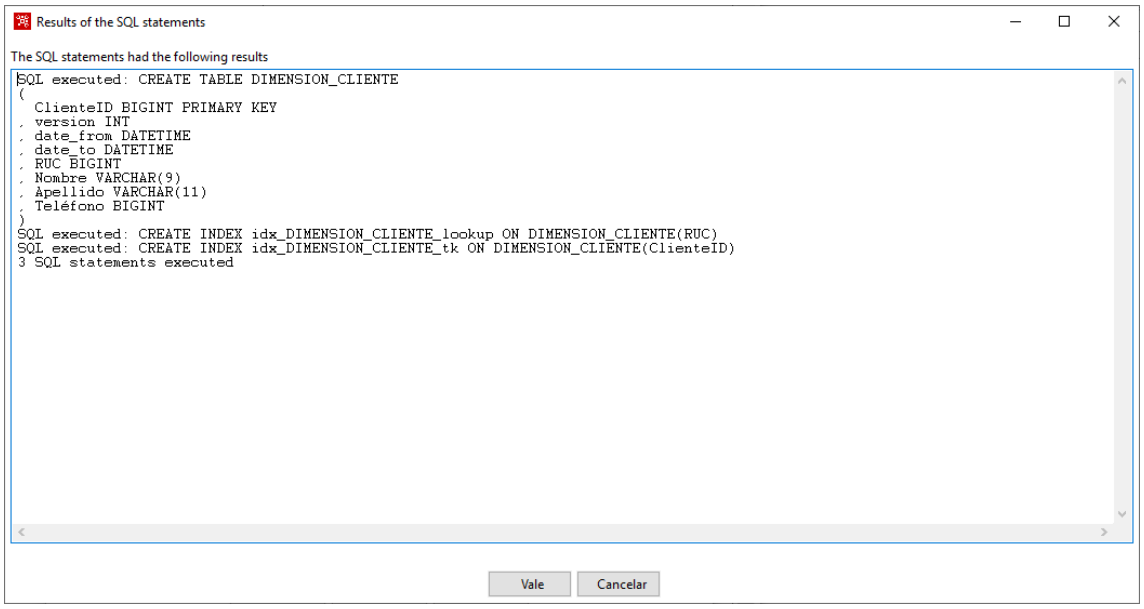


Y los campos solo se tomarán el nombre, apellido y teléfono, los demás se los dejará a lado para las otras dimensiones.



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





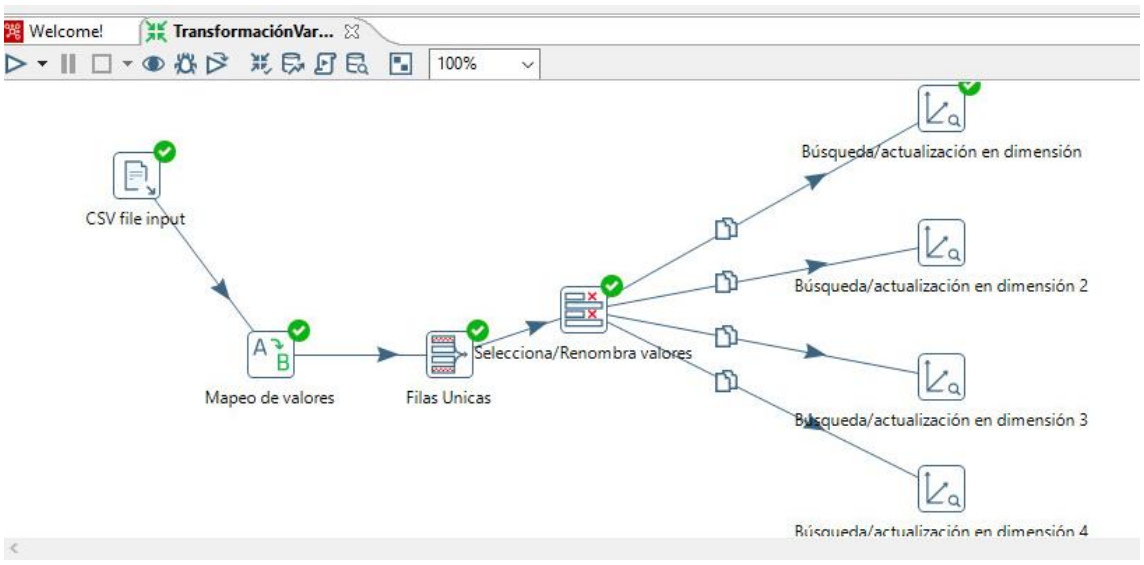
Finalmente se creará un esquema en blanco como el siguiente:

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

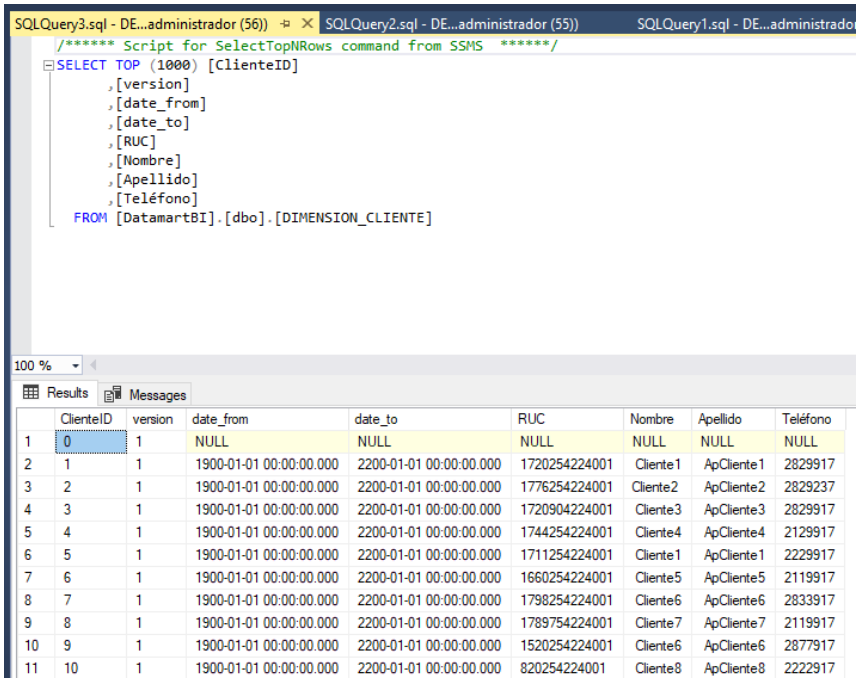
Results | Messages

ClienteID	version	date_from	date_to	RUC	Nombre	Apellido	Teléfono

Al ejecutar veremos un resultado como este.

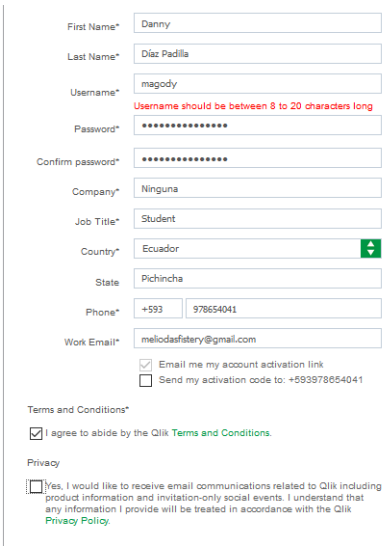


Y en la base de datos se habrá cargado datos especificados para el cliente, la versión indica la actualización realizada sobre esa tupla.



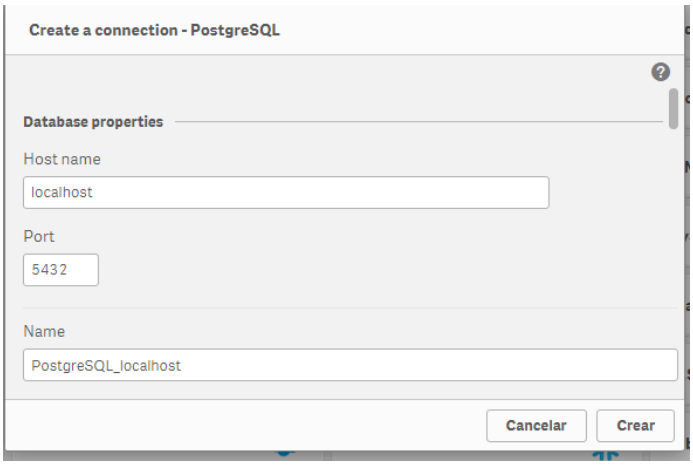
	ClienteID	version	date_from	date_to	RUC	Nombre	Apellido	Teléfono
1	0	1	NULL	NULL	NULL	NULL	NULL	NULL
2	1	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1720254224001	Cliente1	ApCliente1	2829917
3	2	1	1900-01-01 00:00:00.000	2200-01-01 00:00:00.000	1776254224001	Cliente2	ApCliente2	2829237
4	3	1	1900-01-01 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

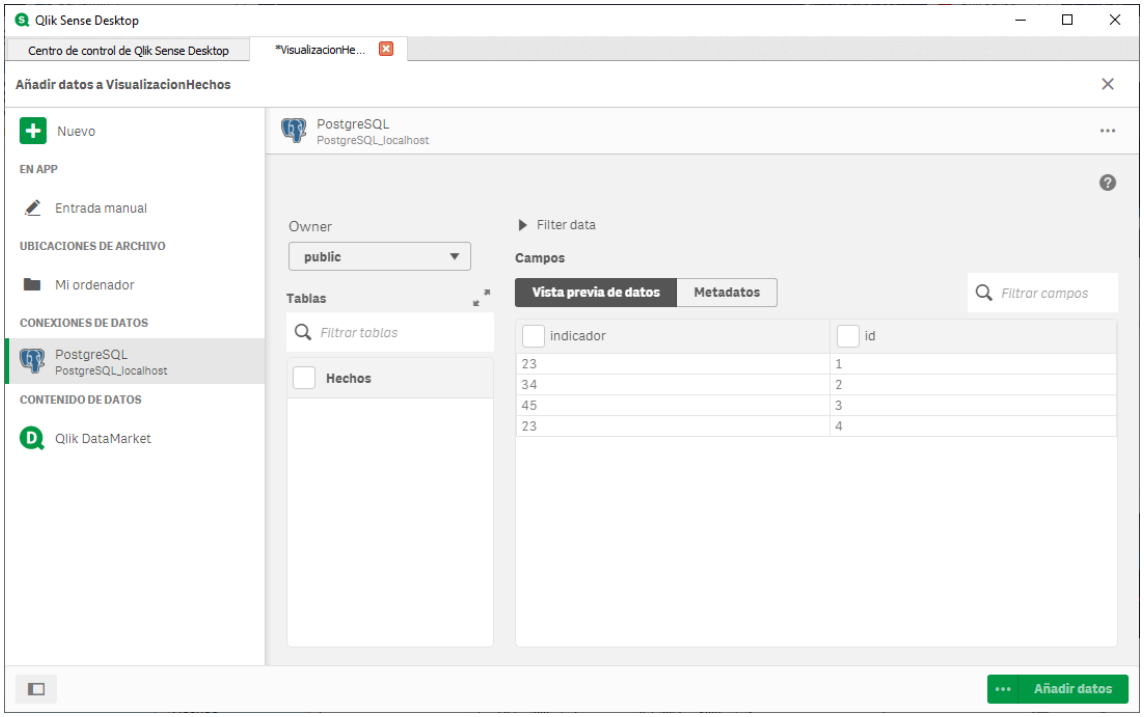


Para utilizar Qlik debe crearse una cuenta y descargarse la aplicación de escritorio.

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

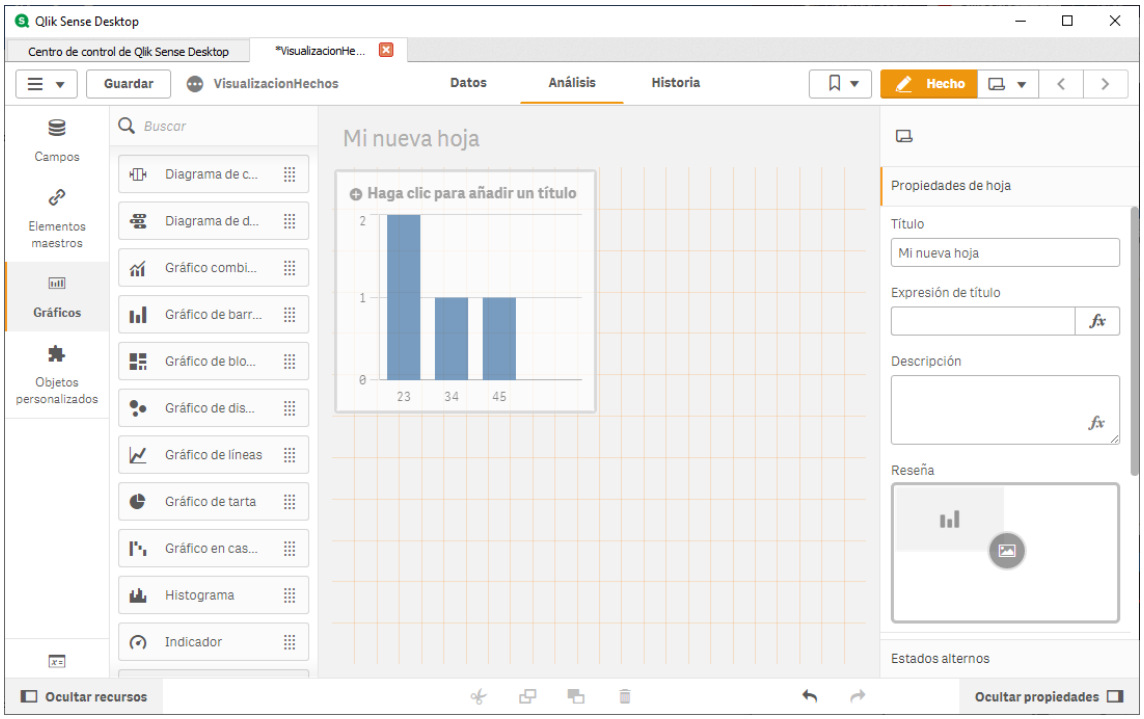


Entonces se marcará la tabla de hechos y automáticamente traerá todos sus datos incluyendo los indicadores.



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)
```

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

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 usa 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