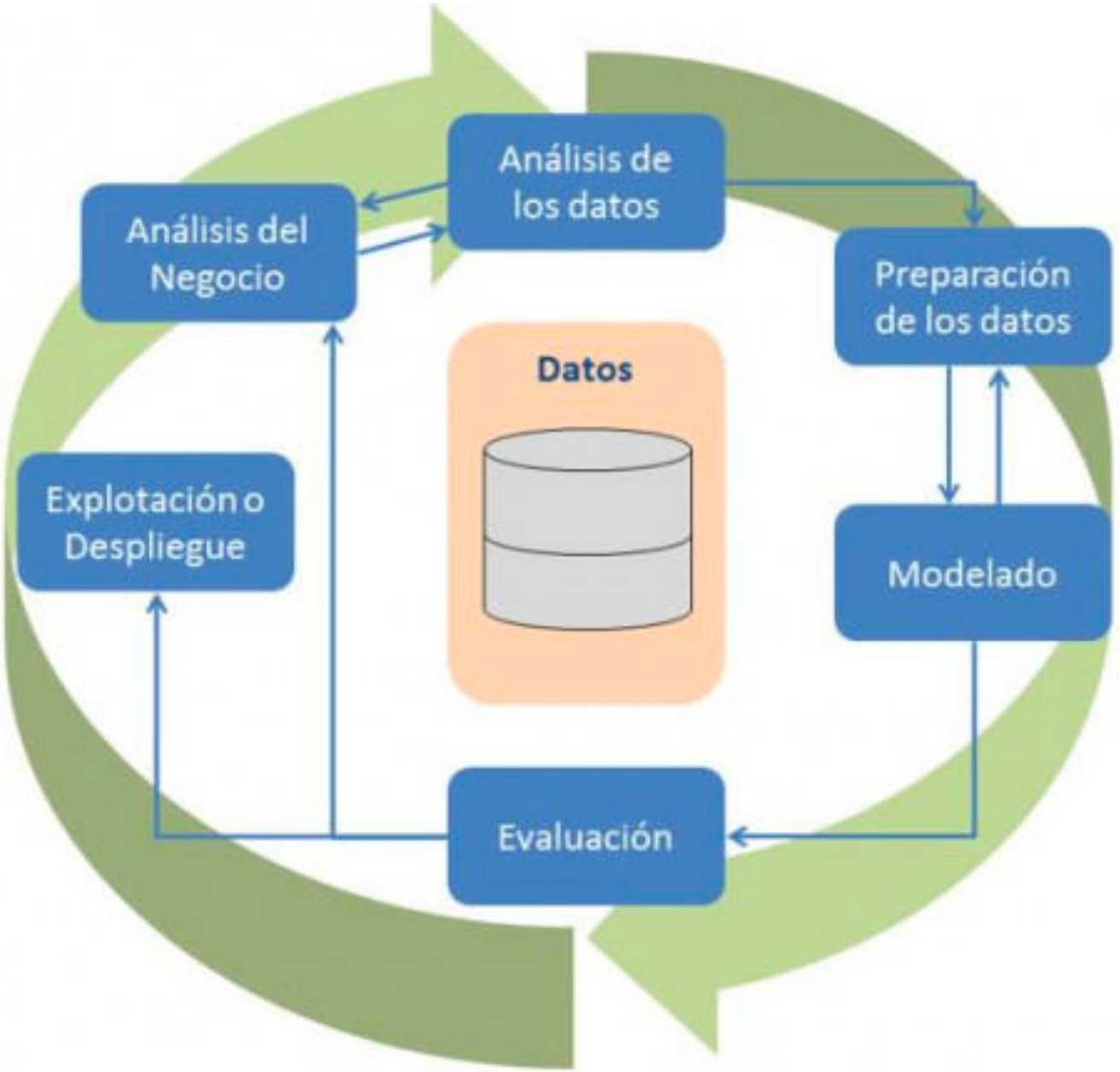
UNIVERSIDAD POLITÉCNICA DE TAPACHULA

MINERÍA DE DATOS

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

FASES DE LA MINERÍA DE DATOS



FASES DE LA MINERÍA DE DATOS

1. **Entendimiento del negocio:**
   - Formulación del problema de negocio (uno de los ya antes mencionados: previsión, gestión de riesgos, segmentación de clientes, etc).

2. **Entendimiento de los datos:**
   - Recolección de datos.

3. **Preparación de los datos:**
   - Transformación de datos: Generalmente, el formato de los datos contenidos en las fuentes de datos no es el idóneo, y la mayoría de las veces no es posible aplicar algún algoritmo de minería sobre los datos iniciales sin que requieran algún cambio (Por ejemplo, transformaciones numéricas).

FASES DE LA MINERÍA DE DATOS

3. Preparación de los datos:

- Limpieza o filtrado de datos: En esta fase se filtran los datos con el objetivo de eliminar valores erróneos o desconocidos, según las necesidades y el algoritmo a utilizar.

- Preprocesado: Se analizan las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos) y se obtienen muestras de los datos en busca de mayor velocidad y eficiencia de los algoritmos o se reduce el número de valores posibles, mediante tareas como: Redondeo, Agrupación y Agregación

FASES DE LA MINERÍA DE DATOS

4. **Modelado:** Creación del modelo.
   - Selección de variables: Después de haber sido preprocesados y realizar la limpieza de datos, se sigue teniendo una cantidad enorme de variables o atributos. La selección de características reduce el tamaño de los datos, eligiendo las variables más influyentes del problema.

   - Extracción de Conocimiento: La extracción del conocimiento es la esencia de la Minería de Datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Estos modelos se representan mediante: reglas, árboles y redes neuronales.

FASES DE LA MINERÍA DE DATOS

5. Evaluación:
Evaluación de la integridad del modelo en el negocio. Una vez obtenido el modelo, se procede a su validación; comprobando que las conclusiones obtenidas son válidas y satisfactorias.

6. Implantación:
Integración en aplicaciones para solucionar el problema de negocio expuesto.

DATAWAREHOUSE

“Es un conjunto de datos orientados por temas, integrados, variantes en el tiempo y no volátiles, que tienen por objetivo dar soporte a la toma de decisiones.” (W. H. Inmon).

“Copia de los datos transaccionales específicamente estructurada para la consulta y el análisis”. (Ralph Kimball)

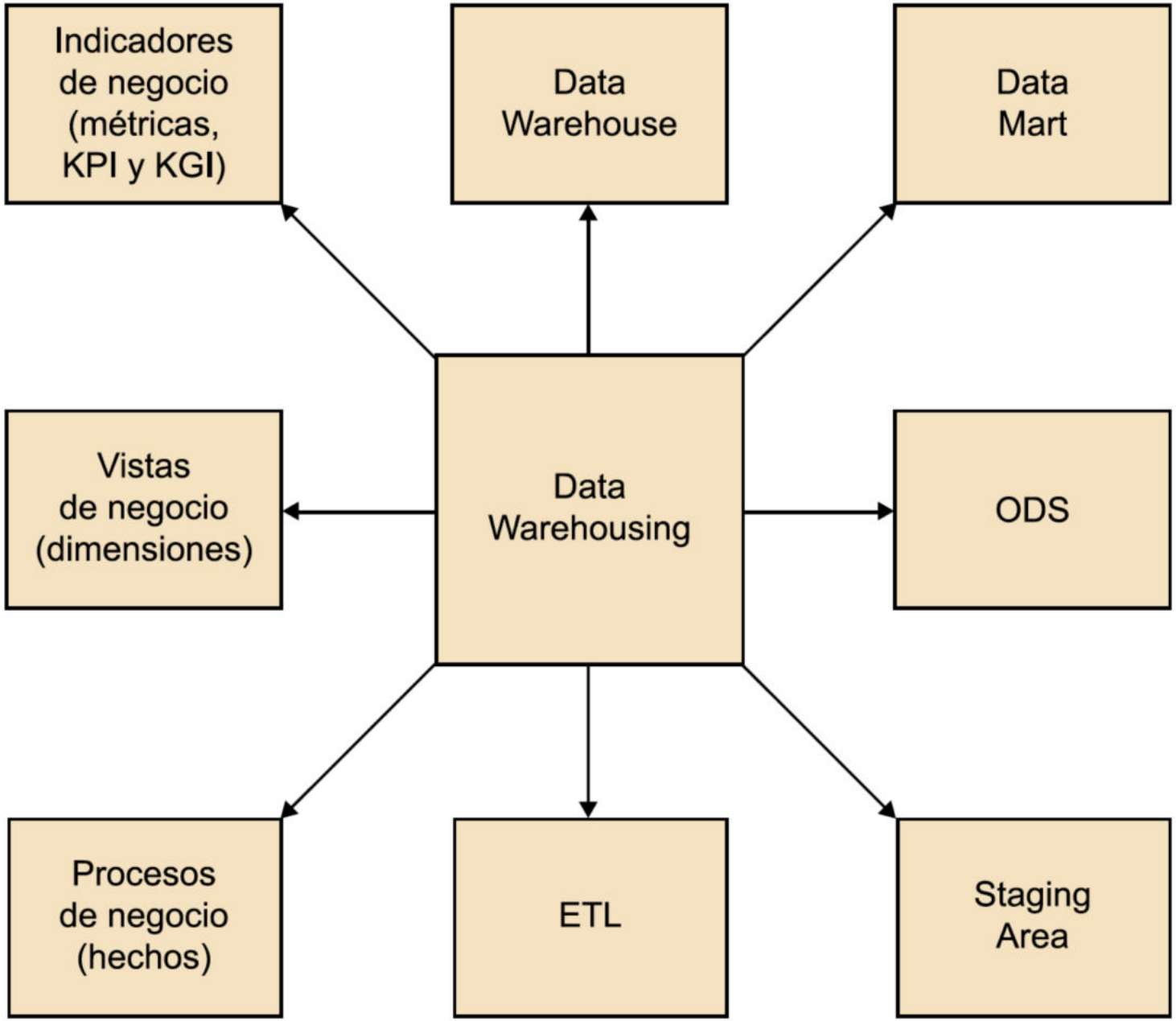
Un data warehouse es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios

Propiedades: estable, coherente, fiable y con información histórica.

DATAWAREHOUSE - CARACTERÍSTICAS

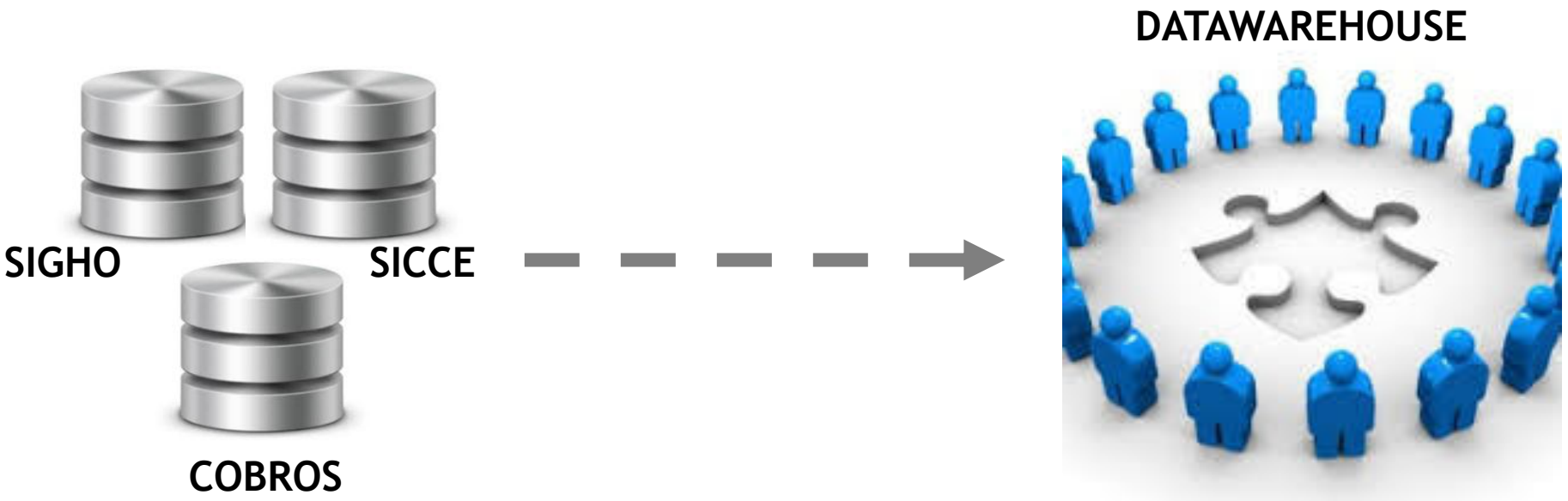
- Orientado a un tema: organiza una colección de información en torno a un tema central.
- Integrado: incluye datos de múltiples orígenes y presenta consistencia de datos.
- Variable en el tiempo: proporciona información histórica de distintos hechos de interés.
- No volátil: la información es persistente y solo de lectura para los usuarios finales.

DATAWAREHOUSE - CONTEXTO



DATAWAREHOUSE - CONTEXTO

a) Datawarehousing: es el proceso de extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información operacionales y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de la organización.



DATAWAREHOUSE - CONTEXTO

b) Datamart: es un subconjunto de los datos del data warehouse con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización.

DATAWAREHOUSE

INGRESOS

EGRESOS

CONSULTAS

DATAWAREHOUSE - CONTEXTO

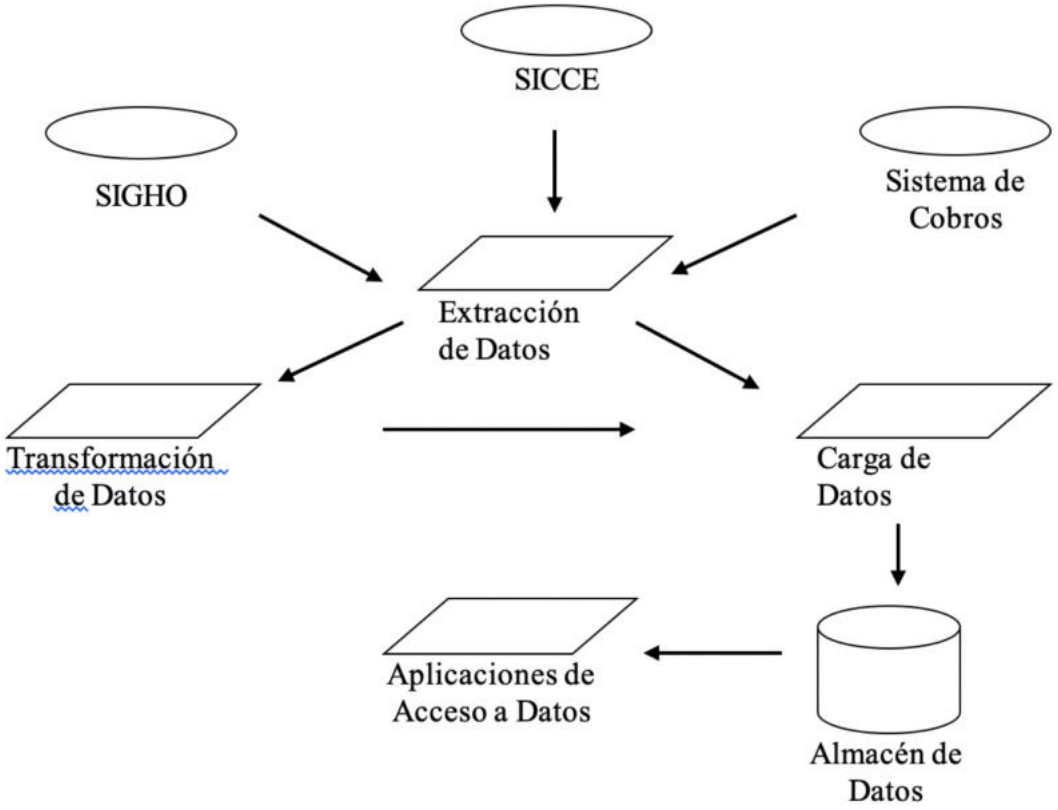
c) Operational data store: es un tipo de almacén de datos que proporciona solo los últimos valores de los datos y no su historial; además, resulta admisible generalmente un pequeño desfase o retraso sobre los datos operacionales.

d) Staging area: es el sistema que permanece entre las fuentes de datos y el data warehouse con el objetivo de:

- Facilitar la extracción de datos desde fuentes de origen con una heterogeneidad y complejidad grande.
- Mejorar la calidad de datos.
- Ser usado como caché de datos operacionales con el que posteriormente se realiza el proceso de data warehousing.
- Acceder en detalle a información no contenida en el data warehouse.

DATAWAREHOUSE - CONTEXXTO

e) Procesos ETL: tecnología de integración de datos basada en la consolidación de datos que se emplea tradicionalmente para alimentar almacenes de datos de cualquier tipo.



f) **Metadatos:** datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas.

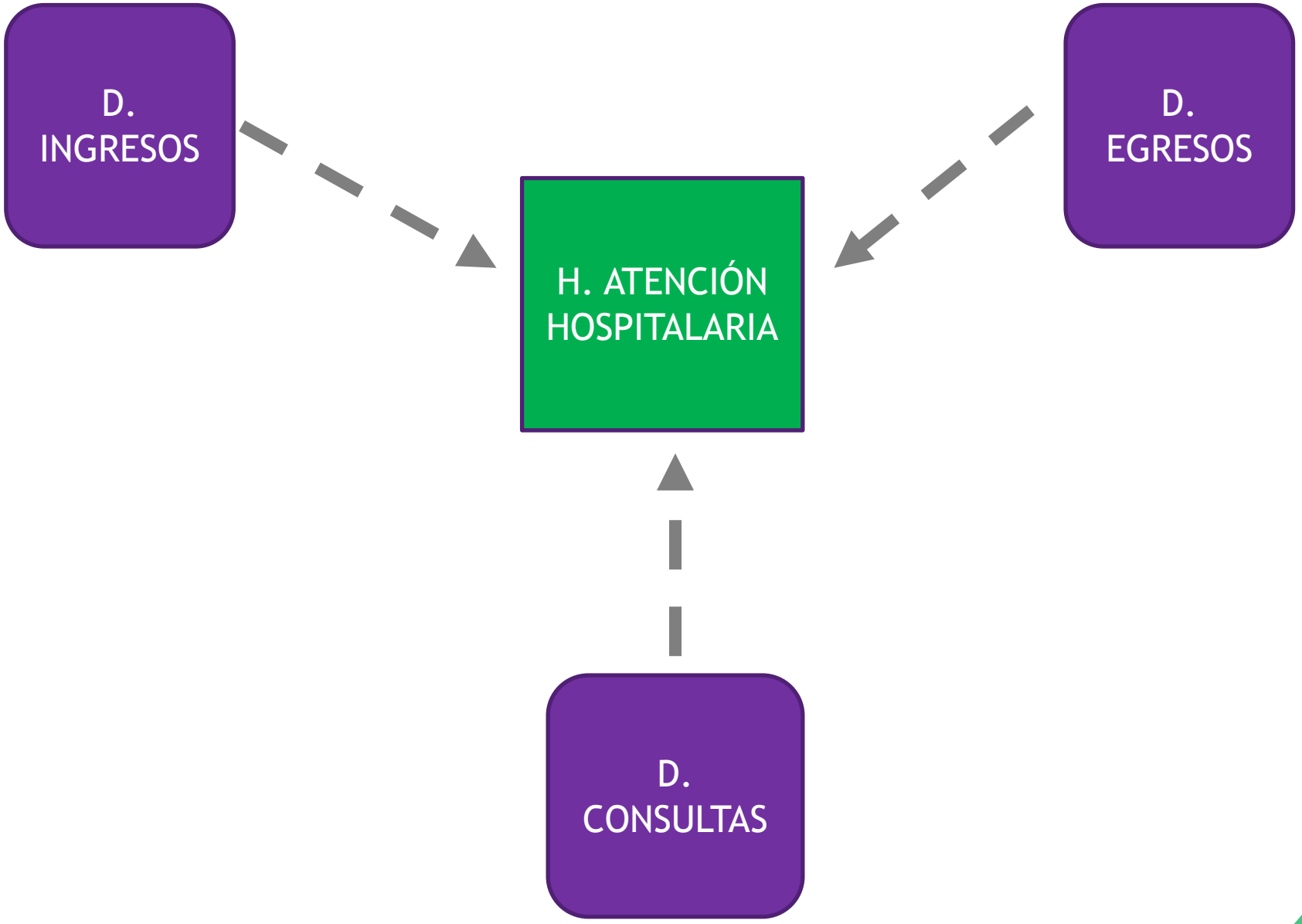
NUM	INDICADOR
1	Numero de consultas otorgadas indicadores (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
2	Numero de consultas otorgadas no indicadores(la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
3	Numero de consultas otorgadas por medico especialista (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
4	Numero de consultas otorgadas por especialidad (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
5	Promedio de consultas por hora medico (resulta de dividir el número de consultas otorgadas en un periodo de tiempo entre el número de horas asignadas en ese mismo periodo de tiempo)
6	Promedio de consultas por consultorio (resulta de dividir el número de consultas otorgadas en un periodo de tiempo entre el número de consultorios laborando en ese mismo periodo de tiempo)

DATAWAREHOUSE - ELEMENTOS

1)Tabla de hecho: es la representación en el data warehouse de los procesos de negocio de la organización. Por ejemplo, una venta puede identificarse como un proceso de negocio, de manera que es factible, si corresponde en nuestra organización, considerar la tabla de hecho ventas.

2) Dimensión: es la representación en el data warehouse de una vista para un cierto proceso de negocio. Si regresamos al ejemplo de una venta, para esta tenemos el cliente que ha comprado, la fecha en la que se ha realizado, etc. Estos conceptos pueden ser considerados como vistas para este proceso de negocio. Puede ser interesante recuperar todas las compras realizadas por un cliente. Ello nos hace entender por qué la identificamos como una dimensión.

DATAWAREHOUSE - ELEMENTOS



DATAWAREHOUSE - ELEMENTOS

3) Métrica: son los indicadores de un proceso de negocio; aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio. Por ejemplo, en una venta tenemos su importe.

NUM	INDICADOR
1	Numero de consultas otorgadas indicadores (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
2	Numero de consultas otorgadas no indicadores(la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
3	Numero de consultas otorgadas por medico especialista (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
4	Numero de consultas otorgadas por especialidad (la solicitud puede ser diaria, quincenal, bimestral, trimestral, semestral, anual)
5	Promedio de consultas por hora medico (resulta de dividir el número de consultas otorgadas en un periodo de tiempo entre el número de horas asignadas en ese mismo periodo de tiempo)
6	Promedio de consultas por consultorio (resulta de dividir el número de consultas otorgadas en un periodo de tiempo entre el número de consultorios laborando en ese mismo periodo de tiempo)

DATAWAREHOUSE - ELEMENTOS

Design Table 'consultas' in 'estadistica' on '(local)'

Column Name	Data Type	Length	Allow Nulls
id_consulta	bigint	8	✓
fecha	datetime	8	✓
espec	varchar	60	✓
medico	varchar	100	✓
hora	datetime	8	✓
consultorio	varchar	50	✓
subsec	tinyint	1	✓
sexo	char	1	✓
diagnostico	nvarchar	300	✓
municipio	varchar	50	✓
edo	varchar	50	✓
dh	varchar	50	✓
expediente	bigint	8	✓
nom_paciente	varchar	80	✓
claveDx	nvarchar	4	✓

Design Table 'egresos' in 'estadistica' on '(local)'

Column Name	Data Type	Length	Allow Nulls
id_egr	int	4	✓
fecha	datetime	8	✓
seccion	varchar	50	✓
sexo	varchar	10	✓
num_exp	varchar	10	✓
espec	varchar	50	✓
nivel	varchar	20	✓
nombre	varchar	30	✓
apellidos	varchar	30	✓

Design Table 'ingresos' in 'estadistica' on '(local)'

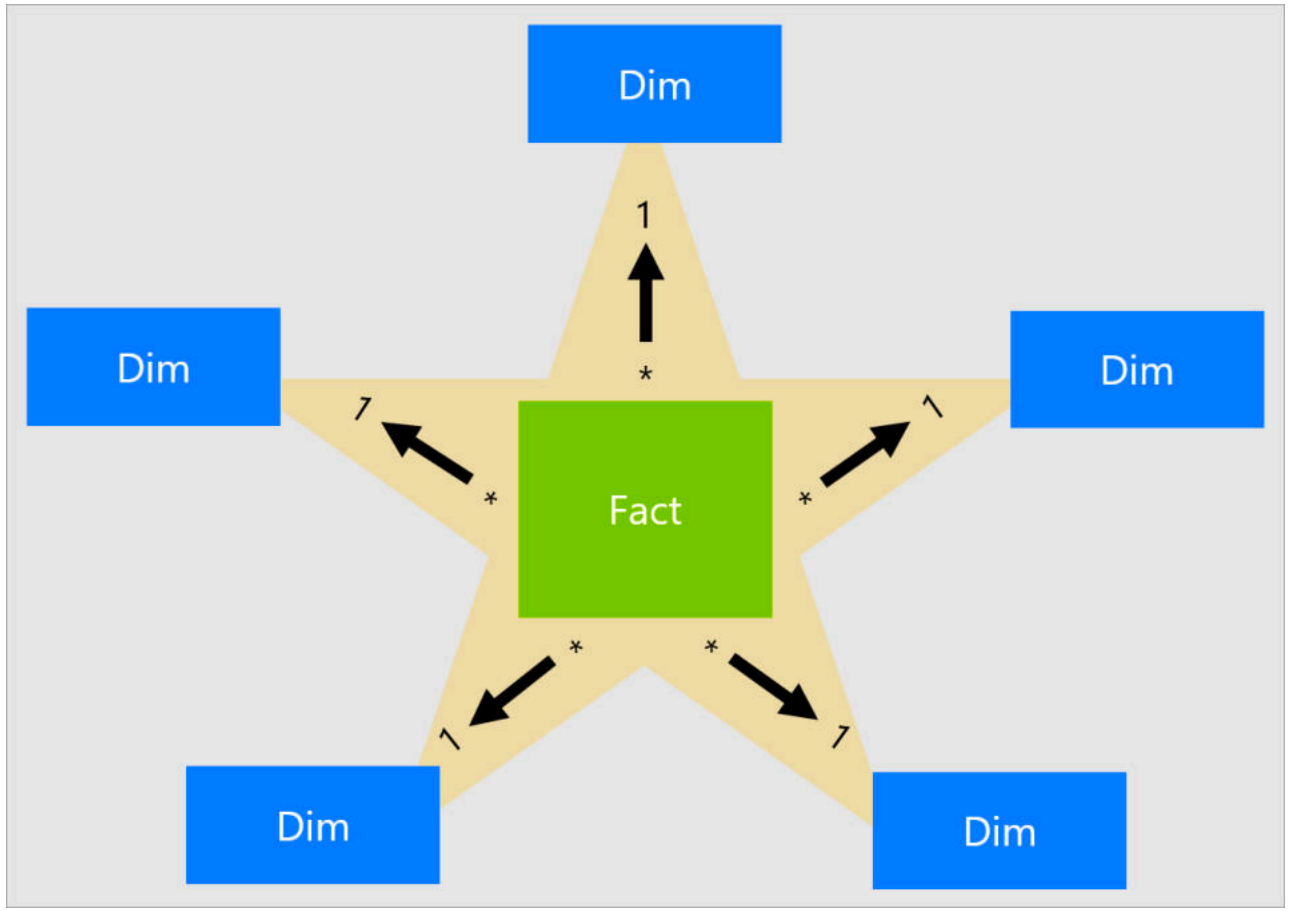
Column Name	Data Type	Length	Allow Nulls
id_ingr	int	4	✓
fecha	datetime	8	✓
seccion	varchar	50	✓
sexo	varchar	10	✓
num_exp	varchar	10	✓
espec	varchar	50	✓
nivel	varchar	20	✓
nombre	varchar	30	✓
apellidos	varchar	30	✓

DATAWAREHOUSE - ELEMENTOS

```
select distinct c.idconsulta,
cast(cast(datepart(yy,c.dfecha) as varchar(5)) + '/' +
cast(datepart(mm,c.dfecha) as varchar(5)) + '/' +
cast(datepart(dd,c.dfecha) as varchar(5)) as datetime) as fecha,
ce.cDescripcion as espec,
p.cNombre + ' ' + p.cPaterno + ' ' + p.cMaterno as medico,
cast(cast(datepart(hh,c.dfecha) as varchar(20)) + ':' +
cast(datepart(n,c.dfecha) as varchar(20)) + ':' +
cast(datepart(ss,c.dfecha) as varchar(20)) as datetime) as hora,
co.cdescripcion as consultorio,/*a.NSUBSECUENTE,*/cp.cSexo,
cc.cDescripcion as diagnostico,m.cNombre as Municipio,e.cNombre as Estado,
td.cDescripcion as DerechoHabiencia,exped.IDExpediente as Expediente,
cp.cNombre + ' ' + cp.cPaterno + ' ' + cp.cMaterno as paciente,
cc.IDDiagnostico as IdDiagnostico, dd.nPrimeraVez as PrimeraVez,
substring(c.cMotivo,1,6) as Motivo, substring(c.cPadecimiento,1,6) as Padecimiento,
substring(c.cPlanTer_ResEsp,1,6) as Plan_Terapeutico, dc.cValor
from Consultas c INNER JOIN Personal p ON p.IDPersonal=c.IDMedico
INNER JOIN Det_Personal_Especialidad dpe ON dpe.IDPersonal=p.IDPersonal
INNER JOIN ctl_Especialidades ce ON ce.IDEspecialidad = dpe.IDEspecialidad
LEFT JOIN hgc_agenda a ON a.cFolioAgenda=c.IdFolioAgenda
LEFT JOIN ctl_consultorios co ON co.CCVE_CONSULTORIO=a.CCVE_CONSULTORIO
                        and co.IDUMedica=a.IDUMedica
INNER JOIN CTL_Pacientes cp ON cp.IDPaciente=c.IDPaciente
INNER JOIN Det_DiagnosticosConsulta dd ON dd.IDConsulta=c.IDConsulta
INNER JOIN Ctl_CIE10 cc ON cc.IDDiagnostico=dd.IDDiagnostico
INNER JOIN Det_DomicilioPacientes dp ON dp.IDPaciente=cp.IDPaciente
INNER JOIN Ctl_Municipios m ON m.IDMunicipio=dp.IDMunicipio
                        and m.IDEstado=dp.IDEstado and m.IDJurisdiccion=dp.IDJurisdiccion
INNER JOIN Ctl_Estados e ON e.IDEstado=m.IDEstado
INNER JOIN Det_Pacientes_Derechohabiencia pd ON pd.IDPaciente=cp.IDPaciente
INNER JOIN Ctl_TipoDerechohabiencia td
                        ON td.IDTipoDerechohabiente=pd.IDTipoDerechohabiente
INNER JOIN Expedientes exped ON exped.IDPaciente=cp.IDPaciente
LEFT JOIN Det_Consulta dc ON c.IDConsulta=dc.IDConsulta
where c.dFecha between '2021/05/01' and '2021/05/23' and dc.iddiccionario=6
order by c.idconsulta
```

DATAWAREHOUSE – TIPOS ESQUEMA

a) Esquema en estrella: consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella.



DATAWAREHOUSE – TIPOS ESQUEMA

b) Esquema en copo de nieve: es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón, la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas, y aparecen nuevas uniones.