PROCESO ETL (EXTRACCIÓN TRANSFORMACIÓN Y CARGA)



Autor: ing.Alejandro Ramos

> REQUISITOS PREVIOS A LA INSTALACIÓN

Requisitos mínimos de hardware

- Procesador Procesador x64: AMD Opteron, AMD Athlon 64, Intel Xeon compatible con Intel EM64T, Intel Pentium IV compatible con EM64T
- 1 GB de memoria RAM
- Disco Duro con al menos 2 GB libres

Requisitos de software

- Java run Time Enviroment 5 o posteriores
- MySQL version 5 o posteriores
- Pentahoo Data Integration(PDI)
- Power Designer

Sistema Operativo

El sistema operativo que se ocupó para este desarrollo fue Windows, las herramientas tienen soporte para cualquier versión desde Windows 7 en adelante.

DISEÑO DE LA BASE DE DATOS CON POWER DESIGNER

Ejec	cutamos Power De	signer		(Abrimos File v
취 Po	owerDesigner				
File	Edit View Repository Tools	Window Help			seleccionamos New
2	New Model	Ctrl+N	' 🎯 🚬 🚨 🗟 🖓 🗣 🔒 🛃		Model
1	New Project	Ctrl+Alt+J)
6	Open	Ctrl+O			
	Save	Ctrl+S			
	Save As				
	Save As New Model				
	Close	Ctrl+Alt+F4			
	Open Workspace	Shift+F2			
	Save Workspace	Shift+F3			
	Save Workspace As				
	Close Workspace	Shift+F4			
ø	Save All	Ctrl+F3			
B,	Print Preview	Ctrl+Alt+P			
3	Print	Ctrl+P			
	Print Selection				
	Import	•			
	Reverse Engineer	•			
	Recent Files	•			
	Recent Workspaces and Projects	•			
	Exit	Alt+F4			

A continuación seleccionamos crear el modelo lógico.

Hie Let View Reportery Dols Window Help	di PowerDesigner								
Image:	hie Edit View Kepository Tools Window Help								
Dight Browser 0 File: Category: Ca	<u>``</u> ₩≌⊌₩₩43×4157,	🖞 New Model							
File: Image: Second	Object Browser · 부 ×	Category:	Category items:						
Local Repository Jutput OK	File: Construction of the second seco	Categories Image: Categories Model types Image: Categories Model types Image: Categories Application Image: Categories Model types Image: Categories Application Image: Categories Model types Image: Categories Application Image: Categories Application Image: Categories Application Image: Categories Application Image: Categories Model types Image: Categories Application Image: Categor	n n 2027 perte and Planning model (LDM) helps you analyze the structure of an inf lede ontry identification and information and information re available in the more concrete physical data mode	Deta Flow Diagram Deta Mo Deta Flow Diagram Deta Mo UML Class Diagram	vement Model vement Model KML MultiDimensional Dat secific physical database implement atio allow you to model views, indexes and of	a n.An			
	Cal Local Cal Repository			C	OK Cancel	Help			

Se nos presenta una nueva ventana donde podemos empezar con el diseño lógico de la base de datos

	MANUAL PROCESO ETL - Microsoft Word
1 PowerDesigner [LDM Prueba, Diagram_1]	
File Edit View Model Symbol Report Repository Tools Window Help	
🗅 🗃 🖆 🖬 🕼 🛆 👙 🛝 🛍 🏝 🗡 🤭 😁 🚰 🌍 📄 🛍 🧏 🖻 🕒 💷	/ () A 11 () · · ·) II I
Object Browser • # × Diagram_1 Result List	Ţ X Toolbox ▼ Ţ X
File: Workspace saludvoll saludvoll Physical Data_2 saludvoll Physical Data_2 saludconceptu Dice_ModeFinico saludiconceptual Logical Data_1 Logical Data_1 Logica	Entity_2
🖾 Local 🔯 Repository	•
Output	▲ Ĥ X
General (Check Model) Generation) Reverse /	······································



Damos doble click en la entidad para ingresar su nombre y atributos



Una vez definimos las entidades y atributos procedemos a generar su relación.

Recordemos que en este caso es una base de datos Dimensional tipo estrella, la misma que cuenta con tablas Dim relacionadas a una tabla de hechos o tabla fac.

All PowerDesigner [LDM Prueba, Diagram_1]	
File Edit View Model Symbol Report Repository Tools Window Help	
1 😺 🖆 🚽 🕼 👌 시 역 🐍 × 🤊 한 국 🦁 🛔 🖺 🕲 🖗 🖉 🖢 🗌 🖉 🏈 🛔 한 수 🗛 12 수 주 🛔 🗹 표 🗊 12 12 12 13 17 19 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
Object Browser • # × Diagram_1 Result List	Ţ X Toolbox ▼ # X
Filter Workspace Stadvoll Stadvoll PryciseD stadvoll Dim_pageiente # Logical Data Image: Compage and	 Image: Standard Standard S
🖸 Local 🙀 Repository	•
Output	- ₫ ×
Ceneral (Check Model), Ceneration), Reverse / "	•

Ahora ya tenemos las relaciones entre las entidades Dim y la tabla Fac o tabla de hechos.

Luego de crear todas las entidades y relaciones obtendremos nuestro diseño de la base de datos dimensional.



Una vez que tenemos todo el diseño de la base de datos, ya podemos generar el modelo físico.

୶ PowerDesigner [LDM Logical Data _1, Diagram_1 - C:\Users'	s\Alejandro\Desktop\PROYECTO TRABAJO\Ale	ex infor\hospitalizacion_tungurahua_2016_nivel5.ldm	1
File Edit View Model Symbol Report Repository	Tools Window Help		
	Complete Links	Ctrl+F5	Vamos a Tools y
Object B V 4 X Diagram_1 Result List	Refresh All Related Terms		
Filter:	Check Model	F4	seleccionamos
Norkspac 🔺	Impact and Lineage Analysis	Ctrl+F11	Generar modelo
saludy	Compare Models	Ctrl+F6 # ID parts	Físico
Saludy Physic	Merge Model	Shift+F6 o Edad	FISICO
saludy	Extended Generation	o Edad_a o Edad_d o Etnia	
saludy	Spell Checking Options	o Discapa o Naciona	
Disca_	Execute Commands	o Naciona o Pais	
E-Sei saludic DIM_TIEMPO_EGRE	Mapping Editor	115	
in tiempo egreso Ir in tiempo egreso Ir in tiempo egreso Ir o anio Ir	Generate Objects	•	
o mes lr o dia lr	Generation Links	•	
	Generate Logical Data Model	Ctrl+Shift+L	
	Generate Physical Data Model	Ctrl+Shift+P cie10 hospit	a
	Generate Conceptual Data Model	Ctrl+Shift+C er	-
	License Parameters		
	d Resources	•	
o Especialidad Variable chara	Apply User Profile	les	
	Customize Menus and Tools		
prueba	Display Preferences	ospital	
	Model Options		
	General Options		



A continuación escogemos el tipo de BDD, en nuestro caso seleccionamos Mysql 5.0

Con el modelo físico ya podemos generar el Script para la BDD





A continuación nos presenta la pantalla de Generate Databas

CREACIÓN, BASE DE DATOS EN MYSQL

El gestor de base de datos que usaremos para subir el script anteriormente generado es Mysql Workbench 6.3.

Abrimos Mysql Workbench.	Damos click en la opción Create
MySQL Workbench MySQL MySQL	new schema. Para generar una base de datos.
INSTANCE Startup / Shutdown Server Logs Options File PERFORMANCE Schema @ Performance Reports Schema SCHEMAS %* Output Image: Comparison of Compa	Click en Apply

Listo ahora tenemos nuestra base de datos, creada con el nombre dw_prueba.

En el siguiente proceso tenemos que cargar el archivo con el script generado en el diseño de la base de datos.

En la siguiente imagen observamos el script cargado y damos en ejecutar

MySQL Workbench							
▲ Local instance MySQL57 ×							
<u>File Edit View Query Database Server To</u>	ols <u>S</u> cripting <u>H</u> elp						
	3						
Navigator	dw_prueba - Schema DW6 ×						
MANAGEMENT	🗀 🖬 🐓 🕵 😥 🔂 🥥 🛞 📕 Limit to 1000 rows 🔹 🎭 😻 🔍 👖 🖃						
💿 Server Status	5						
👤 Client Connections							
👤 Users and Privileges	<pre>/ drop table if exists DIM_CIEI0; 8</pre>						
🕎 Status and System Variables	9 • drop table if exists DIM_ESPECIALIDAD;						
📥 Data Export							
📥 Data Import/Restore	11 • drop table if exists DIM_HOSPITAL;						
INSTANCE	13 • drop table if exists DIM_PACIENTE;						
Startup / Shutdown	14 15 B. drop table if evicts DIM IIENDO EGDESO:						
A Server Logs	16						
🖉 Options File	17 drop table if exists DIN_TIEMPO_INGRESO;						
	18 19 B drop table if exists EACT HOSPITALIZACION 2016:						
PERFORMANCE	20						
🖤 Dashboard	21 /**/						
Or Performance Reports	22 /* Table: DIM_CIE10 */						
Or Performance Schema Setup	24 • create table DIM CIE10						
SCHEMAS 😵 🖉	25 📮 (
Filter objects	26 ID_CIE10 varchar(20) not null, 27 GAPITURO varchar(50)						
▶ 🖻 dw	28 primary key (ID_CIE10)						
▶ 🗐 dw6	29 L);						
▼ 🗐 dw_prueba							

Listo ahora tenemos la base de datos cargada con todas las tablas.



CARGA DE LA BASE DE DATOS CON PENTAHO

Pentaho es una herramienta de Business Intelligence de software libre, es una plataforma compuesta de diferentes programas. En nuestro caso usaremos Pentaho Data Integration GeoKettle

Link de descarga: http://community.pentaho.com/

En trasformación 😚 GeoKettle's GUI - Welcome! File Edit View Repository Transformation Job Wizard Help damos click derecho y Ca Welcomel 🔀 Q Â file:///C:/Users/Alejandro/Desktop/PROYECT0%20TRABAJO/Sistemas/Pen 🗎 🧀 🔶 👘 damos en NEW View E ** Explorer Contrar New GeoKettle Spatialytics.org ETL Tool GeoKettle is a "spatially-enabled" version of Pentaho Data Integration (Kettl the use of geospatial (GIS) data. GeoKettle README file • GeoKettle home page Spatialytics.org community Spatialytics forums Spatialytics ETL / GeoKettle documentation Pentaho Data Integration original start page

Descomprimimos el archivo y ejecutamos



Ahora ya tenemos nuestro espacio para trabajar con nuestra primera trasformación.



Damos doble click en

📓 🔒 🐬 🖑 🕫 GeoKettle's GUI - Transformation 1 (changed) Eile Edit View Repository Transformation Job Wizard Help 💥 Transformation 1 😒 읂 Welcomel Q Δ -21 m 191 N 0.0.0 _ 0 X View Design 💦 Excel input Steps Step name Excel Input 🗀 Input Add Field(s) Access Input CSV file input IFiles ISheets Content Error Handling IFields File or directory ♦ <u>A</u>dd <u>B</u>rowse... De-serialize from file Regular Expression Selected files: # ^ File/Directory Wildcard (RegExp) Required 🐴 Excel Input Fixed file input 1 Delete Edit 🧐 Generate random value 📄 Get File Names Accept filenames from previous steps Accept filenames from previous 📃 🗒 Get Files Rows Count 🔒 Get SubFolder names 🖥 Get System Info Get data from XML KML File Input LDAP Input LDIF Input Show filename(s)... OK Cancel Preview rows Mondrian Input 🔓 Property Input 🔂 RSS Input S3 CSV Input

En la opción Browse seleccionamos nuestro archivo en formato xls y añadimos

Marcamos la siguiente pestaña ¡Sheets





En la pestaña Fields cargamos los atributos de la tabla



Una vez que tenemos la entrada de los datos, ahora seleccionamos el destino que en este caso será una base de datos.



En la carpeta Output seleccionamos Table output y arrastramos a nuestra hoja de trabajo

Creamos la conexión con el archivo de entrada a la base de datos, mantenemos Ctrl y click izquierdo arrastramos a la salida de la bdd





Buscamos Mysql para crear la conexión con la base de datos, configuramos la conexión y probamos con un Test

📲 🖬 🕤 - 🗇 -		MANUAL PROCESO ETL - Microsoft Word
🞊 GeoKettle's GUI - Transform	ation 1 (changed)	
Eile Edit View Repository	<u>Transformation</u> Job <u>W</u> izard <u>H</u> elp)
View Design	🔋 Database C	connection 23
tent Design	General	
Steps	Advanced	Connection Name:
input	Deptions	Prueba
Output	Clustering	
Access Output		Connection Type: Settings
🙀 CSW Output 🖉		KingbaseES A Host Name:
🔕 Delete	Message:	LucidDB localhost 🏶
🏄 Excel Output 👔		MS Access Database Name:
📲 GML File Output	Connection to database [Prueba]	is OK. MaxDB (SAP DB)
🕰 Insert / Update	Hostname : localhost	MonetDB
📲 KML File Output	Port : 3300 Database name : DW/ T	MySQL Port Number:
📲 OGR Output	Deceber Harrie . Dw_r	Neoview 3306
🏓 Properties Output		Oracle User Name:
🛐 RSS Output	OK	Oracle RDB
🖉 SQL File Output		Palo MOLAP Server
Serialize to file	2	PostgreSQL Password:
Shapefile File Output	t	SAP R/3 System
Synchronize after me	erge	Access:
Table output		Native (IDRC)
C Undate		ODBC
ML Output		JNDI
Transform		
Dtility		
Flow		
Cripting		Test Feature List Explore
😑 Lookup		
🗀 Joins		
😑 Data Warehouse		OK Cancel
🗀 Validation		
Statistics	<u></u>	

dench				
GUI - Transformation 1 (changed)				
jew <u>R</u> epository <u>T</u> ransformation <u>J</u> ob	<u>W</u> izard <u>H</u> elp			En la onción target table
A	🔓 Welcome! 🛛 💥 Transformati	on 1 🛛		Li la opcion la get lable
Design		r 🎋 🖻 🔯 🎨 🦺 🥵 🔚 100%	▼	ingresamos en Browse
	📁 Table output			
		Step name cie10		
Database explorer on connection (Prueba		Connection Prueba	- Edit. New.	
🛭 宿 Prueba	E	rget schema		
Catalogs	Preview first 100 rows of [?]	Target table		
din, cietu	Preview first rows of [?]		Browse	
📰 dim_especialidad	Number of rows of [?]	Commit size 1000	•	
dim_hospital		runcate table		
dim_tiempo_eqreso	Channel and (2)	tabase fields		
dim_tiempo_ingreso	Show layout or [1]	se fields		
fact_hospitalizacion_2016	Generate DDL			
012403	Generate DDL for other connection	ita over tables 📃		
	Open SQL for [?]	itioning field	- 4	Seleccionamos la tabla
		ita per month @		que queremos cargar
	Truncate table [?]	h data per day 🔘		que que en los cargai
		ate for inserts 📝		
		ame of table:	- 4	
		blanama field	¥	
OK Cance	I Refresh	generated key 📃		
varenouse	J	Pated key field		
:ion		OK Cancel SQI		
0				
na –				

A lado izquierdo seleccionamos los campos de la base de datos, los campos de la derecha contienen datos de las tablas de Excel.

🗐 🗖 🦘 🖑 🕫	MANUAL PROCESO ETL - Microsoft Word	
GeoKettle's GUI - Transformation 1 (changed)		
jile <u>E</u> dit <u>V</u> iew <u>R</u> epository <u>T</u> ransformation <u>J</u> ob	<u>W</u> izard <u>H</u> elp	
	😭 Welcome! 🔀 Transformation 1 🛛	
View Design	B Table output	_ 0 X
eps 🛛 🕞		
Contract in the second	Step name cie10	
Output	Connection Prueba	T Edit Nam
Access Output		Euron Inewin
CSW Output	Target schema	*
🔯 Delete	Target table dim cie10	Browse
🚁 Excel Output		
🗿 GML File Output	Commit size 1000	
🖴 Insert / Update	Truncate table 📃	
🖓 KML File Output	Ignore insert errors	
🖓 OGR Output	Specify database fields 🔽	
🍻 Properties Output	Main options Database fields	
SS Output		
SQL File Output	Fields to insert:	
Serialize to file	# Table field Stream field 🔨	Get fields
Shapefile File Output	1 ID_cie10 - ID_cie10	Enter field manning
Synchronize after merge	2 ID_CIE10 Capitulo	
Table output	CAPITULO	
C Undate		
A XML Output		
Transform		
🔁 Utility		
Flow		
😑 Scripting		
💼 Lookup		
😑 Joins		
💼 Data Warehouse		
Calidation	OK Cancel SOL	
C Statistics		
😑 Job		

Cargamos los campos de la base de datos

Para verificar que los datos están correctamente relacionados procedemos con ejecutar la trasformación.



M 🗄 🔊	💥 Exec	ute a transforn	nation							
😥 GeoKettle's G	-Local.	remote or clus	tered exec	ution						
<u>File E</u> dit <u>V</u> iev	O Loc	al execution			Execute remotely		0	Execute clustere	d	
					Remote host			Doct transform	nation	
							*	Prenare ever	ition	
View					Pass export to remote server			Start executio	n	
Steps								Show transfo	rmations	
📄 Innut										
Coutput	Detail:	5								
🔊 Acces						📝 Ena	able safe mode			
🖏 CSW						🔽 Cle	ar the log before execution			
🙆 Deleti	Logle	vel				Basic	logging			*
🖉 Excel	Replay	/ date (yyyy/M	M/dd HH:	mm:ss)						
SIVIL										
🖓 KML I	Parame	ters				Variables				
🗿 OGR (#	Parameter	Value	Default value		#	Variable		Value	
🌽 Prope	1					1	Internal.Job.Filename.Dire	ctory	Parent Job File Directory	
🔊 RSS C						2	Internal.Job.Filename.Nam	ne	Parent Job Filename	
SQL F						3	Internal.Job.Name		Parent Job Name	
Serial						4	Internal. Transformation. Re	epository.Dire	/	
Synch										
a Table										
a) Text f										
😋 Upda	Argume	nts								
	# ^	Argument	Value							
Iranstorr	1									
Elower	-									
💼 Scripting										
🗀 Lookup										
😑 Joins										
😑 Data War										
Validation										
loh						Launch				
Mapping					Launch	C	ancel			
							-			



En la pestaña de resultados verificamos que no exista ningún error en la trasformación.

A continuación verificamos los datos que se encuentran en la base de Mysql

MySQL Workbench	
Local instance MySQL57 ×	
<u>File Edit View Query Database Server Tools Scripting Hel</u>	3
Navigator	DW7 dim_cie10 · Table SQL File 6" 🗙
MANAGEMENT **	
Server Status	
Client Connections	select in directed
Users and Privileges	
Status and System Variables	
A Data Export	
A Data Import/Restore	
INSTANCE 🖏	
🚦 Startup / Shutdown	
🛕 Server Logs	
🖋 Options File	Result Grid 🛄 🚷 Filter Rows: Edit: 💋 🐯 🔂 Export/Import: 📳 🔯 Wrap Cell Content: III
PERFORMANCE	ID_CIE10 CAPITULO A00_R00_L_LCiertos opformediates infessionas u parasthatiss
🚳 Dashboard	COD P48 II Neolasias
🐻 Performance Reports	D50 D53 III Enfermedades de la sanare v de los organos
👌 Performance Schema Setup	E00 E90 IV Enfermedades endocrinas, nutricionales y me
	CON CONTRACTOR DE CONSTITUTION DE CONSTITUTION DE CONTRACTOR DE CONSTITUTION DE CONSTITUCION DE CONSTITUCICON DE CONSTITUCICON DE CONSTITUCICON DE CONSTITUCICON DE CO
SCHEMAS 😻 🖬	H00 H59 VII Enfermedades del oio vius anexos
R Filter objects	H60 H95 VIII Enfermedades del oido v de la apofisis mast
▶ 🥃 dw 🔹	IO0 I99 IX Enfermedades del sistema reconstructorio
▼ 🗏 dw_7	K00 V93 X Enformedades del sistema diaestivo
Tables	L00 L99 XII Enfermedades de la piel y el tetido subcutaneo
dim_cisto	M00 M99 XIII Enfermedades del sistema osteomuscular v
dim_hospital	000 099 XV Embarazo, parto y puerperio
dim_paciente	P00 P96 XVI Ciertas afecciones originadas en el periodo
dim_tiempo_egreso	000 O99 XVII Malformaciones congenitas, deformidades
dm_tiempo_ingreso fact bospitalizacion 2016	R00 R99 XVIII Sintomas, signos y hallazgos anormales cli
	dim_deto 1 X Apply Revent