INTEGRACIÓN DE INFORMACIÓN DE FUENTES DISTINTAS

II CONGRESO INTERNACIONAL DE INGENIERÍA INFORMÁTICA

Analíticos

Por qué integrar datos?

¿Qué quiere saber el ejecutivo de hoy?

¿Quienes son mis clientes y que productos prefieren?

¿Cuál es el método de fiscalización más efectivo?

¿Cuales son los clientes con mayor potencial de irse con la competencia?

Necesita Información

¿Debo revisar todos los contenedores en aduanas?

¿Cuales son los clientes con el mayor y menor margen de utilidad?

¿Cual es el impacto del nuevo producto / servicio lanzado, sobre las utilidades?

Por qué integrar datos?

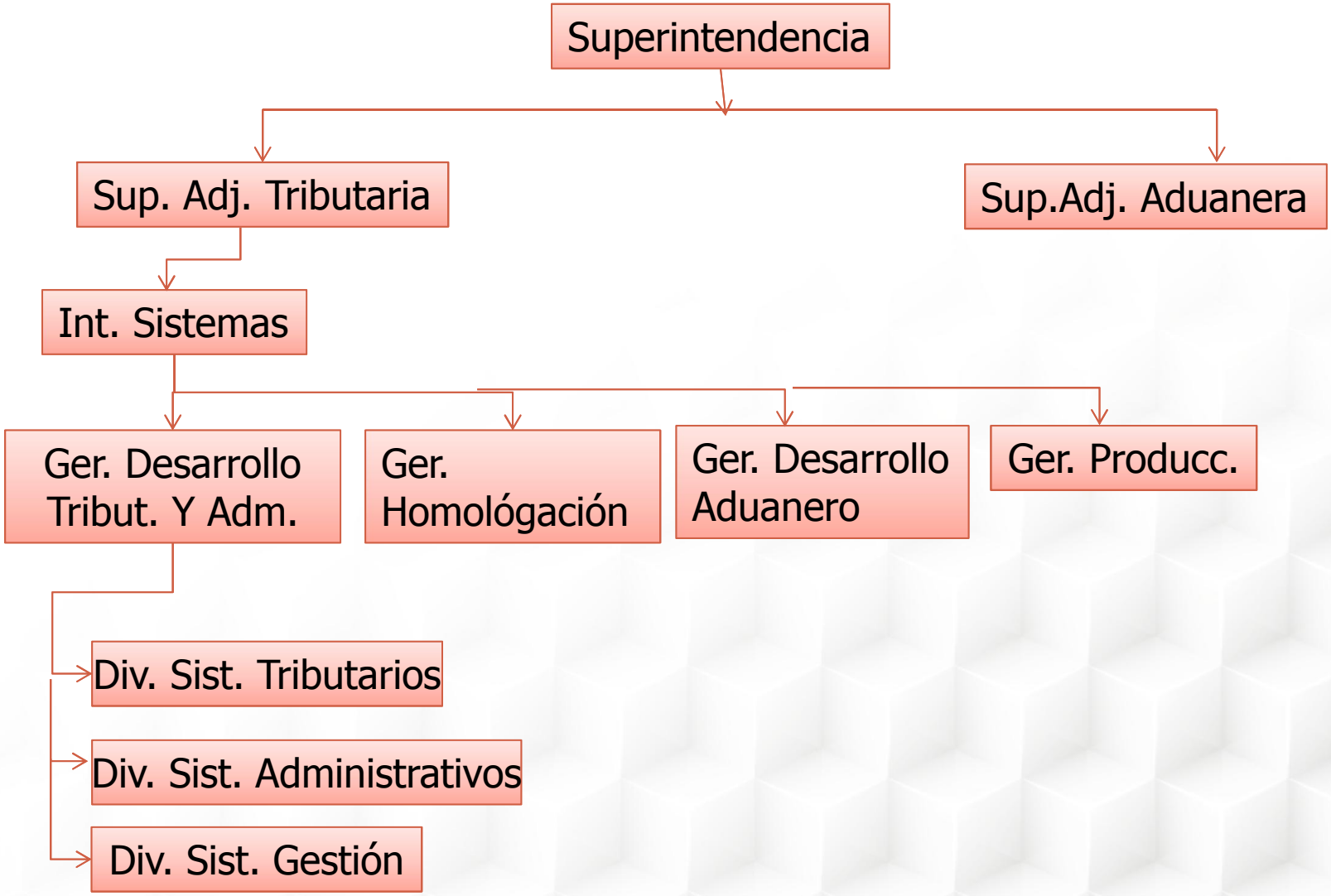
La realidad los frustra debido a que ...

- Existe muchos datos pero poca información
- Tienen problemas de acceso a las fuentes de datos
- Los datos están desintegrados.
- No pueden hacerse preguntas complejas del negocio.
- Incurren en alto costos para responder preguntas.
- No hay análisis histórico de información.

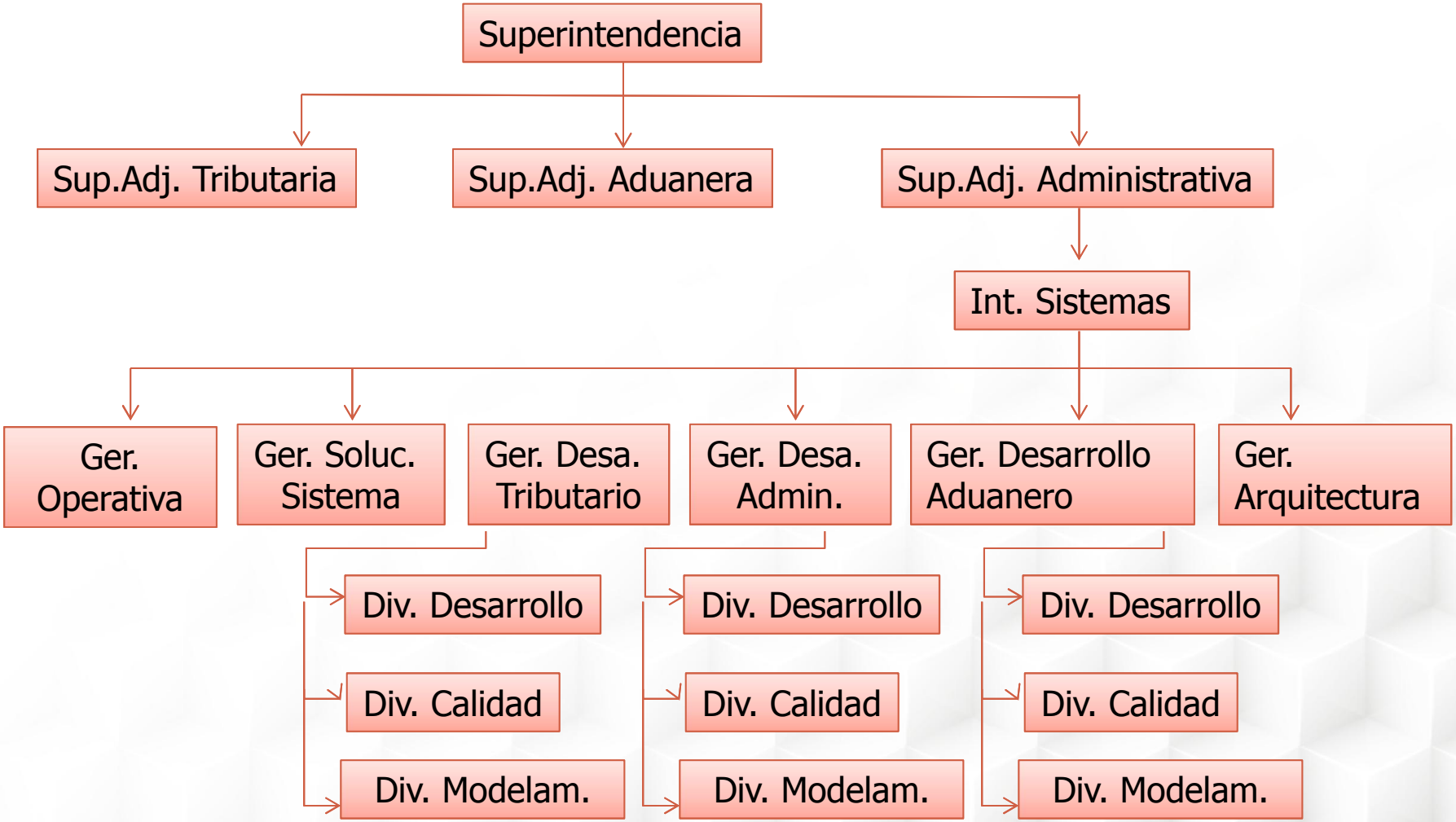
Se requiere una decisión institucional

Analíticos

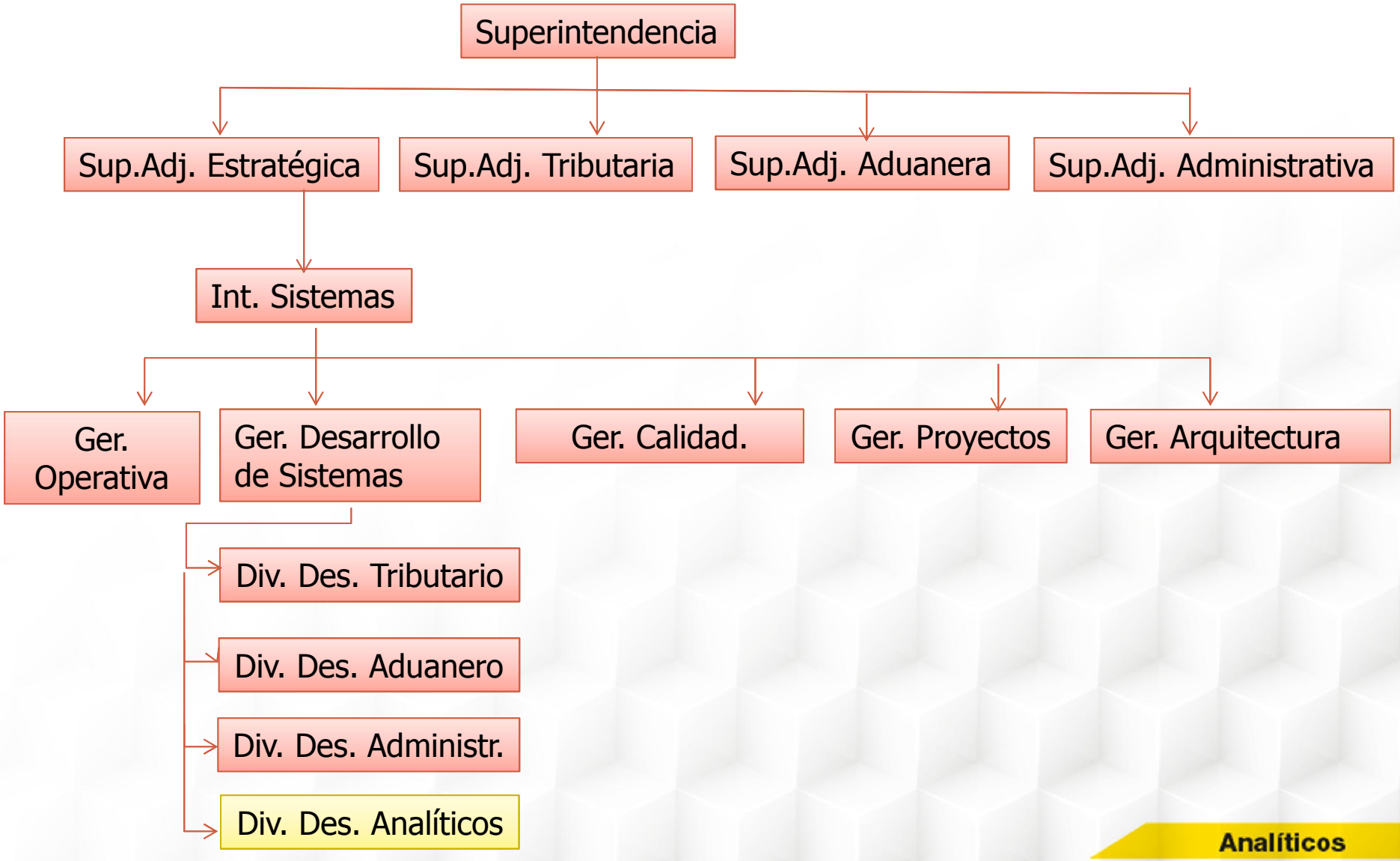
Cómo se adaptan las instituciones?



Cómo se adaptan las instituciones?



Cómo se adaptan las instituciones?



Analíticos

Cómo integrar los Datos?

El Data Warehouse

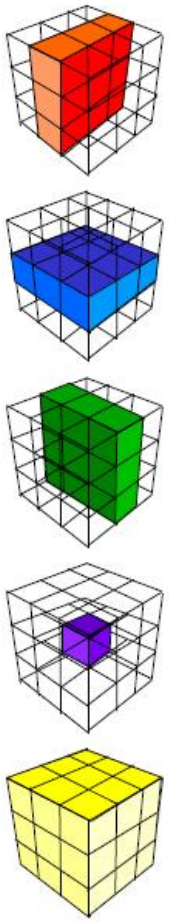
"El DWH es una **colección de datos integrada** en una Base de Datos, no volátil, orientada según un tema, diseñadas para soportar un Sistema de Soporte a las Decisiones (DSS), donde cada unidad de dato es relevante en algún momento del tiempo."(Bill Inmon)

"El más importante reto, que hoy en día, los administradores de negocios deben encarar es cómo integrar y maximizar sus datos para obtener ventajas competitivas" (Bob Sanguedolce CIO eBay Inc)

El Data Warehouse

Ventajas del DW

- Plataforma optimizada para analizar la información.
- Facilita la integración de datos dispersos en distintas Bases de Datos.
- Permite transformar la data en información.
- Aumenta la productividad en los procesos de consulta.

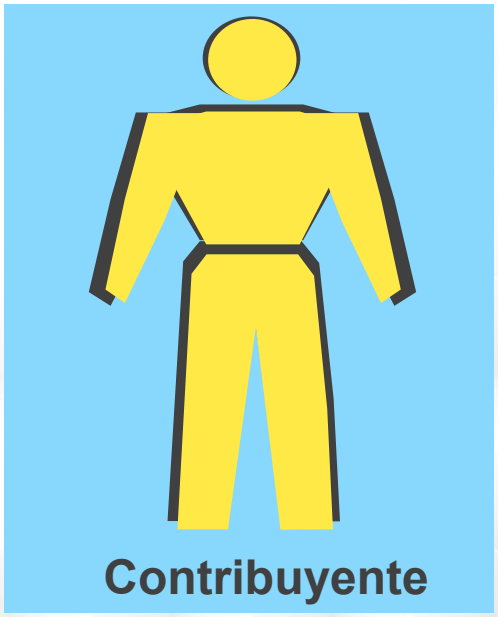


Características del DW

Integrada

La data es definida como única.

Aplicaciones Transaccionales



Data Warehouse

El aspecto más importante del DW es que la información encontrada al interior está siempre integrada.

Características del DW

Orientado a un Tema

La data es categorizada y almacenada por áreas de negocio en lugar de aplicaciones.

Aplicaciones Transaccionales

- Ahorros
- Vehículos
- Préstamos
- Inmuebles
- Acciones e Inversiones

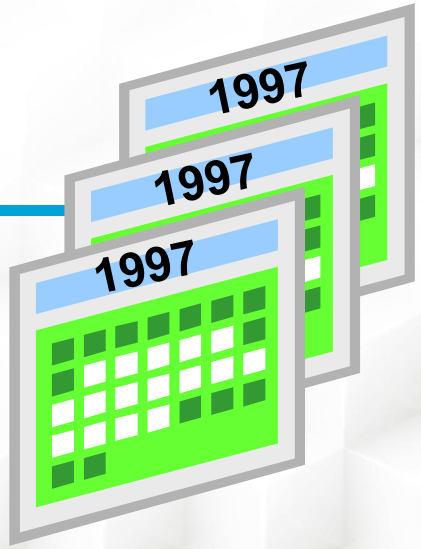
Tema del Data Warehouse

Incremento Patrimonial

Características del DW

Variante en el tiempo

La data es almacenada como serie de fotos asociadas al tiempo.



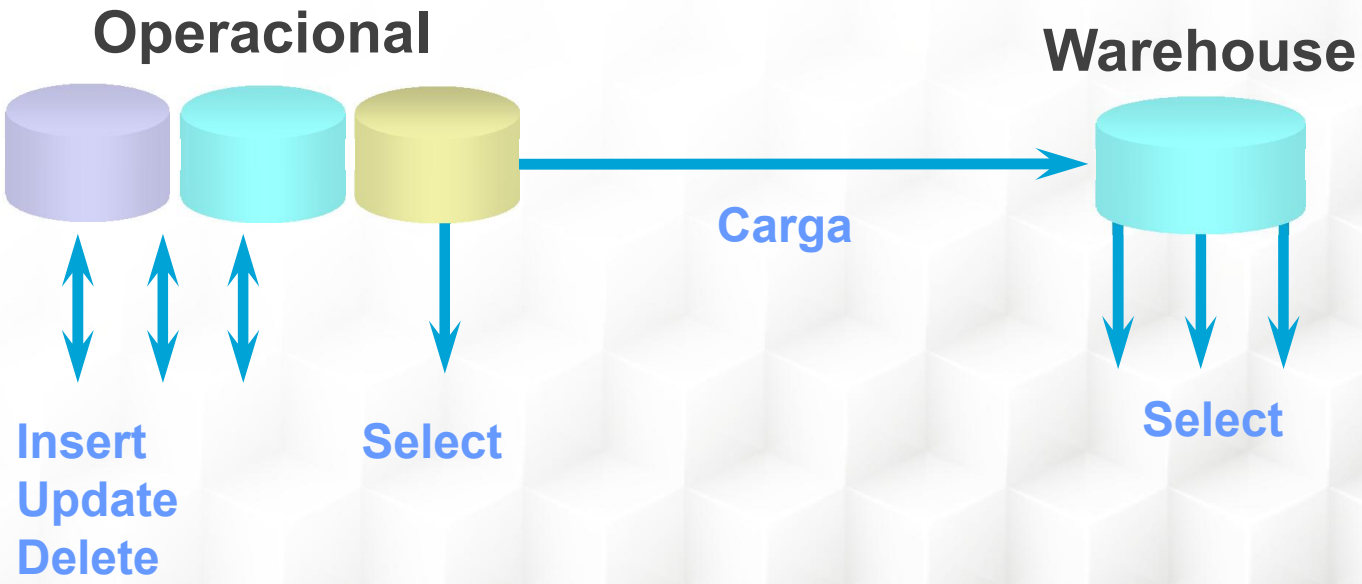
Periodo Tributario	Declaración IGV
01/97	Enero
02/97	Febrero
03/97	Marzo

Data Warehouse

Características del DW

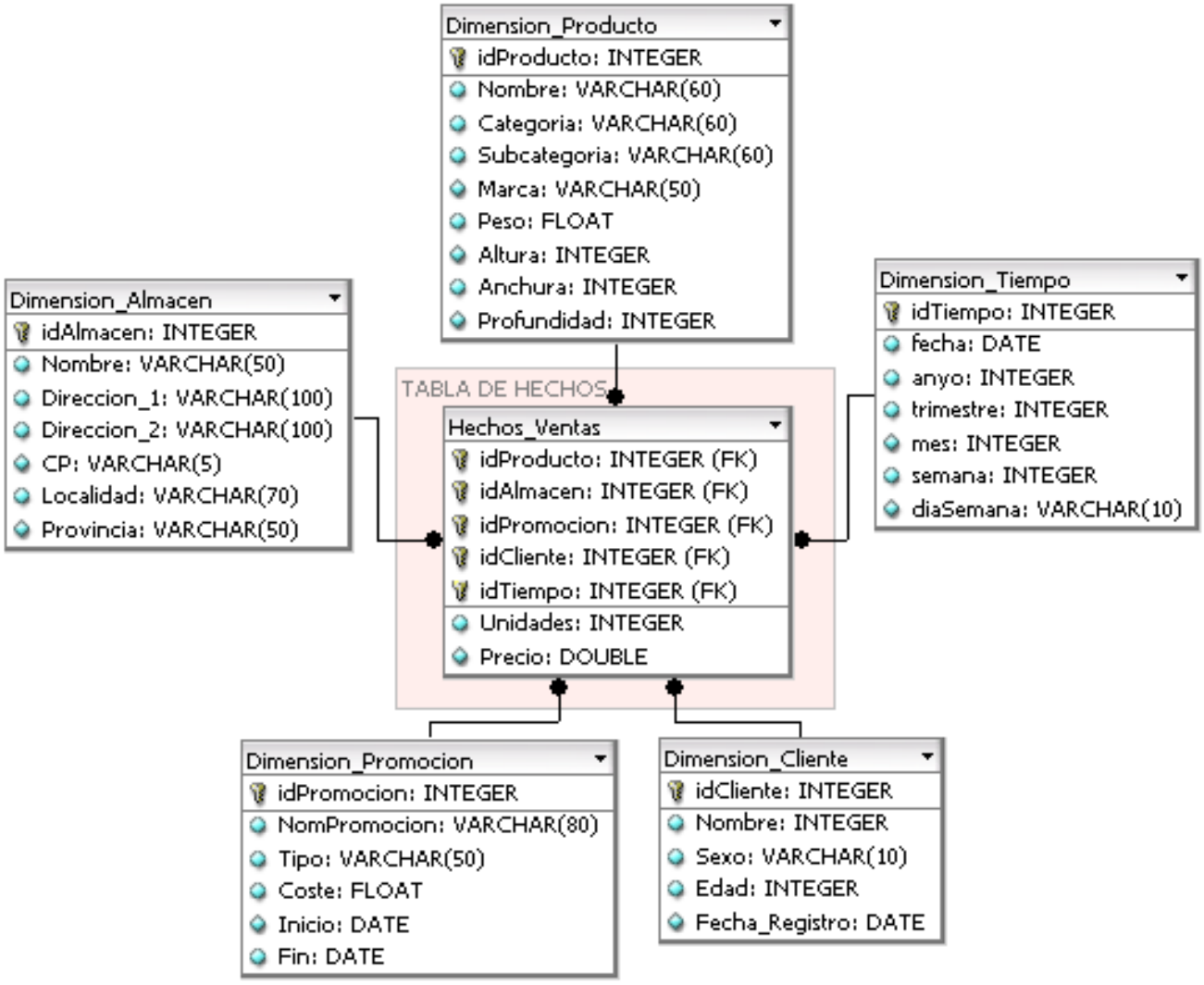
No volátil

La data en el DW típicamente No cambia.

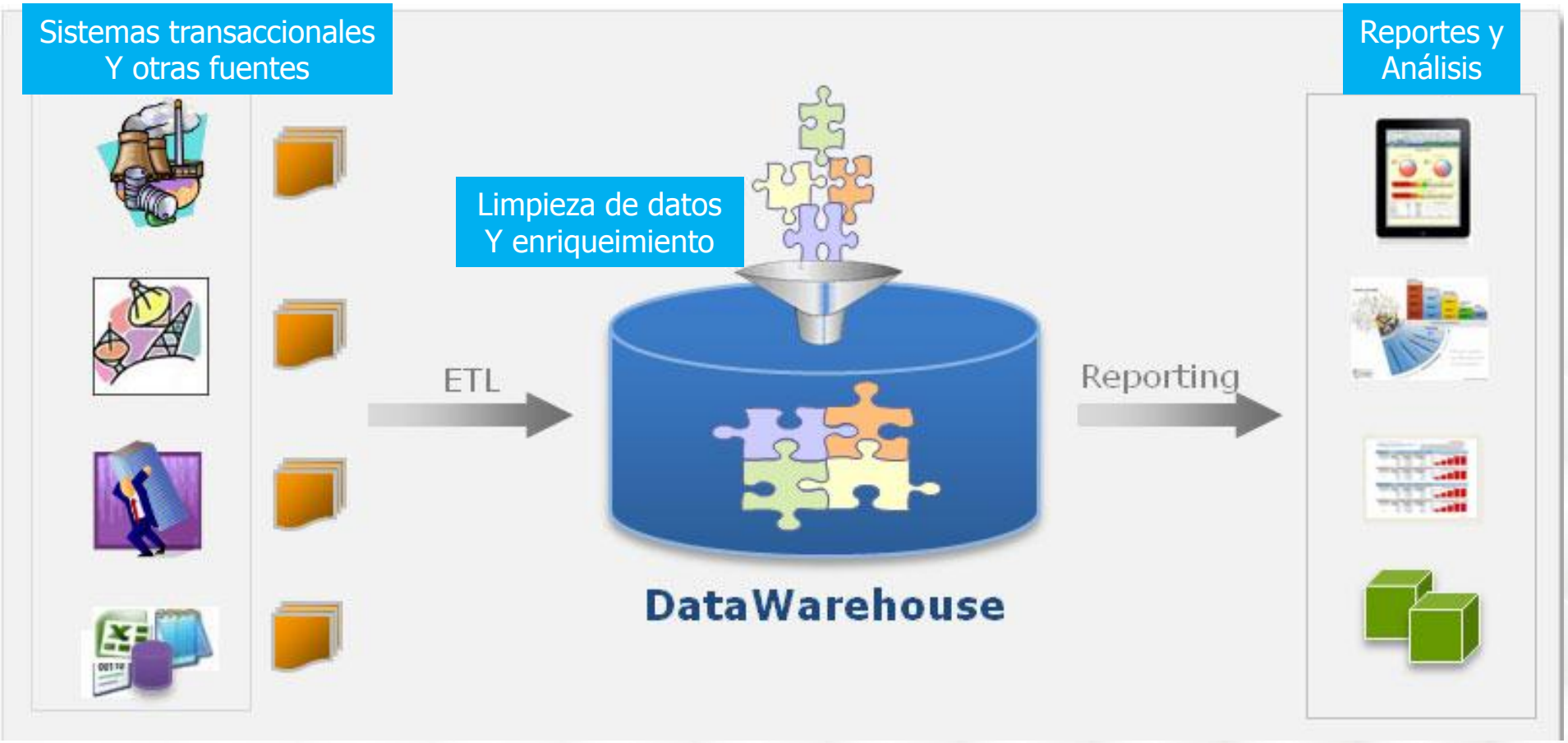


Diseño de DW

Modelo Estrella



El Data Warehouse

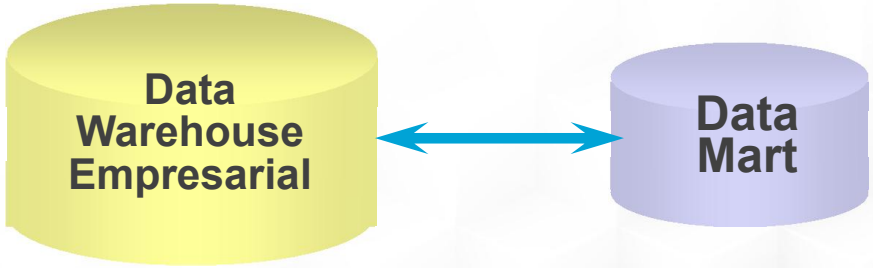


Analíticos

El Data Warehouse

DWE vs. Data Mart

- Implementación a escala grande
- Alcance de todo el negocio
- Datos desde todos los Subject Areas
- Niveles de datos atómicos
- Usuarios de toda la organización
- Punto de distribución de los Data Marts dependientes



- Subconjunto de un Data Warehouse Empresarial
- Optimizado para consultas específicas
- Altamente sumarizado
- Atiende funciones específicas del negocio
- Data Histórica
- Orientada a un grupo de usuarios

Analíticos

Las Fuentes de datos

Hoy en día tenemos un universo muy grande de fuentes de datos

- Nuestros sistemas transaccionales
- Sistemas de otras instituciones
- Páginas Web
- Redes sociales

¿Cuáles escojo?

Analíticos

Las Fuentes de datos

Tenemos que respondernos determinadas preguntas…..

- ¿Cuál está mejor alineada con la visión y estrategia del negocio?
- ¿Qué tipo de decisiones de negocio podré tomar?
- ¿Cuál es la más simple de implementar?
- ¿Qué debo hacer para obtenerla?

Pero; ahí no acaba el problema……

Las Fuentes de datos

Tipos de obtención de datos

- Propias
- Gratuitas
- A la venta
- A través de convenios

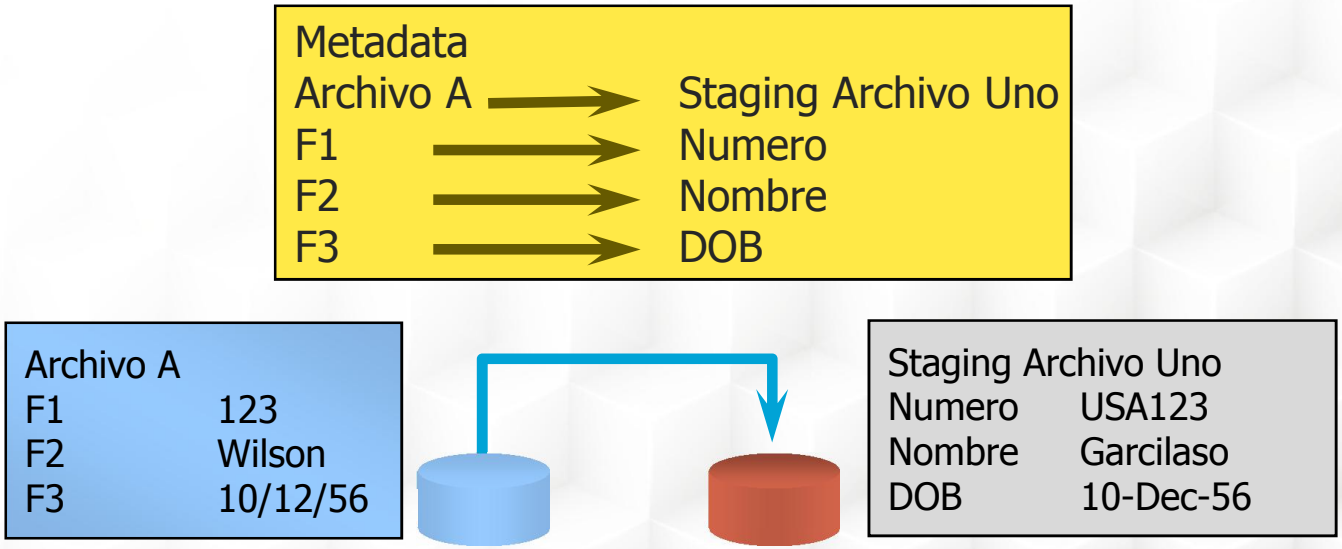
Problemas de las Fuentes de datos

- Calidad del dato
- Disponibilidad y acceso de los datos
- Criterios y estándares diversos utilizados
- Falta de instituciones rectoras de determinados datos

Por tanto, debemos integrar datos…

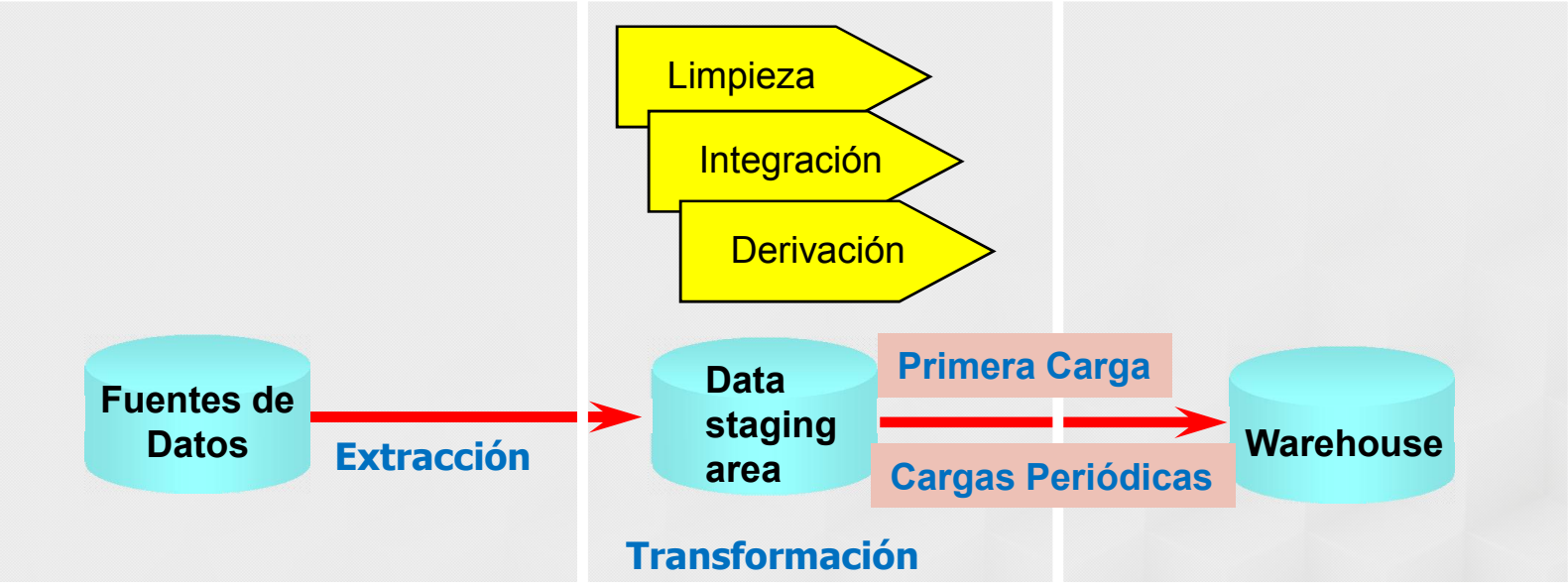
Mapeo de Datos

- Definir los atributos operacionales que se usarán
- Establecer mapeo de los requerimientos del negocio.
- Mapear requerimientos a las necesidades de datos
- Crear la matriz de Mapeo de los datos



Construcción del DW

Extracción, Transformación y Carga (ETL)



La transformación incluye:

- Limpieza: Eliminar anomalías en las Fuentes de datos
- Integración: Consolidar la información
- Derivación: Generar variables derivadas, agregadas, totales, etc..

Estamos listos para atender a los gerentes !

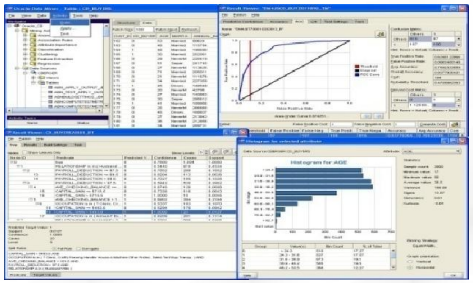
Explotación del DW

Análisis OLAP

- Navegación Ad-hoc
- Reportes y Dashboards

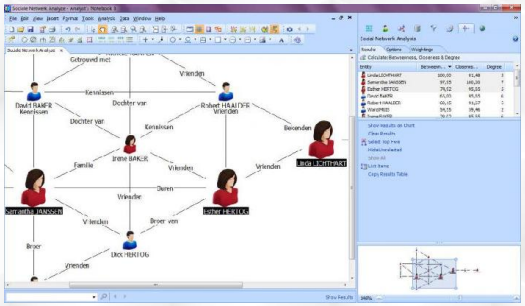
Minería de Datos

- Modelos Predictivos
- Modelos Descriptivos



Análisis de Vinculación

- Definición y Vinculación entre entidades
- Relaciones de grafos



Explotación del DW

¿Cuáles son los contenedores fraudulentos?

Sistema de Minería de Datos predictivos aplicando algoritmos de Redes Neuronales

¿Cómo los agrupo para fiscalizar?

Sistema de Minería de Datos descriptivo aplicando algoritmos de Arboles de decisiones

Monitoreo del DW

- Del proceso de carga
- Del uso del Warehouse
- De la calidad de los datos
- De la performance del sistema

- Cual es el tamaño del DW ahora?
- Cual es el crecimiento ocurrido en el DW?
- Qué usuarios acceden?
- Que perfiles existen?
- Cual es la calidad de los datos?
- Qué controles de seguridad y control existen?

Conclusiones

Es necesario darle una importancia estratégica al análisis de la información dentro de las empresas.

Desarrollar la infraestructura de tecnología que soporte el análisis de la información, dando vital importancia a la integración de los datos.

Preparar al personal técnico y del negocio para afrontar los nuevos retos de análisis de la información.

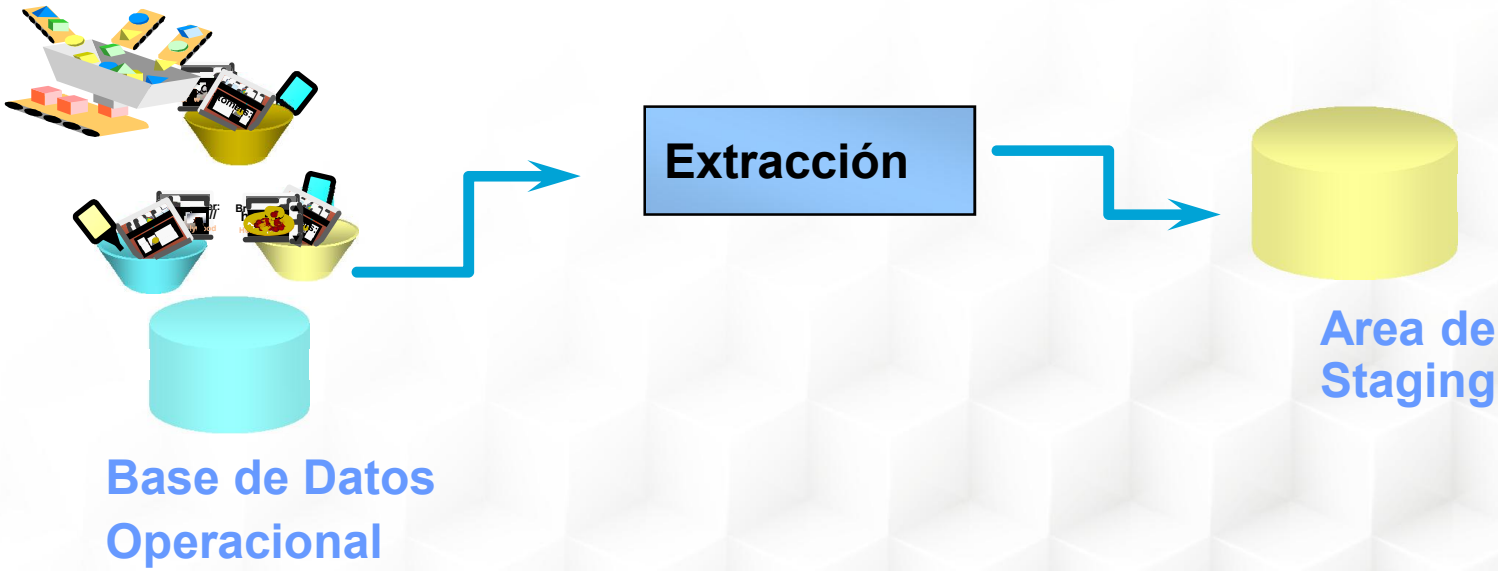
GRACIAS

ameza@sunat.gob.pe

Extracción de los datos

Escoger los datos a extraer desde las múltiples fuentes

- Múltiples fuentes de datos agregan complejidad
- Empezar simple



Limpieza de los Datos

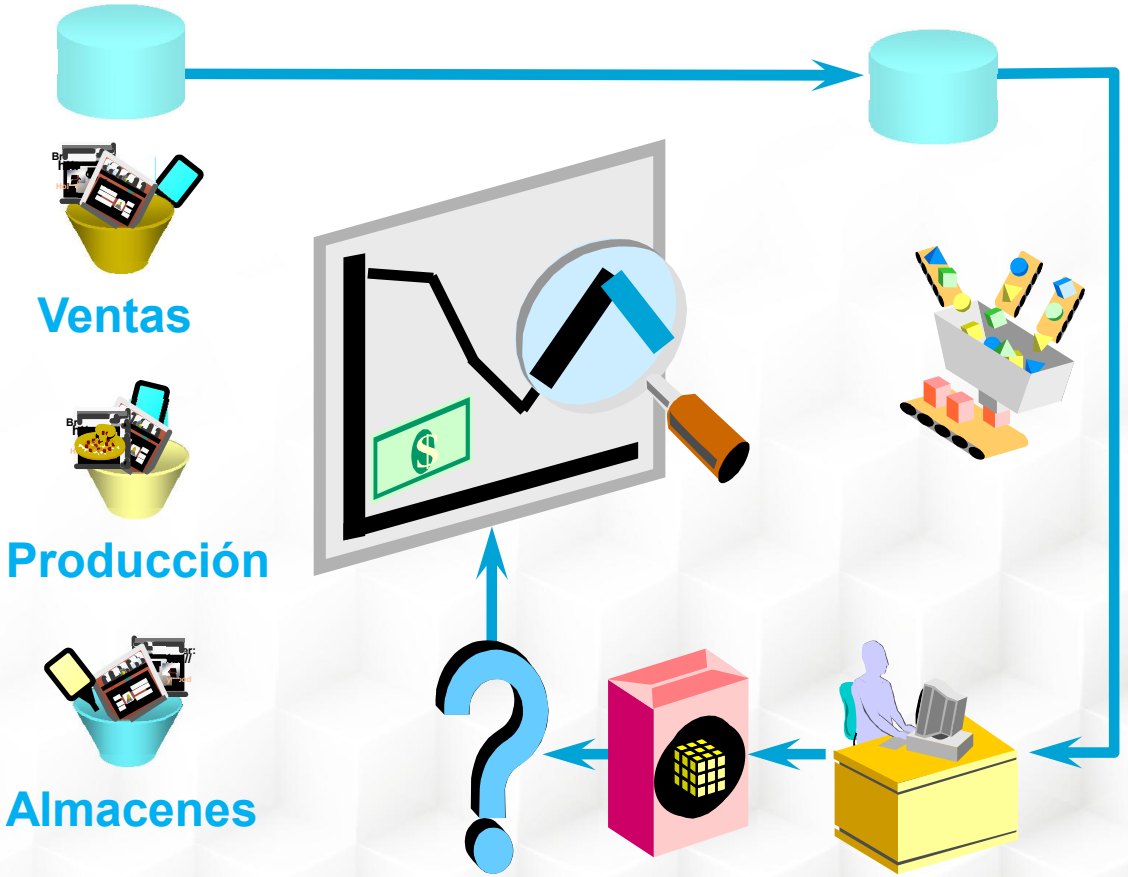
Contar con datos de calidad aceptable para el negocio

- **Determinar la calidad de los datos**
  - Evaluar el dato: Exactitud, completitud, consistente, único y oportuno.
  - Reglas de negocio para identificar las inconsistencias
  - Establecer métricas de calidad mínima y medir.
- **Limpiar los datos inconsistentes**
  - En el Warehouse
  - En las fuentes
    - Identificar y corregir la causa de los defectos
    - Programar limpiezas periódicas de las fuentes de datos

La Calidad de los Datos

Características de la calidad de datos

- Exactitud
- Completo
- Consistente
- Único
- Oportuno



La Calidad de los Datos

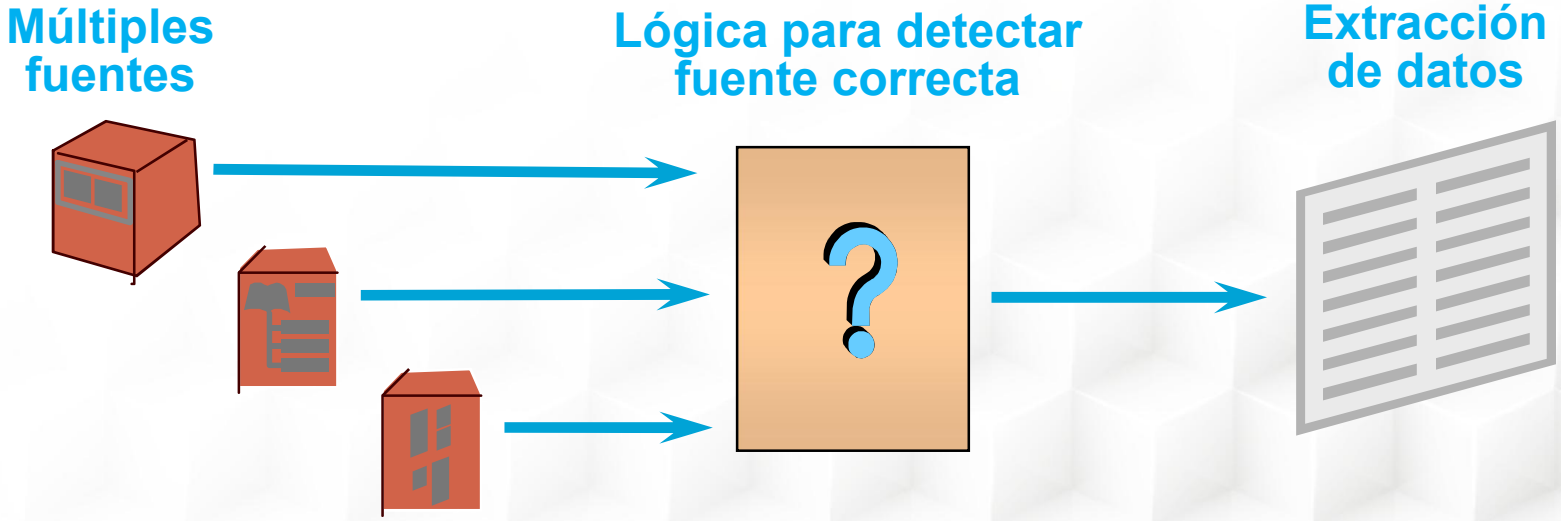
Mejora de la Calidad

- Crear un programa de calidad de datos.
- Documentar las fuentes.
- Diseñar claramente los procesos de limpieza.
- Definir Estándares de Calidad
- Las reglas de limpieza las define el negocio.
- Considerar mejoras en los sistemas transaccionales.

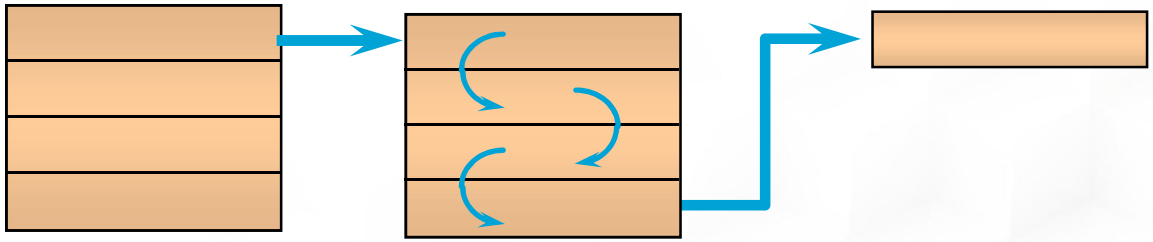
Integración de los datos

Problemas de Múltiples fuentes

- Múltiples fuentes de datos agregan complejidad
- Empezar simple



Transformando Datos: Problemas y Soluciones



Código de producto = 12M65431345

Código país	Territorio venta	Número Producto	Código Vendedor

Problemas de formato

Múltiple codificación

m , f

1 , 0 → m, f

male, female

Múltiples estándares

cm → cm

pulgadas

DD/MM/YY → DD-Mon-YY

MM/DD/YY

S/. 1,000 → USD 600

Euros 9,990

Múltiples idiomas

 → ACME Co.

Problemas de Nombres y Direcciones

Formato de campo simple

Sr. J. Sanchez, Las Gardenias 415 interior 3 Surco, Lima

Formato múltiple de campos

Nombre	Sr. J. Sanchez
Calle	Las Gardenias
Provincia	Lima
Distrito	Surco
Número	415 interior 3

Integración de los datos

Integración de personas

DNI
Pasaporte
Carnet de Extranjería
RUC
Brevete

Jose Antonio Flores Díaz
Jose Flores Díaz
Flores Díaz Jose
J. A. Flores Díaz

El CIC

Integración de direcciones

La Georeferenciación:

Av. Wilson 1402, Lima →

Av. Garcilazo de la Vega 1402, Lima →

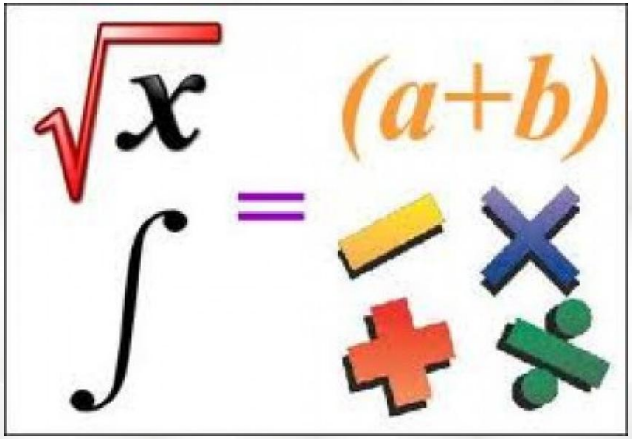
Inca Garcilazo de la Vega 1402, Lima →

Coordenadas: (X, Y, Z)

Datos Derivados

Nuevas variables

- Datos Agregados
- Balances acumulativos
- Totales y subtotales
- Totales de Dimensión
- Ratios
- Variables calculadas



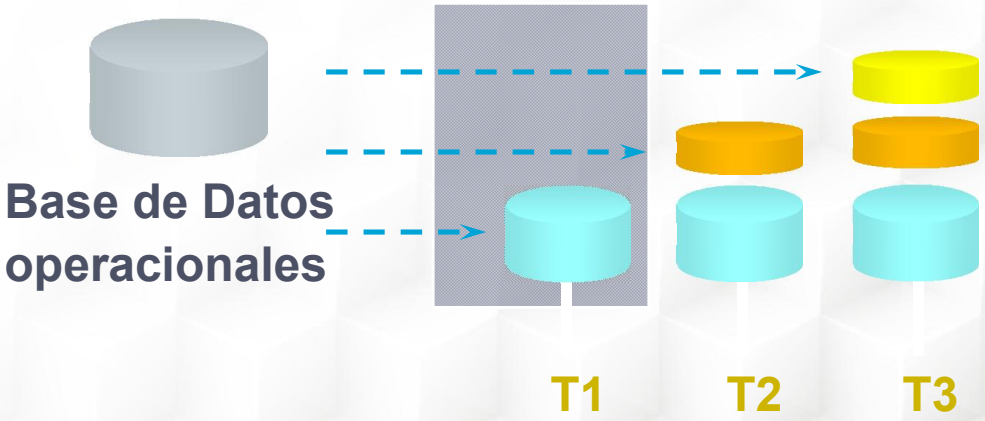
Regresar

Analíticos

Carga del DW

Primera Carga

- Evento simple que puebla el Data Warehouse con la data histórica
- Envuelve grandes volúmenes de dato
- Envuelve una gran cantidad de procesos antes de la carga

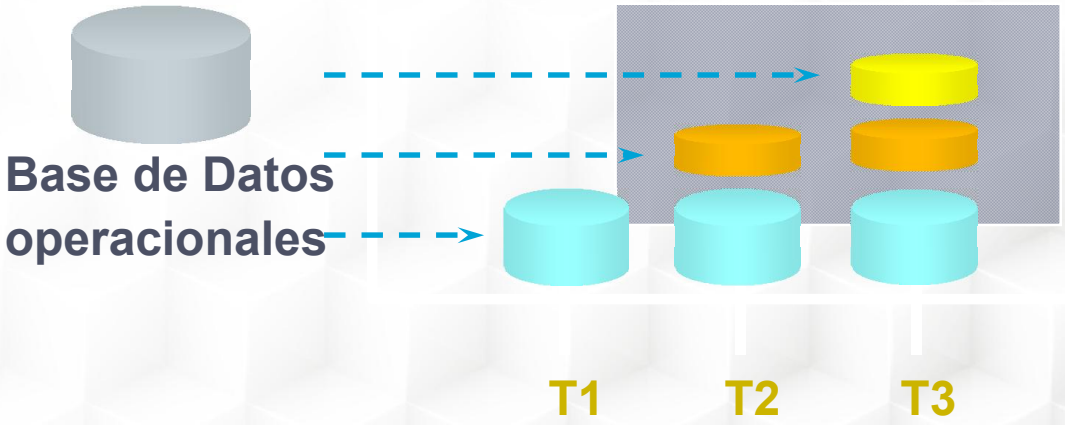


Regresar

Carga del DW

Refrescos

- Ejecución de acuerdo a los ciclos de negocio
- Tareas mas cortas
- Mucho menos data a cargar que la primera vez
- Procesos ETT menos complejos



Carga del DW

Capturar la data que cambia para el refresco

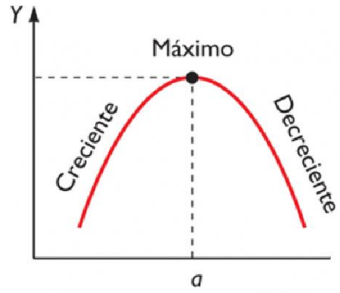
- Capturar nuevos datos de los hechos
- Capturar cambios de los datos de las dimensiones.
- Determinar métodos de captura:
  - Reemplazo de toda la data
  - Comparación de las instancias de las BDs
  - Marcas de tiempo (Time stamping)
  - Triggers en la Base de Datos
  - Log en las Base de Datos
- Considerar técnicas híbridas

Regresar

¿Cuántos contenedores deben ser revisados?

Facilitar comercio vs. Control

- No detección del fraude
- Perdidas de impuestos



- Costos operativos elevados
- Perjuicio y molestia al importador

Revisar los contenedores con fraude…

¿Cuáles son los contenedores fraudulentos?

Analíticos

¿Qué método de fiscalización usar?

¿Un único método para todos? **NO**

Analíticos

¿Cómo fiscalizar a los contribuyentes?

Clasificarlos en grupos homogéneos

Diferentes estrategias de fiscalización por grupos homogéneos de contribuyentes

¿Cómo los agrupo para fiscalizar?

Analíticos