**Seis claves para elegir la arquitectura ETL**

Este artículo describe seis decisiones clave que deben incluirse durante la elaboración de la arquitectura ETL para un almacén de datos dimensional. Estas decisiones tienen impactos importantes en el coste inicial y coste corriente y en la complejidad de la solución ETL y, finalmente, en el éxito de toda la solución BI/DW en global. Lee el consejo del grupo Kimball sobre realizar las elecciones correctas.

1. **¿Debemos utilizar una erramienta ETL?**

Una de las primeras y más fundamentales decisiones que debes tomar es si personalizar tu sistema ETL o utilizar un pack suministrado por un proveedor. Dejando de lado las cuestiones técnicas y los costes de de licencia, no debes adoptar un método que a tus empleados y directores no les sea familiar sin considerar seriamente las implicaciones a largo plazo de estas decisiones. Esta decisión tendrá un mayor impacto en el contexto del ETL, guiando decisiones respecto al personal, diseñando enfoques, estrategias de metadatos, y implementando cronologías a largo plazo.

En el contexto actual, la mayoría de organizaciones deben usar un pack suministrado por un proveedor por regla general. Sin embargo, esta decisión debe tomarse en base a los recursos disponibles para construir y dirigir el sistema. Las herramientas ETL construyen entornos que usan iconos, flechas y propiedades en lugar de escribir código para construir la solución ETL. Ten cuidado, si tu equipo de desarrollo del ETL está integrado por cierto número de programadores de la vieja escuela, es posible que no se adapten bien a una herramienta ETL. Sólo por esta razón, algunas organizaciones piensan que un ETL personalizado es una solución razonable.

Si decides utilizar una herramienta ETL, no esperes una gran amortización en tu primera iteración. Las ventajas aparecerán a medida que realices iteraciones adicionales y empieces a aprovechar las ventajas de utilizar la herramienta en las siguientes implementaciones. También experimentarás los beneficios de la mejorada capacidad de mantenimiento, documentación más completa y soporte mejorado para los metadatos.

1. **¿Dónde y cómo debe tener lugar la integración de datos?**

La integración de datos es un gran tópico para la IT porque, básicamente, se trata de hacer que todos los sistemas trabajen juntos eficientemente. La visión de 360º de la empresa, es un objetivo comúnmente discutido que realmente implica la integración de datos. En muchos casos, la integración de datos debe realizarse entre los sistemas de transacciones primarios de la organización antes de que los datos lleguen al almacén de datos. Sin embargo, antes que enfrentarse a la integración en el entorno operacional, estos requisitos son normalmente devueltos al almacén de datos y al sistema ETL.

La mayoría de nosotros entendemos el concepto principal de que la integración permite que bases de datos dispares funcionen juntas de una manera útil. Todos sabemos que lo necesitamos, simplemente no tenemos una idea clara de cómo desglosar las en partes manejables.

Integración significa que todas las partes de grandes organizaciones estén de acuerdo con cada dato o sólo con alguno de ellos? Este es el punto crucial de la decisión que se debe tomar. En qué nivel de datos la dirección empresarial está de acuerdo/ insiste que suceda? Están dispuestos a establecer definiciones comunes entre la organización y atenerse a esas definiciones?

Fundamentalmente, integración significa llegar a un acuerdo en el sentido de los datos desde la perspectiva de dos o más bases de datos. Con la integración, los resultados de dos bases de datos pueden ser combinados en un solo análisis del almacén de datos. Sin un acuerdo, las bases de datos permanecerán aisladas y no se podrán enlazar en una aplicación. En el contexto de nuestro entorno ETL, la integración de datos toma la forma de las dimensiones acordadas y de los hechos establecidos en el almacén de datos. Las dimensiones acordadas hacen referencia a establecer atributos dimensionales comunes entre tablas de hechos separadas de forma que informes de navegación puedan ser generados utilizando estos atributos. Los hechos establecidos hacen referencia a acuerdos en métricas comunes de negocios como los indicadores de actividad (KPIs) entre bases de datos separadas, para que estos números puedas ser comparados matemáticamente para calcular diferencias y ratios.

1. **¿Qué mecanismo de recopilación de cambios de datos debemos escoger?**

Durante la carga inicial del histórico de datos del almacén de datos, recopilar cambios contenidos en las fuentes de datos no es importante ya que estás cargando datos en de un punto futuro. Sin embargo, la mayoría de tablas del almacén de datos son tan extensas que no pueden ser actualizadas durante cada ciclo ETL. Debes tener la capacidad de transferir sólo los cambios relevantes para la fuente de datos desde la última actualización. Aislar la última fuente de datos se denomina cambiar la recopilación de datos (CDC). La idea detrás de cambiar la recopilación de datos es bastante simple: simplemente transferir los datos que han sido cambiados desde la última actualización.

Pero construir un buen sistema de recopilación de cambio de datos no es tan fácil como suena. Démonos cuenta que el mecanismo seleccionado debe ser totalmente hermético-todos los datos cambiados deben ser identificados. Encontrar la estrategia más completa puede ser evasivo, muchas veces la actualización de los sistemas de tablas de la fuente puede ocurrir fuera de la misma aplicación. Un error aquí, provocará resultados que no podrán ser explicados fácilmente; frecuentemente esto deriva en la recuperación de datos para identificar el culpable.

Los problemas pueden significar elevados costes en términos de rehacer el trabajo-sin mencionar la situación embarazosa. En resumen, recopilar los cambios de datos está lejos de ser una tarea trivial, y debes entender muy bien los sistemas de fuente de datos. Este conocimiento ayudará al equipo ETL a evaluar las fuentes de datos, identificar problemas en la recopilación de los cambios de datos y determinar la estrategia más apropiada.

1. **Cuando debemos organizar los datos?**

En el entorno actual de almacenamiento de datos, es bastante posible para las herramientas ETL establecer una conexión directa con la fuente de la base de datos, extraer y agrupar los datos a través de la herramienta ETL para aplicar cualquier transformación requerida en la memoria, y finalmente escribirla, sólo una vez, en la tabla objetivo del almacén de datos. Desde el punto de vista de la ejecución, esta es una gran capacidad ya que escribe, aquellos registrados específicamente escriben en la RDBMS, pero son muy caros; minimizarlos es un buen objetivo a diseñar.

Sin embargo, a pesar de las ventajas en la ejecución, este no debe ser el mejor planteamiento. Hay diversas razones que una organización debe decidir para físicamente organizar los datos (por ejemplo, escribirlos en un disco) durante el proceso ETL:

* El método CDC más apropiado requiere comparar la copia actual de la tabla fuente con la copia precedente de la misma tabla.
* La organización ha elegido organizar los datos inmediatamente después de la extracción con el propósito de archivar-posiblemente para cumplir los requerimientos de conformidad y control.
* El punto de recuperación/reinicio se require cuando el ETL falla a mitad del trabajo.
* Los procesos ETL que se desarrollan durante largos períodos pueden abrir una conexión don el sistema fuente que crea problemas con bloqueos de la base de datos y sobrecarga el sistema de transacciones.

1. **¿Dónde debemos corregir los datos?**

Los usuarios de negocios son conscientes de que la calidad de los datos es un problema serio y caro. Así que, a la mayoría de organizaciones les gusta apoyar iniciativas para mejorar la calidad de los datos. Pero la mayoría de los usuarios probablemente no tienen ni idea de dónde se originan los problemas de calidad de los datos o que se debe hacer para mejorar la calidad de los datos. Deben pensar que la calidad de los datos es un simple problema de ejecución del equipo ETL. En este entorno, el equipo ETL necesita ser ágil y proactivo: la calidad de los datos no puede ser mejorada solamente por el ETL. La clave está en que el equipo de ETL colabore con la empresa y los equipos de soporte del sistema fuente del IT.

La decisión clave es dónde corregir los datos. Claramente, la mejor solución es en primer lugar, tener los datos recopilados cuidadosamente. Por supuesto, este no es siempre el caso, pero no obstante, en la mayoría de casos. Los datos deben ser corregido volviendo al sistema fuente.

Desafortunadamente, es inevitable que datos de poca calidad lleguen al sistema ETL. En este caso, hay tres opciones:

1) Detener el proceso de carga.

2)Enviar los registros dañados a un archivo temporal para su posterior procesamiento.

3) Únicamente etiquetar los datos y continuar

La tercera opción es por mucho la mejor, siempre que sea posible. Detener el proceso, es obviamente doloroso porque requiere intervención manual para diagnosticar el problema, reiniciar o reanudar el trabajo, o abortar completamente. Enviar registros a un archivo temporal siempre es una solución pobre porque no está claro cuándo o si estos registros serán fijados y reintroducidos en el canal. Hasta que los registros son restaurados al flujo de datos, la integridad global de la base de datos es cuestionable ya que estos registros están perdidos. Recomiendo no utilizar el archivo temporal para infracciones menores de datos. La tercera opción de etiquetar los datos con la condición de error normalmente funciona bien. La tabla de hechos errónea puede ser etiquetada con una dimensión de control que describe la condición de calidad global de la fila del hecho problemático. Los datos mal dimensionados también pueden ser etiquetados con un valores únicos de error en el mismo campo. En este punto, las capacidades en la calidad de los registros pueden indentificar el hecho problemático y las filas de dimensión, indicando la necesidad para resolución y reparar los datos en el sistema fuente.

# 6. Con que rapidez debe estar disponible la fuente de datos vía el sistema DW/BI?

La latencia de los datos describe con qué rapidez el sistema de datos debe ser proporcionado a los usuarios de negocios vía el sistema DW/BI. La latencia de los datos obviamente tiene un gran efecto en el coste y la complejidad de tu entorno ETL. Procesadores inteligentes de algoritmos, paralelización y hardware potente pueden acelerar flujos de datos tradicionalmente orientados a lotes. Pero en algún momento, si el requerimiento de la latencia de datos es suficientemente urgente, el sistema de arquitectura ETL debe pasar de estar orientado a lotes a agrupaciones. Este cambio no es gradual ni evolutivo, es un mayo paradigma de cambio donde casi cada paso en el canal de entrega de datos debe ser re-implementado.

Habitualmente, las agrupaciones ETL requieren latencia de datos en la mayoría de organizaciones lo que enlaza con el ritmo natural del negocio. Encontramos que en la mayoría de organizaciones esta práctica habitual, se traduce en actualizaciones diarias para la mayoría de agrupaciones ETL and actualizaciones semanales o mensuales para otras agrupaciones ETL.

Sinembargo, en algunas circunstancias, actualizaciones más frecuentes o incluso actualizaciones en tiempo real se ciñen mejor al ritmo del negocio. La clave es reconocer que sólo un puñado de procesos empresariales dentro de una organización son apropiados para realizar actualizaciones en tiempo real. No hay ninguna razón de peso para convertir todo el proceso ETL a tiempo real. El ritmo de la mayoría de procesos empresariales simplemente no demandan un tratamiento en tiempo real.

Hay que tener cuidado: preguntando a los usuarios de negocios si quieren la entrega de datos en tiempo real es una invitación abierta para los problemas. Por supuesto, la mayoría de usuarios de negocios responderán positivamente a tener los datos actualizados más frecuentemente, independientemente de si entienden el impacto de su petición. Claramente este tipo de requerimiento de latencia de datos puede ser peligroso. Recomendamos dividir la dificultad del tiempo real en tres categorías: diaria, frecuente e instantánea. Necesitas obtener los usuarios finales para describir sus requerimientos en cuanto a latencia de datos en términos similares y después diseñar tu solución ETL apropiadamente para respaldar cada set de requerimientos:

* Instantaneo significa que los datos visibles en la pantalla del usuario final representan el estado real del sistema de fuente de transacción en cada instante. Cuando el estado del sistema de la fuente cambia, la pantalla online debe también responder instantáneamente.
* Frecuentemente significa que los datos visibles por el usuario final, se actualizan muchas veces al día pero no se garantiza que representen el valor correcto absoluto en cada instante.
* Diariamente significa que los datos visibles en la pantalla son válidos como una descarga de archivo por lotes o armonización del sistema fuente al final del día anterior laborable.

En este artículo, hemos discutido cierto número de claves de decisión que deben ser evaluadas durante la creación de una arquitectura ETL para mantener un almacén de datos dimensional.

Pocas veces hay una sola elección correcta para cualquiera de estas decisiones, como siempre, la decisión correcta será conducida por las necesidades y características de tu empresa.