Septiembre 2014

Integración Big Data y Hadoop

Mejores prácticas para minimizar riesgos y maximizar el retorno de la inversión (ROI) en iniciativas Hadoop.

IBM

Introducción

La tecnología Apache Hadoop está transformando la economía y la dinámica de las iniciativas Big Data al dar soporte a nuevos procesos y arquitecturas que pueden ayudar a reducir costos, aumentar los ingresos y crear ventajas competitivas. Al ser un proyecto de software de fuente abierta que habilita el procesamiento distribuido y el almacenamiento de grandes conjuntos de datos entre clusters de servidores commodity, Hadoop puede aumentar la escala de un solo servidor a miles de servidores, adaptándose a cambios en la demanda. Los componentes primarios de Hadoop incluyen el Hadoop Distributed File System para almacenar grandes archivos y el marco de procesamiento paralelo distribuido Hadoop (conocido como MapReduce).

Sin embargo, por sí misma, la infraestructura Hadoop no presenta una solución completa de integración Big Data, y se plantean tanto retos como oportunidades que es preciso abordar antes de poder cosechar sus beneficios y maximizar el retorno sobre la inversión (ROI).

La importancia de integración Big Data para iniciativas Hadoop

El rápido surgimiento de Hadoop está impulsando un cambio paradigmático en la forma en que las organizaciones ingieren, administran, transforman, almacenan y analizan Big Data. Es posible acceder a una analítica más profunda, mayores conocimientos, nuevos productos y servicios, y mayores niveles de servicio a través de esta tecnología, lo cual le permite reducir costos significativamente y generar nuevos ingresos.

Los proyectos Big Data y Hadoop dependen de recopilar, mover, transformar, depurar, integrar, gobernar, explorar y analizar volúmenes masivos de distintos tipos de datos de muchas fuentes diferentes. Para lograr todo eso se necesita una solución de integración de la información resiliente, extremo a extremo, que sea masivamente escalable y proporcione la infraestructura, las capacidades, los procesos y la disciplina que se necesitan para dar soporte a proyectos Hadoop.

"En la mayoría de los casos, el 80% del esfuerzo de desarrollo en un proyecto Big Data corresponde a integración de datos y solo el 20%, a análisis de datos."

—Intel Corporation, "Extraer, transformar y cargar Big Data con Apache Hadoop"[1]

Una solución efectiva de integración Big Data ofrece simplicidad, velocidad, escalabilidad, funcionalidad y gobernanza para extraer los datos consumibles del lago Hadoop. Sin una integración eficaz, "entra basura, sale basura"… esta no es una buena receta para obtener datos confiables, y mucho menos conocimientos exactos y completos, o resultados transformacionales.

Tras haber estudiado la evolución del mercado Hadoop, los analistas de tecnología líderes coinciden en que la infraestructura Hadoop por sí sola no es una solución de integración de datos completa o eficaz. Para complicar aún más la situación, algunos proveedores de software Hadoop han saturado el mercado con exageraciones, mitos e información engañosa o contradictoria.

Para poner las cosas en claro y desarrollar un plan de adopción para su proyecto Big Data Hadoop, le recomendamos seguir un enfoque con mejores prácticas que tenga en cuenta a las tecnologías emergentes, a los requisitos de escalabilidad y a los actuales recursos y niveles de habilidades. El desafío: crear un enfoque y una arquitectura de integración Big Data optimizada, y mientras tanto evitar escollos en la implementación.

Escalabilidad masiva de datos: El requisito por excelencia

Si su solución de integración Big Data no puede respaldar una escalabilidad masiva de datos, tal vez no alcance a cumplir las expectativas. Para obtener todo el valor de negocios de las iniciativas Big Data, la escalabilidad masiva es esencial para la integración Big Data en la mayoría de los proyectos Hadoop. La escalabilidad masiva de datos significa que no hay límites en los volúmenes de datos procesados, la capacidad de procesamiento o la cantidad de procesadores y nodos de procesamiento utilizados. Usted puede procesar más datos y lograr una capacidad de procesamiento mayor simplemente agregando más hardware. La misma aplicación luego funcionará sin modificación y con un desempeño superior, conforme se agreguen recursos de hardware (ver la Figura 1).

Factor crítico de éxito: Evitar las exageraciones, para distinguir la realidad de la ficción

Durante estas etapas emergentes del mercado Hadoop, haga su propio análisis de todo lo que se dice sobre la capacidad de Hadoop. Existe una brecha significativa entre los mitos y las realidades en la explotación de Hadoop, particularmente en lo que se refiere a la integración Big Data. Se suele afirmar que cualquier herramienta de extraer, transformar y cargar (ETL) no escalable más Hadoop equivale a una plataforma de integración de datos de alto desempeño y alta escalabilidad.

En realidad, MapReduce no se diseñó para el procesamiento de alto desempeño de volúmenes masivos de datos, sino para la tolerancia a fallas de granularidad fina. Esa discrepancia puede bajar el desempeño y la eficiencia general en una orden de magnitud o más.

Hadoop Yet Another Resource Negotiator (YARN) **toma las capacidades de gestión de recursos que estaban en MapReduce y las presenta de tal modo que pueden utilizarlas otras aplicaciones que necesitan ejecutarse dinámicamente en todo el cluster Hadoop. Como resultado, este enfoque hace posible implementar masivamente motores escalables de integración de datos como aplicaciones Hadoop nativas, sin las limitaciones de desempeño de MapReduce. Todas las tecnologías empresariales que buscan ser escalables y eficientes en Hadoop deberán adoptar YARN como parte de su hoja de ruta de producto.**

Antes de iniciar su travesía de integración, asegúrese de comprender las limitaciones de desempeño de MapReduce y la forma en que distintos proveedores de integración de datos las abordan.

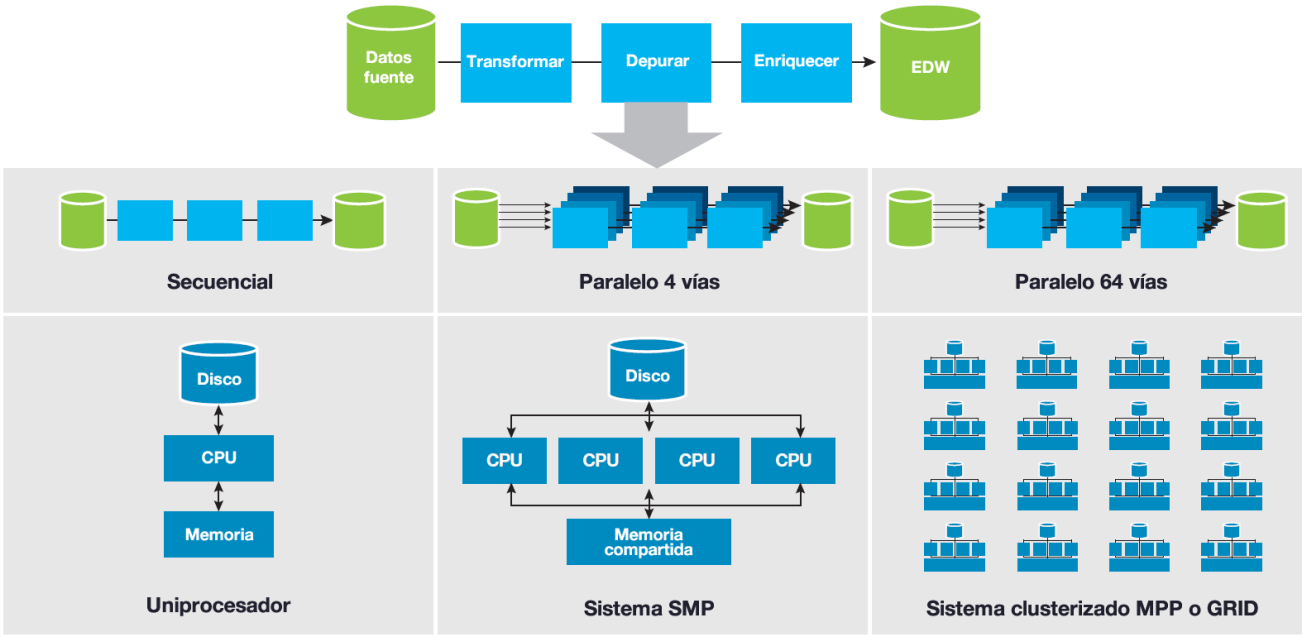

Figura 1. La escalabilidad masiva de datos es un requisito obligatorio para la integración Big Data. En la era Big Data, las organizaciones deben poder dar soporte a un sistema clusterizado MPP a escala.

Factor crítico del éxito: Las plataformas de integración Big Data deben dar soporte a tres dimensiones de escalabilidad

- Escalabilidad lineal de datos: Un sistema de hardware y software ofrece aumentos lineales en la capacidad de procesamiento, con aumentos lineales en recursos de hardware. Por ejemplo, una aplicación ofrece escalabilidad lineal si puede procesar 200 GB de datos en cuatro horas con 50 procesadores, 400 GB de datos en cuatro horas con 100 procesadores, etc.
- Scale-up de aplicaciones: Una medición de la efectividad con la que el software alcanza la escalabilidad lineal de datos entre procesadores dentro de un sistema multiprocesador simétrico (SMP).
- Scale-out de aplicaciones: Una determinación de cuán bien el software alcanza la escalabilidad lineal de datos entre nodos SMP en una arquitectura en donde nada se comparte.

Los requisitos para dar soporte a la escalabilidad masiva de datos no sólo se vinculan con el surgimiento de la infraestructura Hadoop. Los proveedores líderes de data warehouse, como IBM y Teradata, y las plataformas líderes de integración de datos, como IBM® InfoSphere® Information Server, ha proporcionado plataformas de software masivamente paralelo donde nada se comparte durante años, en algunos casos durante casi dos décadas.

Con el tiempo, estos proveedores convergieron en cuatro características de arquitectura de software comunes que dan soporte a una escalabilidad masiva de datos, tal como se muestra en la Figura 2.

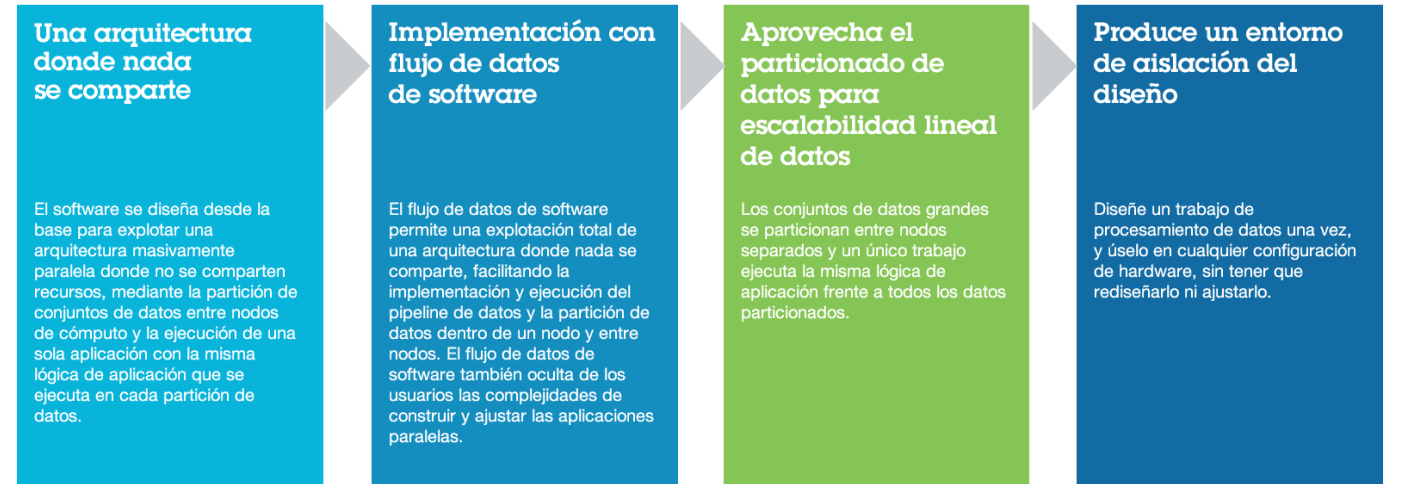

Figura 2. Las cuatro características de la escalabilidad masiva de datos.

La mayoría de las plataformas comerciales de software de integración de datos nunca fueron diseñadas para dar soporte a una escalabilidad masiva de información, lo cual significa que no fueron construidas desde la base para explotar una arquitectura masivamente paralela sin compartir. Dependen de la tecnología de multihilo de memoria compartida en lugar del flujo de datos de software.

Asimismo, algunos proveedores no dan soporte a la partición de grandes conjuntos de datos entre nodos y a la ejecución de un único trabajo de integración en paralelo frente a particiones de datos separadas, o la capacidad de diseñar un trabajo una vez y usarlo en cualquier configuración de hardware sin tener que rediseñarlo y reajustarlo. Estas capacidades son críticas para reducir costos, al ganar eficiencia. Sin ellas, la plataforma no podría funcionar con grandes volúmenes de datos.

La cartera de integración de datos InfoSphere Information Server da soporte a las cuatro características de arquitectura de escalabilidad masiva de datos.

Optimizando las cargas de trabajo de integración Big Data: Un enfoque equilibrado

Como casi todos los casos de uso y escenarios Big Data Hadoop primero requieren integración Big Data, las organizaciones deben determinar cómo optimizar estas cargas de trabajo a lo largo de la empresa. Uno de los casos de uso líderes para Hadoop e integración Big Data es descargar las cargas ETL del data warehouse de la empresa (EDW) para reducir costos y mejorar los acuerdos de nivel de servicio (SLA) de consultas. Ese caso de uso plantea las siguientes preguntas:

- ¿Deberían todas las organizaciones descargar las cargas ETL del EDW?
- ¿Deberían todas las cargas de trabajo de integración de datos enviarse a Hadoop?
- ¿Cuál es el rol continuo para las cargas de trabajo de integración de datos en un grid ETL sin un sistema de gestión de base de datos relacional paralelo (RDBMS) y sin Hadoop?

La respuesta correcta a estas preguntas depende de los requisitos Big Data exclusivos de una empresa. Las organizaciones pueden elegir entre un RDBMS paralelo, Hadoop y un grid ETL escalable para ejecutar cargas de trabajo de integración Big Data. Pero más allá del método que elijan, la infraestructura de información debe cumplir un requisito común: soporte completo para procesamiento masivamente escalable.

Algunas operaciones de integración de datos se ejecutan de manera más eficiente dentro o fuera del motor RDBMS. Del mismo modo, no todas las operaciones de integración de datos se adaptan al entorno Hadoop. Una arquitectura bien diseñada debe ser suficientemente flexible como para apalancar las fortalezas de cada entrono en el sistema (ver la Figura 3).

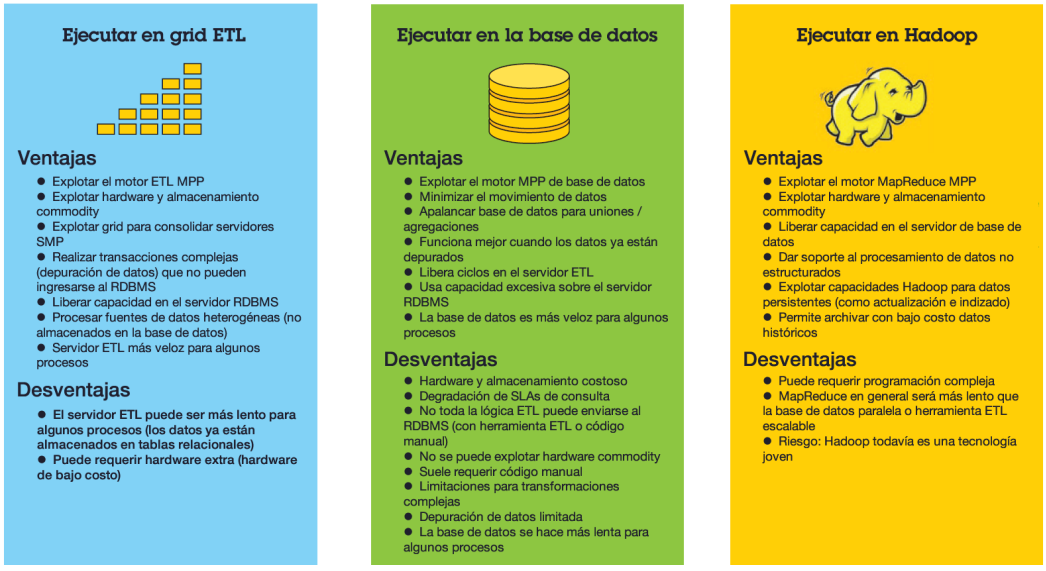

Figura 3. La integración Big Data requiere un enfoque equilibrado que pueda apalancar la fortaleza de cualquier entorno.

Hay tres pautas importantes a seguir cuando se optimizan las cargas de trabajo de integración Big Data:

1. **Llevar el procesamiento e integración Big Data a los datos, en lugar de llevar los datos al procesamiento:** Especificar procesos apropiados que puedan ejecutarse ya sea en RDBMS, en Hadoop o en grid ETL.
2. **Evitar el desarrollo de código manual:** El desarrollo de código manual es caro y no da soporte en forma eficaz a cambios rápidos o frecuentes. Tampoco da soporte a la recopilación automatizada de metadatos operativos y de diseño que son críticos para la gobernanza de datos.
3. **No mantener silos separados de desarrollo de integración para RDBMS, Hadoop y grid ETL:** Esto no cumple ningún fin práctico y se hace muy costoso de soportar. Se debería poder construir un trabajo una vez y ejecutarlo en cualquiera de los tres entornos.

Procesos que se adaptan mejor a Hadoop

Las plataformas Hadoop comprenden dos componentes primarios: un sistema de archivos distribuido, tolerante a fallas, llamado Hadoop Distributed File System (HDFS), y un marco de procesamiento paralelo llamado MapReduce.

La plataforma HDFS es muy buena para procesar grandes operaciones secuenciales, en donde una "porción" de datos leída suele ser 64 MB o 128 MB. En general, los archivos HDFS no se particionan ni ordenan a menos que la aplicación que carga los datos gestione esto. Aunque la aplicación pueda partir y ordenar las porciones de datos resultantes, no hay forma de garantizar que esa porción sea colocada en el sistema HDFS. Esto significa que no hay una buena manera de administrar la colocación de datos en este entorno. La colocación de datos es crítica porque asegura que los datos con las mismas claves de unión (join keys) terminen en los mismos nodos, de modo que el proceso tenga un alto desempeño y al mismo tiempo sea exacto.

Si bien hay formas de acomodar la falta de soporte para colocación de datos, esto tiende a tener un alto costo, que en general requiere procesamiento extra y/o reestructuración de la aplicación. Los archivos HDFS también son inmutables (solo lectura) y procesar un archivo HDFS es similar a ejecutar un scan de tabla completa, ya que con frecuencia se procesa la totalidad de los datos. Esto debería disparar un indicador de alerta de inmediato para operaciones como la unión de dos tablas muy grandes, ya que es probable que los datos no sean colocados en el mismo nodo Hadoop.

MapReduce Versión 1 es un marco de procesamiento paralelo que no fue diseñado específicamente para procesar grandes cargas de trabajo ETL con alto desempeño. Por omisión, puede hacerse una re-partición o re-colocación de datos entre la fase de mapeo y de reducción del procesamiento. Para facilitar la recuperación, los datos aterrizan en el nodo que ejecuta la operación de mapeo antes de ser mezclados y enviados a la operación de reducción.

MapReduce contiene facilidades para mover estructuras de datos de referencia menores a cada nodo de mapeo para algunas operaciones de validación y optimización. Por lo tanto, el archivo completo de referencia se pasa a cada nodo de mapeo, lo cual lo hace más apropiado para las estructuras de datos de referencia menores. Si se está desarrollando código manual, se deben tener en cuenta estos flujos de procesamiento, de modo que es mejor adoptar herramientas que generan código para impulsar la lógica de integración de datos a MapReduce (también conocido como *pushdown ETL*).

El uso del procesamiento pushdown ETL en Hadoop (sin importar qué herramienta lo realiza) puede crear una situación en la que una porción no trivial del procesamiento de integración de datos debe seguir ejecutándose en el motor ETL y no en MapReduce. Esto es válido por varios motivos:

- Una lógica más compleja no se puede enviar a MapReduce
- MapReduce tiene importantes limitaciones de desempeño
- Los datos en general se almacenan en HDFS en forma secuencial aleatoria

Todos estos factores sugieren que la integración Big Data en el entorno Hadoop requiere tres componentes para el procesamiento de cargas de trabajo de alto desempeño:

1) Una distribución Hadoop
2) Una plataforma ETL masivamente escalable en la que no se comparte nada (como la ofrecida por IBM InfoSphere Information Server)
3) Capacidad pushdown ETL en MapReduce

Los tres componentes son necesarios porque un gran porcentaje de la lógica de integración de datos no puede enviarse a MapReduce sin código manual, y porque MapReduce tiene limitaciones conocidas en cuanto a desempeño.

Factor crítico del éxito: Considerar las velocidades de procesamiento de cargas de trabajo de integración de datos

La arquitectura masivamente paralela, donde nada se comparte, de InfoSphere Information Server es optimizada para procesar grandes cargas de trabajo de integración de datos con alto desempeño. IBM InfoSphere DataStage®—como parte de InfoSphere Information Server, que integra datos a través de múltiples sistemas utilizando un marco paralelo de alto desempeño – puede procesar cargas de trabajo de integración típicas de 10 a 15 minutos más rápido que MapReduce.[2]

InfoSphere DataStage también ofrece una optimización balanceada para el entorno Hadoop. La optimización balanceada genera código Jaql para ejecutarse nativamente en el entorno MapReduce. Jaql viene con un optimizador que analiza el código generado y lo optimiza en un componente de mapeo y un componente de reducción. Esto automatiza una tarea de desarrollo tradicionalmente compleja y libera al desarrollador, que ya no tiene que preocuparse por la arquitectura MapReduce.

InfoSphere DataStage puede funcionar directamente en los nodos Hadoop en lugar de en un nodo separado en la configuración, que es lo que requieren algunas implementaciones de proveedores. Esta capacidad ayuda a reducir el tráfico de red, cuando se acopla con IBM General Parallel File System (GPFS™)-FPO, que proporciona un subsistema de almacenamiento que cumple con POSIX en el entorno Hadoop. Un sistema de archivos POSIX permite que los trabajos ETL accedan directamente a datos almacenados en Hadoop, en lugar de requerir el uso de la interfaz HDFS. Este entorno da soporte a la transferencia de la carga de trabajo ETL al entorno de hardware en el que Hadoop se está ejecutando, y ayuda a mover el procesamiento adonde los datos están almacenados y a apalancar el hardware tanto para Hadoop como para procesamiento ETL.

Los sistemas de gestión de recursos, tales como IBM Platform™ Symphony también pueden usarse para administrar cargas de trabajo de integración de datos tanto dentro como fuera del entorno Hadoop.

Esto significa que a pesar de que InfoSphere DataStage tal vez no se ejecuta en el nodo exacto de los datos, se ejecute en el mismo plano de fondo de alta velocidad, eliminando la necesidad de sacar los datos del entorno Hadoop por conexiones de red más lentas.

Requisitos de escalabilidad ETL para dar soporte a Hadoop

Muchos proveedores de software Hadoop difunden la idea de que cualquier herramienta ETL no escalable con pushdown a MapReduce proporcionará un desempeño excelente y scale-out de aplicaciones para la integración de grandes volúmenes de datos, pero eso simplemente no es así.

Sin un motor ETL masivamente escalable en donde nada se comparte, como InfoSphere DataStage, las organizaciones experimentarán limitaciones funcionales y de desempeño. Cada vez más organizaciones advierten que las herramientas ETL no escalables de la competencia con pushdown a MapReduce no son capaces de proporcionar los niveles requeridos de desempeño en Hadoop. Están trabajando con IBM para abordar esta cuestión porque la solución de integración Big Data de IBM da un soporte distintivo a una escalabilidad masiva de datos para integración Big Data.

Algunos de los efectos negativos acumulativos de una dependencia excesiva de pushdown ETL:

- ETL comprime un gran porcentaje de la carga de trabajo EDW. Debido a los costos asociados, EDW es una plataforma muy costosa para ejecutar cargas de trabajo ETL.
- Las cargas de trabajo ETL causan degradación en SLAs de consulta, y con el tiempo requieren invertir en capacidad EDW adicional y de alto costo.
- Los datos no son depurados antes de ser volcados al EDW y nunca se depuran en el entorno EDW, lo cual promueve una mala calidad de datos.
- La organización sigue dependiendo demasiado del desarrollo de código manual de scripts SQL para transformaciones de datos.
- Agregar nuevas fuentes de datos o modificar los scripts ETL existentes es costoso y lleva mucho tiempo, lo cual limita la capacidad de responder rápidamente a nuevos requisitos.
- Las transformaciones de datos son relativamente simples porque no es posible enviar una lógica más compleja al RDBMS utilizando una herramienta ETL.
- Sufre la calidad de datos.
- Las tareas críticas, tales como perfiles de datos, no están automatizadas y en muchos casos no pueden realizarse.
- No se implementa una gobernanza de datos significativa (data stewardship, linaje de datos, análisis de impacto), lo cual dificulta y encarece la respuesta a los requisitos regulatorios y la confianza en información crítica para el negocio.

Por el contrario, las organizaciones que adoptan plataformas de integración de datos masivamente escalables que optimizan las cargas de trabajo de integración Big Data minimizan los potenciales efectos negativos, y quedan mejor posicionadas para transformar su negocio con Big Data.

Mejores prácticas para integración Big Data

Una vez que usted decide adoptar Hadoop para sus iniciativas Big Data, ¿cómo implementa los proyectos de integración de grandes volúmenes de datos, protegiéndose de la variabilidad de Hadoop?

Gracias al trabajo con muchos adoptantes tempranos de la tecnología Hadoop, IBM identificó cinco mejores prácticas de integración Big Data fundamentales. Estos cinco principios representan enfoques de excelencia para iniciativas exitosas de integración Big Data:

1. Evitar el desarrollo de código manual en cualquier parte, por cualquier motivo
2. Una plataforma de integración y gobernanza de datos para la empresa
3. Integración de datos masivamente escalable, disponible donde sea necesario ejecutarla
4. Gobernanza de datos de clase mundial en toda la empresa
5. Administración robusta y control de operaciones en toda la empresa

"Ante la duda, utilice las herramientas de mayor nivel siempre que sea posible."

— "ETL a gran escala con Hadoop," Presentación en Strata+Hadoop World 2012, dada por Eric Sammer, Arquitecto Principal de Soluciones de Cloudera[4]

Mejor práctica #1: Evitar el desarrollo de código manual en cualquier parte, por cualquier motivo

Durante las dos últimas décadas, las grandes organizaciones reconocieron las numerosas ventajas de reemplazar el código manual por herramientas comerciales de integración de datos. El debate entre código manual versus herramientas de integración de datos ya quedó aclarado, y muchos analistas de tecnología resumieron las importantes ventajas en cuanto al retorno de la inversión (ROI)[3] que se obtienen con la adopción de software de integración de datos de clase mundial.

La primera práctica recomendada es evitar el desarrollo de código manual en cualquier parte, para cualquier aspecto de la integración Big Data. En cambio, se recomienda aprovechar interfaces gráficas de usuario para actividades como:

- Acceso y movimiento de datos en toda la empresa
- Lógica de integración de datos
- Reunir trabajos de integración de datos de objetos lógicos

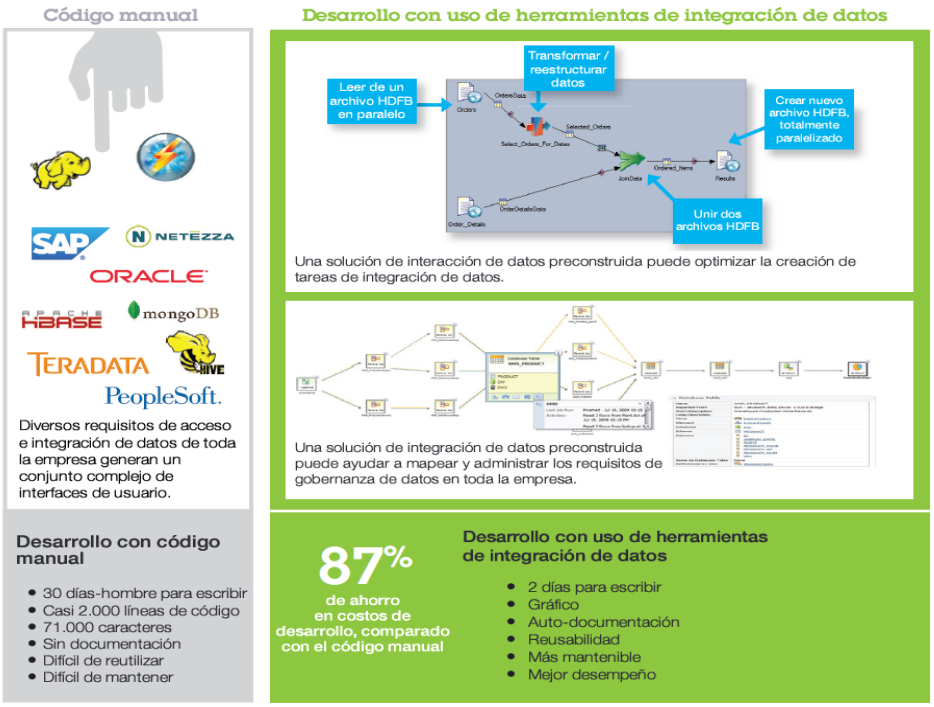

Figura 4. El software de integración de datos proporciona múltiples GUIs para dar soporte a diversas actividades. Estas GUIs reemplazan el desarrollo de código manual complejo y ahorran a las organizaciones altas sumas en costos de desarrollo.

- Reunir flujos de trabajo mayores
- Gobernanza de datos
- Gestión operativa y administrativa

Al adoptar esta práctica, las organizaciones pueden explotar la productividad comprobada, el costo, por el valor del tiempo y las robustas ventajas de control operativo y administrativo del software comercial de integración de datos, y evitar, así, el impacto negativo del desarrollo de código manual (ver la Figura 4).

Mejor práctica #2: Una plataforma de integración y gobernanza de datos para la empresa

Una excesiva dependencia del pasaje de ETL al RDBMS (debido a falta de herramientas de software de integración de datos escalables) ha impedido a muchas organizaciones reemplazar script y código SQL y establecer una gobernanza de datos significativa en toda la empresa. Sin embargo, reconocen que hay grandes ahorros de costos asociados a pasar grandes cargas de trabajo ETL del RDBMS a Hadoop. Pero pasar de un silo de codificación manual ETL en el RDBMS a un nuevo silo de codificación manual de ETL y Hadoop solo duplica los altos costos y los tiempos de entrega.

Implementar una única plataforma de integración de datos ofrece la oportunidad de la organización transformacional, mediante la habilidad para:

- **Construir un trabajo una vez y ejecutarlo en todas partes,** en cualquier plataforma en la empresa, sin modificación
- **Acceder, mover y cargar datos** entre una variedad de fuentes y destinos en toda la empresa
- **Dar soporte a una variedad de paradigmas de integración de datos,** incluso el procesamiento batch; federación; captura de datos de cambio; habilitación SOA de tareas de integración de datos; integración en tiempo real con integridad transaccional y/o integración de datos autoservicio para usuarios de negocio.

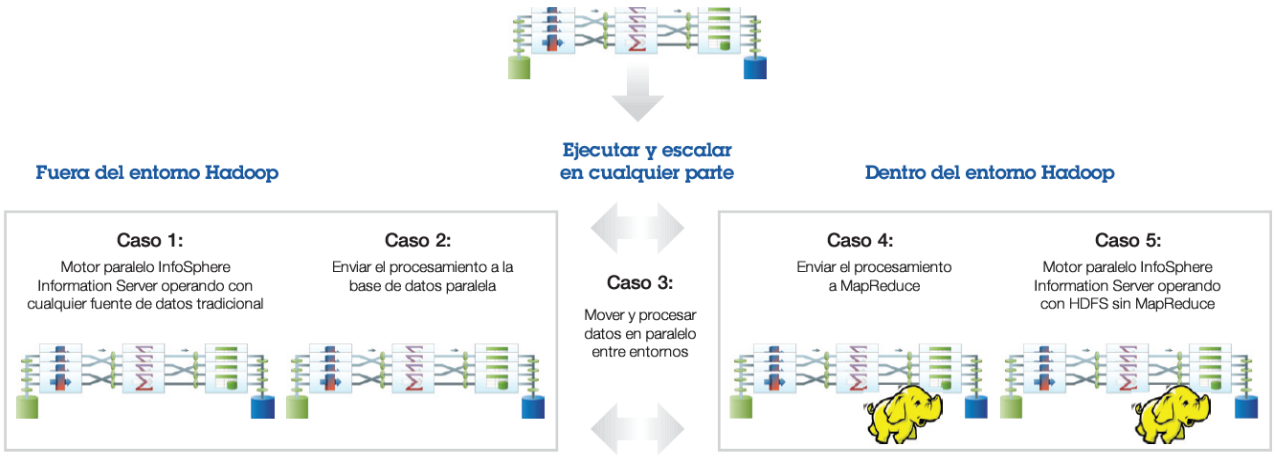

Figura 5. La integración escalable Big Data debe estar disponible para cualquier entorno.

También ofrece la oportunidad de establecer una gobernanza de datos de clase mundial, que incluye data stewardship, linaje de datos y análisis de impacto entre herramientas.

Mejor práctica #3: Integración de datos masivamente escalable, disponible donde sea necesario ejecutarla

Hadoop ofrece un potencial significativo para el procesamiento distribuido a gran escala de cargas de trabajo de integración, a un costos sumamente bajo. Sin embargo, los clientes necesitan una solución de integración masivamente escalable, para materializar las ventajas potenciales que Hadoop puede ofrecer.

Los escenarios para ejecutar la carga de trabajo de integración de datos pueden incluir:

- RDBMS paralelo
- Grid sin RDBMS o Hadoop
- En Hadoop, con o sin pushdown en MapReduce
- Entre el entorno Hadoop y el entorno externo, extraer volúmenes de datos de un lado, procesar y transformar los registros en vuelo y cargar los registros por el otro lado

Para alcanzar el éxito y la sustentabilidad -y mantener bajos costos- una solución de integración Big Data eficaz debe dar soporte flexible a cada uno de estos escenarios. Sobre la base de la experiencia de IBM con clientes Big Data, InfoSphere Information Server actualmente es la única plataforma de software de integración de datos comercial que da soporte a estos escenarios, incluso con un pushdown de lógica de integración de datos en MapReduce.

Hay muchos mitos que circulan dentro de la industria acerca de la ejecución de herramientas ETL en Hadoop para integración Big Data. La sabiduría popular parece indicar que combinar cualquier herramienta ETL no escalable y Hadoop proporciona todo el procesamiento de integración de datos masivamente escalable. En realidad, MapReduce sufre varias limitaciones para el procesamiento de cargas de trabajo de integración de datos a gran escala:

- No toda la lógica de integración de datos puede enviarse a MapReduce utilizando la herramienta ETL. Sobre la base de experiencias con sus clientes, IBM estima que aproximadamente el 50% de la lógica de integración de datos no puede enviarse a MapReduce.
- Los usuarios deben realizar código manual complejo para ejecutar una lógica de datos más compleja en Hadoop, o restringir el proceso para que ejecute transformaciones relativamente simples en MapReduce.
- MapReduce tiene limitaciones de desempeño conocidas para el procesamiento de grandes cargas de integración de datos, ya que fue diseñado para dar soporte a tolerancia a fallas de granularidad fina, a expensas del procesamiento de alto desempeño.

Mejor práctica #4: Gobernanza de datos de clase mundial en toda la empresa

A la mayoría de las grandes organizaciones les resulta difícil, sino imposible, establecer una gobernanza de datos en toda la organización. Hay varios motivos para ello. Por ejemplo, los usuarios de negocio administran datos utilizando terminología de negocios que les es familiar. Hasta hace poco, no se disponía de mecanismo para definir, controlar y administrar esta terminología de negocios y vincularla con activos de TI.

Además, ni los usuarios de negocio ni el personal de TI tienen un alto grado de confianza en sus datos, e incluso tal vez no conozcan con certeza su origen ni su historia. La tecnología para crear y administrar la gobernanza de datos a través de

capacidades tales como linaje de datos y análisis de impacto entre herramientas no existía, y los métodos manuales implican una complejidad abrumadora. Los requisitos regulatorios de la industria no hacen más que agregar más complejidad a la gestión de gobernanza. Finalmente, una confianza excesiva en el desarrollo de código manual para la integración de datos dificulta la implementación de gobernanza de datos en toda la organización.

Es esencial establecer una gobernanza de datos de clase mundial –con un ciclo de vida de datos totalmente gobernado para todos los activos de datos clave– que incluyen el entorno Hadoop pero no se limitan a él. Aquí hay algunos pasos sugeridos para un ciclo de vida de datos integral:

- **Encontrar:** Aprovechar términos, etiquetas y colecciones para encontrar fuentes de datos gobernadas y curadas
- **Curar:** Agregar etiquetas, términos y propiedades personalizadas a activos relevantes
- **Recopilar:** Usar colecciones para captar activos para un análisis específico o esfuerzo de gobernanza
- **Colaborar:** Compartir colecciones para realizar curaduría y gobernanza adicional
- **Gobernar:** Crear y referenciar políticas y reglas de gobernanza de información; aplicar calidad datos, enmascaramiento, archivado y depuración de datos
- **Descargar:** Copiar datos en un clic a HDFS para su análisis en la aumentación de warehouse
- **Analizar:** Analizar los datos descargados
- **Reutilizar y confiar:** Comprender cómo están siendo usado los datos, considerando su linaje para análisis e informes

Con una iniciativa integral de gobernanza de datos implementada, usted puede construir un entorno que ayude a asegurar que todos los datos Hadoop sean de alta calidad, seguros y adecuados para su uso. Permite a los usuarios de negocio responder preguntas del tipo:

- ¿Entiendo el contenido y el significado de estos datos?
- ¿Puedo medir la calidad de esta información?
- ¿De dónde vienen los datos de mi informe?
- ¿Qué se está haciendo con los datos en Hadoop?
- ¿Dónde estaban antes de llegar a nuestro lago de Hadoop?

Mejor práctica #5: Administración robusta y control de operaciones en toda la empresa

Las organizaciones que adoptan Hadoop para integración Big Data deben esperar una administración y gestión de operaciones robusta, de clase mainframe, que incluya:

- Una interfaz de consola de operaciones que ofrece respuestas rápidas para todos los que operen en aplicaciones de integración de datos, desarrolladores y otros participantes, mientras hacen un seguimiento del entorno
- Gestión de cargas de trabajo para asignar prioridad de recursos a ciertos proyectos en un entorno de servicios compartidos y poner en cola cargas de trabajo cuando el sistema está ocupado
- Análisis de desempeño para obtener conocimiento del consumo de recursos para identificar cuellos de botella y determinar cuando los sistemas pueden necesitar más recursos
- Construir cargas de trabajo que incluyen actividades basadas en Hadoop, definidas a través de Oozie directamente en la secuencia de trabajos, así como otras actividades de integración

La administración para la integración Big Data debe incluir:

- Un instalador basado en la web, integrado, para todas las capacidades
- Configuraciones de alta disponibilidad para requisitos de reuniones 24/7
- Opciones de implementación flexible para implementar nuevas instancias o expandir las instancias existentes en sistemas de hardware expertos y optimizados
- Autenticación, autorización y gestión de sesiones centralizadas
- Registro de auditoría para eventos relacionados con seguridad, para promover el cumplimiento de la ley Sarbanes-Oxley

• Certificación de laboratorio para diversas distribuciones Hadoop

Las mejores prácticas en integración Big Data sientan las bases del éxito

Las organizaciones están recurriendo a las iniciativas Big Data para reducir costos, aumentar los ingresos y obtener las ventajas que implica ser el primero en mover las fichas. La tecnología Hadoop apoya nuevos procesos y arquitecturas que habilitan la transformación de negocios, pero existen algunos retos y oportunidades en el área de Big Data que deben ser abordados antes de que dicha transformación pueda materializarse.

IBM recomienda construir una arquitectura de integración Big Data que sea lo suficientemente flexible como para apalancar las fortalezas de los RDBMS, grid ETL y entornos Hadoop. Los usuarios deben poder construir un flujo de trabajo de integración una vez y luego ejecutarlo en cualquiera de los tres entornos.

Las cinco mejores prácticas de integración Big Data delineadas en este documento representan enfoques de excelencia que permiten posicionar a sus proyectos para el éxito. Seguir estas pautas puede ayudar a su organización a minimizar los riesgos y los costos y a maximizar el ROI para sus proyectos Hadoop.

Más información

Para conocer más sobre mejores prácticas de integración y soluciones de integración IBM, contacte a su representante IBM o Asociado de Negocio IBM, o visite:
ibm.com/software/data/integration

Además, IBM Global Financing puede ayudarlo a adquirir las capacidades de software que su empresa necesita en la forma más económica y estratégica posible. Trabajamos con clientes con calificación crediticia para personalizar una solución de financiación que se adapte a los objetivos de su empresa, permita una administración eficaz del efectivo y mejore el costo total de propiedad. Financie sus inversiones críticas en TI e impulse su negocio hacia adelante con IBM Global Financing. Más información:
ibm.com/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2014

[1] Intel Corporation. "Extraer, transformar y cargar Big Data con Apache Hadoop". Julio de 2013. http://intel.ly/UX1Umk

[2] Mediciones producidas por IBM durante su trabajo in-situ con una implementación de cliente.

[3] International Technology Group. "Caso de estudio para estrategia de integración de datos: Comparación de IBM InfoSphere Information Server y Herramientas de Fuente Abierta." Febrero de 2013. **ibm.com**/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=XB&htmlfid=IME14019USEN

[4] "Large-Scale ETL With Hadoop," Strata+Hadoop World 2012 presentación de Eric Sammer, Principal Solution Architect, Cloudera. www.cloudera.com/content/cloudera/en/resources/library/hadoopworld/strata-hadoop-world-2012-large-scale-etl-with-hadoop.html

Please Recycle

IMW14791-USEN-00