

2 Diseño de Almacenes y bancos de datos



2.1 Diseño de Almacenes de datos (Data Warehouse)



2.2 Referencia Arquitectural de los Data Warehouse



2.3 Funcionamiento de los almacenes de Datos



2.4 Funciones Extracción, Transformación y Limpieza (ETL)

2.1 Diseño de Almacenes de datos (Data Warehouse)

- Para abordar un proyecto de DW es necesario hacer un estudio de algunos temas generales de la empresa, los cuales se describen a continuación:
 - **Situación actual de partida** - Cualquier solución propuesta de DW debe estar muy orientada por las necesidades del negocio y debe ser compatible con la arquitectura técnica existente y planeada de la compañía.
 - **Tipo y características del negocio** - Es indispensable tener el conocimiento exacto sobre el tipo de negocios de la organización y el soporte que representa la información dentro de todo su proceso de toma de decisiones.


- **Entorno técnico** - Se debe incluir tanto el aspecto del hardware (mainframes, servidores, redes) así como aplicaciones y herramientas, dando énfasis a los DSS, cómo operan, etc.
- **Expectativas de los usuarios** - Un proyecto de DW no es únicamente un proyecto tecnológico, es una forma de vida de las organizaciones y como tal, tiene que contar con el apoyo de todos los usuarios y su convencimiento sobre su bondad.
- **Etapas de desarrollo** - Con el conocimiento previo, ya se entra en el desarrollo de un modelo conceptual para la construcción del DW.

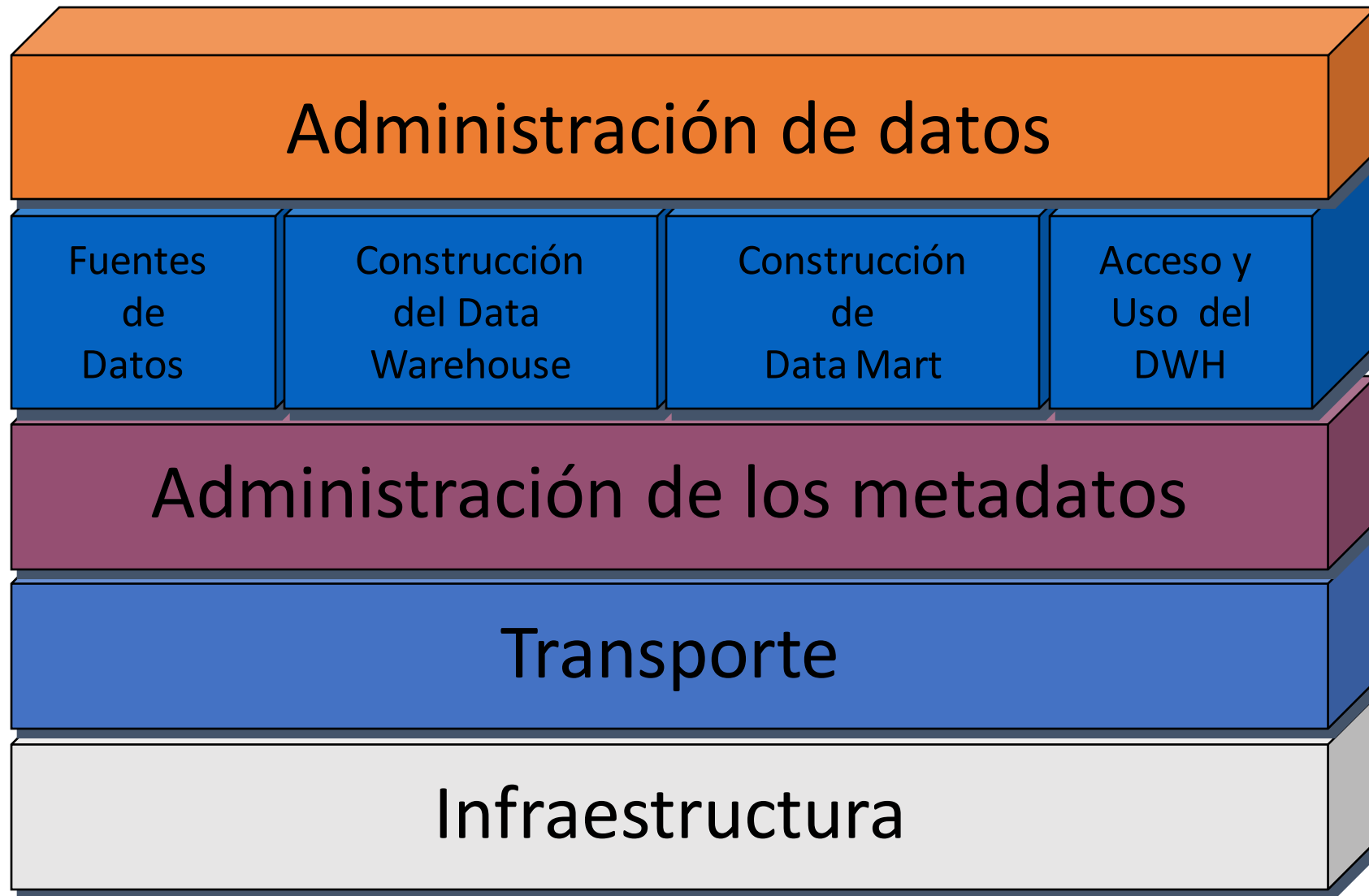
- **Prototipo** - Un prototipo es un esfuerzo designado a simular tanto como sea posible el producto final que será entregado a los usuarios.
- **Piloto** - El piloto de un DW es el primero, o cada uno de los primeros resultados generados de forma iterativa que se harán para llegar a la construcción del producto final deseado.
- **Prueba del concepto tecnológico** - Es un paso opcional que se puede necesitar para determinar si la arquitectura especificada del DW funcionará finalmente como se espera.

-
- Un DW se puede implementar de acuerdo con diferentes enfoques de diseño: de arriba hacia abajo, de abajo hacia arriba o mixto.
 - **De arriba hacia abajo.** La metodología de arriba hacia abajo se basa en el diseño general del DW y, por lo tanto, es más sistemática. Sin embargo, implica tiempos de desarrollo más largos y mayores riesgos de no completarse dentro de lo programado, ya que todo el DW se está desarrollando realmente.
 - **De abajo hacia arriba.** El método ascendente se basa en el uso de prototipos y, por lo tanto, las extensiones del sistema se realizan de acuerdo con un esquema paso a paso. Este enfoque suele ser más rápido, proporciona resultados más tangibles, pero carece de una visión global de todo el sistema que se va a desarrollar.
 - **Mixto.** La metodología mixta se basa en el diseño general del DW, pero luego procede con un enfoque de creación de prototipos, mediante la aplicación secuencial de diferentes partes de todo el sistema. Este enfoque es muy práctico y generalmente preferible, ya que permite dar pasos pequeños y controlados teniendo en cuenta el panorama completo.

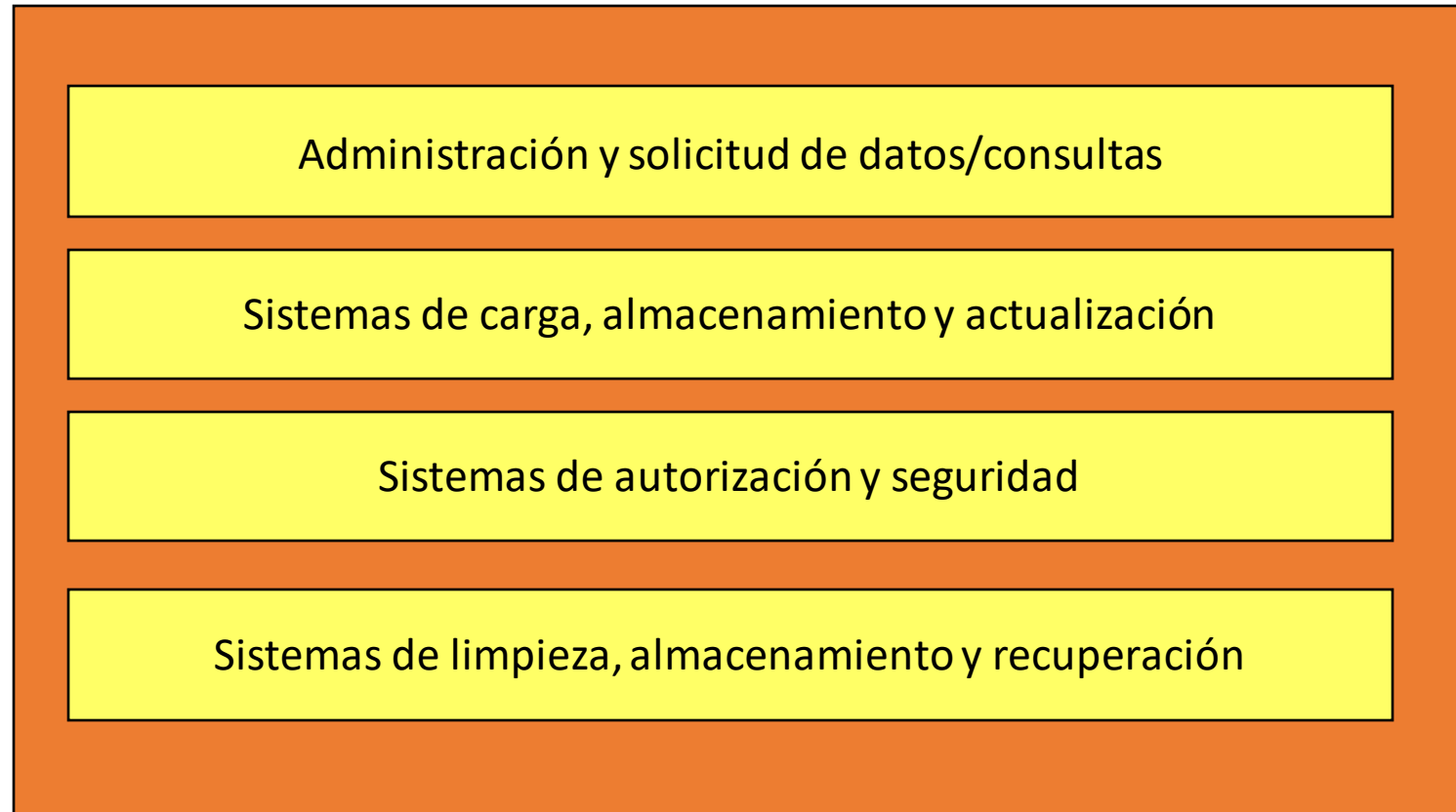
2.2 Referencia Arquitectural de los DW

-
- Es una forma unificada y común de ver los componentes de una solución de Data Warehouse, basada en el framework de Zachman y ampliamente aceptada.
 - Proporciona un diagrama conceptual para analizar las diversas opciones de implementación, así como de su planeación.
 - A través de esta arquitectura referencial podemos:
 - Evaluar las inversiones actuales.
 - Analizar costos y beneficios
 - Administración y análisis del riesgo
 - Evaluar herramientas de diversos proveedores
 - Administración y planeación del proyecto
 - Administración y análisis de las habilidades
 - Simulación del Proyectos
 - Arquitectura y Diseño

- 
- La arquitectura de referencia se describe del siguiente modo:
 - Un conjunto de datos extraídos de bases de datos operacionales.
 - Un software que prepara los datos para que puedan ser accedidos por los usuarios.
 - Un conjunto de aplicaciones y herramientas que ejecutan un conjunto de consultas y análisis complejos.



Administración de Datos



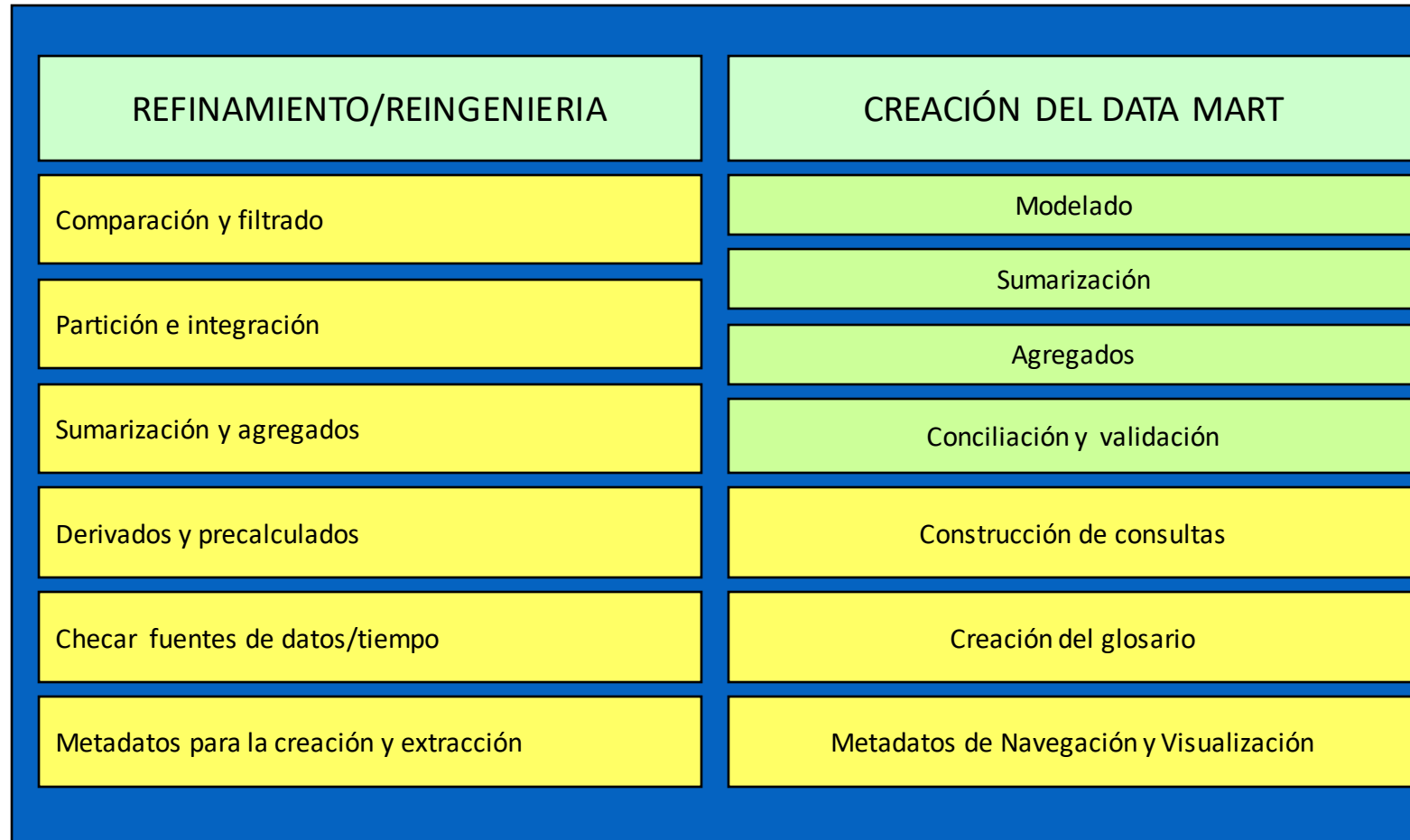
Fuentes de Datos



Construcción del DW

REFINAMIENTO	REINGENIERÍA	DATA WAREHOUSE
Estandarización	Partición e integración	Modelado
Comparación y filtrado	Sumarización y agregados	Sumarización
Limpieza	Derivados y precalculados	Agregación
Checar fuentes de datos/tiempo	Traducción y Formateo	Conciliación y validación
Verificar calidad de datos	Transformación y Mapeo	Construcción de consultas
Metadatos de Extracciones	Metadatos de Reingeniería	Creación del glosario
		Metadatos de Visualización y Navegación

Construcción del DM(s)



Acceso y uso del DWH

ACCESO Y RECUPERACIÓN	ANÁLISIS Y REPORTES
Acceso directo al Data Warehouse	Reporteadores
Acceso al Data Mart	Herramientas de análisis y de DSS
Reingeniería	Herramientas de modelación empresarial
Transformación a estructura multidimensional	Herramientas de Minería de Datos
Creación de almacenamiento local	Herramientas OLAP
Visualizador y navegador de los metadatos	Aplicaciones
Metadatos del Data Warehouse	Metadatos para los reportes

Administración de los Metadatos

Administración del glosario y de los esquemas del DW y DM

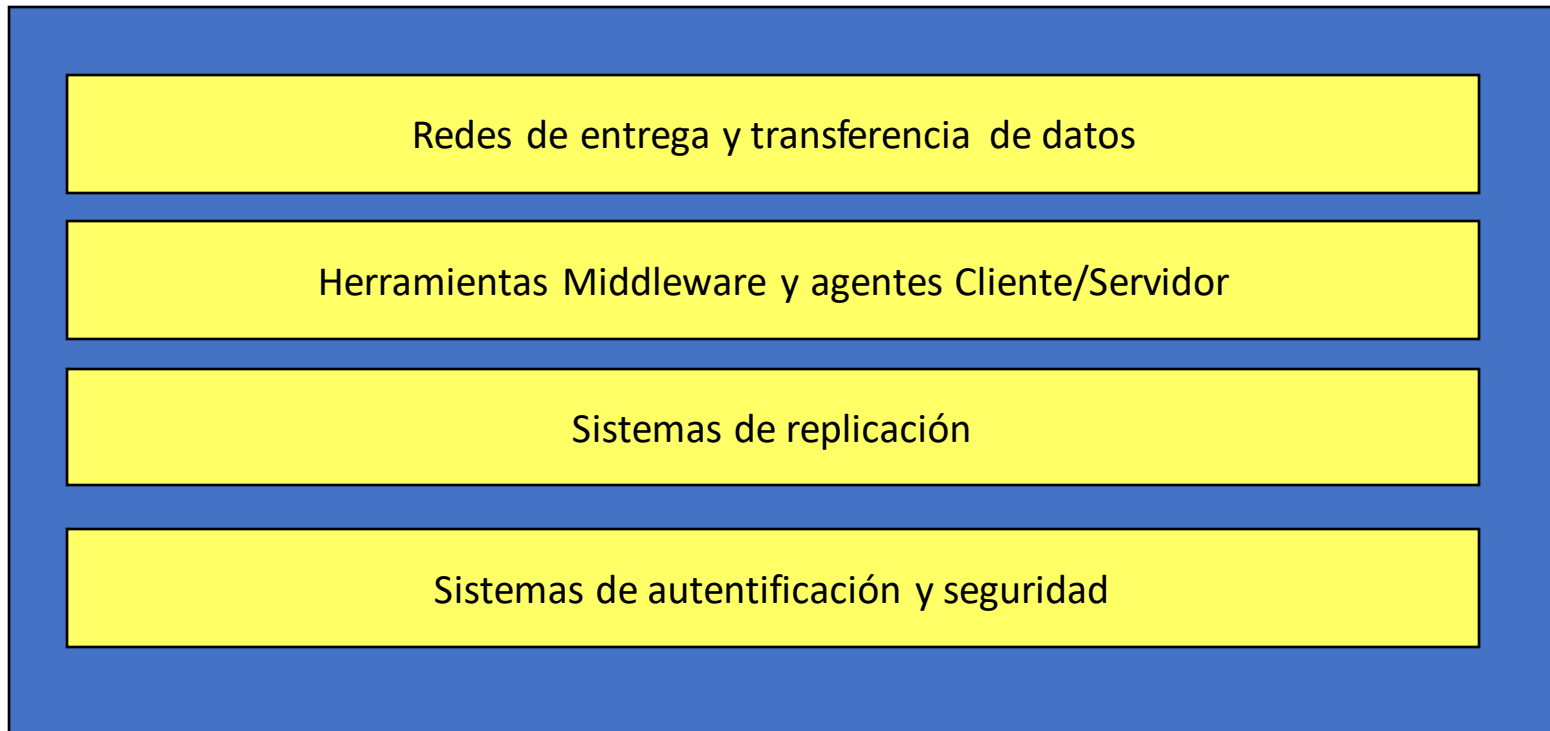
Administración en la creación de los metadatos, su almacenamiento y su actualización

Administración de consultas predefinidas, índices y reportes

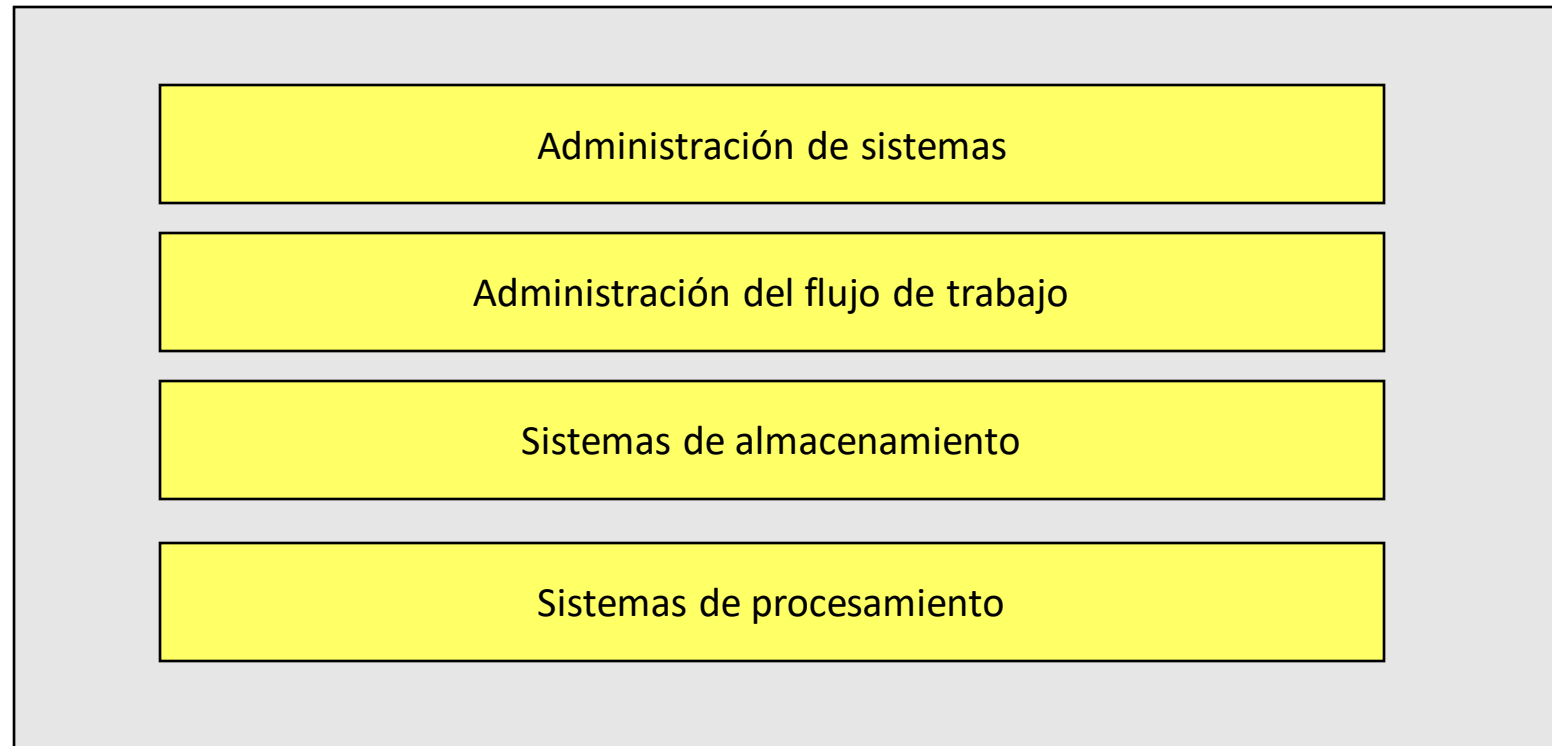
Administración de actualizaciones y replicaciones

Administración de usuarios, de respaldo y recuperación

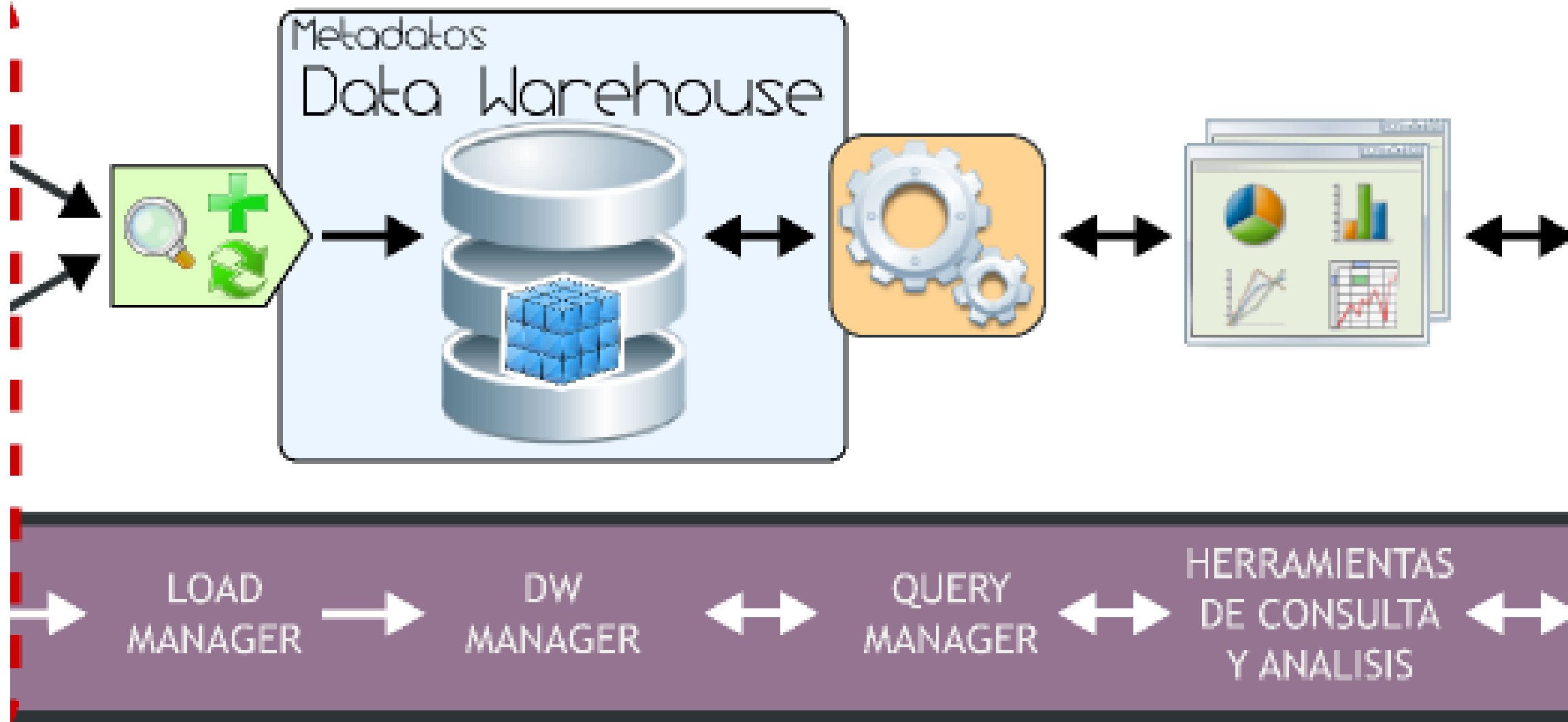
Transporte



Infraestructura



2.3 Funcionamiento de los almacenes de Datos



2.4 Funciones Extracción, Transformación y Limpieza (ETL)

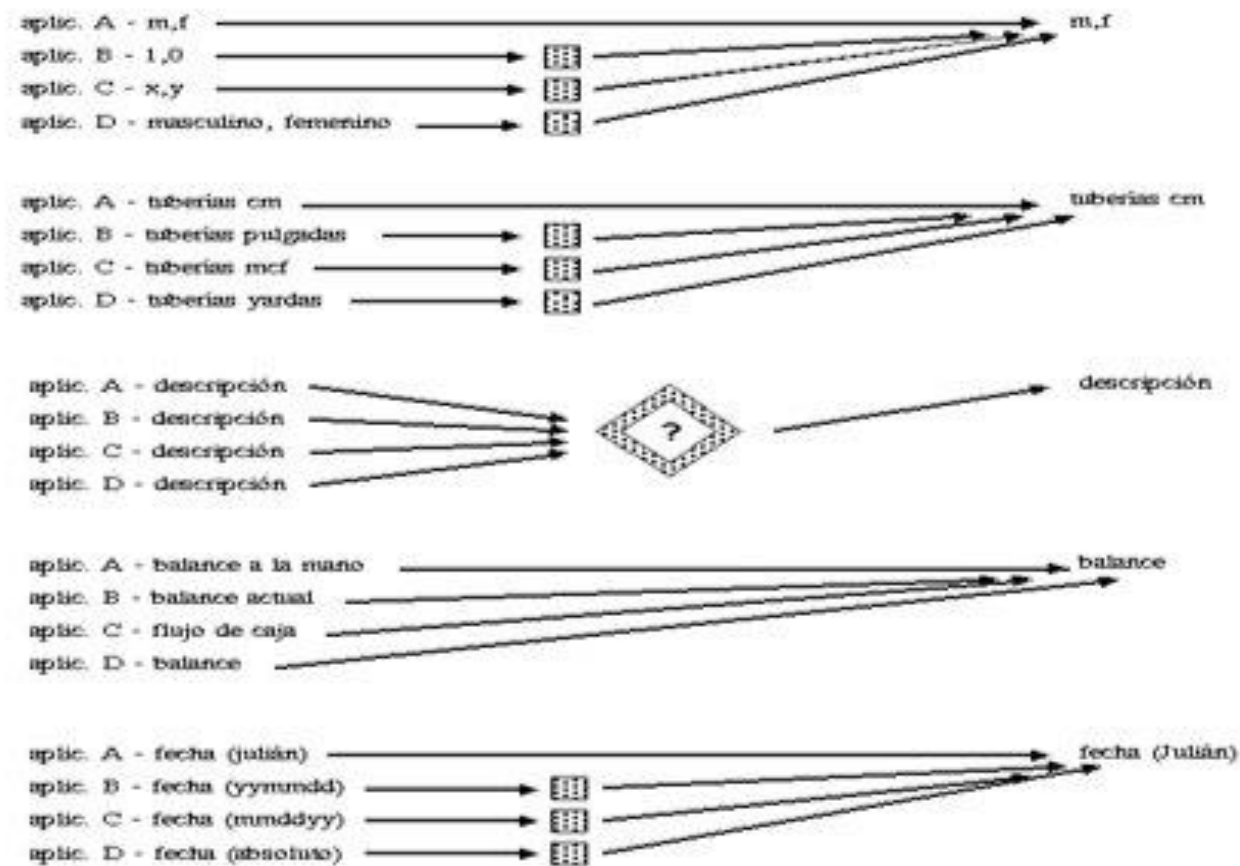
ETL

- ETL se refiere a las herramientas de software que se dedican a realizar de forma automática tres funciones principales: extracción, transformación y carga de datos en el DW
 - **Extracción.** Durante la primera fase, los datos se extraen de las fuentes internas y externas disponibles. La selección de los datos que se van a importar se basa en el diseño del almacén de datos, que a su vez depende de la información necesaria para los análisis de inteligencia empresarial y los DSS que operan en un dominio de aplicación específico.

-
- **Transformación.** El objetivo de la fase de limpieza y transformación es mejorar la calidad de los datos extraídos de las diferentes fuentes, a través de la corrección de incoherencias, inexactitudes y valores faltantes. Algunas de las principales deficiencias que se eliminan durante la fase de limpieza de datos son:
 - incoherencias entre los valores registrados en diferentes atributos que tienen el mismo significado;
 - duplicación de datos;
 - datos faltantes;
 - existencia de valores inadmisibles

-
- Durante la fase de limpieza, se aplican reglas automáticas preestablecidas para corregir los errores actuales. En muchos casos, los diccionarios con términos válidos se utilizan para sustituir los términos supuestamente incorrectos, en función del nivel de similitud.
 - Durante la fase de transformación, se realizan conversiones de datos para garantizar la homogeneidad y la integración con respecto a las diferentes fuentes de datos.
 - Se realizan la agregación y consolidación de datos con el fin de obtener los resúmenes que reducirán el tiempo de respuesta necesario para las consultas y análisis posteriores.

Transformación de datos



Cuando los datos se mueven al data warehouse desde las aplicaciones orientadas al ambiente operacional, los datos se integran antes de entrar al depósito.



-
- **Carga.** Por último, después de ser extraídos y transformados, los datos se cargan en las tablas del DW para ponerlos a disposición de los analistas y las aplicaciones de apoyo a la toma de decisiones. Se debe considerar la carga de datos de la siguiente manera:
 - Replicar las tablas de dimensiones
 - Replicar la tabla de hechos
 - Realizar una carga automática
 - Indexar los datos para un mayor rendimiento

Integridad de los datos: problemas, causas y remedios

PROBLEMA	CAUSA	REMEDIO
datos incorrectos	datos recopilados sin el debido cuidado datos introducidos incorrectamente modificación incontrolada de datos	comprobación sistemática de los datos de entrada automatización de la entrada de datos Implementación de un programa de seguridad para el acceso y las modificaciones
datos no actualizados	La recopilación de datos no coincide con las necesidades del usuario	actualización oportuna y recopilación de datos recuperación de datos actualizados de la web
faltan datos	falta de recopilación de los datos adquiridos	identificación de los datos necesarios mediante análisis preliminar y estimación de los datos que faltan

-
- De forma más general, podemos identificar los siguientes factores principales que pueden afectar a la calidad de los datos.
 - **Precisión.** Los datos deben ser altamente precisos. Por ejemplo, es necesario comprobar que los nombres y las codificaciones están correctamente representados y los valores están dentro de intervalos admisibles.
 - **Integridad.** Los datos no deben incluir un gran número de valores que faltan. Sin embargo, hay técnicas de aprendizaje y minería de datos son capaces de minimizar de forma sólida los efectos de la incompletitud parcial en los datos.
 - **Coherencia.** La forma y el contenido de los datos deben ser coherentes en las diferentes fuentes de datos después de los procedimientos de integración, con respecto a la moneda y las unidades de medida.
 - **Puntualidad.** Los datos deben actualizarse con frecuencia, basándose en los objetivos del análisis. Es costumbre organizar una actualización del DW regularmente en tiempos determinados.

-
- **Relevancia.** Los datos deben ser pertinentes para las necesidades del sistema de inteligencia empresarial con el fin de añadir un valor real a los análisis que se realizarán posteriormente.
 - **Interpretabilidad.** El significado de los datos debe ser bien entendido e interpretado correctamente por los analistas, también sobre la base de la documentación disponible en los metadatos que describen a un DW.
 - **Accesibilidad.** Los analistas y las aplicaciones DSS deben tener fácil acceso a los datos.
 - **No redundancia.** Se debe evitar la repetición de datos y la redundancia para evitar el desperdicio de memoria y las posibles incoherencias.

Metadatos

- Para documentar el significado de los datos contenidos en un DW, se recomienda establecer una estructura de información específica, conocida como metadatos.
- Los metadatos indican para cada atributo de un DW la fuente original de los datos, su significado y las transformaciones a las que han sido sometidos.
 - La documentación proporcionada por los metadatos debe mantenerse constantemente actualizada, con el fin de reflejar cualquier modificación en la estructura del DW.
 - La documentación debe ser accesible a los usuarios del DW, de acuerdo con los derechos de acceso correspondientes a las funciones de cada analista.
 - En particular, los metadatos deben realizar las siguientes tareas informativas:
 - documentación de la estructura del DW: diseño, vistas lógicas, dimensiones, jerarquías, datos derivados, localización de cualquier DM;
 - documentación de la genealogía de los datos, obtenida mediante el etiquetado de las fuentes de datos de las que se extrajeron los datos y la descripción de cualquier transformación realizada en los propios datos;
 - documentación del significado general del DW con respeto al dominio de aplicación, proporcionando la definición de los términos utilizados y describiendo completamente las propiedades de los datos, la propiedad de los datos y las políticas de carga.

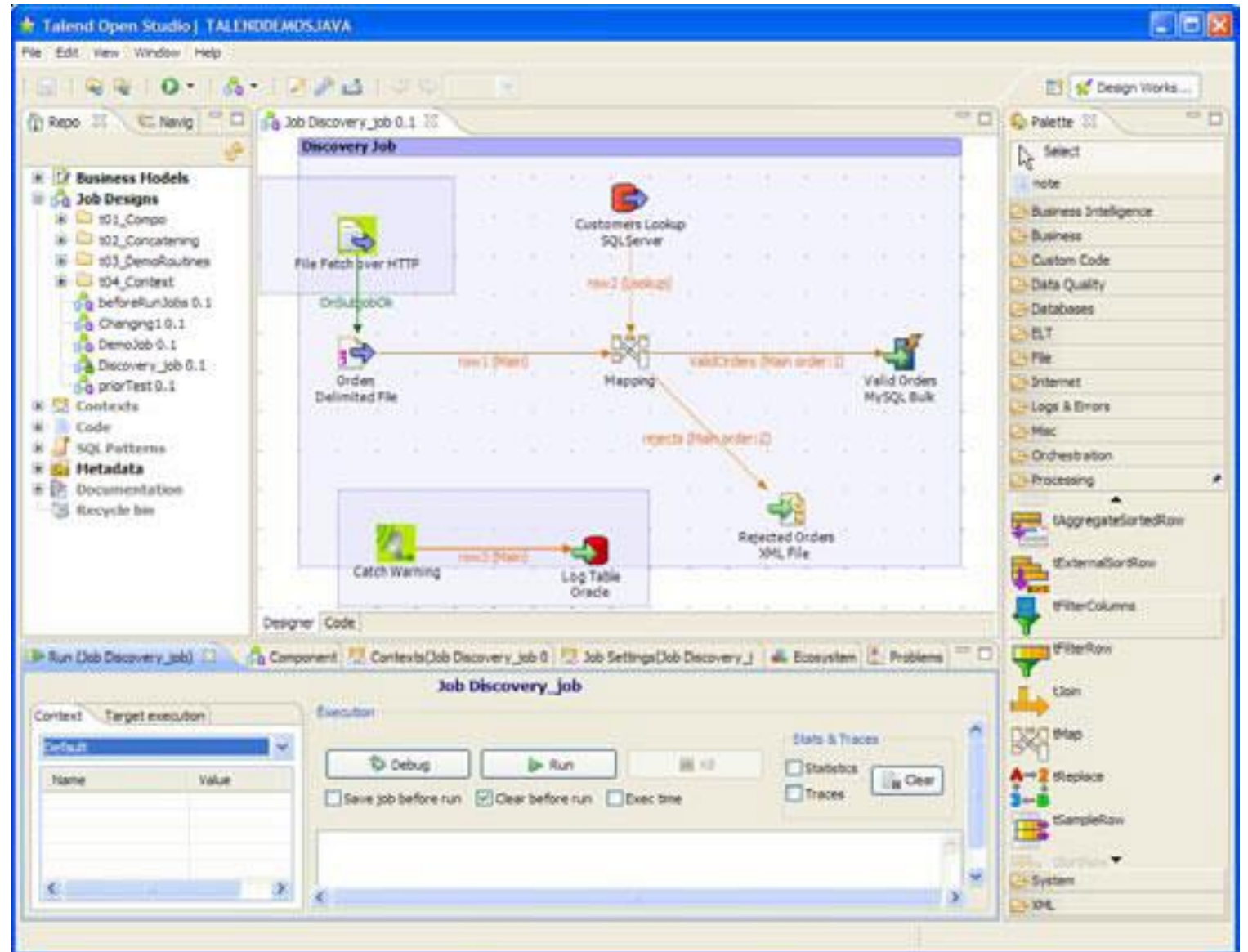
Herramientas ETL

- Se componen de diversos motores:
 - a) Motor de extracción: utiliza adaptadores como ODBC, JDBC, JNDI, SQL nativo, adaptadores de archivos planos u otros.
 - b) Motor de transformación: proporciona una librería de objetos que permite a los desarrolladores transformar los datos de origen para adaptarse a las estructuras de datos de destino, permitiendo, por ejemplo, la sumarización de los datos en destino en tablas resumen.
 - c) Motor de carga: utiliza adaptadores a los datos de destino, como el SQL nativo, o cargadores masivos de datos para insertar o actualizar los datos en las bases de datos o archivos de destino.
 - d) Servicios de administración y operación: permiten la planificación, ejecución y monitorización de los procesos ETL, así como la visualización de eventos y la recepción y resolución de errores en los procesos.

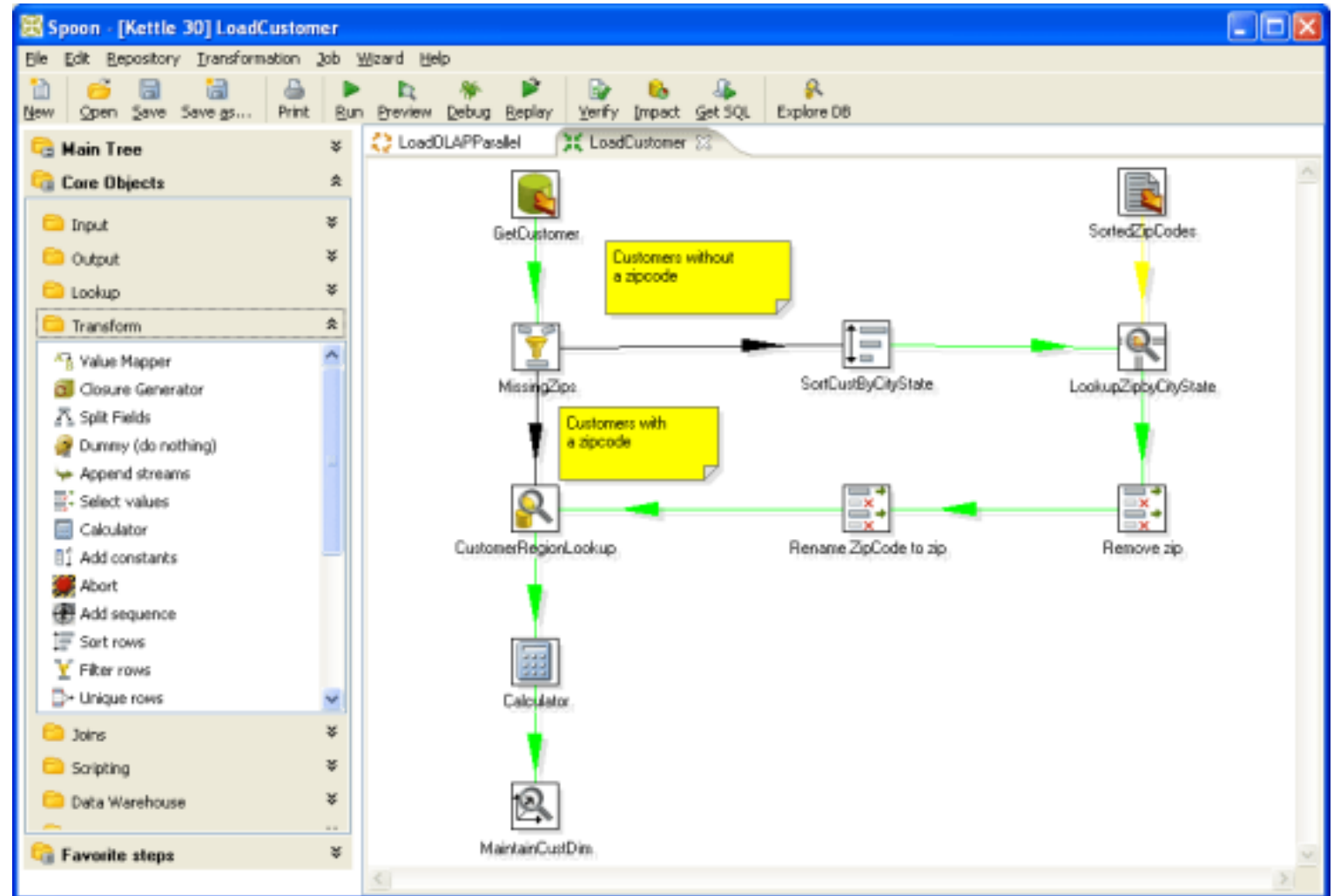
Software ETL

- Ab Initio
- Barracuda Software
- Bitool
- Cognos Decisionstream
- Genio, Hummingbird
- IBM Websphere DataStage (Previously Ascential DataStage)
- Informática PowerCenter
- metaWORKS (Document Tools)
- Microsoft DTS (incluido en SQL-Server 2000)
- Microsoft SQL Server Integration Services (SSIS) (a partir de MS SQL Server 2005)
- MySQL Migration Toolkit
- Oracle Data Integrator
- Oracle Warehouse Builder
- SAP Data Services
- Stratio Sparta
- Kettle (ahora llamado Pentaho Data Integration).
- Talend Open Studio.

Talend



Kettle



Consideraciones del proceso ETL

Muy apoyado en los metadatos

Dos tipos de herramientas

- Genéricas
 - Cubren casi todo
- Propietarias (ad-hoc)
 - En ocasiones hay partes del proceso no automatizables

Dos tipos de carga

- Carga inicial
- Cargas incrementales

Frecuencia de carga incremental

- Tiempo real
- Detección de cambios
- Batch