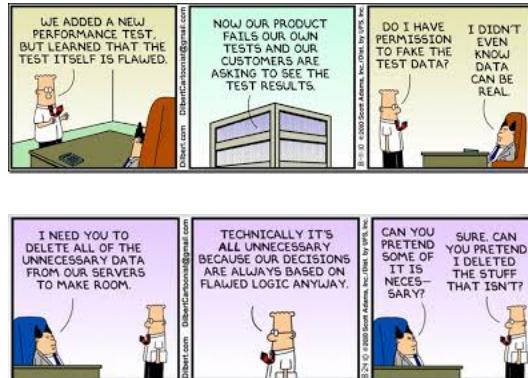




Data Warehousing



Karim Pichara
 Computer Science Department
 Pontificia Universidad Católica de Chile

Data everywhere: Problem



- I can't **get** the data I need
 - need an expert to get the data
- I can't **find** the data I need
 - data is scattered over the network
 - many versions, subtle differences
- I can't **understand** the data I found
 - available data poorly documented
- I can't **use** the data I found
 - results are unexpected
 - data needs to be transformed from one form to other

Algunos Tipos de Bases de Datos Actuales

Operacionales



Your order qualifies for FREE Super Saver Shipping ([Restrictions apply](#)). Choose this option at checkout.

Subtotal (1 item): \$109.00

This order contains a gift

Proceed to Checkout

or

[Sign in](#) to turn on 1-Click ordering.

Analíticas



Karim Pichara B.

PUC Chile

Bases de Datos Operacionales (Transaccionales)

- Cubren los aspectos operacionales de una organización
- Son las más comunes hoy en día
- On-Line Transaction Processing systems (OLTP)
 - Compras
 - Inventarios
 - Pagos
 - Registros de clientes
 - etc...

Karim Pichara B.

PUC Chile

Bases de Datos Operacionales

United Confirmation Number: I4WS3K

FLIGHT	DEPARTING	ARRIVING	AIRCRAFT	DURATION
UA1122	6:04 a.m. Wed., Jun. 25, 2014 Boston, MA (BOS)	9:44 a.m. Wed., Jun. 25, 2014 San Francisco, CA (SFO)	Boeing 737-900	Flight Time: 6 hr 40 mn

Fare Class:
United
Economy (V)

Meals: Meals for Purchase
No Special Meal Offered.

Traveler Information:

Mr. KARIME PICHARA

Seat Assignments: BOS - SFO: 24D

Karim Pichara B.

PUC Chile

Bases de Datos Analíticas (Data Warehouses)

- Cubren los aspectos estratégicos de una organización
- Se utilizan para analizar la información en búsqueda de conocimiento relevante
- Son cada vez más comunes
- On-Line Analytical Processing systems (OLAP)
 - Business Intelligence
 - Data Mining
 - Customer Relation management (CRM)

Karim Pichara B.

PUC Chile

Bases de Datos Analíticas

Página principal Informes estándar Informes personalizados

Mi panel

Visitas diarias

1 ene 8 ene 15 ene 22 ene

Tipos de tráfico

Tipo de tráfico	Porcentaje
feed	25.70%
orgánico	24.90%
referencias	23.05%
directo	14.85%
correo electrónico	7.35%

Duración de la visita por país

País/Territorio	Visitas	Duración media de la visita
Estados Unidos	67.445	00:01:54
Reino Unido	18.948	00:01:37
India	8.882	00:00:58
Canadá	6.371	00:01:02
Alemania	5.845	00:00:32
Francia	5.243	00:00:38

Karim Pichara B.

Data Warehousing

- **¿Qué es un Data Warehouse?**

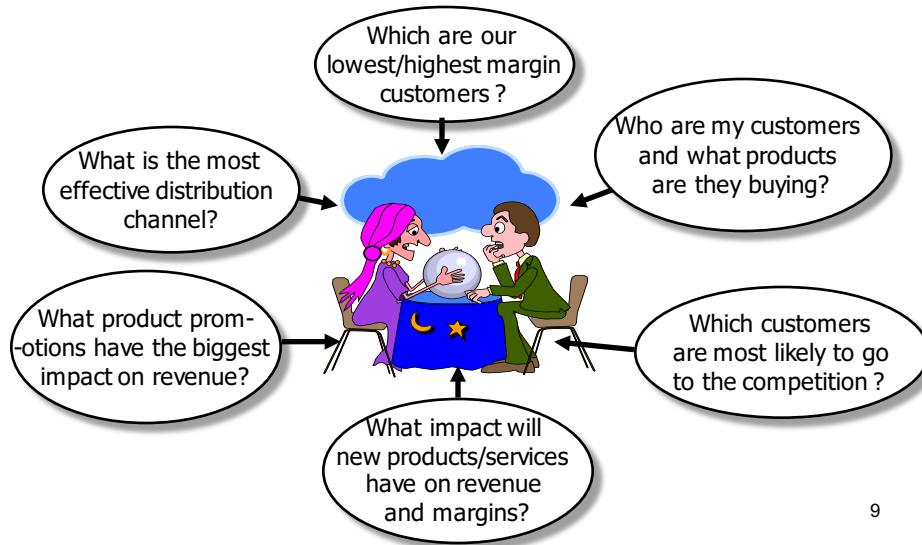
“A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.” Barry Devlin

Welcome to
Data Warehouse
ASU

Student Database

Karim Pichara B. PUC Chile

Why Data Warehouses?



9

- Características:

- **Orientada a temas específicos:** No concentra información tan detallada como transacciones diarias, más bien guarda info sobre temas más generales como cliente, proveedor, producto y ventas.
- **Integrada:** Integra varias fuentes de datos heterogéneas, como bases de datos relacionales, archivos planos, registros de transacciones, etc.
- **Varía en el tiempo:** Como contiene información histórica se actualiza cada cierto tiempo y guarda las fechas a las cuales corresponde la información

- **No volátil:** Un DW está siempre separada físicamente de los datos guardados en bases de datos operacionales, debido a esto un DW no requiere herramientas de proceso de transacciones, recuperación o mecanismos de control de recurrencia. Sólo requiere dos herramientas de acceso, carga *initial* de datos y acceso a la información.

Karim Pichara B.

PUC Chile

Más características:

- El acceso a la información permite sólo la lectura de datos.
- Contiene mucha información.
- La dinámica es lenta.
- Puede contener información redundante.
- Contiene metadata (datos sobre los datos)

DW de Wal-Mart (DW de retail más grande del mundo)

- 320 Gb 1990, 1 TB en 1992, hoy ~ 2.5 PB
- Datos históricos, 65 semanas
- Decenas de Miles de usuarios
- más de 1000 millones de transacciones diarias

Karim Pichara B.

PUC Chile

Ejemplo de una vista para un DW

The screenshot shows the SAP Enterprise Portal 5.0 interface. At the top, there's a blue header bar with the SAP logo and the text "SAP Enterprise Portal 5.0 - Microsoft Internet Explorer proporcionado por LAN". Below the header is a menu bar with items like "Home", "SAP Business Warehouse", "Mi Intranet", "Presupuesto", "Directorio Personas", "Autoservicio Personas", "Aplicaciones", "Políticas y Procedimientos", "Correo", "Documentación SAP", "Navegación en Portal", "Procedimientos Comerciales", and "Sist de Int". A search bar and links for "Personalizar Página", "Portal", and "Agregar a Favoritos" are also present. The main content area is titled "Bienvenido" and "SAP Business Warehouse". On the left, there's a navigation tree under "Sistemas de Información" with categories like "Controlling", "Gerencia General Pax", "Gerencia Comercial", "Gerencia Lan Express", "Gerencia Experiencia de Viaje", "Gerencia General Carga", "Control de Gestión Soporte", "Control de Gestión y Planificación VPT", "Contraloría Corporativa", "Sistemas de Información", "Gerencia de Negocios Aeroportuarios", and "Otros Reportes Corporativos". The right side of the screen shows a large blue background area with the SAP logo in the top right corner.

Karim Pichara B. PUC Chile

Ejemplo de una vista para un DW

Dotaciones Responsable Jerárquico

* La Estructura de Reportable jerárquico se carga de HR una vez al mes, con fecha de corte último día del mes anterior. Esta información aparece actualizada después del cierre contable (aprox los 15 de cada mes). Cualquier modificación en la estructura o responsabilidad de centros de costo contactar a Gestión Estructura (gestion.estructura@lan.com).

Blóque navegación:

Centro de coste	Funció	Mes natural	Nodo Responsable
País	Sociedad	Tipo Contrato	Estructura

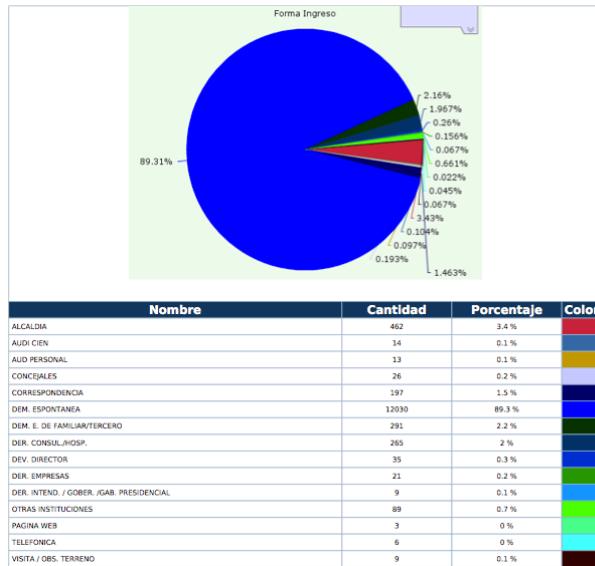
Dotaciones Responsable Jerárquico

Numeros

This screenshot shows a detailed view of the "Dotaciones Responsable Jerárquico" report. It features a hierarchical navigation pane on the left with sections for "Mes natural" (July, June, May, April) and "Nodo Responsable" (ENRIQUE CUETO, IGNACIO CUETO, ARMANDO VALDIVIESO, FRANCISCO QUIMPERT, CRISTIAN URETA, VLAMIR DOMIC, DAMIAN SICKIN, BRUNO ARDITO, MARCO JOFFRE, MARCO JOFFRE, FRANCISCO SOTOMAYOR, JORGE INHEN, CRISTIAN LEON, ANGELA CORRALES, GERENCIA REPRESENTACION TECNICA, MAGDALENA SPATE). The main area displays a grid of financial data for these nodes across the four months. The grid cells are color-coded in shades of red, green, and yellow, with some cells containing numerical values. A large watermark reading "Numeros" is overlaid across the grid area.

Karim Pichara B. PUC Chile

Ejemplo de una vista para un DW



Karim Pichara B.

Ejemplo de una vista para un DW



Karim Pichara B.

Bienvenido Felipe Castillo

Gerencia Informe por Canal

Ambas Marcas MOBIL ESSO

Resumen Gerencia - Ambas Marcas

Informe Enero a Enero de 2010

Periodo	CANAL	Acumulado
Periodo '10	Periodo '09 % Crec. POA Comp.POA	Accum.'10 Accum.'09 % Crec. POA Comp.POA
Con. (MM \$)		
	Estaciones de Servicio	
	Distribuidores	
	Industrial	
	Competencia y Exportación	

Información Detallada - Oficina - Contribución (MM \$)

Cen. 2010	Cen. 2009	% Crec.	POA Com	Comp.POA	Accum. Cen. 2010	Accum. Cen. 2009	% Crec.	POA Com	Comp.POA
Oficina					Dis. Norte-Est				
					Dis. Santiago-Centro				
					TCT				
					Total				

Canal: Distribuidores

Margen (\$...)

Competencia y Exportación

Total País

Agua Detalle de Reventa
 Cuidado Automotriz Detalle E/S

Bienvenido Felipe Castillo

INICIO · DOCUMENTOS · NOTICIAS · PRODUCTOS · LABORATORIO · REPORTES
COPEC Mobil Planner © 2010 Todos los derechos reservados.

Lista

Bienvenido Felipe Castillo

Reporte-Informe de Mix - Windows Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Correo Gmail - Redes INICIO ... Mobil Planner Reporte... Reporte-Cu... Reporte-Cu... Reporte-Cu... Página Herramientas

Cuidado resumen

Ambas Marcas

Informe de Enero a Enero 2010

Z- 110 (claudio Rojas) - Informe Mix

Doble click en Mix para ver aplicaciones

Mix	Volumen			Contribución			Margen				
	Per.2010	Per.2009	Crec. %	Mix/Tot. %	Per.2010	Per.2009	Crec. %	Mix/Tot. %	Per.2010	Per.2009	Crec. %
A - Flagship	14.327	17.719	-19.3%	3.0%	1.594.833	38.370.653	-95.9%	1.4%	111	2.165	-94.9%
B - Premium	43.367	26.945	49.8%	9.0%	5.839.409	10.797.295	-45.5%	5.2%	134	373	-63.9%
C - Estándar	87.495	122.003	-28.3%	18.1%	23.397.427	24.992.004	-6.4%	20.8%	267	204	30.5%
D - Competitivo	52.154	57.655	-9.5%	10.8%	15.901.873	30.124.969	+65.3%	14.2%	306	532	+41.4%
Otros	295.514	302.459	-2.7%	55.1%	65.693.259	84.391.967	+22.1%	50.4%	230	278	+17.5%
Total General	492.897	528.981	-6.7%	100.0%	112.507.094	188.666.791	+49.4%	100.0%	233	356	+24.7%

Evolución de Mix año 2010 (Lts)

Lineas... Rubros...

Bienvenido Felipe Castillo

INICIO Microsoft SQL... Fras_Verka Reporte-Info... Misiones Trek... Downloads SAP Logon 710 compardia http://apus.c... http://apus.c... http://apus.c... Venta Cliente... http://apus.c... Facturación L... Microsoft Pow...

100% 11:28 Martes

ComScore

comSCORE | MyMetrix

Cross Visiting Key Measures Cross Visiting (1) Key Measures (1) Key Measures (3)

	Media	Total Unique Visitors (000) ~	% Reach	% Composition Unique Visitors	Composition Index UV	Composition Index PV	Average Daily Visitors (000)	Total Minutes (MM)	Total Pages Viewed (MM)	Total Visits (000)
1	Total Internet - Total Audience	7,114	100.0	100.0	100	100	3,751	10,059	12,376	324
1	News/information	5,151	72.4	100.0	100	100	1,089	197	254	39
1	LUN	1,727	24.3	100.0	100	100	307	89	103	12
2	LATERCERA.COM	1,148	16.1	100.0	100	100	122	24	22	5
3	Medios Regionales	672	12.3	100.0	100	100	82	9	21	2
4	TVN Online	751	10.6	100.0	100	100	48	5	9	1
5	ELMUNDO.ES	711	10.0	100.0	100	100	42	7	9	1
6	El Mercurio	600	8.4	100.0	100	100	49	6	13	1
7	METEODECHILE.CL	597	8.4	100.0	100	100	54	3	5	1
8	Emol Noticias	573	8.1	100.0	100	100	50	3	3	1
9	CLARN.COM	489	6.9	100.0	100	100	33	1	2	1
10	COOPERATIVA.CL	443	6.2	100.0	100	100	38	9	4	1
11	Grupo La Nación	438	6.2	100.0	100	100	22	4	6	
12	Terra News	392	5.5	100.0	100	100	28	1	2	
13	LANACION.CL	341	4.8	100.0	100	100	19	3	4	
14	LASEGUADA.COM	278	3.9	100.0	100	100	25	4	5	
15	MSN News	272	3.8	100.0	100	100	12	1	1	
16	20MINUTOS.ES	234	3.3	100.0	100	100	10	0	1	
17	Vilka	228	3.2	100.0	100	100	9	0	0	
18	ATTNACHEL.CL	199	2.8	100.0	100	100	8	0	0	
19	CNN Network	199	2.8	100.0	100	100	20	1	2	
20	ELPAIS.COM	199	2.8	100.0	100	100	10	0	1	
21	ELSUR.CL	149	2.1	100.0	100	100	18	3	5	
22	Que es Sites	145	2.0	100.0	100	100	6	1	1	

News and information V. U

Google Analytics

Analítica web para empresas

Página principal Informes estándar Informes personalizados

Visitas diarias Tipos de tráfico Duración de la visita por país

Herramientas de medición para su empresa

Estadísticas de varios canales Soluciones para móviles Informes sociales

¿ Y qué es entonces Data Warehousing?

Data Warehousing → Proceso de construir un Data Warehouse

Este proceso requiere:

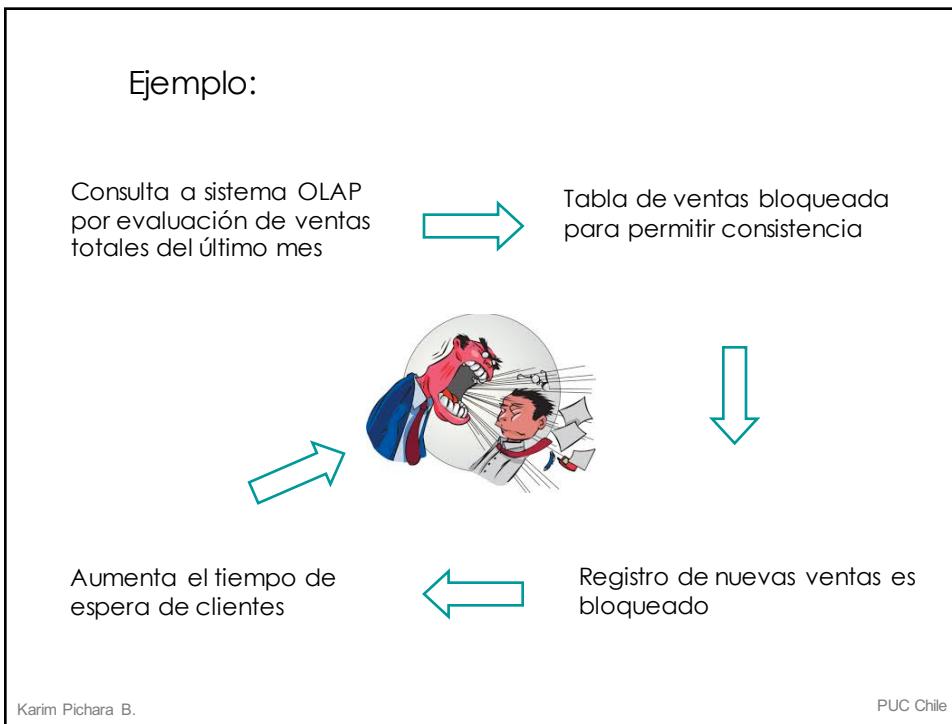
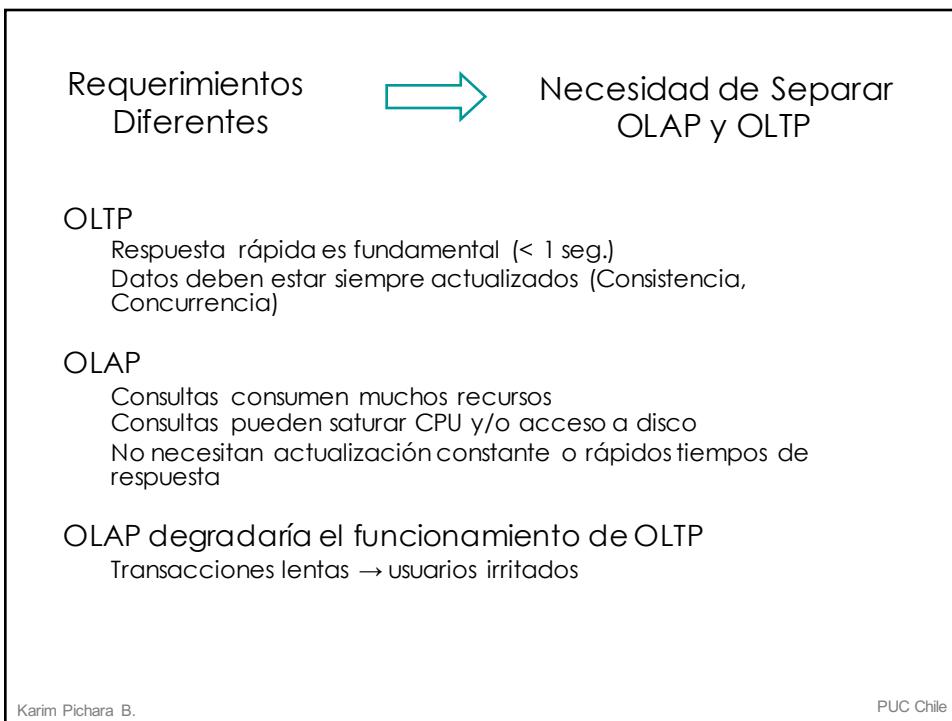
- Filtrado de los datos (Data Cleaning)
- Integración de los datos
- Consolidar los datos

Karim Pichara B. PUC Chile

OLTP vs. OLAP

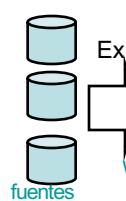
- **OLTP: On-Line Transaction Processing**
 - Muchas transacciones cortas (consultas y actualizaciones)
 - Ejemplos:
 - Registrarse en un sitio web
 - Registrar movimientos de cuenta Bancaria
 - Ingresar la compra de un cliente en una tienda
 - Preguntas simples, poca información (sólo un par de registros)
 - Actualizaciones frecuentes
 - Conurrencia es un gran problema
- **OLAP: On-Line Analytical Processing**
 - Transacciones largas y complejas
 - Ejemplos:
 - Mostrar el reporte de ventas de cada departamento este mes
 - Mostrar los clientes que movieron más de \$1US Mill en un mes
 - Identificar los productos más vendidos en la semana
 - Preguntas complejas, requieren procesar mucha información
 - Actualizaciones son poco frecuentes
 - Cada consulta puede consumir muchos recursos

Karim Pichara B. PUC Chile

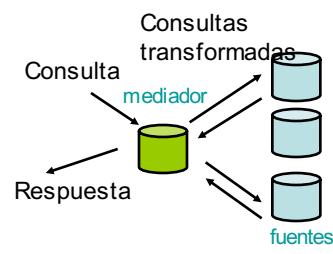


Tipos de Integración

- Eager (LAV) Integration: El DW crea copias condensadas de los datos y ejecuta comandos sobre esta copia
- Lazy (GAV) Integration: no se integra previamente a cada consulta, pregunta directamente a bases de datos distribuidas



Local as View



Global as View

Karim Pichara B.

PUC Chile

Eager vs Lazy Integration

- Ventajas de integración floja:
 - Evitan copias redundantes de la información
 - Otorga mayor flexibilidad a sistemas de seguridad
- Desventajas de integración floja:
 - Carga extra al sistema de consultas
 - Información histórica puede no estar disponible
 - Los interpretes (“wrappers”) requieren de gran complejidad
- Ventajas Integración activa:
 - Es mucho más común en la práctica
 - Mejor rendimiento
 - Menor complejidad
 - Datos antiguos se manejan mejor

Karim Pichara B.

PUC Chile

Data Mart

- Es como una DW pero más pequeña y específica
- Generalmente dedicada a un departamento particular de la compañía
- Problema frecuente es heterogeneidad de cada departamento que dificulta la integración de las data marts

Karim Pichara B.

PUC Chile

Ej: Data Mart

PREGUNTAS_PPROV		PRTAL_AA7ERAL												
D_AY_MV	NEODC_M	INTERNAL			TICKET			Real			Virtual			
Total Avión Mtro	Total Avión Mtro	02424577.00	103.571.700,00	115.855.20,00	99.626.50,77	95.479,40	94.589.02,31	120.983.45,27	103.912.07,51	100.862.80,00	18.763.059,77	120.041.55,83	124.242.45,40	147.245.47,14
Total LanChile	Total LanChile	98.438.761,32	81.159.821,00	85.510.32,72	77.025.763,60	72.703.52,77	70.520,76,43	87.468.32,77	79.634.51,39	82.041.00,77	87.238.44,30	86.361.01,88	86.839.45,27	93.501.207,17
Total INT LA	Total INT LA	72.876.886,53	59.688.934,00	62.176,30,34	56.873.51,27	54.210,95,5	52.201.73,89	66.557.69,72	58.932.27,17	69.979.73,28	62.318.76,10	65.822.97,24	70.222.13,11	78.945.28,69
Internacional Lan Chile	Internacional Lan Chile	72.876.886,53	59.688.934,00	62.176,30,34	56.873.51,27	54.210,95,5	52.201.73,89	66.557.69,72	58.932.27,17	69.979.73,28	62.318.76,10	65.822.97,24	70.222.13,11	78.945.28,69
INTL	INTL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
BUE	BUE	6.194.388,36	5.598.430,00	6.914.559,96	6.451.262,64	6.948.590,00	6.051.000,71	7.237.074,19	6.886.244,67	7.033.446,04	7.022.944,64	8.951.78,81	8.690.005,72	7.899.440,62
CDS	CDS	72.478.20,79	115.248,40	117.208,69	105.213,52	96.745.15	99.322,94	130.446,07	145.570,24	152.623,57	125.957,29	139.006,30	157.785,36	169.850,62
CDR	CDR	73.478.20,79	115.248,40	117.208,69	105.213,52	96.745.15	99.322,94	130.446,07	145.570,24	152.623,57	125.957,29	139.006,30	157.785,36	169.850,62
CUN	CUN	903.353,04	891.392,97	65.647,55	72.912,37	728.158,79	760.073,07	931.495,08	984.475,25	931.34,47	947.245,54	827.899,2	-	948.427,25
DLI	DLI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EUR	EUR	0.20.471,01	9.902.79,00	10.071.44,46	9.747.00,07	9.839.01,02	9.952.60,3	9.402.20,46	10.74.31,0	12.00.36,67	12.45.63,14	10.943.56,13	11.91.20,34	11.91.20,34
GBO	GBO	2.556.10,00	2.556.10,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00	1.420.00,00
HAY	HAY	2.09.376,81	153.927,69	140.977,68	152.425,06	160.521,40	159.246,69	139.765,01	127.775,83	165.095,72	126.649,33	645.823,53	2.599.846,7	2.599.846,7
HPC	HPC	1.19.01,14	1.051.69,50	943.427,13	996.02,37	21.613,71	191.05,00	617.525,24	700.317,57	541.12,68	752.697,15	1.19.239,57	1.076.199,31	1.32.630,61
LAX	LAX	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
LIM	LIM	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
MCI	MCI	9.76.72,85	4.11.00,01	636.42,57	542.194,98	9.747.00,07	9.839.01,02	10.952.60,3	9.402.20,46	10.74.31,0	12.00.36,67	12.45.63,14	10.943.56,13	11.91.20,34
MDC	MDC	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
MEX	MEX	3.727.444,42	3.000.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00	3.780.00,00
MIA	MIA	5.951.777,76	7.146.742,01	6.591.0319	5.393.774,62	5.174.297,60	5.018.047,04	7.85.779,13	5.945.79,60	6.659.83,36	4.648.06,28	6.028.774,04	8.908.952,07	10.870.002,30
MVO	MVO	1.094.432,02	1.00.37,07	987.34,47	737.954,33	784.79,60	892.43,20	768.82,16	877.81,21	945.987,31	1.097.606,42	1.122.768,71	1.344.452,19	1.344.452,19
MVO	MVO	1.094.432,02	1.00.37,07	987.34,47	737.954,33	784.79,60	892.43,20	768.82,16	877.81,21	945.987,31	1.097.606,42	1.122.768,71	1.344.452,19	1.344.452,19
NYC	NYC	7.045.834,94	5.502.430,00	8.627.055,88	6.851.413,83	6.031.680,37	5.72.187,29	8.829.70,52	7.243.95,34	7.013.08,09	7.022.88,28	7.047.945,87	7.300.793,7	7.694.595,47
PPT	PPT	2.209.528,65	1.841.23,47	1.801.04,78	1.209.857,24	1.388.497,69	1.314.50,21	1.648.632,46	1.504.50,78	1.241.94,36	1.494.85,28	1.688.68,30	1.825.873,30	2.082.395,65
POD	POD	1.287.970,14	1.250.805,80	507.086,87	622.965,96	595.089,95	746.37,95	1.080.374,62	810.025,20	707.493,26	711.759,07	591.002,40	1.045.770,98	-
ROS	ROS	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
SSA	SSA	406.769,28	510.721,43	24.561,04	197.950,44	195.38,65	163.95,45	264.42,28	45.09,07	70.27,14	65.24,19	277.79,03	79.452,63	222.035,54
SRY	SRY	3.164.004,27	3.032.504,42	3.811.70,21	3.280.428,63	2.830.01,71	2.848.93,07	3.067.413,72	2.978.83,06	3.093.714,7	3.318.41,44	3.603.32,63	3.812.42,44	4.032.51,44
UJO	UJO	1.453.625,24	1.240.073,71	1.400.73,29	1.341.493,26	1.341.393,37	1.362.700,94	1.361.493,77	1.258.59,46	1.351.137,41	1.422.39,78	1.377.50,98	1.44.393,95	1.44.393,95
LGM	LGM	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
LBN	LBN	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
ECU	ECU	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
LNB	LNB	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
SON	SON	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
CAT	CAT	0.00.960,05	0.00.16,00	6.649.40,15	8.46.008,57	0.790.76,40	5.794.24,04	0.562.24,04	6.32.30,21	8.04.12,41	6.56.12,41	0.870.42,02	0.76.00,00	0.76.00,00
REVISA	Producción_Total	X Tickets	X Rechazos	X Caducos	X Otros_Inv_violados	X Ex_Equip_Inv_UPass	X INGOUTDI	X Arribaciones	X Otros_Inv_no_violado	X COMISIONESPAX	X C...			

Karim Pichara B.

PUC Chile

Diseño de un DW

- **Enfoque Top-Down:**

- Comienza con el diseño completo y planificación total.
- Es útil cuando la tecnología es bien conocida, madura, estable y cuando los problemas de negocios a resolver se conocen bastante bien.

- **Enfoque Bottom-Up:**

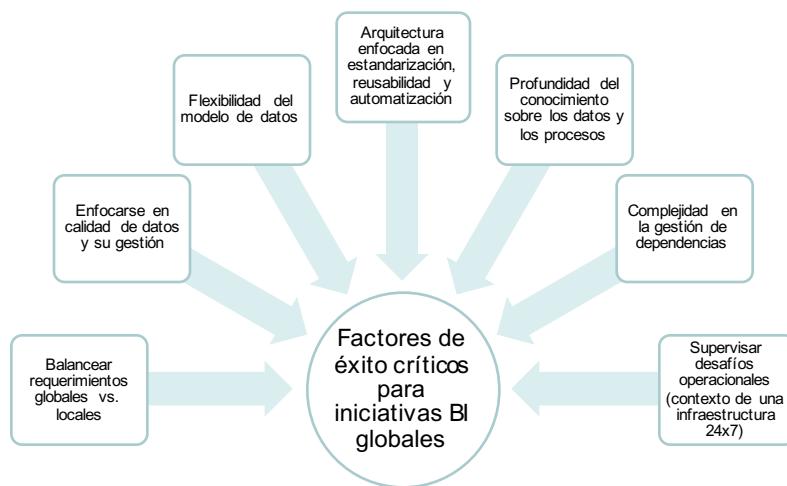
- Comienza con experimentos y prototipos
- Permite avanzar de a poco asegurando que cada inversión que se hace es necesaria
- Es útil cuando se requiere una rápida y oportuna implementación.

- Combinando los dos anteriores

Karim Pichara B.

PUC Chile

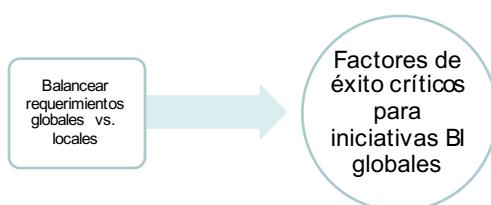
Iniciativas de BI globales: 7 factores de éxito



Fuente: Murali, 2012

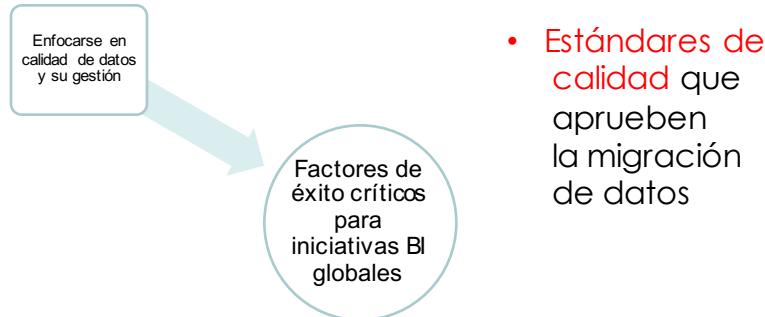
1: Balancear requerimientos globales vs. locales

- Requerimientos globales de una iniciativa BI deben estar alineados con los requerimientos de usuarios locales
- El entendimiento de la información debe ocurrir tanto en niveles corporativos como en las unidades de negocio



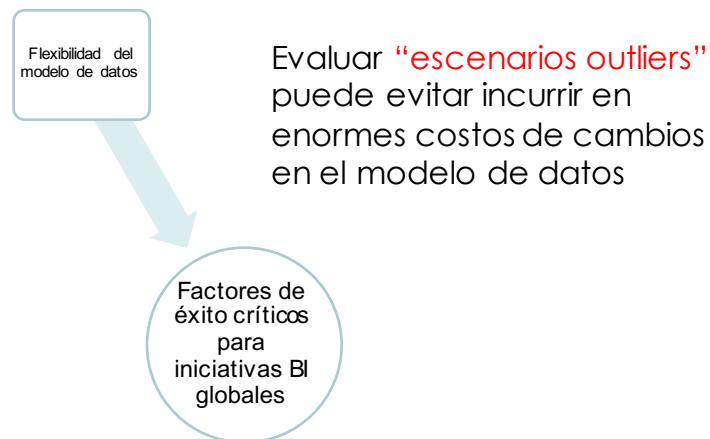
2: Enfocarse en calidad de datos y su gestión

- El atributo más complicado en una solución BI a gran escala es la calidad de los datos
- Deben crearse:
 - datos maestros específicos por región
 - team centralizado de alineamiento corporativo



3: Flexibilidad del modelo de datos

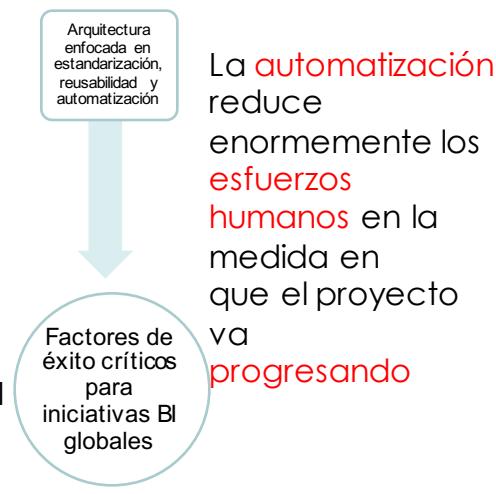
El modelo de datos de un proyecto BI debe ser capaz de asegurar **escalabilidad y adaptabilidad**



4: Arquitectura enfocada en estandarización, reusabilidad y automatización

La arquitectura de una iniciativa BI debe ser **estándar para facilitar la re-usabilidad y la automatización**.

Siempre debe mantenerse el **foco en la re-usabilidad** durante el desarrollo de un sistema BI (evita ineficiencias).



5: Profundidad del conocimiento sobre los datos y los procesos

El **equipo de desarrollo** debe entender los **procesos** de negocios, los **matices** de los procesos y el **significado** de los **datos**

Profundidad del conocimiento sobre los datos y los procesos

Factores de éxito críticos para iniciativas BI globales

Las **ramificaciones** de los cambios que se hacen en **etapas tardías** resultan en **grandes cantidades de re-validaciones** de datos que deben volver a ejecutarse.

6: Complejidad en la gestión de dependencias

En general existen **altos grados de dependencia** entre las iniciativas BI. Todas estas interdependencias deben ser rigurosamente gestionadas.

Se deben analizar los cambios planificados para el futuro y **evaluar el impacto** de esos cambios.

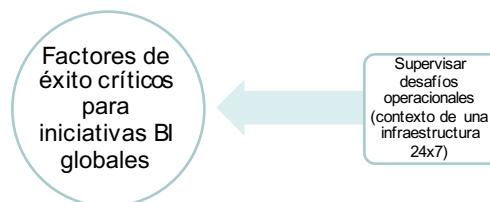
Complejidad en la gestión de dependencias

Factores de éxito críticos para iniciativas BI globales

Se debe asegurar que no se necesitarán **cambios mayores** en las **fases críticas**

7: Supervisar desafíos operacionales (contexto de una infraestructura 24x7)

Deben generarse **evaluaciones de tolerancia a fallas** para asegurar la disponibilidad del sistema de reporting (**actualizado**) en casos en que sea necesario hacer un “**reloading**” de la información en alguna **localidad**.



Arquitecturas de modelamiento de los datos

- Se debe modelar la información de tal forma de facilitar las labores del Data Warehouse (Responder consultas, generar reportes, aplicar algoritmos de Data Mining)

Algunos Conceptos

- Diseño Lógico:

- Organización conceptual de la base de datos
 - Etapa de Modelación

- Diseño Físico:

- Seleccionar Estructuras (tablas, índices, hardware, etc.)
 - Organizar estructura en disco

Karim Pichara B.

PUC Chile

Objetivos del diseño Lógico

- Simplicidad

- Usuarios deberían entender el diseño
 - Datos y su organización deberían estar de acuerdo con el modelo conceptual de usuarios
 - Modo consulta debería ser fácil e intuitivo

- Expresividad

- Incluir suficiente información para responder las consultas principales
 - Incluir datos relevantes, filtrar los irrelevantes

- Rendimiento

- Debe permitir un diseño físico posible y eficiente

Karim Pichara B.

PUC Chile

Objetivos del diseño Físico

Satisfacer los requerimientos que el diseño lógico impone en forma óptima



No malgastar recursos



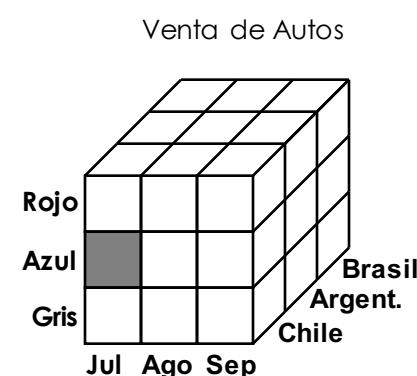
Permitir acceso eficiente a la información

Karim Pichara B.

PUC Chile

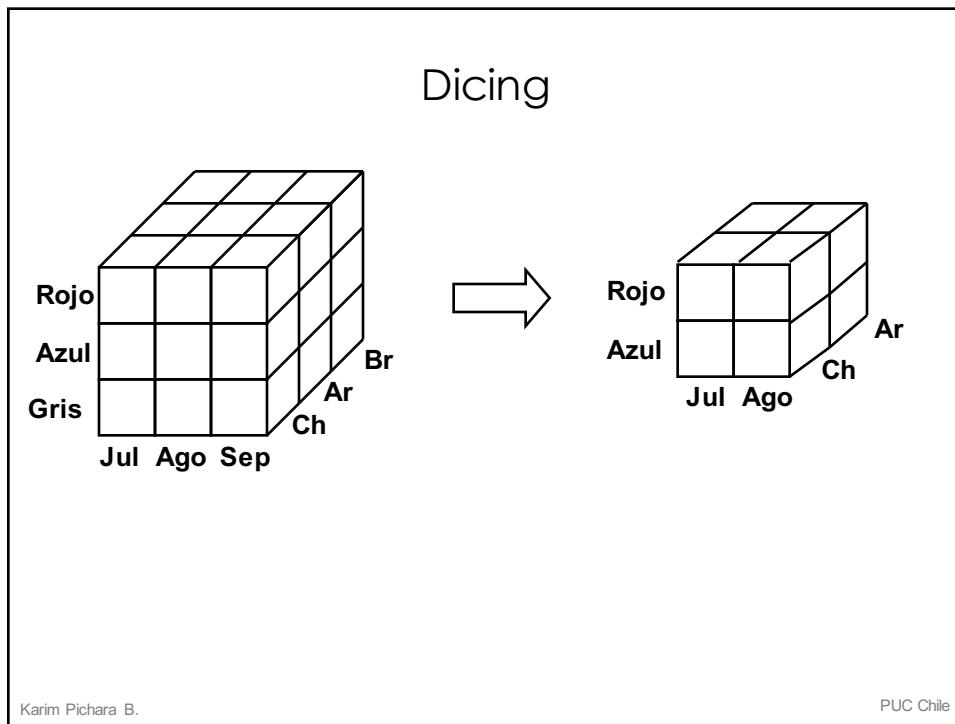
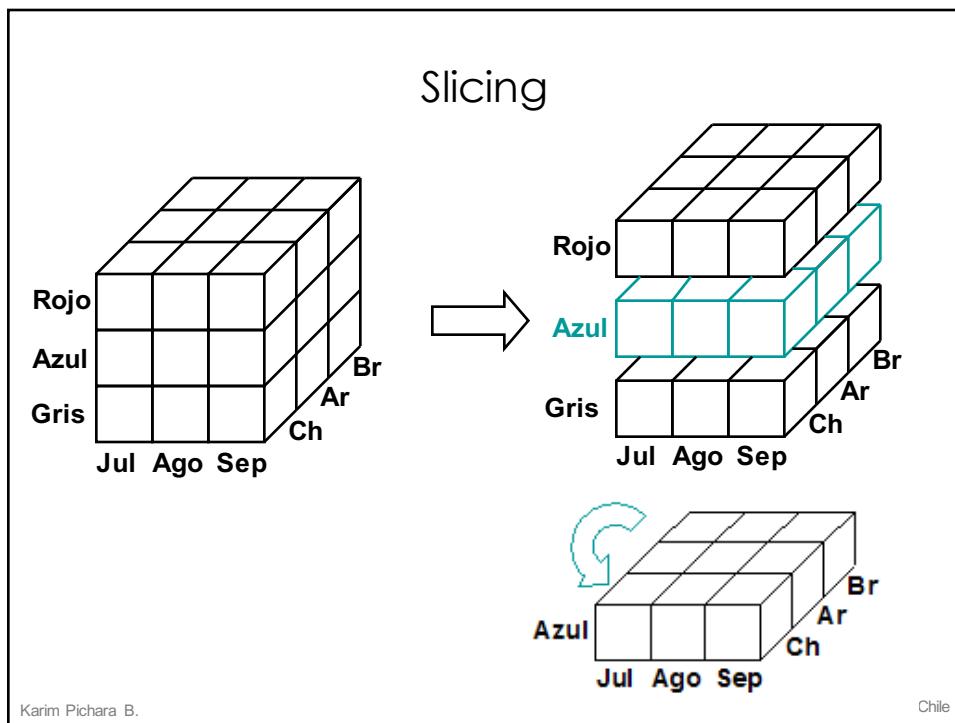
Data Cube

- Ejes representan atributos
 - Generalmente discretos
 - Ej. Color, mes, lugar, etc.
 - También llamados dimensiones
- Celdas guardan información agregada
 - Ej. Ventas totales de autos
- En la práctica datacubes tienen mucho más de 3 dimensiones



Karim Pichara B.

PUC Chile



Roll Up and Drill Down

Número de autos vendidos

	Ch	Ar	Br	Total
Jul	45	33	30	108
Ago	50	36	42	128
Sep	38	31	40	109
Total	133	100	112	345

Número de autos vendidos

Ch	Ar	Br	Total
133	100	112	345

Roll Up
por mes

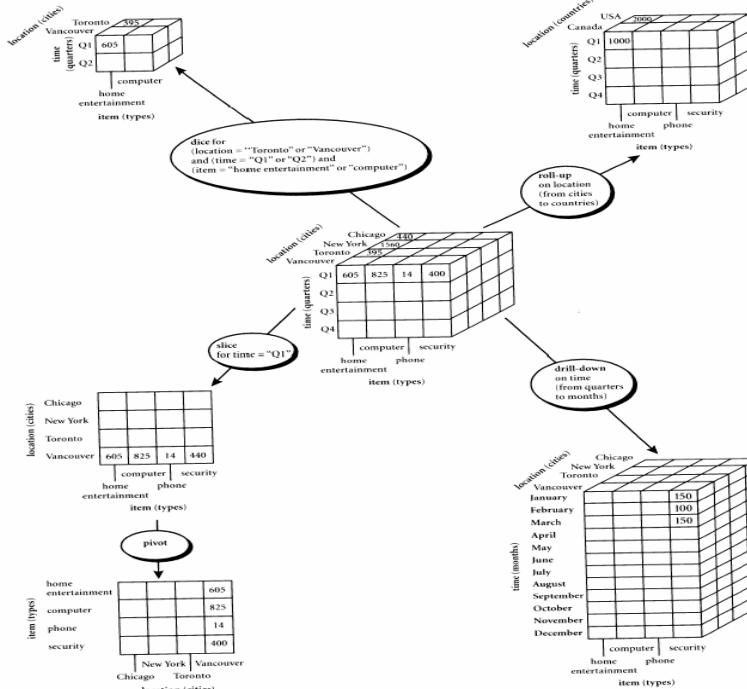
Drill down
por color

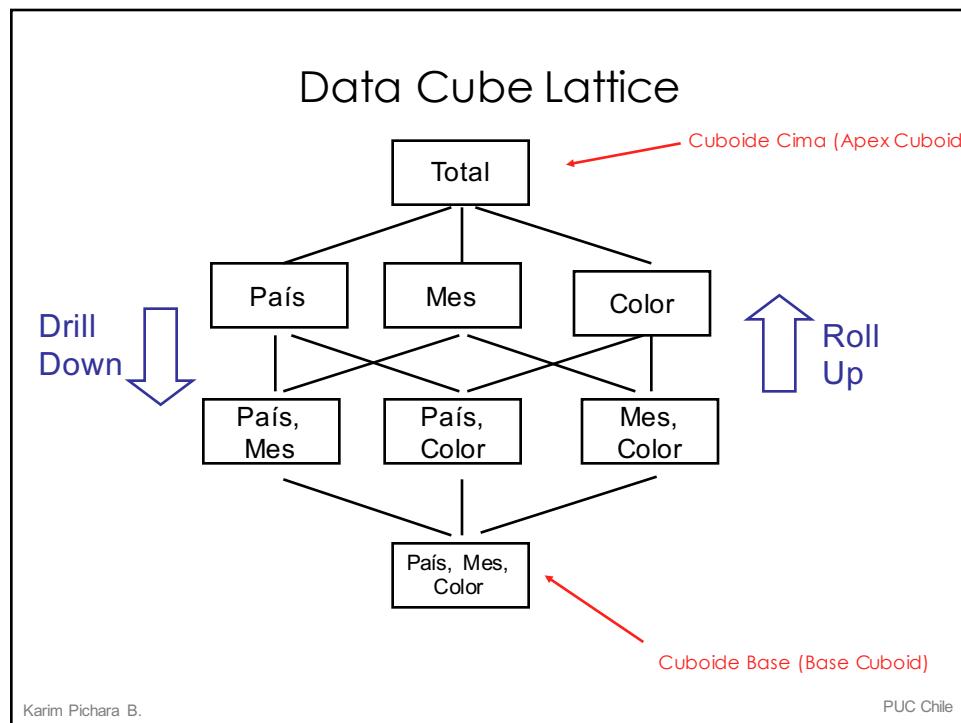
Número de autos vendidos

	Ch	Ar	Br	Total
Rojo	40	29	40	109
Azul	45	31	37	113
Gris	48	40	35	123
Total	133	100	112	345

Karim Pichara B.

PUC Chile





- Un Data Cube es un Lattice de Cuboides
- Celda Base: es una celda en el cuboide Base
- Celda Agregada: es una celda en cualquier cuboide distinto del base (agrega una o más dimensiones). Para las dimensiones agregadas usaremos la notación “*”.

- Ejemplo:

(Enero, *, *, 2800) Celdas 1D
 (*, Toronto, *, 1200) Celdas agregadas
 (Enero, *, Business, 150) Celda 2D
 (Enero, Toronto, Business, 45) Celda 3D

Karim Pichara B.

PUC Chile

- En un cubo n-dimensional:

 $a = (a_1, a_2, \dots, a_n, \text{value}_a)$ i_D cell

 $b = (b_1, b_2, \dots, b_n, \text{value}_b)$ j_D cell

“a” es ancestro de “b” $\Leftrightarrow (i \leq j) \text{ y } (a_m = b_m \quad \forall 1 \leq m \leq n, a_m \neq “*”)$

Ej: $a = (\text{Ene}, *, *, 2800)$

$b = (\text{Ene}, *, \text{Business}, 150)$

$c = (\text{Ene}, \text{Toronto}, \text{Business}, 45)$

b es parente de c

a y b son ancestros de c

Karim Pichara B.

PUC Chile

Tablas de Dimensión

- ¿Qué es una tabla de dimensión ?
 - Tabla que corresponde a un objeto o concepto del mundo real
 - Ejemplo: consumidor, producto, día, empleados, regiones, tiendas, promociones, vendedores, etc.

- Propiedades

- Contienen varias columnas descriptivas
 - En general son tablas anchas (docenas de columnas)
- Generalmente no tienen muchas filas
 - Al menos en comparación con las tablas de hechos
 - Usualmente < 1 millón de filas
 - Relativamente estáticas

Karim Pichara B.

PUC Chile

Tabla de Hechos

- ¿Qué es una tabla de hechos ?
 - Tabla que contiene mediciones acerca de un evento en un proceso de interés. Ej: venta, insumo, etc.
- Cada fila contiene 2 tipos de datos:
 - Columnas con valores numéricos o mediciones
 - Llaves a tablas de dimensiones
- Propiedades
 - Gigantes: A menudo millones o billones de filas
 - Angostas: A menudo pocas columnas
 - Cambian frecuentemente
 - Nuevos eventos en el mundo producen nuevas filas en la tabla
 - Típicamente las nuevas filas son sólo agregadas (no hay un ordenamiento especial)

Karim Pichara B.

PUC Chile

Uso de tablas

- De Dimensión

- La información se filtra en base a los atributos de cada dimensión
- Tablas de hechos son referenciadas a través de sus tablas de dimensión
- Agrupamientos son realizados a través de las columnas de atributos de cada dimensión

- De Hechos

- Guarda la info clave para el análisis

Karim Pichara B.

PUC Chile

Tablas de Hechos vs Tablas de Dimensión

Tabla de hechos

- Angostas
- Gigantes (muchas filas)
- Numéricas
- Crecen en el tiempo

Tabla de dimensión

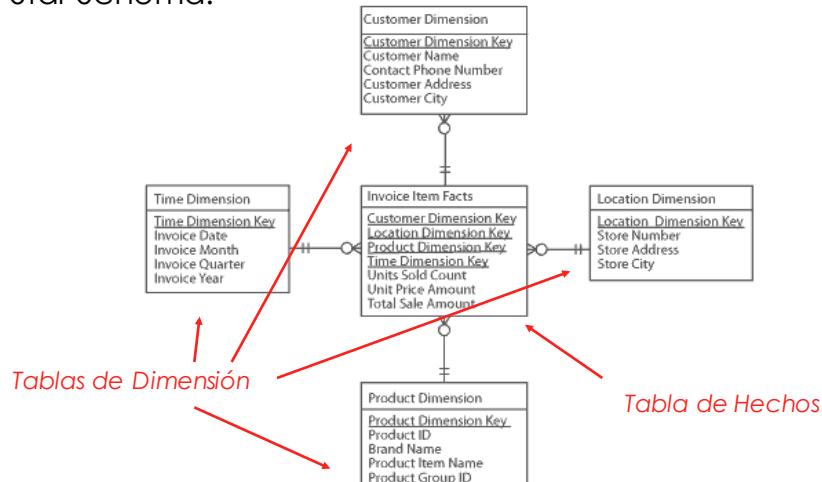
- Anchas
- Pequeñas (pocas filas y columnas)
- Descriptivas
- Relativamente estáticas

Karim Pichara B.

PUC Chile

Modelamiento de los Datos

Star Schema:

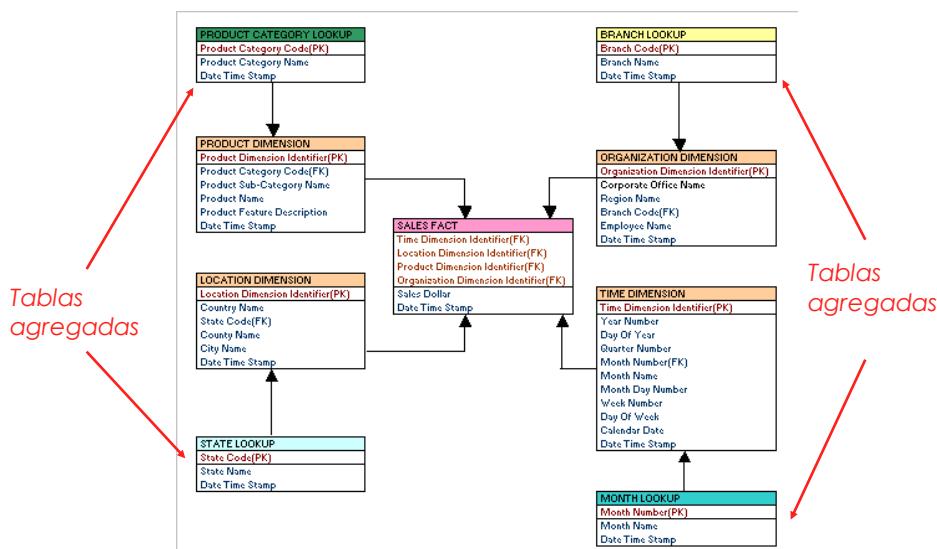


Karim Pichara B.

http://dmreview.com/editorial/dmdirect0302/030802_schraml_1.gif

PUC Chile

Snowflake Schema: Variante del Star Schema, algunas tablas de dimensiones se normalizan, agregándose tablas adicionales.

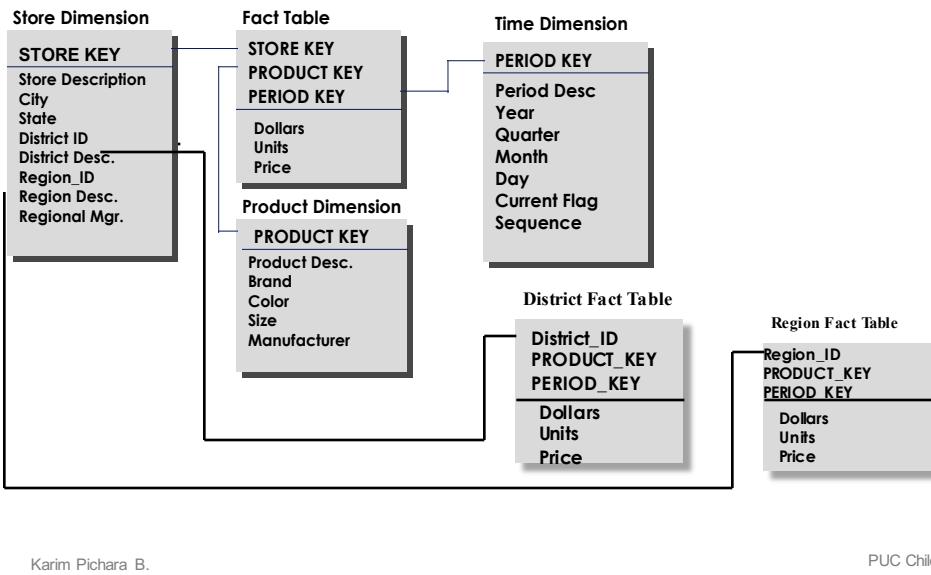


Karim Pichara B.

http://www.learndatamodeling.com/images/datamodels/snow_flake.gif

PUC Chile

Modelo de Constelación (Fact Constellation): Existe más de una tabla de hechos



Ejemplo:

	<i>location = "Chicago"</i>				<i>location = "New York"</i>				<i>location = "Toronto"</i>				<i>location = "Vancouver"</i>							
	<i>item</i>				<i>item</i>				<i>item</i>				<i>item</i>							
<i>time</i>	<i>home</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>
Q1	854	882	89	623		1087	968	38	872		818	746	43	591		605	825	14	400	
Q2	943	890	64	698		1130	1024	41	925		894	769	52	682		680	952	31	512	
Q3	1032	924	59	789		1034	1048	45	1002		940	795	58	728		812	1023	30	501	
Q4	1129	992	63	870		1142	1091	54	984		978	864	59	784		927	1038	38	580	

Karim Pichara B.

PUC Chile

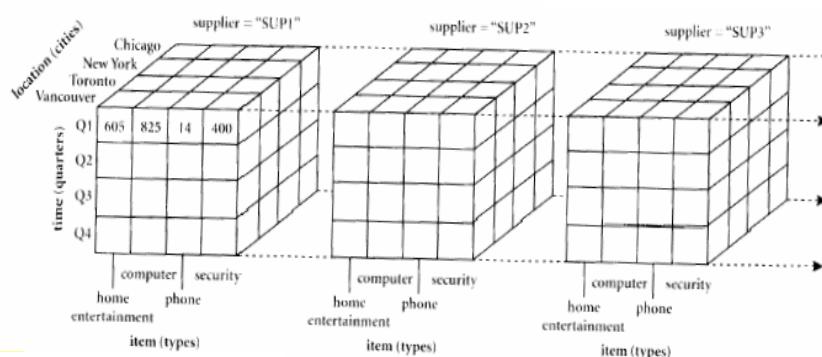
Data Cube

				Chicago	854	882	89	623	
				New York	1087	968	38	872	
				Toronto	818	746	43	591	
				Vancouver					
				Q1	605	825	14	400	
				Q2	680	952	31	512	
				Q3	812	1023	30	501	
				Q4	927	1038	38	580	
				computer					
				home					
				phone					
				entertainment					
item (types)									

Karim Pichara B.

PUC Chile

Representación 4-D del cubo de datos

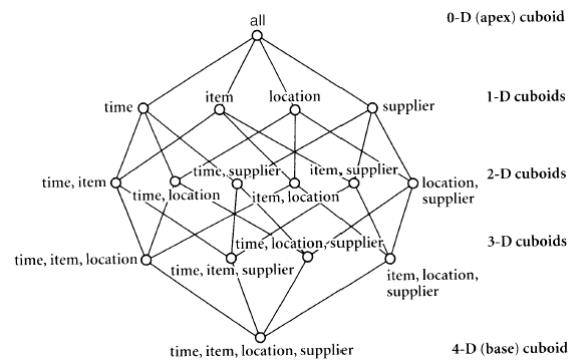


Se agregó la dimensión "supplier"

Karim Pichara B.

PUC Chile

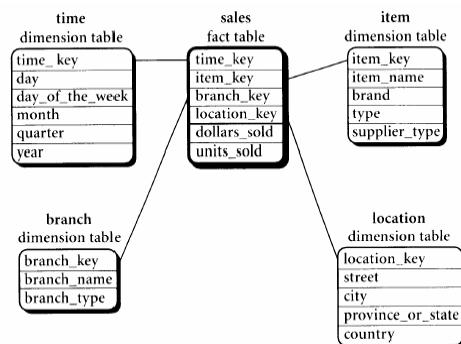
Data Cube Lattice



Karim Pichara B.

PUC Chile

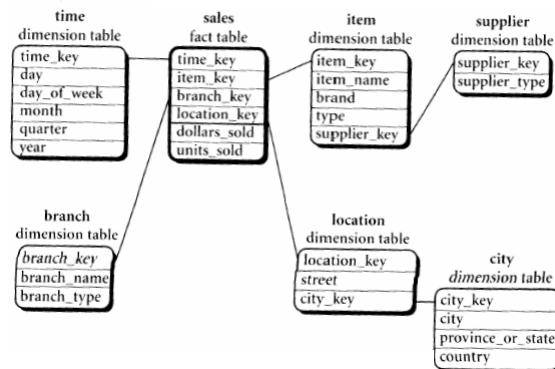
Star Schema



Karim Pichara B.

PUC Chile

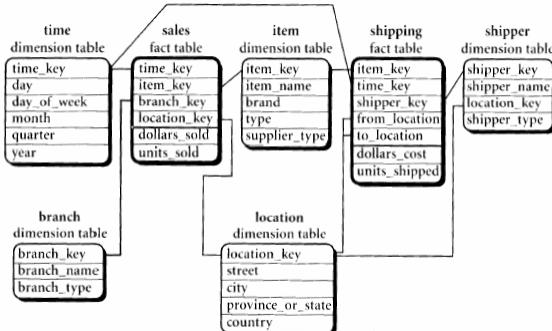
Snowflake Schema



Karim Pichara B.

PUC Chile

Fact Constellation Schema



Karim Pichara B.

PUC Chile

Pasos en la modelación de dimensiones

1. Identificar el proceso a ser modelado
2. Determinar la resolución con la cual los hechos serán almacenados (grain)
3. Elegir las dimensiones
4. Identificar los valores numéricos para los hechos

Karim Pichara B.

PUC Chile

Preguntas relevantes

- ¿Cuál es el impacto de una promoción?
– Requiere calcular ventas históricas del producto
- ¿Cuál es la acumulación de inventario del cliente?
– Requiere calcular compra histórica del cliente
- ¿Cuál es la canibalización?
– Requiere detectar ventas históricas de productos similares

Karim Pichara B.

PUC Chile

Preguntas relevantes

- ¿Cuál es la venta cruzada de productos?
 - Requiere detectar ventas de otros productos que sean complementarios
 - Pañales y cerveza (fomentar compra compulsiva)
- ¿Cuál es la ganancia neta de la promoción?
 - Considera costos de la promoción, descuentos, inventarios de cliente, canibalización y ventas cruzadas

Karim Pichara B.

PUC Chile

1. Identificar el proceso a ser modelado

- Ej: Datos en Supermercado
 - Datos adquiridos por cajeras mediante códigos de barras
 - 100 tiendas en 5 ciudades
 - ~60.000 productos
 - Algunos tienen UPCs (Universal Product Codes)
 - Otros no (por ejemplo, pan, carne, flores)
- Objetivo: entender el impacto del precio y promociones en las ganancias
 - Promociones = cupones, descuentos, anuncios
 - impacto del precio -> Ventas, Precios
 - impacto en ganancia -> ingresos

Karim Pichara B.

PUC Chile

2. Resolución de la tabla de hechos

- Objetivo: determinar el máximo nivel de detalle del DW
- Ejemplo:
 - Una fila de la tabla de hechos puede representar:
 - Un ítem de una de las cajas de un supermercado ó todos los ítems de ese tipo vendidos por esa caja en el día
 - Un ticket para abordar un avión o el total de tickets vendidos por vuelo
 - Resumen diario del inventario de un producto o venta semanal
 - Un estudiante en un curso o un estudiante en un área

Karim Pichara B.

PUC Chile

2. Resolución de la tabla de hechos

- Mayor resolución implica:
 - Mayor expresividad
 - Mayor número de filas
- Trade-off entre rendimiento y expresividad
 - Recomendación: ante la duda preferir expresividad
 - Información agregada pre-calculada puede resolver problemas de rendimiento

Karim Pichara B.

PUC Chile

3. Elegir dimensiones

- Determinar candidatos dependiendo del significado de las filas de la tabla de hechos

-Ejemplo:

- filas de tabla de hechos representan alumnos en un curso
- Dimensiones posibles pueden ser curso, estudiante, semestre, etc.

Karim Pichara B.

PUC Chile

4. Identificar valores numéricos para hechos

- Útil para el análisis de datos
- Identificar rangos y unidades
- Datos continuos o discretos
- Elegir unidades de la forma adecuada
- Unificar criterios para todo el sistema

Karim Pichara B.

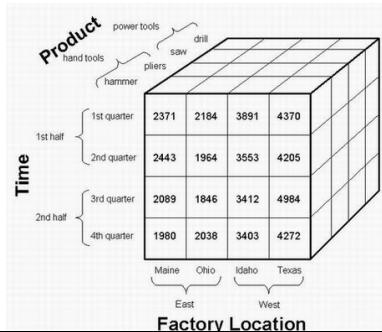
PUC Chile

Almacenamiento de la información

- **MOLAP** (Multidimensional On Line Analytical Processing)

– Almacena el cubo multidimensional de datos como un arreglo multidimensional

– Problemas con la densidad de los datos y redundancia



Karim Pichara B.

PUC Chile

Densidad de los Datos

- Imagine un DW de una cadena de tiendas
- Dimensiones: consumidores, productos, tiendas y días
- Suponga que hay 100.000 clientes, 10.000 productos, 1.000 tiendas y un período de 1.000 días
- Cubo de datos tiene 1,000,000,000,000,000 de celdas

Karim Pichara B.

PUC Chile

Densidad de los Datos

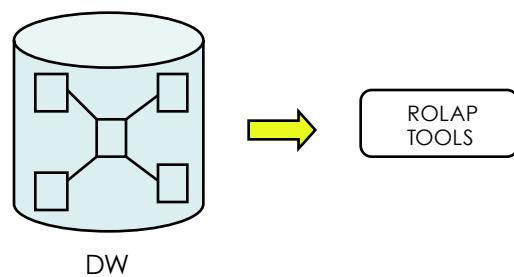
- Afortunadamente la mayoría están vacías
- Una tienda determinada no vende cada producto cada día
- Un consumidor no visita las 1.000 tiendas diariamente, quizás nunca visita más de 2 o 3 de las tiendas
- Un consumidor no compra todos los productos
- ¿Será esto un inconveniente para el uso de arreglos multidimensionales ?

Karim Pichara B.

PUC Chile

Almacenamiento de la información

- **ROLAP** (Relational On Line Analytical Processing)
 - Almacena el cubo multidimensional de datos en una base de datos relacional (ej. Star Schema)



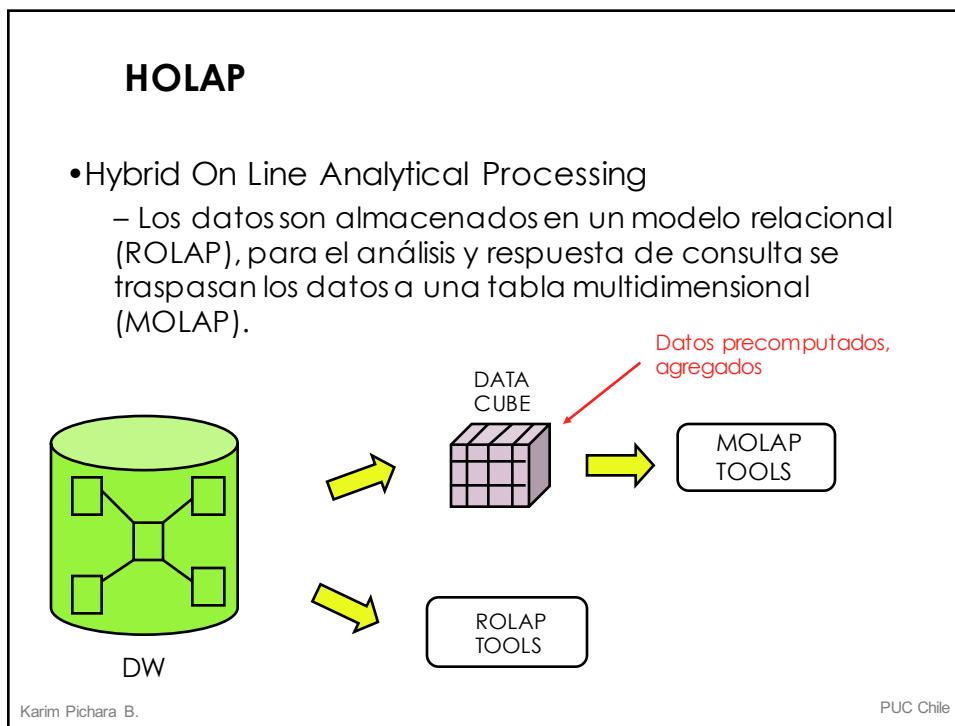
Karim Pichara B.

PUC Chile

•MOLAP	•ROLAP
<ul style="list-style-type: none"> -Usualmente precalcula valores agregados -Acceso eficiente a los datos, respuestas rápidas -No escala bien a muchas dimensiones 	<ul style="list-style-type: none"> -Mejor expresividad para consultas -Escala bien con la dimensionalidad -Escala bien a muchos datos -Densidad de Datos no es un problema -Tecnología madura (bases operacionales) -Respuesta a consultas no tan buena como MOLAP -Necesita construir los índices relacionales

Karim Pichara B.

PUC Chile



Número de Dimensiones

- ¿Modelar dos conceptos como dimensiones separadas o dos aspectos de la misma dimensión?
- Ejemplo: diferentes tipos de promociones
 - Anuncio, descuento, cupones, mejor posición en estantes
 - Opción A: 4 dimensiones
 - Separar cada promoción en una dimensión
 - Opción B: 1 dimensión
 - Cada fila captura una combinación de las 4 dimensiones

Karim Pichara B.

PUC Chile

Después de crear el Data Warehouse

Variación temporal de datos

- Tablas de hechos muy dinámicas pero dimensiones varían lentamente
 - Nuevas ventas a cada minuto
 - No hay nuevos productos cada día
 - No se abren nuevas sucursales a menudo
- ¿Qué significa un cambio para una dimensión?
 - Clientes se cambian de dirección
 - Agrupamiento de tiendas cambia con crecimiento
- ¿Cómo tomamos en cuenta estos cambios en nuestra DW?
 - Opción 1: actualizar la información
 - Opción 2: preservar la historia

Karim Pichara B.

PUC Chile

Actualizar la Información

- Ejemplo:

– El tamaño del producto es incorrecto no es 1 sino 3 mts.

- El error se actualiza en sistema OLTP
- ¿Qué hacemos en el DW?
 - Arreglemos el dato en la tabla de dimensión
 - Problema: pueden haber datos precomputados o agregados que contengan la info con error

– ¿Qué pasaría en este caso ?

- Juan Pérez vivía en Iquique en 2000
- Juan Pérez se cambió a Santiago en 2003
- ¿Qué hacemos?
 - Ok, actualicemos
 - Nueva consulta : ¿Cuales fueron las ventas en Iquique el 2000 ?

Karim Pichara B.

PUC Chile

Preservar Historia

• Histórial sin errores puede ser importante para un DW

• ¿Cómo podemos capturar cambios y preservar la historia ?

Crear una nueva entrada en tabla de dimensión

Dimensión cliente

	Key	Nombre	Sexo	Ciudad	Año
	457	Juan P.	M	Iquique	2000

	784	Juan P.	M	Stgo	2003

Nueva entrada

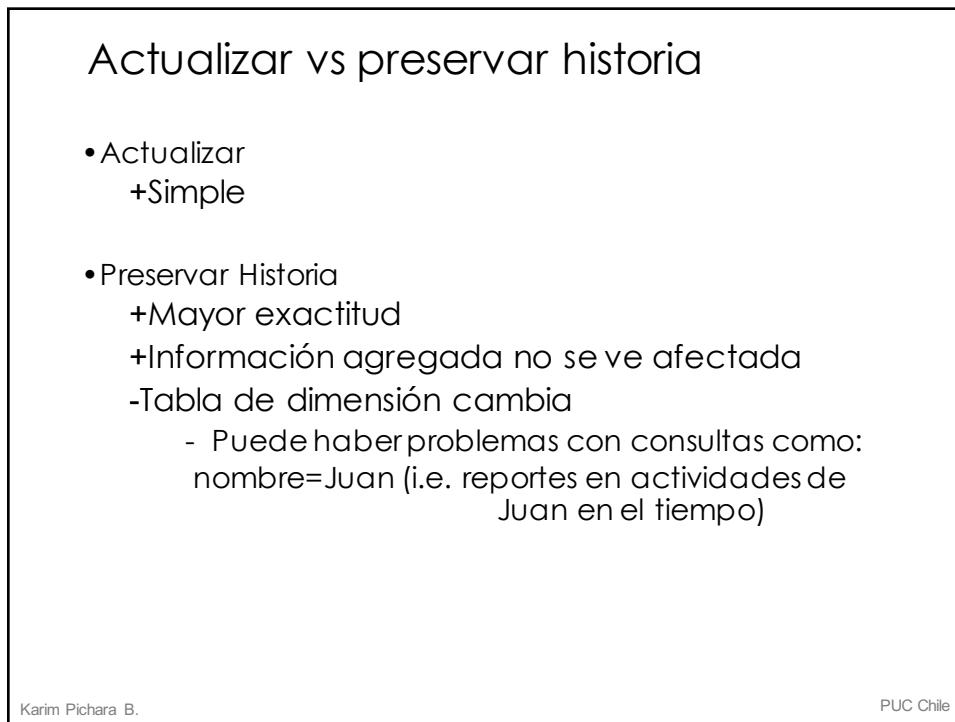
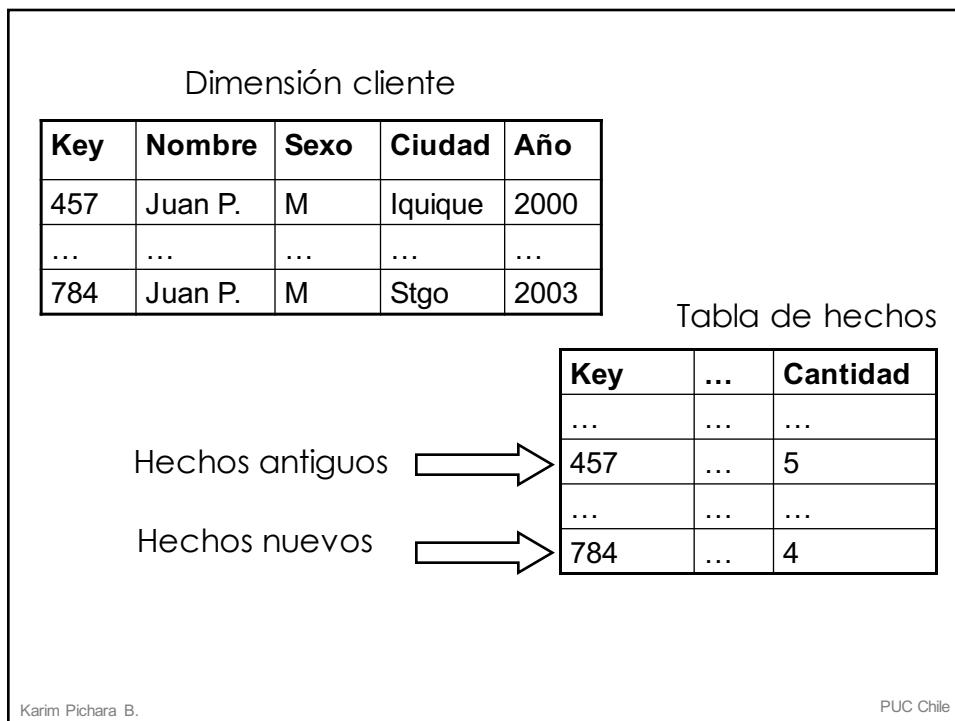


– Hecho antiguo apunta a 457

– Nuevos hechos apuntan a 784

Karim Pichara B.

PUC Chile



Operación y Mantención

- Soporte de aplicaciones
 - Conocer todas las aplicaciones disponibles
 - Conocer sistemas de seguridad, controles, menús, relaciones entre aplicaciones, etc.
- Soporte de herramientas de análisis
 - Debe ayudar a encontrar información deseada
 - Debe entender lo que los usuarios desean desde la perspectiva del negocio
- Soporte de entrenamiento
 - Enseñar a los usuarios de la DW como utilizarla, sus herramientas, sus datos

Karim Pichara B.

PUC Chile

Operación y Mantención (cont..)

- Soporte de atención de usuarios (help desk)
 - Manuales, mensajes, etc.
- Soporte de operación
 - Verificar el correcto funcionamiento de la DW
- Soporte de mantenimiento y actualización de DW
 - Preocuparse de mantener la base de datos actualizada (extracción, transformación, almacenaje)
- Soporte de evolución de la DW
 - Estudiar comportamientos de uso (apoyo a áreas de baja utilización del DW)

Karim Pichara B.

PUC Chile

¿Cómo justificar un DW?

- Razones más comunes:

- Ahorrar dinero siendo más eficientes
- Agilizar proceso de extracción de información
- Ser más competitivo
- Mejorar productividad
- Mejorar toma de decisiones

- Razones específicas:

- Reducir costos de acceso masivo a la información
- Mejorar relación con clientes
- Identificar oportunidades de negocio ocultas
- Ejecutar un marketing más efectivo

Karim Pichara B.

PUC Chile

Crecimiento de Wal Mart (500%)



Karim Pichara B.

PUC Chile

Valor de la acción de 3M



Karim Pichara B.

PUC Chile

Ganancias de Fed-Ex



Karim Pichara B.

PUC Chile

Métricas de Usabilidad

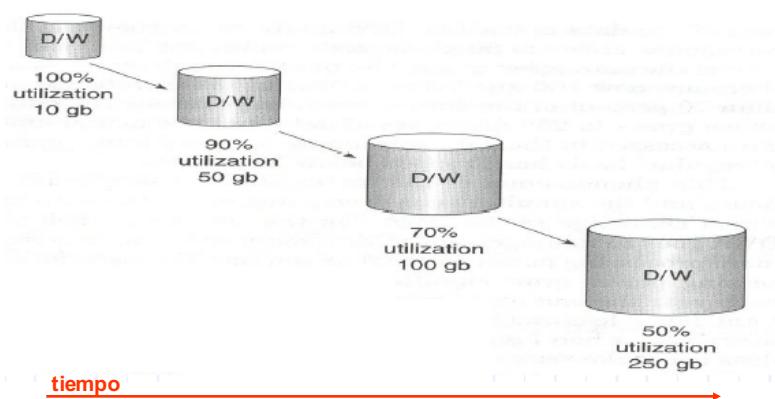
- Ejemplos:

- Número de usuarios activos de la DW
- Frecuencias de uso
- Tiempo de las sesiones
- Número y tipos de preguntas de los usuarios

Karim Pichara B.

PUC Chile

Usabilidad decrece con el tiempo



Con el tiempo el porcentaje de uso decrece por lo cual es necesario saber cuales son los datos que se utilizan para evitar almacenar información irrelevante

Karim Pichara B.

PUC Chile

Preguntas Importantes

- ¿Qué datos se usan?
- ¿Quién está usando el DW ?
- ¿Quién no está usando el DW?
- ¿Cuál es el tiempo de respuesta?