DATA WAREHOUSE & BUSINESS INTELLIGENCE

Ing. Julio Paciello
juliopaciello@cds.com.py

Contenido

Data Warehouse & Business Intelligence

- Business Intelligence
- OLAP vs OLTP
- ODS, Staging
- DWH, Data Marts
- Software de BI y visualizaciones
- Conceptos de Data Mining

Contenido

Diseño multidimensional

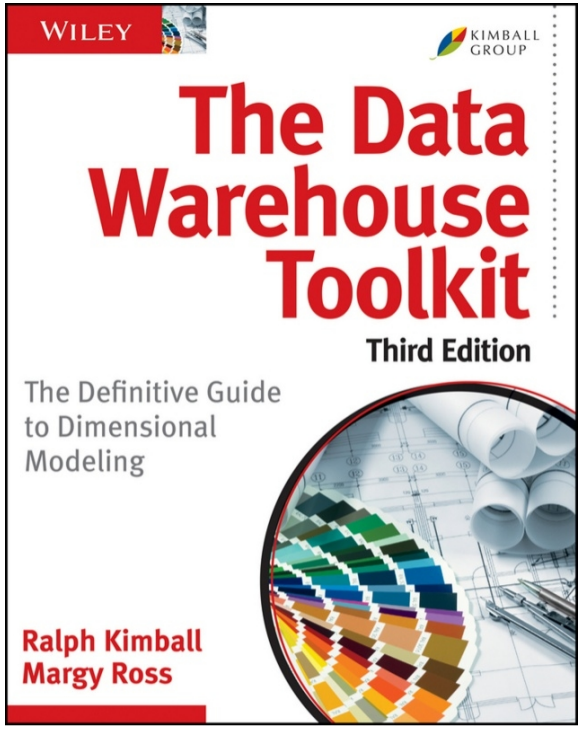
- Hechos y dimensiones
- Modelo en estrella y copo de nieve
- Tipos de Dimensiones
- Estrategias: *Slowly changing, Rapidly changing dimensions*
- Tablas de hechos: *Transactional*, *Periodic Snapshot*, *Accumulating Snapshot*
- Análisis de casos de estudio

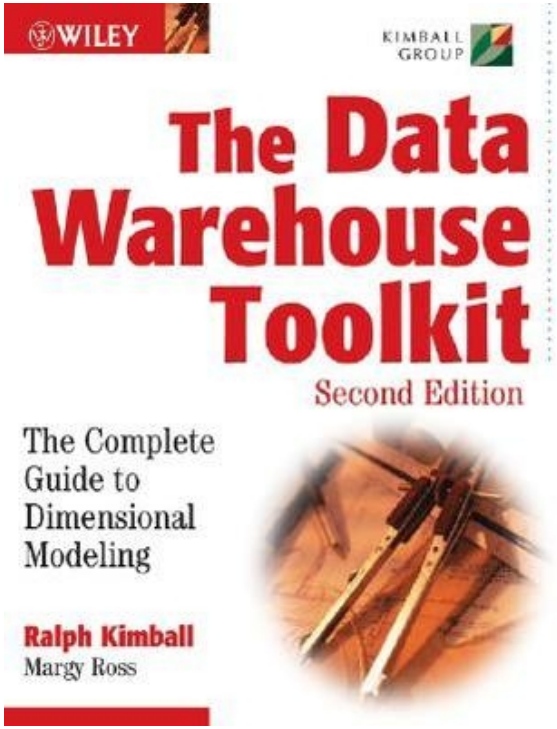
Caso Práctico 1: Datos del PGN

- Staging Area: Pentaho Data Integration
- Visualizaciones: Power BI

Bibliografía

Ralph Kimball et al., The Data Warehouse Toolkit





Mi experiencia previa

Personal

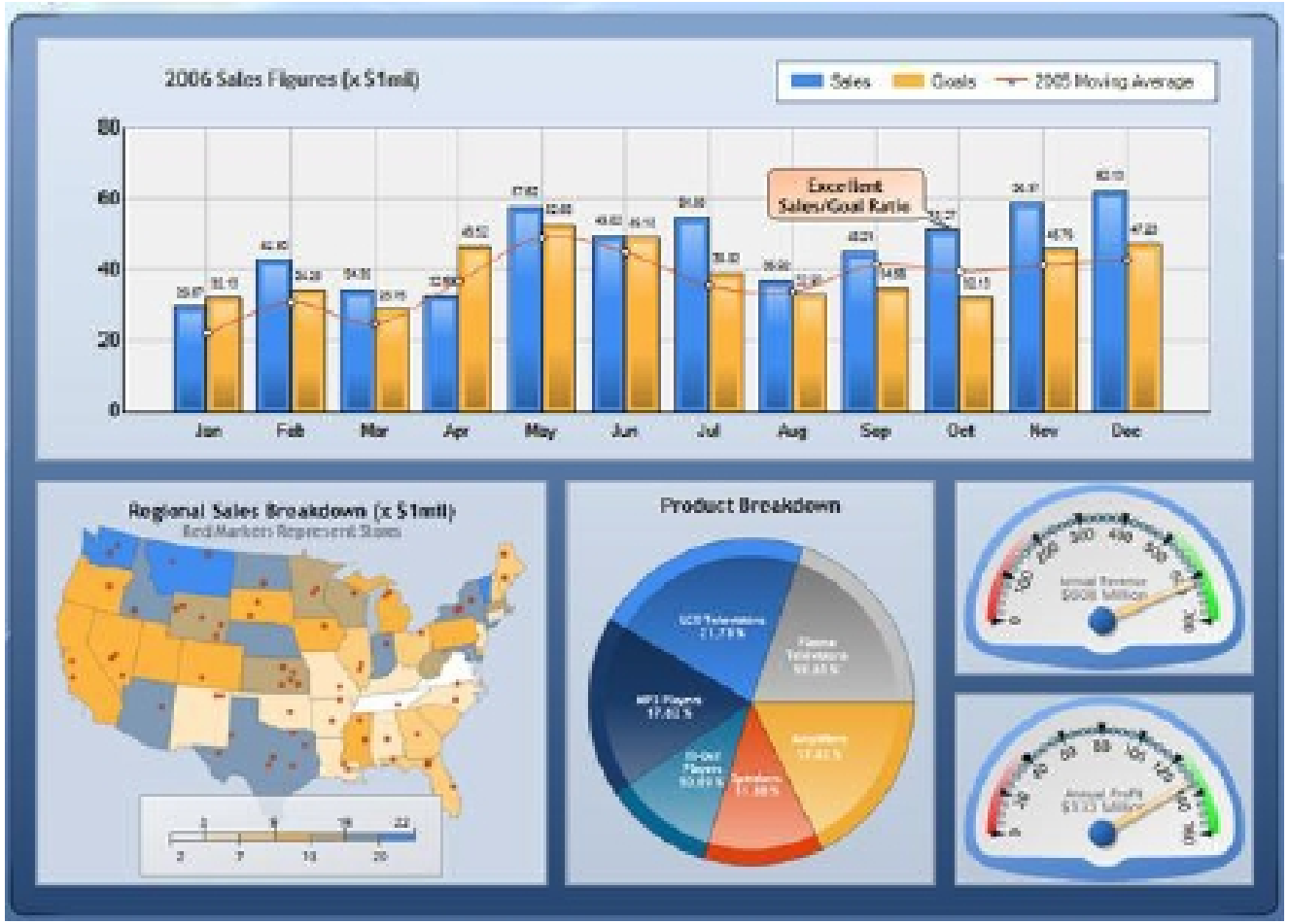
financiera el comercio

¡El poder de hacer!

VUE

VENTANILLA ÚNICA DE EXPORTACIÓN

BI convencional



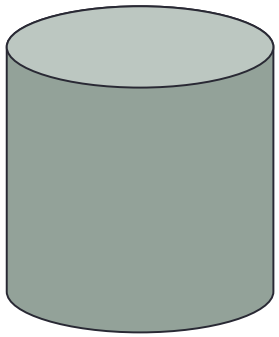
Business Intelligence

- Infraestructura tecnológica para obtener la máxima información de los datos disponibles para la mejora continua de los procesos de negocio

- Sistemas de BI
  - OLAP (Online Analytical Processing)
  - CRM (Customer Relationship Management)
  - GIS (Geographic Information System)
  - KDD (Knowledge Discovery in Databases)
  - …

Business Intelligence

Aggregate Data



Database, Data Mart, Data Warehouse, ETL Tools, Integration Tools

Present Data

Reporting Tools, Dashboards, Static Reports, Mobile Reporting, OLAP Cubes

Enrich Data

Add Context to Create Information, Descriptive Statistics, Benchmarks, Variance to Plan or LY

Inform a Decision

Decisions are Fact-based and Data-driven

Amazon.com and NetFlix

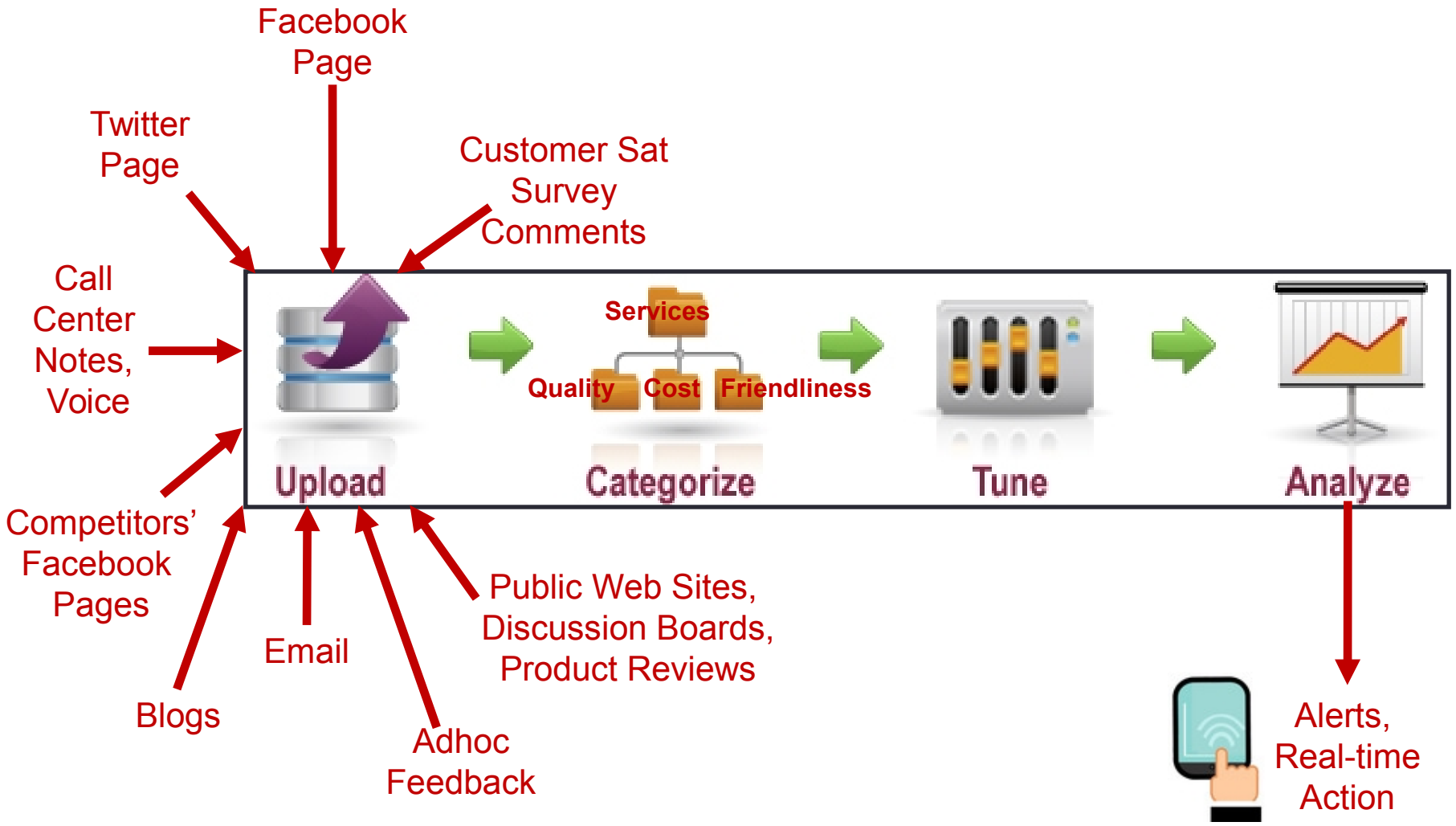
Collaborative Filtering

Intentan predecir otros ítems que un cliente desea comprar en base a lo que hay en sus shopping cart y wish lists y el comportamiento de compra de otros clientes

House of cards

Referencias:
https://www.cio.com/article/3207670/big-data/how-netflix-built-a-house-of-cards-with-big-data.html
https://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html

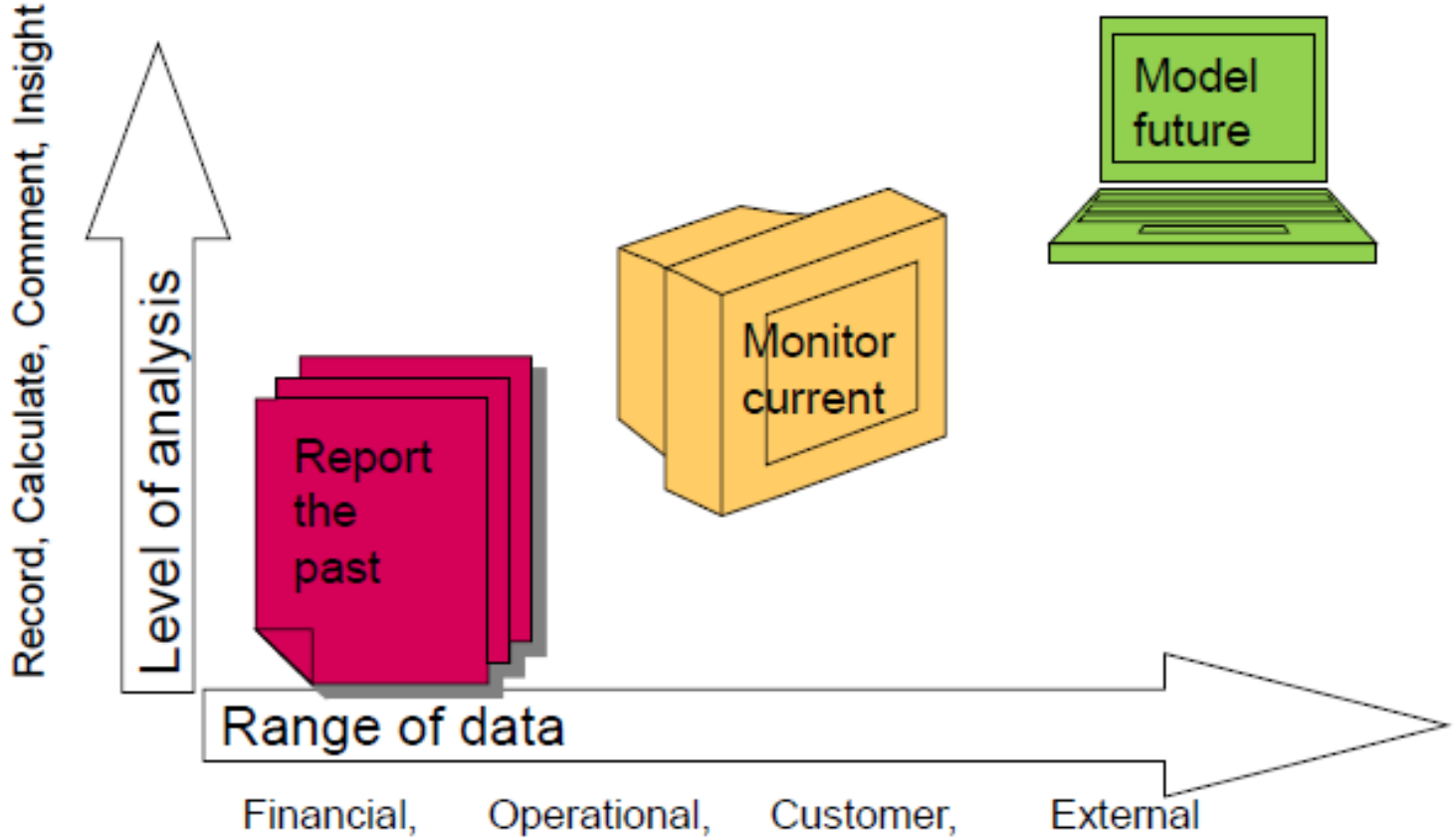
Unstructured Text Processing



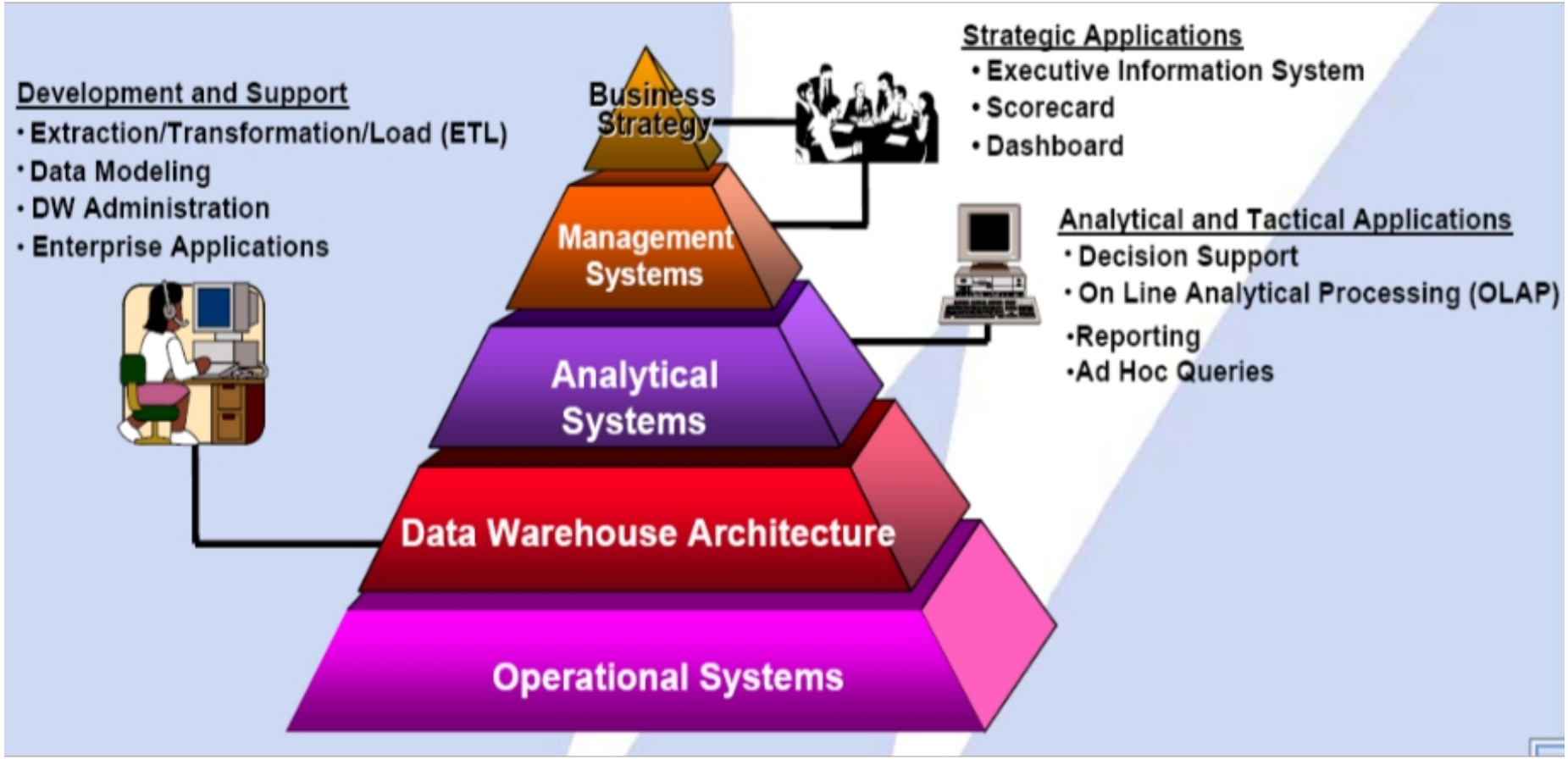
Unstructured Text Processing

https://www.predictiveanalyticstoday.com/lexalytics/

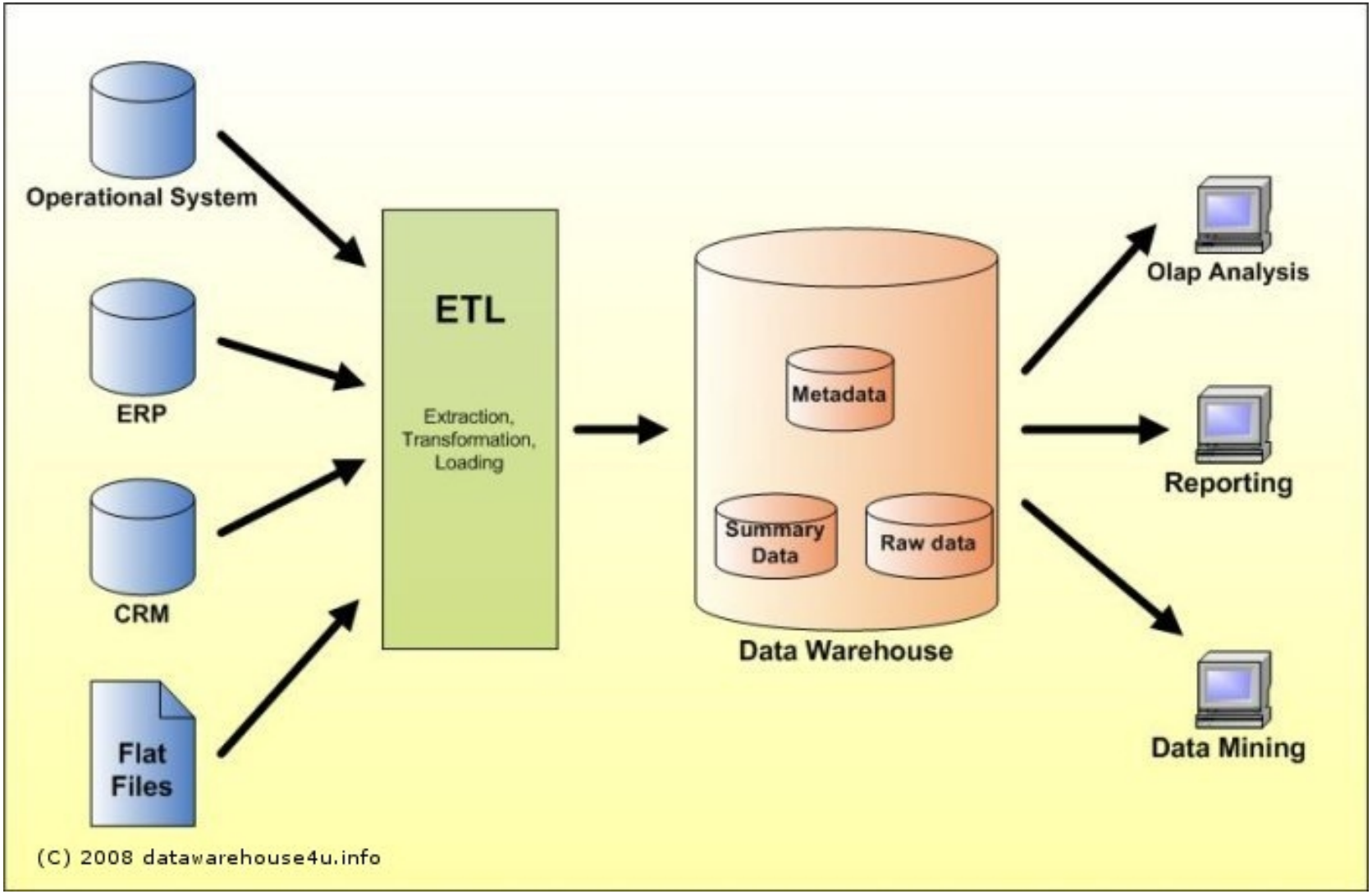
Niveles de análisis (madurez)



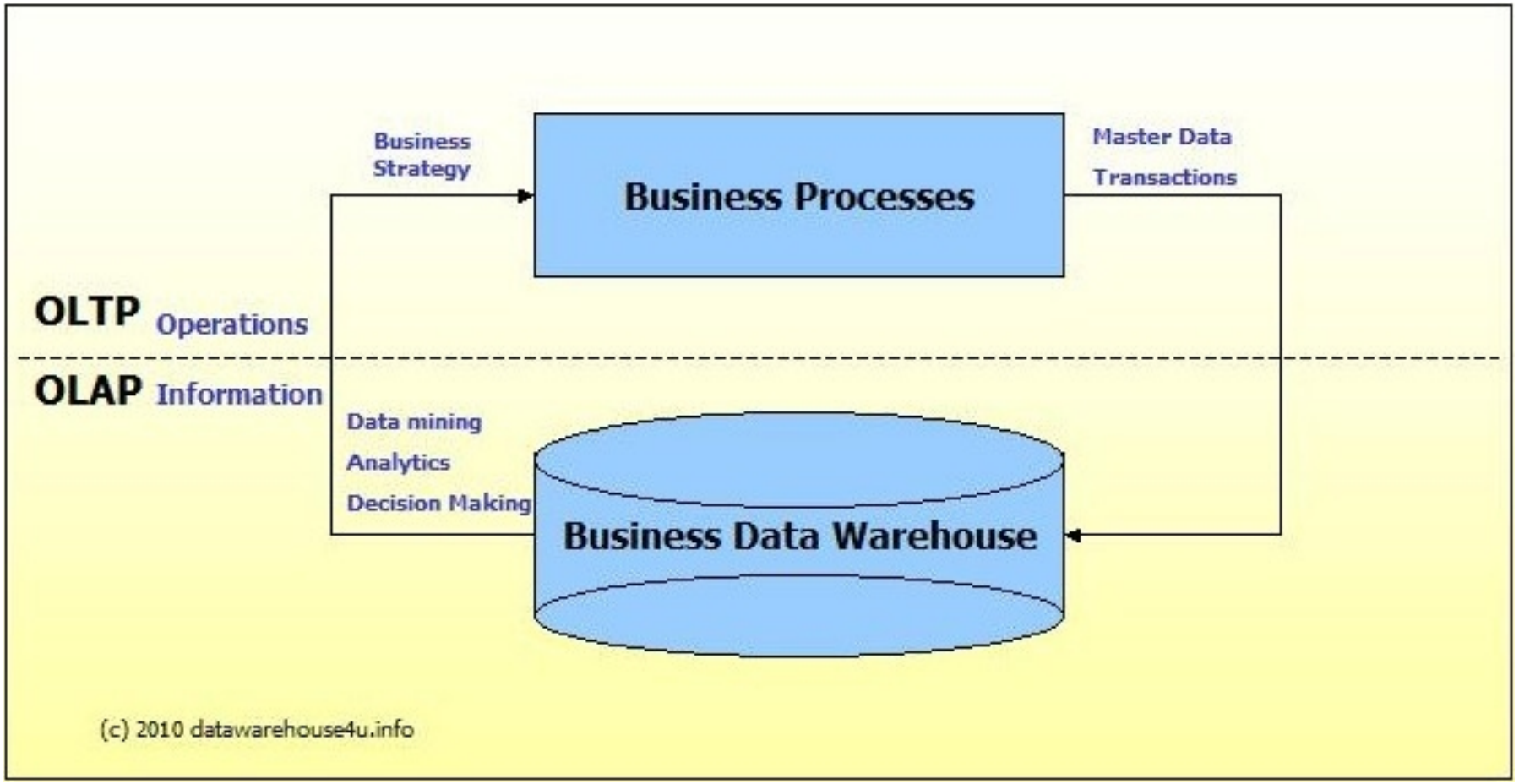
Jerarquía de BI



Arquitectura DWH



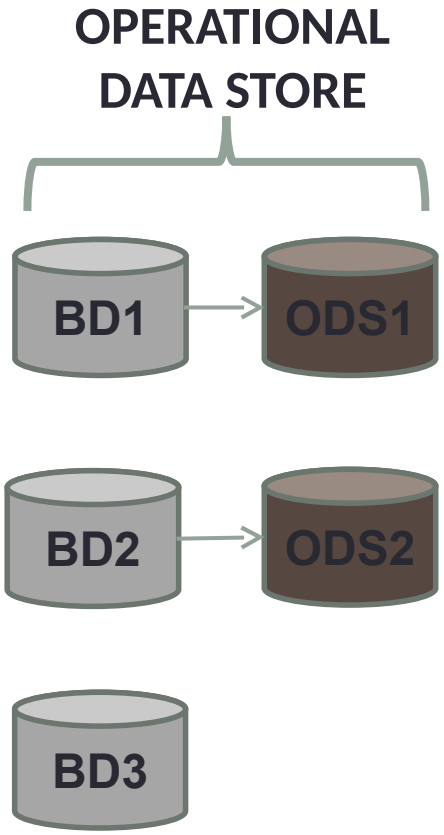
OLAP vs OLTP


Business Strategy
Business Processes
Master Data
Transactions
OLTP Operations
OLAP Information
Data mining
Analytics
Decision Making
Business Data Warehouse
(c) 2010 datawarehouse4u.info

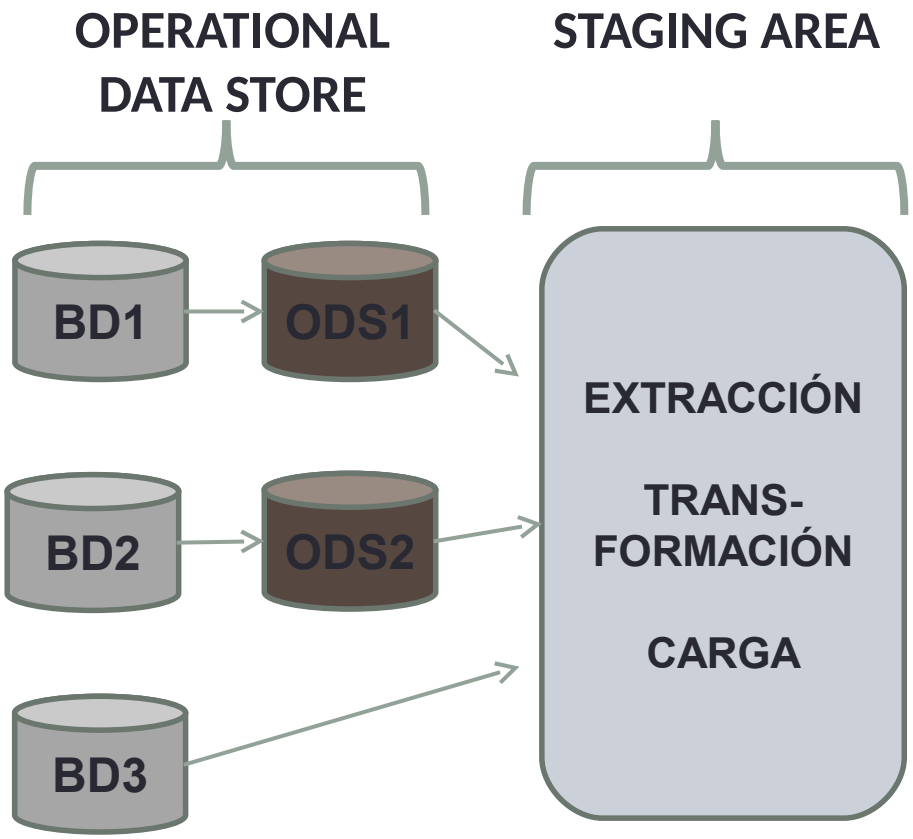
OLAP vs OLTP (2)

CARACTERISTICA	OnLine Transaction Processing OLTP	OnLine Analytical Processing OLAP
ORIGEN DE DATOS	***Bases de datos transaccionales***	Consumen ***múltiples orígenes OLTP***
PROPOSITO	***Operativa*** de procesos de negocio	***Toma de decisiones*** estratégicas
REPRESENTACION DE DATOS	Típicamente ***transacciones***	***Cubos*** multi-dimensionales
INSERTS & UPDATES	Frecuentes, ***por cada transacción***	Periódico, ***por lotes***
CONSULTAS	***Simples***, involucran pocos registros	***Complejas***, involucran agregaciones
VELOCIDAD AL PROCESAR	***Rápida***	***Lenta***, gran volumen de datos
ALMACENAMIENTO	***Controlable***, archivando históricos	***Grande***, acumula los datos
DISEÑO	Altamente ***normalizado***	***Desnormalizado*** esquema en estrella
BACKUP & RECOVERY	***Crítico***, típicamente diario	***No crítico***, puede recargarse con los datos OLTP

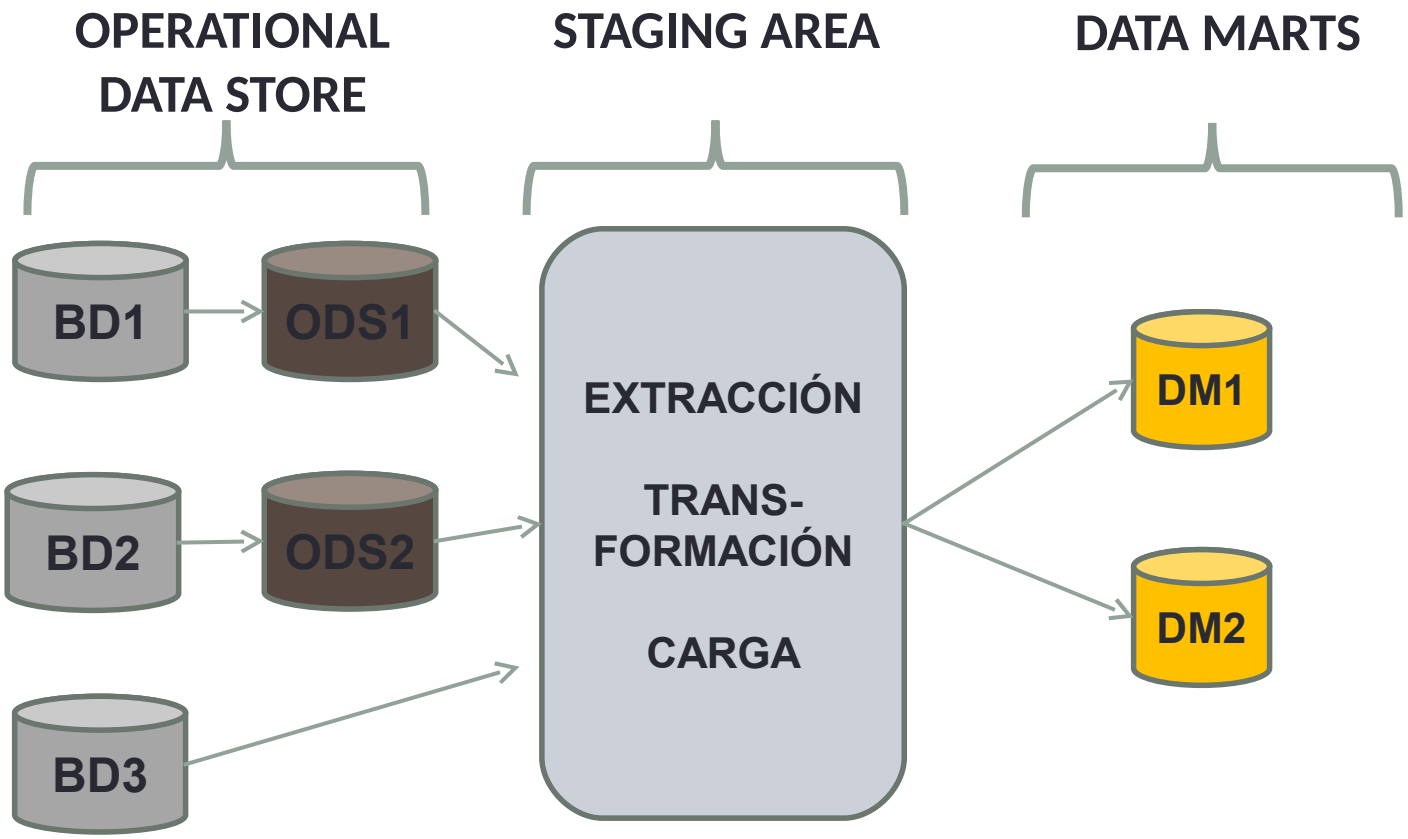
ODS, Staging, Data Marts



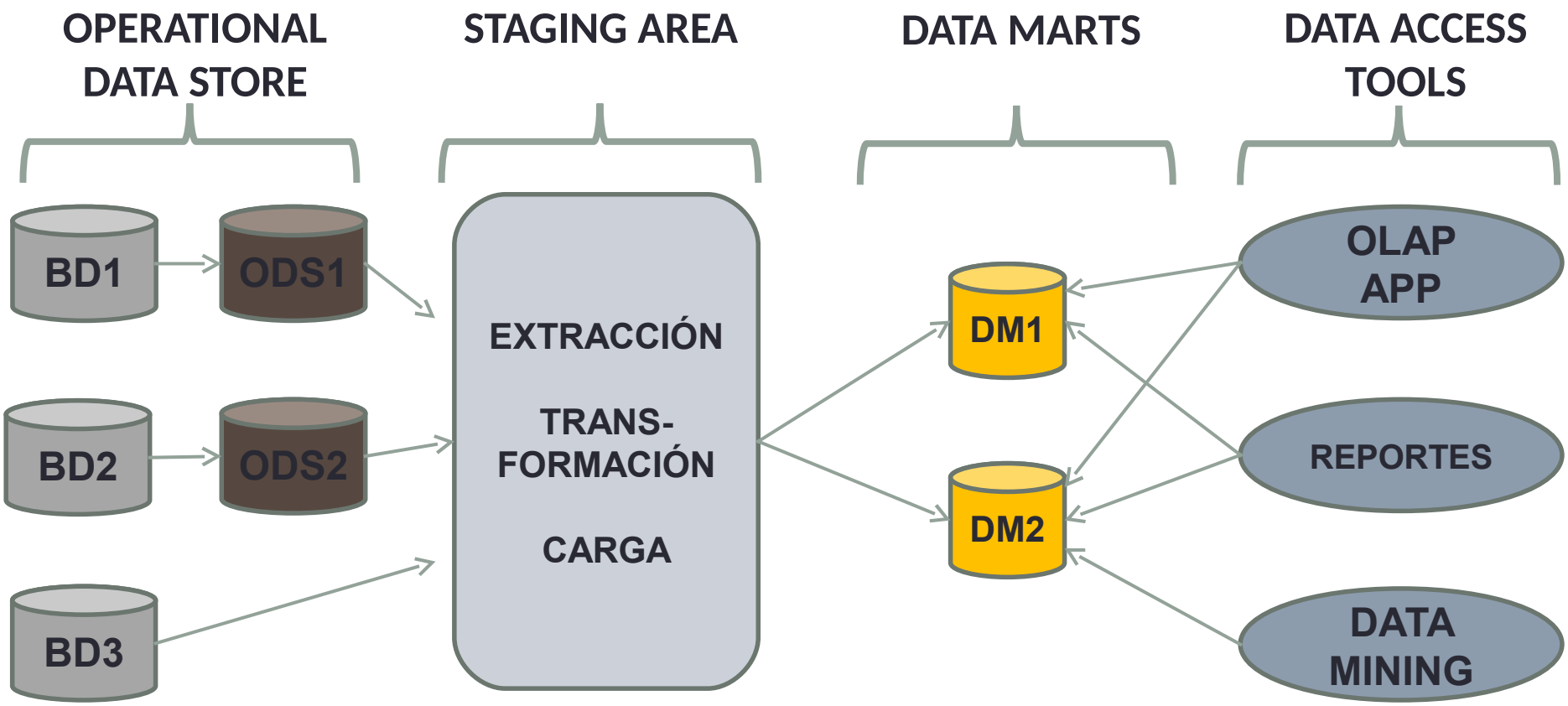
ODS, Staging, Data Marts



ODS, Staging, Data Marts



ODS, Staging, Data Marts



Operational Data Store (ODS)

Características:

- Instantánea de la base de datos operacional
- No almacena datos históricos
- Al pasar del RDBMS al ODS podrían realizarse operaciones de transformación
- Actualización no se realiza en tiempo real

Propósitos:

- Tomar un punto de corte del RDBMS donde ya no habrán actualizaciones
- No utilizar la base operativa para operaciones de carga del DWH

Staging Area

ETL (Extracción, Transformación y Carga)

- Obtiene registros de los orígenes de datos, los procesa y carga al DW.
- ***Extracción:*** Obtener datos de múltiples orígenes
- ***Transformación:*** Procesar los datos para la carga
  - Limpieza/Reemplazo de valores (NULL => 0)
  - Agrupación en clases, ej. Edad: 0-20, 20-40, 40-60, > 60
  - Filtrado, seleccionar cuáles columnas utilizar
  - Separar una columna en múltiples columnas
  - Validaciones
- ***Carga:*** Append de los datos al DW

Herramientas de ETL

- *Open source*
  - Pentaho Data Integration
  - Talend
- Comerciales:
  - IBM Infosphere Datastage
  - Oracle Data Integrator
  - Microsoft SQL Server Integration Services
  - SnapLogic (Tableau)
  - QlikView Data Files (QVD)
  - ….

Data Marts

- **Subconjunto del DW**, orientado a un tema de análisis, normalmente asociado a un departamento de la empresa. *Ej: clientes, créditos*
- **Diseño multidimensional**, cada objeto de análisis es una tabla de hechos enlazada con diversas tablas de dimensiones.
- **Desnormalizado**, esquema en Estrella propone una tabla para cada dimensión

Software BI / OLAP

- Power BI
  - Microsoft SQL Server Analysis Services
- IBM Cognos Anaytics
- Oracle BI
- Tableau
- QlikView
- *Open Source*:
  - Pentaho
  - JasperBI
  - SpagoBI

Gartner Magic Quadrant 2016

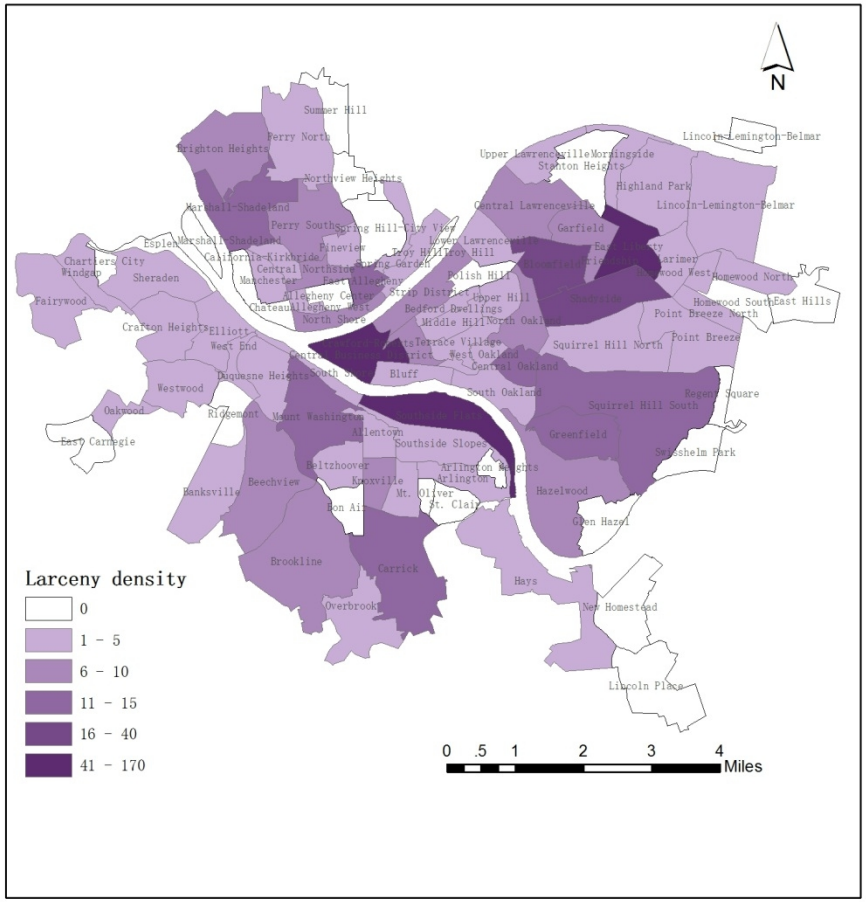
Gartner Magic Quadrant 2017

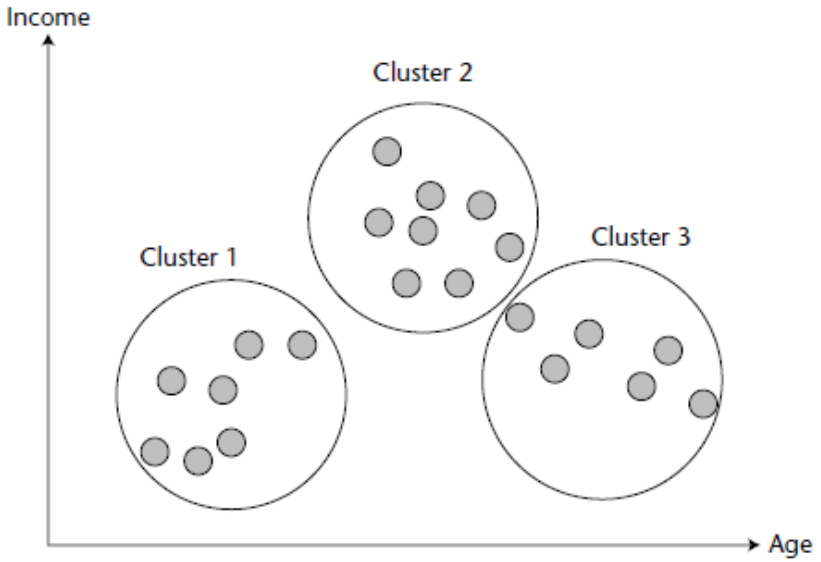
Data Mining

- **Estadísticas:** regresión, análisis multivariable, análisis clúster
- **Simbólicas:** árboles de decisión, reglas
- **Técnicas de inteligencia artificial:** redes neuronales, algoritmos predictivos

Data Mining

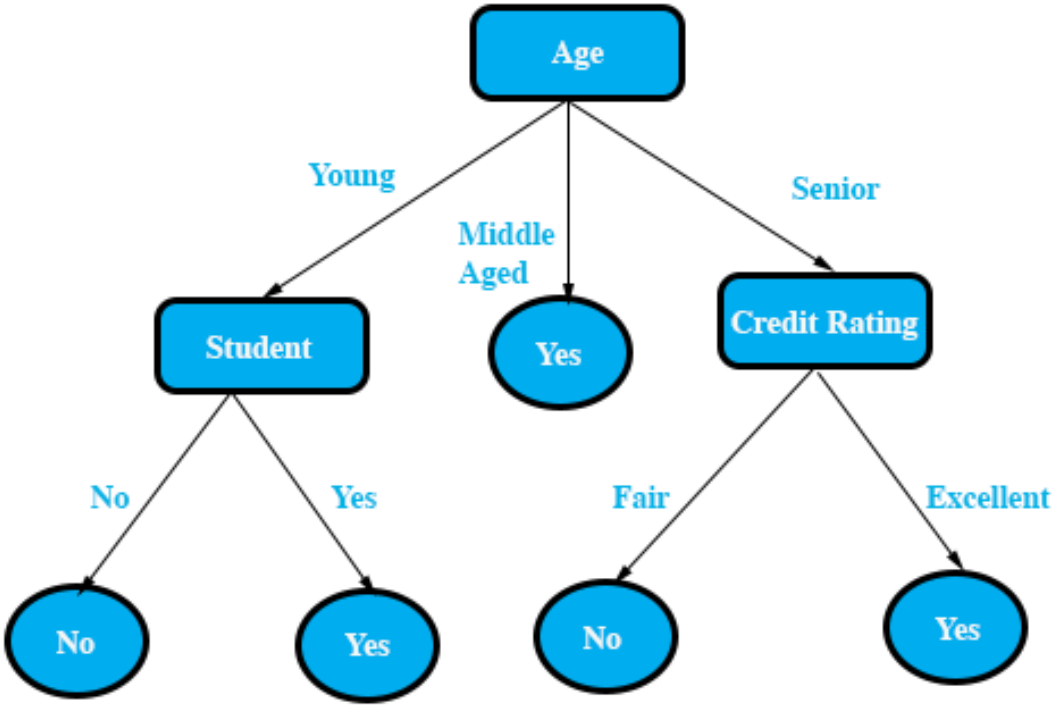
- Clustering





Data Mining

- Decision Trees
  - Ej: Aprobación de crédito



Data Mining

- Red Neuronal

https://www.youtube.com/watch?v=0F5w6q0ZpBI

Conclusiones

- **Sobre el Data Warehouse:**
  - Se requiere enriquecer el DW con la mayor cantidad de hechos
    - *OLTP*
    - *CRM*
    - *Social Networks*
    - *Server logs*

Conclusiones (2)

- **Sobre el Software de BI:**
  - Se pueden mezclar soluciones
    - Reporting
    - Cubos OLAP, tablas dinámicas Excel
    - GIS
  - Se puede complementar con otras herramientas
    - Por ej: un motor de Data Mining para calcular pronósticos y tendencias sobre los datos
  - Se requiere una constante revisión de manera a validar que no exista una mejor opción

Conclusiones (3)

- **Sobre la operativa del DW**
  - El DW soportará la carga con pocos Data Marts
  - El almacenamiento se debe controlar y optimizar su rendimiento
    - Típicamente se utiliza hardware de Storage dedicado
  - En la medida en que el desempeño del proceso de ETL disminuya (demore horas), debe plantearse particionar los datos (particiones anuales, mensuales, diarias)
  - En la medida en que el desempeño de las *apps* de consultas disminuya, debe plantearse la creación de índices y re-validar que la solución de DW continue como una buena opción

Análisis interno

- ¿Que hechos tenemos actualmente en el Data Ware?

Análisis interno

- ¿Que hechos tenemos actualmente en el Data Ware?
- ¿Hacemos análisis del CRM y Social media?

Análisis interno

- ¿Que hechos tenemos actualmente en el Data Ware?
- ¿Hacemos análisis del CRM y Social media?
- ¿Tenemos analytics georreferenciado? Recolectamos los datos en los sistemas transaccionales?

Análisis interno

- ¿Que hechos tenemos actualmente en el Data Ware?
- ¿Hacemos análisis del CRM y Social media?
- ¿Tenemos analytics georreferenciado? Recolectamos los datos en los sistemas transaccionales?
- ¿En que nivel de madurez estamos?
  - Reportes de lo sucedido? monitoreo de lo que está sucediendo? predicción de lo que sucederá?

Referencias

- Kimball, R. Ross, M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2nd Ed, Wiley. 2002
- Datawarehouse4u.info. URL: http://datawarehouse4u.info/
- Dataprix.com. URL: http://www.dataprix.com/arquitectura-data-warehouse-areas-datos-nuestro-almacen-corporativo