

Minería de Datos

IIC2433

Data Warehouse
Vicente Domínguez

Data everywhere: Problem



Data everywhere: Problem



Data everywhere: Problem



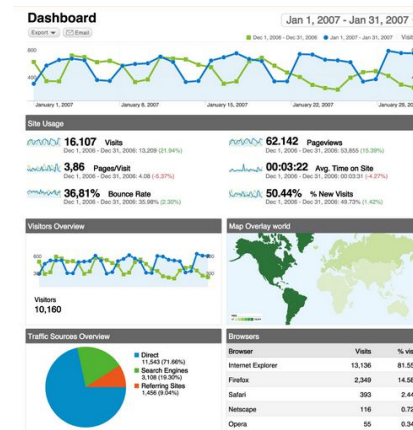
- I can't **get** the data I need
 - need an expert to get the data
- I can't **find** the data I need
 - data is scattered over the network
 - many versions, subtle differences
- I can't **understand** the data I found
 - available data poorly documented
- I can't **use** the data I found
 - results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

Algunos Tipos de Bases de Datos Actuales

Operacionales



Analíticas



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...

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 Fare Class: United Economy (V) Meals: Meals for Purchase No Special Meal Offered.	Flight Time: 6 hr 40 mn

Traveler Information:

Mr. KARIME PICHARA

Seat Assignments: BOS - SFO: 24D

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)

Bases de Datos Analíticas

 **Página principal**

Informes estándar

Informes personalizados

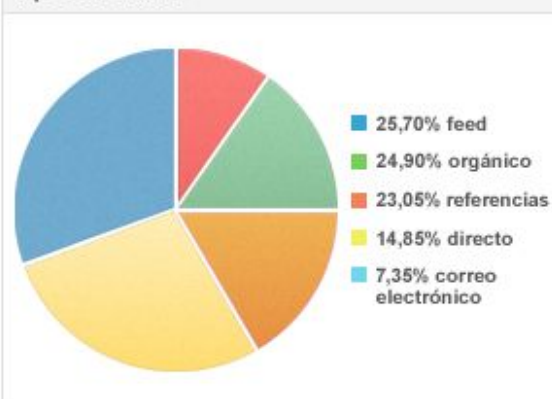


Mi panel

Visitas diarias



Tipos de tráfico



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

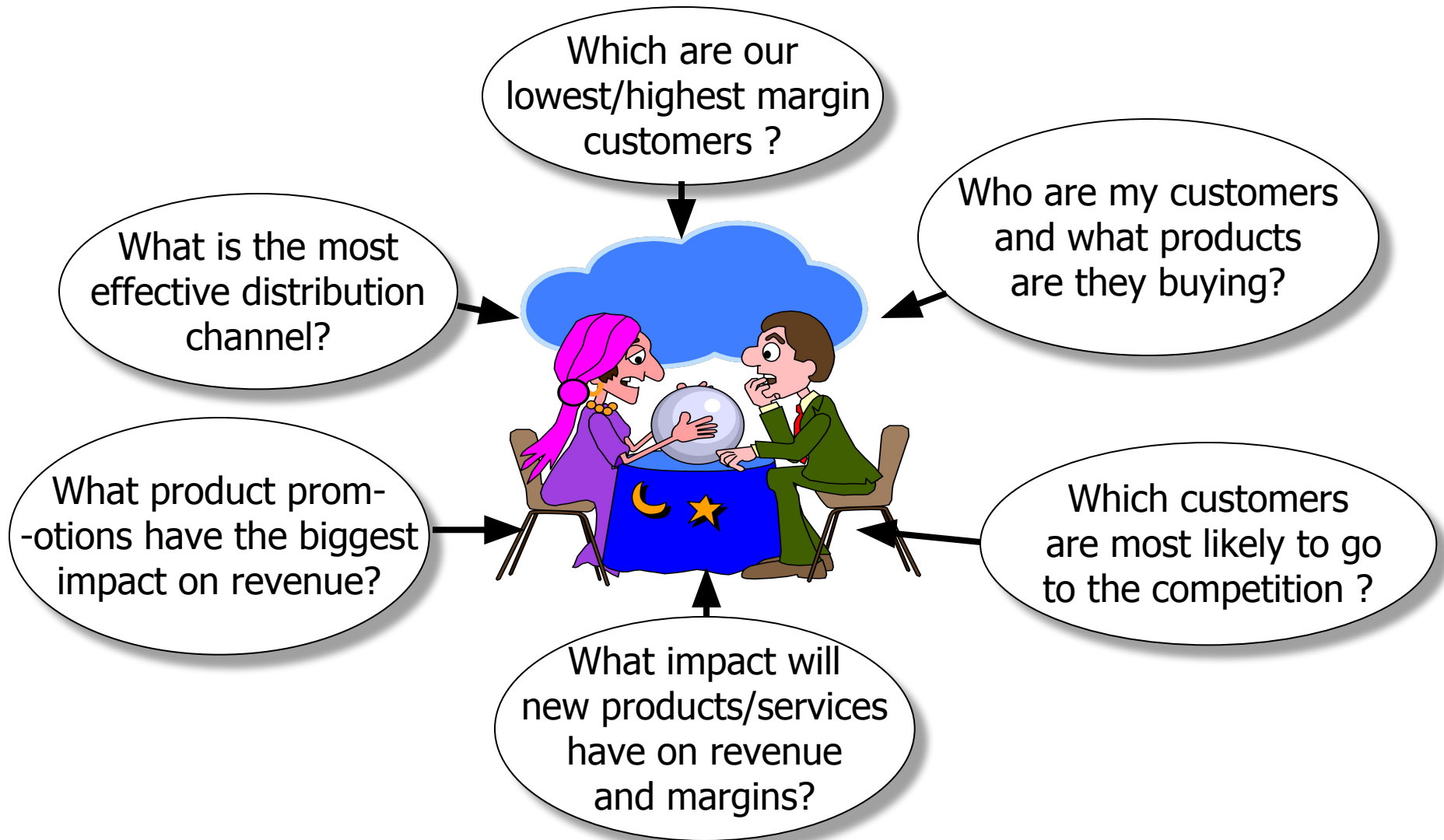
Data Warehousing

- ¿Qué es un Data Warehouse?

“Un **Datawarehouse** es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.”



Why Data Warehouses?



Data Warehousing

- **Qué es un Data Warehouse?**

- Base de datos orientada a la toma de decisiones
- Guarda información generalizada y consolidada (de un largo horizonte de tiempo)
- La información la integra desde diferentes bases de datos
- Provee una arquitectura y herramientas para organizar, entender y utilizar la información de tal forma de tomar mejores decisiones estratégicas.
- Está optimizada para responder preguntas complejas de ejecutivos de la compañía

- 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 inicial de datos y acceso a la información*.

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

Ejemplo de una vista para un DW

SAP Enterprise Portal 5.0 - Microsoft Internet Explorer proporcionado por LAN

Bienvenido

Buscar Personalizar Página | Portal Agregar a Favoritos

Home	SAP Business Warehouse	Mi Intranet	Presupuesto	Directorio Personas	Autoservicio Personas	Aplicaciones	Políticas y Procedimientos	Correo	Documentación SAP	Navegación en Portal	Procedimientos Comerciales	Si de Int
------	-------------------------------	-------------	-------------	---------------------	-----------------------	--------------	----------------------------	--------	-------------------	----------------------	----------------------------	-----------

SAP Business Warehouse

Bienvenido

SAP

- ▼ Sistemas de Información
 - ▼ Controlling
 - Costo Fijo Responsable Jerárquico
 - Panel Proyecto ECO (Costo Fijo)
 - Dotaciones por Responsable Jerárquico
 - ▶ Control de Gestión y Planificación Corporativa
 - ▼ Gerencia General Pax
 - ▶ Gerencia Comercial
 - ▶ Gerencia Lan Express
 - ▼ Gerencia Experiencia de Viaje
 - Panel Experiencia de Viaje
 - ▶ Control de Gestión y Planificación
 - ▶ Gerencia de Servicios Pax
- ▶ Gerencia General Carga
- ▶ Control de Gestión Soporte
- ▶ Control de Gestión y Planificación VPT
- ▶ Contraloría Corporativa
- ▶ Sistemas de Información
- ▶ Gerencia de Negocios Aeroportuarios
- ▶ Otros Reportes Corporativos

Ejemplo de una vista para un DW

Dotaciones Responsable Jerarquico

* La Estructura de Responsable Jerarquico 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).

Bloque navegación:

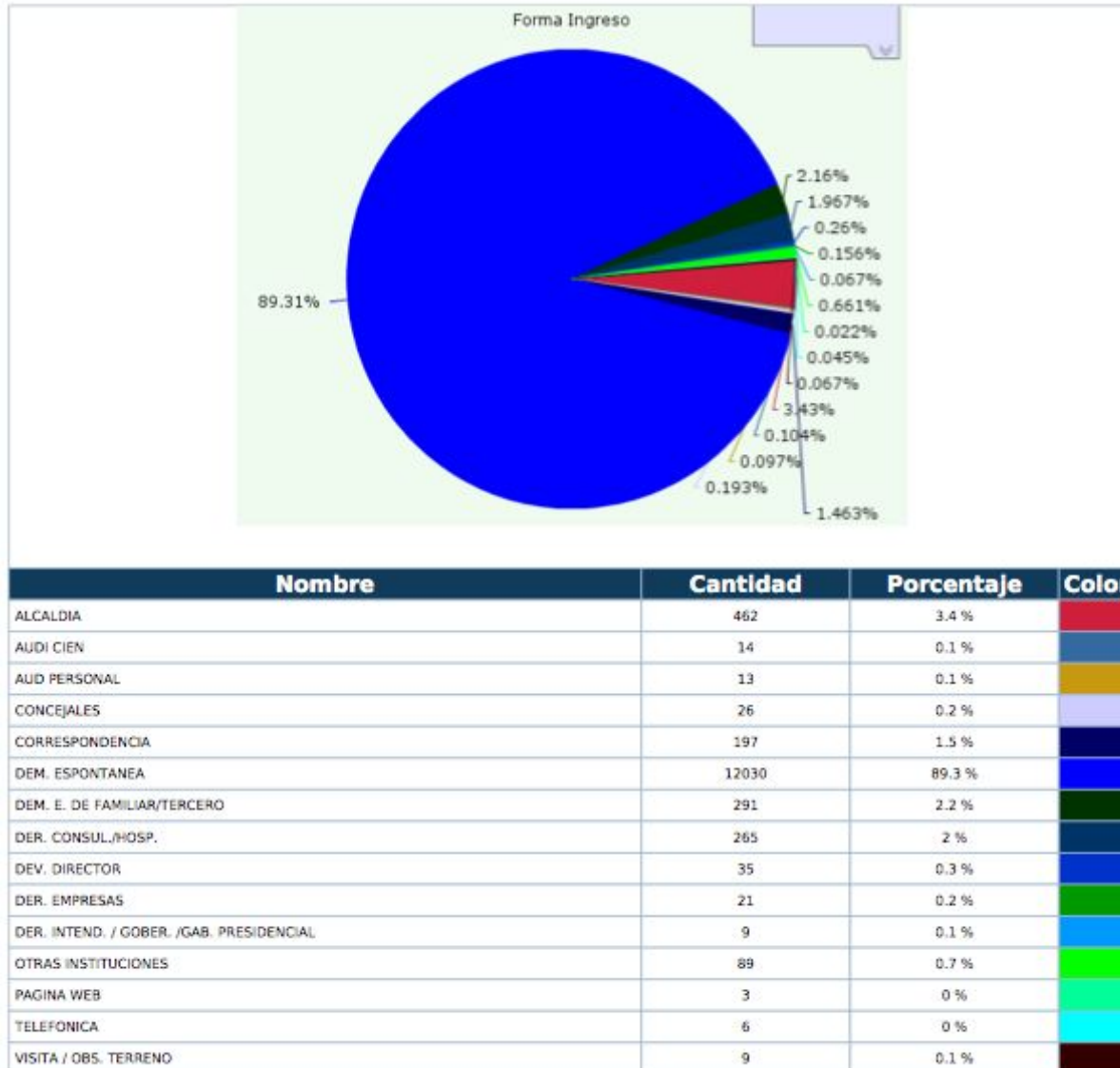
Centro de coste		Función		Mes natural		Nodo Responsable	
Pais		Sociedad		Tipo Contrato		Estructura	

Dotaciones Responsable Jerarquico

Mes natural	JUL					JUN					MAY					ABR
Nodo Responsable	R 2005	PPTO	R 2006	PPTO - 2006	% Var	R 2005	PPTO	R 2006	PPTO - 2006	% Var	R 2005	PPTO	R 2006	PPTO - 2006	% Var	R 200
▼ ENRIQUE CUETO	15.131	15.758	15.809	15.758	100	15.025	15.721	15.250	15.721	100	15.133	15.890	15.835	15.890	100	15.133
▶ ENRIQUE CUETO	5	5	5	5	100	5	5	5	5	100	5	5	5	5	100	5
▼ IGNACIO CUETO	10.594	11.413	11.515	11.413	100	10.605	11.340	10.979	11.340	100	10.287	11.415	11.476	11.415	100	10.287
▶ IGNACIO CUETO	5.233	5.515	5.255	5.515	100	5.255	5.512	5.718	5.512	100	5.692	5.914	5.731	5.914	100	5.692
▶ ARMANDO VALDIVIESO	245	251	234	251	100	244	254	227	254	100	239	254	242	254	100	239
▶ FRANCISCO GUIMPERT	222	2753	2.764	2.764	100	2.715	2.763	2.763	2.763	100	2.561	2.754	2.713	2.754	100	2.561
▶ CRISTIAN URETA	1.437	1.563	1.505	1.563	100	1.431	1.555	1.555	1.555	100	1.403	1.425	1.415	1.425	100	1.403
▶ DAMIAN SCOKIN	521	1.512	1.514	1.514	100	515	1.512	1.512	1.512	100	502	1.512	1.513	1.512	100	502
▶ BRUNO ARDITO	1.5	1.5	1.5	1.5	100	1.5	1.5	1.5	1.5	100	1.5	1.5	1.5	1.5	100	1.5
▼ MARCO JOFRE	1.25	2.140	2.179	2.140	100	1.25	2.140	2.140	2.140	100	1.25	2.140	2.140	2.140	100	1.25
▶ MARCO JOFRE	3	3	3	3	100	3	3	3	3	100	3	3	3	3	100	3
▶ FRANCISCO SOTOMAYOR	621	807	842	807	100	623	807	827	807	100	623	807	819	807	100	623
▶ JORGE IHENEN	3.45	928	928	928	100	3.45	928	928	928	100	3.45	928	928	928	100	3.45
▶ CRISTIAN LEON	310	241	244	244	100	300	235	235	235	100	291	238	238	238	100	291
▶ ANGELA CORRALES	33	23	15	15	100	21	21	15	15	100	35	21	15	15	100	35
▶ GERENCIA REPRESENTACION TECNICA	1	18	10	10	100	15	16	10	10	100	15	15	10	10	100	15
▶ MAGDALENA SPATE	1	12	15	15	100	12	13	15	15	100	12	13	15	15	100	12
▶ OSCAR GONZALEZ	1	1	1	1	100	1	1	1	1	100	1	1	1	1	100	1

Números

Ejemplo de una vista para un DW



Ejemplo de una vista para un DW





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	Cump.POA		Acum.*10	Acum.*09	% Crec.	POA	Cump.POA
Con. (MM \$)	724	663	9%	608		Estac. de Servicio					
						Distribuidores					
						Industrial					
						Compet y Exportación					

Información Detallada - Oficina - Contribución (MM \$)

Con 2010	Con 2009	% Crec.	POA Con	Cump.POA		Acum. Con 20	Acum. Con 20	Acum. % Cre	POA Con	Cump.POA
					Dis. Norte-Sur					
					Dis. Santiago-Centro					
					TCT					
1,147	45	938%	382	100%	Total	1,147	45	938%	382	100%

Canal: Distribuidores

Margen (\$...					Compet y Exportación					
					Total País	111	111	100%	111	100%

☐ Agua

☒ Detalle de Reventa

☐ Cuidado Automotriz

☐ Detalle E/S



Vol Mix

Ambas Marcas (P)

Ambas Marcas (A)

MOBIL (P)

MOBIL (A)

ESSO (P)

ESSO (A)

Informe de Enero a Enero 2010

Z- 130 (Jose Avila) - Volúmen Mix

	COMPRAS			VENTAS			DELTA STOCK	
Clasificación	Per.2010	Per.2009	Crec. %	Per.2010	Per.2009	Crec. %	Per.2010	Per.2009
Diesel Competitivo	31.292	21.213	47%	41.934	23.004	82%	10.642	1.791
Diesel No Competitivo	22.874	9.787	133%	26.940	15.478	74%	4.066	5.691

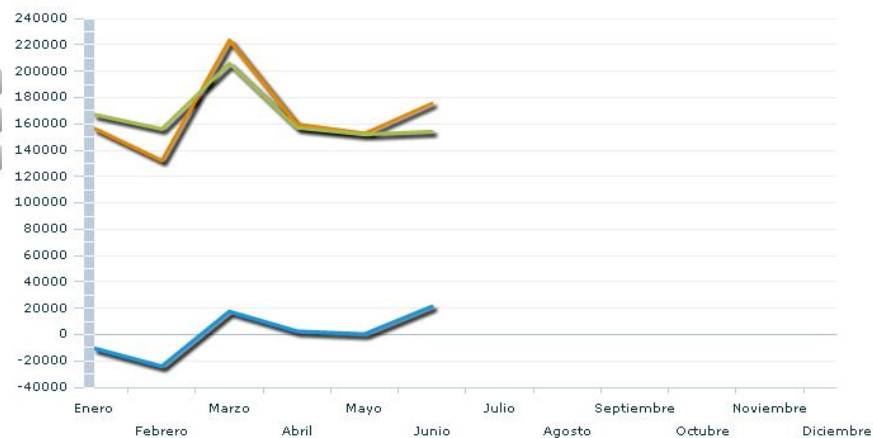
☒ Compra 2010
☒ Venta 2010
☒ Stock 2010

Mobil

Esso

Ambas

Gráfico Compra Vs Venta

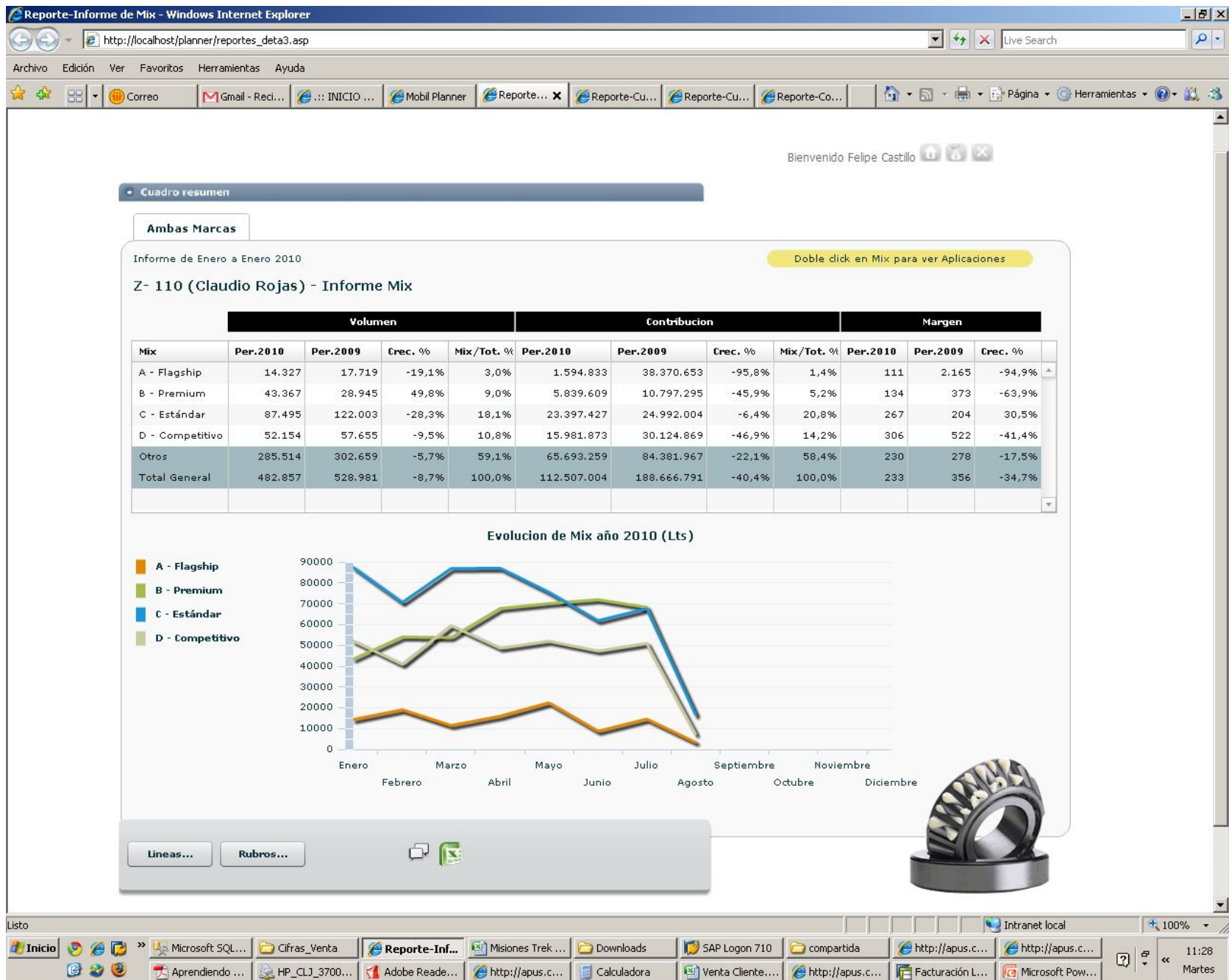


	Clasificación	Per.2010	Per.2009	Crec. %	Per.2010	Per.2009	Crec. %	Per.2010	Per.2009
Ambas Marcas	Diesel Competitivo	46.678	30.958	50%	56.091	35.296	58%	9.413	4.338
	Diesel No Competitivo	31.659	13.686	131%	32.716	20.515	59%	1.057	6.829
	Gasolina Competitivo	27.878	33.050	-15%	32.240	34.464	-6%	4.362	1.414
	Gasolina Estándar	19.292	17.527	10%	18.367	16.582	10%	-925	-945
	Gasolina Premium	23.952	15.476	54%	15.539	15.179	2%	-8.413	-297
	Gasolina Sintético	836	396	111%	692	420	64%	-144	24
	Industrial Competitivo	32.898	22.869	43%	34.158	36.471	-6%	1.260	13.602
	Industrial No Competitivo	46.003	29.611	55%	45.257	37.729	19%	-746	8.118
	TOTAL AMBAS	229.196	163.573	40%	235.060	196.656	19%	5.864	33.083

☐ Agua☐ Cuidado Automotriz

TODOS





Analítica web para empresas

Disfrute de ella en la plataforma líder de Google. [Más información](#)

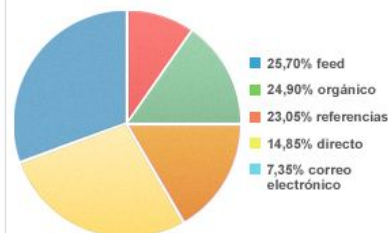
[Página principal](#)[Informes estándar](#)[Informes personalizados](#)

Mi panel

Visitas diarias



Tipos de tráfico



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

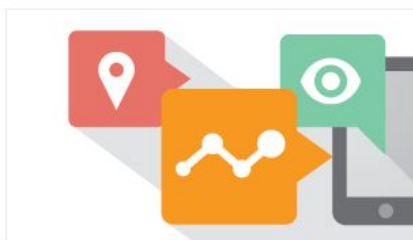
Herramientas de medición para su empresa



Estadísticas de varios canales

Consulte la ruta completa de la conversión con los embudos multicanal

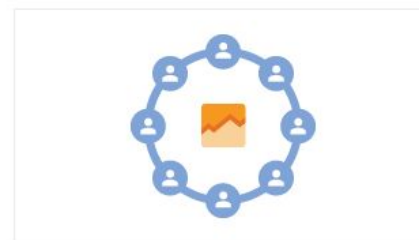
[Más información](#)



Soluciones para móviles

Mida los resultados en smartphones, sitios web para móviles y aplicaciones para móviles

[Más información](#)



Informes sociales

Mida el impacto de las redes sociales en los objetivos de su negocio y las conversiones

[Más información](#)

¿ Y qué es entonces Data Warehousing?

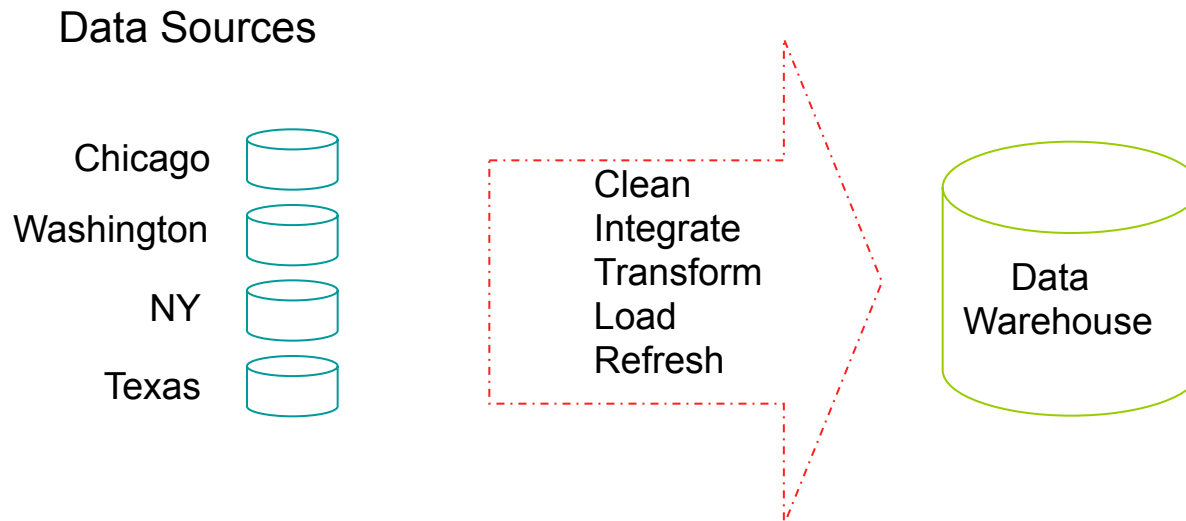
Data Warehousing



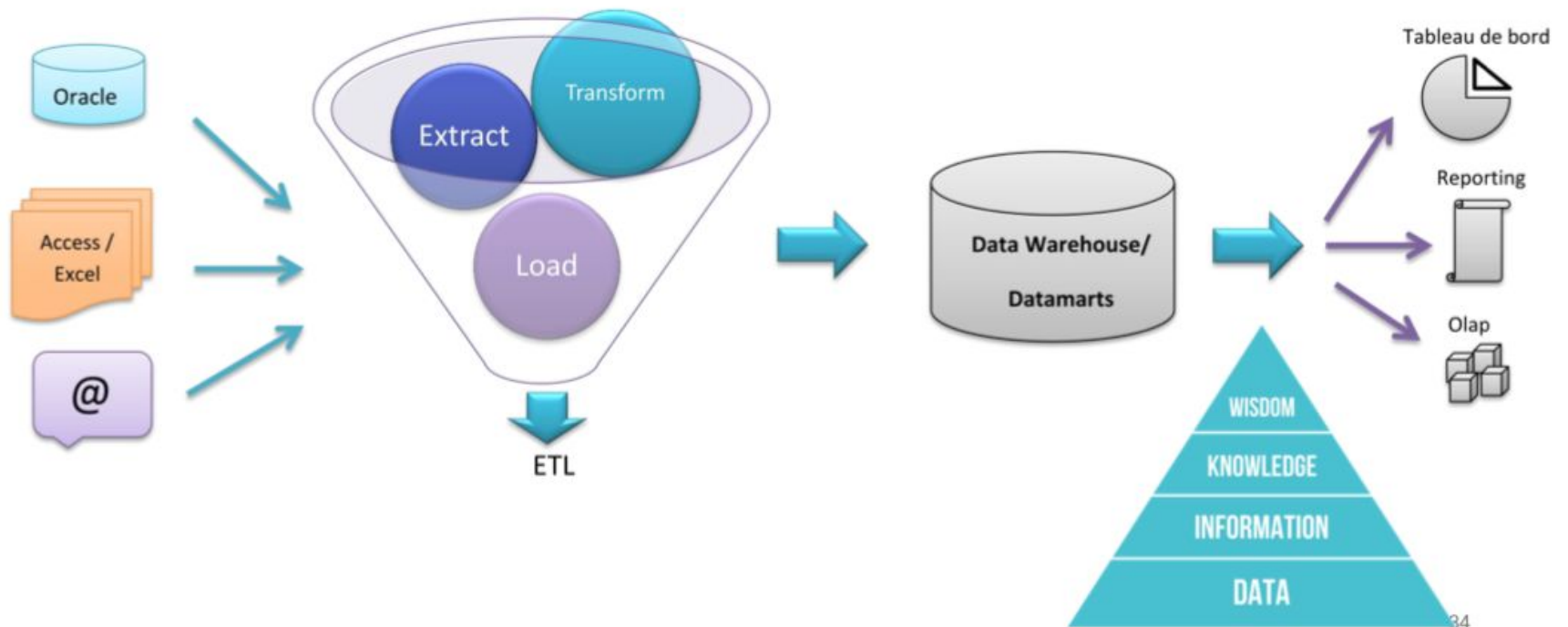
Proceso de construir un Data Warehouse

Este proceso requiere:

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



Integración de datos



Bases de Datos Analíticas

- 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)

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
- Concurrencia 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 \$1 US 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

Requerimientos
Diferentes



Necesidad de Separar
OLAP y OLTP

OLTP

Respuesta rápida es fundamental (< 1 seg.)

Datos deben estar siempre actualizados (Consistencia, Concurrencia)

OLAP

Consultas consumen muchos recursos

Consultas pueden saturar CPU y/o acceso a disco

No necesitan actualización constante o rápidos tiempos de respuesta

OLAP degradaría el funcionamiento de OLTP

Transacciones lentas → usuarios irritados

Ejemplo:

Consulta a sistema OLAP
por evaluación de ventas
totales del último mes



Tabla de ventas bloqueada
para permitir consistencia



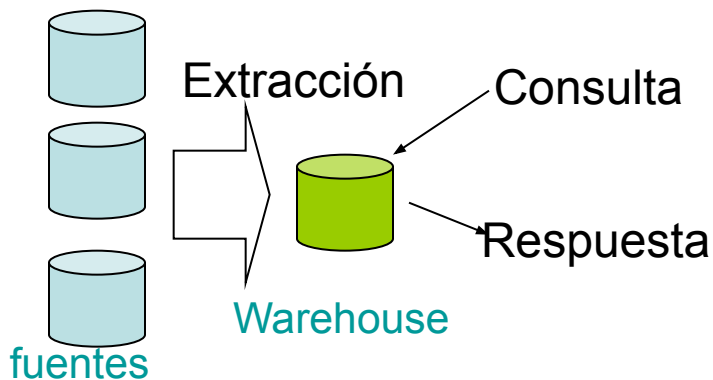
Aumenta el tiempo de
espera de clientes



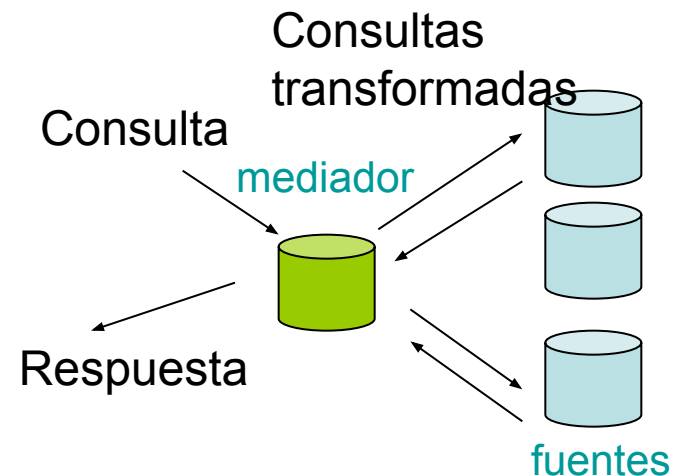
Registro de nuevas ventas es
bloqueado

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

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 intérpretes (“wrappers”) requieren de gran complejidad
- Ventajas Integración activa:
 - Es mucho más común en la práctica
 - Mejor rendimiento
 - Menor complejidad
 - Datos antiguos se manejan mejor

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

Ej: Data Mart

XPGESTIONPAX/PROY
D_AV_MXV
NEGOC_M
MATERIAL
CUENTA
SCENARIO

PX
TOTAL MATERIAL
Tickets
Real

	200501	200502	200503	200504	200505	200506	200507	200508	200509	200510	200511	200512	200601
Total Avión Mixto	126.249.977,08	103.873.711,03	111.906.293,18	99.026.517,77	95.292.779,45	94.590.833,31	120.863.438,27	113.913.037,81	111.800.962,50	116.783.559,00	129.043.669,81	134.242.654,40	147.245.679,40
Total LanChile	99.438.761,32	81.159.821,03	85.311.922,72	77.025.763,60	72.703.524,77	70.530.796,43	87.468.327,95	79.534.534,98	82.641.100,77	87.230.442,30	96.381.011,88	98.939.453,27	109.501.287,17
Total INT LA	72.876.986,53	59.088.594,53	62.117.101,34	56.873.512,27	54.211.215,52	53.261.735,80	66.557.687,22	58.183.271,73	60.978.733,26	62.318.762,10	65.822.971,24	70.222.133,81	78.545.288,89
Internacional Lan Chile	72.876.986,53	59.088.594,53	61.721.804,93	55.716.846,55	54.211.215,52	53.261.735,80	66.557.687,22	58.016.531,52	60.301.319,80	61.828.173,84	65.822.971,24	70.222.133,81	78.545.288,89
BRA	5.085.952,54	4.556.985,19	4.344.014,61	4.134.135,30	3.815.016,75	4.038.546,18	5.855.210,03	4.731.492,64	4.976.452,13	4.606.107,77	4.819.516,36	4.813.456,84	5.632.391,04
BUE	6.194.388,36	5.395.490,87	6.984.589,96	6.451.262,64	6.948.591,09	6.061.090,71	7.237.074,19	6.986.244,67	7.033.445,04	7.502.944,66	9.159.718,61	8.610.005,72	7.899.440,82
CCS	1.572.077,78	1.136.248,10	1.117.296,09	1.052.931,52	967.654,15	908.392,94	1.301.446,07	1.462.570,24	1.502.624,57	1.255.857,29	1.319.006,30	1.677.856,38	1.805.697,17
COR	731.479,59	650.320,94	715.377,22	711.972,87	739.139,69	760.576,07	933.491,89	894.417,95	789.742,47	847.245,54	827.809,24	913.954,09	956.125,78
CUN	903.353,04	881.382,97	65.647,55	-	-	-	-	-	-	-	-	948.427,76	138.801,62
EUR	12.391.471,21	9.901.279,98	11.067.644,46	10.477.102,80	9.747.030,78	9.939.013,82	10.952.603,93	9.409.526,46	10.714.361,19	12.069.365,87	11.234.583,14	10.943.561,59	11.811.328,34
GBD	2.565.722,21	1.749.120,44	2.064.402,92	1.425.307,13	1.657.602,54	1.844.267,10	2.220.196,89	1.813.021,97	1.683.266,18	1.805.411,77	2.068.514,29	2.519.846,73	2.748.990,85
HAV	2.018.376,61	1.939.275,69	1.409.774,86	1.521.425,86	1.607.521,43	1.592.148,60	1.918.765,01	1.277.575,83	1.651.857,92	1.126.649,33	645.523,53	-	1.302.665,41
IPC	1.118.013,14	1.051.686,50	943.427,31	390.102,37	231.617,21	191.135,00	613.525,24	700.917,97	547.125,88	752.607,15	1.116.239,57	1.075.199,31	1.628.630,61
LAX	8.017.950,05	6.111.818,98	6.649.961,75	6.148.086,57	5.790.736,38	5.794.249,43	6.562.057,92	6.284.102,34	6.679.324,17	6.565.725,41	6.566.938,00	6.870.430,21	7.522.272,31
LIM	-	-	-	-	-	-	-	-	-	-	-	-	167.447,02
MDZ	515.763,85	443.538,01	636.162,57	542.194,98	667.856,03	747.169,61	695.298,65	718.273,14	662.404,19	577.053,91	739.652,63	722.035,54	713.698,28
MEX	5.046.596,65	3.727.545,42	4.142.610,49	3.888.120,04	3.708.822,12	3.819.719,11	4.894.527,63	3.998.186,49	4.123.503,05	5.132.423,44	5.028.128,60	5.314.926,79	6.001.436,82
MIA	9.561.777,76	7.146.742,01	6.581.281,91	5.393.774,62	5.174.297,60	5.101.847,04	7.852.791,13	5.145.751,60	5.699.833,36	4.648.016,28	6.028.774,06	8.908.952,07	10.878.002,10
MYD	1.094.433,02	1.010.737,08	967.342,47	797.954,33	784.718,22	722.847,86	892.433,20	768.821,29	817.881,21	945.987,91	1.097.606,42	1.122.765,71	1.344.492,19
NNS	-	-	-	-	-	-	-	-	-	-	-	-	-
NYC	7.045.834,14	5.502.493,82	6.827.055,86	6.185.413,93	6.031.860,37	5.721.857,28	6.828.701,52	7.243.956,94	7.021.868,09	7.122.186,63	7.047.849,87	7.300.759,14	7.694.556,47
PPT	2.209.526,65	1.841.293,47	1.801.044,78	1.200.857,24	1.389.497,69	1.114.582,21	1.404.632,14	1.504.507,78	1.241.847,96	1.494.853,28	1.686.686,30	1.835.879,30	2.082.359,65
PUJ	1.297.871,14	1.250.809,50	507.086,87	622.905,18	685.009,85	746.879,95	1.680.374,62	691.029,20	707.493,26	711.159,07	561.002,40	540.980,50	1.405.770,86
ROS	-	-	-	-	-	-	-	-	-	-	-	-	-
SSA	488.769,28	518.721,43	244.561,84	197.150,44	185.738,65	163.958,45	264.427,38	45.018,17	-	-	-	-	377.607,31
SYR	3.564.004,27	3.032.924,42	3.161.786,21	3.280.428,63	2.830.011,71	2.849.534,07	3.067.431,72	2.979.635,07	3.189.734,47	3.311.441,12	4.453.112,14	4.785.564,58	5.017.024,95
UIO	1.453.625,24	1.240.179,71	1.490.735,20	1.295.720,10	1.248.493,26	1.143.930,37	1.382.708,06	1.361.481,77	1.258.554,66	1.353.137,41	1.422.309,78	1.317.531,55	1.416.559,29
LGM	-	-	-	-	-	-	-	-	-	-	-	-	-
LGN	-	-	-	-	-	-	-	-	-	-	-	-	-
ECU	-	-	-	-	-	-	-	-	-	-	-	-	-
LNS	-	-	-	-	-	-	-	-	-	-	-	-	-
SGN	-	-	-	-	-	-	-	-	-	-	-	-	-
CAT	-	-	-	-	-	-	-	-	-	-	-	-	-

REVISAR Producción Total Tickets Rechazos IB Caducos Otros Ing volados Ex Equip Ing LPass INGDUTY ArrAviones Otros Ing no volado COMISIONESPAX Cc

Calcular

Inicio 2006 FORECAST 2005 REALES ACUERDO QF... ACUERDO QF... INGRESOS_2... Inbox - Micros... Windows Live ... ES 12:29 PM

Diseño de un DW

- Enfoque **Top-Down**:

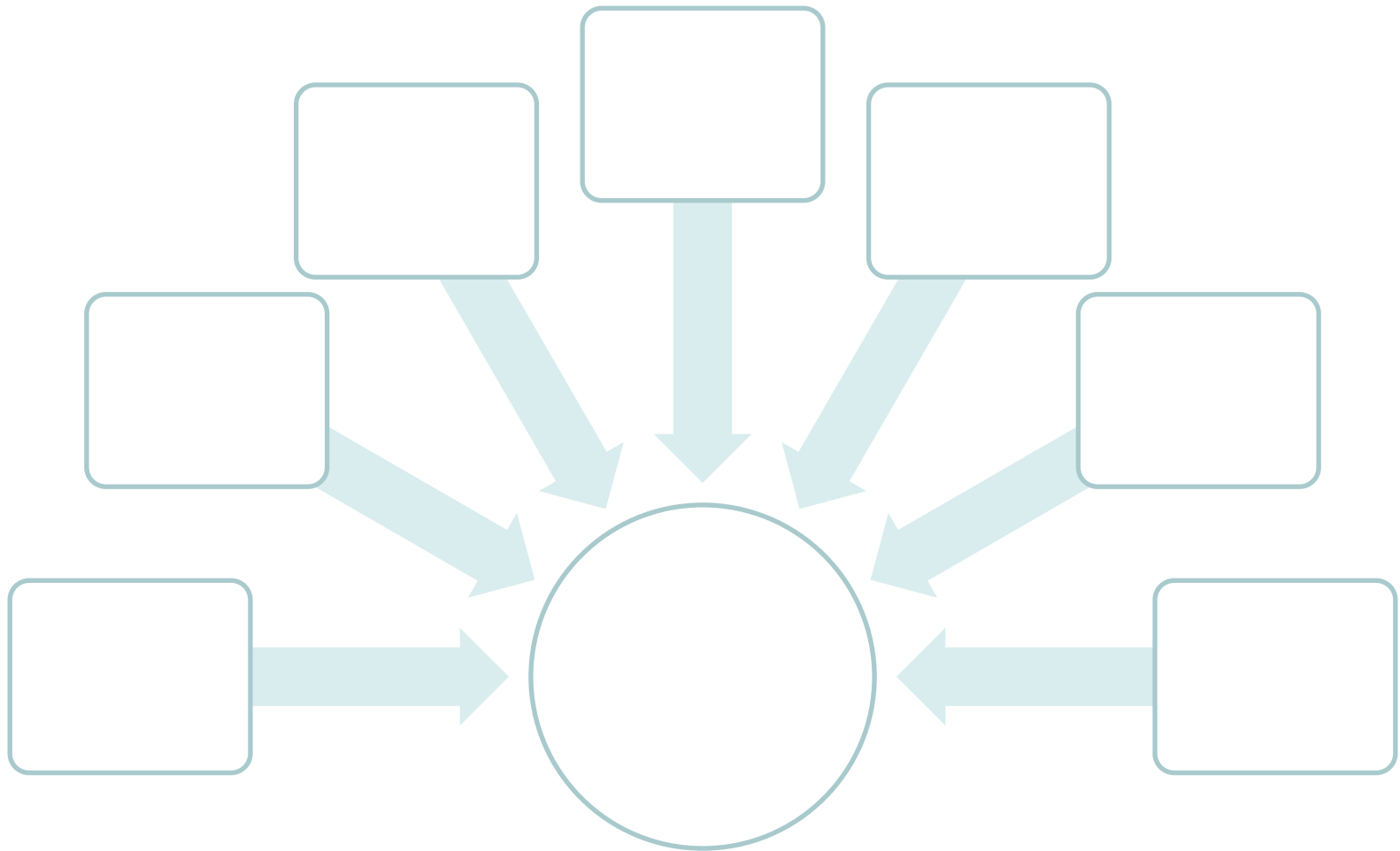
- Comienza con el diseño completo y planificación total.
- Es útil cuando la tecnología es bien conocida, madura, estable y cuando los problemas de negocios a resolver se conocen bastante bien.

- 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.

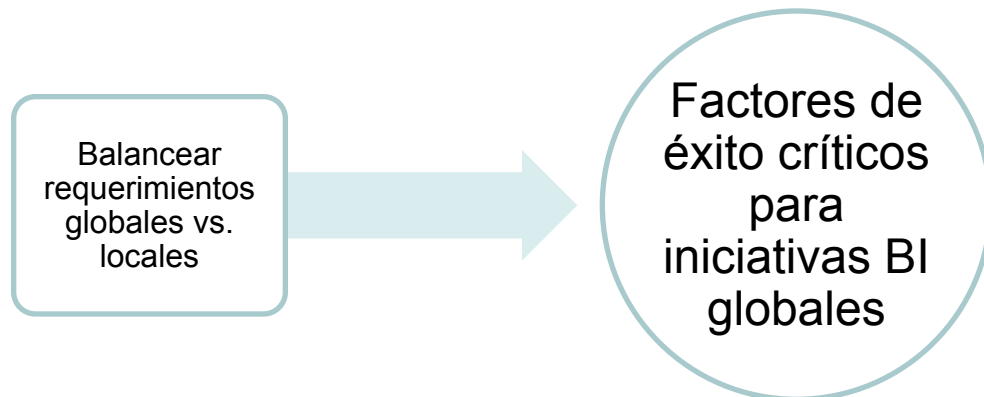
- Combinando los dos anteriores

Iniciativas de BI globales: 7 factores de éxito



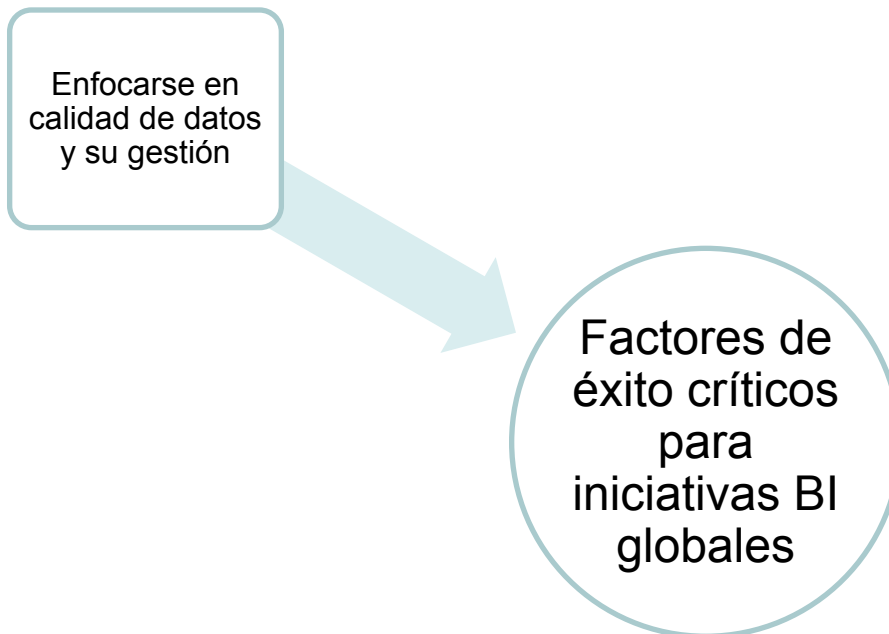
1: Balancear requerimientos globales vs. locales

- **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

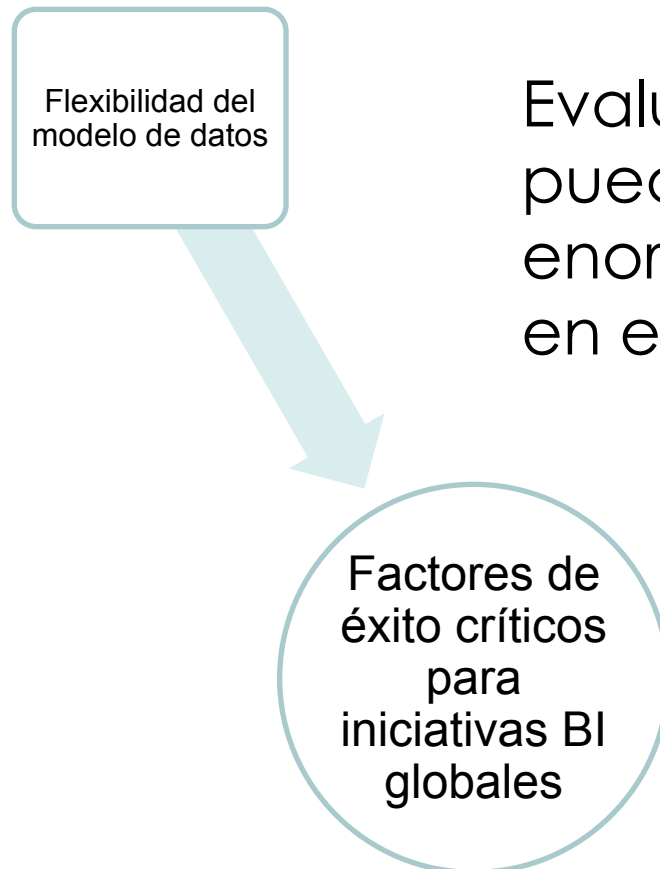


- **Estándares de calidad** que aprueben la migración de datos

3: Flexibilidad del modelo de datos

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

Flexibilidad del
modelo de datos



```
graph TD; A[Flexibilidad del modelo de datos] --> B((Factores de éxito críticos para iniciativas BI globales));
```

A light blue arrow points from the rectangular box containing 'Flexibilidad del modelo de datos' to the circular box containing 'Factores de éxito críticos para iniciativas BI globales'.

Evaluar “**escenarios outliers**” puede evitar incurrir en enormes costos de cambios en el modelo de datos

Factores de
éxito críticos
para
iniciativas BI
globales

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).

Arquitectura enfocada en estandarización, reusabilidad y automatización



Factores de éxito críticos para iniciativas BI globales

La **automatización** reduce enormemente los **esfuerzos humanos** en la medida en que el proyecto va **progresando**

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**

```
graph TD; A[Profundidad del conocimiento sobre los datos y los procesos] --> B((Factores de éxito críticos para iniciativas BI globales));
```

Profundidad del conocimiento sobre los datos y los procesos

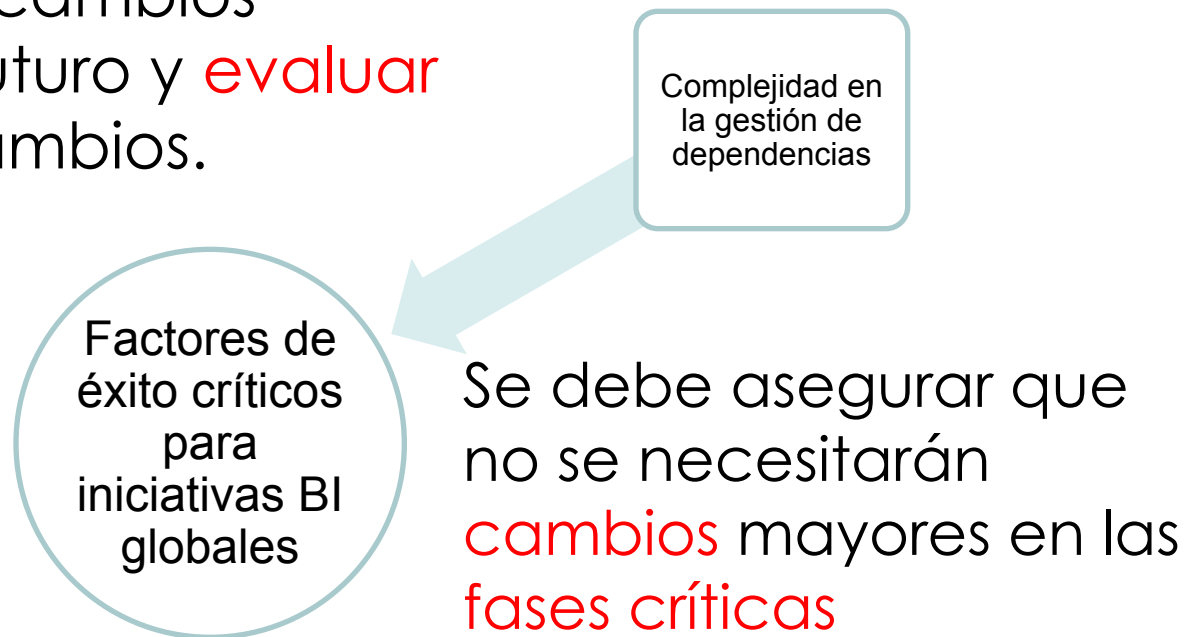
Factores de éxito críticos para iniciativas BI globales

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

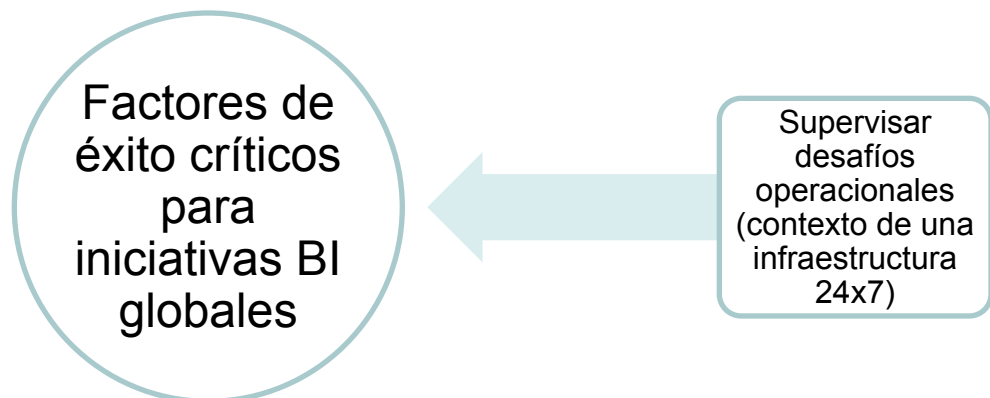
En general existen **altos grados de dependencia** entre las iniciativas BI. Todas estas interdependencias deben ser rigurosamente gestionadas.

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



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

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

Objetivos del diseño Físico

Satisfacer los requerimientos que el diseño lógico impone en forma óptima



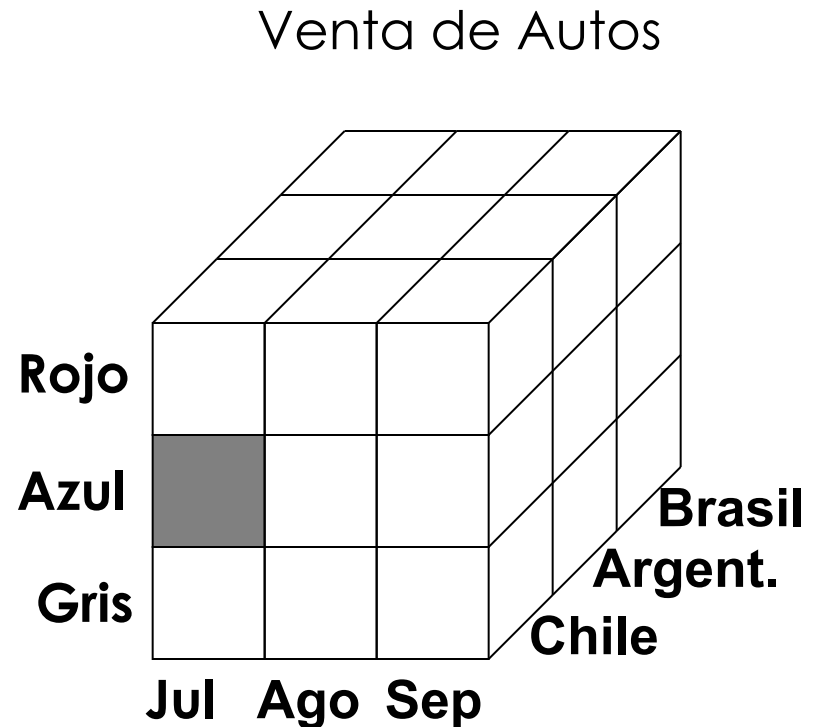
No malgastar recursos



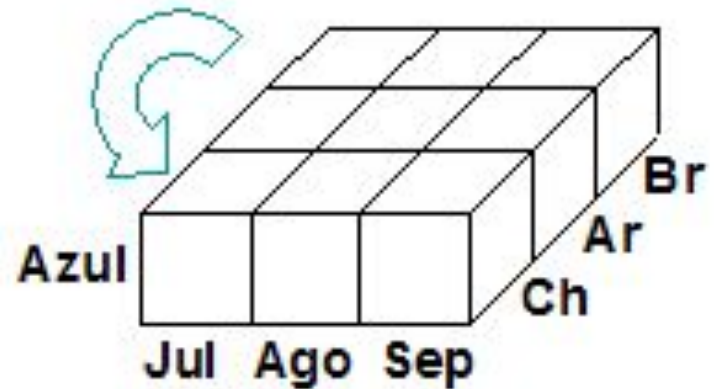
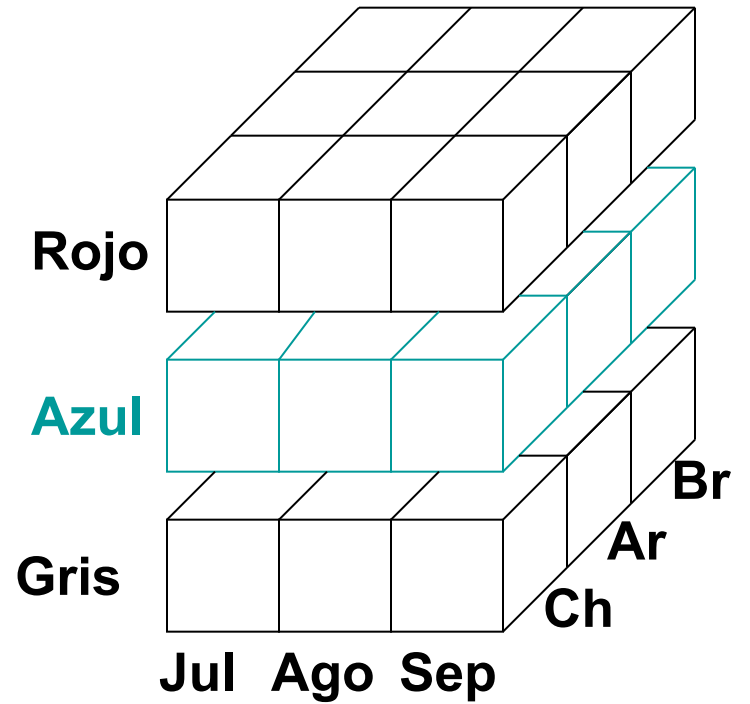
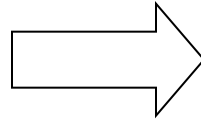
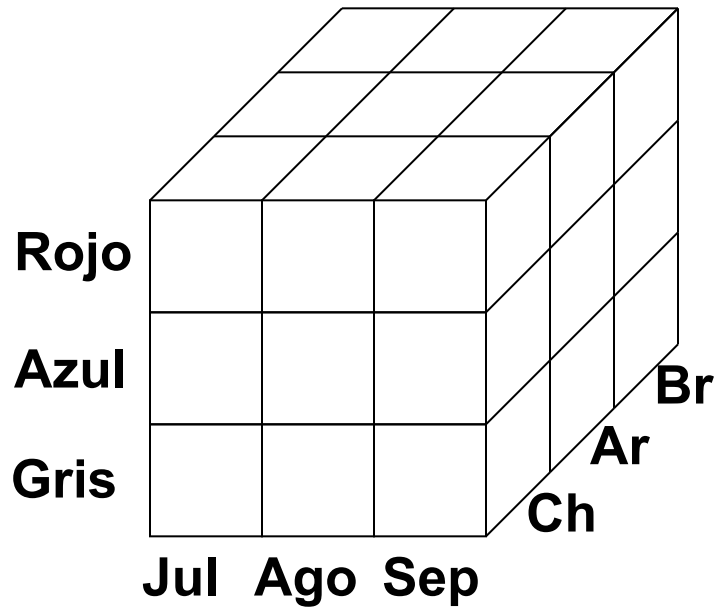
Permitir acceso
eficiente a la
información

Data Cube

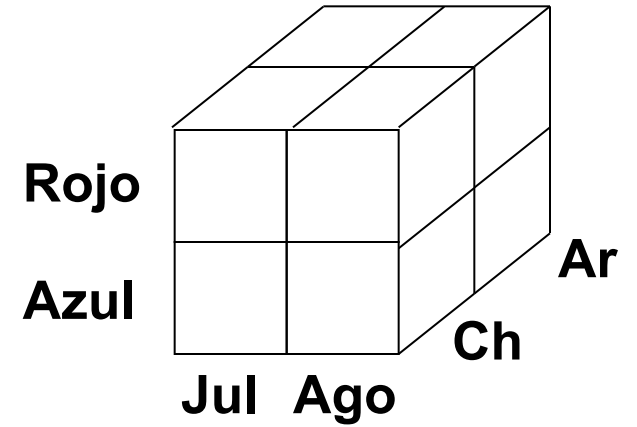
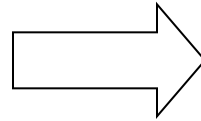
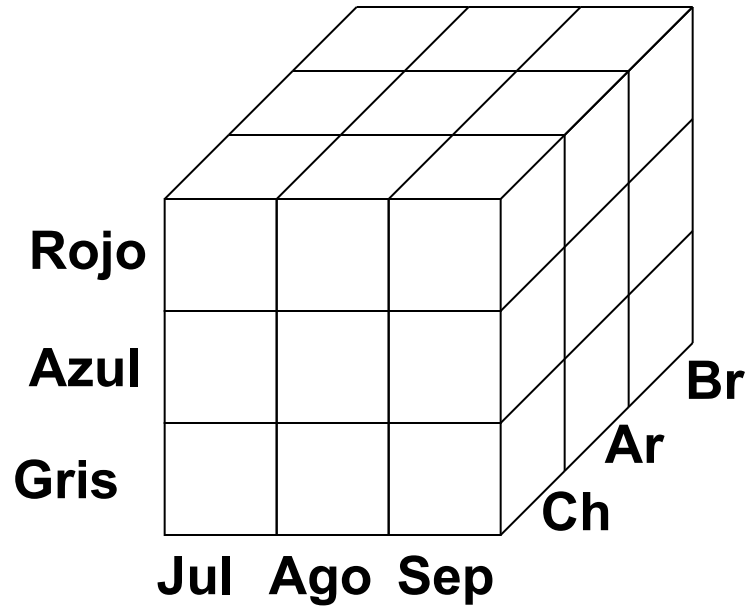
- Ejes representan atributos
 - Generalmente discretos
 - Ej. Color, mes, lugar, etc.
 - También llamados dimensiones
- Celdas guardan información agregada
 - Ej. Ventas totales de autos
- En la práctica datacubes tienen mucho más de 3 dimensiones



Slicing



Dicing



Roll Up and Drill Down

Número de autos vendidos

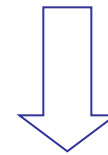
	Ch	Ar	Br	Total
Jul	45	33	30	108
Ago	50	36	42	128
Sep	38	31	40	109
Total	133	100	112	345



Roll Up
por mes

Número de autos vendidos

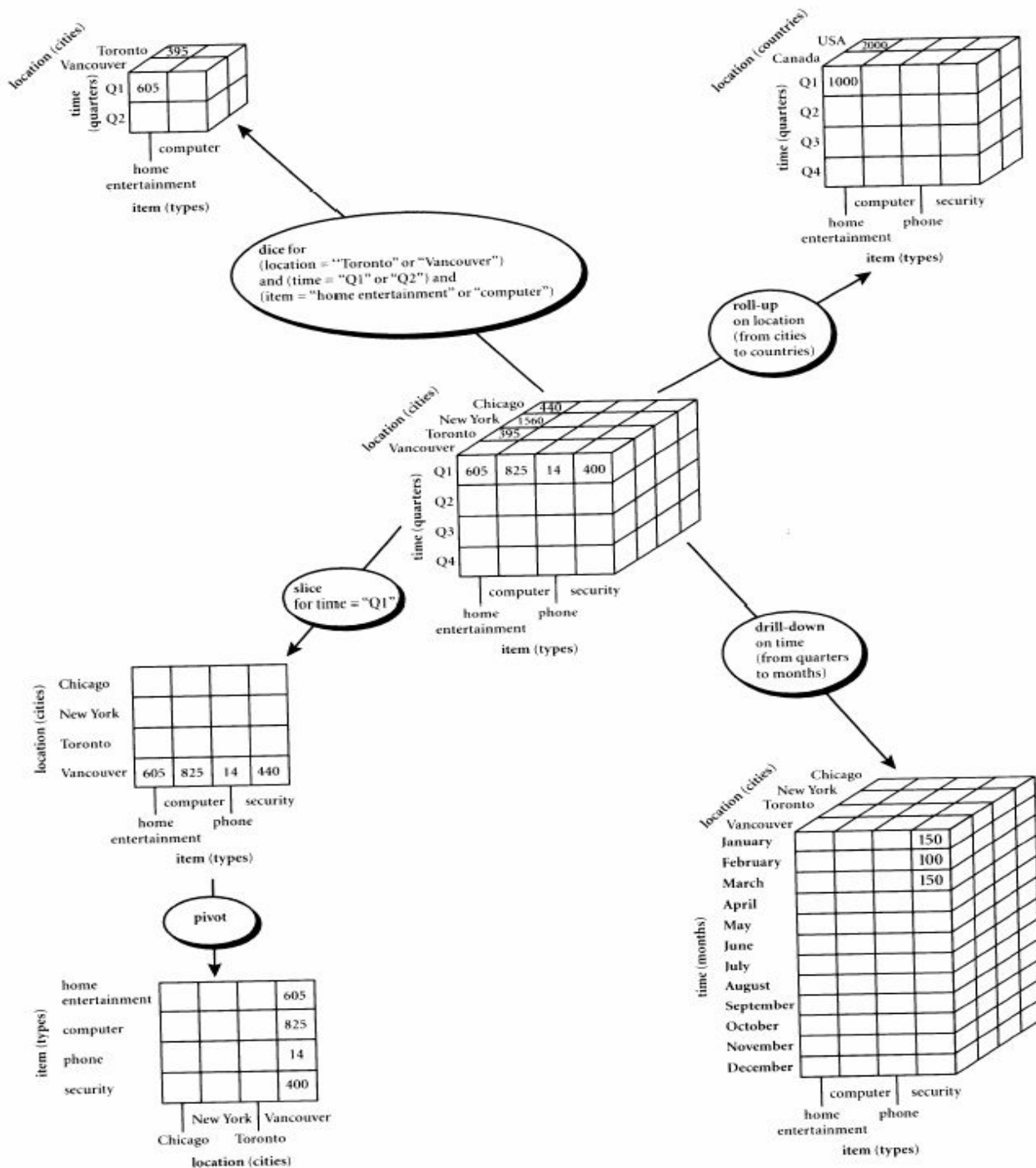
Ch	Ar	Br	Total
133	100	112	345



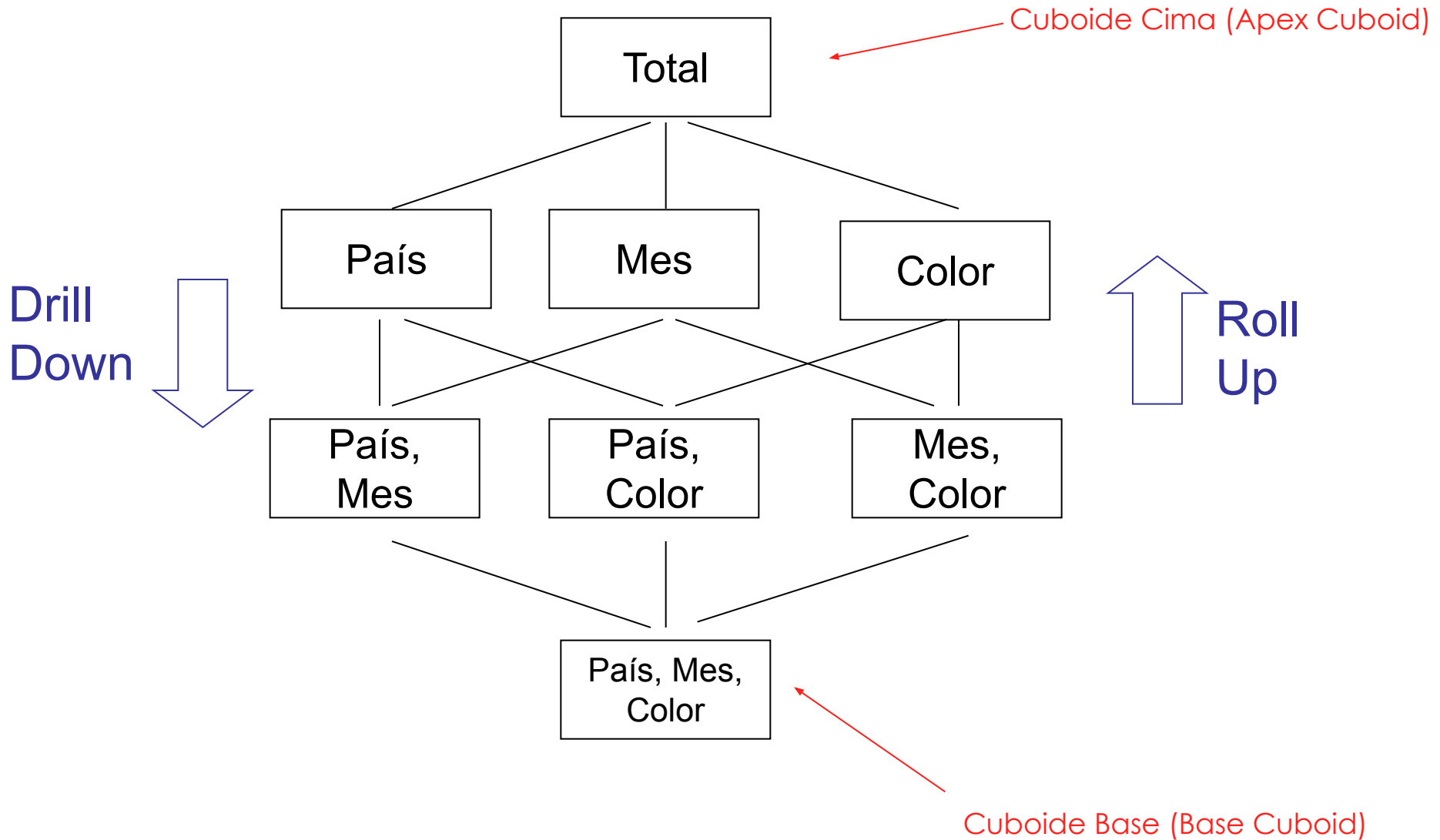
Drill down
por color

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



Data Cube Lattice



- 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)

(*,Toronto, *,1200)

(Enero,*,Business,150)

(Enero, Toronto, Business,45)



Celdas 1D

Celda 2D



Celdas agregadas

Celda 3D

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, proveedores, 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

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)

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

Tablas de Hechos vs Tablas de Dimensión

Tabla de hechos

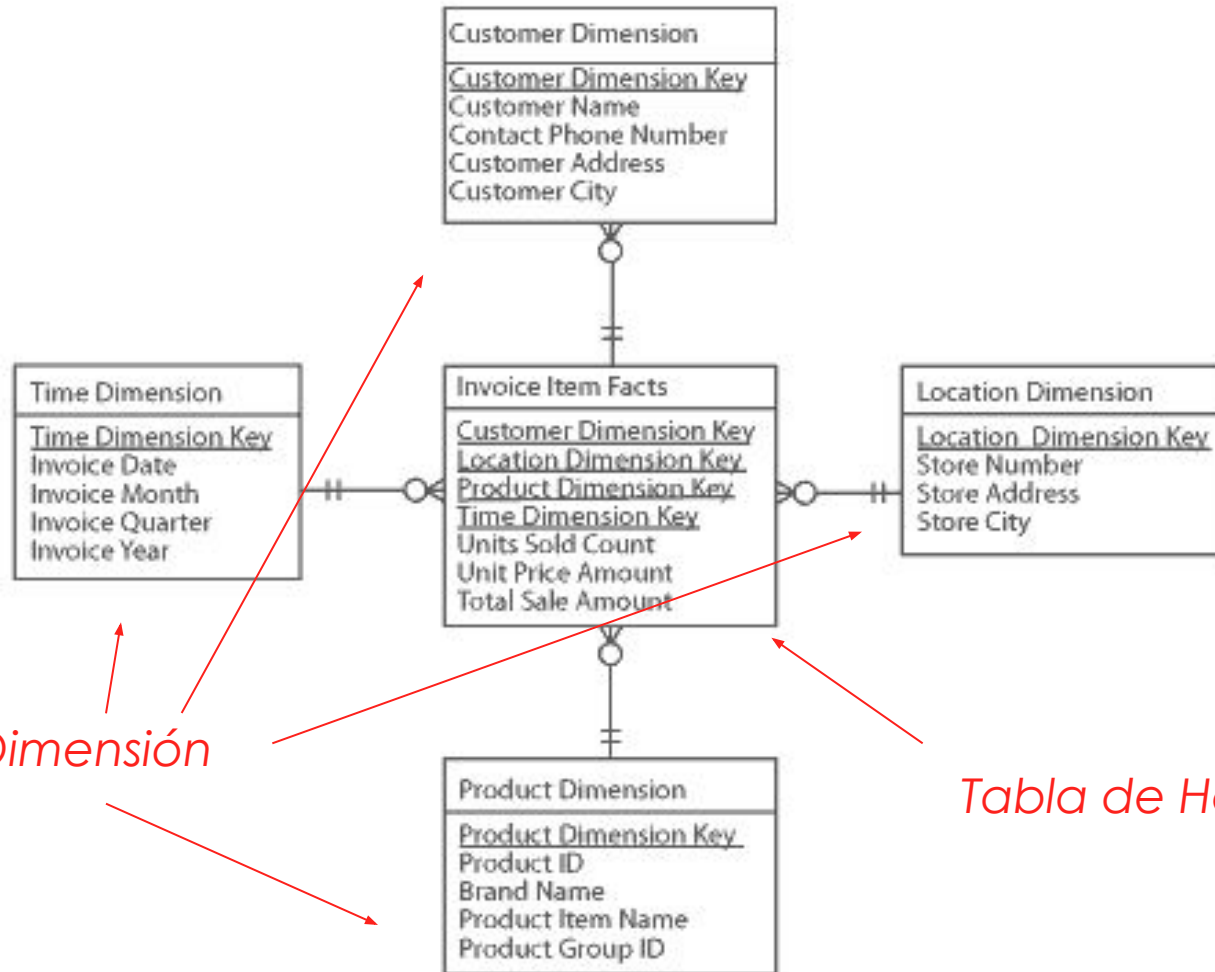
- Angostas
- Gigantes (muchas filas)
- Numéricas
- Crecen en el tiempo

Tabla de dimensión

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

Modelamiento de los Datos

Star Schema:

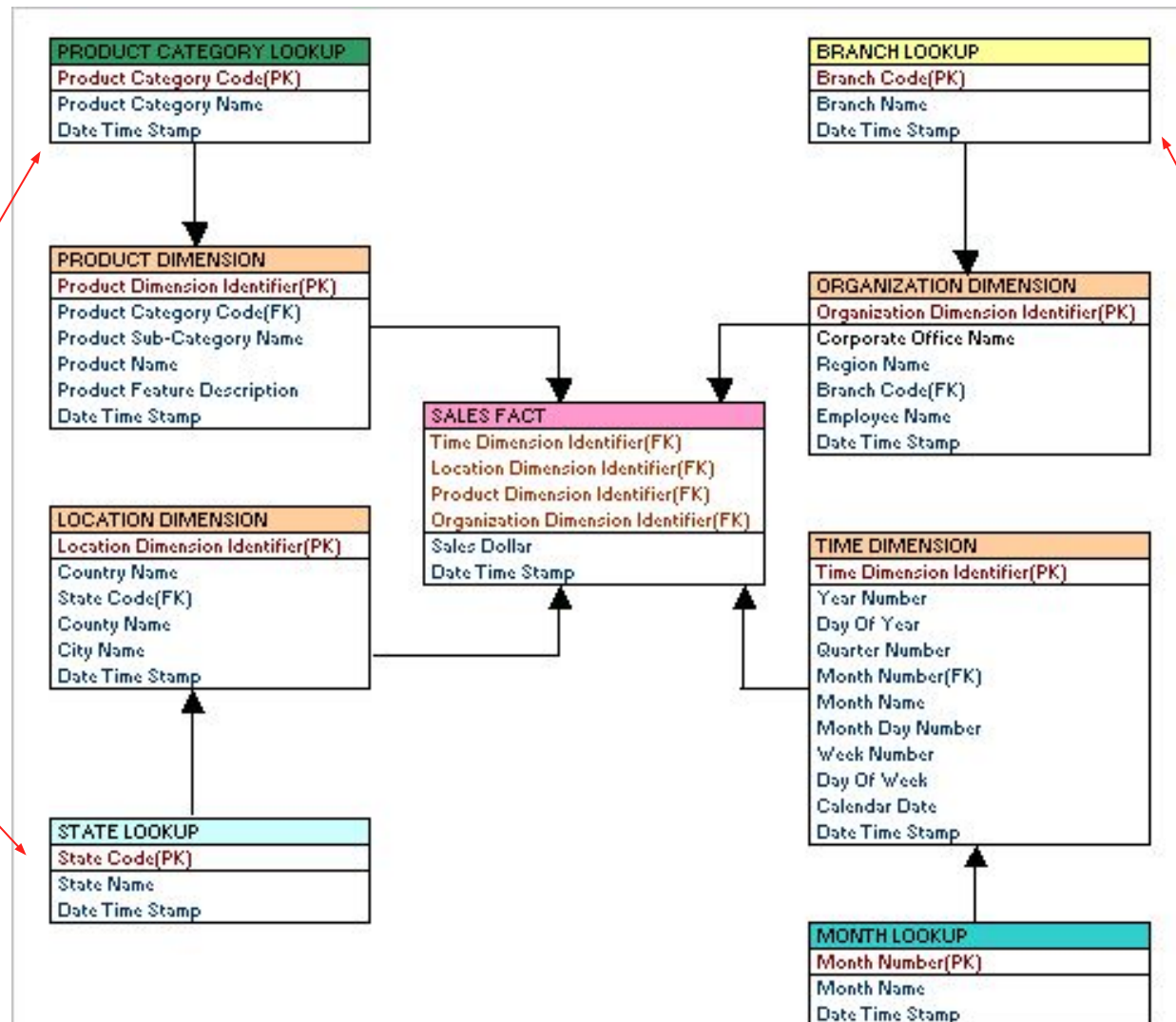


Tablas de Dimensión

Tabla de Hechos

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

Tablas agregadas



Tablas agregadas

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

Store Dimension

STORE KEY

Store Description
City
State
District ID
District Desc.
Region_ID
Region Desc.
Regional Mgr.

Fact Table

STORE KEY PRODUCT KEY PERIOD KEY

Dollars
Units
Price

Product Dimension

PRODUCT KEY

Product Desc.
Brand
Color
Size
Manufacturer

Time Dimension

PERIOD KEY

Period Desc
Year
Quarter
Month
Day
Current Flag
Sequence

District Fact Table

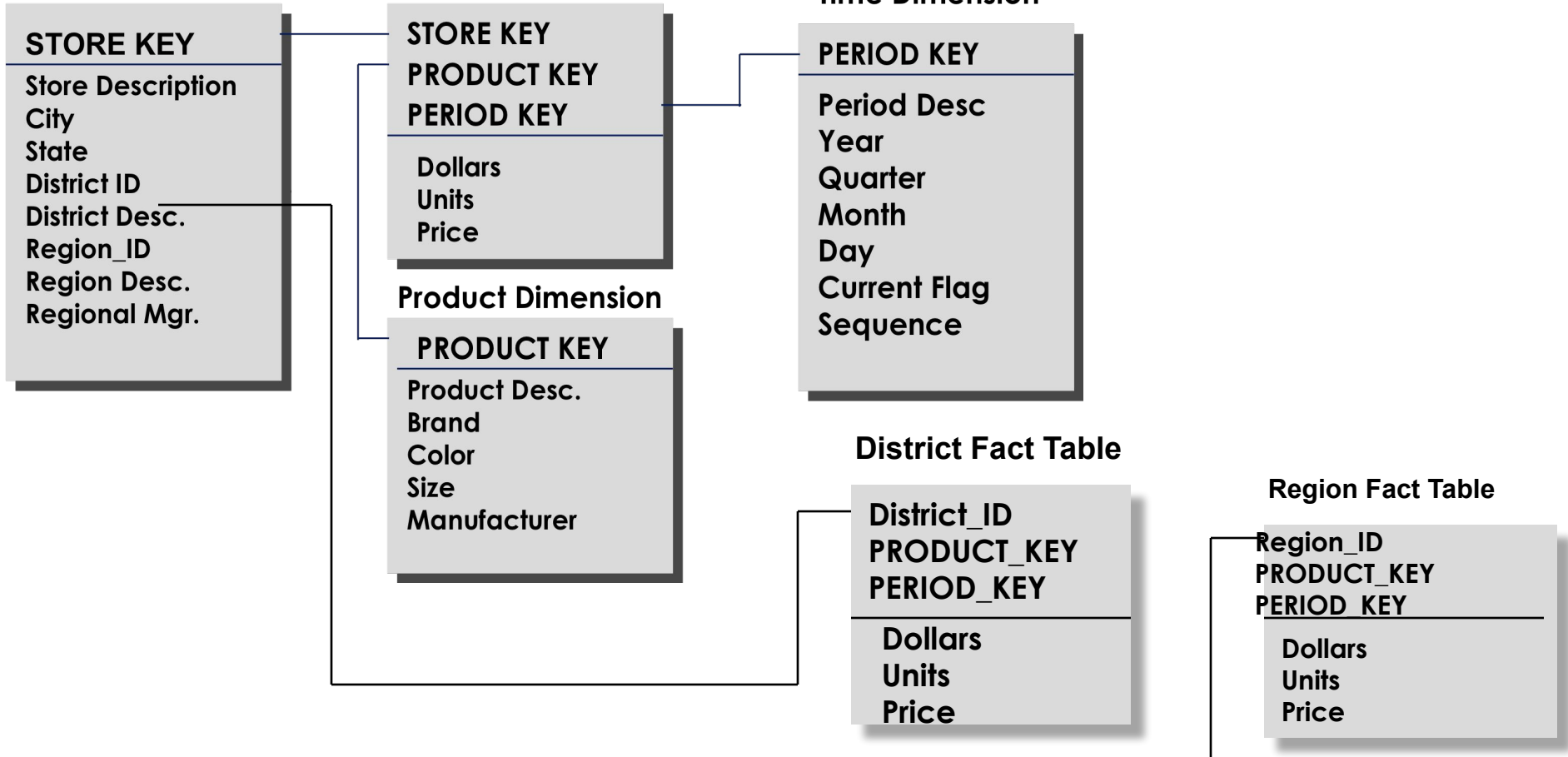
District_ID PRODUCT_KEY PERIOD_KEY

Dollars
Units
Price

Region Fact Table

Region_ID PRODUCT_KEY PERIOD_KEY

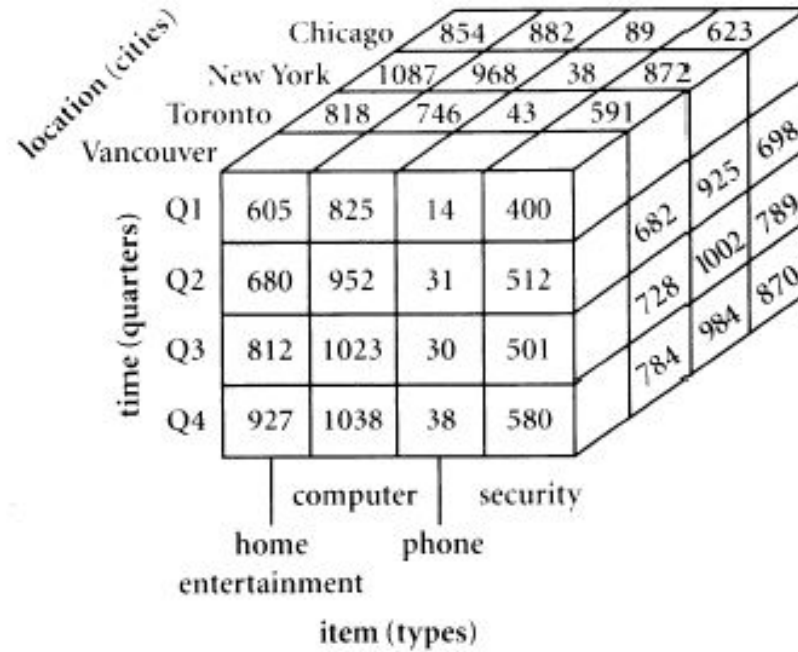
Dollars
Units
Price



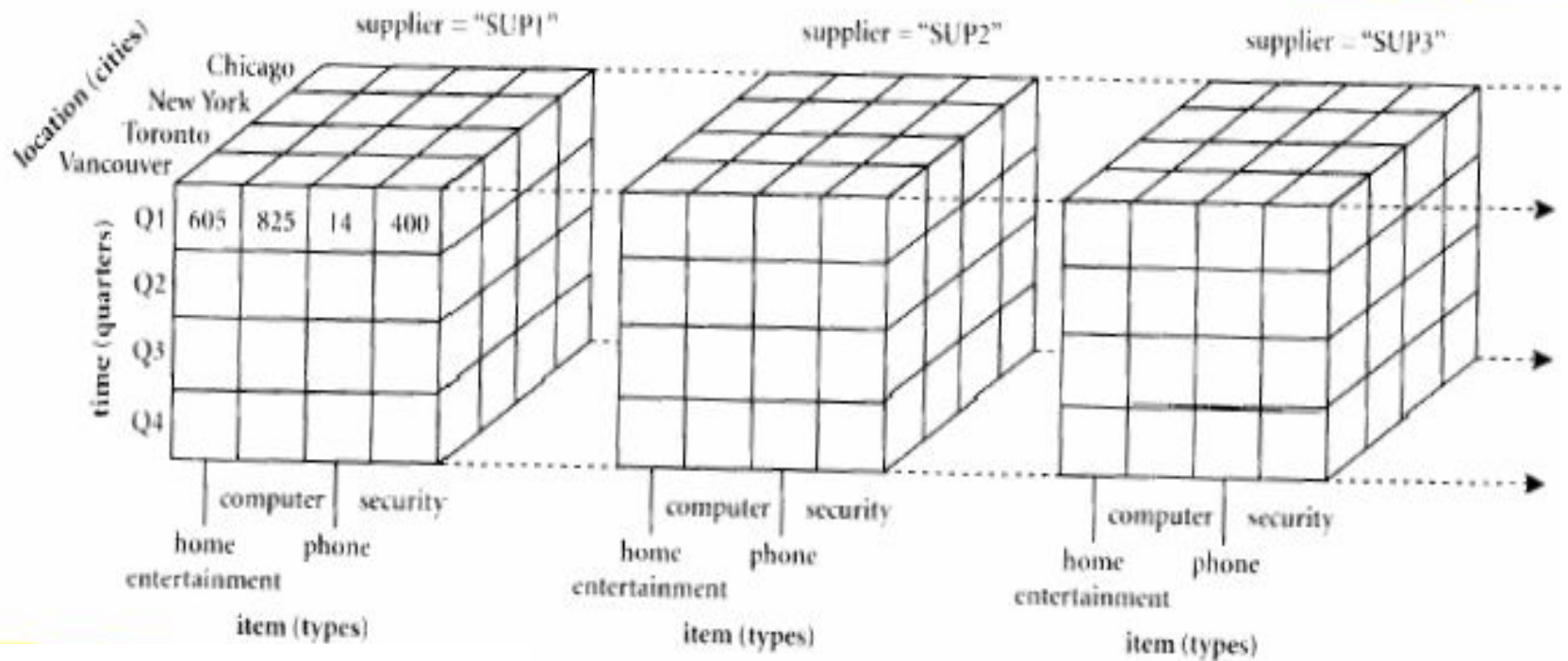
Ejemplo:

	<i>location</i> = "Chicago" <i>item</i>				<i>location</i> = "New York" <i>item</i>				<i>location</i> = "Toronto" <i>item</i>				<i>location</i> = "Vancouver" <i>item</i>			
<i>time</i>	<i>home ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>home 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

Data Cube

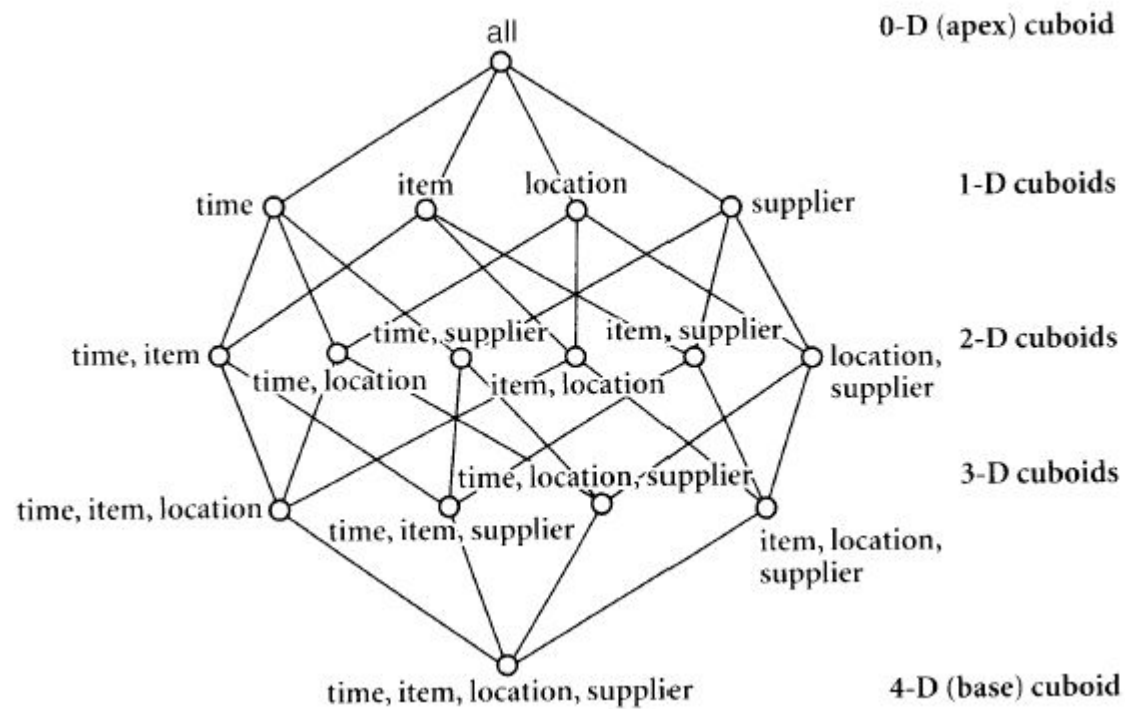


Representación 4-D del cubo de datos

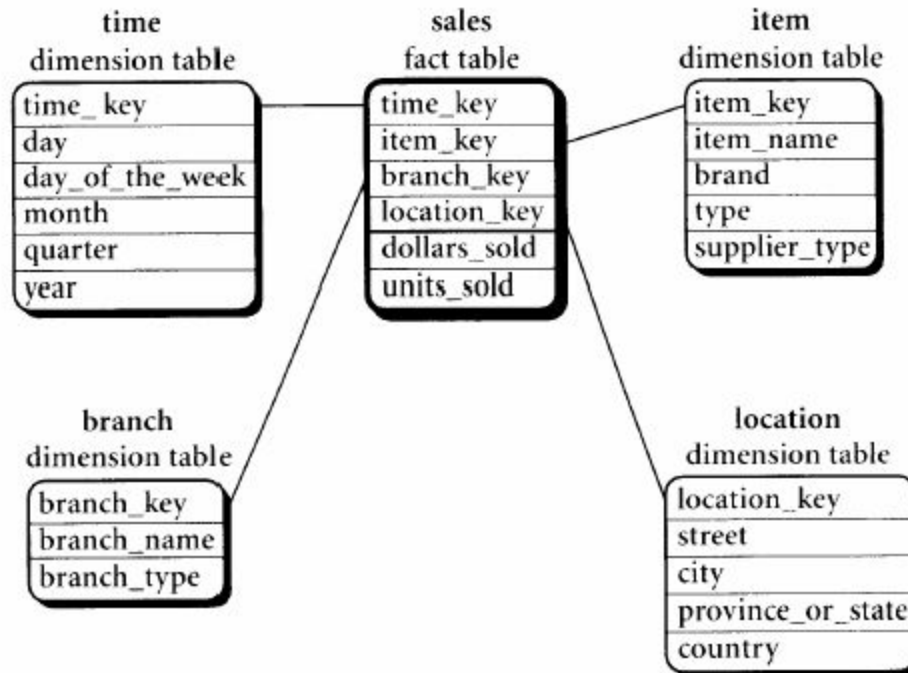


Se agregó la dimensión "supplier"

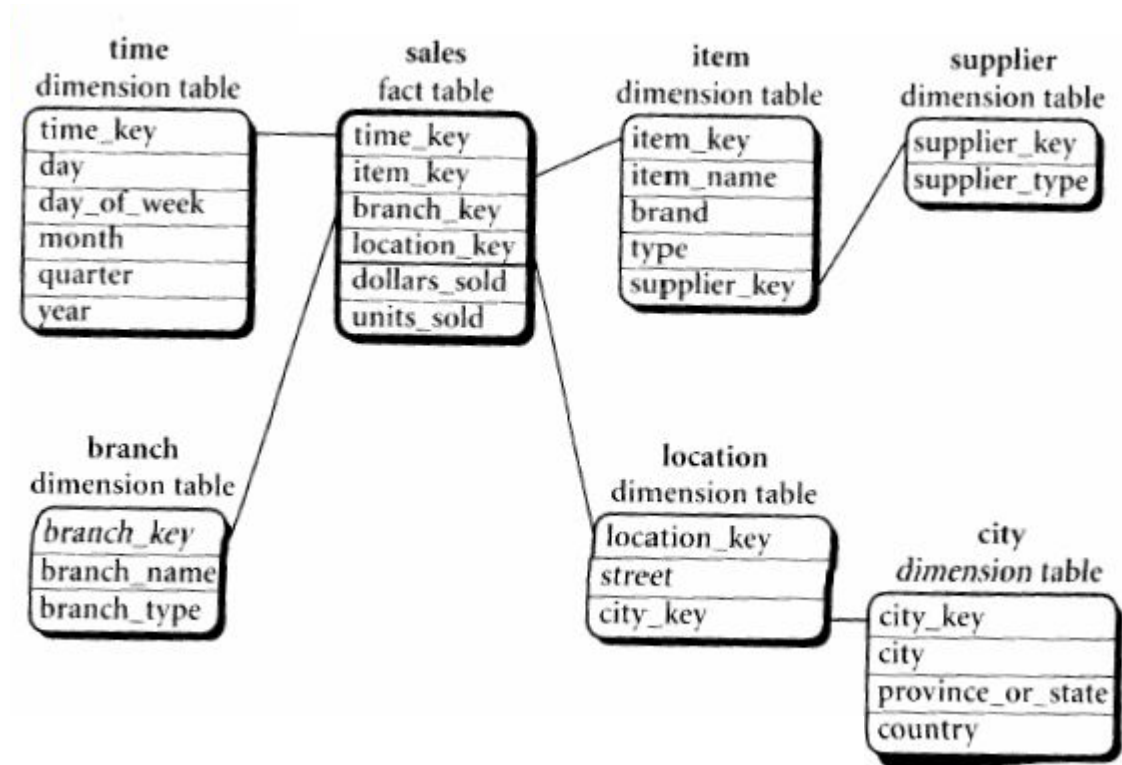
Data Cube Lattice



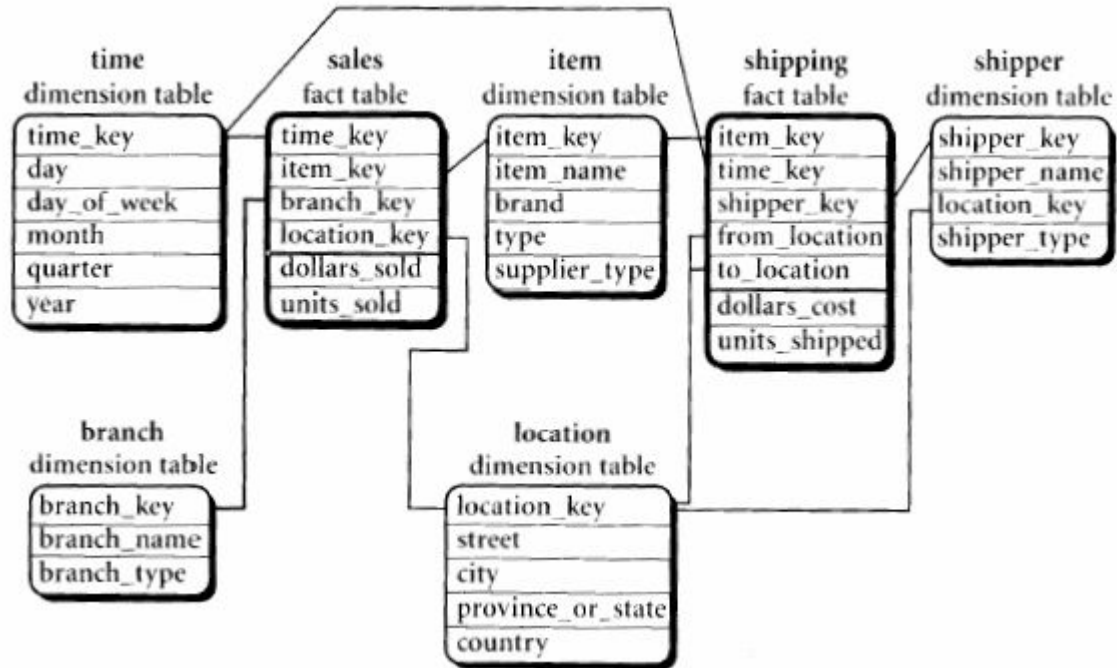
Star Schema



Snowflake Schema



Fact Constellation Schema



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

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

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

1. Identificar el proceso a ser modelado

- Ej: Datos en Supermercado

- Datos adquiridos por cajas 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

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

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

3. Elegir dimensiones

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

–Ejemplo:

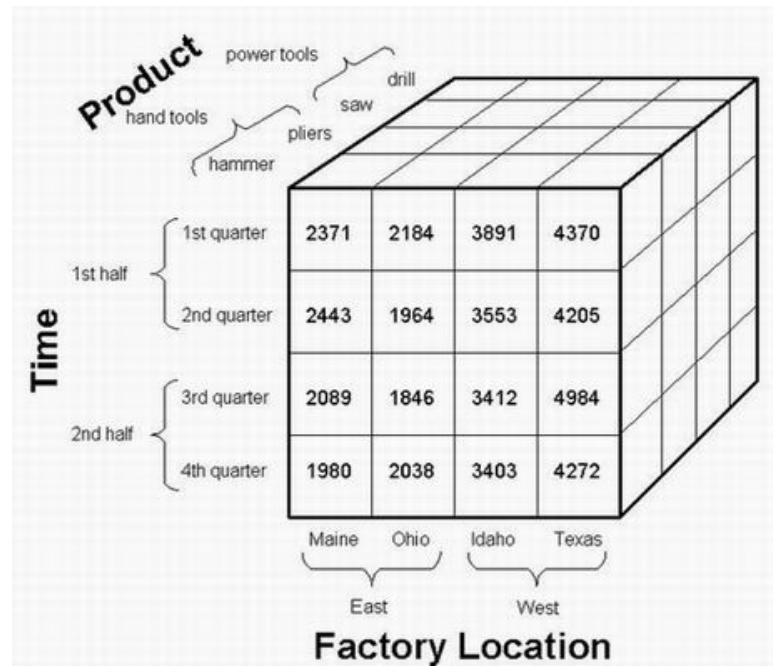
- filas de tabla de hechos representan alumnos en un curso
- Dimensiones posibles pueden ser curso, estudiante, semestre, etc.

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

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



Densidad de los Datos

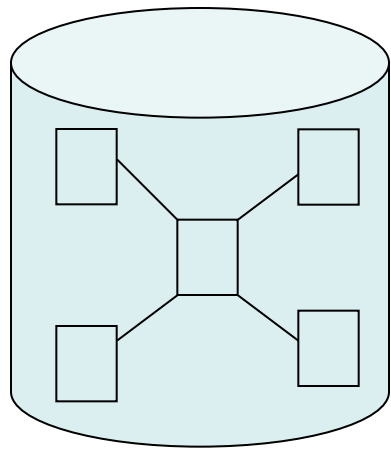
- Imagine un DW de una cadena de tiendas
- Dimensiones: consumidores, productos, tiendas y días
- Suponga que hay 100.000 clientes, 10.000 productos, 1.000 tiendas y un período de 1.000 días
- Cubo de datos tiene 1,000,000,000,000,000 de celdas

Densidad de los Datos

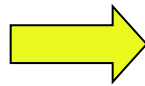
- Afortunadamente la mayoría están vacías
- Una tienda determinada no vende cada producto cada día
- Un consumidor no visita las 1.000 tiendas diariamente, quizás nunca visita más de 2 o 3 de las tiendas
- Un consumidor no compra todos los productos
- ¿Será esto un inconveniente para el uso de arreglos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

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)



DW



ROLAP
TOOLS

• **MOLAP**

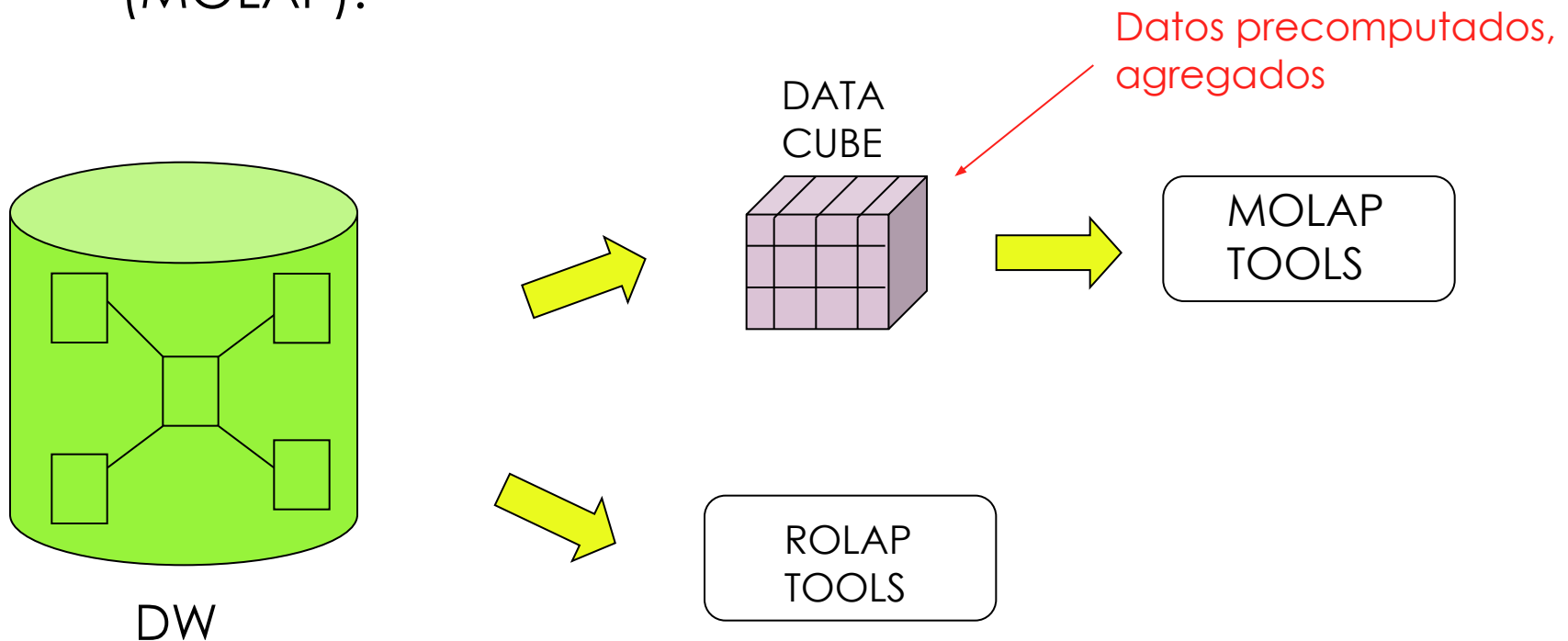
- Usualmente precalcula valores agregados
- Acceso eficiente a los datos, respuestas rápidas
- No escala bien a muchas dimensiones

• **ROLAP**

- 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

HOLAP

- Hybrid On Line Analytical Processing
 - Los datos son almacenados en un modelo relacional (ROLAP), para el análisis y respuesta de consulta se traspasan los datos a una tabla multidimensional (MOLAP).



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

Número de Dimensiones

–Factores a considerar

- ¿Cómo los usuarios piensan acerca de los datos?
- ¿Es un anuncio y un cupón promociones separadas o dos aspectos de la misma promoción?

–Menos tablas es bueno pues implica un diseño más simple

–Considerar rendimiento...

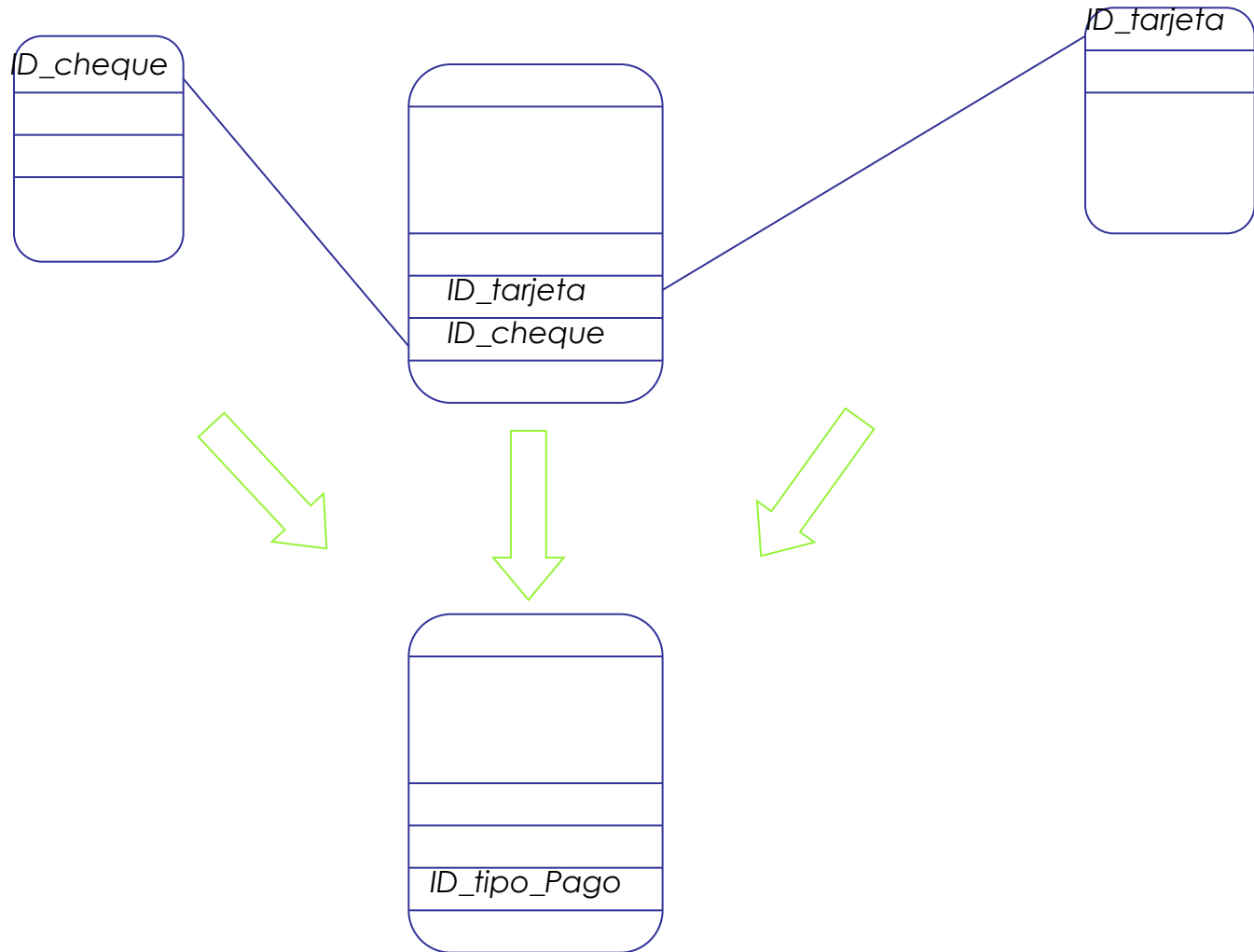
Número de Dimensiones

- Factores de rendimiento
 - Aspectos dimensionales afectan significativamente los requerimientos de disco
 - Consultas OLAP
 - Intensivas en acceso a datos no en cálculos
 - Lectura a disco es el cuello de botella
 - Planear para consultas típicas en hardware típico (en general se repiten ciertas consultas en ciertos PC's y discos)
 - Tamaño del disco
 - Menos tamaño puede ser importante

Atributos problemáticos

- Algunos atributos no son modelados eficientemente por ninguna dimensión
 - Método de pago (efectivo vs. tarjeta vs. cheque)
 - Bolsa (papel, plástico, etc.)
- Crear una nueva dimensión y ponerlos a todos
 - No importa que se pierda relación lógica
 - Reduce número de tablas de dimensión y ancho de tabla de hechos
- Alternativas
 - Crear nueva dimensión para atributo
 - Eliminar el atributo

Atributos problemáticos



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

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 ?

Preservar Historia

- Historial 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

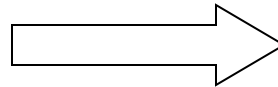
- Hecho antiguo apunta a 457
- Nuevos hechos apuntan a 784

Dimensión cliente

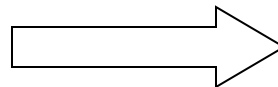
Key	Nombre	Sexo	Ciudad	Año
457	Juan P.	M	Iquique	2000
...
784	Juan P.	M	Stgo	2003

Tabla de hechos

Hechos antiguos



Hechos nuevos



Key	...	Cantidad
...
457	...	5
...
784	...	4

Actualizar vs preservar historia

- Actualizar
 - +Simple
- Preservar Historia
 - +Mayor exactitud
 - +Información agregada no se ve afectada
 - Tabla de dimensión cambia
 - Puede haber problemas con consultas como:
nombre=Juan (i.e. reportes en actividades de
Juan en el tiempo)

Actualizar vs preservar historia

- Ambas formas son usadas
- En general usar mezclas
 - Preservación para algunos atributos
 - Actualizar otros
- Cosas a considerar:
 - En general preservar historia es mejor
 - ¿ Consultas necesitarán antiguo o nuevo valor ?
 - ¿ El beneficio de preservar la información es mayor que tener que agregar filas a la tabla de dimensión correspondiente?
 - En casos como nuevo número de teléfono probablemente no, aunque de todas formas depende del caso

Operación y Mantenición

- 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

Operación y Mantenición (cont..)

- Soporte de atención de usuarios (help desk)
 - Manuales, mensajes, etc.
- Soporte de operación
 - Verificar el correcto funcionamiento de la DW
- Soporte de mantención 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)

¿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

Crecimiento de Wal Mart (500%)



Fuente Greg McMullen – Ford Motor Company

Wal Mart - Usos

- Tipos de información y consulta:
 - Información resumida de ventas diarias
 - Análisis de ventas regionales
 - Análisis de ventas por producto
 - Análisis de impacto de promociones
 - Análisis de tendencias
 - Etc.
- Wal Mart usa esta información al negociar con:
 - Proveedores
 - Agencias de promociones
 - Sistema de transporte y distribución
 - Etc.

Valor de la acción de 3M



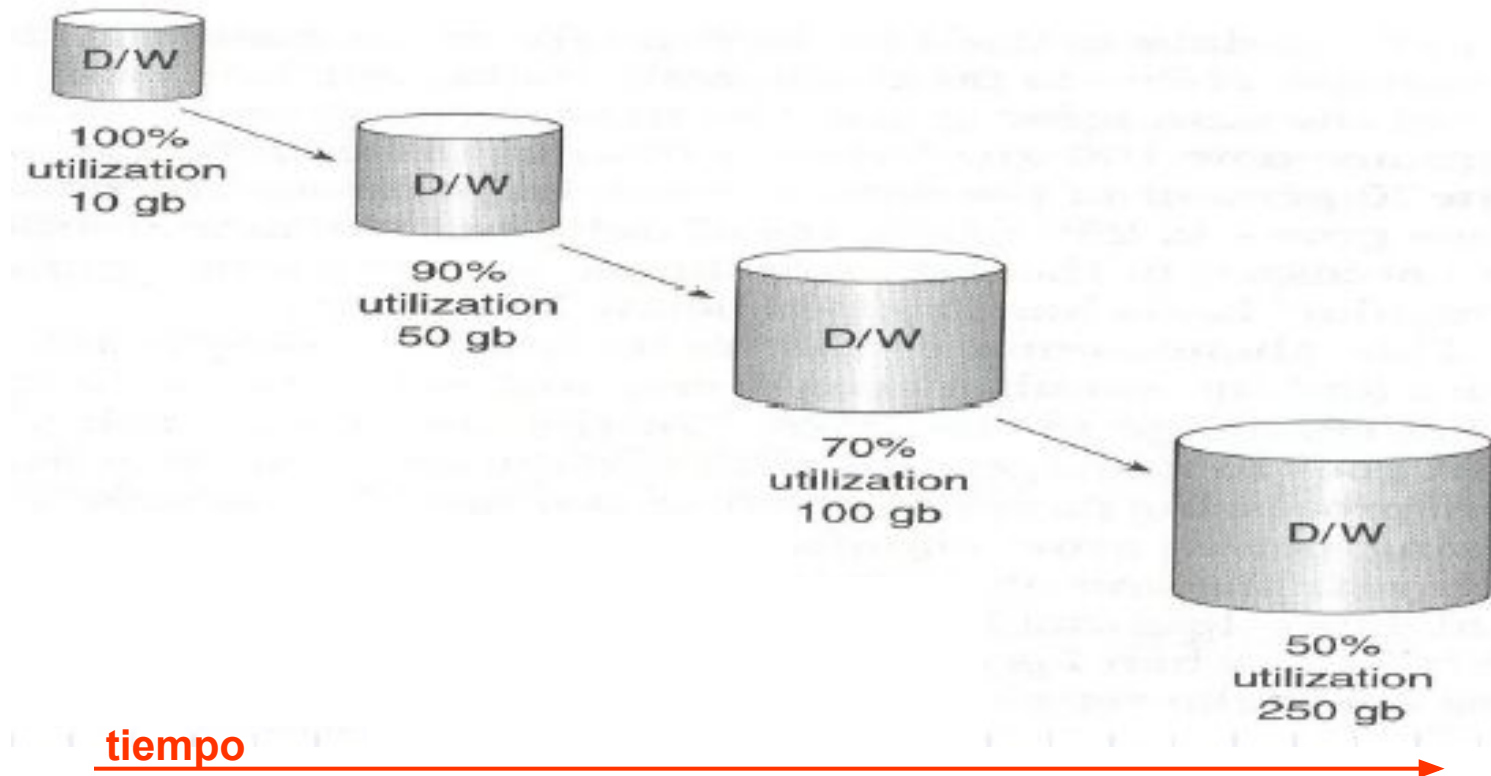
Ganancias de Fed-Ex



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

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

Preguntas Importantes

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

Referencias

- Extracto de material del curso Minería de Datos IIC2433 del profesor Karim Pichara