Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Campus LATAM

Nombre del trabajo:

4.3 Fase 3a. Preprocesamiento de datos, integración de datos

Curso:

Análisis de Grandes Volúmenes de Datos

Alumnos:

Juan Carlos Alvarado Carricarte, A01793486

Bryan Rodolfo Alvarado Cruz, A01793670

Eduardo Gabriel Arévalo Aguilar, A01793897

Profesor:

Alberto De Obeso Orendain

Tutor:

Luis Angel Lozano Medina

Fecha de entrega:

05 de febrero

Resumen

El objetivo de este trabajo es realizar la integración de datos demográficos de la zona de Jalisco, México, a través de procesos de Extracción, Transformación y Carga (ETL) con la herramienta Knime, la cuál tiene el aval del profesor para realizar las prácticas con ETLs. Para ello, se utilizaron distintas fuentes de información, como el Directorio Estadístico Nacional de Unidades Económicas (DENUE), datos de hogares, mortalidad, natalidad y educación. Se realizó un join para relacionar los municipios con su respectiva zona o región en Jalisco. Se sigue una metodología basada en el diseño de un grafo acíclico dirigido, partiendo de los diccionarios de datos de cada fuente. Se muestra el modelo de datos integrados y se explican las relaciones entre las fuentes de datos.

Palabras clave: ETL, Knime, integración de datos, limpieza de datos, fuentes de datos, integración de datos, México, Jalisco, DENUE.

Preprocesamiento de datos, integración de datos

El análisis de datos es una herramienta vital para la toma de decisiones actualmente, y hemos visto cómo ha tomado relevancia en cualquier entorno empresarial, gubernamental, universidades, deportes, centros de investigación, deportes, etc. Con el crecimiento desmedido de los datos, es necesario el uso de herramientas que nos permitan procesar estos grandes volúmenes de datos, así mismo, hacer una correcta extracción, transformación y carga de los datos para su posterior análisis. Entendiendo la problemática, Knime es una herramienta de análisis de datos que permite la integración de información de distintas fuentes y la transformación de datos de manera rápida y sencilla con una interfaz gráfica muy intuitiva.

En el presente trabajo, se realiza un proceso de Extracción, Transformación y Carga (ETL) de datos utilizando Knime. Se han elegido como fuentes de datos los datos estadísticos de la demografía de la zona de Jalisco en México, que incluyen inicialmente el Directorio Estadístico Nacional de Unidades Económicas (DENUE), hogares, mortalidad, natalidad y educación. A partir de estas fuentes, se realiza una integración de datos con un archivo que contiene el CVE de todos los municipios de Jalisco y la zona a la cuál pertenecen, para obtener un modelo de datos integrados que permita su análisis y exploración por zonas de manera más efectiva.

En este trabajo, se presenta el proceso de ETL usando la herramienta Knime, detallando las distintas etapas de extracción, transformación y carga de información para cada una de las fuentes seleccionadas.

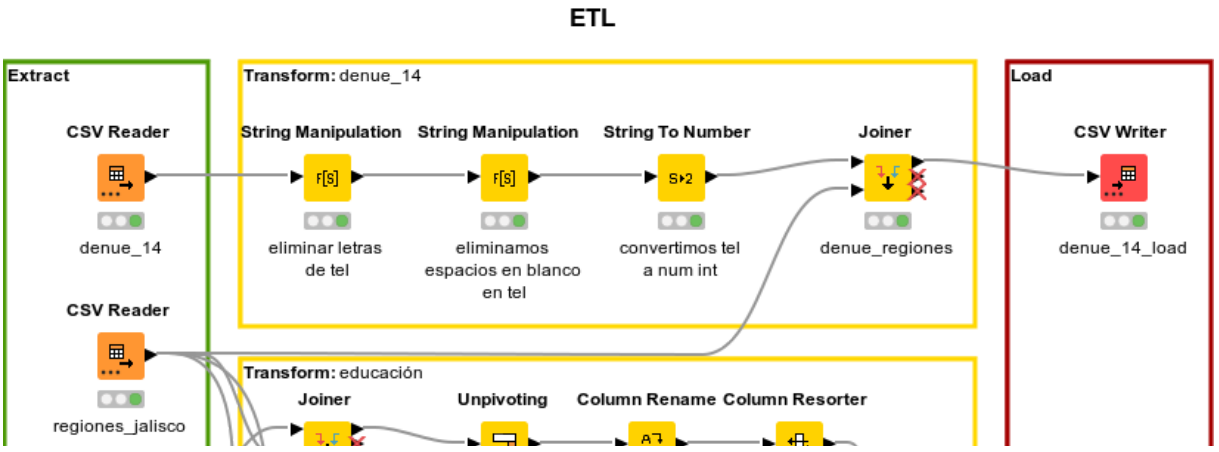
Elaboración del ETL

Para la creación del ETL se hizo uso de la herramienta de análisis de datos llamada Knime. Las fuentes seleccionadas fueron descargadas de la web del INEGI (Instituto Nacional de Estadística y Geografía), obtuvimos datos de: 1. DENUE (Directorio Estadístico Nacional de Unidades Económicas), que ofrece información detallada como identificación, ubicación, empleados, etc., sobre los negocios activos. Posteriormente en la sesión de demografía descargamos datos como: 2. Hogares (indica cantidad de hogares y población de los mismos), 3. Mortalidad (tiene indicadores de defunciones y % de las mismas), 4. Natalidad (información acerca de nacimientos, segmentación por niños y niñas, promedio de nacidos por edades de las madres y grupo de edades, etc.) y 5. Educación (contiene muchos indicadores como población que asiste a las escuela por grupo de edades, tasa de alfabetización por grupo de edades, % de población que no asiste a la escuela por edades, etc.). El diseño del ETL se elaboró examinando las características y detalles de las fuentes de datos, logrando una estructura más amigable que nos permitiera conocer mejor la naturaleza de estos datos. Todas las fuentes fueron integradas con el archivo que contiene el CVE por municipios y sus zonas en Jalisco, posteriormente se agregaron los nodos para realizar todo el proceso de transformación adecuado para cada fuente de datos.

DENUE (Directorio Estadístico Nacional de Unidades Económicas)

Este archivo tiene una extensión .csv, extraemos sus datos desde la fuente usando un nodo que lee archivos csv, se identifica que la columna de teléfono tenía letras, se procedió eliminando las letras de ese campo, igualmente se eliminan los espacios en blanco y se convierte ese campo a **int.** Luego se cruza con el archivo de referencia **regionesjalisco.csv** que es el que nos sirve para identificar en cuál zona queda algún municipio de acuerdo a su

CVE, de esta forma terminamos la transformación. Procedemos a cargar el resultado usando un nodo de escritura csv, el resultado final es denue_14_load.csv



Hogares

Explorando el archivo hogares_00.xlsx, como nos percatamos es tipo Excel, repetimos el procedimiento de extracción, en este caso usamos un nodo de lectura de archivos de Excel, procedemos a integrar con el archivo de regiones mediante un join, posteriormente vemos que esta fuente tiene referenciada la información de los años en columnas distintas, por lo que decidimos transformar esas columnas en filas y al lado una columna nueva que represente sus valores, luego renombramos estas nuevas columnas, las ordenamos, como la fuente tiene mucha información de los valores en blanco, recurrimos a eliminar esos valores nulos ya que no nos aportan información a nuestro proceso, luego eliminamos las letras de la columna de valores, para quitar resultados como NA, no aplica, sin registro, etc. Nos vuelven a quedar algunos valores nulos, por lo que los eliminamos para limpiar más los datos, finalmente convertimos nuestra columna de años y valores en enteros.

En la siguiente tabla podemos ver cómo vienen los datos originalmente, con una columna por cada año de registro:

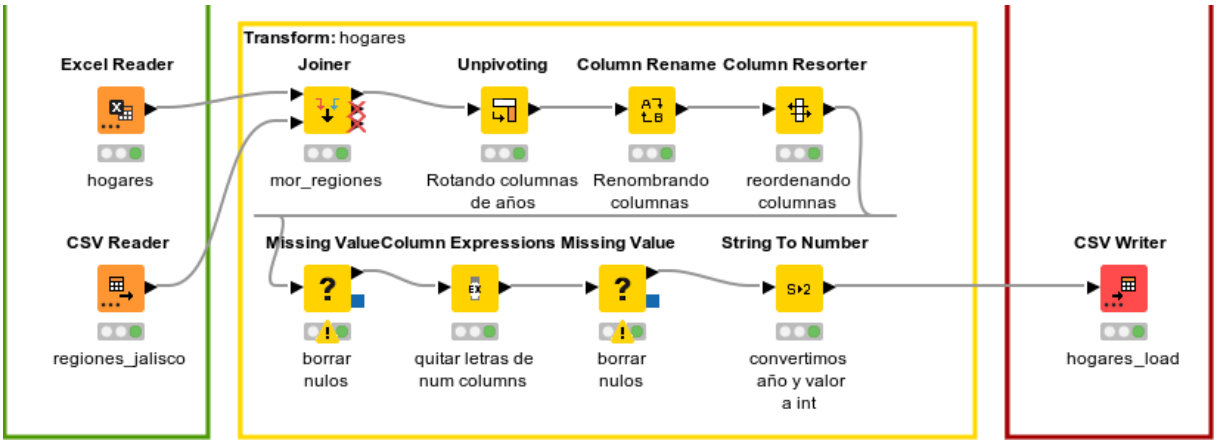
File Table - 3:53 - Excel Reader (hogares)

File Edit Hilite Navigation View

Table "default" - Rows: 5005 Spec - Columns: 12 Properties Flow Variables

Row ID	cve_en...	desc_entidad	cve_mu...	desc_municipio	id_indic...	indicador	2000	2005	2010	2015	
Row0	0	Estados Unidos Mexicanos	0	Estados Unidos Mexicanos	1002000014	Población en hogares	95380242	100221103	110610075	119530753	1
Row1	0	Estados Unidos Mexicanos	0	Estados Unidos Mexicanos	1002000018	Hogares	22268916	24803625	28159373	31949709	3
Row2	1	Aguascalientes	0	Estatal	1002000014	Población en hogares	936920	1048311	1178123	1312544	1
Row3	1	Aguascalientes	0	Estatal	1002000018	Hogares	208167	248905	289575	334589	3
Row4	1	Aguascalientes	1	Aguascalientes	1002000014	Población en hogares	637834	709280	791370	?	9
Row5	1	Aguascalientes	1	Aguascalientes	1002000018	Hogares	147147	173948	201071	?	2
Row6	1	Aguascalientes	2	Asientos	1002000014	Población en hogares	37581	40347	45423	?	5
Row7	1	Aguascalientes	2	Asientos	1002000018	Hogares	7561	8834	10007	?	1
Row8	1	Aguascalientes	3	Calvillo	1002000014	Población en hogares	50928	49899	54041	?	5
Row9	1	Aguascalientes	3	Calvillo	1002000018	Hogares	10760	11350	12834	?	1
Row10	1	Aguascalientes	4	Cosío	1002000014	Población en hogares	12574	13658	15036	?	1
Row11	1	Aguascalientes	4	Cosío	1002000018	Hogares	2593	2968	3465	?	3
Row12	1	Aguascalientes	5	Jesús María	1002000014	Población en hogares	63807	81467	99211	?	1

ETL para hogares:



En la tabla final para hogares se aprecia toda la transformación que se le realizó a la fuente:

Transformed input - 3:50 - String To Number (convertimos)

File Edit Hilite Navigation View

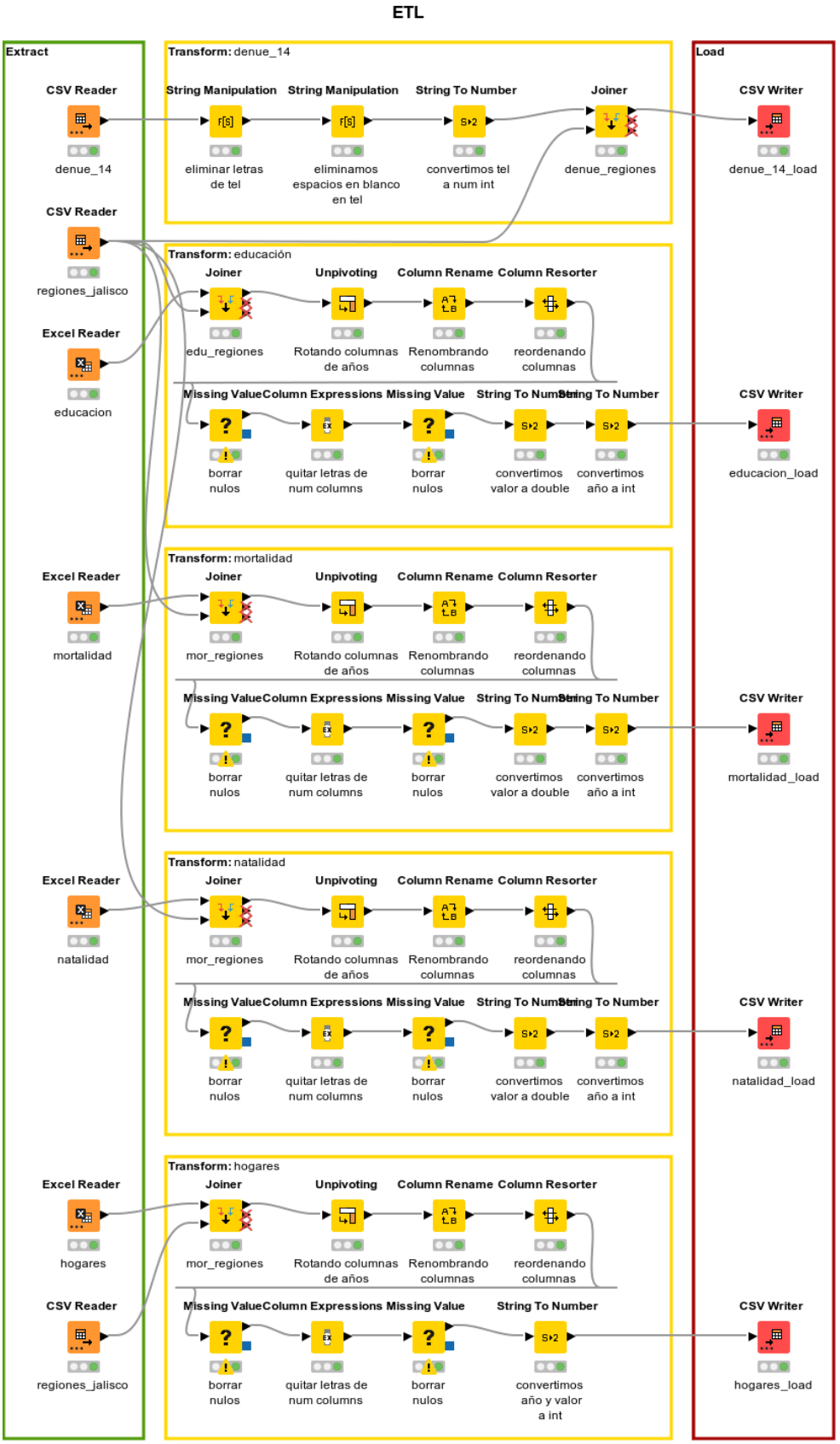
Table "default" - Rows: 14653 Spec - Columns: 11 Properties Flow Variables

Row ID	RowIDs	cve_en...	desc_en...	cve_mu...	desc_m...	id_indic...	indicador	unidad...	ano	valor	REGION
Row0	Row4_Row77	1	Aguascalientes	1	Aguascalientes	1002000014	Población en hogares	Personas	2000	637834	altos sur
Row1	Row4_Row77	1	Aguascalientes	1	Aguascalientes	1002000014	Población en hogares	Personas	2005	709280	altos sur
Row2	Row4_Row77	1	Aguascalientes	1	Aguascalientes	1002000014	Población en hogares	Personas	2010	791370	altos sur
Row4	Row4_Row77	1	Aguascalientes	1	Aguascalientes	1002000014	Población en hogares	Personas	2020	945505	altos sur
Row5	Row5_Row77	1	Aguascalientes	1	Aguascalientes	1002000018	Hogares	Hogares	2000	147147	altos sur
Row6	Row5_Row77	1	Aguascalientes	1	Aguascalientes	1002000018	Hogares	Hogares	2005	173948	altos sur
Row7	Row5_Row77	1	Aguascalientes	1	Aguascalientes	1002000018	Hogares	Hogares	2010	201071	altos sur
Row9	Row5_Row77	1	Aguascalientes	1	Aguascalientes	1002000018	Hogares	Hogares	2020	266778	altos sur
Row10	Row6_Row0	1	Aguascalientes	2	Asientos	1002000014	Población en hogares	Personas	2000	37581	centro
Row11	Row6_Row0	1	Aguascalientes	2	Asientos	1002000014	Población en hogares	Personas	2005	40347	centro
Row12	Row6_Row0	1	Aguascalientes	2	Asientos	1002000014	Población en hogares	Personas	2010	45423	centro

Mortalidad, Natalidad y Educación

Estas 3 fuentes tienen una estructura muy similar al archivo de hogares, donde los años tienen una columna independiente con sus registros, por lo cual, el procedimiento que ejecutamos en todos fue el mismo, a excepción que la columna de valores al tener múltiples

tipos de resultados, la convertimos en decimal y no entero, del resto, todo el proceso de ETL es el mismo. Documentamos cómo se ve la perspectiva general de nuestro proceso de ETLs:



Todos los archivos se exportaron en extensión .csv para su posterior importación u análisis en cualquier herramienta de analítica y visualización de datos.

Al finalizar estos pasos para cada una de nuestras fuentes, se garantiza la calidad de los mismos, ya que se realizó una depuración con limpieza y transformación para retirar los datos que no aportan al proceso.

En conclusión, podemos evidenciar que el proceso de ETL permitió integrar datos de diferentes fuentes, limpieza y transformación para obtener un panorama completo de la demografía de la región de Jalisco y que los resultados obtenidos podrían ser utilizados para analizar, visualizar y así facilitar la tomar decisiones en diversos ámbitos. Como opción de mejora, podemos seguir sacando tablas de los resultados obtenidos, ya que en los archivos de mortalidad, natalidad, hogares y educación, en los valores hay datos de porcentajes, valores absolutos y decimales que representan distintos indicadores, estos datos se pueden seguir segmentando en la sesión de ETLs, pero también al momento de generar las visualizaciones, las tablas que los importen pueden generar otras subtablas, o en su defecto, crear medidas que solo den resultados a los valores que indiquemos.

Referencias

ETL Data Manipulation. (s. f.). KNIME. https://www.knime.com/nodeguide/etl-data-manipulation

Hitachi (actualización 2021, octubre 8). *A Pentaho Data Integration Tutorial. Hitachi Vantara*. https://help.hitachivantara.com/Documentation/Pentaho/9.1/Setup/Pentaho_Data_Integration_(PDI)_tutorial

Geografía, E. D. N. I. Y. (s. f.-a). Directorio Estadístico Nacional de Unidades Económicas. DENUE. *https://www.inegi.org.mx/app/mapa/denue/default.aspx*

Regiones de Jalisco | Gobierno del Estado de Jalisco. (s. f.). https://www.jalisco.gob.mx/es/jalisco/regiones

ETL. (s. f.). KNIME. https://www.knime.com/etl-software