UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en
microelectrónica y sistemas distribuidos

Analítica de Datos para Ingeniería

Jose Aguilar

CYTED 516RT05112
CADING

CYTED
PROGRAMA IBEROAMERICANO DE CIENCIA
Y TECNOLOGÍA PARA EL DESARROLLO

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Profesor: Jose Aguilar

Contacto: aguilar@ula.ve

Información del curso: www.ing.ula.ve/˜aguilar

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

OBJETIVO

Formarse en el área de Analítica de Datos, y en las áreas afines a la misma, a través del paradigma aprender haciendo.

OBJETIVOS ESPECIFICOS

- Adquirir los conceptos claves en Analítica de Datos
- Vincular la Analítica de Datos a las ingenierías
- Desarrollar la capacidad de realizar tareas de analítica de datos en sus contextos de trabajo
- Desarrollar proyectos científicos-tecnológico en analítica de datos para diferentes ámbitos de las ingenierías

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Contenido

1. Generalidades de la Analítica de Datos
2. Metodologías para realizar Analítica de Datos
3. Ciencia de los Datos para Ingeniería.
4. Modelo de Datos
5. Tipos de Tareas de Analítica de Datos.
6. Técnicas de Analítica de Datos
7. Algunos Conceptos Vecinos:
   - Minería de Datos, Grafos y Procesos,
   - BigData

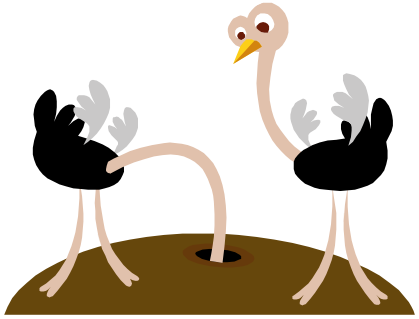
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

BIBLIOGRAFIA

- David Loshin, “Business Intelligence: The Savvy Manager's Guide”, The Morgan Kaufmann Series on Business Intelligence, 2010
- Stephan Kudyba , Richard Hoptroff , “Data Mining and Business Intelligence: A Guide to Productivity”, IGI Publishing, 2011
- José Hernández, José Ramírez Quintana, César Ferri Ramírez, "Introducción a la Minería de Datos" Editorial Pearson, 2004
- Peter F. Drucker, “Gestión del Conocimiento”, Deusto S.A. ediciones, 2009
- Ralph Kimball, Margy Ross “The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling”, Wiley; 2 edition, 2009.
- Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, “Big Data For Dummies,”, Wiley, 2013
- I. H. Witten, E. Frank & M. A. Hall "Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Third Edition". Morgan Kaufmann Publishers. San Francisco, California, 2011.
- Michael Milton, “Head First Data Analysis”, O'Reilly Media, 2009

Introducción a la Analítica de Datos y a la Ciencia de los Datos

ACTITUD hacia la vida



Actitud Pasiva

Actitud Preactiva

Actitud Reactiva

Actitud Proactiva

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

CONOCIMIENTO

"En los últimos 10 años se han producido más conocimientos que en los 10.000 años anteriores".

Bill Gates

Estamos en la Civilización del Conocimiento

La sociedad del Conocimiento

- **Ausencia de fronteras,** porque el conocimiento viaja aun con menos esfuerzo que el dinero
- **Disponible para todos,** en virtud de que la información cada día es más fácil de adquirir.
- La mayoría de los empleados cada día serán **menos de tiempo completo para la organización.**
- Nacimiento de nuevas instituciones teorías, problemas, a un **ritmo vertiginoso.**

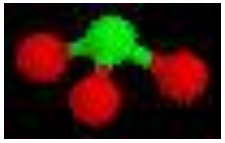
CONOCIMIENTO

ERA INDUSTRIAL

VALORES PREDOMINANTES

- Poder
- Control
- Disciplina
- Especialización
- Estructura jerarquizada

ERA DEL CONOCIMIENTO



VALORES PREDOMINANTES

- Descentralización
- Información
- Innovación
- Calidad
- Trabajo en equipo

Los trabajadores trabajan más con sus mentes que con sus manos (Knowledge Worker)

SOCIEDAD DEL USO DE CONOCIMIENTO

Embudo del Conocimiento

Administración Inteligente

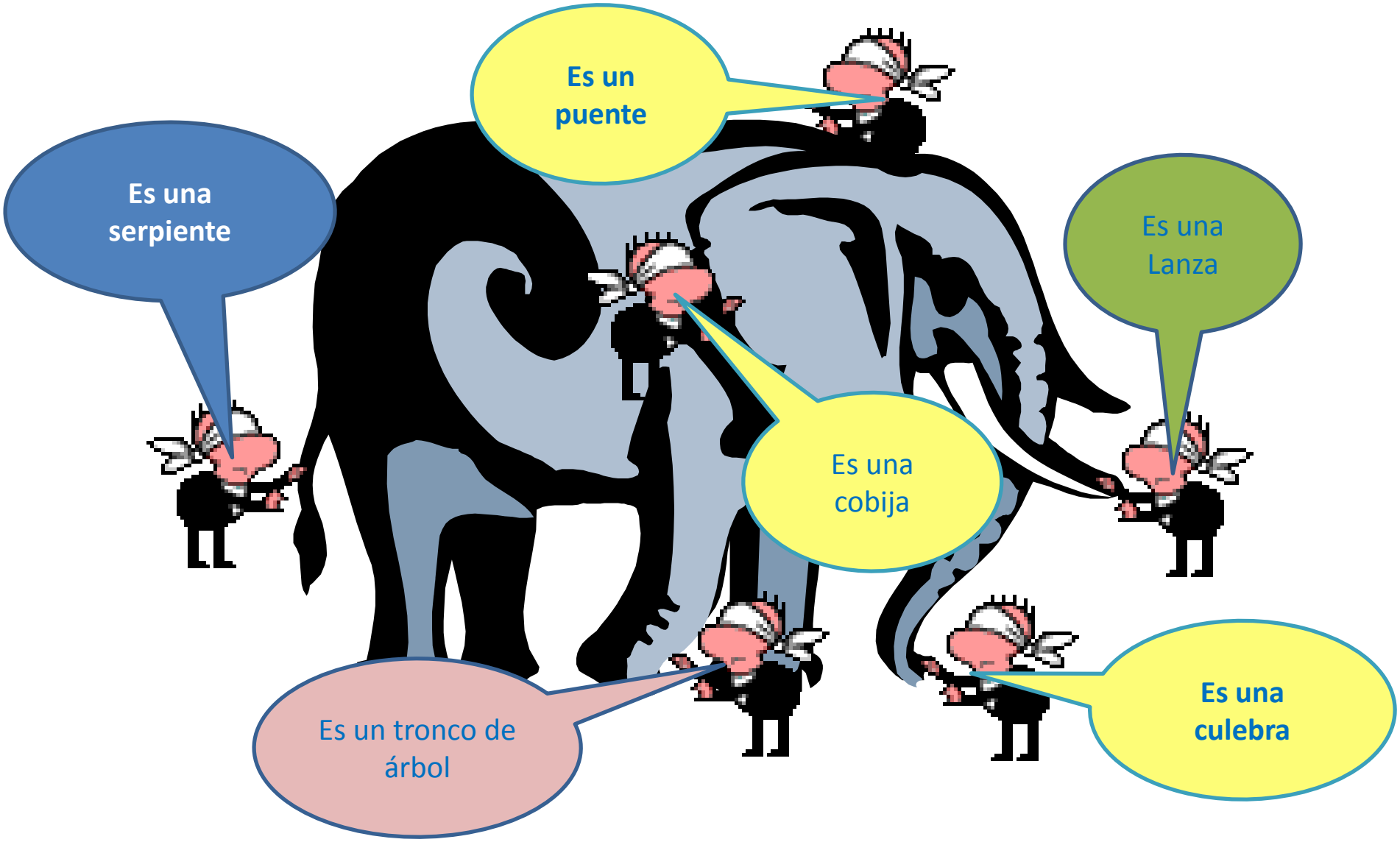
No se puede administrar lo que no se puede ver

La capacidad de ver toda la organización es el aspecto más importante de la **administración inteligente.**

Las organizaciones tienen que ser **flexibles y adaptables** a cambios en el proceso y modelo de los negocios, así como también a nuevas tecnologías.

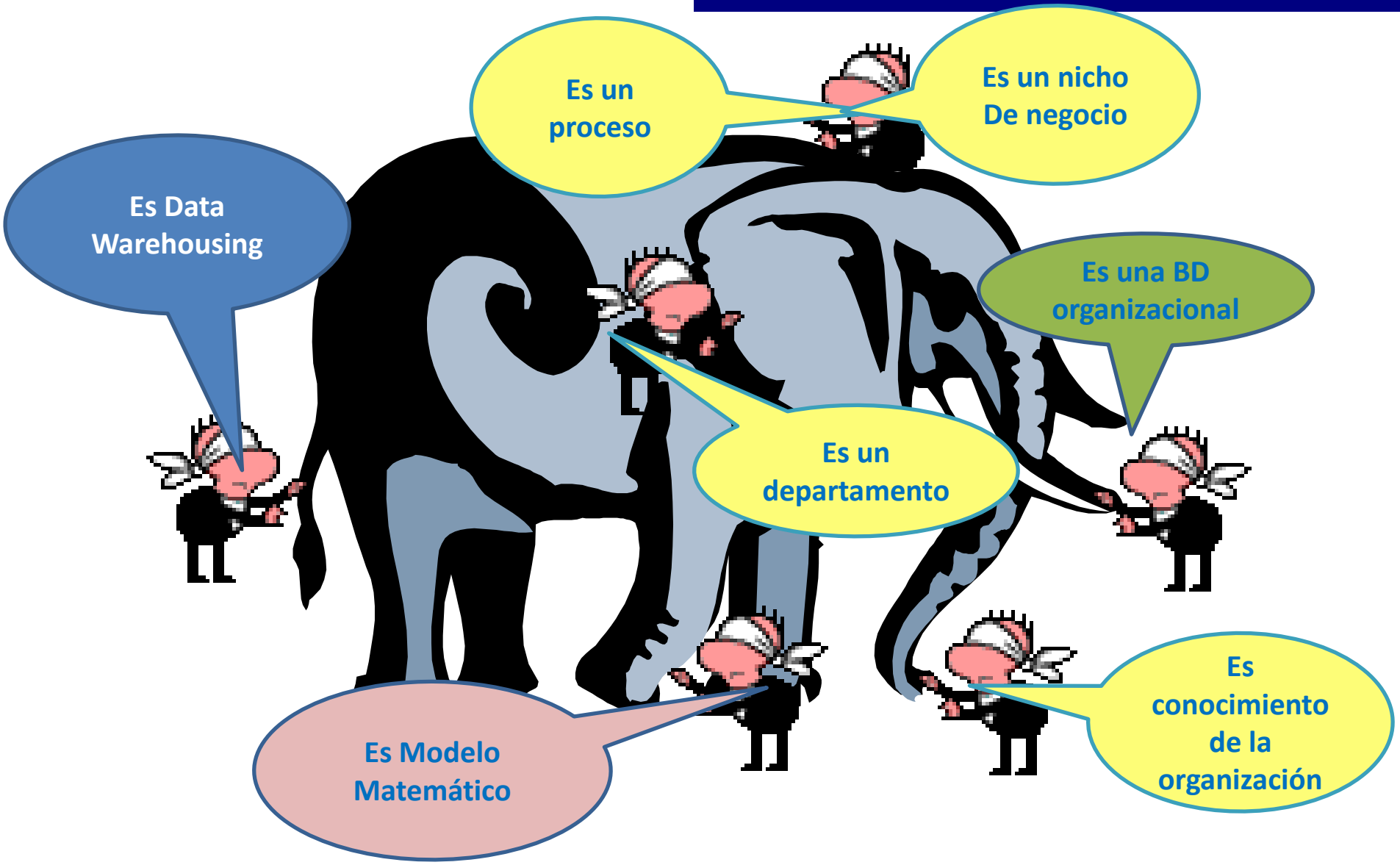
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Un ciego describiendo un Elefante



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Un gerente describiendo su organización Sin A`dD



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Dato, información, conocimiento e inteligencia en una organización

Dato, información, conocimiento e inteligencia en una organización

Sistemas Operacionales

Eventos

Aprendizaje

Datos

Decisiones estratégicas

Planes

Reglas

Conocimiento

Tendencias, patrones y excepciones

Información

Extracción e Integración

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Analítica de Datos

Es la ciencia de la recogida, almacenamiento, extracción, limpieza, transformación, agregación y análisis de datos, **con el fin de descubrir información y conocimiento.**

El **alto grado de datificación incrustado en la sociedad** exige nuevas herramientas y mecanismos para la manipulación y la representación de los datos que facilitan la **extracción de conocimiento significativo para las organizaciones**.

Analítica de Datos

Los datos son el nuevo petróleo de la economía

Es la ciencia que examina datos en bruto con el propósito de buscar conocimiento, sacar conclusiones, generar información, entre otras cosas.

Es usado en muchos ámbitos:

- La industria para tomar mejores decisiones empresariales
- Las ciencias para verificar o reprobar modelos o teorías existentes.
- …

Analítica de Datos

Con las grandes cantidades de datos disponibles, las organizaciones **deben centrarse en la explotación de los datos** para obtener una **ventaja competitiva**.

- **Las computadoras** son más poderosas,
- **el trabajo en red** es omnipresente,
- se han desarrollado **algoritmos que pueden conectar conjuntos de datos**

esto permite **análisis más amplios y profundos** que antes era imposible.

Los objetivos principales de AdD son:

- Ayudar a ***ver los problemas de la Organización desde una perspectiva de los datos***, y
- Comprender los principios de ***extracción de conocimiento útil a partir de los datos***.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

La gestión usando AdD

El éxito de la analítica sólo puede medirse en términos de lo bien que ayudan a lograr objetivos estratégicos

Por lo tanto, se debe:

- Identificar los objetivos de la organización
- Recoger los datos necesarios para medir sus objetivos
- Analizar los datos
- Sacar conclusiones basado en los datos

Ejemplo de Análisis de los datos

Datos disponibles para un agricultor:

1. Los patrones climáticos históricos
2. Los datos de cultivo de plantas y la productividad de cada cepa
3. Las especificaciones de los Fertilizantes
4. Las especificaciones de los Plaguicidas
5. Los datos de productividad del suelo
6. Los datos del ciclo de plagas
7. Los costos, la fiabilidad, etc., de las máquinas
8. Los datos de conducción del agua
9. Los datos históricos de oferta y demanda
10. Los precios del Mercado

Ejemplo de Análisis de los datos

Los datos se pueden usar para responder a:

¿Cuál es el patrón de siembra agrícola para obtener el mejor precio?

¿cuáles son los productos agrícolas con mejor rendimiento?

Usar Analítica Datos para obtener conocimiento, desde sus datos.

- En cuanto a los **datos del tiempo y plagas**, podría establecer **correlaciones entre un cierto tipo de hongo cuando el nivel de humedad** alcanza un cierto punto.
- Si las **futuras proyecciones meteorológicas** para los próximos meses **predicen un bajo nivel de humedad,** por lo tanto, **habrá riesgo bajo de ese hongo**.

Para el agricultor, esto podría significar ser capaz de plantar **un determinado tipo de producto agrícola no resistente a ese hongo**, con un mayor rendimiento y precio en el mercado, **sin tener que comprar un determinado fungicida**...

Analítica de Datos

Los datos pueden "hablar“

El análisis de datos contiene aspectos del razonamiento científico:

Define
Interpreta
Evalua
Ilustra
Discute
Explica
Clarifica
Compara
Contrasta

Data Analytical

http://www.youtube.com/watch?v=-xR5erOhkXo

Society for Industrial and Applied Mathematics

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Objetivo de un análisis:

- **Explicar** los fenómenos de causa y efecto
- **Encontrar** respuestas a un problema particular
- **Concluir** acerca de eventos del mundo real basado en el problema
- **Aprender** de un problema
- **Predecir/pronosticar** en el mundo real fenómenos
- **Interpretar/Analizar** una situación
- …

La ciencia de los datos

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Ciencia de los datos

¡Todos estamos generando datos!

tomando el tren

abastecimiento de combustible el automóvil

comprando un café

ajustando la temperatura

poniendome una multa por exceso de velocidad

enviando un correo electrónico

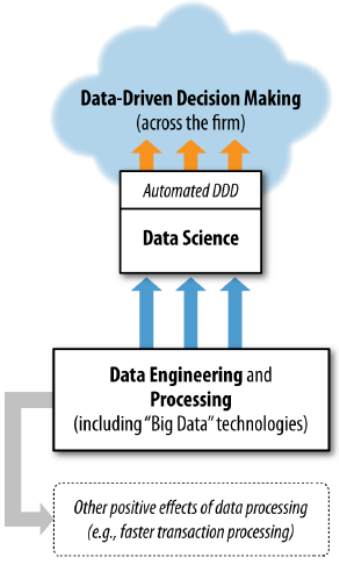
haciendo una cita

viendo esta clase

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

La ciencia de los datos

La ciencia de datos requiere de *principios, procesos y técnicas* para la comprensión de los fenómenos para la *extracción automatizada* de los datos.



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

La ciencia de los datos

Combinación de las matemáticas, estadísticas, etc., para resolver el problema de captura de datos, además de la limpieza, la preparación y la alineación de los datos.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

estadísticas
estocástica

conocimiento del
dominio

Ingeniería
industriales

comportamiento /
social

**minería
datos**

**minería
proceso**

Ciencias de los datos

bases de datos
Algoritmos

inteligencia
proceso
negocio

computación distribuida
a gran escala

visualización
datos

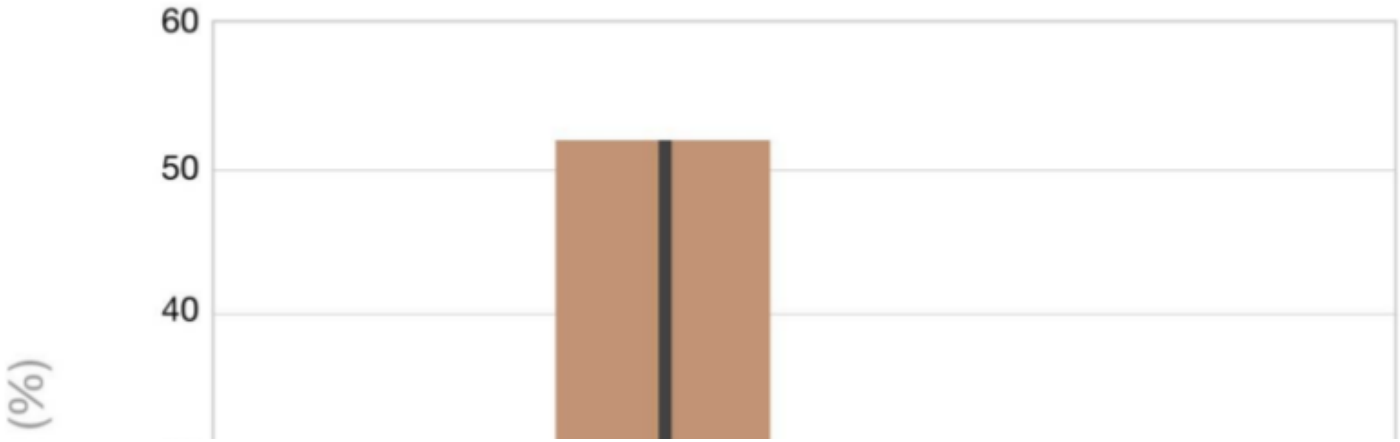
Conocer los datos

La ciencia de datos es un procedimiento que **consume tiempo y requieren mucho trabajo**, pero que es absolutamente necesario para la AdD con éxito.

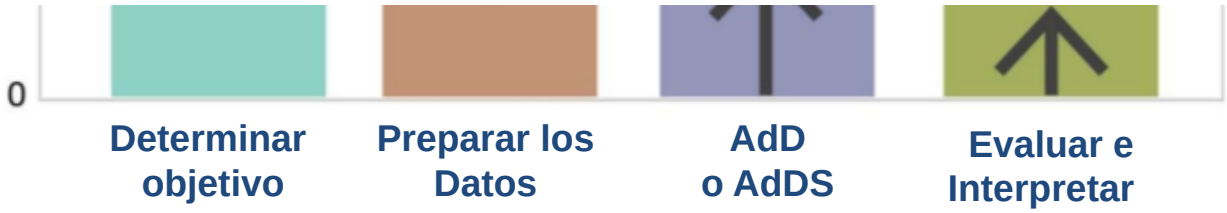
Los datos deben pasar por **procesos de ensamblaje, integración, limpieza, agregación y preparación general**.

Expertos de dominio deben ser consultados para explicar las anomalías, los valores perdidos, el significado de los números enteros que representan categorías en lugar de cantidades numéricas, y así sucesivamente.

¿Qué etapa lleva mas esfuerzo?

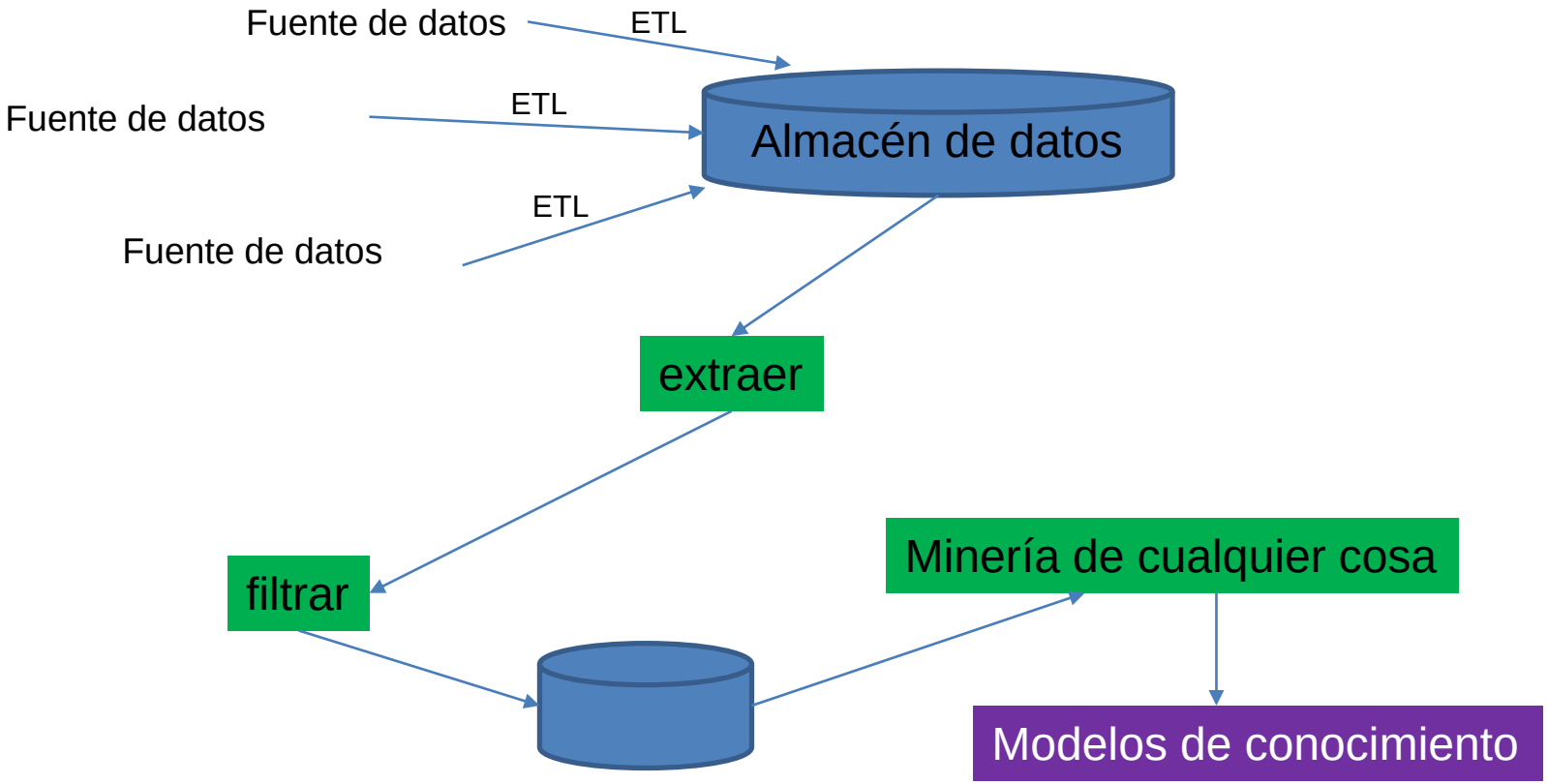


Preparación de la entrada para un proceso de AdD suele **consumir la mayor parte del esfuerzo invertido en el proceso**.



La ciencia de datos es un procedimiento que consume tiempo y requieren mucho trabajo, pero que es absolutamente necesario para la AdD con éxito.

Descripción del flujo de trabajo de obtener desde fuentes de datos heterogéneas para procesar datos



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Problemas: Calidad de los datos

- valores faltantes

cómo interpretar? ¿no disponible? usar el medio?

- valores duplicados: incluyendo partidos parciales (Jon Smith = John Smith?)
- inconsecuencia:

múltiples direcciones para la persona

- datos desactualizados
- uso inconsistente:

Significa "destino" de la primera etapa o de vuelo?

- valores atípicos:

sueldos que son negativos,

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Problemas: Interoperabilidad

- ¿Cómo pueden compararse o combinarse datos de una base de datos con otra?
- ¿Qué pasa si los campos no son lo mismo, o no están presentes, o se usan de manera diferente?
- traducción / mapeo de términos estándares

  unidades como metros, o galones, etc.
  identificadores como SSN, UIN, ISBN
- Bases de datos "federadas": consultas que combinan información en varios servidores

La ciencia de los datos

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas mencionados como datos mal capturados, anómalos y vacíos, etc.

- normalización de formatos,
- remoción de anomalías,
- corrección de errores
- eliminación de duplicados.

La ciencia de los datos

Transformación

En esta etapa se **transforman las variables de entrada en nuevas variables de interés**, a través de diversos métodos.

Una transformación de variables puede ser la combinación entre variables

- concatenación de cadenas,
- multiplicación entre variables,
- otras operaciones aritméticas, etc.

La ciencia de los datos

Reducción

Consiste en **decidir qué datos deben ser utilizados para el análisis**. El criterio que se sigue para realizar reducción de variables incluye la relevancia con respecto a los objetivos que se persiguen, y limitaciones técnicas tales como los volúmenes máximos de datos o tipos de datos concretos.

Así que en este paso se reduce la cantidad de variables **a sólo las necesarias para modelar el proceso en estudio**.

- **Realizar análisis estadísticos** para reducir variables que posean una alta relación lineal, como por ejemplo un análisis de correlación.
- **Identificar las posibles variables** que se pueden reducir.
- **Justificar la reducción** de las mismas
- **Construir la nueva vista minable** con las nuevas variables reducidas

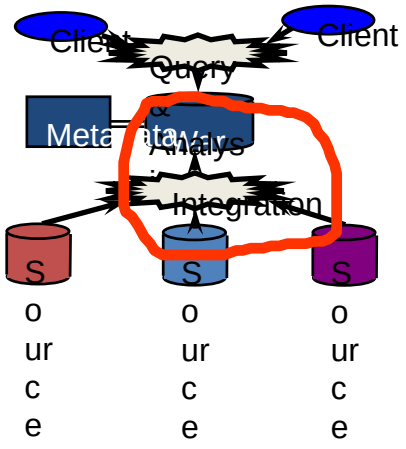
Proceso ETL

ETL (Extracción, Transformación y Carga)

Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

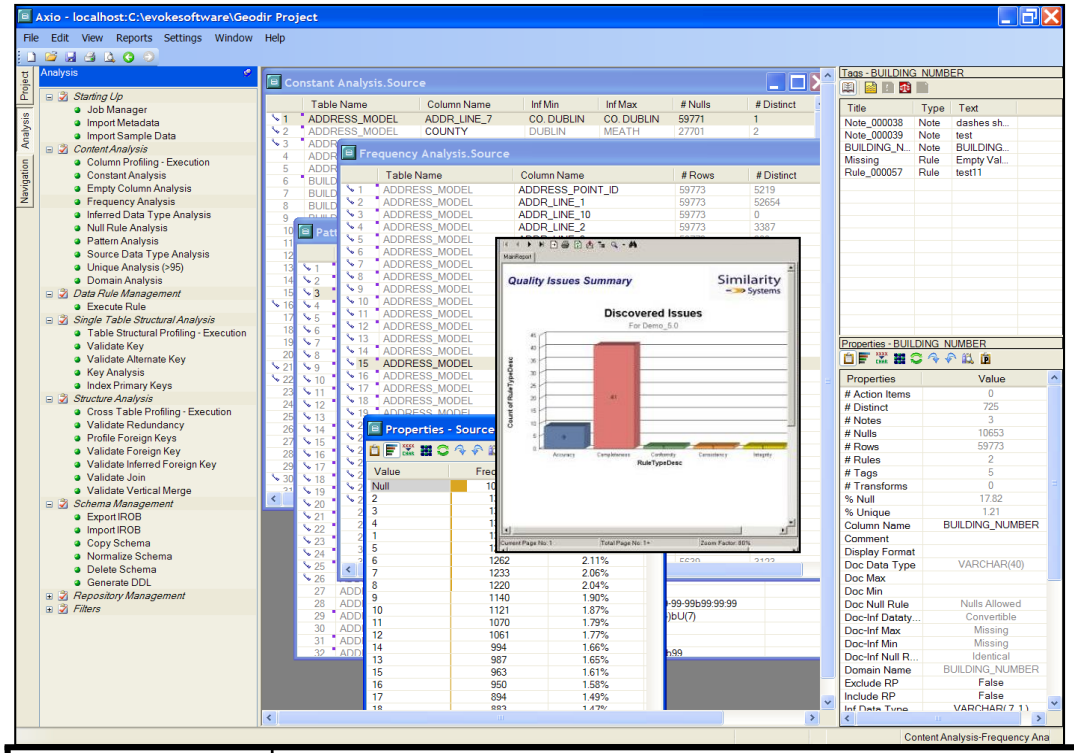
Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.



Preparación de los Datos

Perfil de los datos



Experto IN 	Perfila y etiqueta anomalías
Ùsuario 	Revisa anomalías

Catálogo de Problemas

- Completos
- Conformes
- Consistentes
- Precisos
- Duplicados
- Dependencias
- Correctas especificaciones y transformación
- **Íntegros**

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Peligros en AdD

Ciencias de los Datos

- Privacidad
- Seguridad
- Decisiones basado en datos incompletos
- Decisiones con datos inexactos
- Usando sólo los datos que apoyan nuestras decisiones
- Llegar a la conclusión errónea de los datos: por ejemplo, los precios de las acciones

Modelado de Datos

Introducción a Data Warehouse

Analizando la información de una empresa

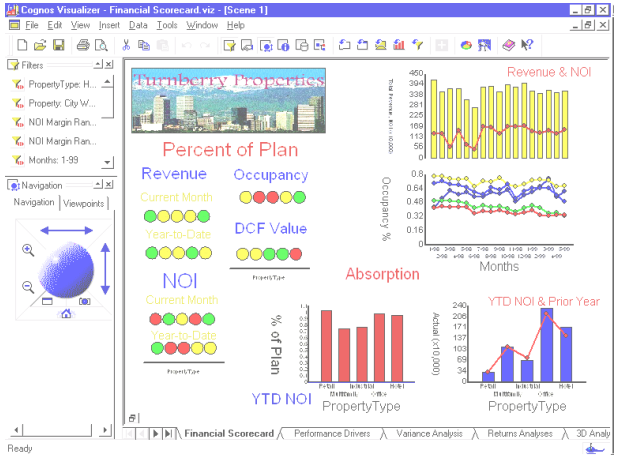
- ✓ Información periódica de las ventas
- ✓ Información del esfuerzo comercial
- ✓ Información sobre los pedidos a los proveedores

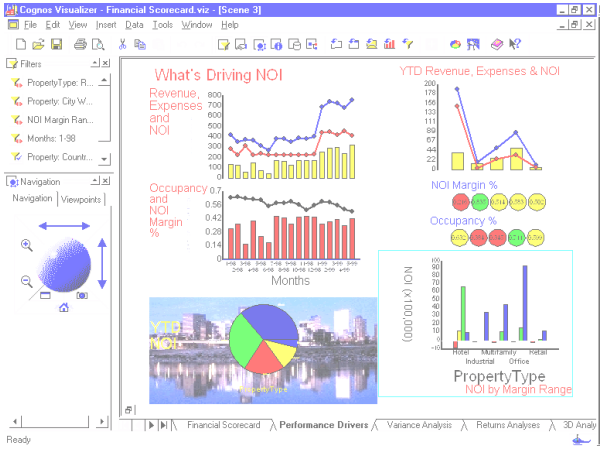
Por qué no integrarla y cruzarla para obtener:

- ✓ ¿En qué zonas se está vendiendo más de cada línea de productos?
- ✓ ¿Quienes son los clientes más rentables?
- ✓ ¿Cuál es la relación entre el esfuerzo comercial y las operaciones cerradas?
- ✓ ¿De qué proveedores se está comprando la mayor parte de los productos vendidos ?

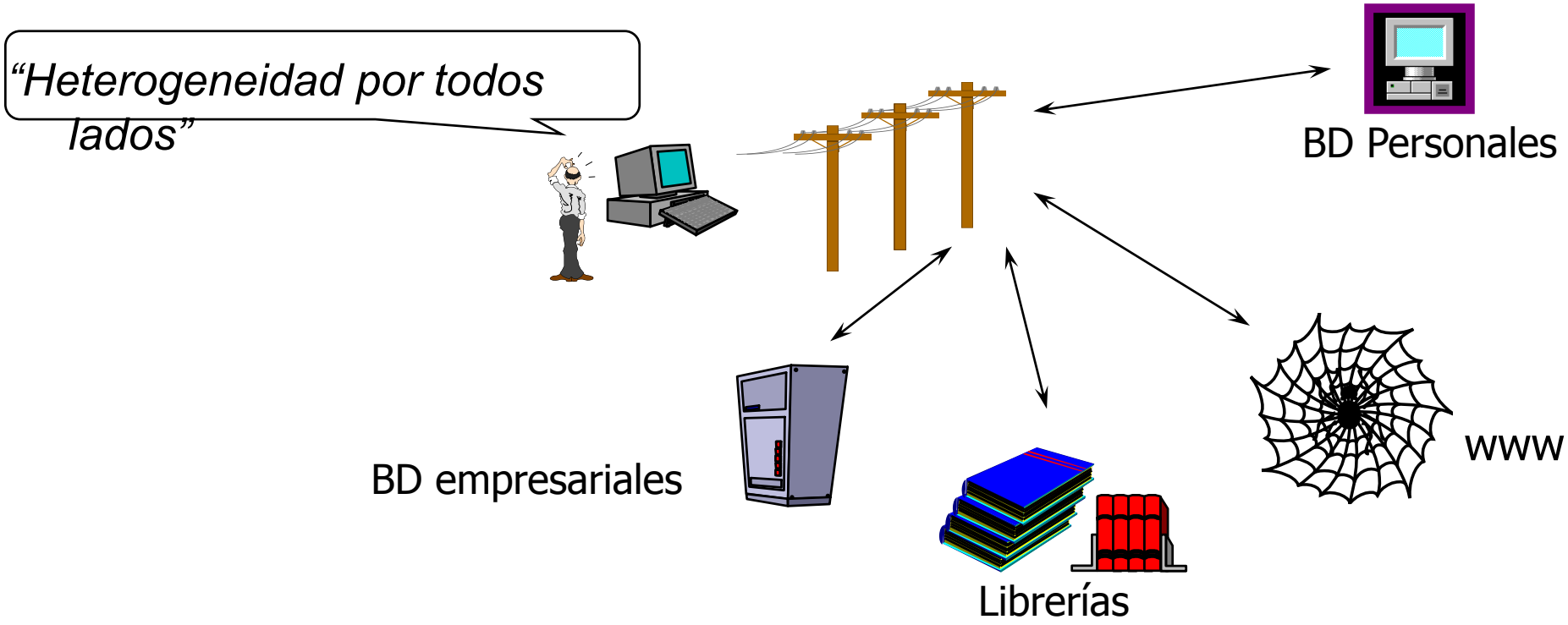
Introducción a Data Warehouse

- ✓ Se necesita entender no solo QUÉ está pasando, sino CUÁNDO, DÓNDE, QUIÉN, CÓMO Y POR QUÉ.
- ✓ Requerimientos de información con OPORTUNIDAD .
- ✓ ESCALAR, ENRIQUECER Y COMPARTIR a todos los usuarios en la organización





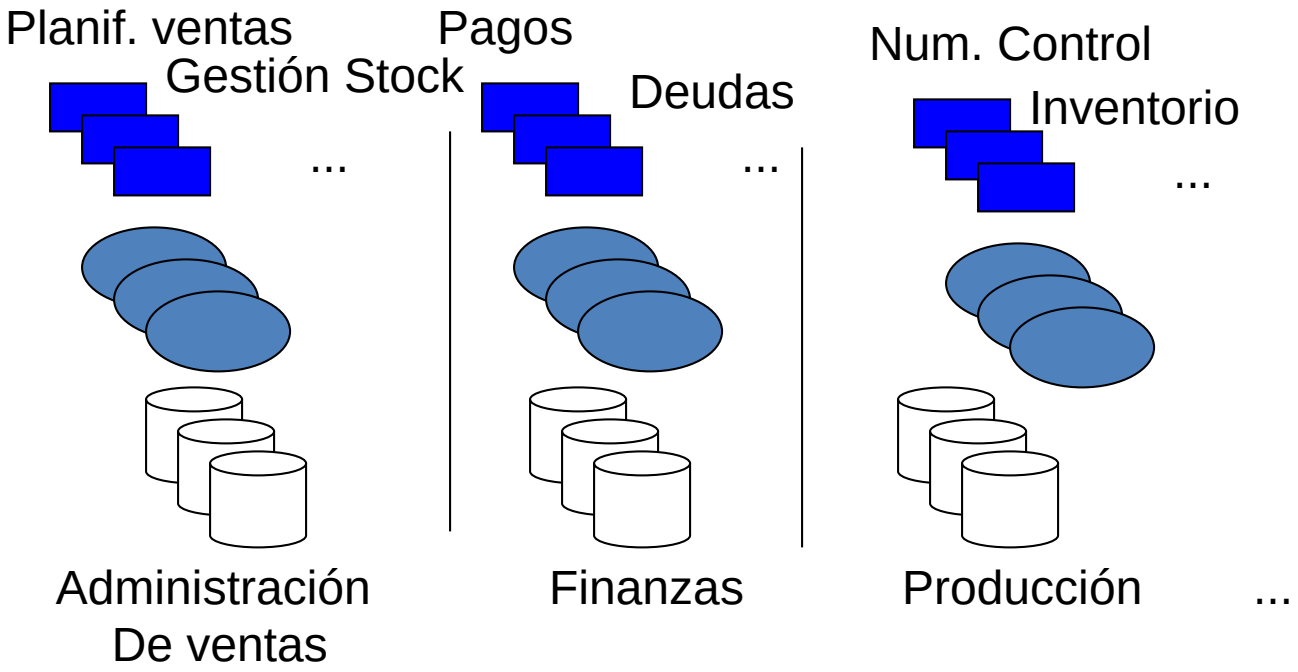
Problemas: Heterogeneidad de las fuentes de Información



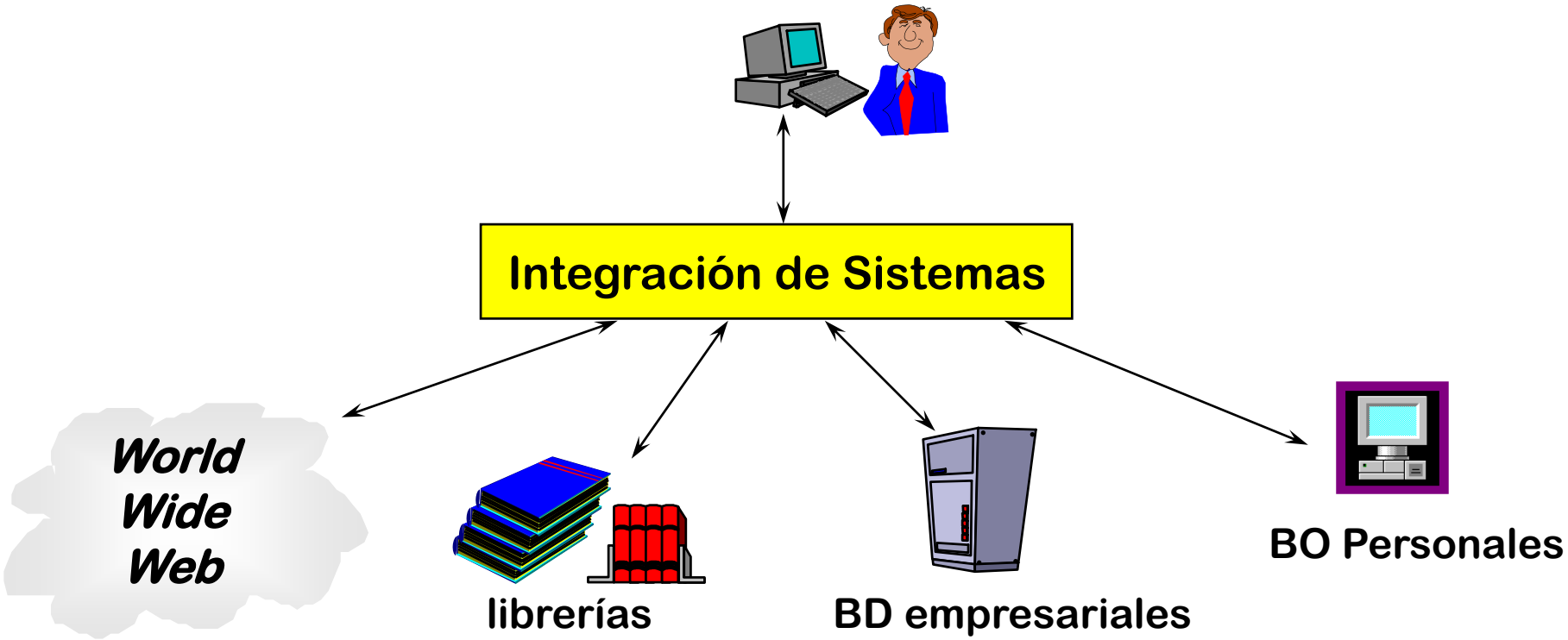
- Diferentes interfaces
- Diferentes representaciones de datos
- Información duplicada e inconsistente

Problemas: Islas de Gestión de datos en grandes empresas

- **Fragmentación vertical** de los sistemas de información
- Desarrollo de las **aplicaciones guiadas por los sistemas operativos**



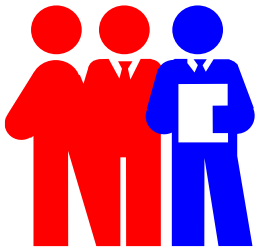
Objetivo: Unificar Acceso a los Datos

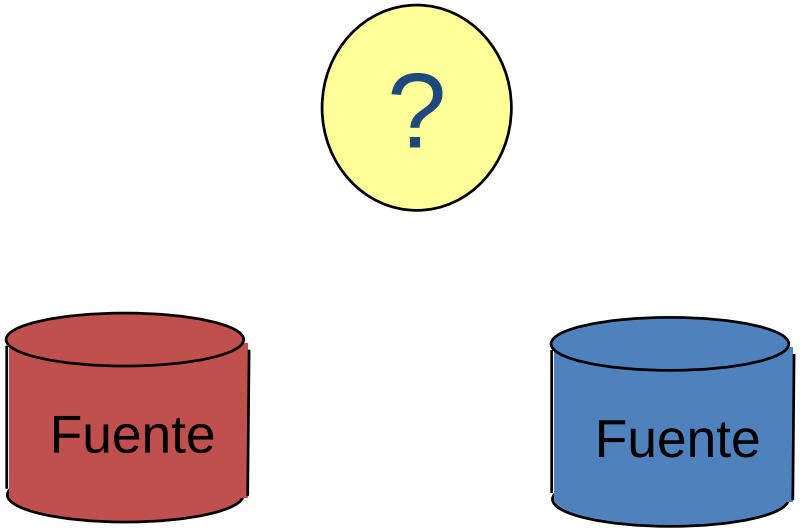


- Recopilar y combinar la información
- Proporcionar visión integrada, en una interfaz de usuario uniforme
- Soportar el intercambio

Objetivo: Unificar Acceso a los Datos

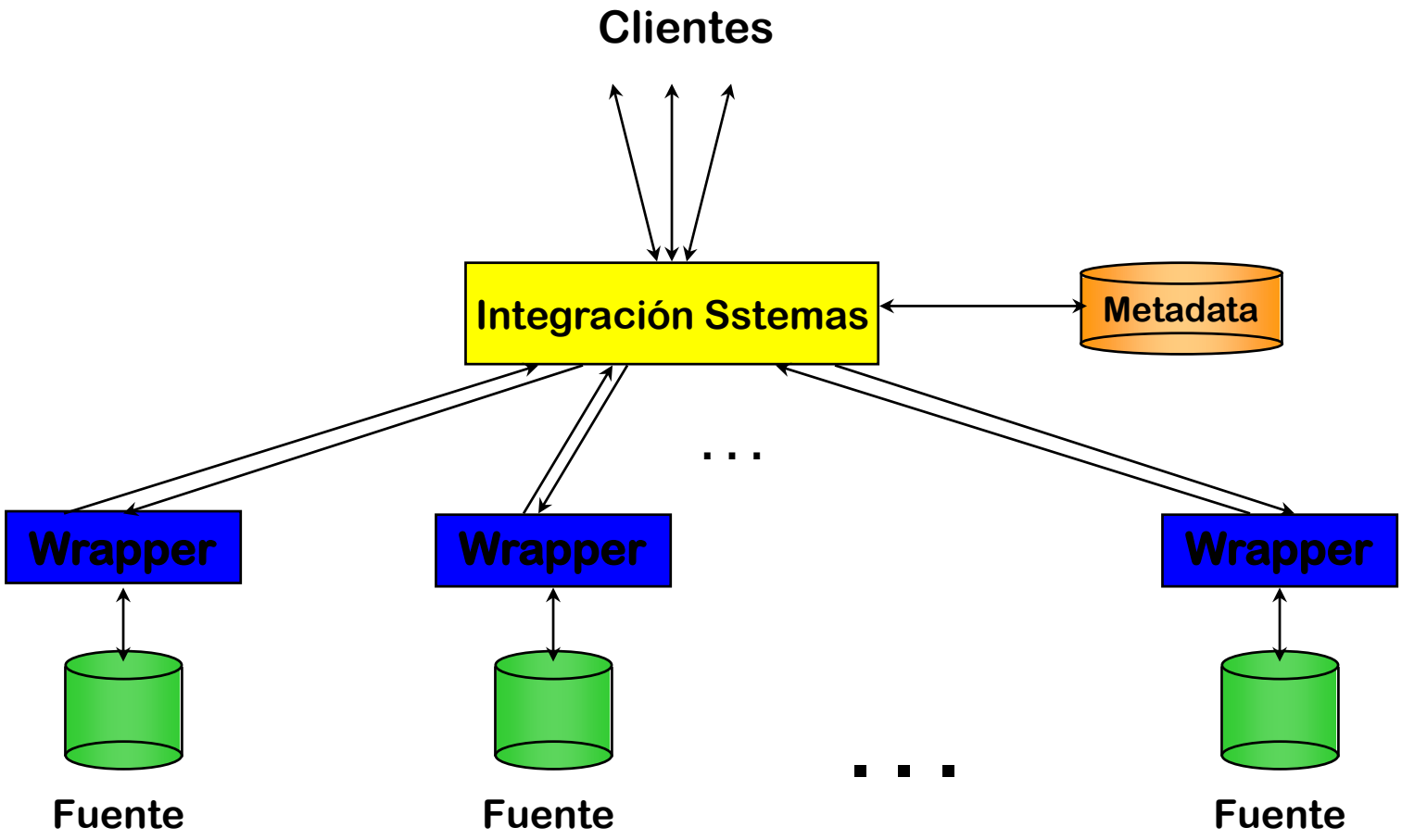
- Dos enfoques:
  - Guiado por la consulta (perezoso)
  - Warehouse (ansioso)





Enfoque tradicional

Guiado por la consulta (perezoso, on-demand)



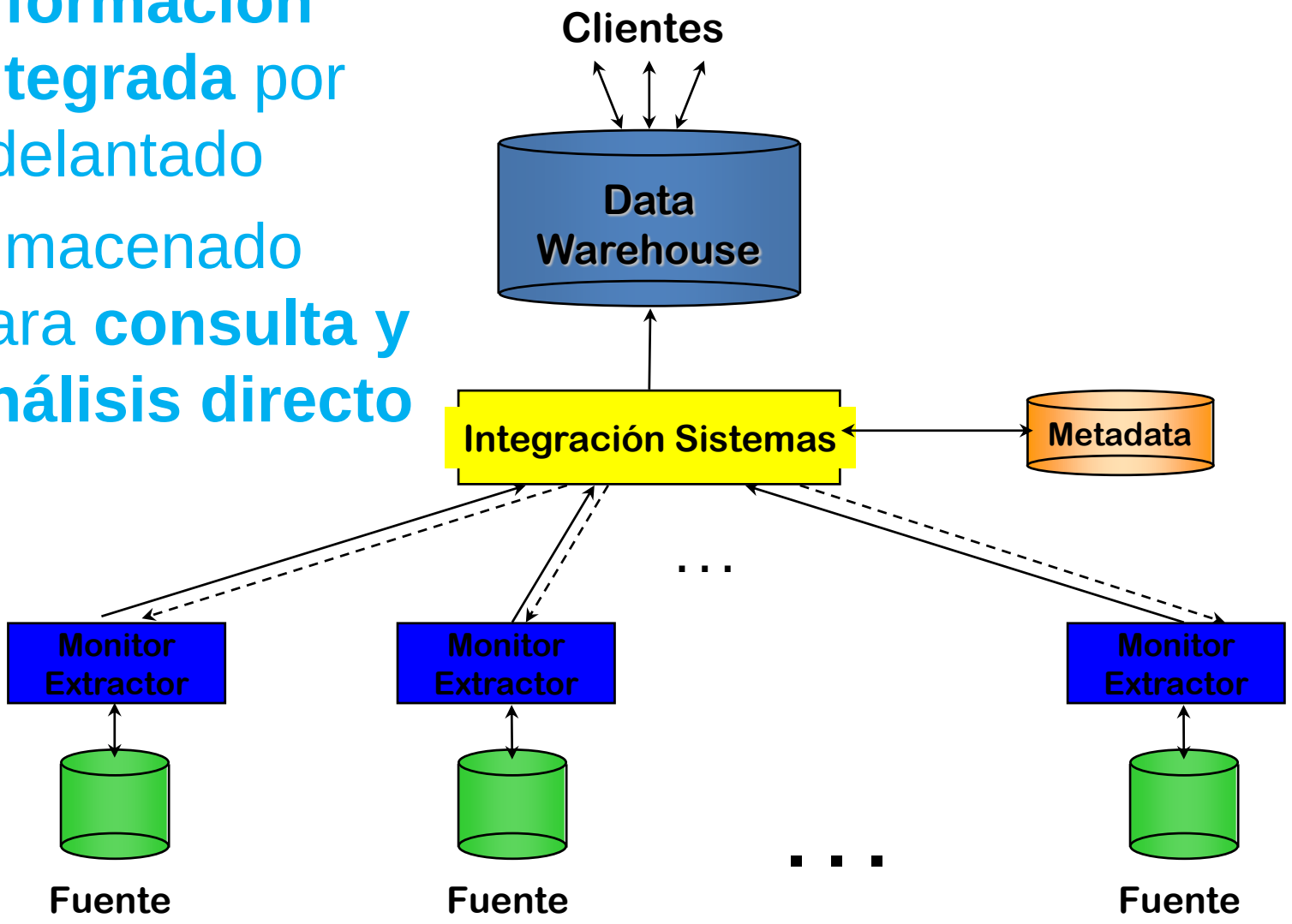
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Problema del Enfoque tradicional

- El **retraso** en el procesamiento de consultas
- Fuentes de información **lentas o no disponibles**
  - Filtrado e integración son complejas
  - Ineficiente y potencialmente costoso para las consultas frecuentes
- **Compite con el procesamiento local** en el sitio fuente

Enfoque Warehousing

- **Información integrada** por adelantado
- Almacenado para **consulta y análisis directo**



Ventaja Enfoque Warehousing

- **Alto rendimiento de la consulta**
  - Pero no necesariamente la información más actualizada
- **No interfiere con el procesamiento local en el sitio origen**
  - Las consultas complejas en warehouse
  - OLTP en las fuentes de información
- **Información copiada en el almacén**
  - Se puede modificar, anotar, resumir, reestructurar, etc.
  - Puede almacenar información histórica
  - Seguridad, sin auditoría
- **Usada en la industria**

Concepto Data Warehousing

Es un **gran almacén de datos** para consultar

Es un **repositorio de datos** de muy **fácil acceso**, alimentado de **numerosas fuentes**, transformadas en grupos de información sobre **temas específicos de negocios**, para permitir **nuevas consultas, análisis, toma de decisiones**.

Orientado hacia temas

Integración de Datos

… variante en el tiempo ...

No volátiles

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Por qué Data Warehouse?

Diferentes funciones y datos:

- **Datos que faltan**: apoyo a las decisiones requiere datos históricos que BDs operacionales no suelen tener
- **Consolidación de datos**: Se requiere de la consolidación (agregación, resumen) de los datos de fuentes heterogéneas
- **Calidad de los datos**: las diferentes fuentes suelen utilizar representaciones de datos inconsistentes, códigos y formatos que deben ser conciliados, etc.

Funcionamiento de Data warehouse

Tres funciones esenciales:

1. Recopilación de los datos desde Bases de Datos
2. Gestión de los datos en el almacenamiento
3. Análisis de datos para toma de decisiones

Sistemas de Soporte a Decisiones (DSS)

herramientas para hacer consultas o informes

Sistemas de información ejecutiva (EIS)

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Modelos dimensionales

Es una técnica de **diseño lógico** comúnmente utilizada para Data Warehouses, que busca presentar los datos en una arquitectura estándar y permita una **alta performance de acceso** a los usuarios finales.

El modelo se basa en **esquemas estrella**, conformados por **Tablas de Hechos y Tablas Dimensionales** (p.ej. cubos).

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Modelos dimensionales

- Un **modelo relacional desnormalizado**
  - Compuesto por tablas con atributos
  - Las relaciones son definidas por claves nuevas y claves externas
- Organizado para l**a comprensibilidad y facilidad de presentación** de informes, en lugar de facilitar la actualización
- Consultado y mantenido por **herramientas especiales de gestión analítica**

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Esquema en estrella: Componentes

- Datos (hechos)
- Dimensiones
- Atributos
- Jerarquías de atributos

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Diseño de Esquemas

Los tipos de Esquema

- **En estrella**
- **Constelación**
- **Copo de nieve**

1. **Aislar Datos a tener en cuenta**
   - Esquemas de las Tablas de hechos
2. **Definir las dimensiones**
   - Ejes de análisis
3. **Estandarizar dimensiones**
   - Dividir en varias tablas unidas por referencias
4. **Integrar todo**
   - Varias tablas de hechos comparten algunas tablas de dimensiones (constelación de la estrella)

Diseño de Esquemas

- **Los datos en las dimensiones se organizan por temas:**

  Los clientes, los productos, las ventas, ...

- **Tema = datos + dimensiones**
  - **Recopilación de datos** útiles sobre un tema

    Ejemplo: ventas
  - **Sintetizar** una visión única de los temas a analizar

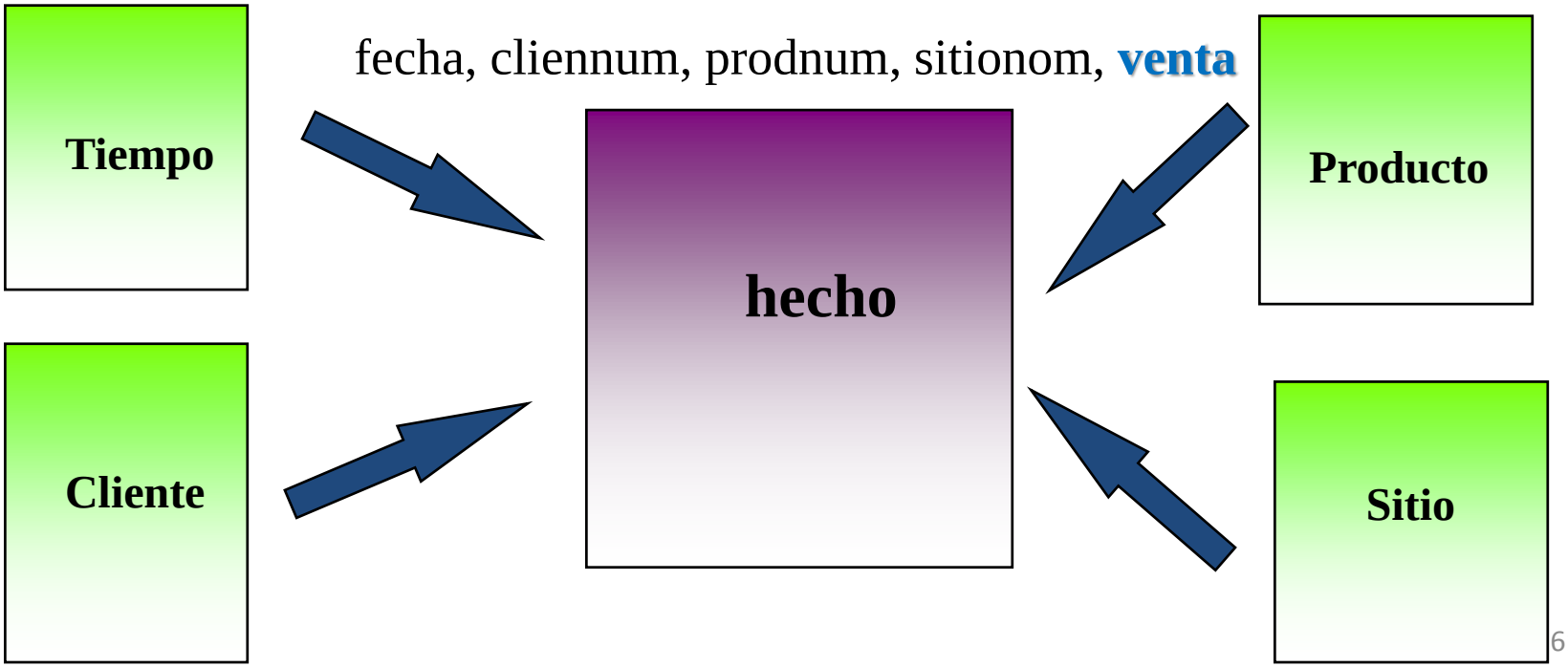
    Ejemplo: Ventas (producto, período, tienda, número)
  - **Detallar** la vista según dimensiones

    Ejemplo:

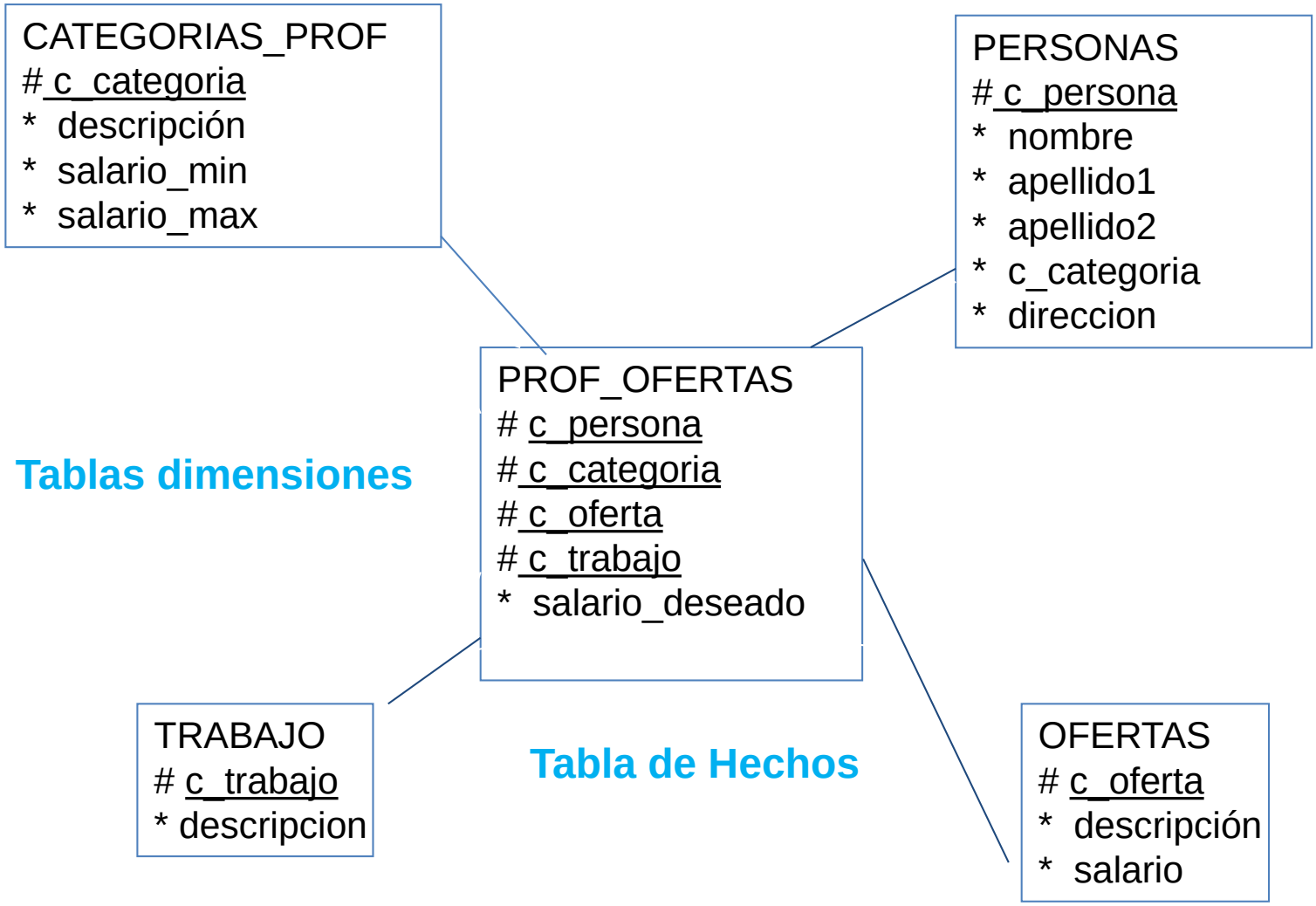
    Productos (IDprod, descripción, color, tamaño ...)
    Tiendas (IDmag, nombre, ciudad, departamento, país)
    Periodo (IDper, año, trimestre, mes, día)

Esquema en estrella

- Una sola tabla de hechos y para cada dimensión una tabla de dimensiones
- No captura jerarquías directamente



Esquema en estrella



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

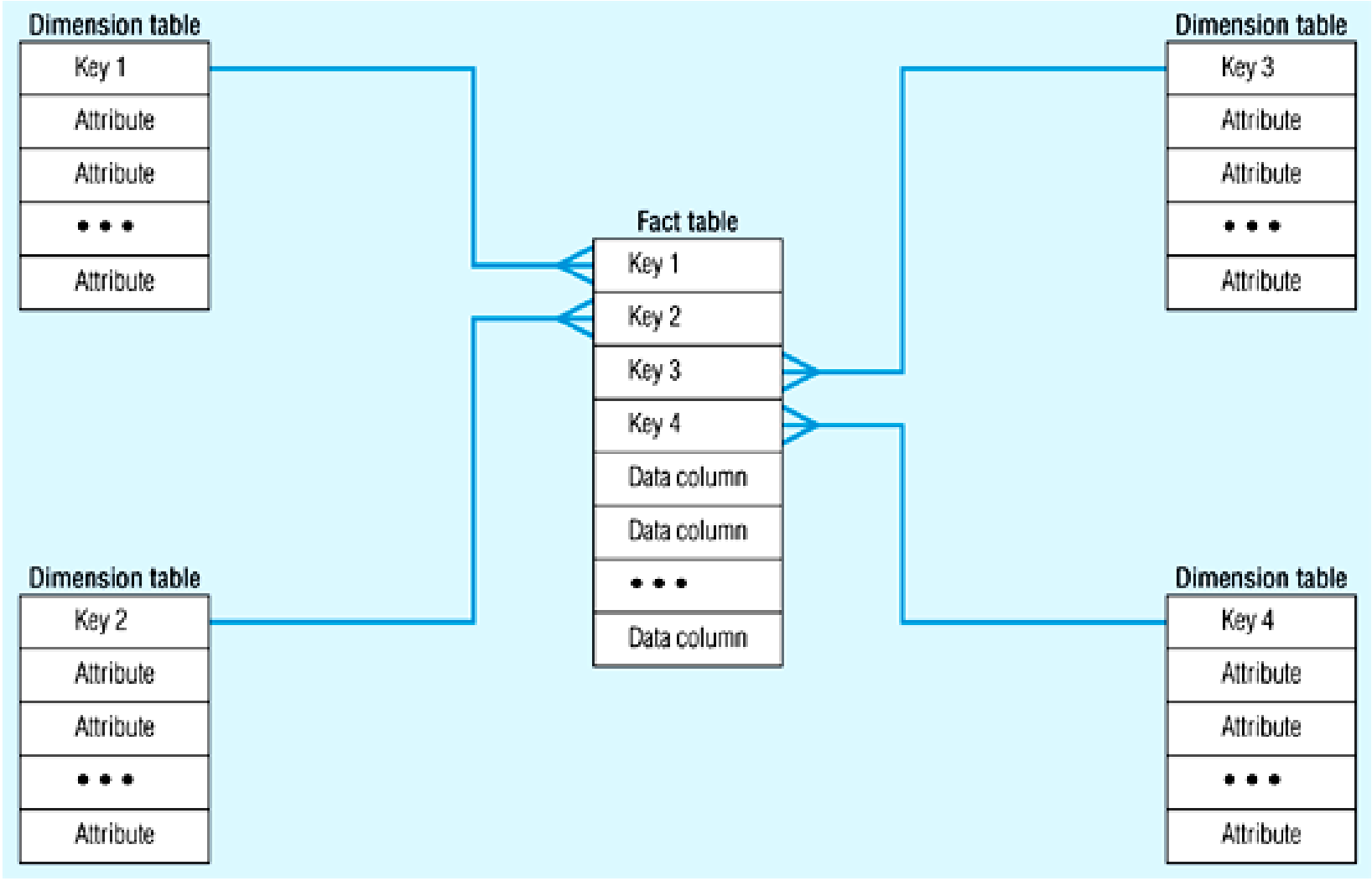
Esquema en estrella

- El modelo estrella es una representación de una vista de la organización.
  - Ventas
  - Mercadeo

- El modelo estrella consolida hechos en relación a dimensiones o filtros.

- Esquema en estrella
  - Hecho rodeado de varias dimensiones (4-15)
  - Las dimensiones se de-normalizan

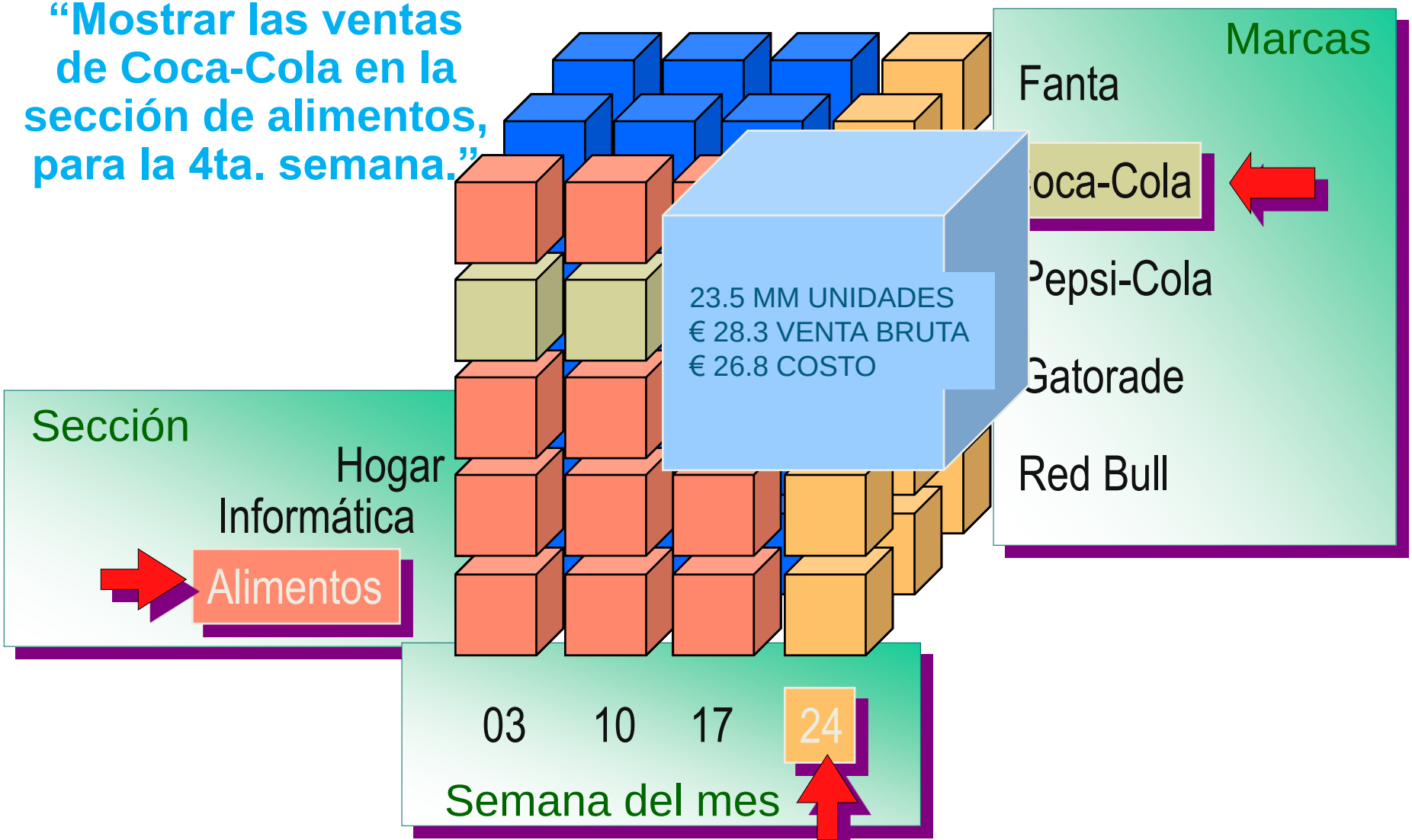
Esquema en estrella: Atributos



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Cubo Multidimensional

"Mostrar las ventas de Coca-Cola en la sección de alimentos, para la 4ta. semana."



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Herramientas para explotación del Datawarehouse

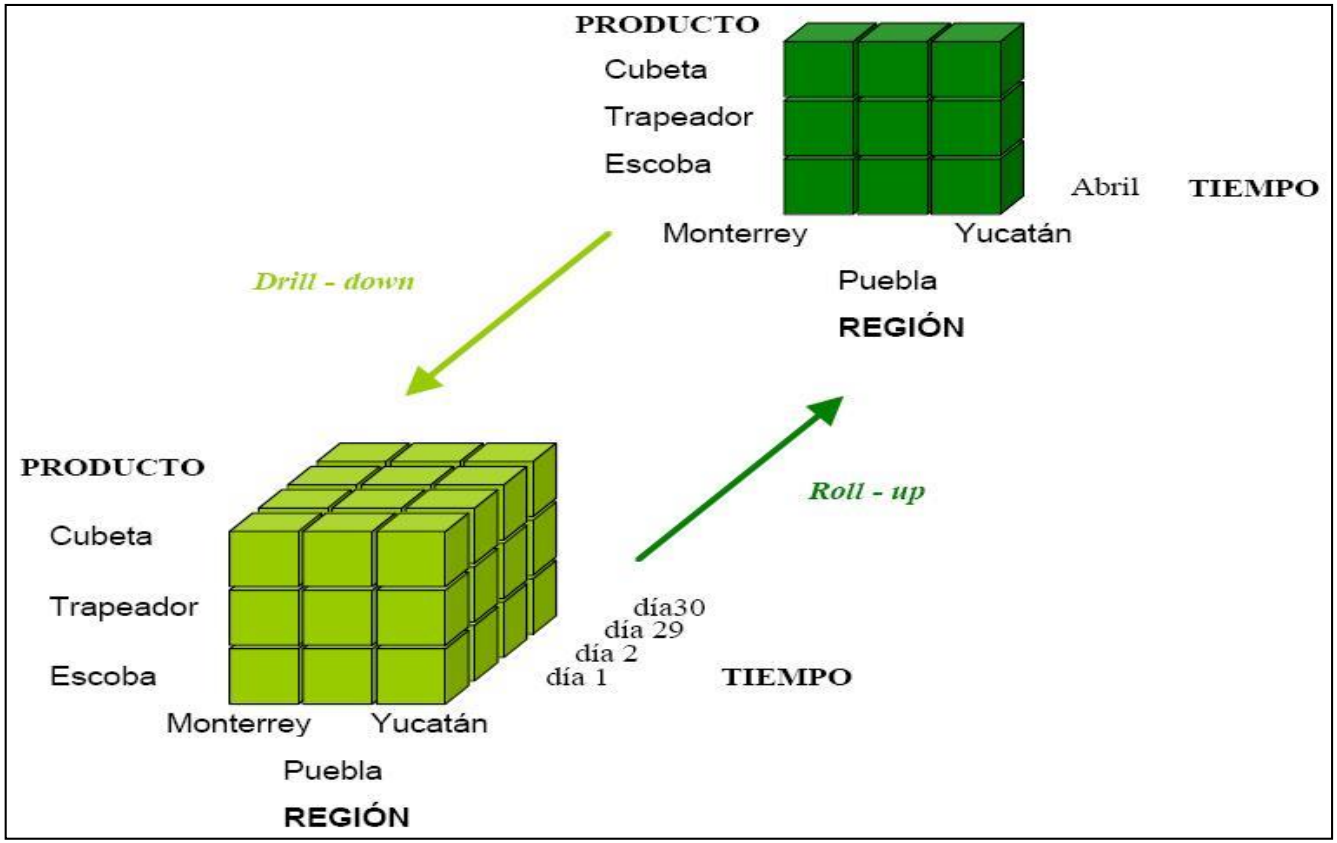
Análisis multidimensional (OLAP online analytical processing)

Facilitan el análisis de datos a través de dimensiones y jerarquías, uutilizando consultas rápidas predefinidas

Operaciones clásicas OLAP

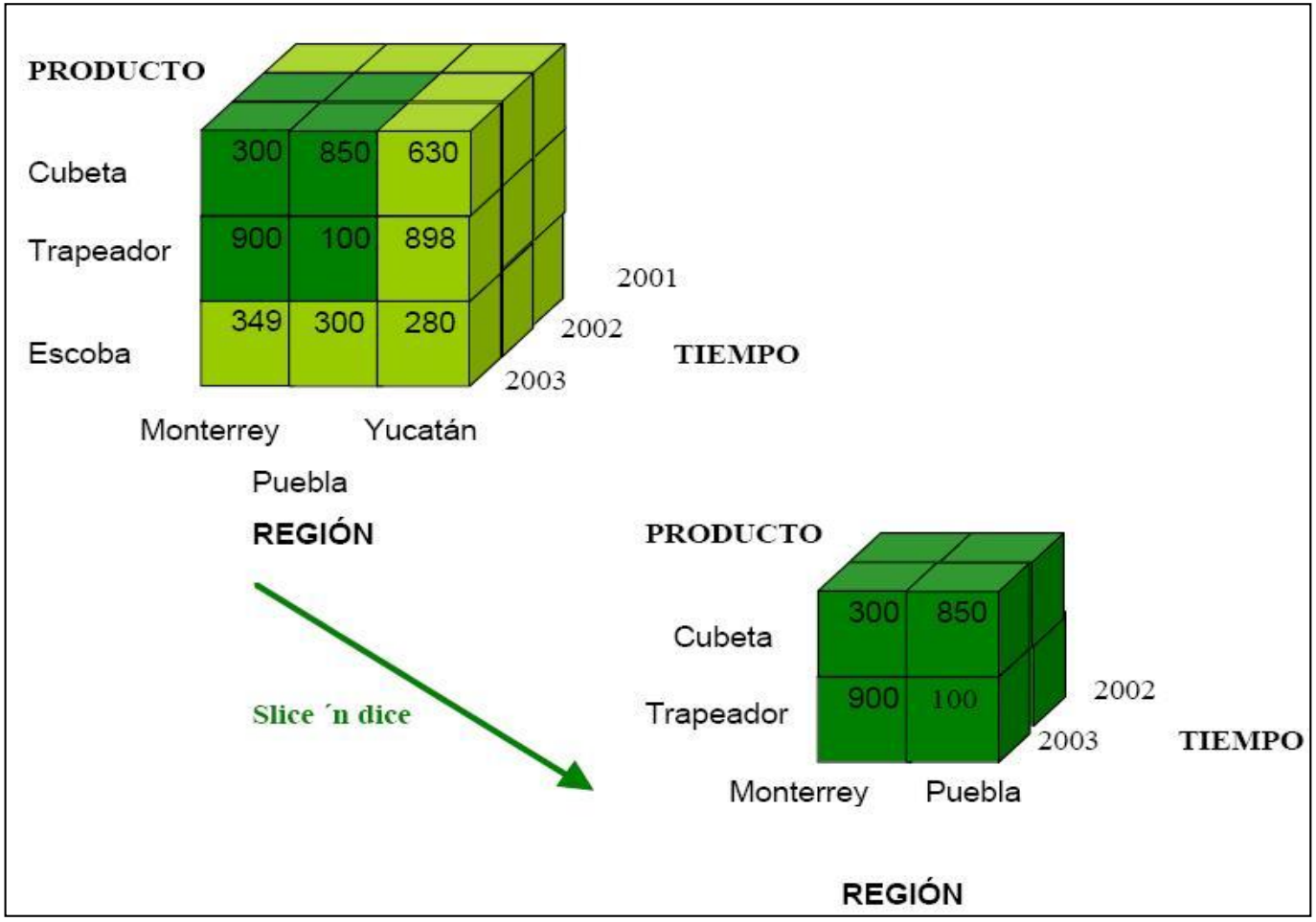
Roll up (drill-up): agrega medidas que van de un nivel Ni a un nivel más general Nj de una dimensión.

Drill down (roll down): es la operación inversa. A partir de un nivel superior este operador permitir bajar de nivel.



Operaciones clásicas OLAP

Slice and dice: permite restringir los valores asociados a una o varias dimensiones del cubo, es decir, toma un subconjunto de dimensiones y de niveles seleccionados del DW.



Otras operaciones

drill across
navegar a través de más de una tabla de hechos

drill through
navegar a través del nivel inferior del cubo a tablas relacionales

Pivote (rotar)
Rotar el cubo

Tipos de tareas de Analítica de Datos.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

¿Qué es la AD?

MODELOS de conocimiento!!!

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

¿Qué es la AD?

- **Métodos Descriptivos**

Encontrar patrones interpretable que describen los datos.

- **Métodos de Predicción**

Utilizar algunas variables para predecir los valores desconocidos o futuros de otras variables.

MODELOS!!!

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Modelos de Analítica

	Descriptivo	Predictivo	Prescriptivo
Preguntas	**Qué paso?** **Qué está pasando?** **Cuál es el problema?** **Qué acciones son necesarias?**	**Por qué esta pasando?** **Qué se producira?** **Por qué se producira?**	**Qué debería hacerse?** **Por qué debería hacerse?** **Qué pasa si se intenta eso?**
Habilitadores	• Reportes • Dashboards • Data Warehousing • Alertas	• Data Mining • Text Mining • Web/Media Mining • Forecasting	• Optimización • Simulación • Modelos de Decisión
Resultados	Bien definidos los problemas y oportunidades	Proyección de los futuros estados y condiciones	Mejores posibles decisiones y transacciones

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Modelos de Analítica

	Optimización	Identificación	Diagnóstico
Preguntas	**Qué puedo mejorar? Cómo mejorarlo?**	**Cómo es el modelo? Qué caracteriza a esos modelos?**	**Por qué sucede? Cuáles son las causas?**
Habilitadores	• Reportes • Modelos de mejora • Simulación	• Simulación • Formulas matemáticas	• Optimización • Simulación • Modelos de Decisión
Resultados	Mejores en la organización	Caracterización	Mejores posibles decisiones y transacciones

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Las estrategias analíticas básicas:

Describiendo

Factorización

Agrupación

Comparando

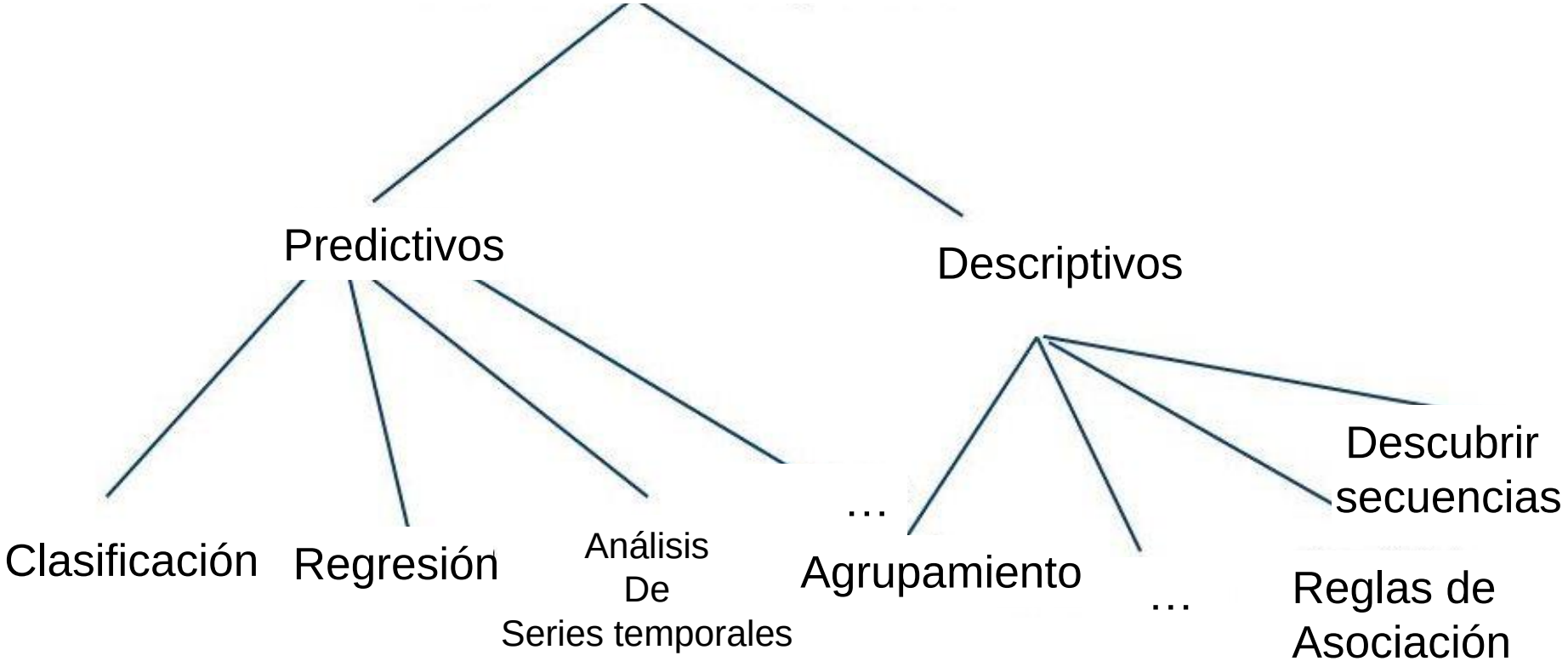
Clasificación

encontrar puntos comunes

encontrar covarianza

Descartar alternativas

Las estrategias de analíticas



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Clasificación

Examinar las características de un nuevo objeto y **asignarle** una clase o categoría de acuerdo a un conjunto de tales objetos previamente clasificados.

- Ejemplos:
  - Clasificar los estudiantes en categorías según sus rendimiento: bajo, medio y alto
  - Detectar los estados operacionales de un sistema: con falla, seguro, inactivo.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Agrupación o segmentación

Dividir una población en un número de grupos más homogéneos

- No depende de clases pre-definidas a diferencia de la clasificación

- Ejemplo:
  - Dividir la base de clientes de acuerdo con los hábitos de consumo
  - Establecer los grupos de estudiante según sus estilos de aprendizaje

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Asociación

Determinar cosas u objetos que van juntos

- Ejemplo:
  - Determinar **qué productos se adquieren conjuntamente en un supermercado**

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Pronóstico

Predecir un valor futuro con base a valores pasados

Prognosis **Predicción**

- Ejemplos:
  - Predecir cuánto efectivo requerirá un cajero automático en un fin de semana
  - Pronóstico incluye la duración esperada, la función y la descripción del curso de la enfermedad, como el declive progresivo, la crisis intermitente o una crisis repentina e impredecible.

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en
microelectrónica y sistemas distribuidos

Técnicas de Analítica de Datos

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Everything Mining

- Datos espaciales
- Espacio-temporal
- objetivos en movimiento
- datos multimedia
- flujos de datos

Minería de datos

Minería de procesos

Minería de dominios: salud, control del tráfico aéreo, alimentos, energía,

- Minería de texto
- Minería web
- Minería de la Web Semántica
- Ontología Minería
- Minería de grafos
- Datos Vinculados

Minería Semántica

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Ejemplos de Técnicas

- El análisis estadístico

Dos categorías principales:
* Estadísticas descriptivas
* Estadística inferencial

- El análisis predictivo
- La Correlación
- La Regresión
- Computación Inteligente (machine learning)

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Técnicas de Aprendizaje Automático:

- Es imposible prever todos los problemas desde el principio
- Un **sistema es inteligente** si es capaz de observar su entorno y aprender de él
- La **autentica inteligencia** reside en adaptarse, tener capacidad de integrar nuevo conocimiento, resolver nuevos problemas, aprender de los errores, etc.

Aprendizaje Automático (Machine Learning en inglés) es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender.

Crear algoritmos capaces de generalizar comportamientos y reconocer patrones.

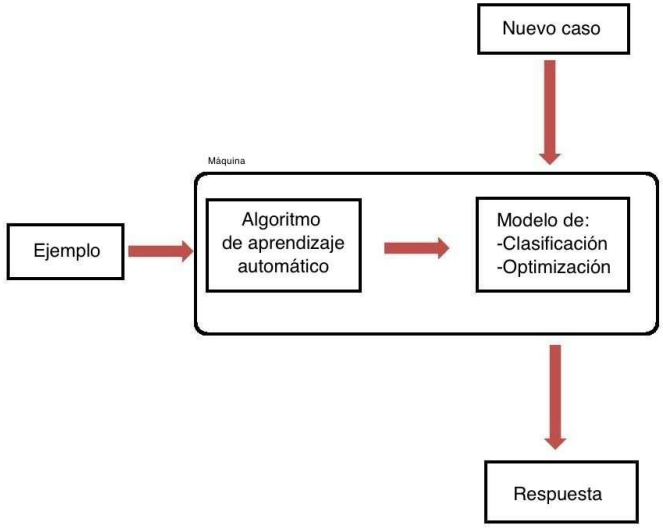
Dar a los programas la capacidad de adaptarse sin tener que ser reprogramados

Inducción del conocimiento

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

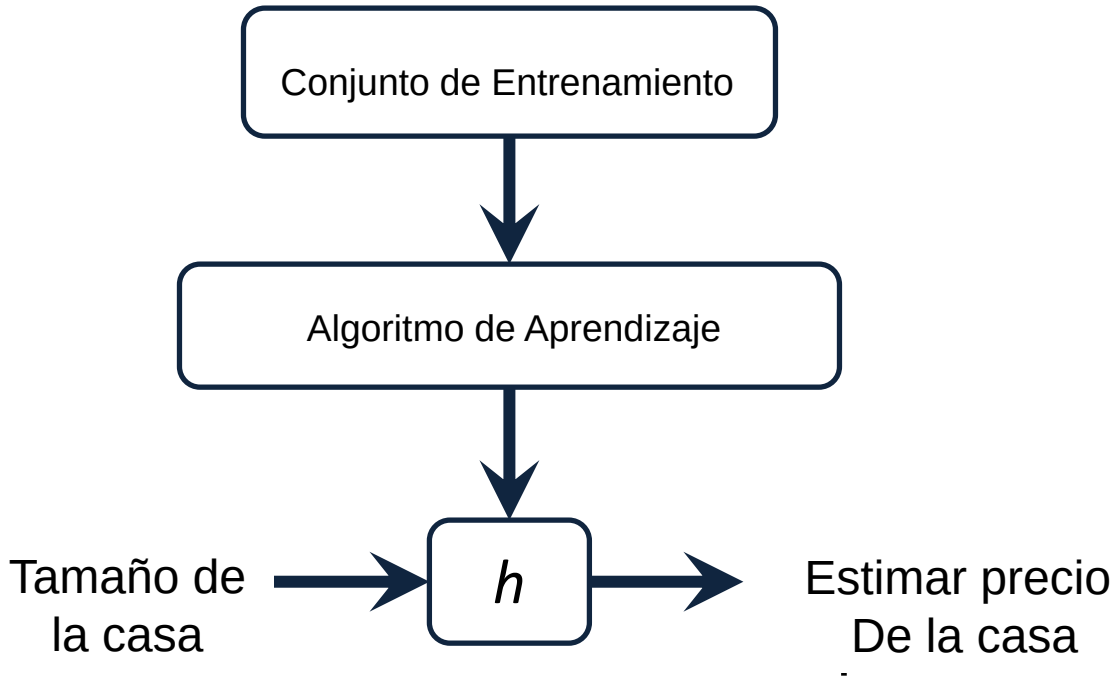
Algunas Técnicas de Aprendizaje Automático:

- Árboles de decisión,
- Reglas de asociación,
- Redes Neuronales Artificiales,
- Tablas de decisión
- Algoritmos Evolutivos
- Y muchos más (algoritmos bio-inspirados, etc.)



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Construcción de modelos



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

ALGORITMOS DE APRENDIZAJE

1. SUPERVISADOS: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. Produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema.

- Clasificación, Predicción

2. NO SUPERVISADOS: descubren patrones y tendencias en los datos, sin tener ningún tipo de conocimiento previo acerca de cuales son los patrones y categorías buscadas.

- Clustering, Análisis de enlace, Análisis de frecuencia

3. OTROS: Aprendizaje semisupervisado, Aprendizaje por refuerzo, Transducción, Aprendizaje multi-tarea, etc.

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en
microelectrónica y sistemas distribuidos

Metodologías para realizar Analítica de Datos en una organización

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

MIDANO

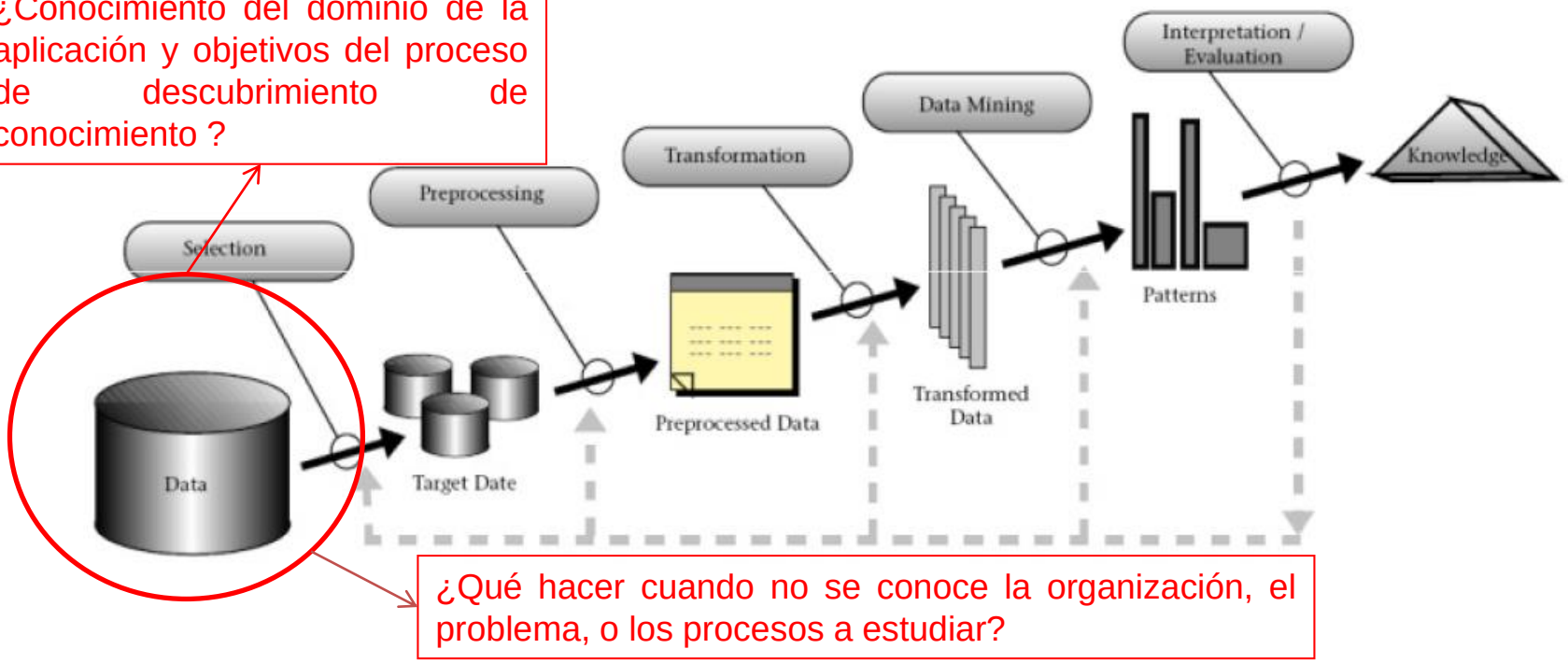
“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

Extendida para ser usado en el análisis de datos

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

MIDANO

¿Conocimiento del dominio de la aplicación y objetivos del proceso de descubrimiento de conocimiento ?



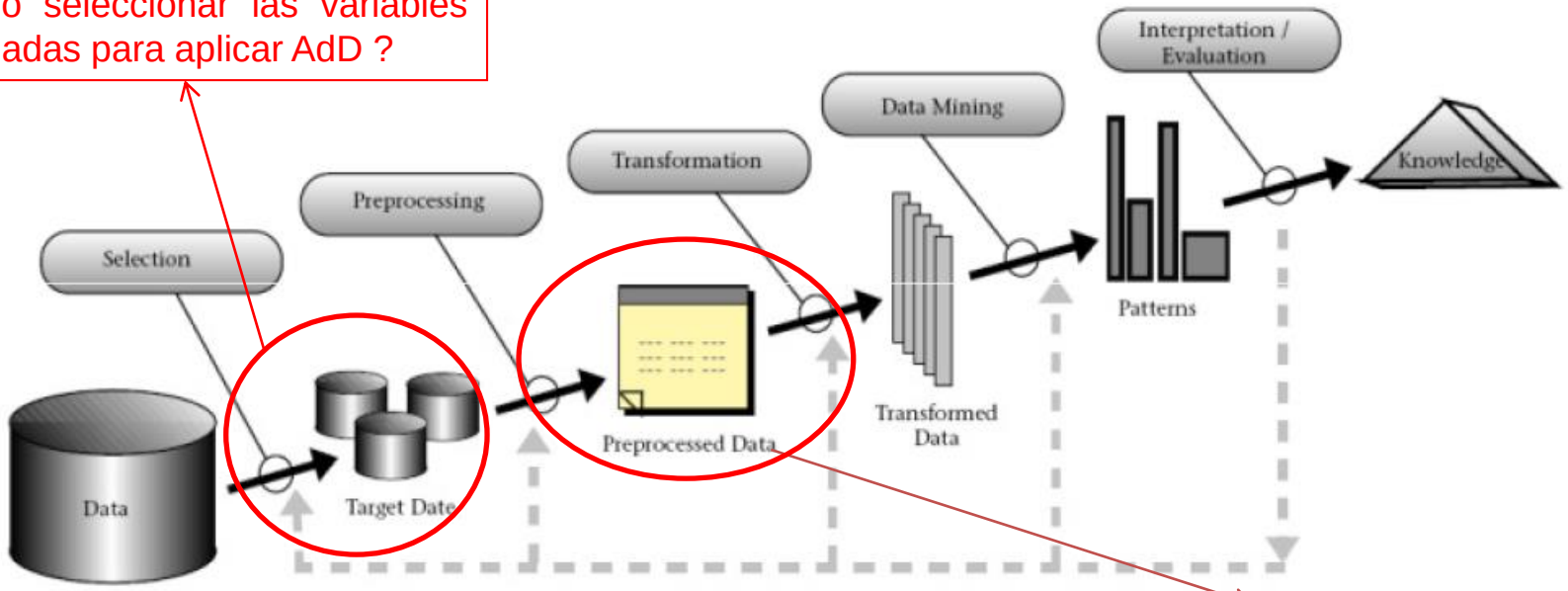
¿Qué hacer cuando no se conoce la organización, el problema, o los procesos a estudiar?

"Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional"

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

MIDANO

¿Cómo seleccionar las variables adecuadas para aplicar AdD ?



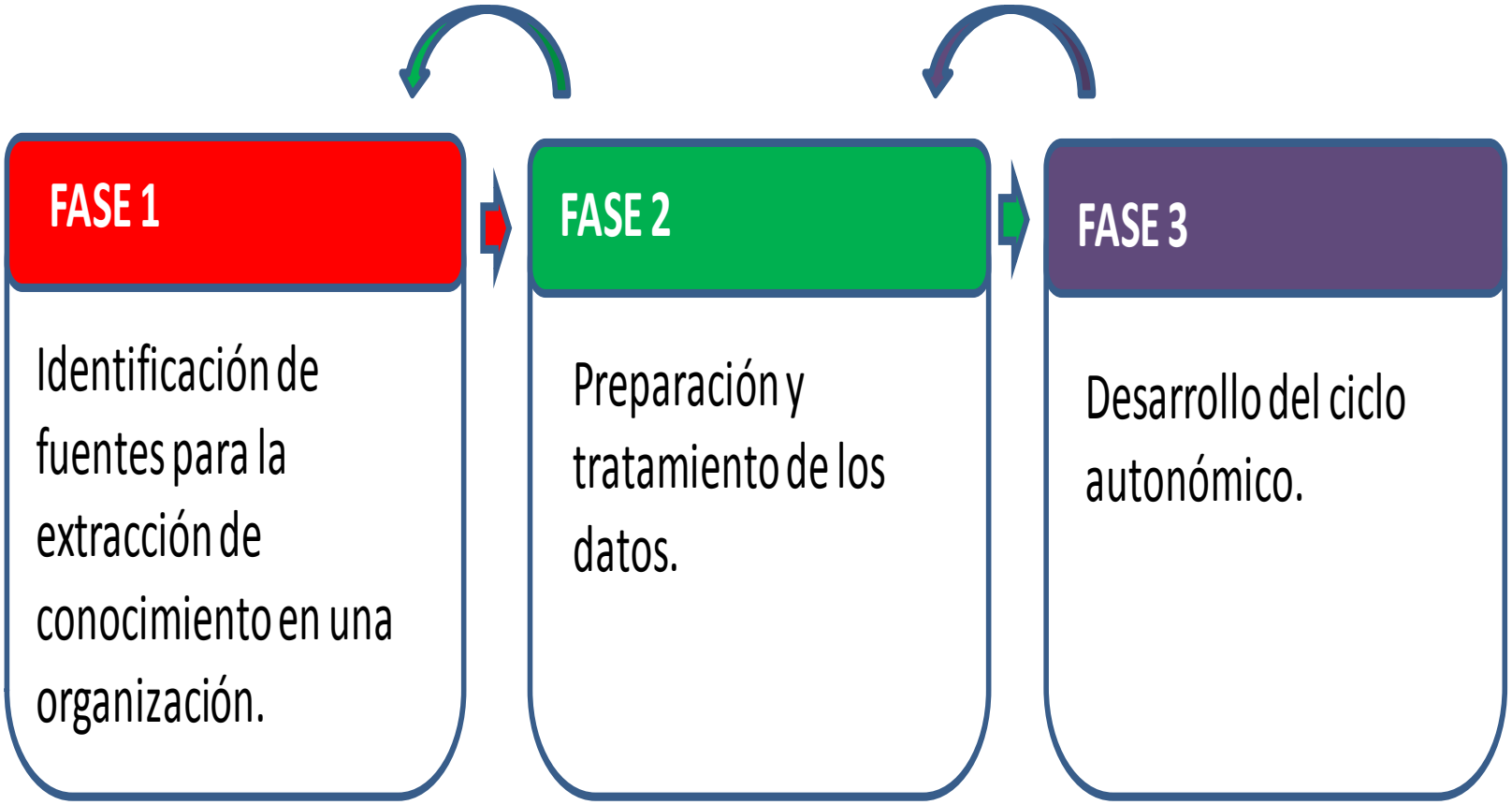
¿Cómo realizar el procesamiento de datos?

“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

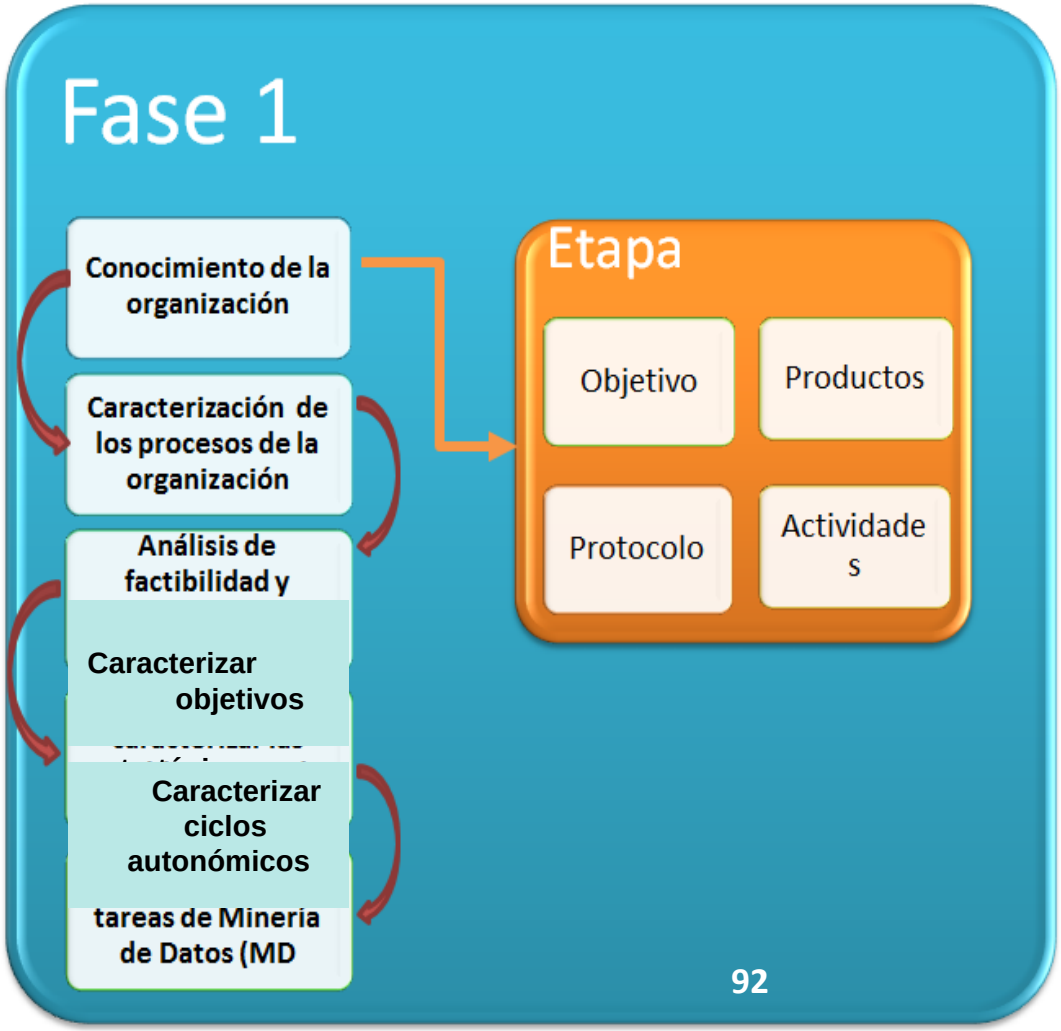
MIDANO-AdD

MIDANO consta de tres fases.



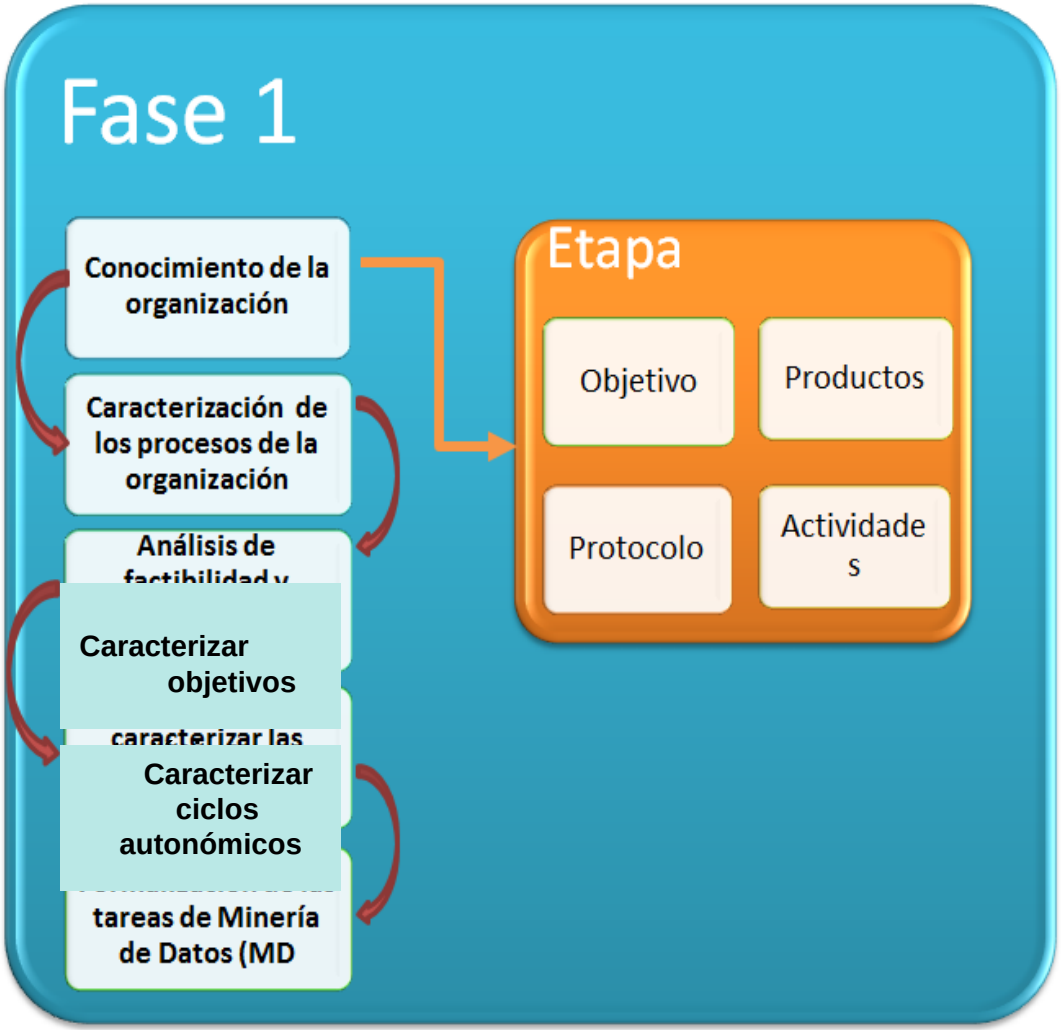
Fase 1: Identificación de fuentes para la extracción de conocimiento en una organización

Esta fase tiene como finalidad realizar un proceso de ingeniería de conocimiento, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s), o los procesos a estudiar.



Fase 1: Identificación de fuentes para la extracción de conocimiento en una organización

Esta fase tiene como finalidad realizar un **proceso de ingeniería de conocimiento**, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s), o los procesos a estudiar.



Etapa 1: Conocimiento de la Organización

1. Objetivo

- Conocer la organización/empresa, sus objetivos, procesos, objetos y actores

2. Protocolo de la Fase:

- Descripción de los elementos de la institución/empresa y sus características. Objetivos, Procesos , Objetos y Actores.
- Descripción de las relaciones entre estos elementos.
- Organización de estos elementos.

Etapa 1: Conocimiento de la Organización

Preguntas y ejemplos para determinar los elementos de la institución/empresa

Elemento	Preguntas	Ejemplos
Objetivos	¿Cuál es la razón de ser de la institución?	Conocer, determinar, establecer, la finalidad de la institución/empresa.
Procesos	¿Cuales son las actividades que permiten alcanzar los objetivos de la institución?	Procesos de producción o administrativos.
Objetos	¿Qué cosas o entidades se manipulan en los procesos de la institución?	Pueden ser físicos o abstractos, departamentos, documentos, herramientas, plantas.
Actores	¿Quiénes ejecutan los procesos?	Personas, sistemas, máquinas, etc.

Etapa 2: Caracterización detallada de los procesos de la organización

1. Objetivo

- Conocer los procesos sobre los cuales se puede enfocar el proyecto de AdD.

2. Protocolo de la Fase:

- Familiarización con los procesos sobre los cuales se puede realizar la ingeniería de conocimiento
- Identificación de la fuente de conocimiento
- Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso.

Etapa 2: Caracterización detallada de los procesos de la organización

1. Familiarización con los procesos sobre los cuales se puede realizar la extracción de conocimiento

- ¿Qué productos generan esos procesos?
- ¿Qué beneficios proporcionan esos procesos a la organización?
- ¿Qué problemas tienen actualmente?
- ¿Importancia de esos procesos para la organización, o impacto sobre otros procesos?
- ¿Qué impacto generaría la mejora de esos procesos o el estudio de los mismos?

2. Identificar la fuente del conocimiento

- ¿Cuáles son los actores o personas que intervienen en los procesos?
- ¿Quién o quiénes son las personas expertas en los procesos?
- ¿Existen documentos que permitan conocer esos procesos?
- ¿Existen sistemas computacionales que intervengan o interactúen en el proceso?

3. Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso explicado

- ¿Dónde se encuentra los datos almacenados del proceso en cuestión?
- ¿Cómo se almacenan los datos del proceso?
- ¿Qué variables son observadas del proceso?
- ¿Cuáles son las variables más importancia de esos datos para la organización?

Etapa 3: Análisis de factibilidad y selección de los procesos

1. Objetivo

- Analizar los procesos con la información proporcionada/recogida, con la finalidad de conocer la factibilidad de la aplicación de la AdD sobre cada uno de ellos

2. Protocolo de la Fase:

- Revisión de los procesos propuestos por los expertos
- Disponibilidad del experto o grupo de expertos
- Análisis de las fuentes de información sobre los procesos

Etapa 3: Selección de los Procesos

Ejemplo de Tabla para selección de procesos

Peso	Criterios	Proceso 1	Proceso 2
	Importancia para la organización		
	Interacciones entre procesos		
	Procesos dependientes		
	Importancia de la calidad del producto		
	Seguridad Industrial		
	Proposito de la tarea de AdD		
	Replicabilidad de la herramienta a desarrollar		
	Cantidad de Expertos		
	Fuentes de información		
	Confidencialidad de la información		
	¿Qué información se recoge del proceso para ser almacenada?		
	Con que frecuencia se recoge la información almacenada		
	¿Qué herramientas se cuentan, para recolectar y manipular la información?		
	Total sin ponderación		
	Total ponderado		

Criterios vinculados a la importancia del proceso para la organización

Criterios vinculados a la factibilidad de hacer una Tarea de Análitica de Datos

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

1. Objetivo

- Caracterizar las posibles objetivos estratégicos a alcanzar, con las tareas de AdD, en los procesos seleccionados

2. Protocolo de la Fase:

- Descripción de los escenarios actuales de los procesos seleccionadas en la institución/empresa.
- Especificación de los objetivos estratégicos a alcanzar en esos procesos, y posibles escenarios futuros detrás de ellos.
- Especificación de los indicadores (modelos de conocimiento, medidas estadísticas, etc.) para el análisis e interpretación de los objetivos estratégicos
- Especificación de los requerimientos para los posibles escenarios futuros (donde se puedan aplicar tarea(s) de AdD)
- Elaboración de los casos de uso para los requerimientos funcionales

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Para los procesos seleccionados

Descripción del escenario actual

Resultados que se obtienen	Actor(es) asociado(s)	Variables Asociadas	Actividades que se realizan

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Para los procesos seleccionados:

todos sus posibles escenarios futuros

Escenarios futuros deben estar orientados a lograrlos

Métricas estadísticas, modelos de conocimiento, …

Descripción del escenario futuro

Objetivos Estratégicos a alcanzar	Actor(es) asociado(s)	Variables Asociadas	Actividades de AdD que se realizarían	Funcionalidades nuevas	Resultados que se desean obtener (indicadores de logro)

Descripción del escenario futuro: < xxx >

El conjunto de escenarios futuros define una **planificación estratégica tecnológica organizacional**

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Priorización de los escenarios futuros

Criterios	Escenario 1	Escenario 2	Escenario 3
Importancia del resultado que se espera del escenario para la empresa/institución			
Utilidad del escenario para la empresa/institución			
Cantidad de expertos asociados al escenario			
Seguridad Industrial (si aplica)			
Fuentes de información requeridas por el escenario			
Confidencialidad de la información			
¿Con que frecuencia se recogen los datos almacenados asociados a la información de interés?			
¿Con qué herramientas se cuenta para recolectar y manipular los datos?			
Replicabilidad de la herramienta a desarrollar en otros escenarios			

Vinculados a los **objetivos estratégicos** y su importancia

Vinculados a los **datos**

Etapa 5: Caracterización de los ciclos autonómicos de AdD para cada Objetivo Estratégico

1. Objetivo

- Especificación de los Ciclos Autonómicos (CA) para cada escenario futuro (objetivo estratégico) priorizado

2. Protocolo de la Fase:

- Determinación de las tareas de AdD que deben caracterizar a c/ciclo por sus roles
  - Tareas de monitoreo
  - Tareas de análisis
  - Tareas de toma de decisión
- Especificación de las relaciones entre ellas
- Especificación general de las fuentes de datos requeridas por cada tarea

Especificación del Ciclo Autonómico

Objetivo: Definir un objetivo válido de supremo interés para el proceso a estudiar.

Procedimiento General

Paso 1 Tareas de Monitoreo: Se identifican, capturan, pre-procesan, las variables del proceso bajo estudio, para poder tener una **observación** clara del proceso bajo estudio

Paso 2: Tareas de análisis: Se **interpretan** las situaciones que va aconteciendo en el proceso que se está estudiando, para comprenderlo, diagnosticarlo, analizarlo, entre otras cosas.

Paso 3 : Toma de decisiones: Se definen **acciones a tomar** sobre el proceso, con el fin de alcanzar el objetivo definido para el ciclo.

Etapa 5: Caracterización de los ciclos autonómicos de AdD para cada Objetivo Estratégico

Por cada ciclo autonómico

Objetivo estratégico a alcanzar: < … >

	Nombre	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre el objetivo estratégico
Tareas de AdD de Observación				
Tareas de AdD de Análisis				
Tareas de AdD de Toma de decisión				

Métricas estadísticas, modelos de conocimiento, … que produce

Usado en el futuro como métrica de calidad del CA

Relaciones entre las tareas del CA de AdD

	Tarea AdD1	Tarea AdD2	Tarea AdD13
Tarea AdD1			
Tarea AdD2			
Tarea AdD3			

Etapa 6: Especificación de las tareas de AdD

1. Objetivo

- Caracterizar general de las tareas de AdD a realizar en los CA especificados en la fase anterior (objetivos, requerimientos, etc.).

2. Protocolo de la Fase:

- Selección y descripción de los actores y componentes necesarios para hacer cada tarea de AdD.
- Especificación de los requerimientos de c/tarea de AdD: tecnológicos, de datos, organizacionales, etc.
- Especificación de las fuentes de datos requeridas por cada tarea

Etapa 6: Especificación de las tareas de AdD

Tabla para describir tareas de AdD

Nombre de la tarea	<nombre de la tarea>
Descripción	<La finalidad de esta tarea>
Fuente de datos	<BD, historicos>
Tipo de tarea de analítica de datos	<Asociacion, Agrupamiento, Clasificacion, Predicción, reglas de asociación, etc.>
Técnicas de analítica de datos	<Define las posibles tecnicas a usar, por ejemplo: regresión, redes neuronales artificiales, algoritmo K-NN, etc.>
Tipo de Modelo de Conocimiento	<modelo descriptivo, modelo prescriptivo, modelo de optimizacion, modelo predictivo, etc.>
Tareas relacionadas de analítica de datos	<Con que otras tareas de AdD se relaciona>
Tipo de tarea del ciclo autonómico (rol)	<Pueden ser para observar, analizar/interpretar, o actuar sobre el proceso>

Etapa 6: Especificación de las tareas de AdD

Tabla para especificación detallada de las tareas de AdD

Macro-Algoritmo	Especificar Tipo de Tarea de Minería
<paso a paso del código>	< Debe indicarse de manera concreta la tarea a realizar>
…	Por ejemplo, calcular una medida de centralidad de minería de grafo, realizar un agrupamiento de tales datos según tales criterios de similitud, etc.)
…	

Esta tabla es particularmente importante para las tareas de AdDS

Fase 2: Preparación y tratamiento de los Datos

- En esta fase se plantea realizar la preparación de los datos desarrollando dos etapas. Los productos más resaltantes de esta fase son las vistas minables (conceptual y operativa) y las variables objetivos, y el modelo de datos multidimensional.

Fase 2: Preparación y tratamiento de los Datos

Para aplicar AdD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema en estudio.

Para realizar tareas de AdD es necesario tener los datos integrados en una sola vista, la cual comúnmente se conoce como *Vista Minable*. Existen dos tipos de vista minable:

- **Vista Minable Conceptual (VMC):** describe en detalle cada una de las variables a tomar en cuenta para c/tarea de AdD, en cada CA (proveniente de la primera fase de MIDANO).
- **Vista Minable Operativa (VMO):** Es el resultado de cargar los datos del historial y de realizar la etapa de tratamiento de datos, basado en la información de la VMC. La VMO se traduce a lo que se conoce como Vista Minable en la literatura, para realizar tareas de MD.

Con esas vistas se construye el modelo de datos multidimensional de c/CA

Etapa 1: Definición del modelo de datos

a. Objetivos

- Ubicar y comprender los datos asociados a cada tarea de AdD
- Construir una VMC que tenga las variables de interés para el caso de estudio
- Construir una VMO inicial
- Definir la(s) variable(s) objetivo(s) asociadas a los objetivos estratégicos o a responder con las tareas de AdD
- Definir el modelo de datos multidimensional de cada CA

b. Protocolo de la etapa

- Comprender la fuente de datos de entrada
- Generar la VMC y la VMO inicial
- Integración de los datos de entrada
- Generar las tablas del modelo de datos multidimensional de cada CA

Etapa 1: Definición del modelo de datos

VMC

Variable	Descripción	Procedencia	Observaciones

modelo de datos multidimensional (tipo estrella)

Nombre	Nombre de la tabla de hecho
Claves a las tablas de dimensiones	Todas las claves a las tablas de dimensiones
Variables Objetivos	Variables que describen o se asocian al conocimiento extraído (predicciones, etc.)
Otras variables	Variables requeridas por la tarea de AdD, por ejemplo, derivadas de operaciones de procesamiento de las dimensiones o de OLAP

Nombre	Nombre de la tabla de dimensión
Claves de la dimensión	Clave de la dimensión
Atributos de la dimensión	Atributos que describen el tema asociado a esa dimensión

Etapa 1: Definición del modelo de datos

c. Productos principales

- Documento que describe las características de los repositorios donde se encuentran los datos
- Documento que describe la VMC, la cual es presentada en una tabla descriptiva.
- Vista minable operativa (modelo)
- Archivo donde esta almacenada la VMO
- Documento que describe las características de la(s) variable(s) objetivo(s)
- Modelo de datos multidimensional de cada CA
- Modelo de datos multidimensional (Constelación) del Data Warehouse

Etapa 2: Caracterización de los datos del dominio de la aplicación

a. **Objetivos**
   - Identificación de las variables en la VMC con las operaciones de:
     - (E)xtracción, (T)ransformación y Carga (L), para el caso de datos organizacionales
     - (C)olección, (C)uración y (A)nálisis para el caso de datos externos
   - Instanciación/Alimentación de las tablas (Cargar los datos)

b. **Protocolo de la etapa**
   - Integración de los datos de entrada en el DW

c. **Productos principales**
   - Tablas ETL y CCA

Etapa 2: Caracterización de los datos del dominio de la aplicación

Tabla ETL

Variable	Extracción	Transformación	Carga
Nombre de la variable	De que fuente de datos organizacional se extraerá	Especificación del proceso de pre-procesamiento de los datos (estudios de dependencia, limpieza, cambio de formatos, etc.)	A que dimensión del modelo de datos irá

Tabla CCA

Variable	Colección	Curación	Análisis
Nombre de la variable	Identificación de fuentes externas para su obtención	Preparación de las operaciones para su obtención (limpieza, calculo, etc.)	Determinación de criterios sobre la calidad del dato (verificar si mide fenómeno deseado) y a que dimensión irá

Etapa 3: Tratamiento de datos (ciencias de los datos)

a. Objetivos

Esta etapa se centra en generar datos de calidad, es decir, sin anomalías, sin inconsistencias de formato, sin capturas erróneas, sin campos vacíos; aplicando métodos de limpieza, transformación y reducción sobre la vista minable operativa.

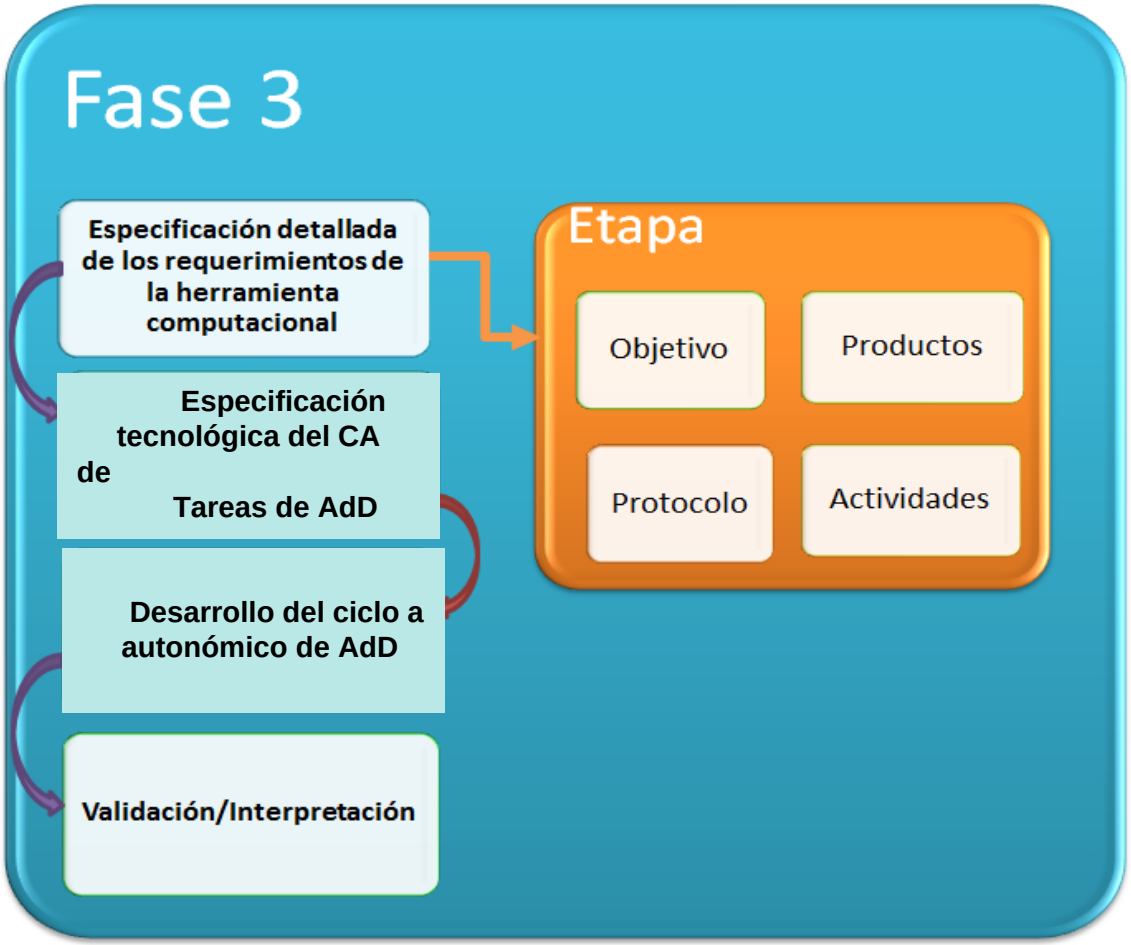
b. Protocolo de la etapa

- Limpieza
- Transformación
- Reducción
- Cálculos …

c. Productos principales

- VMO depurada
- DW implementada funcionalmente
- Documento descriptivo de los tratamientos realizados usando tablas descriptivas con información pertinente.

Fase 3: Desarrollo del ciclo autonómico de tareas de AdD



Etapa 1: Especificación detallada de los requerimientos de la herramienta computacional

a. Objetivos

captar los requerimientos no funcionales.

b. Protocolo de la etapa

- Requisitos de interfaz de usuario,
- Interfaces de software,
- Requerimientos de desempeño,
- Adicionalmente se pueden mencionar: de portabilidad, costos, rendimiento, accesibilidad, entre otros.

c. Productos principales

- Informe de requerimiento no funcionales

Etapa 2: Especificación tecnológica del ciclo autonómico de Tareas de AdD

a. Objetivos

Caracterización la implementación tecnológica del ciclo autonómico de tareas de AdD.

b. Protocolo de la etapa

- Escoger las técnicas de AdD para las tareas en el CA.
- Selección del Software para realizar c/tarea de AdD
- Definir cuáles son los datos de entrenamiento y de prueba contenidos en el DW a usar
- Definir las interfaces entre las tareas del CA
- Definir una estrategia para la validación de las técnicas seleccionada (cruzada, etc.).

c. Productos principales

- Documento con la especificación tecnológica del ciclo

Etapa 2: Especificación tecnológica del ciclo autonómico de Tareas de AdD

Tabla para especificación técnica de las tareas de AdD

Macro-Algoritmo	Especificar Tipo de Tarea de Minería	Herramienta
<paso a paso del código>	< Debe indicarse de manera concreta la tarea a realizar>	<Instrumento tecnológico a usar a utilizar para dicho calculo >
…	Por ejemplo, calcular una medida de centralidad de minería de grafo, realizar un agrupamiento de tales datos según tales criterios de similitud, etc.)	Por ejemplo, Netgraph o Netlogo para minería de grafo, o k-means para agrupamiento (indicando valor de k)
…		

Esta tabla es particularmente importante para las tareas de AdDS

Etapa 3: Desarrollo del ciclo autonómico de AdD

a. Objetivos

Realizar la herramienta de toma de decisiones usando el ciclo autonómico de tareas de AdD.

b. Protocolo de la etapa

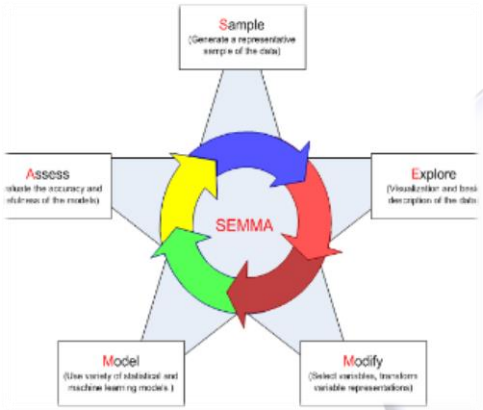
- Construcción del modelo de conocimiento generado por cada tarea de AdD
- Repetir el procedimiento de ser necesario, hasta que el modelo cumpla los errores de entrenamiento establecidos
- Integrar las tareas de AdD en el CA

c. Productos principales

- Prototipo del CA

En esta etapa, se puede usar cualquier metodología de desarrollo de tareas de MD, para desarrollar las tareas de AdD.

Etapa 3: Desarrollo del ciclo autonómico de AdD
Desarrollo de las tareas de AdD



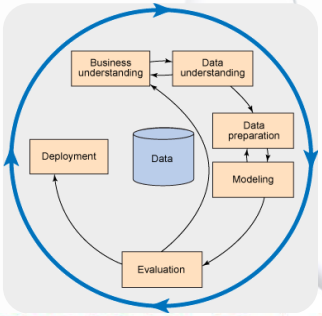
SEMMA

- Orientado a la parte técnica
- Carece de un análisis del problema.

Se puede usar cualquier metodología de desarrollo de MD para esta fase de desarrollo de tareas de AdD,

CRISP-DM

- Proceso continuo y progresivo del proceso de creación
- Más utilizado por empresas que trabajan con DM



CATALYST

- Estructura en “boxes”
- Primer Modelo: Analiza el problema.
- Segundo Modelo: Solución en el aspecto técnico.

Etapa 4: Validación/Interpretación

a. Objetivos

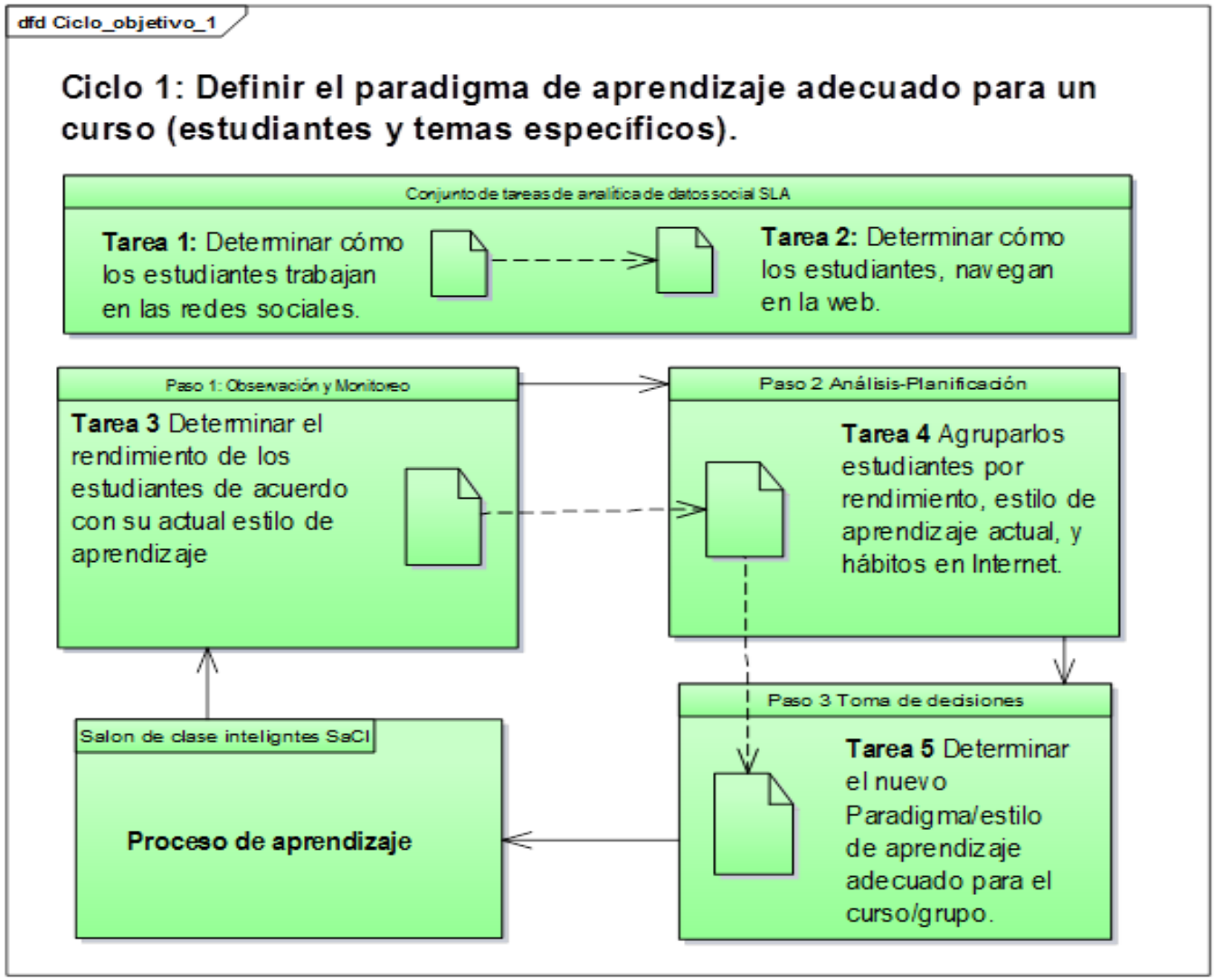
Validar la herramienta de toma de decisiones.

b. Protocolo de la etapa

- Validar el modelo de conocimiento generado por cada tarea de AdD usando los datos de prueba, y siguiendo la estrategia de validación establecida (aplicarla y observar el rendimiento).
- Realizar las correcciones necesarias
- Repetir el procedimiento de ser necesario, hasta que el modelo cumpla los errores de prueba establecidos
- Validar el comportamiento del CA, usando los criterios definidos en la etapa 1.5
- Validar el comportamiento del CA, en el sistema de toma de decisión organizacional

Ejemplos de uso de Midano

Ciclo 1: Determinar el paradigma de aprendizaje adecuado para un curso



Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 1: Conocimiento de la organización:

Se trata de una empresa que se encarga de la exploración, extracción, producción, mejoramiento y comercialización de crudo extrapesado.

Etapa 2: Caracterización de los procesos de la organización

La cadena de valor de la empresa se muestra en la siguiente figura, donde el proceso principal objeto de estudio se concentra en la tercera etapa de la cadena de valor.

Para el grupo de expertos, una de las etapas más importantes para obtener el producto deseado es la refinación, llevada a cabo en lo que se conoce como "complejo mejorador".

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 3: Selección del Proceso

Se estudió cada uno de los subproceso (objetivos, actividades, productos, etc.), y se obtuvo la interacción entre ellos.

En la tabla se ilustra este proceso de priorización y selección, considerando sólo los dos procesos que resultaron mejor ponderados en este caso de estudio.

Criterios	CDU	DCU
Importancia para la organización	5	5
Propósito de la MD	5	5
Interacciones entre procesos	2	4
Procesos dependientes	5	3
Importancia de la calidad del producto	4	4
Seguridad Industrial	4	5
Replicabilidad de la herramienta desarrollada	5	4
Cantidad de Expertos	5	5
Fuentes de información	5	5
Confidencialidad de la información	3	3
¿Qué información se recoge del proceso para ser almacenada?	5	5
Con que frecuencia se recoge la información almacenada	4	4
¿Qué herramientas se cuentan, para recolectar y manipular la información?	4	4
Total sin ponderación	56	56
Total ponderado	83	76

Descripción del escenario futuro

El escenario futuro seleccionado es para **predecir la calidad del producto y optimizar la cantidad de nafta a la salida de la columna destilador atmosférico.**

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 4: Análisis para caracterizar las posibles tareas de Minería de Datos (MD)

Descripción del escenario actual

Resultados que se obtienen	Actor(es) asociado(s)	Variables Asociadas	Actividades que se realizan
Gasoil directo (SRGO), nafta pesada y residuo atmosférica.	• Expertos asociados al proceso • Ingenieros de Procesos • Operadores • Unidad de destilación atmosférica	• Tren de precalentamiento: temperatura de la carga. • Desaladores: tiempo para el asentamiento y separación del agua del petróleo, presión. • Hornos de crudo: temperatura • Columna de crudo: presión, temperatura, rata de vapor de despojamiento.	• Carga del crudo. • Precalentamiento del crudo diluido. • Desalado. • Precalentamiento del crudo desalado. • Generación de cortes de crudo en la columna.

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Descripción del escenario futuro

Resultados que se desean obtener	Actor(es) asociado(s)	Variables Asociadas	Actividades de MD que se realizarían	Funcionalidades nuevas
Predicción de la calidad del producto, para optimizar el proceso	• Expertos asociados al proceso • Operadores • Columna de crudo	Presión, temperatura de tope y rata de vapor de despojamiento de la columna de crudo.	Predicción	• Predicción de las características del producto, según las condiciones de funcionamiento de la torre de crudo. • Ayudar a optimizar el proceso de producción, generando información para orientar a los actores en la toma de decisiones con la predicción (es) resultante(s).

Descripción del escenario futuro

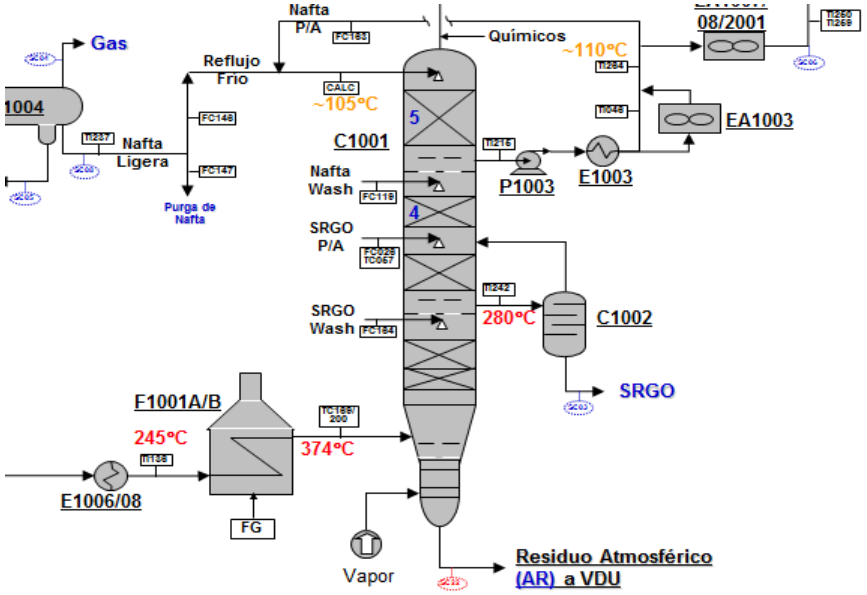
El escenario futuro seleccionado es para **predecir la calidad del producto y optimizar la cantidad de nafta a la salida de la columna destilador atmosférico.**

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Comprensión de los datos de entrada***

  A través de los diagramas de instrumentos de la planta, se determinaron cuáles son los datos asociados a las variables, se tomaron las variables más importantes asociadas al escenario futuro



Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMC***

  A partir del escenario futuro escogido, y con apoyo del grupo de expertos, se construyó una VMC con las características de la Tabla 1. Debido a que la misma cuenta con más de cien variables, sólo se presenta una pequeña muestra.

Tabla 1. Muestra de la VMC del escenario seleccionado

Variable	Descripción	Dependencia	Observaciones
11FC900	Flujo de nafta de lavado para la preflash	Identificar relación con FY119	Controlada
11PI1005	Presión de entrada de nafta wash	–	No es relevante para el estudio
11PI001A	Presión tope de la columna preflash	–	–

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMO***

  Se carga el historial en un archivo con las variables obtenidas en la VMC.

  Los datos proporcionados por la empresa fueron entregados en formato Excel, donde todos los datos están integrados en un documento menos una variable de laboratorio, ya que la misma, es tomada con una frecuencia diferente a las demás variables.

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Integración de los datos de entrada***

Fecha	11_FI158T_PNT	....	11_FC010_MEAS
01.01.2009 0:00:00	320.139504	....	39.03201294
01.01.2009 0:05:00	318.8554796		39.03201294
01.01.2009 0:10:00	315.9257853		39.03201294
01.01.2009 0:15:00	316.9394877		39.03201294
01.01.2009 0:20:00	316.2324899		39.03201294
01.01.2009 0:25:00	318.2673392		39.03201294
01.01.2009 0:30:00	311.0020414		39.03201294
01.01.2009 0:35:00	314.7039024		39.03201294
.	.	.	.
.	.	.	.
.	.	.	.

(a) Formato de la tabla de datos con las variables asociadas a los sensores de la planta

Fecha	[°API]
02.07.2008 05:00:00	45.9
03.07.2008 05:00:00	46.1
04.07.2008 05:00:00	46.1
05.07.2008 05:00:00	46.2
06.07.2008 05:00:00	46.4
07.07.2008 05:00:00	45.8
08.07.2008 05:00:00	46
09.07.2008 05:00:00	45.6
10.07.2008 05:00:00	45.1
11.07.2008 05:00:00	45.4
12.07.2008 05:00:00	45.3
13.07.2008 05:00:00	45.6
.	.
.	.
.	.

(b) Formato de la tabla de datos de gravedad API (medición de laboratorio)

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMO***

	11_FI158T_PNT	51_FT006_PNT
02.01.2009 23:30:00	314.7672055	1393.200177
02.01.2009 23:35:00	313.6730738	1396.853361
02.01.2009 23:40:00	317.3760808	1391.633283
02.01.2009 23:45:00	314.5253747	1391.253645
02.01.2009 23:50:00	315.1430386	1398.356516
02.01.2009 23:55:00	311.9205457	1400.912088
03.01.2009 0:00:00	312.5063793	1392.555884
03.01.2009 0:05:00	312.7566352	1394.478128
03.01.2009 0:10:00	312.8069345	1399.825388
03.01.2009 0:15:00	312.0659453	1401.837267

	Gravedad API a 60 °F
	[°API]
02.01.2009 05:00:00	47
03.01.2009 05:00:00	46.7
04.01.2009 05:00:00	46.8
05.01.2009 05:00:00	47.1
06.01.2009 05:00:00	48.6
07.01.2009 05:00:00	46.9

Vista minable operativa(VMO)

	11_FI158T_PNT	51_FT006_PNT	[°API]
02.01.2009 23:30:00	314.7672055	1393.200177	47
02.01.2009 23:35:00	313.6730738	1396.853361	47
02.01.2009 23:40:00	317.3760808	1391.633283	47
02.01.2009 23:45:00	314.5253747	1391.253645	47
02.01.2009 23:50:00	315.1430386	1398.356516	47
02.01.2009 23:55:00	311.9205457	1400.912088	47
03.01.2009 0:00:00	312.5063793	1392.555884	46.7
03.01.2009 0:05:00	312.7566352	1394.478128	46.7
03.01.2009 0:10:00	312.8069345	1399.825388	46.7
03.01.2009 0:15:00	312.0659453	1401.837267	46.7

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Definir las variables objetivo***

  Observar el(los) objetivo(s) de cada una de las variables en el escenario futuro seleccionado. Con el escenario futuro seleccionado se obtiene lo siguiente:

  - Escenario futuro: Producir la mayor cantidad de Nafta a 46 API
  - Funcionalidades nuevas: Predicción del API del producto, según las condiciones de funcionamiento de la torre de destilación.

Tabla 2. Variables objetivos

Variables objetivo	Observaciones
API NAFTA	Predecir el api de la nafta
11FC158	Maximizar el flujo de nafta producto a 46 api

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Limpieza***

Se ubicaron las variables con más errores en la VMO. Los resultados obtenidos son reflejados en la Tabla 3.

Tabla 3. Variables con mas errores en la VMO

PERIODO	Ene-Mar 2008	Abril-Jul 2008	Jul-Sept 2009	Oct-Dic 2010	Abril-Jul 2011	Ener-Marzo 2012	Oct-Dic 2013
NOMBRE DE LA VARIABLE	% Error	% Error	% Error	% Error	% Error	% Error	% Error
11_FC158_MEAS	0,0063593	92,43361416	100	100	100	100	100
11_FC044_MEAS	15,42766296	60,76037112	22,43088159	100	100	0,004852014	100
11_FC300_MEAS	0,0063593	18,76932921	4,478724197	100	100	100	100
11_FC119_MEAS	0,0063593	39,43158793	4,478724197	100	100	100	100
11_FC133_MEAS	99,9936407	99,7493868	100	0,0054638	14,629017772	33,915574996	0,007715454

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Limpieza***

La Tabla fue evaluada con los expertos del proceso con la finalidad de definir acciones que se podrían tomar. Las acciones tomadas son resumidas en la tabla 4, donde se describe la justificación de cada acción realizada.

Tabla 4. Acciones tomadas con las variables con más anomalías en la VMO

NOMBRE	JUSTIFICACIÓN	ACCIÓN
11_FC158_MEAS	Se puede eliminar porque es el mismo registro que el 11_FI158T_PNT	Eliminar de la VMO
11_FC044_MEAS	Se puede eliminar del estudio (es una línea de arranque o usada en operaciones muy puntuales).	Eliminar de la VMO
11_FC300_MEAS	Preferiblemente incluirla, si esta 11_FT300_PNT, se puede eliminar	Estudiar
11_FC119_MEAS	Preferiblemente incluirla, si esta FC119, se puede eliminar	Estudiar
11_FC133_MEAS	Se puede eliminar	Eliminar de la VMO

Fase 2: Preparación de datos

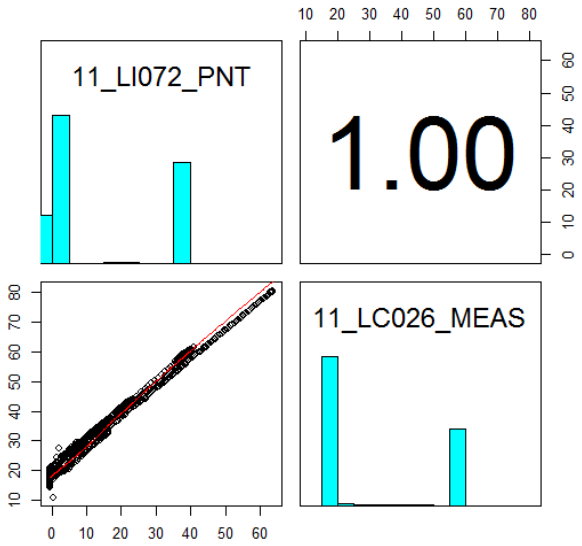
Etapa 2. Tratamiento de Datos

- ***Limpieza***

Para las variables que tienen dependencias con otras variables, se construyeron modelos lineales para sustituir los datos dañados por el resultado de estas relaciones.

Dichas dependencias pueden ser apoyadas y justificadas usando un gráfico de dispersión. En este caso, se seleccionó un diagrama de dispersión con las siguientes características:

- El histograma y nombre de cada variable (ver la diagonal de la Figura).
- La distribución de los puntos entre las dos variables y la curva regresada (parte inferior izquierda de la Figura).
- El coeficiente de correlación entre parejas de variables (parte superior derecha de la Figura).



- ***Transformación***

En este estudio no aplica el proceso de transformación, debido a que toda la data se encuentra en un formato consistente de unidades y magnitudes en las variables.

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Reducción***

Se realizaron análisis estadísticos entre variables que el experto identificó con dependencias en la VMC y además se construyó una matriz con la correlación entre todas las variables, con la finalidad de identificar las variables altamente correlacionadas.

	11_FI158T_PNT	11_FIC011_PNT	11_FC010_MEAS	11_FC159_MEAS	11SC01_BSW_NUM	11SC24_BSW_NUM
11_FI158T_PNT	1	0.368111972669728	0.484318010844852	-0.204051244869186	0.144851822147737	-0.311201934354127
11_FIC011_PNT	0.368111972669728	1	0.210678189668377	0.180986820619784	0.270251411195036	-0.067983791041037
11_FC010_MEAS	0.484318010844852	0.210678189668377	1	0.0075735692671944	-0.066805623663014	0.0505010255310379
11_FC159_MEAS	-0.204051244869186	0.180986820619784	0.0075735692671944	1	0.373898847616421	0.865146726620298
11SC01_BSW_NUM	0.144851822147737	0.270251411195036	-0.066805623663014	0.373898847616421	1	0.340144245505393
11SC24_BSW_NUM	-0.311201934354127	-0.067983791041037	0.0505010255310379	0.865146726620298	0.340144245505393	1
11_PI1001A_PNT	0.149797893568095	-0.230486068559374	0.0795532368827437	-0.530813442822542	-0.169561318779577	-0.265363516836238
11_PDI1001_PNT	-0.3513783530082	-0.064687446799048	0.0146327207575092	0.898505092755542	0.280231203226639	0.987591989101072
11_FC069_MEAS	0.836782714193593	0.429652170296929	0.412503957853951	-0.246747320588408	0.069211016952507	-0.427705464814407
11_TC167_MEAS	0.540859492533276	0.254940613271293	0.297113202094931	0.136683941679193	0.121901253006149	0.0642377397275748
11_TI168_PNT	0.528217711006828	0.251589322326238	0.294849147655103	0.16455588536907	0.131377400543795	0.0938984439071776
11_FC097_MEAS	0.863314754277501	0.4603093266029	0.453315700501899	-0.146716637712636	0.198073044334871	-0.327059949655631
11_TI205_PNT	0.465570226005517	0.231037998971888	0.128340774528971	-0.247838628436702	-0.262645490028422	-0.465553704306329
11_PI149_PNT	-0.023605088600139	0.24201692083265	0.0128921387555581	0.479588931261072	-0.030646242732858	0.183565008277804
11_PDI148_PNT	0.290003045342975	-0.090345184479632	0.0422040129610041	-0.978816278934173	-0.367572420350681	-0.911512964571754
11_FC164_MEAS	0.520628675400118	0.241198936667788	0.327886765160563	-0.021969684525066	-0.060401548555582	-0.074430658678374
11_FC114_MEAS	0.669735791599634	0.377566765068302	0.232778574068234	-0.481099372744771	0.152729477350047	-0.612534965210986
11_TI206_PNT	0.379966415482969	0.0902448084933138	0.0880606872816109	-0.68915732489563	-0.431286570684153	-0.781755090555928
11_LC016_MEAS	0.210108572264961	0.265240077290013	0.125599091378307	0.366193900688599	0.410047378697821	0.228435524378336
11_FC148_MEAS	-0.55524114311874	-0.255163977546328	-0.316329956267743	-0.083064170826386	-0.187975808467312	-0.041865764739969
11_FI149_PNT	-0.429285166720203	-0.103822072026452	-0.107453649503538	0.80189622301047	0.150822460062886	0.833467964006083
11_TI204_PNT	0.716377153818456	0.303070770590223	0.344980429883973	-0.184537891817327	0.009581408992425	-0.2871877823792

Se realizó la reducción de la VMO sobre las variables con alta correlación. La VMO inicial contaba con 97 variables, después de realizar la limpieza y reducción de datos la VMO final cuenta con 33 variables.

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Reducción***

Tabla 5. Reducción de Variables

Nombre	Resultado	Acción	Justificación
11_PI1001B_PNT	Correlación alta con 11_PI1001A_PNT	ELIMINAR	Variable altamente correlacionada, donde ambas aportan la misma información
11_PI157A_PNT/ 11_PI157B_PNT	Correlación alta con 11_PI149_PNT	ELIMINAR	Variable altamente correlacionada, donde las tres aportan la misma información
11_PI156_PNT	Correlación alta con 11_PI155_PNT	ELIMINAR	Variable altamente correlacionada, donde ambas aportan la misma información
. . .			

Se realizó la reducción de la VMO sobre las variables con alta correlación. La VMO inicial contaba con 97 variables, después de realizar la limpieza y reducción de datos la VMO final cuenta con 33 variables.

Fase 3: Desarrollo de herramientas de MD

Etapa 1: Especificación detallada de los requerimientos de la herramienta computacional

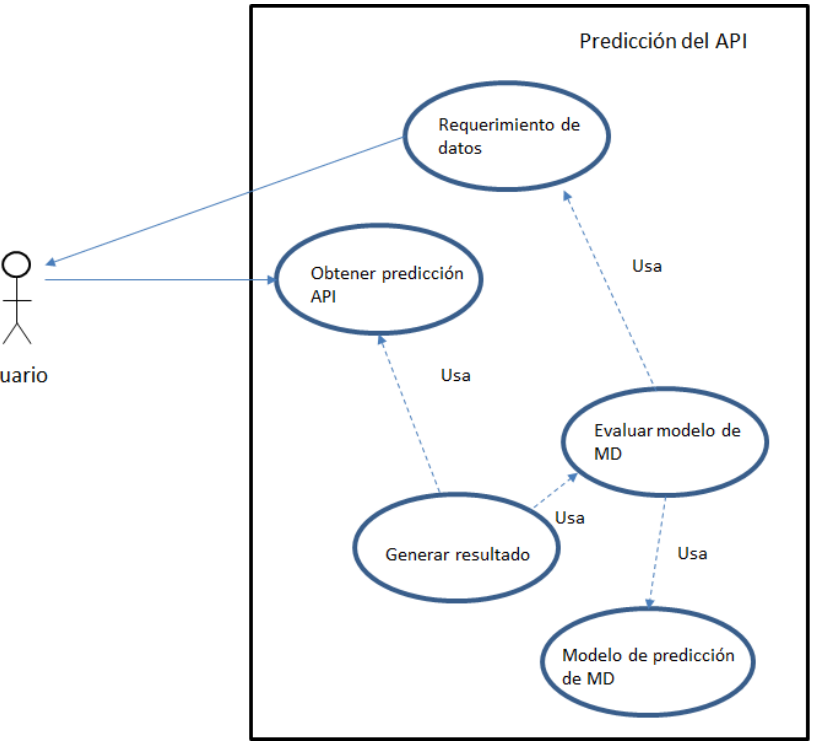
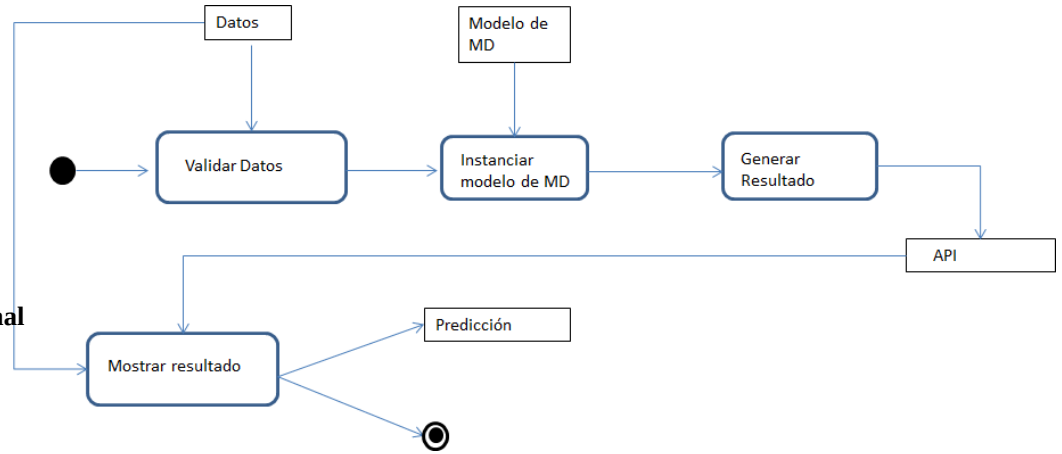


Diagrama de actividades de la herramienta computacional para la predicción del API

Entre los requisitos no funcionales:

- Requisitos de interfaz de usuario.
- Interfaces de software.
- Requerimientos de desempeño.
- Otros: de portabilidad, costos, accesibilidad, etc.

Caso de uso para la predicción del caso de estudio



Fase 3: Desarrollo de herramientas de MD

Etapa 2: Desarrollar el modelo de MD

Predicción del API de nafta:

1. Selección del Software para realizar las tareas de MD: Weka.
2. Escoger la tarea de MD para el escenario futuro: predicción.
3. Definir cuáles son los datos de entrenamiento y de prueba dispuestos en la vista minable: Los datos de entrenamiento son el 70% de los registros de la VMO y el resto será utilizado para realizar una validación cruzada[2] con Weka
4. Selección del algoritmo de MD: Los algoritmos considerados a evaluar son en general algoritmos basados en regresión lineal y redes neuronales.

Tabla 6. Algoritmos evaluados para la predicción del API de nafta

Algoritmo	Mean absolute error	Root mean squared error
LinearRegression	0.3546	0.503
RBFNetwork	0.4964	0.6953
SimpleLinearRegression	0.4422	0.6198
PaceRegression	0.3562	0.5028
IsotonicRegression	0.4293	0.5923

Validar

7. Modelo de MD: modelo que expresa las relaciones, a través de fórmulas y reglas, entre las variables del proceso. La ecuación obtenida para la predicción del API obtenida con *LinearRegression* viene dada por

API = 0.0032 * 11_FI158T_PNT + 0.0006 * 11_FIC011_PNT - 0.004* 11_FC010_MEAS -0.1673 * 11SC01_BSW_NUM +2.8329 * 11_PI149_PNT -0.1857 * 11_FC114_MEAS +0.0481 * 11_TI206_PNT - 0.0027 * 11_LC016_MEAS -0.0087 * 11_FC148_MEAS + 0.0042 * ….

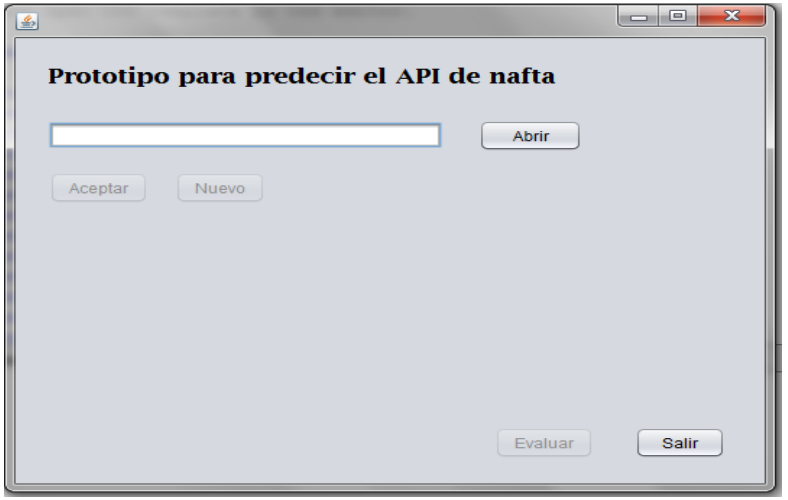
Fase 3: Desarrollo de herramientas de MD

Etapa 3: Implementación usando el modelo de MD

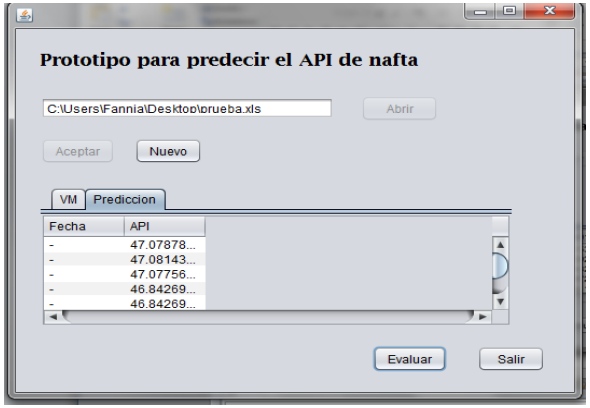
- **Formato de la entrada**

	A	B	C	D	E	F	G	H	I	J
1	Fecha	11_FI158T_PI	11_FIC011_P	11_FC010_M	11SC01_BSW	11SC24_BSW	11_TC167_M	11_TI205_PN	11_PI149_PN	11_FC114_M
2	-	303.373624	156.608638	40.409008	2.23353173	0.03146036	376.358974	370.675293	1.15274024	7.02434635
3	-	303.980151	151.507827	40.409008	2.23392856	0.03146106	376.532866	370.675293	1.15274024	7.02434635
4	-	304.417303	142.997793	40.409008	2.23432538	0.03146176	376.694006	370.675293	1.15274024	7.02434635
5	-	297.576434	141.570275	40.409008	2.23472221	0.03146246	376.845872	370.675293	1.15274024	7.02434635
6	-	301.650863	142.512228	40.409008	2.23511903	0.03146316	377.019765	370.675293	1.15274024	7.02434635
7	-	300.597413	161.583226	40.409008	2.23551586	0.03146386	377.193657	370.675293	1.15274024	7.02434635
8	-	300.120383	240.557088	40.409008	2.23591268	0.03146456	377.367549	370.675293	1.15274024	7.02434635
9	-	295.659939	244.557774	40.409008	2.23630951	0.03146526	377.541442	370.675293	1.15274024	7.02434635
10	-	296.365927	248.684239	40.409008	2.23670634	0.03146596	377.690216	370.675293	1.15274024	7.02434635
11	-	299.257376	243.036145	40.409008	2.23710316	0.03146666	377.587813	370.675293	1.15274024	7.02434635
12	-	297.409396	244.189429	40.409008	2.23749999	0.03146736	377.413921	370.675293	1.15274024	7.02434635
13	-	299.060158	243.92957	40.409008	2.23789681	0.03146806	377.240028	370.675293	1.15274024	7.02434635

Variables de la VMO



- **Salida**

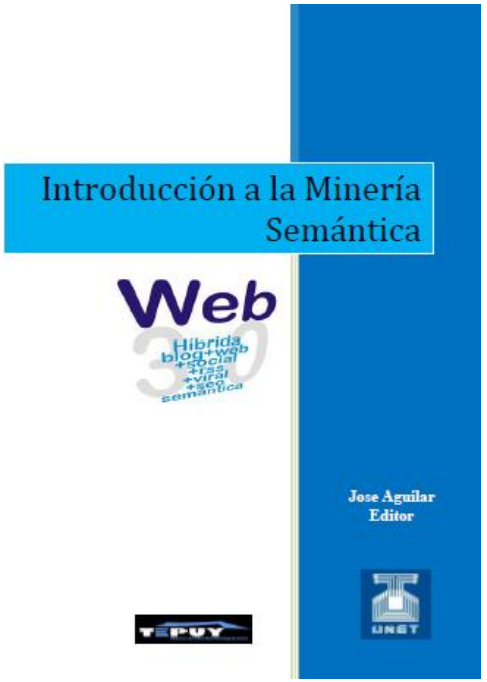


http://users.dsic.upv.es/~cferri/weka/

https://www.cs.waikato.ac.nz/ml/weka/

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

TEPUY Research+Development



"Si buscas resultados distintos,
entonces no hagas siempre lo mismo"

A. Einstein

www.ing.ula.ve/~aguilar
aguilar@ula.ve

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

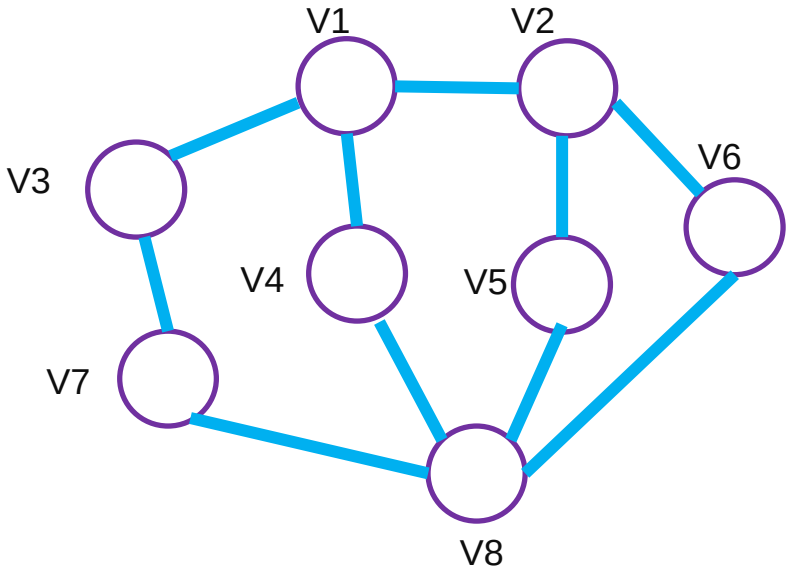
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Algunos Conceptos Vecinos: Inteligencia de Negocios, Minería (de Grafos, etc), BigData.

Grafos

Un grafo **G** es un par ordenado de un conjunto de vértices **V** y un conjunto de aristas **E**

G = (V,E)

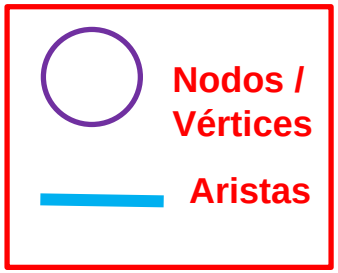


Par ordenado:

(a, b) ≠ (b,a) si a ≠ b

Par No ordenado:

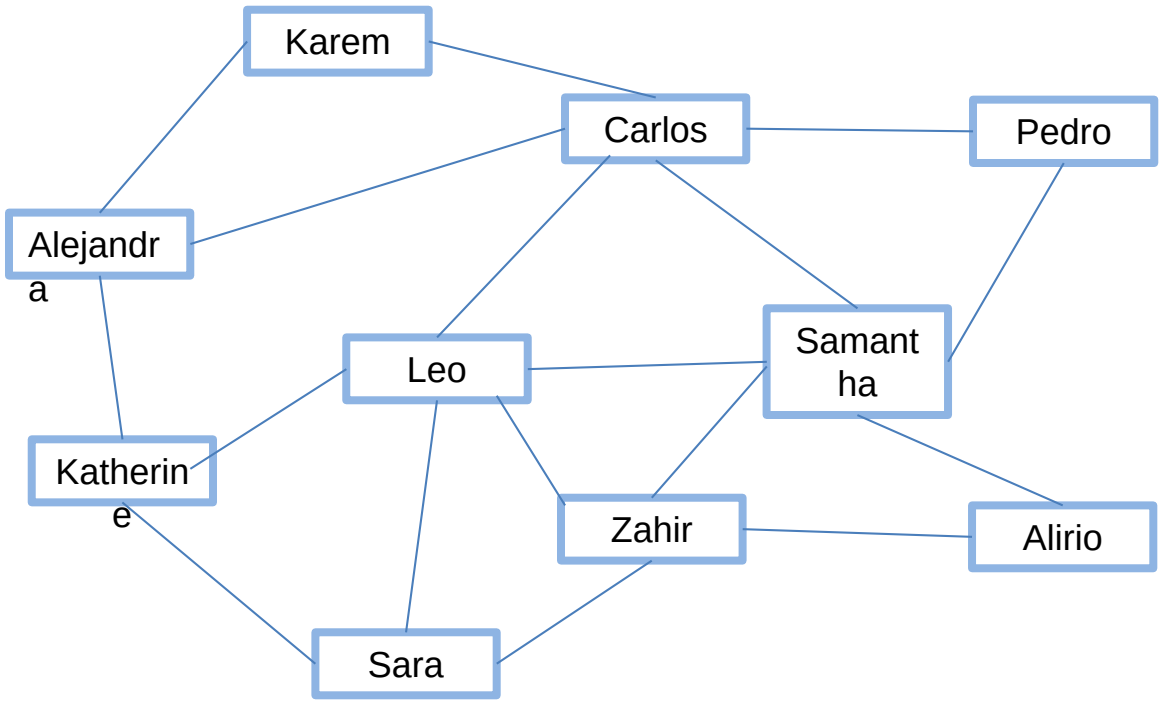
{a,b} = {b,a}



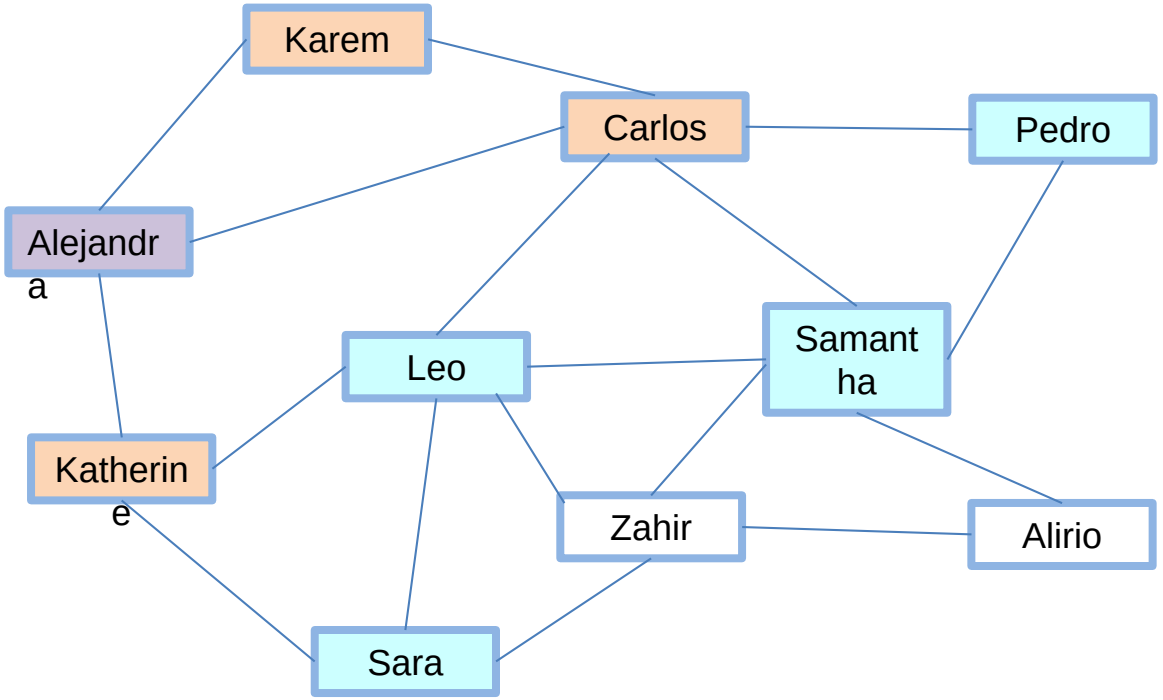
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Grafos

Red Social FACEBOOK



Grafos

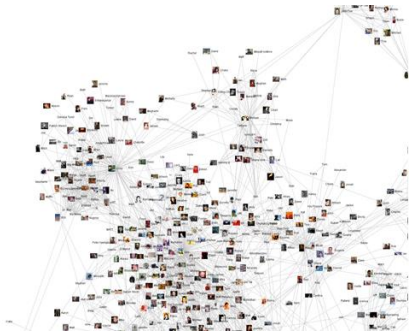


Red Social FACEBOOK

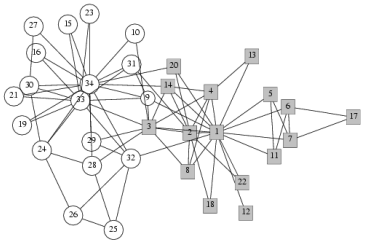
Para Sugerir un amigo a ALEJANDRA hay que encontrar todos los nodos que tengan longitud del camino igual a 2.

Redes en el mundo real

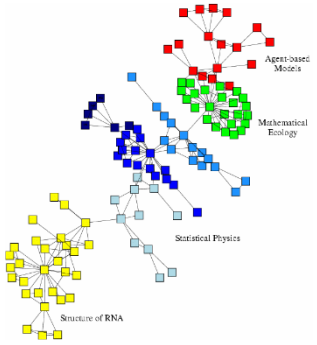
- **Redes de información:**
  - World Wide Web: hyperlinks
  - Redes de citación
  - Redes de Noticias y Blogs
- **Redes sociales**
  - Organizativas
  - Comunicativas
  - Colaborativas
  - Contactos sexuales
- **Redes tecnológicas:**
  - Energéticas
  - Transporte (aéreo, carreteras, fluviales,…)
  - Telefónicas
  - Internet
  - Sistemas Autónomos



Redes de amistad



Karate club network



Redes de colaboración

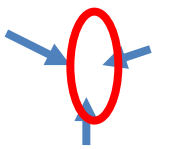
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Herramientas

- **Gephi** (visualization and basic network metrics)
- **NetLogo** (modeling network dynamics)
- **Pajek:** amplia funcionalidad basada en menús, incluyendo muchas, muchas métricas de red y manipulaciones
  - pero ... no extensible
- **Guess:** extensibles, herramientas de secuencias de comandos de análisis exploratorio de datos, pero la selección más limitada de métodos incorporados en comparación con Pajek
- **NetLogo:** plataforma general agente basado en la simulación con el apoyo de modelado excelente red
  - muchos de los demos en este curso fueron construidos con NetLogo
- **IGRAPH:** utilizado en la versión de nivel de doctorado. bibliotecas se puede acceder a través de R o Python. Rutinas escalan a millones de nodos. (for programming assignments)

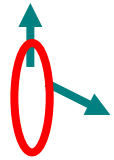
Métricas:
Propiedades de los nodos de la Red

- Conexiones
  - indegree
    cuantos arcos estan dirigidos al nodo
  - outdegree
    arcos que salen del nodo
  - degree (in or out)
    todos los arcos del nodo, entrada y salida



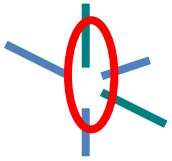
indegree=3

$$\sum_{i=1}^{n} A_{ij}$$



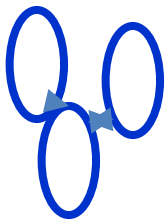
outdegree=2

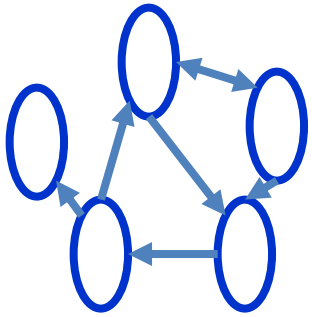
$$\sum_{j=1}^{n} A_{ij}$$



degree=5

- Degree sequence: Lista ordenada de los grados de cada nodo
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]

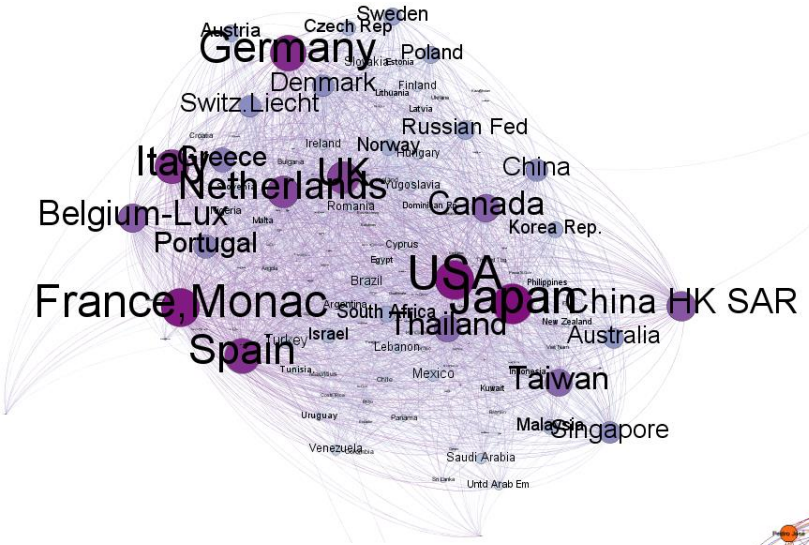




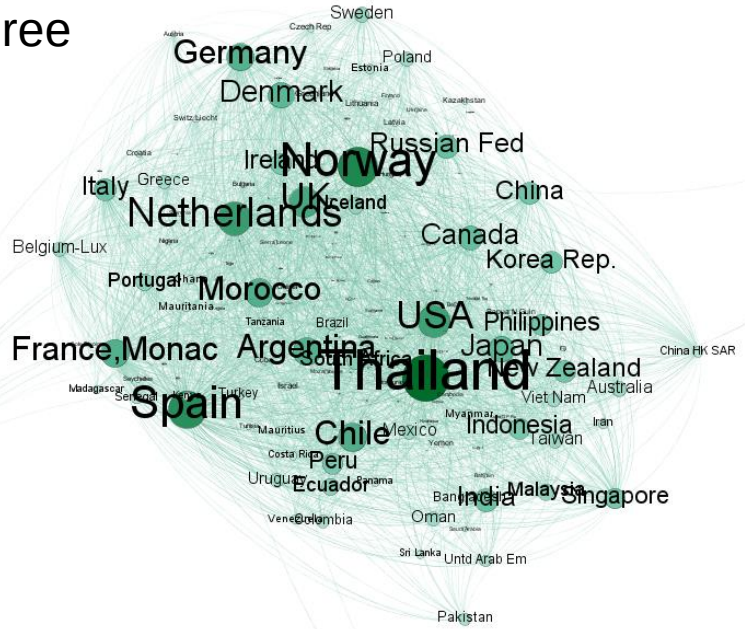
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

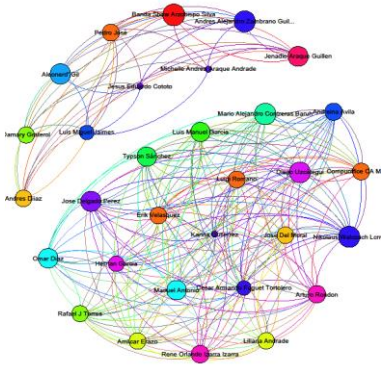
Métricas

InDegree



OutDegree





UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Métricas de redes

- **Cada métrica de red da respuesta a las siguientes preguntas:**

- pregunta: **¿Quién es más central?**

  1) METRICA DE RED: centralidad

  a)Centralidad de grado (degree centrality).
  1) Indegree o grado de entrada
  2) Outdegree o grado de salida

  b) Centralidad de cercanía (closeness centrality).

  c) Centralidad de intermediación (Betweenness centrality).

- pregunta: **¿Todo está conectado?**

  2) METRICA DE RED: los componentes conectados

  - Componentes fuertemente conectados:
  -Componentes Débilmente conectados:

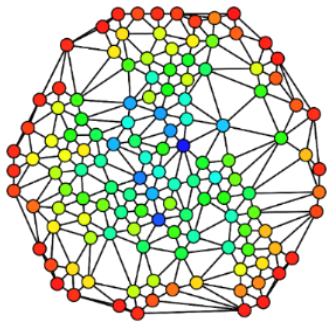
  3) METRICA DE RED: tamaño de componente gigante(giant component)

- pregunta: **¿A qué distancia están las cosas?**

  4) METRICA DE RED: rutas más cortas

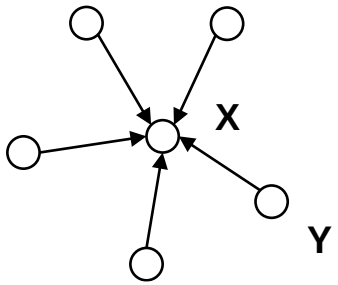
- pregunta: **¿Cómo densa son?**

  5) METRICA DE RED: densidad grafo

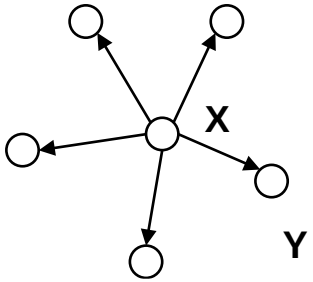


Métricas: Centralidad

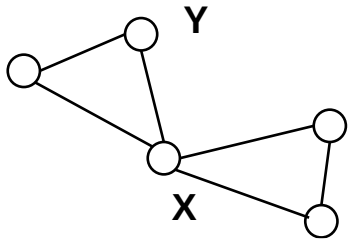
medida posible de un vértice en un grafo, que determina su importancia relativa dentro de éste



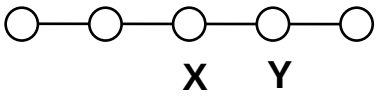
indegree



outdegree

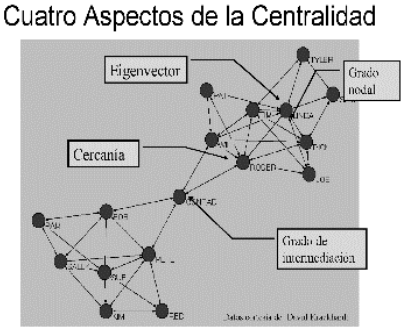


Betweenness (intermediación)



Closeness (cercanía)

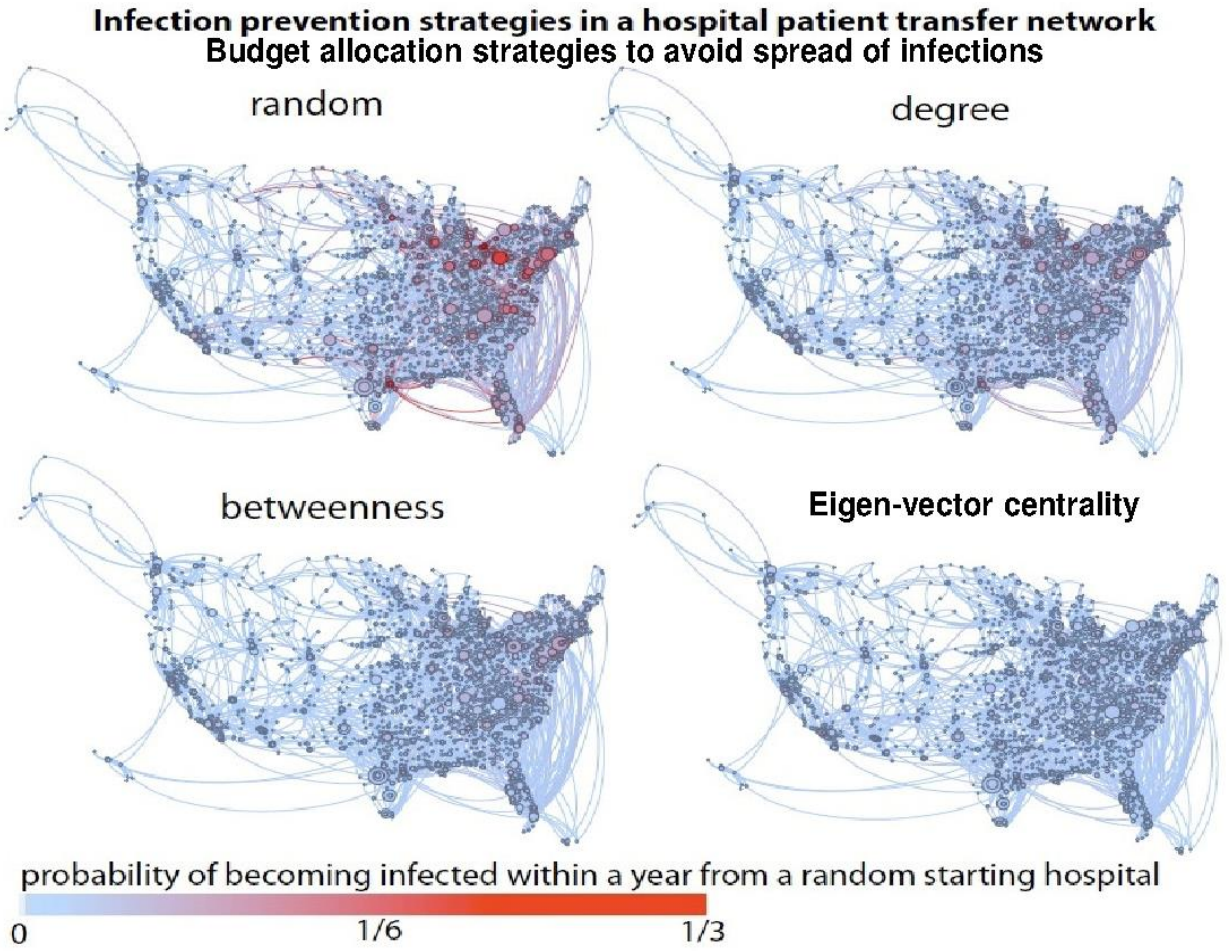
La centralidad de vector propio («eigenvector centrality»).



UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

APLICACIONES

Infecciones en hospitales EEUU por transferencia de pacientes (4)



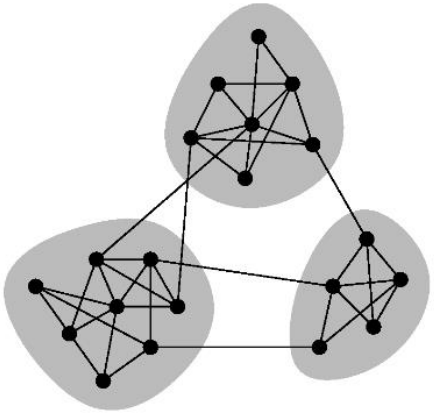
UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

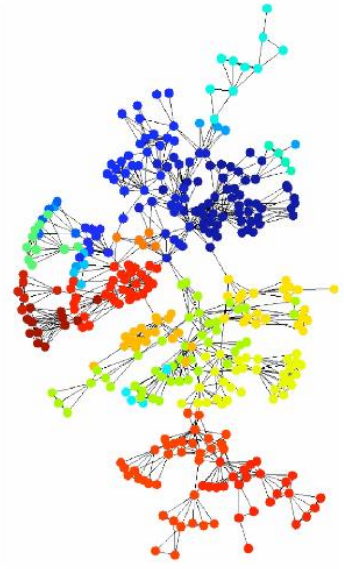
Métricas: Comunidades

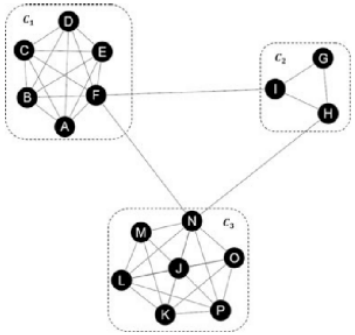
- Mutualidad
  - Cada miembro conoce a todos los miembros
- Frecuencia
  - Cada miembro conoce al menos k miembros del grupo
- Cercanía
  - Los miembros están separados por máximo de n saltos

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Comunidades - Clusters - Módulos







- En esta red, hay **tres comunidades**: C_1, C_2 y C_3
- Cada comunidad está formada por un grafo completo (un **clique**) de tamaño variable ($C_1 = K_6$, $C_2 = K_3$ y $C_3 = K_7$)
- La densidad de enlaces entre las comunidades es muy baja. Los pocos enlaces que existen son **puentes**

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Minería de Gráfos y Redes

- Minería de Patrón de Gráfo
- Modelado estadístico de Redes
- Agrupación y clasificación de grafos y redes homogéneas
- Agrupación, clasificación de las Redes heterogéneos
- Descubrimiento y Predicción de Enlace en Redes de Información
- Búsqueda de Similitud en Redes de Información
- Evolución de las redes de información social

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Minería de Grafos

Objetivo: Desarrollar algoritmos para extraer y analizar grafos.

- Búsqueda de patrones en ellos
- Búsqueda de grupos de grafos similares (clustering)
- Construcción de modelos de predicción para las grafos (clasificación)

- Aplicaciones
  - descubrimiento motivo estructural
  - reconocimiento de proteínas
  - ingeniería inversa en VLSI
  - Mucho más …

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Por qué Minería de Grafos?

- **Los grafos son ubicuos**
  - Compuestos químicos (quimio-informática)
  - Estructuras de las proteínas, las vías/redes biológicas (Bioinformactica)
  - Flujo de programas, flujo de tráfico, flujo de trabajo
  - bases de datos XML, Web, de redes sociales
- **Grafos es un modelo general**
  - Árboles, secuencias, lazos, etc.
- **Diversidad de grafos**
  - Dirigidos vs. no dirigidos, etiquetados vs. no etiquetados (arcos y vértices), ponderados, con ángulos y geometrías (topológicos en 2-D/3-D)
- **La complejidad de los algoritmos: muchos problemas son de alta complejidad**

Explosión de Datos

Air Bus A380

-1 billon de código
- cada motor genera 10 TB c/30 min

640TB por vuelo

Twitter generaba aproxim. 12 TB de datos/día

New York Stock intercambiaba 1TB de datos/día

Capacidad de almacenamiento se ha duplicado aproximadamente cada tres años desde la década de 1980

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Big Data

"Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son **demasiado grandes y difíciles de procesar** con las bases de datos y el software tradicionales" (ONU, 2012)

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Big Data

Los **grandes datos** permiten una mayor **inteligencia de negocios** mediante el **almacenamiento, el procesamiento y el análisis de datos** que se **ha ignorado** con anterioridad debido a las **limitaciones de las tecnologías tradicionales de gestión de datos**

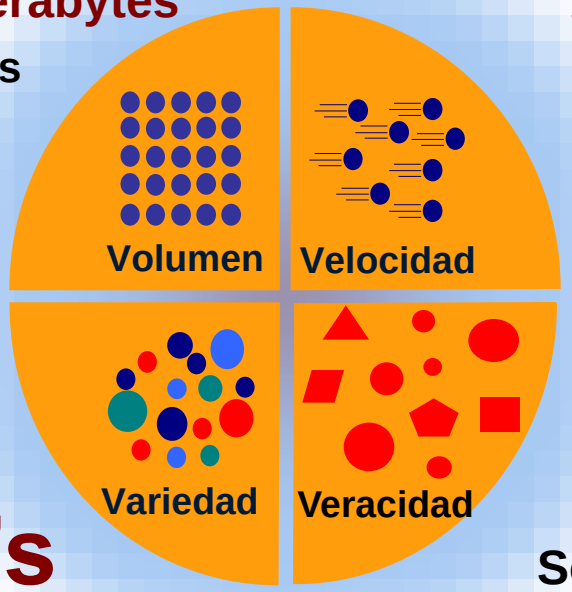
Source: *Harness the Power of Big Data: The IBM Big Data Platform*

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Big Data: Nueva Era de la Analítica

12+ terabytes
de Tweets por día

5+ million
eventos comerciales por segundo.



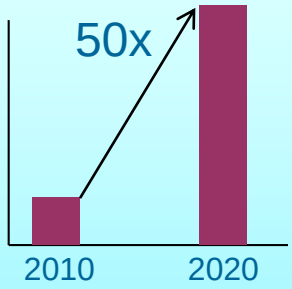
100's
de diferente tipos de datos.

Solo **1 de 3**
tomadores de decisiones confían en su información.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Características de Big Data

Eficiente procesamiento cada vez mayor de grandes **Volumenes**



35 ZB

En respuesta a la creciente **Velocidad**

30 Billones sensores RFID , etc.

Analizar la amplia **Variedad**

80% e los datos del mundo es no estructurado

Establecer la **Veracidad** de las fuentes de datos

1 de 3 líderes de negocios no confían en información que usan para tomar decisiones

AdBD

Darle **Valor** a los datos

Que se pueda tomar la mejor Decisión en base a los datos

Visualizar los datos adecuadamente

legibles y accesibles, para encontrar patrones y claves oculta

MIDANO

Garantizar **Viabilidad** del proyecto basado en datos

capacidad de una organizació para hacer un uso eficaz del gran volumen de datos

Los 5 clásicos Casos de uso en Big Data

Exploración
Encontrar, visualizar, comprender todos los grandes volúmenes de datos para mejorar la toma de decisiones

Tener una vista del cliente de 360o
Extender puntos de vista de los clientes existentes, mediante la incorporación de fuentes de información internas y externas
icionales

Extensión Inteligente de la Seguridad
minimar riesgo, detectar fraudes y supervisar la seguridad informática en tiempo real

Análisis de operaciones
Analizar una variedad de datos para mejorar resultados comerciales

Aumentar capacidades de Procesamiento de Datos
Integrar capacidades de big data y data warehouse para aumentar la eficiencia operativa