Modelos y Persistencia de Datos

Alejandro Sierra

Agenda

- EII (Integración de Información Empresarial)
  - Descubrimiento
  - Perfilamiento
  - Limpieza
  - Transformación
  - Replicación
  - Federación
  - Flujos de Datos (Streaming)
  - Despliegue
- ETLs

EII (Integración de Información Empresarial)

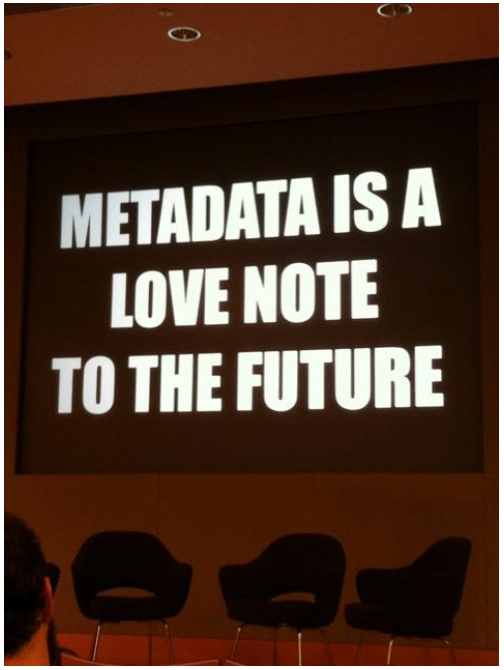
- Soportar temas como
  - Master Data Management
  - Metadata Management
  - Data Warehousing
- Movimiento de Datos
- Diferentes capacidades
  - Posiblemente una necesidad de negocio requiera de la combinación de varias.

Escenarios de EII

- Migración y conversión
- Consolidación
- Compartir datos
- Distribuir datos geográficamente
- Archivar datos
- Obtener datos externos
- Integrar datos estructurados y no estructurados

Capacidades de EII

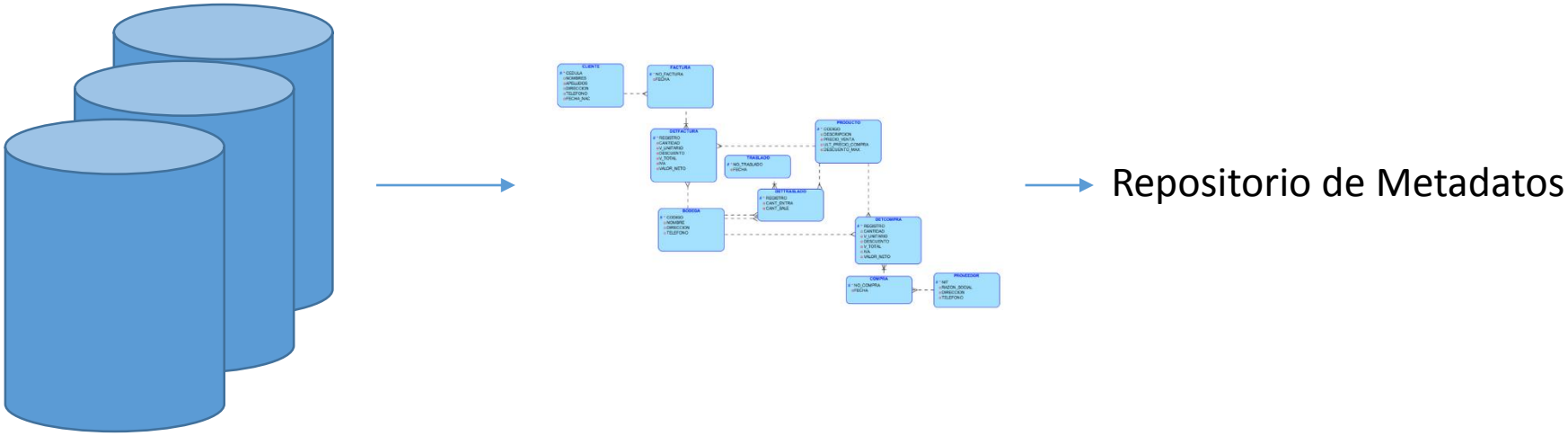
Capacidad de Descubrimiento



- Entender
  - Fuentes de datos
  - Reglas de negocio
  - Definiciones de negocio
  - …. Metadatos
- No nos enfocamos en los datos como tal sino en los metadatos.
- Descubrimiento automático de metadatos
  - Mucho más fácil si tenemos un repositorio de metadatos (o varios unificados mediante federación)
- Unificar metadatos técnicos con metadatos de negocio.
- Es un proceso constante de descubrimiento-mantenimiento de los metadatos.
  - Servicio de Descubrimiento que puede ser invocado periódicamente o por demanda

Capacidad de Descubrimiento

- Escenario
  - A través de una interfaz de usuario, un analista invoca el Servicio de Descubrimiento para obtener los metadatos de un conjunto de bases de datos.
  - El Servicio de Descubrimiento lee la estructura de la base de datos mediante los DDLs. Con esto genera metadatos en forma de Diagramas ER (ingeniería inversa)
  - Estos metadatos alimentan el Repositorio de Metadatos.



Capacidad de Perfilamiento

- Nos enfocamos en los datos
- Buscar relaciones implícitas en los datos
- Buscar problemas de calidad
  - Duplicados
  - Datos que no cumple las reglas de negocio
  - Columnas que no cumplen un requisito UNIQUE
  - Datos Nulos
- También es un proceso continuo de auditoria sobre los datos.
- Complementa los metadatos del Servicio de Descubrimiento

Capacidad de Perfilamiento

- Escenario
  - El Servicio de Perfilamiento es invocado para un conjunto de bases de datos operacionales
  - Se detectan problemas de calidad en los datos
  - Se genera un reporte para los Responsables de Información de las bases de datos involucradas.
  - Se guarda la información en el repositorio de metadatos

Capacidad de Limpieza

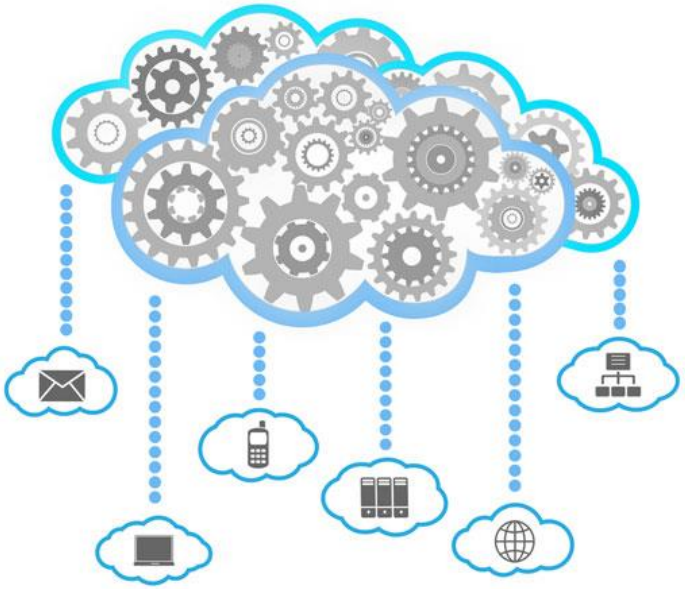
- Muchas fuentes con formatos diferentes
- Sistemas legacy
- Campos usados para un propósito diferente al original.
  - Guardar información estructurada en campos de texto libre.
- El perfilamiento es un insumo para la limpieza
- La limpieza se debe poder hacer por registro o por lotes.
- Se reportan al Repositorio de Metadatos los resultados.
  - Cantidad de registros analizados
  - Cantidad de registros rechazados
  - …

Capacidad de Limpieza

- Investigación: Perfilamiento.
  - Conocimiento específico al dominio (por ejemplo int vs double)
- Estandarización: Definir un formato de destino y las reglas para convertir los formatos origen. (por ejemplo Cll vs Calle, Cr, Kr, Carrera)
- Detectar coincidencias.
  - Determinístico: Coincidencia exacta
  - Probabilístico: Porcentaje. Mayor complejidad.
  - Detectar duplicados y encontrar registros correspondientes en diferentes fuentes de datos
- Determinación:
  - Filtro. Algunos registros se guardan para la revisión del Mayordomo de Información.

Capacidad de Transformación

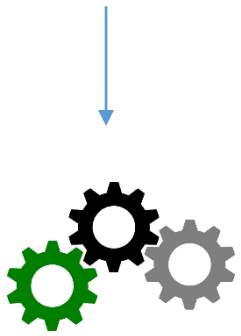
- Cálculos derivados
  - Ej. Precio total de una factura
- Agregaciones
  - Ej. Total de ventas por región
- Procesamiento intenso
  - Paralelismo
- Cambios de formato/estructura
- Codificaciones
- También se generan Metadatos a cerca del estado del proceso y los resultados (errores, # de filas, …)



Capacidad de Transformación

- Escenario: Tablas de resumen.

Cliente	Producto	Fecha	Valor
…	X	15/06/2015	1000
…	Y	15/06/2015	2000
…	X	16/06/2015	1100
…	Y	16/06/2015	2000
…	Z	16/06/2015	3000

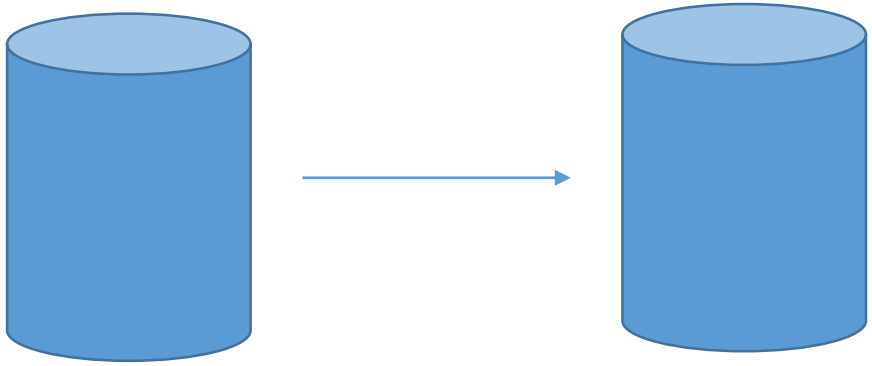


Fecha	Valor
15/06/2015	3000
16/06/2015	6100

Producto	Valor
X	2100
Y	4000
Z	3000

Capacidad de Replicación

- Hay estrategias de HW y de SW
- Se puede usar el Servicio de Transformación de manera trivial.
- En lotes / tiempo real suave / tiempo real
  - Requerimientos
  - Capacidad de Infraestructura
- Metadatos:
  - Estado de replicación.
  - Cuales servidores están sincronizados.

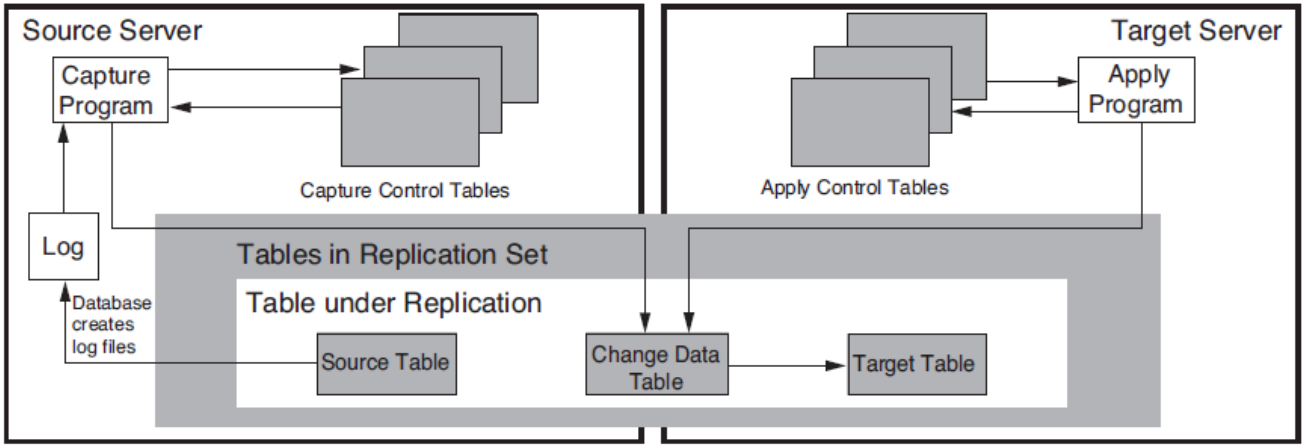


Replicación por Triggers

- Incluir triggers para Insert/Update/Delete
- Ventaja: Soportado por la mayoría de bases de datos relacionales
- Desventaja
  - Degrada el desempeño en la base de datos original (Una operación no culmina hasta que no sea replicada)
- Solo si los servidores tiene muy buena conectividad

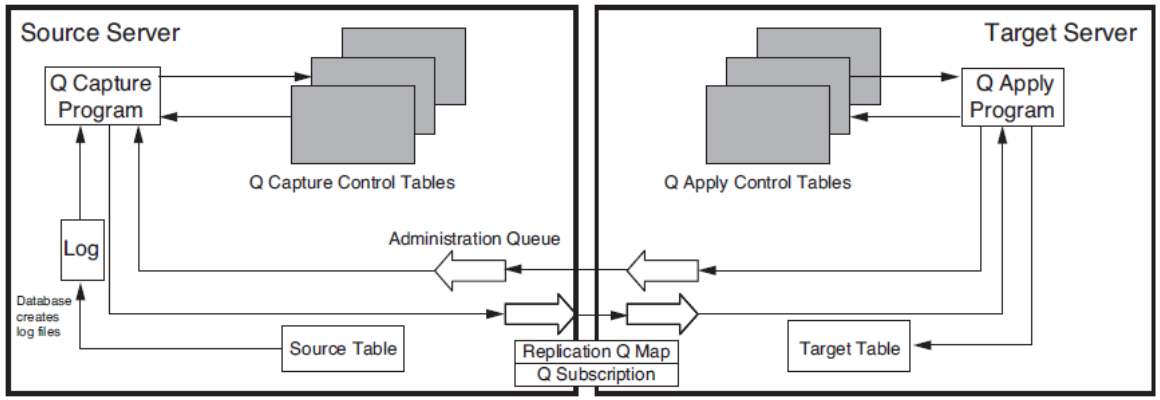
Replicación por SQL

- Lectura de los logs de origen.
- No afecta el desempeño en el origen.
- Limitaciones tecnológicas.



Replicación por Colas

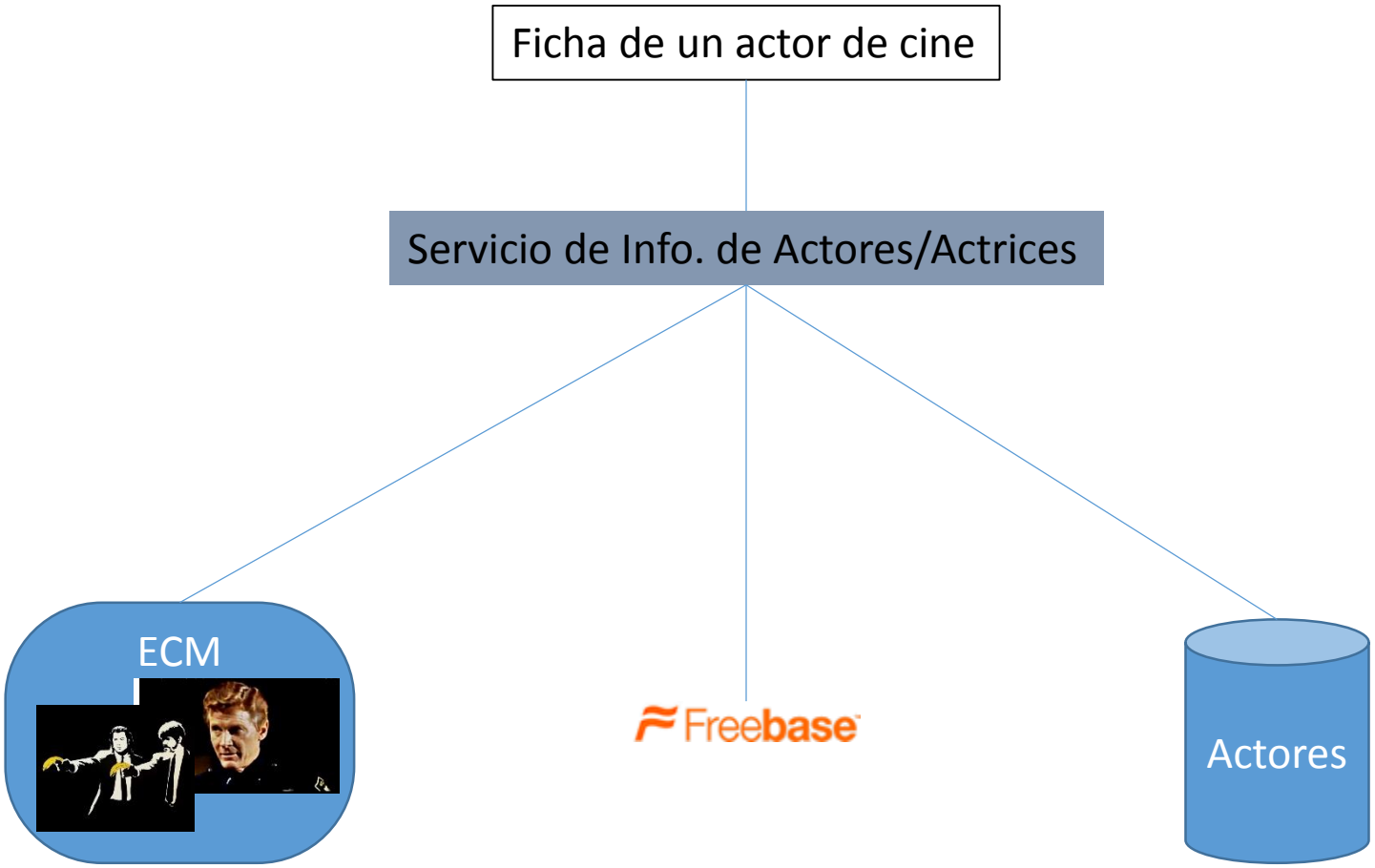
- Utilizar una cola de mensajes (Middleware)
  - Debe asegurar la entrega de los mensajes
- Asíncrono
  - Eventualmente consistentes.
- No depende de el estado del destino.
  - Puede estar apagado.
- Escalable a grandes volúmenes y diferentes configuraciones



Capacidad de Federación

- Provee una vista única de diferentes fuentes de datos (BD, Servicios, …).
- En ocasiones no necesitamos o no podemos mover los datos
  - Restricciones de infraestructura
  - Consultas esporádicas y sin necesidad de alto desempeño.
  - Consultas bajo demanda.
- Se utiliza el repositorio de Metadatos para conocer las diferentes fuentes
  - Ubicación, protocolo de acceso, …
- Se ajusta mejor a bajos volúmenes de datos.
- Puede funcionar como piloto para un proceso más complejo que involucre movimiento de datos
- CACHE!

Capacidad de Federación



Capacidad de Federación.

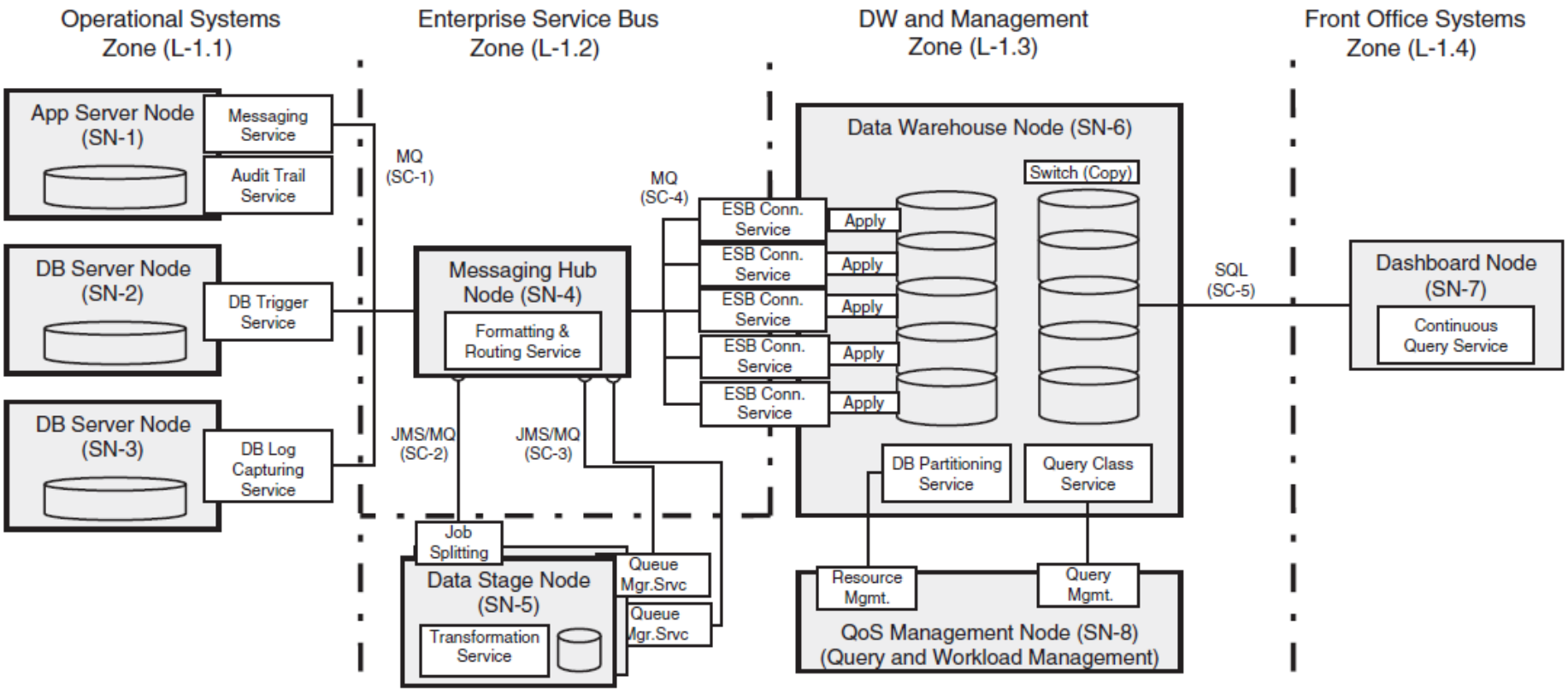
¿Cuándo es pertinente?

Costos de Replicación > Beneficios de Negocio

Flujos de Datos (Streaming)

- VVV (Volumen, Velocidad, Variedad)
- A diferencia del procesamiento en lotes necesitamos procesar los datos a medida que llegan.
- Queremos acercarnos a tiempo real
  - Mejor tiempo de reacción ante eventos
- Paralelismo para las transformaciones.
  - Procesar imágenes
  - Procesar texto
  - ….

Patrón de BI en tiempo real suave



The Art of Enterprise Information Architecture (Godinez et. Al.). Capítulo 6

Patrón de BI en tiempo real suave

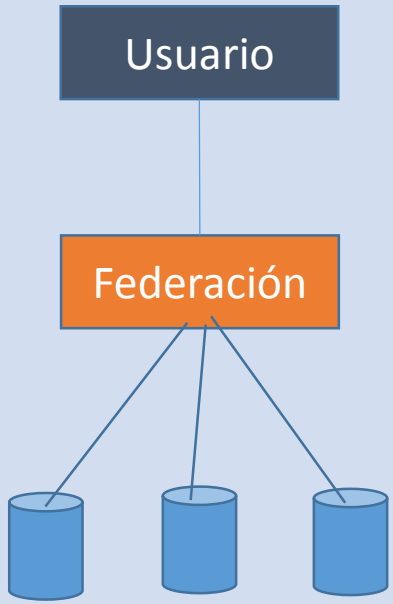
- Capturar datos en el momento que son creados
- Servicio de Transformación en Paralelo
- Paralelismo
  - Particiones de datos de entrada
  - Particionar el procesamiento
  - Esto permite escalabilidad lineal
    - Capacidad de procesamiento proporcional a # de Procesadores.
  - Clusters de computadores
- El Hub de Mensajes debe garantizar la entrega de los mensajes.

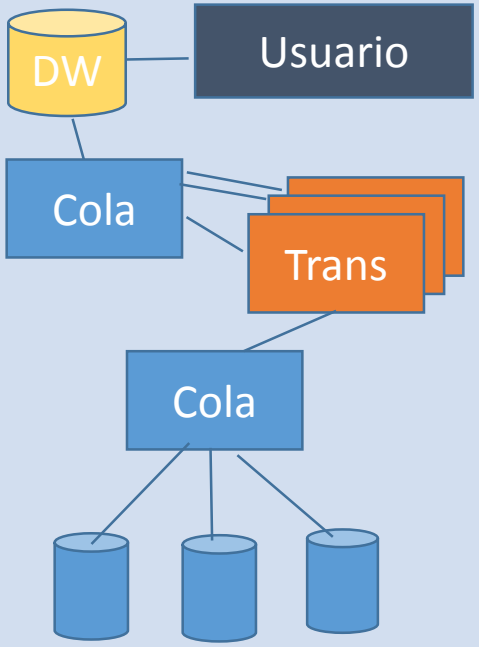
Capacidades de Despliegue

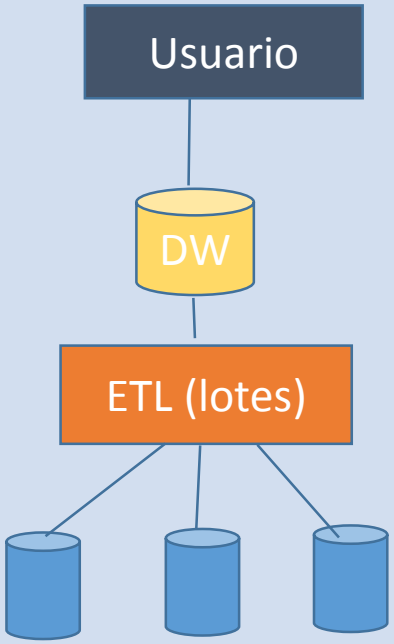
- Todos las capacidades de EII se deben poder desplegar como servicios
- Reusabilidad
  - Por ejemplo componentes de limpieza son altamente reusables
- Escalabilidad
  - Los servicios deben permitir grandes cantidades de datos como entrada
- Estándares
  - SOA, ESB
- Flexible
  - Invocados de diferentes maneras: SOAP/HTTP, REST, JMS, ….
- Reportar resultados
  - Metadatos

Resumen según necesidades de Tiempo

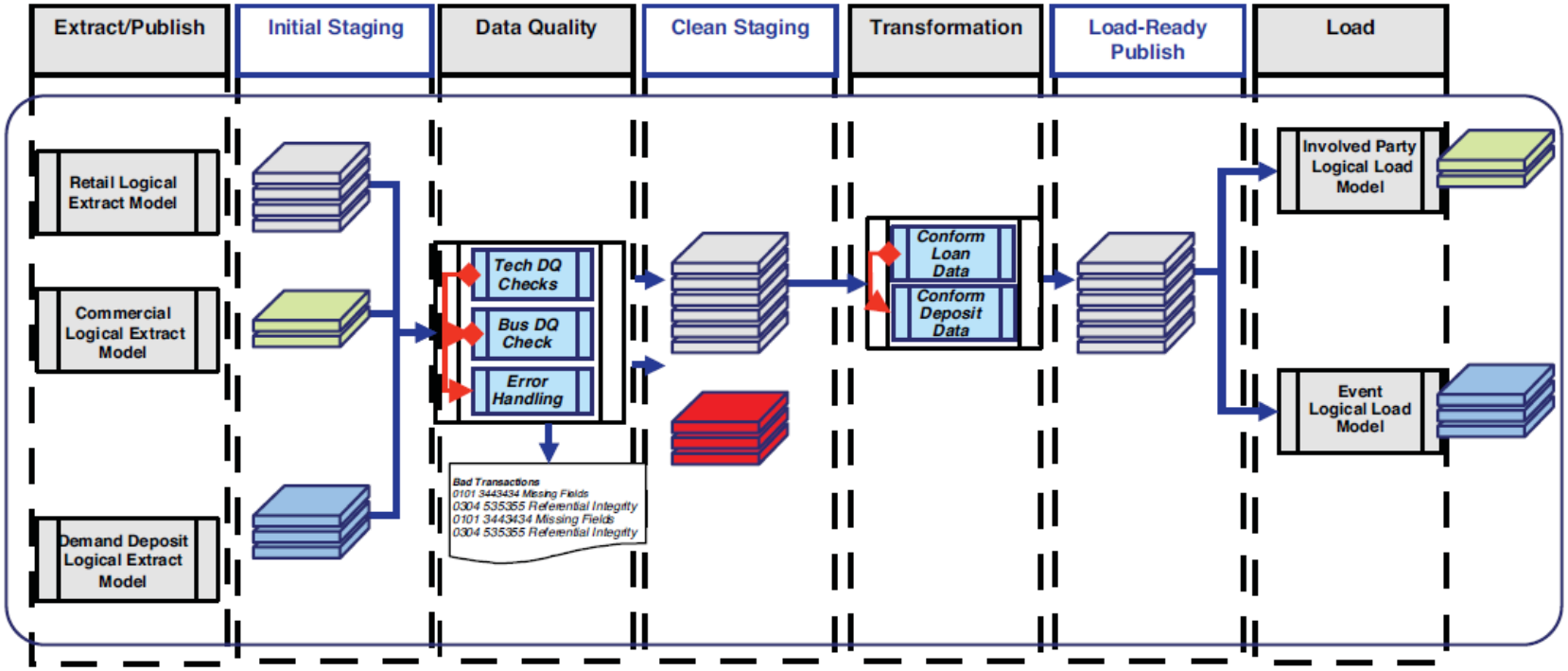
Tiempo Real	Tiempo real suave	Lotes (datos históricos)
Federación	Streaming	ETL clásico





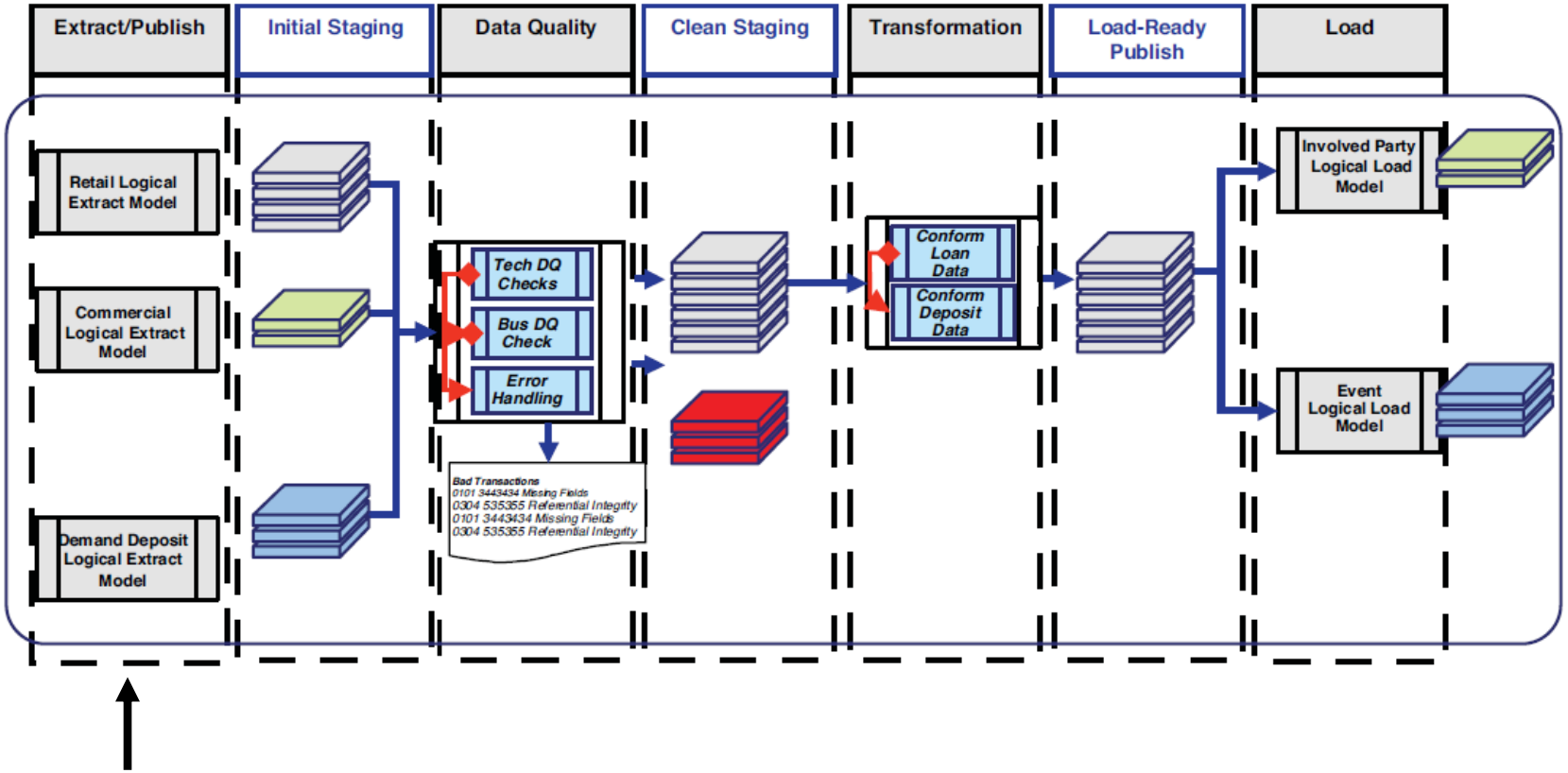


Arquitectura de Referencia para Integración de Datos



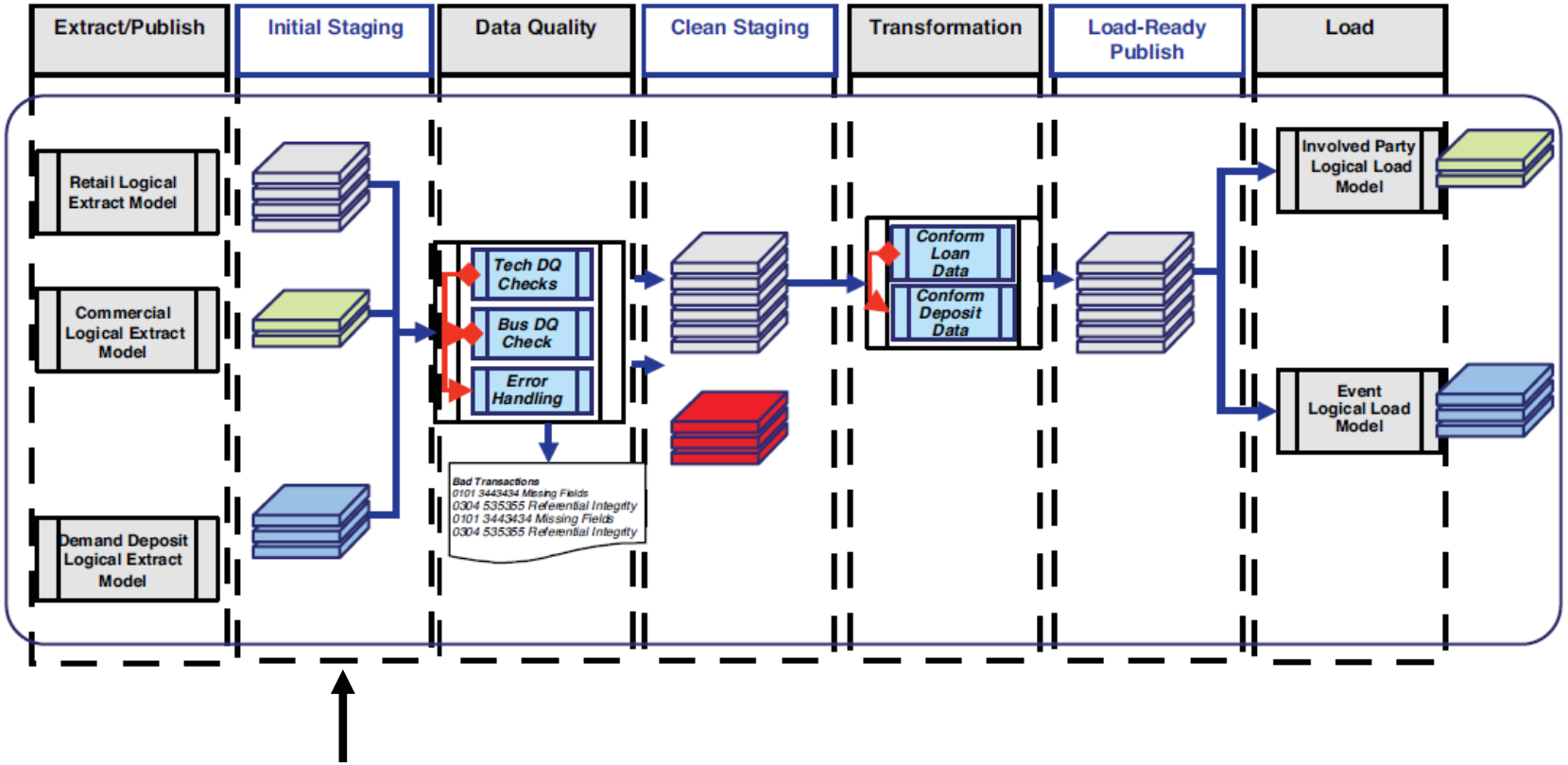
Data Integration Blueprint and Modeling. A. Giordano, 2011. Capítulo 2

Arquitectura de Referencia para Integración de Datos



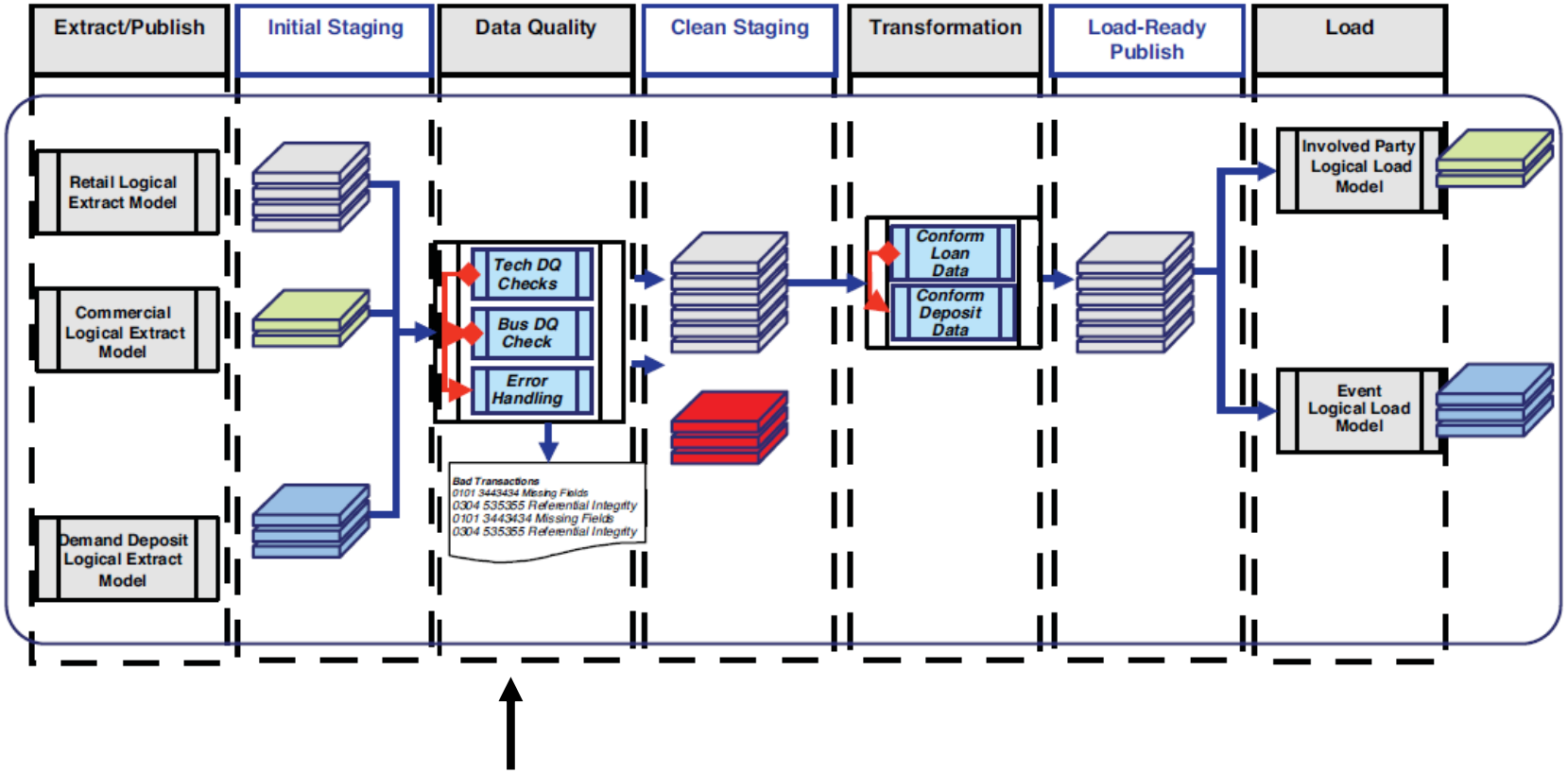
- Leer una sola vez cada fuente de datos y hacer todas las “copias necesarias” (Read once, write many)
- Traer todo pensando en necesidades futuras.

Arquitectura de Referencia para Integración de Datos



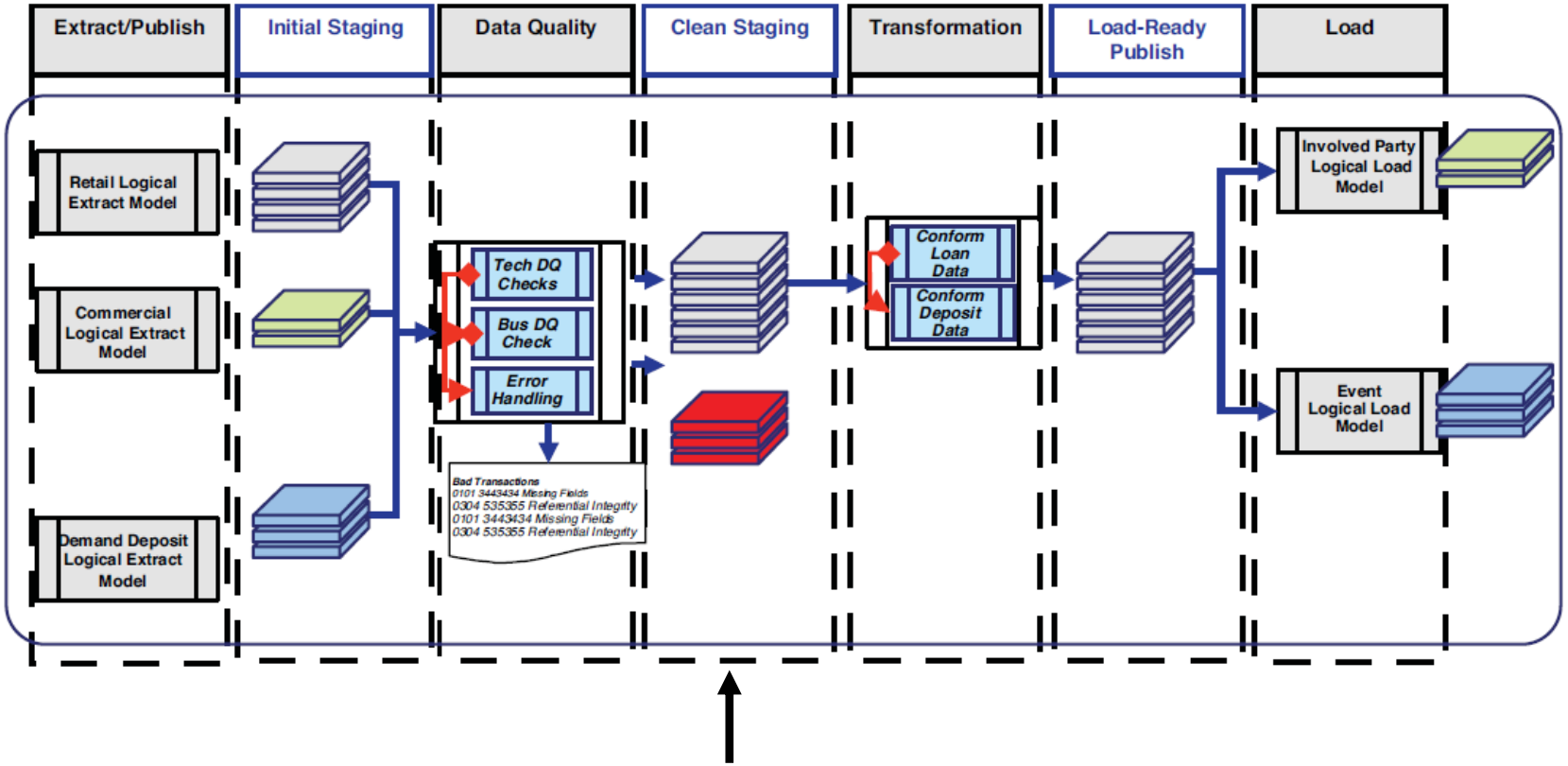
- Almacenamiento no volátil
- Perfilamiento

Arquitectura de Referencia para Integración de Datos



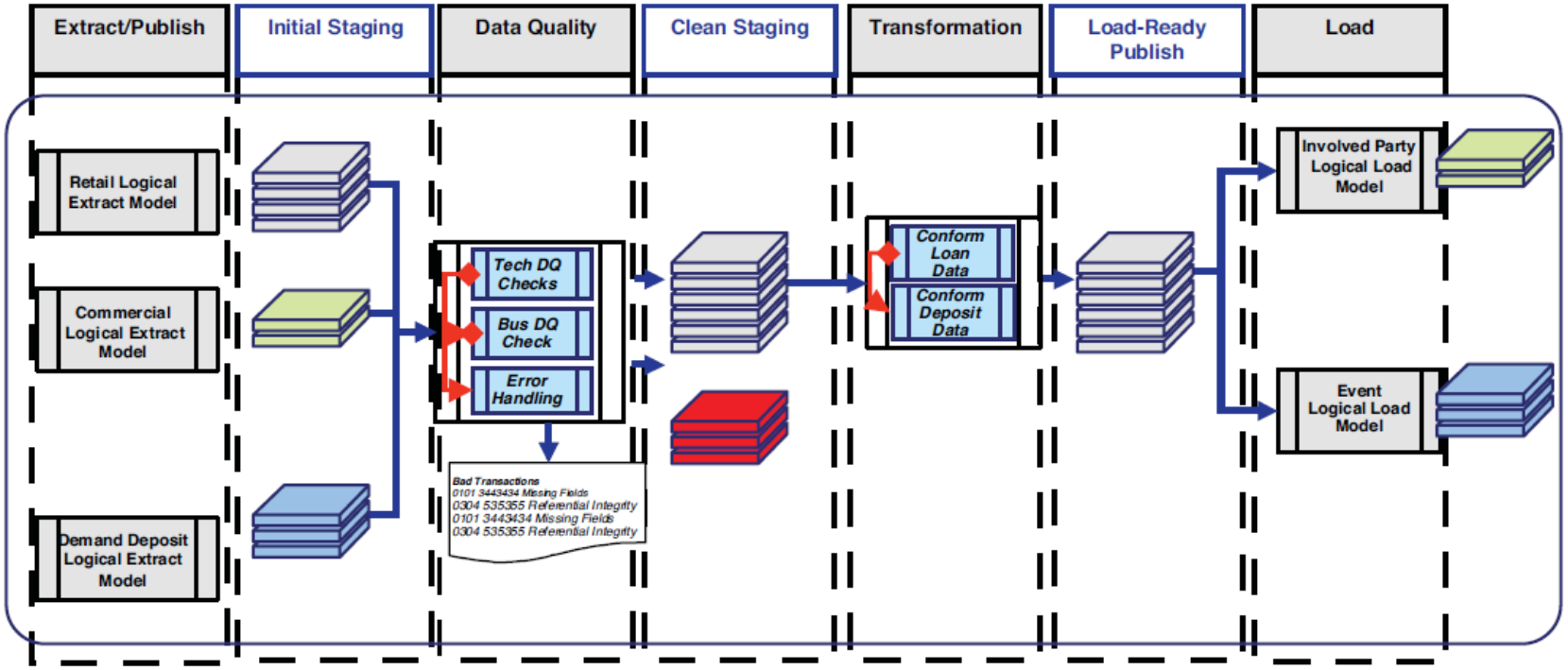
- Calidad de Datos
  - Negocio
    - Definiciones inconsistentes
    - Datos inexactos
- Calidad de Datos
  - Tecnología
    - Datos Faltantes
    - Datos Inválidos

Arquitectura de Referencia para Integración de Datos



- Separar
  - Datos limpios
  - Datos por revisar
  - Datos rechazados

Arquitectura de Referencia para Integración de Datos



- Joins
- Lookups
- Agregaciones

Actividades de EII

Planear y analizar

- Definir requerimientos
- Descubrimiento
- Documentar linaje
- Perfilamiento
- Recolectar reglas de negocio

Diseñar

- Determinar que componentes se han usado para incrementar reutilización
- Diseñar estructuras y servicios necesarios
- Mapeo fuentes/destinos
- Diseñar orquestación

Desarrollar

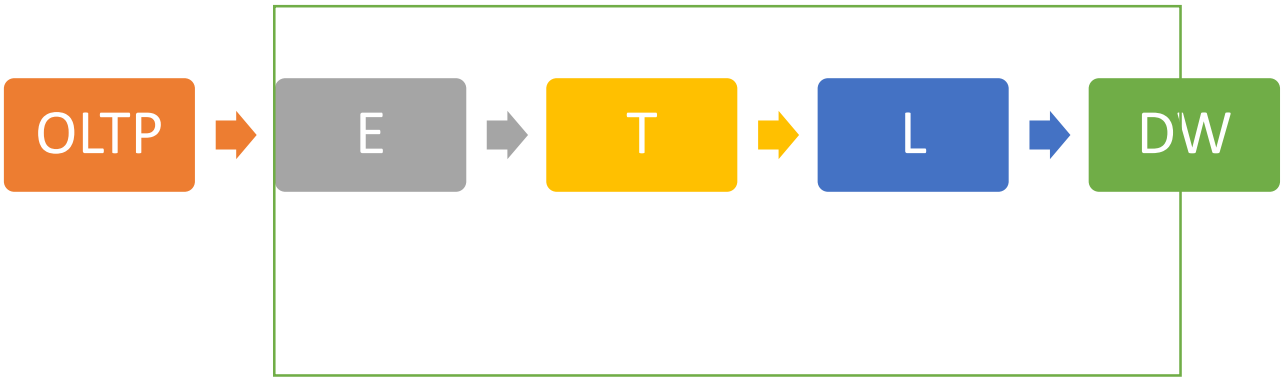
- Servicios
- Flujos de datos
- Mantener metadatos

Implantar / monitorear

- Métricas de desempeño
- Captura de errores
- Notificaciones

ETL

- Proceso para mover datos de bases de datos transaccionales (OLTP) a bodegas de datos (DW)
  - Extracción
  - Transformación
  - Carga (Load)



En ocasiones se separa la limpieza (Cleansing) de la transformación y se le llama ECTL

Referencias

- **The Art of Enterprise Information Architecture**

  A Systems-Based Approach for Unlocking Business Insight
  - Mario Godinez, Eberhard Hechler, Klaus Koenig, Steve Lockwood, Martin Oberhofer, Michael Schroeck
  - IBM Press
  - 2010
- **Data Integration Blueprint and Modeling**

  Techniques for a Sclable and Sustainable Architecture
  - Anthony David Giordano
  - IBM Press
  - 2011
- **DAMA-DMBOK data management body of knowledge**
  - DAMA International
  - Bradley Beach, New Jersey Technics Publications Data Management Association
  - 2017