

Introducción a KNIME



1 . KNIME

¿QUÉ ES KNIME?

[KNIME](#) es una plataforma open source para analítica de Business Intelligence, Machine Learning y ETL mediante un simple proceso de drag and drop (arrastrar y soltar). KNIME proporciona una plataforma con una interfaz gráfica de usuario donde se pueden crear flujos de trabajo rápidamente por individuos sin demasiada experiencia técnica para poder analizar los datos.

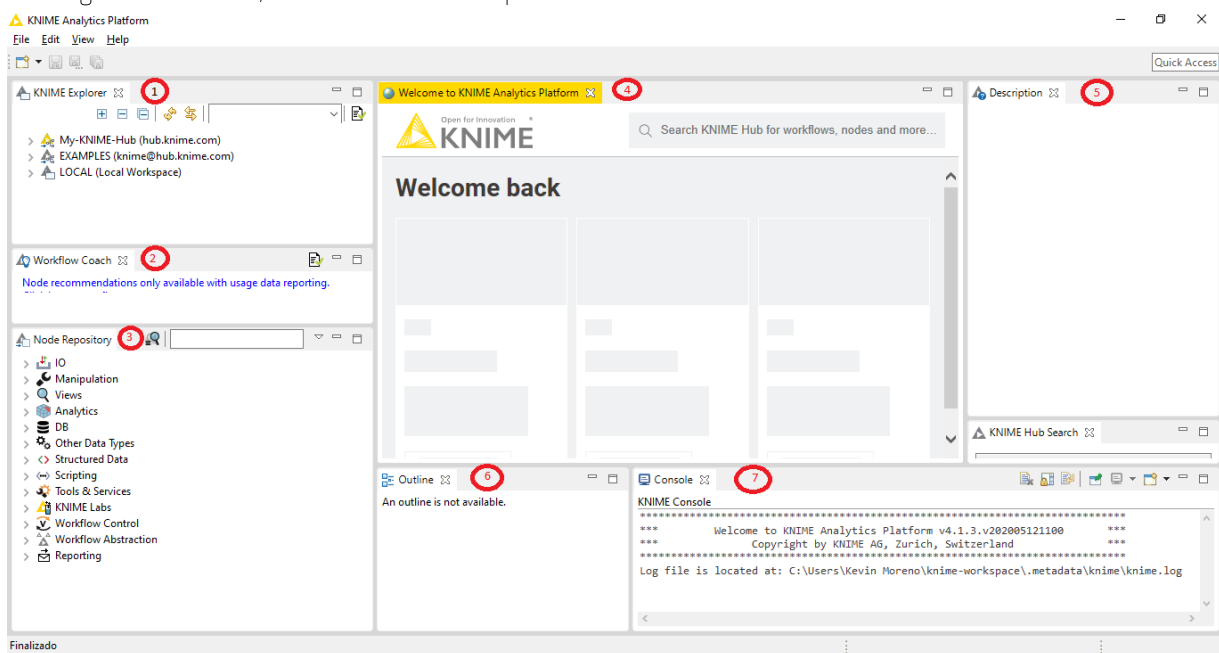
La plataforma de KNIME puede usarse para lo siguiente:

- Procesos ETL.
- Machine Learning de manera sencilla.
- Pueden incluirse modelos de Deep learning.
- Cálculo de analíticas potentes sobre los datos.
- Permite utilizar diferentes tipos de datos como series temporales, imágenes, textos...

KNIME también dispone de una versión de pago. Esta versión facilita el desarrollo conjunto y la organización dentro de una empresa, pero no tiene funcionalidades de análisis adicionales. Todo el análisis que se pueda hacer en la versión de pago, puede hacerse en la versión gratuita.

PRIMERA VISTA DE KNIME

Para descargar KNIME debemos ir a <https://www.knime.com/downloads> e introducir el e-mail. Una vez descargado e instalado, al abrirlo vemos esta pantalla.



1. **KNIME explorer:** Aquí encontramos los archivos del proyecto en el que estamos trabajando, además de algunos ejemplos de KNIME.
2. **Workflow coach:** En esta ventana nos aparecerán sugerencias de nodos mientras estemos trabajando, sugeridas a partir de estadísticas de uso de la comunidad.
3. **Node repository:** Aquí encontraremos todos los nodos. Están organizados en categorías y subcategorías y además se puede buscar un nodo por nombre.
4. **Welcome page / workflow view:** Al abrir KNIME aparecerá la Welcome Page con diversos recursos para crear rápidamente un nuevo proyecto y tutoriales o información de interés, en especial orientados a usuarios nuevos. Tras cerrarla o tras abrir un proyecto, esta será la vista del Workflow con cada uno de sus nodos, el espacio de trabajo.
5. **Description:** Aquí aparecerá información descriptiva, útil para ver la descripción de uso de los nodos.
6. **Outline:** Cuando el workflow aumente de tamaño se puede navegar en esta vista.
7. **Console:** Consola donde aparecen registros y errores de la ejecución del workflow.

Para crear un nuevo proyecto podemos hacerlo desde la Welcome Page o desde File -> New. La vista de KNIME puede configurarse en View, para ocultar o mostrar las diferentes ventanas.

2 . EJEMPLOS DE USO

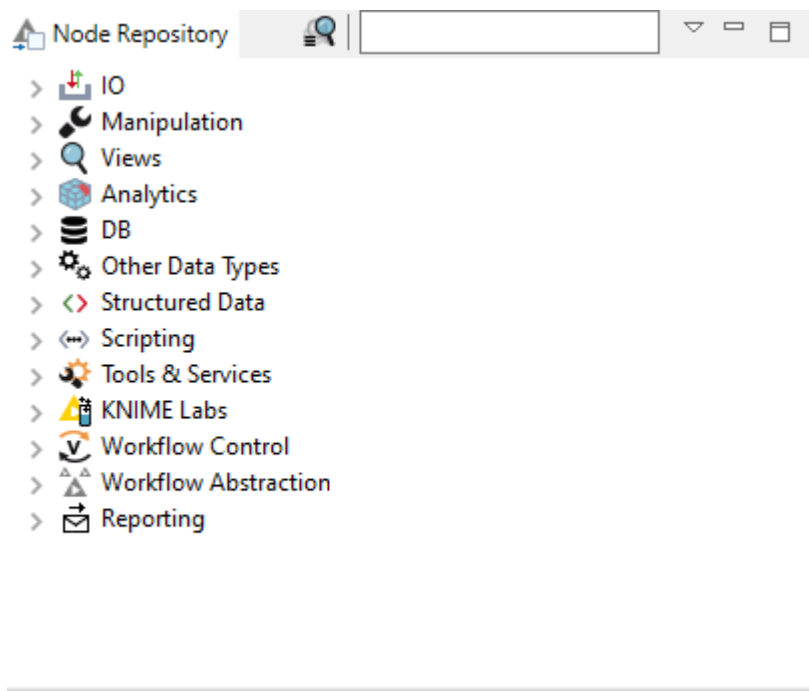
NODOS Y WORKFLOW

En el apartado anterior quizá hayas visto varias veces las palabras “nodo” y “workflow”, pero, ¿qué son exactamente?

Un nodo es la unidad de procesamiento más pequeña con la que trabaja KNIME. Es equivalente a un paso en Pentaho Data Integration, es una operación concreta, desde lectura de archivos, filtrado de filas hasta predicción de un modelo.

Un Workflow es una secuencia de nodos. Los nodos pueden conectarse entre sí mediante sus inputs y sus outputs. Es el análogo a transformación o job en Pentaho.

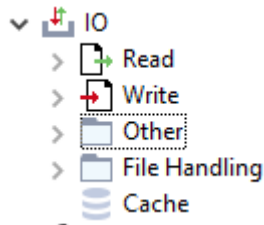
Tipos de nodos:



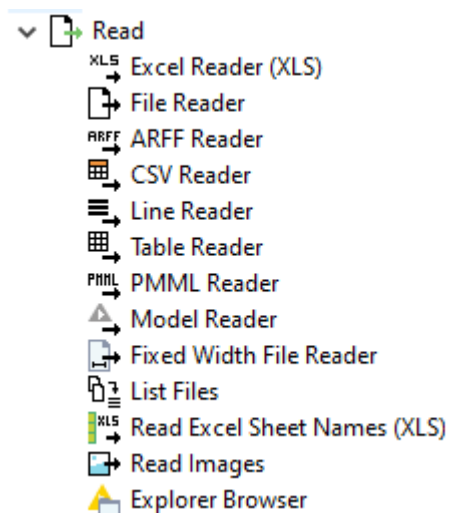
Los nodos los encontramos en el Node Repository. Se pueden ver las principales categorías, aunque cada una puede dividirse a su vez en subcategorías. Podemos navegar por ellas haciendo click en las flechas para desplegarlas, o bien podemos hacer una búsqueda. Si hacemos click sobre la lupa al lado del cuadro de búsqueda, vemos que se pone con un fondo azul. Este fondo azul significa que hace una búsqueda aproximada, y devuelve resultados que contengan las palabras de búsqueda o puedan estar relacionados con ellos.

NODOS PRINCIPALES DE INTERÉS

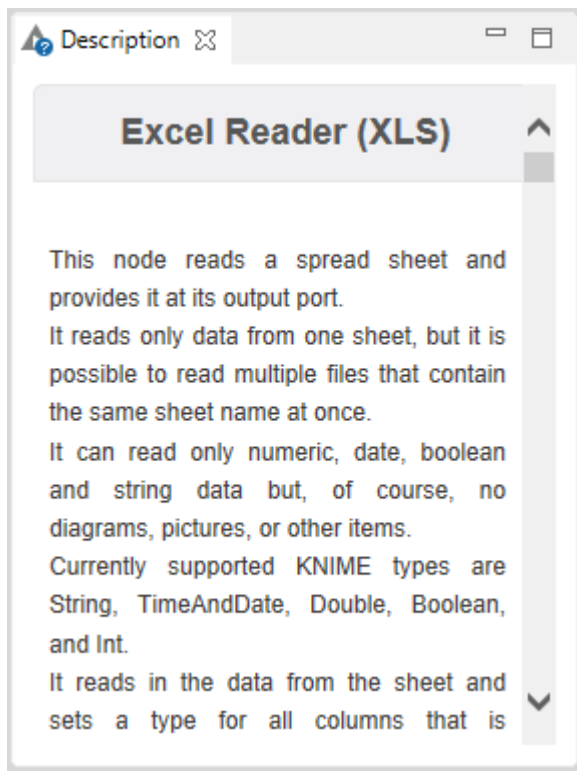
IO (Input-Output):



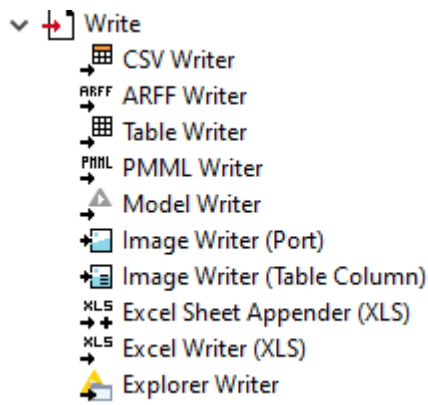
Para poder hacer uso de los datos, primero debemos importarlos a la plataforma de KNIME. En la categoría IO podemos ver varias subcategorías, entre ellas Read (lectura) y Write (escritura). En Read encontramos todas las funcionalidades que proporciona KNIME respecto a la lectura de archivos, en concreto vemos que podemos leer desde CSV y desde archivo Excel.



Si buscamos más información acerca del funcionamiento de un nodo concreto, basta con hacerle click e ir a la pestaña de Description.

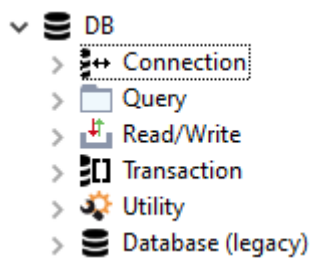


En la subcategoría Write podemos ver las opciones de KNIME como salida de ficheros.

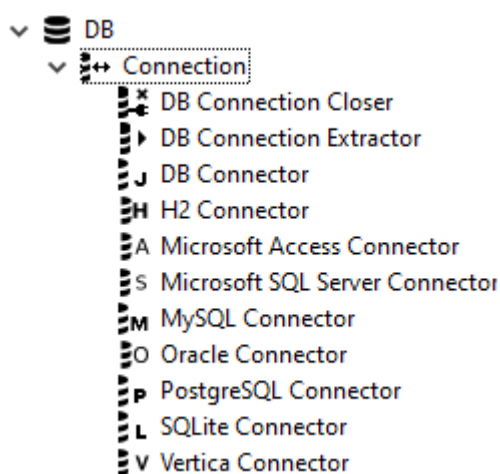


DB (base de datos):

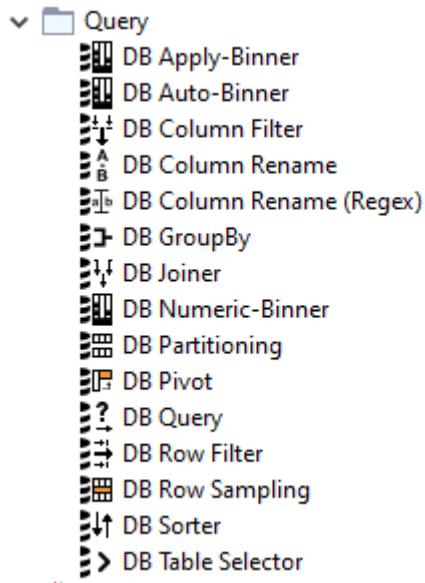
Para entrada y salida de datos en una base de datos, debemos mirar la categoría DB.



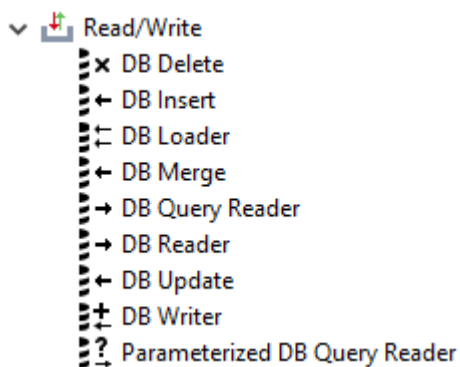
En Connection podemos configurar la conexión a la base de datos y ver los tipos de base de datos que soporta KNIME. Tenemos soporte de Vertica, MySQL, Oracle, SQL Server... Según el tipo de base de datos que se use, será necesario disponer del driver JDBC.



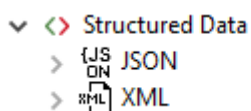
En Query disponemos de muchas funcionalidades de bases de datos para hacer una consulta específica y obtener los datos que queramos. Tenemos DB Query para crear una consulta, DB Row Filter para filtrar, DB GroupBy, DB Joiner...



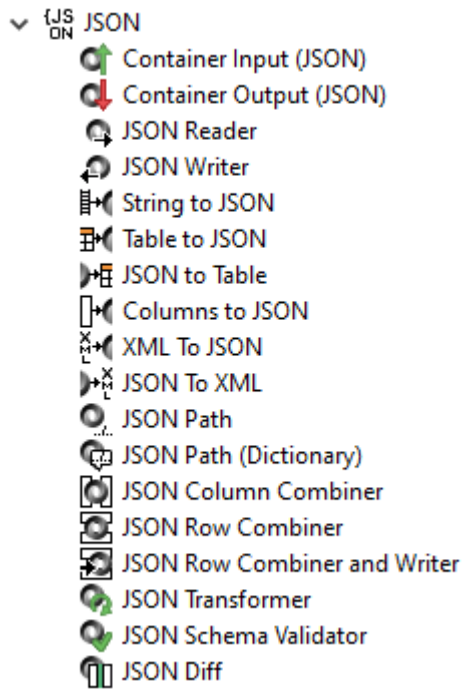
En Read/Write también tenemos opciones para ejecutar comandos usuales de base de datos, DB Delete, DB Merge, DB Insert, DB Query Reader...



Structured Data:

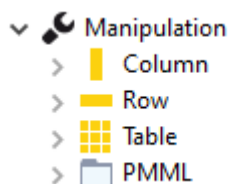


Para datos estructurados como JSON y XML, KNIME tiene diversas funcionalidades adicionales. Podemos ver, por ejemplo, en JSON las siguientes: JSON Reader, JSON Path, XML to JSON...



Manipulation:

Una vez tenemos los datos en la plataforma, el siguiente paso es transformarlos para limpiarlos y quedarnos con los necesarios.



En Manipulation vemos operaciones de manipulación de Columnas y Filas. En Filas hay operaciones de filtrado, de extracción de cabecera, muestreo (para modelos de Machine Learning), agrupado, separación...

En Columnas tenemos operaciones de casteo de tipo de dato, separar o combinar columnas, normalizado...

KNIME tiene muchas posibilidades y vale la pena echar un vistazo junto con la descripción de cada nodo para ver el alcance de la aplicación. Prácticamente todo lo que se desee hacer en un proceso ETL se puede conseguir mediante nodos de KNIME.

Workflow Control

- Workflow Control
 - Automation
 - Variables
 - Loop Support
 - Switches
 - Error Handling
 - Meta Nodes

KNIME también ofrece posibilidad de controlar el flujo de datos. Bucles (Loop Support), condicionales, tratamiento de errores...

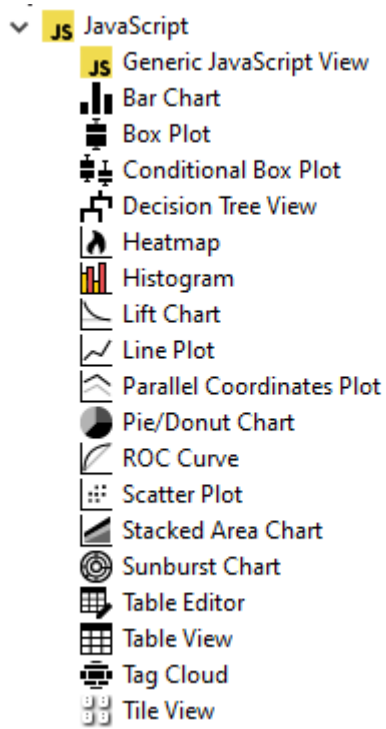
- Switches
 - IF Switch
 - End IF
 - CASE Switch Data (Start)
 - CASE Switch Data (End)
 - CASE Switch Model (Start)
 - CASE Switch Model (End)
 - CASE Switch Variable (Start)
 - CASE Switch Variable (End)
 - Empty Table Switch
 - Java IF (Table)

Views

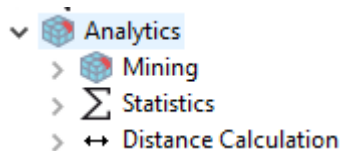
KNIME ofrece varias opciones de visualización de gráficos, aunque el consenso es utilizar una plataforma externa con tal de obtener las mejores visualizaciones, como PowerBI, una vez se han tratado los datos.

- Views
 - JavaScript
 - Property
 - Utility
 - Visualization Column Appender
 - Local (Swing)

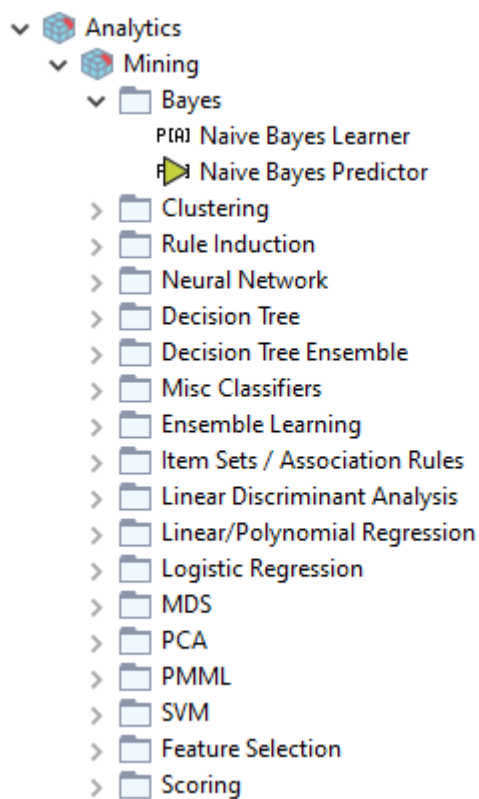
Desplegando JavaScript vemos los tipos de gráficos.



Analytics

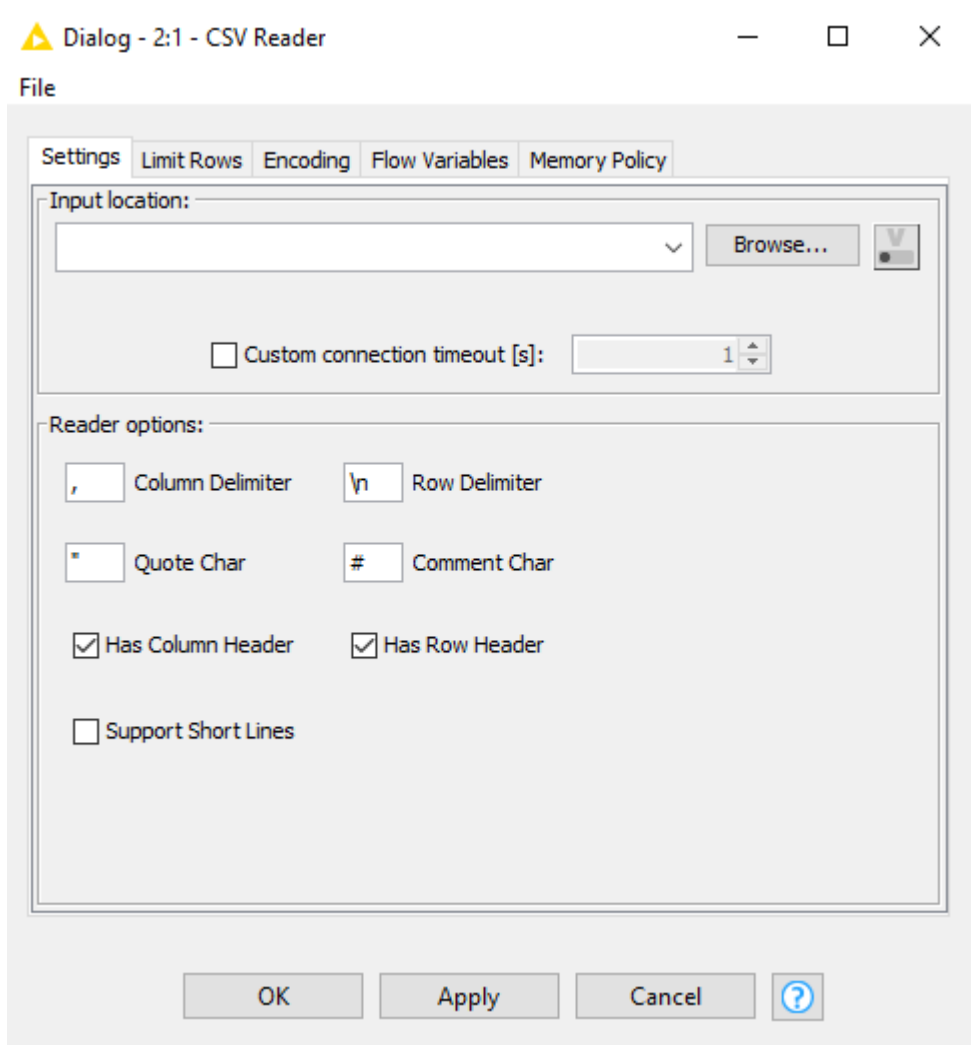


Tenemos la posibilidad de calcular estadísticas potentes o de aplicar procesos de minería de datos. Si hacemos click en Mining, vemos las posibilidades que KNIME ofrece. Por ejemplo, podemos aplicar un clasificador Naive Bayes mediante dos nodos, uno para entrenar el modelo y otro para predecir, sin necesidad de escribir ningún tipo de código.

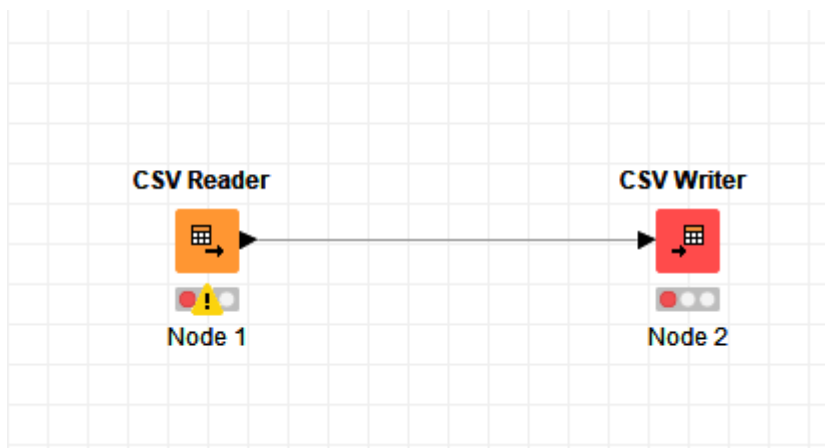


CREACIÓN DE UN WORKFLOW

El funcionamiento de KNIME es sencillo, escogemos el nodo que queramos utilizar y lo arrastramos a la ventana de nuestro workflow. Una vez ahí podemos hacer doble click para configurar el nodo, por ejemplo en un nodo de lectura de CSV tenemos esta ventana, que podemos configurar las opciones de lectura del archivo.






Para conectar dos nodos basta con hacer click en el primer nodo y arrastrar hasta el segundo.



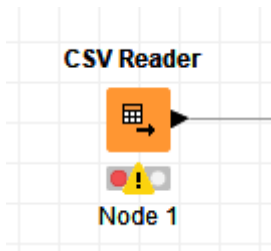
El semáforo de debajo de los nodos indica el estado del nodo. La luz roja indica la falta de configuración, la luz amarilla indica que el nodo está configurado, pero aún no se ha ejecutado. La luz verde indica que se ha ejecutado correctamente y pueden salir además un triángulo con una exclamación indicando que es necesaria la configuración y un símbolo rojo con una cruz indicando que ha habido errores en la ejecución.

En la barra



Vemos las opciones para ejecutar el workflow. Podemos ejecutar los nodos seleccionados con  o ejecutarlos todos . Con  cancelamos los seleccionados o todos.

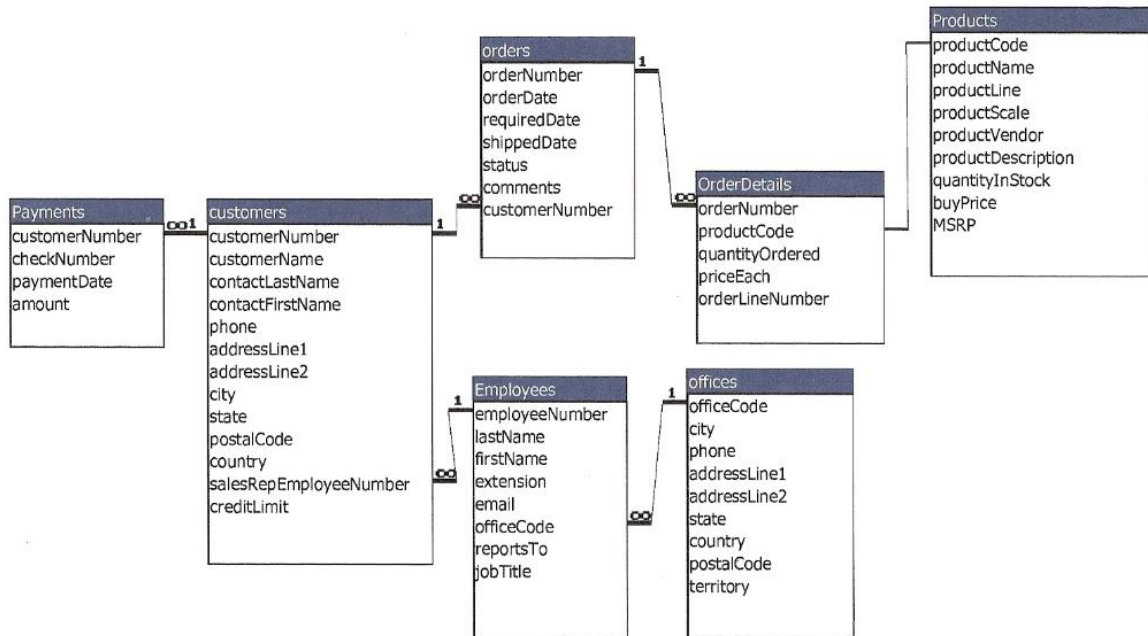
También podemos modificar la descripción de cada nodo haciendo click en "Node 1".



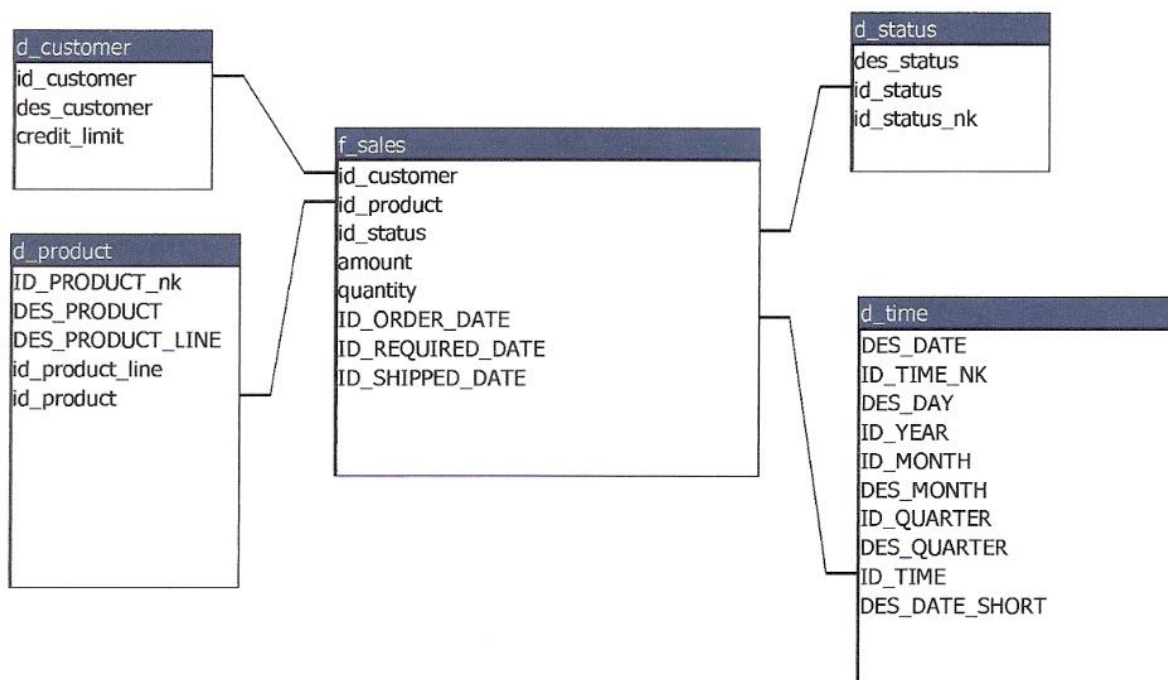
ETL CON KNIME

Vamos a crear un proceso ETL sencillo con KNIME. Partimos de las siguientes tablas de ejemplo.

ClassicModels

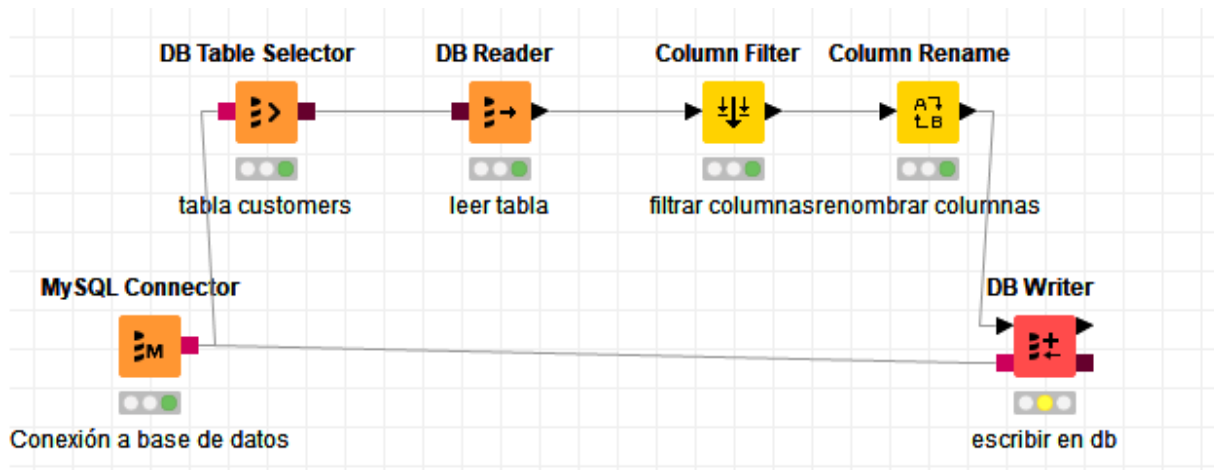


El objetivo es crear un modelo en estrella con la siguiente información:

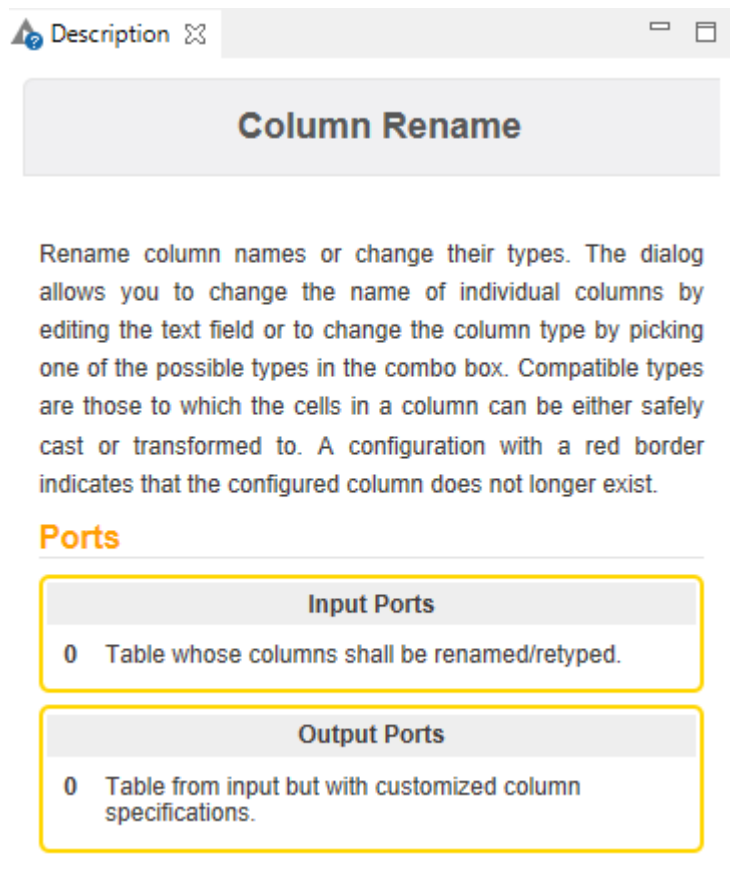


Dimensión customers:

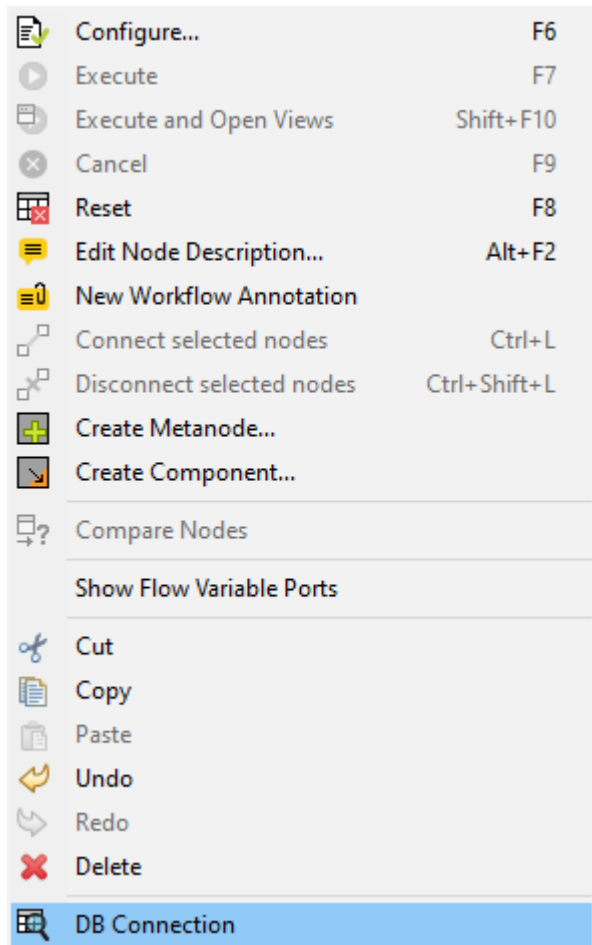
Primero, para la construcción de la dimensión 'customers', tenemos el siguiente Workflow.



Un aspecto a destacar es el tipo de conexiones de entrada y salida de cada nodo. Por ejemplo, el conector MySQL tiene una conexión simbolizada por un cuadrado rojo, el nodo DB Reader tiene una conexión de entrada simbolizada con un cuadrado negro y una de salida con un triángulo. Estos símbolos representan el tipo de dato o metadato de entrada y salida y además proporcionan una guía visual sobre qué conexiones entre nodos son posibles. En el cuadro de Descripciones podemos verlo mejor, en el apartado de Ports.



Si hacemos click derecho en un nodo, se despliega un menú contextual. En la opción inferior podemos ver la salida de cada nodo una vez se haya ejecutado. Esto nos permite ver fácilmente el estado del flujo de datos en cada punto del Workflow.

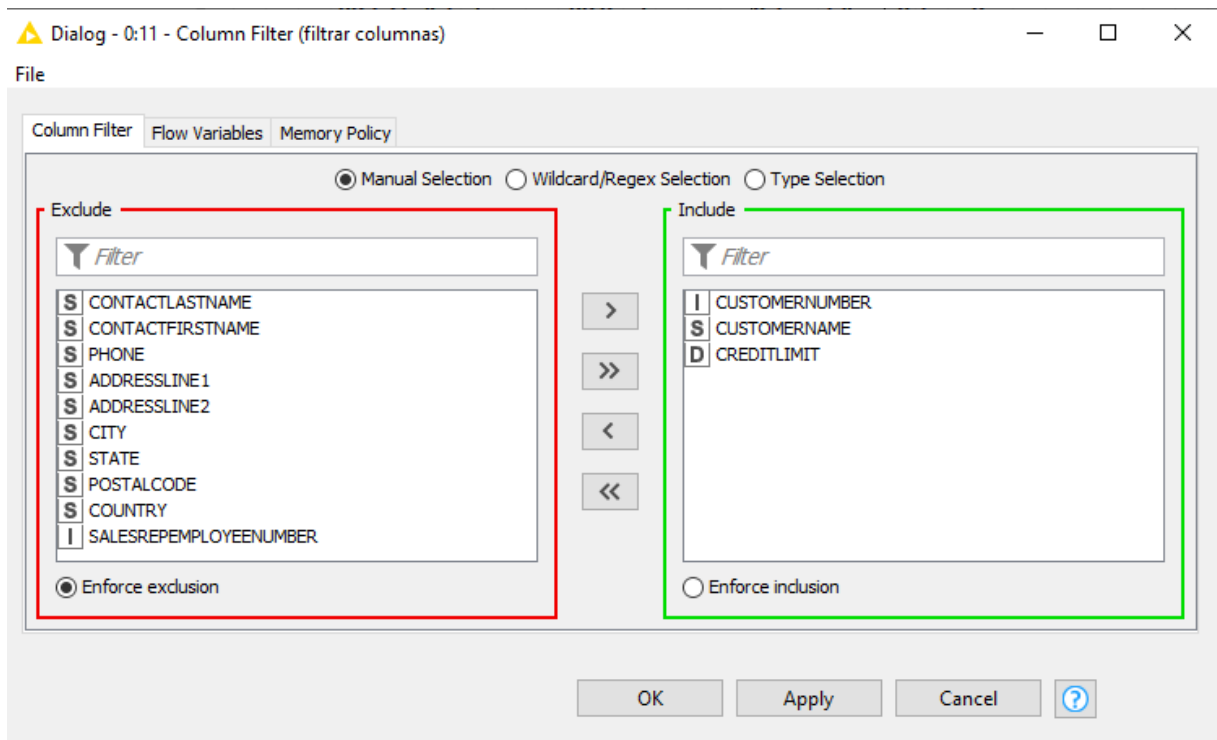


MySQL Connector + DB Table Selector + DB Reader:

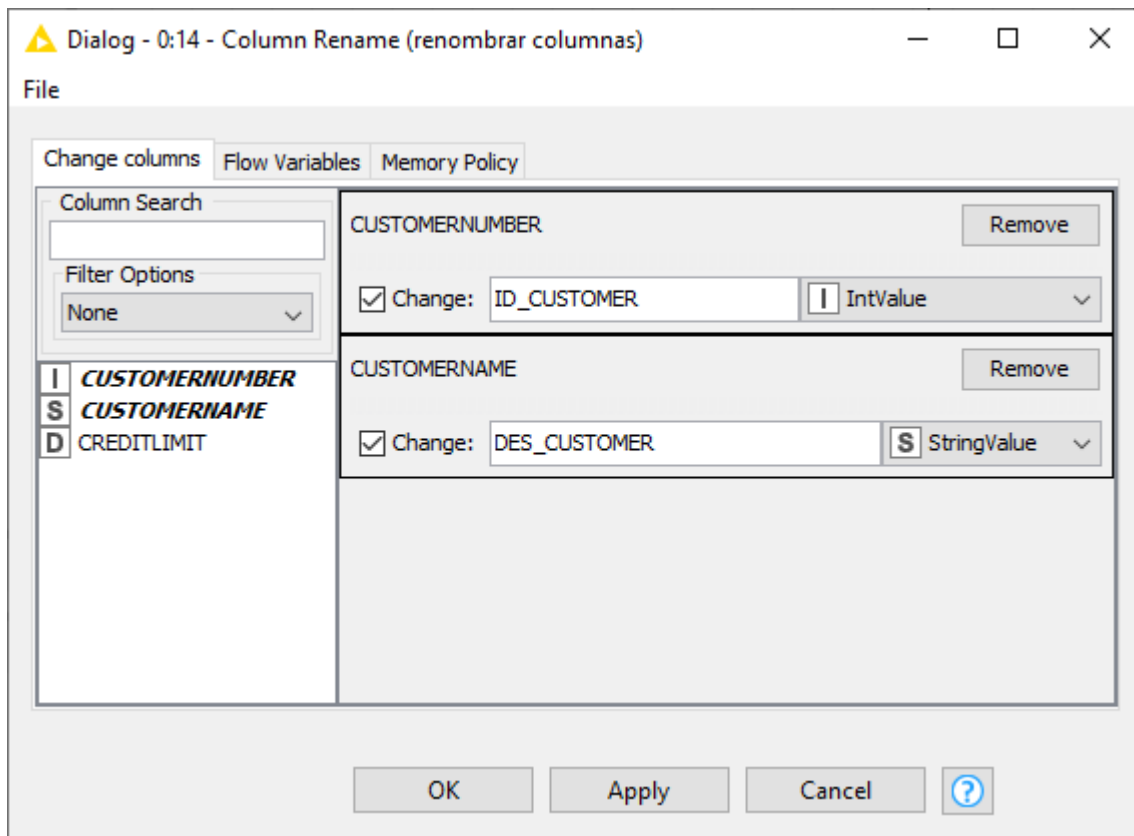
Con estos nodos nos encargamos respectivamente de crear la conexión a la base de datos, seleccionar la tabla que queremos cargar en KNIME y leerla (importar los datos a KNIME). Esta subrutina la emplearemos a lo largo del proceso ETL con el fin de cargar los datos en KNIME, dado que tanto los datos de entrada como los de salida están en una base de datos MySQL. Haciendo doble click configuramos cada uno de los nodos y los ejecutamos.

Column filter + Column rename:

Elegimos las columnas que queremos de la tabla, en este caso:



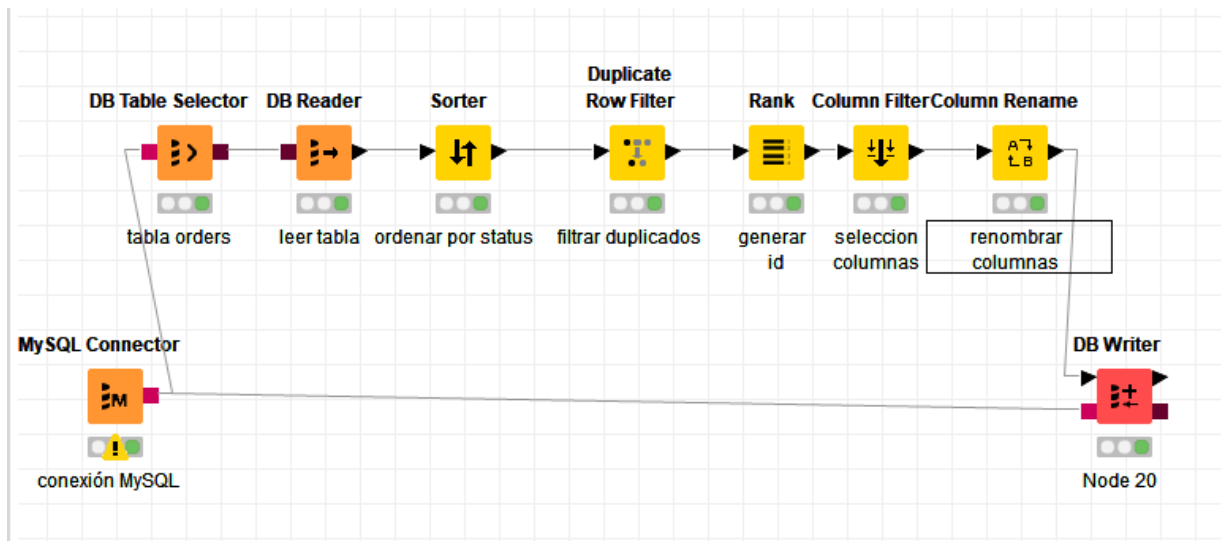
Con columna rename podemos darle el nombre que queremos a las columnas.



DB Writer:

Finalmente, este nodo escribe en una tabla el flujo de datos obtenido, con la conexión a base de datos que se le proporcione

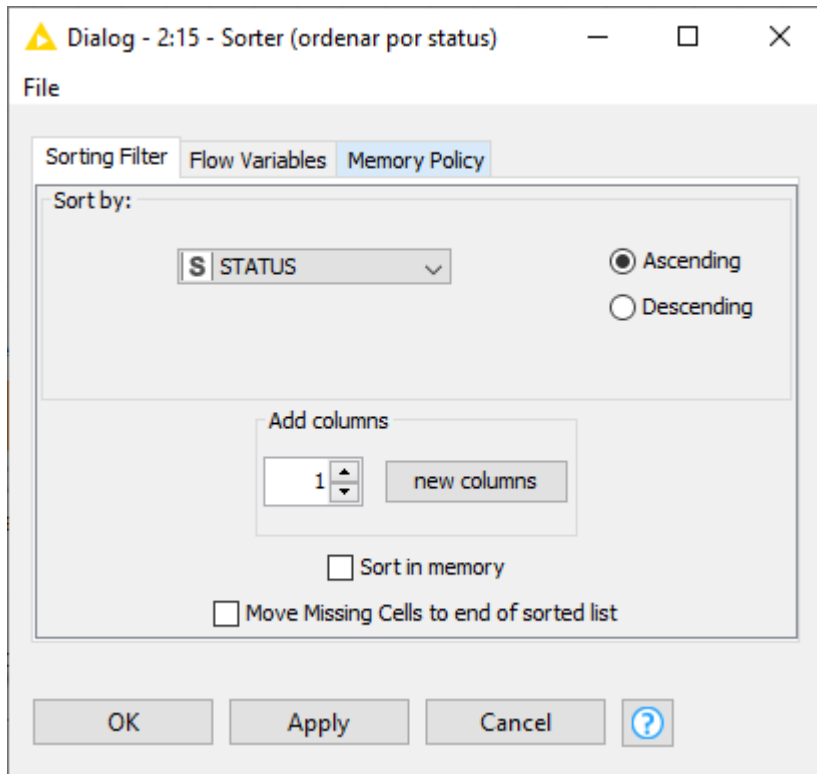
Dimensión status:



MySQL Connector + DB Table Selector + DB Reader

Como en el workflow anterior, conexión a la base de datos y obtenemos la tabla 'orders'.

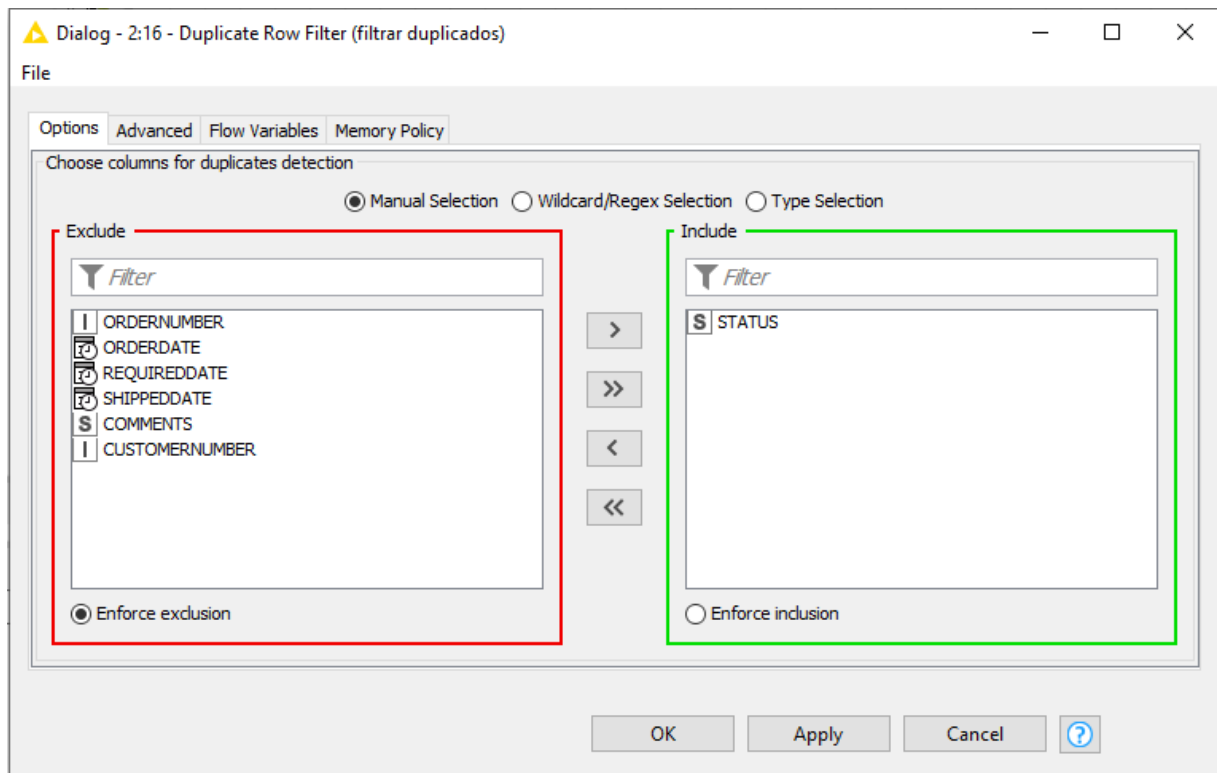
Sorter



Para ordenar las filas, ordenamos por STATUS.

Duplicate Row Filter

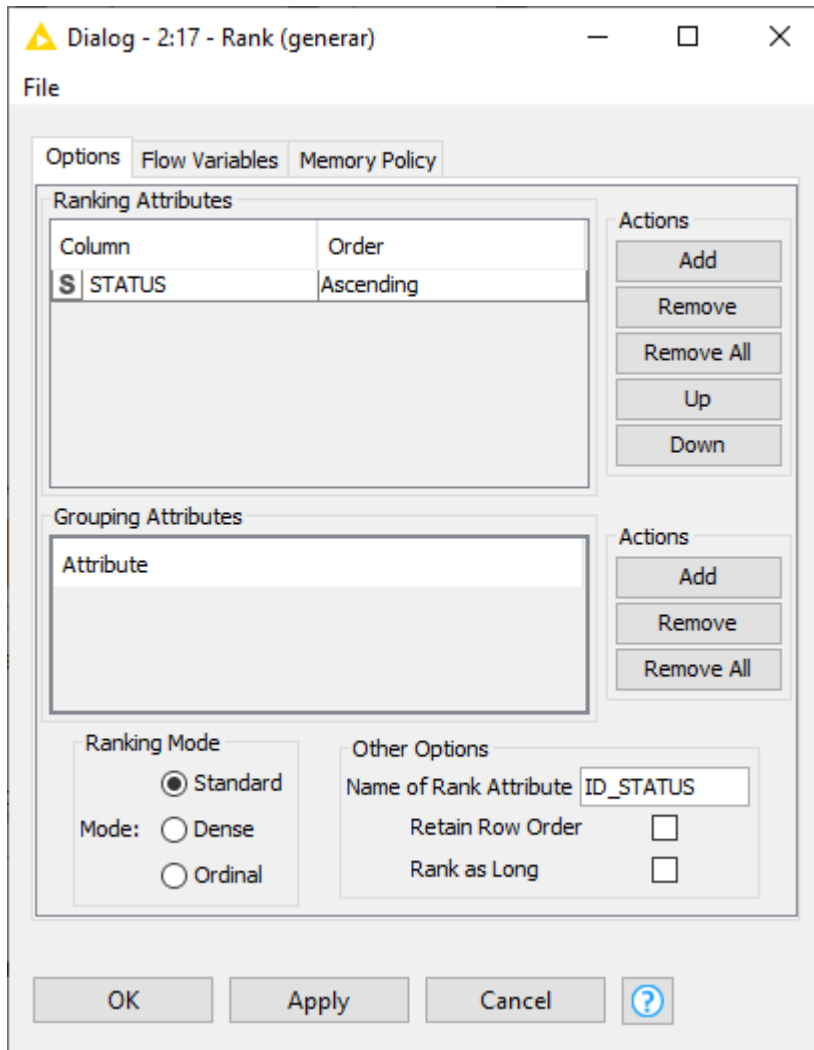
Para el filtrado de filas duplicadas.



Podemos elegir qué columnas determinan que dos filas sean iguales. En este caso buscamos obtener los distintos Status.

Rank

Para generar la clave subrogada (surrogate key), necesitamos generar un entero que determine de manera única cada Status. En este ejemplo lo hacemos mediante el nodo Rank, aunque existen otras posibilidades (mediante un campo autoincremental en la base de datos o crearlo con un nodo de código Java).



En Ranking Attributes usamos Add para añadir campos hasta obtener Status, el campo que decidirá el valor del "Rank". En Other Options podemos darle nombre a la columna que obtenemos, en nuestro caso ID_STATUS.

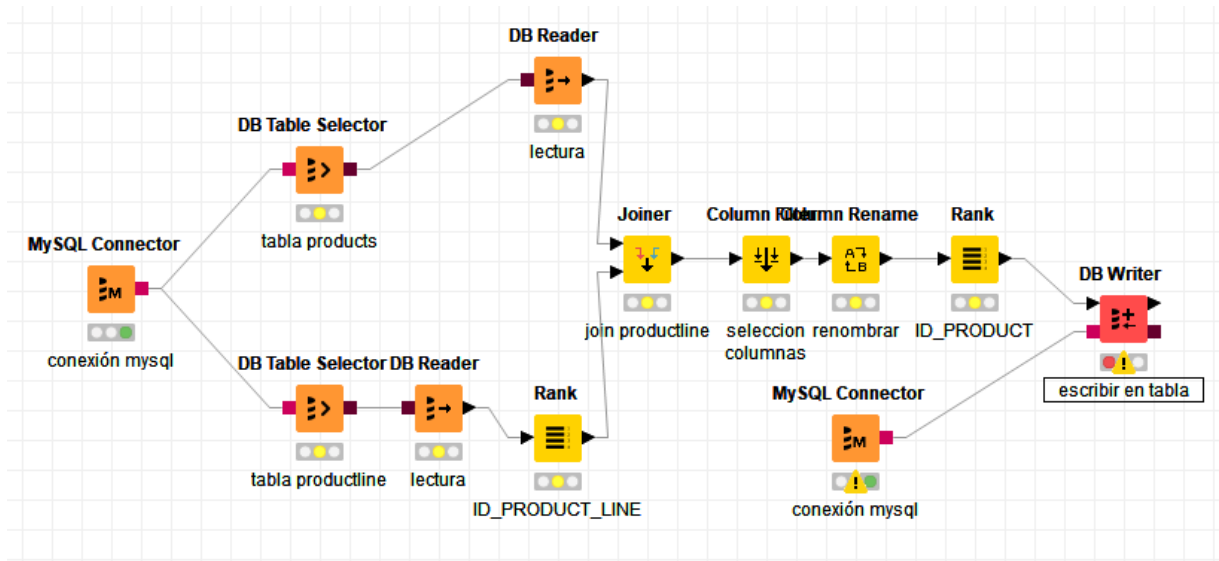
Column filter + Column rename

Seleccionamos las columnas Status e ID_STATUS y renombramos Status a ID_STATUS_NK, como en el anterior Workflow.

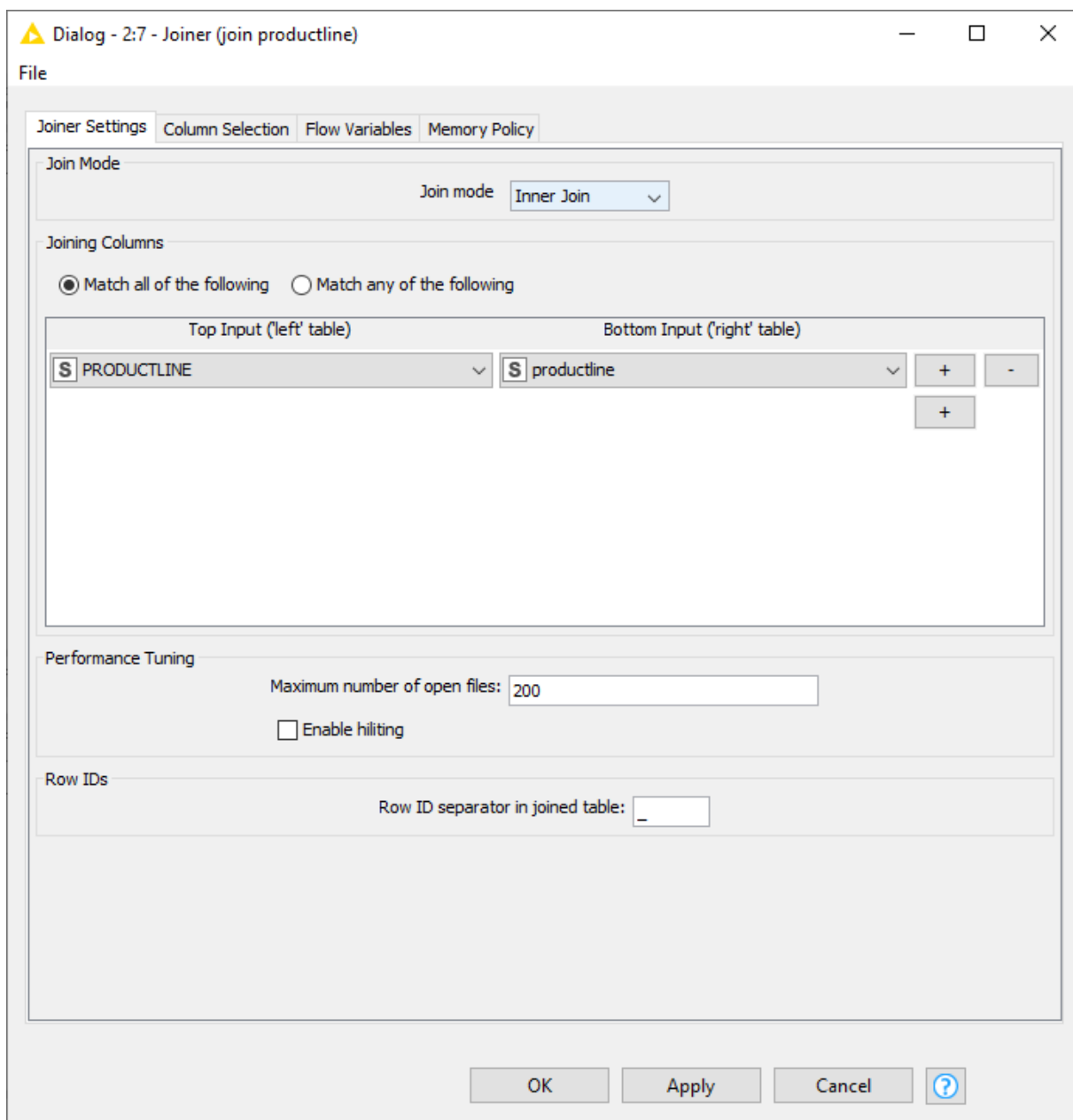
DB Writer

Escribimos en la DB el flujo de datos.

Dimensión product



Para esta dimensión necesitamos dos tablas de la base de datos, por una parte necesitamos la tabla 'products' y por otra parte la tabla 'productline'. Añadimos el ID_PRODUCT_LINE mediante el nodo Rank y unimos los productos y la información de la línea de producto mediante el Joiner.



Dialog - 2:7 - Joiner (join productline)

File

Joiner Settings Column Selection Flow Variables Memory Policy

Join Mode

Join mode Inner Join

Joining Columns

Match all of the following Match any of the following

Top Input ('left' table) Bottom Input ('right' table)

PRODUCTLINE productline

Maximum number of open files: 200

Enable hiliting

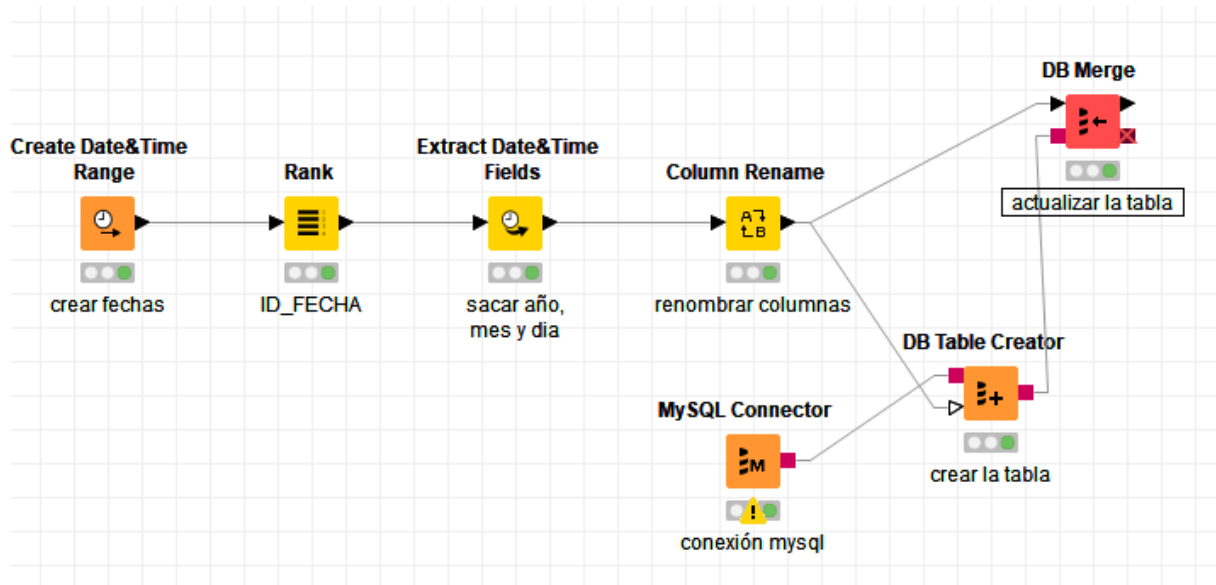
Row IDs

Row ID separator in joined table: _

OK Apply Cancel ?

Y con los últimos nodos seleccionamos y renombramos las columnas que queremos, para finalmente volcarlo en la tabla de salida.

Dimensión time



Create Date&Time Range

Dialog - 2:1 - Create Date&Time Range (crear fechas)

File

Options Flow Variables Memory Policy

Output Settings

Output type: Date

New column name: Fecha

Mode Selection

Number of rows: Fixed: 1.000 Variable

Starting Point

Start: Date: 2000-06-02 Time: 10:33:49

Time Zone: Europe/Paris

Use execution date&time

Ending Point

Interval: +1 day

End: Date: 2022-06-02 Time: 10:33:49

Use execution date&time

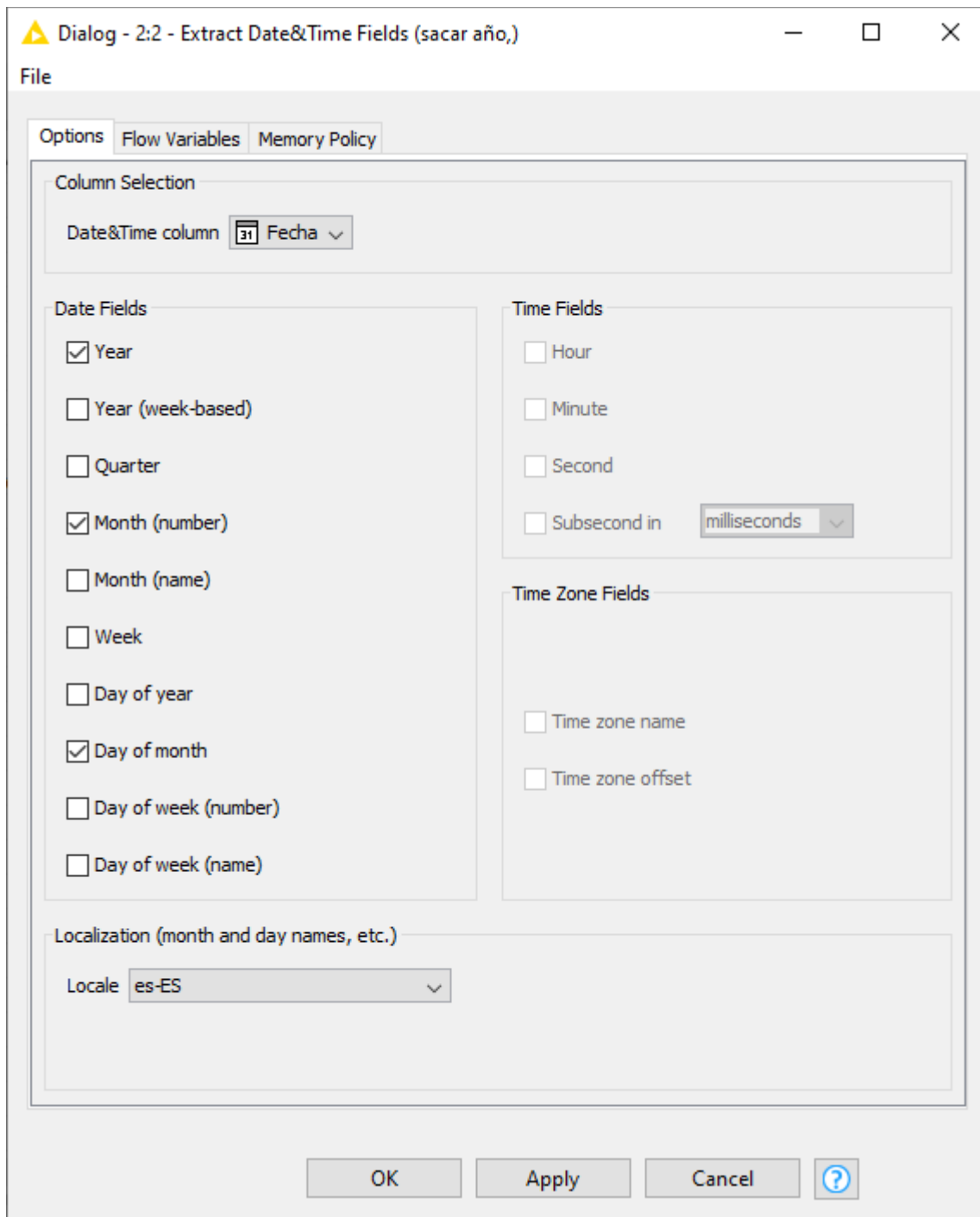
OK Apply Cancel ?

Este nodo nos permite generar intervalos de fecha o fecha y hora. En nuestro caso buscamos generar días desde el 2000 hasta la fecha, por lo tanto marcamos Número de filas como variable y seleccionamos el punto inicial, el final (execution date&time) y el intervalo (una fila por cada día).

Rank

Añadimos el Rank según la fecha para generar un ID_FECHA.

Extract DateTime Fields



Este nodo permite extraer partes de la fecha (o tiempo) que queramos. En nuestro caso buscamos el Año, el Mes y el día del mes.

Rename columns

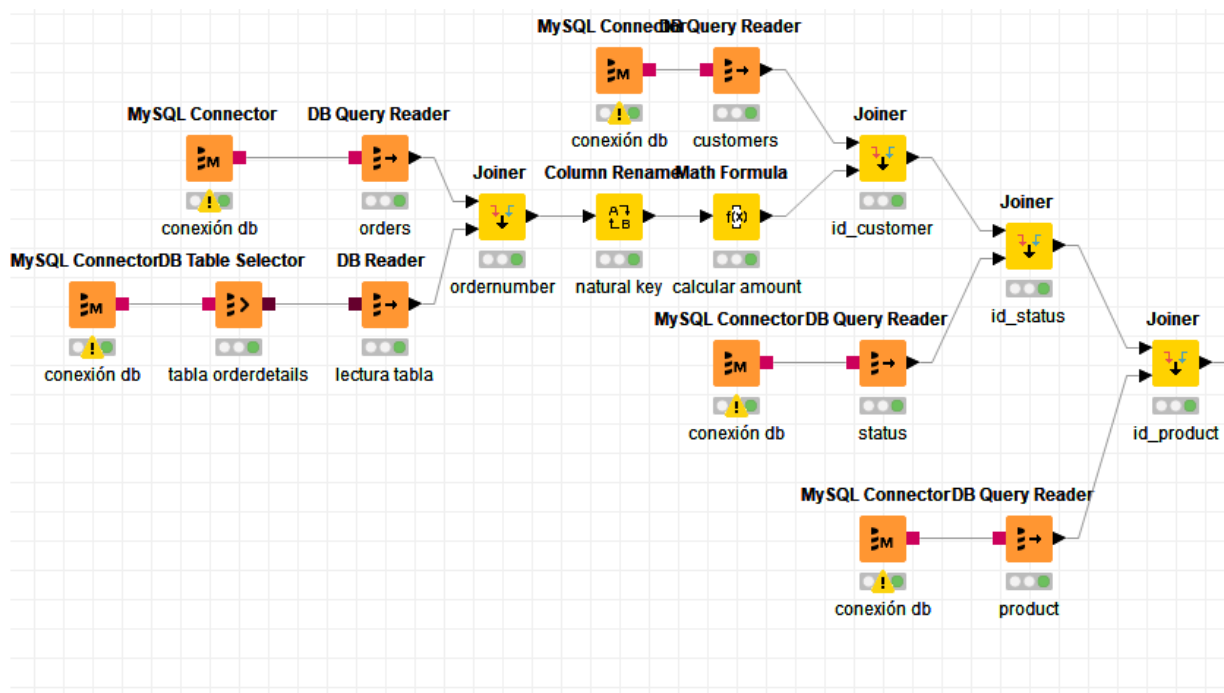
Renombramos las columnas Year, Month, Day of month a sus equivalentes en español.

MySQL Connector + DB Table Creator + DB Merge

La conexión de MySQL se conecta al Table Creator, junto a los datos del flujo. El Table Creator crea la tabla con las especificaciones del flujo de datos (número de columnas y tipos). En caso de existir la tabla, el nodo pasa los metadatos al nodo DB Merge, que se encarga de poblar la tabla por un proceso Merge, con los datos del flujo y los datos que contenga la tabla en ese momento.

Tabla de hechos

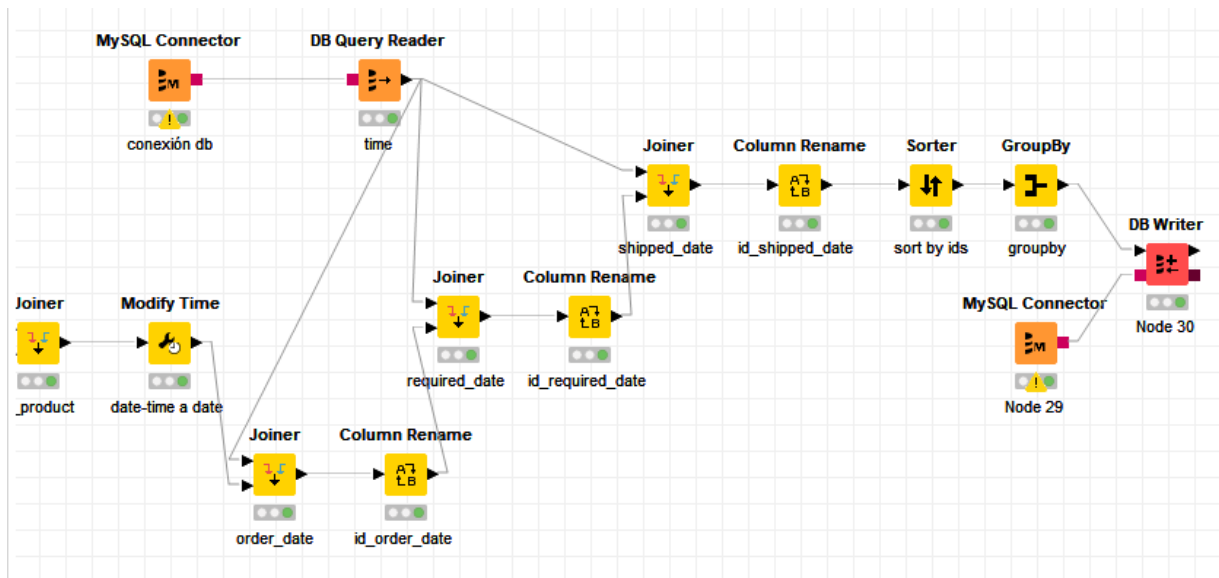
Primera parte del Workflow:



Para la construcción de la tabla de hechos debemos cargar los hechos de la tabla OrderDetails y vamos añadiendo todas las dimensiones. En este caso utilizamos el nodo DB Query Reader que nos permite leer directamente una consulta SQL (select * from table) y nos ahorra el seleccionar la tabla y leerla.

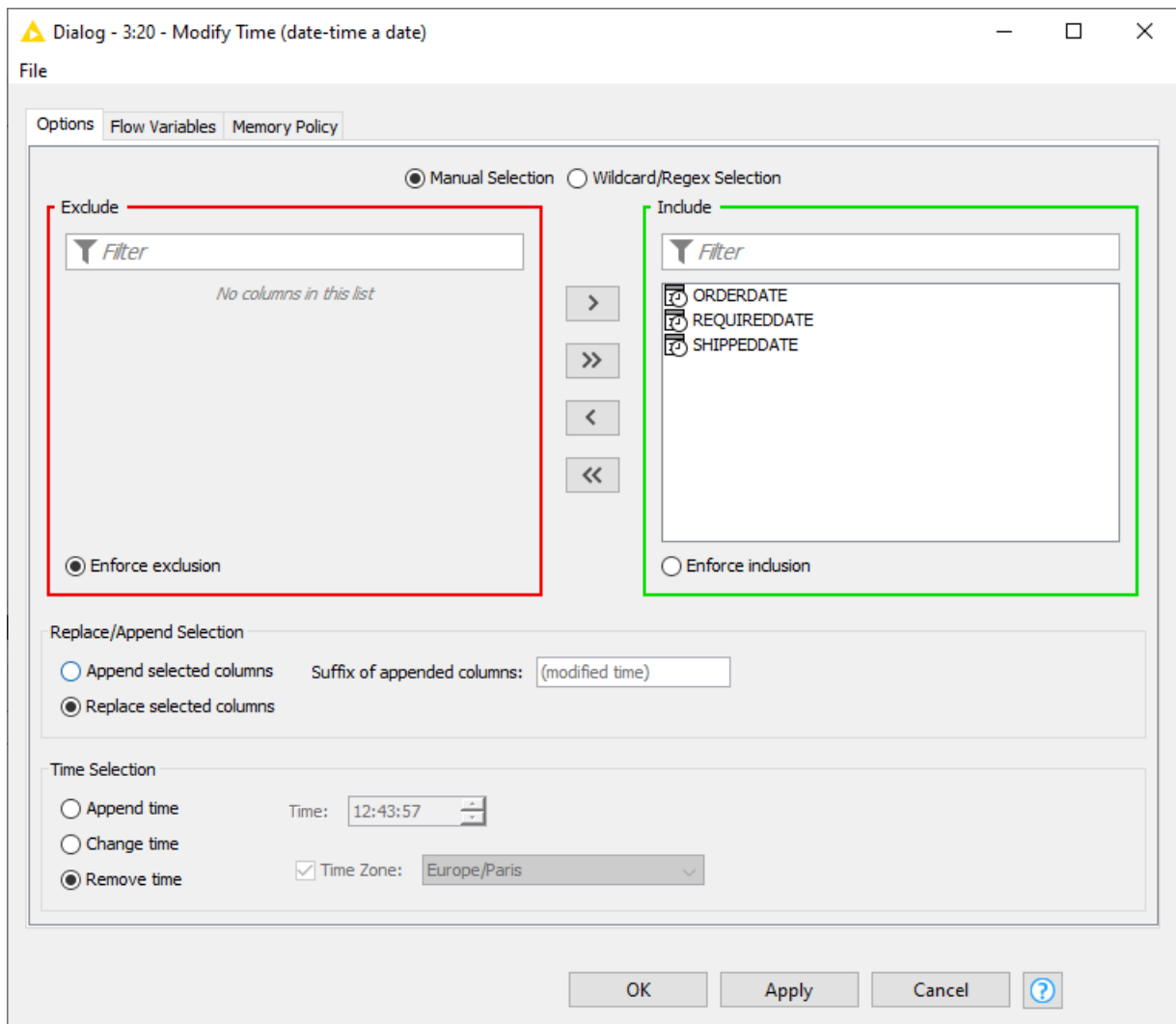
Cada una de las dimensiones las vamos uniendo mediante Joiner. El nodo Math Formula nos permite calcular amount como priceEach*quantity y almacenarlo en una columna.

Segunda parte del Workflow:



La última dimensión a añadir es la de tiempo, que se utiliza para tres campos: order_date, required_date y shipped_date.

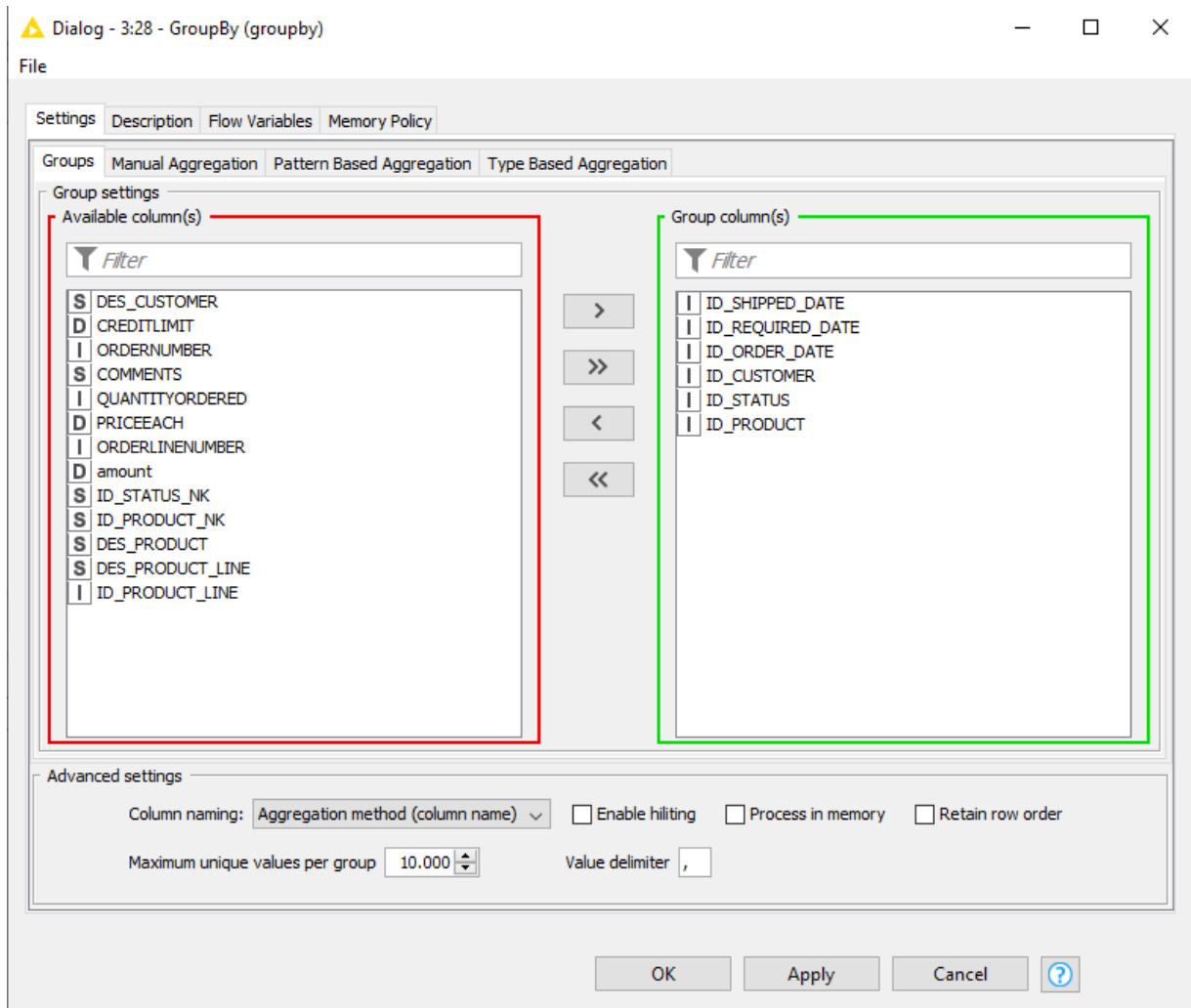
Modify Time

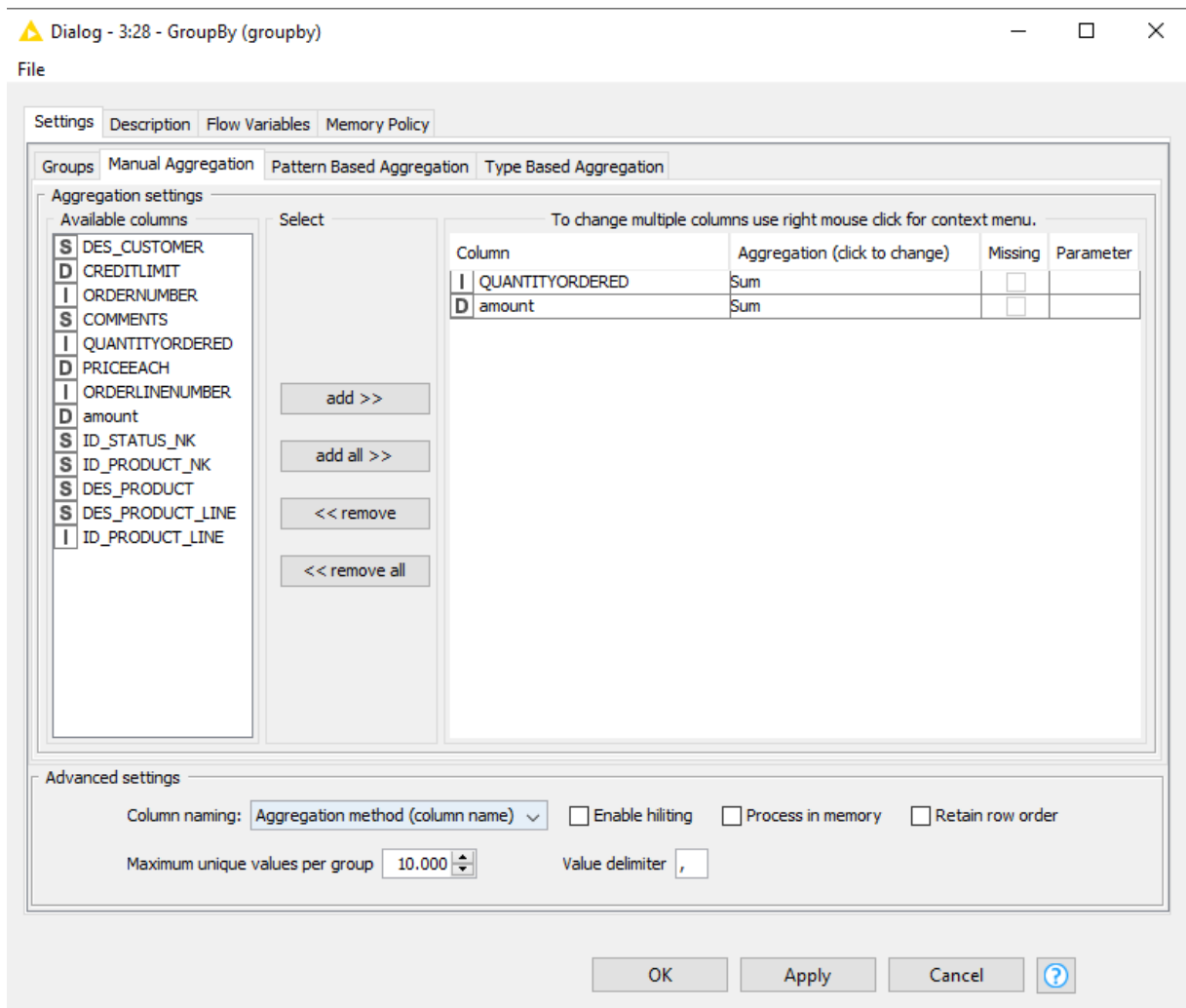


Dado que ORDERDATE, REQUIREDDATE, SHIPPEDDATE vienen en formato DateTime, para pasarlo a formato Date y poder compararlo con la dimensión Time que hemos creado, el nodo **Modify time** nos permite eliminar las horas del dato y dejarlo en formato Date.

Una vez añadimos todas las dimensiones, tenemos el nodo **Sorter** para ordenar las filas según sus ID.

GroupBy





Finalmente, con GroupBy agrupamos las filas según sus ID y calculamos las métricas agregadas por suma, en este caso quantityOrdered (cantidad pedida) y amount (importe).

Una vez tenemos todo el flujo de datos, podemos comprobarlo haciendo click derecho y observando el output del nodo final de Groupby, obtenemos nuestra tabla de hechos, que escribimos en la DB como hemos hecho en pasos anteriores.

Group table - 3:28 - GroupBy (groupby)

File Hilitte Navigation View

Table "default" - Rows: 2996 Spec - Columns: 8 Properties Flow Variables

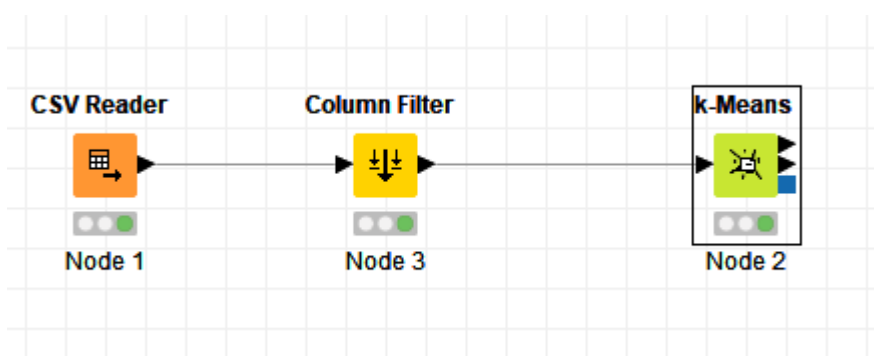
Row ID	ID_SHI...	ID_RE...	ID_OR...	ID_CUS...	ID_STA...	ID_PR...	Sum(QUANTITYOR...	Sum(amount)
Row0	953	956	949	363	6	23	30	5,160
Row1	953	956	949	363	6	27	50	3,400
Row2	953	956	949	363	6	50	22	1,914
Row3	953	956	949	363	6	80	49	1,666
Row4	954	961	952	128	6	29	25	3,775
Row5	954	961	952	128	6	33	26	3,770
Row6	954	961	952	128	6	61	45	1,395
Row7	954	961	952	128	6	64	46	2,484
Row8	957	961	953	181	6	19	39	4,797
Row9	957	961	953	181	6	20	41	2,050
Row10	975	983	974	141	6	11	34	5,950
Row11	975	983	974	141	6	15	41	4,633
Row12	975	983	974	141	6	26	24	3,456
Row13	975	983	974	141	6	28	29	3,770
Row14	975	983	974	141	6	40	23	4,554
Row15	975	983	974	141	6	49	38	5,358
Row16	975	983	974	141	6	57	35	1,925
Row17	975	983	974	141	6	68	44	1,760
Row18	975	983	974	141	6	81	26	2,912
Row19	975	983	974	141	6	88	35	1,680
Row20	975	983	974	141	6	89	49	3,234
Row21	975	983	974	141	6	94	33	3,696
Row22	975	983	974	141	6	95	32	1,696
Row23	976	981	972	121	6	2	26	5,408
Row24	976	981	972	121	6	6	42	5,418

EJEMPLO SENCILLO ML

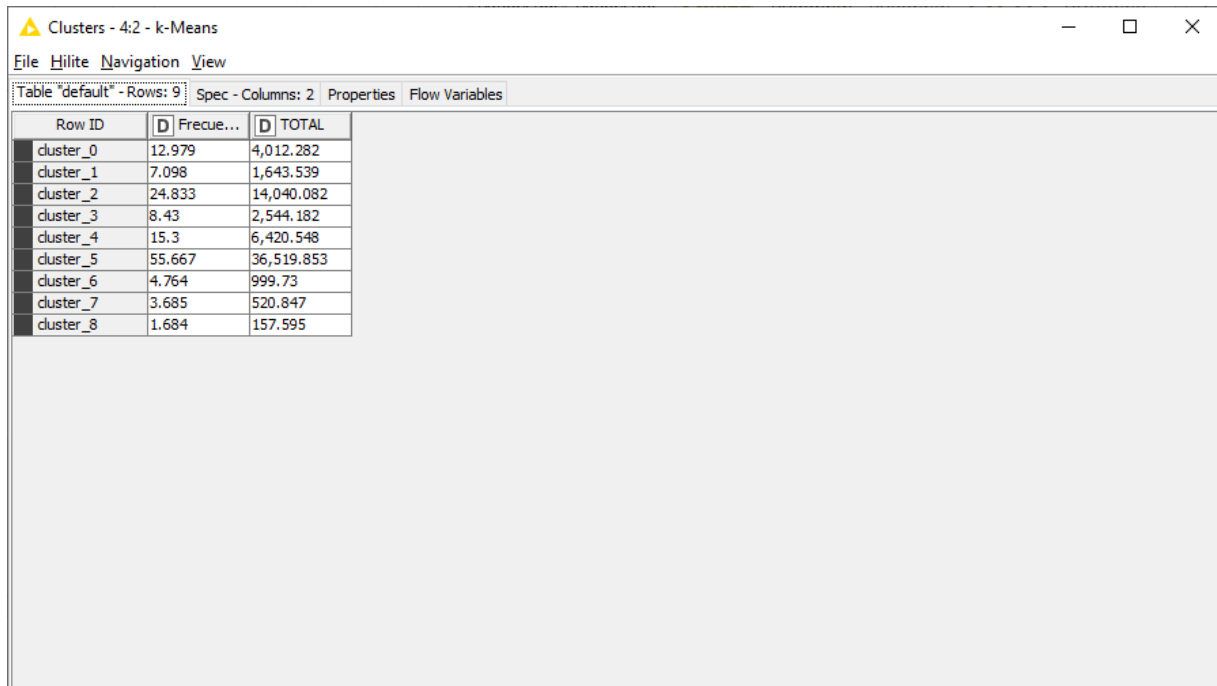
Aplicaremos el algoritmo kMeans con los datos estudiados en el artículo de Vertica-ML-Python.

Los datos de entrada tienen tres columnas: la id del cliente, la frecuencia de compra y el total de gasto. Queremos crear grupos mediante el algoritmo kMeans en función de su frecuencia y su gasto para ver cómo podemos catalogar a un cliente.

El workflow por tanto es el siguiente:



Leemos los datos de entrada, seleccionamos las variables de interés (Frecuencia y Total) y aplicamos el nodo de k-Means. Los resultados del algoritmo en KNIME son los siguientes:



Row ID	Frecue...	TOTAL
cluster_0	12.979	4,012.282
cluster_1	7.098	1,643.539
cluster_2	24.833	14,040.082
cluster_3	8.43	2,544.182
cluster_4	15.3	6,420.548
cluster_5	55.667	36,519.853
cluster_6	4.764	999.73
cluster_7	3.685	520.847
cluster_8	1.684	157.595

Donde cada fila es el centroide de cada cluster. Como referencia, en Python obtuvimos los siguientes resultados:

```
-----CENTROS-----  
[[2.23121387e+00 2.56333232e+02]  
 [2.10000000e+01 1.94942400e+04]  
 [1.57812500e+01 6.33970631e+03]  
 [5.90000000e+01 5.47344170e+04]  
 [8.11764706e+00 2.14206976e+03]  
 [2.56000000e+01 1.29492506e+04]  
 [4.96153846e+00 1.00110822e+03]  
 [1.19285714e+01 3.81115741e+03]  
 [5.40000000e+01 2.74125710e+04]]  
-----INERCIA-----  
Total Sum of Squares                8553241695.926128  
Total Within-Cluster Sum of Squares 155544996.73001078  
Total Between-Cluster Sum of Squares 8397696699.196117  
Between-Cluster / Total SS          0.9818144976770508
```

Donde además podemos ver métricas de rendimiento.

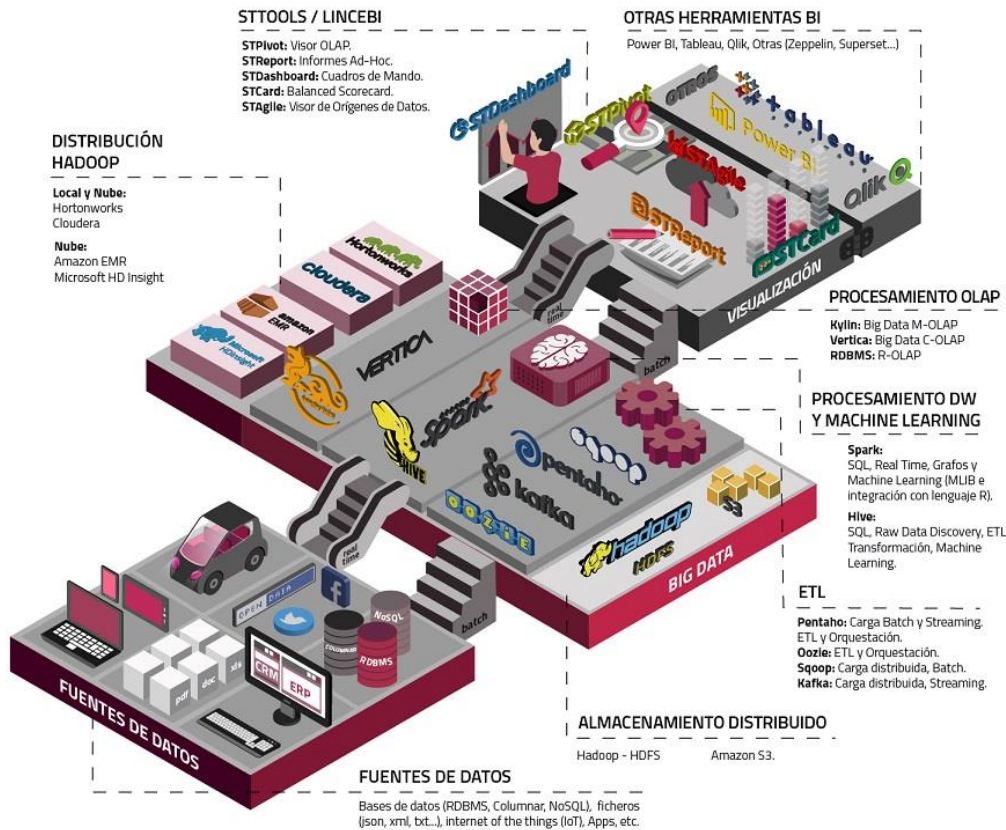
CONCLUSIÓN

KNIME es una plataforma para procesos ETL, con nodos con muchas funcionalidades de manera que se pueda conseguir casi cualquier resultado que se puede conseguir en otras herramientas de ETL como las de Pentaho o Talend. Se diferencia de estas por poder aplicar modelos de Machine Learning. Además de las funcionalidades base que ofrece, existen una serie de extensiones en <https://www.knime.com/knime-extensions> por ejemplo, para el procesamiento de textos. KNIME permite a usuarios no tan expertos en ciencia de datos poder aplicar estos modelos para realizar predicciones o un análisis más detallado de los datos, sin necesidad de escribir código en R o en Python. A diferencia de Alteryx, otra plataforma de analítica de datos y ETL, KNIME es Open Source y completamente gratuita.

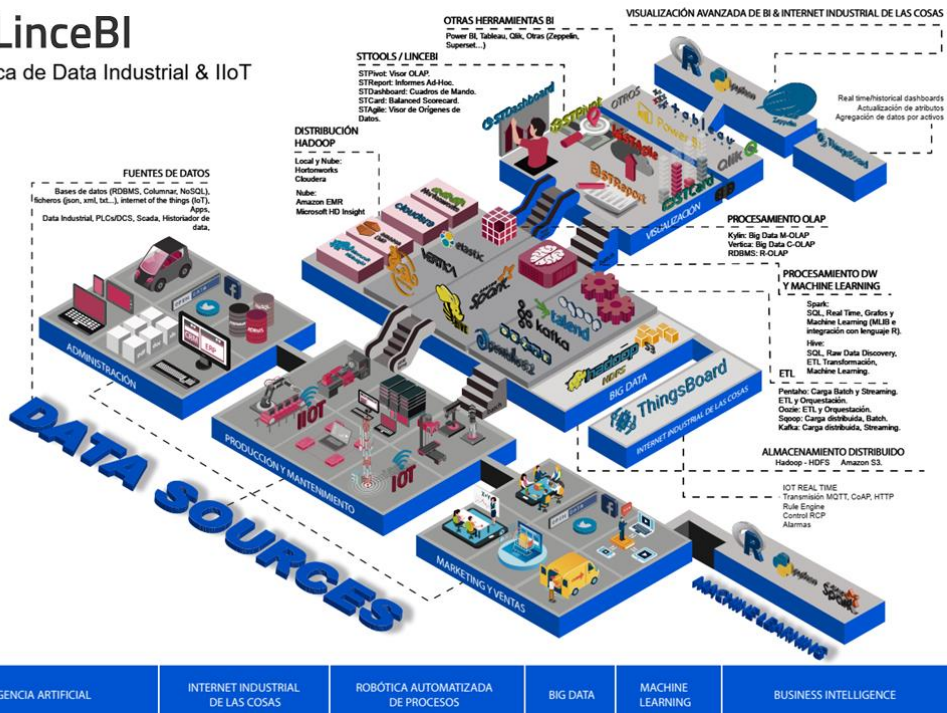
TECNOLOGÍAS

Recientemente, hemos sido nombrados Partners Certificados de Vertica, Talend, Microsoft, Snowflake, Kylligence, Pentaho, etc.





LinceBI
 Analítica de Data Industrial & IIoT



INFORMACIÓN SOBRE STRATEBI



Stratebi es una empresa española, con sede en Madrid y oficinas en Barcelona, Alicante y Sevilla, creada por un grupo de profesionales con amplia experiencia en sistemas de información, soluciones tecnológicas y procesos relacionados con soluciones de Open Source y de inteligencia de Negocio.

Esta experiencia, adquirida durante la participación en proyectos estratégicos en compañías de reconocido prestigio a nivel internacional, se ha puesto a disposición de nuestros clientes.

Somos **Partners Certificados en Microsoft PowerBI** con una dilatada experiencia

Stratebi es la única empresa española que ha estado presente todos los Pentaho Developers celebrados en Europa habiendo organizado el de España.

En Stratebi nos planteamos como **objetivo** dotar a las compañías e instituciones, de herramientas escalables y adaptadas a sus necesidades, que conformen una estrategia Business Intelligence capaz de rentabilizar la información disponible. Para ello, nos basamos en el desarrollo de soluciones de Inteligencia de Negocio, mediante tecnología Open Source.

Stratebi son **profesores y responsables de proyectos** del Master en Business Intelligence de la Universidad UOC, UCAM, EOI...

Los profesionales de Stratebi son los creadores y autores del primer weblog en español sobre el mundo del Business Intelligence, Data Warehouse, CRM, Dashboards, Scorecard y Open Source. Todobi.com

Stratebi es partner de las principales soluciones Analytics: Microsoft Power BI, Talend, Pentaho, Vertica, Snowflake, Kylligence, Cloudera...

Todo Bi, se ha convertido en una referencia para el conocimiento y divulgación del Business Intelligence en español.

OTROS

Trabajamos en los principales sectores y con algunas de las compañías y organizaciones más importantes de España.

SECTOR PRIVADO

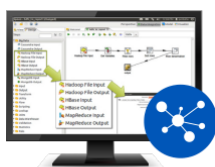


SECTOR PÚBLICO



EJEMPLOS DE DESARROLLOS ANALYTICS

A continuación, se presentan **ejemplos de algunos screenshots** de cuadros de mando diseñados por Stratebi, con el fin de dar a conocer lo que se puede llegar a obtener, así como Demos Online en la web de Stratebi:



Data Ingestion
 Manipulation
 Integration



Enterprise and
 Ad Hoc Reporting



Data Discovery
 Visualization



Predictive
 Analytics

Pentaho Analytics Platform

Hadoop

NoSQL

Analytic
 Databases

Relational



