Informatica
PLATINUM PARTNER

PowerData

WhitePaper

Mejores prácticas de Data Warehousing con Snowflake Data Cloud

Guía para optimizar, escalar y entregar resultados más rápidos de los proyectos de Snowflake con la nube de gestión inteligente de datos de Informatica

La importancia del almacenamiento de datos

El almacenamiento de datos es una tecnología que agrega datos estructurados de una o más fuentes para que puedan compararse y analizarse para una mayor inteligencia empresarial.

Pero con los actuales volúmenes de datos en rápido crecimiento, cargas de trabajo de procesamiento y casos de uso de análisis de datos, los sistemas tradicionales de almacenamiento de datos ya no pueden seguir el ritmo. Por lo tanto, muchas organizaciones están trasladando sus almacenes de datos a soluciones modernas basadas en la nube, el primer paso en un proceso de modernización de almacenes de datos más grande. El objetivo es ser más eficientes, más ágiles y mejor preparados para las exigencias de la era digital

Estas organizaciones tienen varias opciones. Existe el ecosistema tradicional de grandes proveedores de almacenamiento local y pequeños y medianos proveedores de sistemas de almacenamiento de datos especialmente diseñados, la mayoría de los cuales están transfiriendo sus ofertas a la nube. Existen los principales proveedores de infraestructura en la nube, como Amazon Web Services, Microsoft y Google, que ahora ofrecen sus propias soluciones de data warehouse y data lake. Y por último, pero no menos importante, están los nuevos proveedores de soluciones listos para usar, en particular Snowflake, cuya arquitectura nativa de la nube y conjunto de características modernas lo han convertido en una alternativa atractiva para muchas organizaciones, incluidos muchos clientes de Informatica.

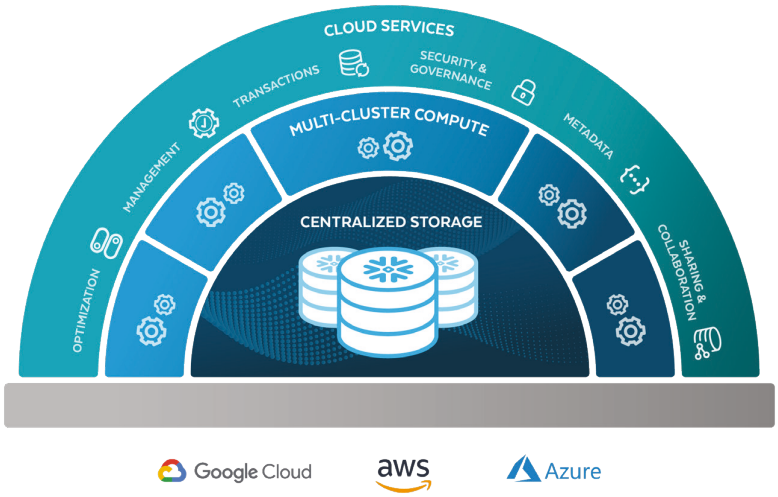
Con Informatica Intelligent Data Management Cloud y Snowflake, puede ingerir, integrar y catalogar datos locales y de múltiples nubes y proporcionar datos limpios, seguros y gobernados para análisis a escala y velocidad en la nube. Puede compartir y monetizar datos relevantes y confiables con proveedores, socios y clientes en tiempo real, todo impulsado por la catalogación y el gobierno de datos en toda la empresa.

Este documento describe las mejores prácticas recomendadas para los clientes de Informatica a medida que incorporan y replican datos en Snowflake, los leen o procesan sus datos en Snowflake Data Cloud utilizando Informatica. Este documento hace referencia principalmente a las funciones de los servicios inteligentes en la nube (IDMC) de Informatica en Informatica Intelligent Data Management Cloud como referencia.

Snowflake también es compatible con el catálogo de datos empresariales de Informatica y otros productos de Informatica. Se pueden encontrar más detalles sobre las funciones de productos específicos de Informatica en las guías de usuario de los productos respectivos.

Conceptos y descripción general de Snowflake Data Cloud

Snowflake ofrece la nube de datos, una red global donde miles de organizaciones movilizan datos con escala, concurrencia y rendimiento casi ilimitados. La nube de datos de Snowflake permite a las organizaciones unificar y conectarse a una sola copia de todos sus datos con facilidad. El resultado es un ecosistema de miles de empresas y organizaciones que se conectan no solo a sus propios datos, sino que también se conectan entre sí compartiendo y consumiendo sin esfuerzo datos y servicios de datos compartidos. La nube de datos hace que las grandes y crecientes cantidades de datos valiosos estén conectados, accesibles y disponibles.



Construido completamente desde cero para la nube, la arquitectura de Snowflake se compone de:

• Almacenamiento centralizado para cantidades prácticamente ilimitadas de datos estructurados y semiestructurados. Una cuenta de Snowflake puede contener una o más bases de datos y hacer referencia a tablas externas para proporcionar una vista única de los datos que residen en Snowflake o en otras fuentes.

• Cómputo de múltiples clústeres para ejecutar múltiples cargas de trabajo sin contención de recursos. Un almacén virtual en Snowflake es un clúster de servidores de bases de datos implementados para ejecutar consultas de usuarios bajo demanda.

• Servicios en la nube para automatizar tareas comunes de administración, seguridad y metadatos.

Tradicionalmente, cada vez que hablamos de un almacén de datos, nos referimos tanto a los datos almacenados en él como a la computación requerida para procesar dichos datos, juntos como uno solo. Snowflake, por otro lado, separa el almacenamiento y la computación requerida para la ejecución, y usa el término "almacén de datos" para referirse específicamente a la computación de múltiples clústeres que se usa para trabajar con los datos.

Esta separación de almacenamiento y cómputo es una distinción conceptual clave y un diferenciador importante para la arquitectura de Snowflake, porque permite una mejor elasticidad y utilización de recursos. Le permite utilizar un pequeño almacén para procesos sencillos independientemente del almacenamiento, mientras reserva tamaños de almacén más grandes y costosos para procesos más complejos que requieren potencia informática adicional. Por ejemplo, para procesar un gran volumen de datos preexistentes iniciales, es posible que primero necesite utilizar un gran almacén; pero a partir de ahí, puede reducir la escala a tamaños más pequeños y rentables para el procesamiento delta en curso, donde cada proceso carga una cantidad de datos relativamente menor.

Esta capacidad de expandir o reducir el poder de cómputo en uso, según los requisitos, brinda verdadera elasticidad y una utilización mucho más eficiente de los recursos del almacén.

Solución conjunta de Informatica y Snowflake

Informatica ofrece la única solución de gestión de datos integral, nativa de la nube y de clase empresarial para data warehouse y data lake, lo que ayuda a los clientes a acelerar su viaje a la nube de manera transparente, sin necesidad de codificación.

Con Informatica Intelligent Data Management Cloud, ofrecemos una cartera completa de productos perfectamente integrados con Snowflake Data Cloud, diseñados para ayudar a las empresas a entregar datos consistentes, confiables y correctamente gobernados en la nube. Los productos y servicios de Informatica están certificados como "preparados para Snowflake", un estándar que valida que las integraciones de los socios se adhieren a las mejores prácticas publicadas por Snowflake.

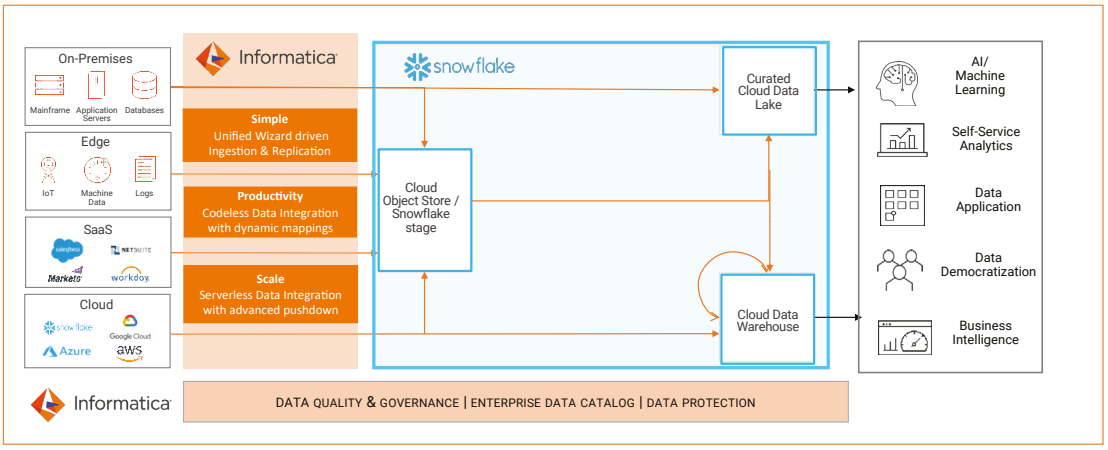
Acelere el paso a Snowflake

Informatica Intelligent Cloud Services Accelerator para Snowflake reduce drásticamente el tiempo necesario para ingerir o integrar datos en Snowflake, lo que permite a los clientes conjuntos de Informatica y Snowflake conectarse rápidamente a fuentes de datos en la nube. El acelerador está disponible para los clientes de Snowflake a través de Snowflake Partner Connect (vea este video para obtener más información sobre el acelerador).

Con la plataforma de Snowflake y las soluciones de Informatica, las empresas pueden centralizar todos los datos en una sola ubicación y admitir muchos casos de uso diferentes, incluido el lago de datos, la ingeniería de datos, el almacén de datos, la ciencia de datos, las aplicaciones de datos y las cargas de trabajo de uso compartido de datos, con una ingestión guiada por asistente. experiencia y soporte para capacidades avanzadas sin servidor.

Sin embargo, para garantizar una iniciativa exitosa de modernización del lago de datos y el almacenamiento de datos en la nube en Snowflake, debe catalogar, ingerir, integrar, limpiar y controlar sus datos (sin importar el tipo de fuente de datos). Y para promover con éxito la adopción de su nueva plataforma, también necesita una base sólida para la gestión continua de datos, que comprende estos tres pilares:

- Gestión de metadatos
- Integración de datos nativos de la nube
- Gobernanza de datos de extremo a extremo



Catálogo de datos empresariales

El descubrimiento inteligente de datos con Informatica Enterprise Data Catalog lo ayuda a administrar los cambios y comprender mejor los datos que se necesita para moverlos al data warehouse o data lake en la nube en Snowflake.

Enterprise Data Catalog permite a los usuarios de negocio y de TI aprovechar todo el potencial de sus activos de datos empresariales al proporcionar una vista unificada de metadatos que incluye metadatos técnicos, contexto comercial, anotaciones de usuarios, relaciones, calidad de datos y uso.

Puede descubrir, clasificar y catalogar activos de datos en toda la empresa y ponerlos a disposición como fuentes para su almacén de datos.

Ingestión masiva de nubes

Las organizaciones suelen incorporar datos en un lago de datos en la nube antes de trasladarlos a **almacenes de datos en la nube** donde pueden estar disponibles para BI y análisis. Esto requiere la ingesta eficiente y precisa de grandes volúmenes de datos de una variedad de fuentes por lotes o en tiempo real. Por ejemplo**:**

- Archivos como archivos estáticos locales, detectores de archivos o archivos en servidores FTP
- Captura de datos masivos y cambios incrementales (CDC) de bases de datos y almacenes de datos
- Fuentes de transmisión como datos de IoT, registros, flujo de clics o redes sociales
- Sistemas de mensajería como Apache Kafka, Amazon Kinesis o JMS

Esto presenta un desafío significativo, un desafío que el servicio Informatica Cloud Mass Ingestion facilita mucho más.

La arquitectura de lago de datos típica implica la ingesta de datos de las fuentes anteriores en lagos de datos en la nube o sistemas de mensajería (como Apache Kafka). Una vez que los datos están disponibles en el lago, se pueden aplicar varias técnicas de integración de datos, como enriquecimiento, transformaciones y agregación, para prepararlos para las iniciativas de análisis o IA.

Cloud Mass Ingestion ayuda a los clientes a ingerir datos de transmisión masiva y en tiempo real desde una variedad de tipos de fuentes de una manera altamente escalable y eficiente, a través de una experiencia unificada y simple impulsada por un asistente. También ofrece monitoreo en tiempo real para que los clientes administren trabajos de ingestión de ejecución prolongada, así como la capacidad de administrar su ciclo de vida.

Cloud Mass Ingestion aborda tres patrones principales de casos de uso.

1. Ingestión del lago de datos en la nube: ayuda a los clientes a ingerir datos de varias fuentes en los lagos de datos en la nube para el procesamiento y análisis posteriores.
2. Migración y sincronización del almacén de datos en la nube: ayuda a los clientes a ingerir y sincronizar sus bases de datos y almacenes de datos en un almacén de datos en la nube proporcionando una carga masiva inicial y una carga de CDC incremental, aplicando cambios mientras aborda las desviaciones del esquema (desviación del esquema) en la fuente de forma automática. manera.
3. Ingestión de Messaging Hub: ayuda a los clientes a ingerir datos de transmisión, IoT y CDC incrementales en Kafka, que se pueden usar para análisis en tiempo real y distribución descendente.

Hay tres componentes de **Cloud Mass Ingestion:**

Ingestión masiva de bases de datos

Este componente lo ayuda a ingerir datos CDC masivos e incrementales de bases de datos y almacenes de datos locales.

Esto permite a los usuarios ingerir datos de miles de tablas (junto con los esquemas de origen) en Snowflake. Puede admitir tanto la carga inicial como las cargas en curso mediante CDC. Este servicio también realiza un seguimiento de los cambios de esquema y, para evitar la desviación del esquema, los aplica automáticamente en las tablas de Snowflake sin necesidad de intervención manual.

Ingestión masiva de archivos

Este componente lo ayuda a ingerir datos de fuentes de archivos, ya sea en las instalaciones o en la nube, lo que permite que los datos basados en archivos se carguen tal cual en una capa de aterrizaje en Snowflake. Invoca internamente el comando de copia en archivos ubicados en el almacenamiento local o en la nube, lo que suele ser mucho más eficiente que usar tareas de integración de datos porque no requiere leer el contenido de los archivos. Si los archivos están en un sistema de almacenamiento en la nube como Amazon S3 o Azure Blob, se cargarán directamente en Snowflake mediante el comando de copia; si son locales, primero se cargarán en el almacenamiento en la nube y luego se copiarán en Snowflake.

El componente de ingesta masiva de archivos también admite la transferencia de archivos desde servidores FTP/SFTP/FTPS remotos. Incluye la opción de activar la tarea de transferencia a través de un detector de archivos cuando ocurre un evento de archivo específico. Puede definir qué archivos recoger de acuerdo con un conjunto de propiedades, como el patrón de nombre de archivo, el tamaño del archivo y los datos. También puede aplicar acciones en estos archivos, como compresión o grabado, antes de escribirlos en Snowflake.

Ingestión masiva de datos de transmisión

Este componente ayuda a ingerir datos de transmisión, mensajería y fuentes de IoT. Puede leer los datos de fuentes como registros, flujo de clics, redes sociales, Apache Kafka, Amazon Kinesis Streams, Azure EventHub, fuentes de JMS y fuentes de IoT como MQTT u OPC UA para ingesta en lagos de datos en la nube. Una vez que los datos están en un lago de datos en la nube, se pueden transferir a Snowflake mediante soluciones de integración de datos.

Integración de datos en la nube

Con Informatica Cloud Data Integration, puede conectarse a cientos de fuentes, incluidas aplicaciones en la nube, locales y SaaS, con automatización inteligente.

La integración de datos en la nube le permite crear tareas complejas de integración de datos utilizando un diseñador de mapas con transformaciones de integración de datos avanzadas listas para usar que se pueden usar para realizar cualquier proceso o transferencia de datos que pueda ser necesario antes de cargarlo en Snowflake y hacerlo. disponible para las necesidades de informes y análisis de los usuarios. Estos flujos de integración se pueden ejecutar a escala en un entorno sin servidor o mediante el procesamiento de Spark.

Elástico de integración de datos en la nube

El servicio Elastic de integración de datos en la nube de Informatica permite que su organización de TI procese tareas de integración de datos sin administrar servidores ni requerir experiencia adicional en big data. en función de la demanda y el consumo.

Utilice el servicio elástico de integración de datos en la nube donde se espera que se procesen grandes volúmenes en un lago de datos (si los datos ya están en Snowflake, le recomendamos que utilice la optimización pushdown avanzada de Informatica).

Gobernanza y calidad de datos

Informatica ofrece múltiples productos para respaldar la necesidad de **gobierno de datos y democratización** de datos. La aplicación de un programa de gobierno de datos empresariales a su Snowflake Data Cloud asegurará que todos sus datos sean confiables y lo suficientemente útiles para impulsar sus iniciativas comerciales. Las siguientes capacidades principales aseguran que maximice el valor de Snowflake Data Cloud.

- Gobierno de datos: identificar el contenido empresarial de los datos; definir procesos, políticas y propiedad; y permitir a los usuarios no técnicos la capacidad de comprender y acceder a los datos.
- Calidad de los datos: mida las métricas y los cuadros de mando de la calidad de los datos.
- Catálogo de datos: descubra qué atributos se están definiendo (por ejemplo, esquemas, tablas, columnas).
- Privacidad de datos: hacer cumplir políticas, informar sobre riesgos, buscar registros de sujetos, realizar análisis de incumplimiento, etc.
- Mercado de datos: proporcione un consumo de datos amplio, consistente, controlado y fácil para todos los empleados

Implementaciones de casos de uso de Data Lake y Data Warehouse

El uso más común de Snowflake para la mayoría de las organizaciones es como almacén de datos empresariales (EDW), el almacén de datos en el que se ejecutan los análisis. En la mayoría de los casos, los datos en un CDW (almacén de datos en la nube) se almacenan mediante un modelo dimensional, aunque también se pueden almacenar mediante una bóveda de datos, un modelo de almacenamiento de datos operativos (ODS) o cualquier otra forma de almacenamiento de datos. A los efectos de este documento, utilizaremos un almacén de datos dimensional como referencia; sin embargo, muchas de las prácticas que se describen a continuación también deberían aplicarse a una bóveda de datos o un ODS.

Los datos en un CDW deben estar listos para el consumo de los productos de informes y análisis que se encuentran encima. Por lo tanto, estos datos deben obtenerse y unirse desde múltiples fuentes, transformarse, limpiarse y estandarizarse antes de que puedan cargarse en el CDW para uso empresarial (Nota: esta es una diferencia clave entre los datos que residen en un lago de datos versus en un CDW).

Para que el proceso de preparación sea más eficiente, le recomendamos que adopte un enfoque de dos etapas para cargar datos en un CDW y que elija un flujo ELT (extraer, cargar, transformar) sobre ETL (transformar y luego cargar):

1. Cargue datos desde su origen en una capa de preparación o aterrizaje en Snowflake. Haga esto de forma masiva, realizando muy pocas (o preferiblemente cero) transformaciones, usando las capacidades de Cloud Mass Ingestion descritas anteriormente.

2. Transforme desde el área de preparación, procese más los datos para transformarlos, limpiarlos y cargarlos en las capas posteriores.

Dependiendo de los requisitos de análisis, puede haber más de dos etapas de preparación de datos. El hecho de tener un enfoque de dos etapas o de tres etapas depende casi por completo del nivel de transformaciones necesarias para preparar los datos para las aplicaciones que los consumen. En función de estos requisitos de consumo (ya sea de análisis o de otras aplicaciones posteriores), es posible que necesite que los datos se almacenen en una estructura que sea más fácil de consumir, o puede necesitar que se limpien, pero también que permanezcan en la misma forma que en la fuente. En tales casos, típicamente habrá un ODS donde los datos se cargan antes de cargarlos en el CDW.

Mejores prácticas para almacenamiento de datos en Snowflake

Estas son las diversas etapas del flujo de datos en Snowflake Data Cloud. Tenga en cuenta que, según su arquitectura, es posible que tenga una o más de las etapas que se describen a continuación. No es obligatorio tenerlos todos; estas son simplemente mejores prácticas y patrones comúnmente observados.

Área de aterrizaje/escenario

Aquí es donde se cargan los datos desde la fuente. Como se describió en secciones anteriores, las estructuras de datos imitan las estructuras de origen y, por lo general, no se realizan transformaciones durante la carga aquí.

Almacén de datos operativos

Esta es la siguiente etapa de los datos. Incluso aquí, los datos se almacenan en estructuras similares a la fuente, pero luego se transforman y limpian para que estén listos para el análisis. Por lo general, esto se hace para respaldar los informes existentes que ya usan los datos en su forma anterior (si no hay consumidores directos de datos en su forma original, entonces este paso no es necesario).

Almacén de datos empresarial

Como se indicó anteriormente, el destino principal de los datos en Snowflake es un EDW, para el cual la mayoría de los clientes eligen un modelo dimensional.

Un almacén de datos que utiliza un modelo dimensional normalmente representa los datos como "hechos" o "dimensiones". En resumen: los hechos representan registros de transacciones que ocurren como parte de su negocio y las dimensiones representan los datos maestros que utiliza para analizar esas transacciones. Puede visualizar un modelo dimensional con una tabla de hechos en el centro y varias tablas de dimensiones a su alrededor, razón por la cual la representación más común se denomina esquema en estrella.

Hay varias variaciones en los detalles de cómo diferentes clientes implementan un modelo dimensional. Algunos ejemplos:

- Basado en clave sustituta
- Basado en clave natural
- Estructuras de Snowflake (tenga en cuenta que "Snowflake" aquí es un término de modelado de datos, no se trata de Snowflake la aplicación)

La forma en que se implementa suele ser una elección arquitectónica.

Otra variación de dicho modelo es una bóveda de datos. Esta es una técnica de modelado donde hay concentradores, satélites y enlaces. Estos usan conceptos similares a un modelo dimensional mientras cargan datos, pero también hay algunas variaciones. Puede encontrar más información sobre el **modelado de bóveda de datos aquí.**

Informatica es compatible con todas estas opciones (consulte las ilustraciones a continuación). Con Informatica, puede elegir el enfoque que mejor se adapte a sus requisitos y configurar la ingesta de datos en consecuencia.

Caso de uso de almacenamiento de datos de Snowflake #1:
Ingesta masiva y replicación en Snowflake

Ingestión masiva y replicación en Snowflake Las organizaciones suelen incorporar datos en un lago de datos en la nube antes de mover los datos a los almacenes de datos en la nube. Es más eficiente transferir los datos de origen a Snowflake tal como están (usando el servicio Cloud Mass Ingestion) y luego procesarlos dentro de Snowflake usando el tamaño de almacenamiento de datos que sea más apropiado. Las siguientes secciones describen estos pasos en detalle.

Estos son algunos escenarios comunes en los que los datos externos se cargan en Snowflake con poca o ninguna transformación de datos:

Migración de un almacén de datos existente a Snowflake

En este escenario, los datos se migran del almacén de datos heredado de la organización (generalmente en las instalaciones) a Snowflake, probablemente como parte de una iniciativa más amplia de modernización del almacén de datos. La migración del almacén de datos suele implicar dos pasos: carga inicial y carga incremental continua o intermedia.

Carga inicial

En la fase de carga inicial, las tablas del almacén de datos normalmente se copian en el nuevo almacén de datos en Snowflake con el mismo formato y estructura que el origen, sin cambios de esquema, lo que puede no requerir la copia en un área de preparación. No obstante, el destino principal de este paso es la capa de almacenamiento de datos en Snowflake.

Este paso es importante para establecer una línea de base después de la migración donde puede confirmar que los datos en Snowflake coinciden con los datos de origen que estaban disponibles al comienzo de la migración. Las estructuras de la tabla en este escenario son similares, si no idénticas, a las del almacén de datos original. Y para la capa de análisis, es más fácil cambiar a Snowflake si las definiciones comerciales subyacentes aún pueden funcionar tal cual.

Carga de datos semiestructurados

Si está utilizando Snowflake como un lago de datos, es probable que esté cargando muchos más datos en la capa de aterrizaje de los que necesita su EDW. Algunos de estos datos pueden existir como archivos de varios tipos semiestructurados, como Parquet, JSON, Avro u ORC. Puede cargarlos mediante la tarea Ingestión masiva de archivos. Snowflake recomienda cargar dichos archivos en una tabla con una columna de tipo VARIANT.

File Mass Ingestion también se puede utilizar para el procesamiento posterior de estos archivos en función de sus requisitos: este artículo describe algunos ejemplos de cómo puede utilizar File Mass Ingestion, combinado con asignaciones, para cargar dichos datos semiestructurados y extraer la información que necesita. están interesadas en.

Cargas en curso

Una vez que se realiza la carga inicial en Snowflake, suele haber un período de varias semanas en el que ambos almacenes de datos se mantienen actualizados. Esta carga dual ayuda a garantizar que toda la funcionalidad anterior siga funcionando correctamente después de la migración y que los datos disponibles en Snowflake sean los mismos que en el almacén de datos anterior. Si algo sale mal con la migración, es más fácil cambiar temporalmente a los usuarios al antiguo almacén de datos mientras se soluciona el problema.

Para mantener este estado, es necesario seguir cargando datos nuevos y modificados desde el antiguo almacén de datos a Snowflake. Se pueden usar varios tipos de herramientas para realizar los escenarios de migración: herramientas de uso masivo que pueden copiar esquemas y datos para una gran cantidad de tablas en una sola configuración, o herramientas de integración de datos que se pueden configurar para definir un flujo y luego ejecutar para cada objeto de origen que se va a migrar.

Como se trata de un flujo de una sola vez o de corto plazo, el manejo de los cambios de esquema no es un requisito necesario (a diferencia de los casos de uso de "derivación del esquema" en los que se necesita dicha capacidad de manejo de cambios de esquema). Cuando ocurren tales cambios, los clientes pueden preferir revisar cuidadosamente y aplicar cambios selectivos a su almacén de datos de Snowflake, en lugar de dejarlos en manos de una herramienta.

Por otro lado, debido a que estos escenarios generalmente involucrarán una cantidad excepcionalmente grande de datos, tener buenas funciones de recuperación de fallas será un requisito clave para que estas herramientas tengan, de modo que una tarea fallida pueda ejecutarse desde el punto de falla en adelante.

Replicación continua de datos de bases de datos y almacenes de datos en Snowflake

En este caso, un escenario diferente al de una migración de almacenamiento de datos, los datos se replican en Snowflake desde su fuente de forma continua. La fuente suele ser transaccional: por ejemplo, ERP, CRM o datos de interacción social. Estos datos se sincronizan con Snowflake a una frecuencia predefinida o de forma continua en tiempo real. También incluye la detección automática y la aplicación de los cambios realizados en el esquema de origen al nuevo esquema.

A diferencia de los escenarios de migración anteriores, que tienen una duración finita, este escenario de replicación es un escenario continuo sin un final planificado.

Cloud Mass Ingestion realiza una sincronización eficiente a escala mediante cargas iniciales e incrementales. También admite el manejo automático de cambios en el esquema de origen, incluida su replicación en las estructuras de destino.

Caso de uso de almacenamiento de datos de Snowflake #2:
Procesamiento de datos cargados

Una vez que los datos sin procesar estén disponibles en la capa de aterrizaje de Snowflake, querrá transformarlos adecuadamente para su consumo. Desde aquí, los datos podrían tomar cualquiera de las siguientes rutas:

- Puede transformar los datos para que se ajusten a su modelo de datos de destino (ya sea un modelo dimensional EDW, una bóveda de datos o un almacén de datos operativos).
- Puede descargarlo y sincronizarlo con otra aplicación fuera de Snowflake.
- Puede aplicar algoritmos de ciencia de datos a escala empresarial y transformaciones de calidad de datos para usar en proyectos de aprendizaje automático (ML).

El cómputo o motor real que utilice para hacer esto dependerá del volumen de datos, la naturaleza del procesamiento y el costo del procesamiento tanto dentro como fuera de Snowflake, así como cualquier otro factor que pueda ser importante según los requisitos de su proyecto. Recomendamos utilizar Cloud Data Integration Mapping Designer de Informatica como herramienta principal para diseñar estos flujos de datos o "mapeos". Luego, en tiempo de ejecución, puede elegir entre cualquiera de los motores según su elección.

1. Utilice el motor nativo de Informatica para aplicar transformaciones. Esta es la opción predeterminada que lee los datos de origen y los procesa fila por fila, aplicando transformaciones agregadas o de nivel de fila sobre ellos y escribiendo en el destino. Este es el escenario ETL más común.

2. Empuje la lógica de mapeo a Snowflake. Al obtener los datos en la capa de aterrizaje, ha realizado la E y la L (extracción y carga) del ELT. Ahora bien, este "empuje hacia abajo" implementa la T (transformación). Puede diseñar el flujo de datos gráficamente mediante asignaciones de Informatica y configurar la optimización pushdown avanzada (APDO) como opción de tiempo de ejecución. Cuando lo ejecuta, Informatica traduce la lógica de asignación a los comandos de Snowflake. Si toda la lógica de mapeo se puede traducir, puede ejecutarla completamente dentro de Snowflake, utilizando el almacén que ha configurado con Informatica.

3. Si bien la mayoría de los clientes parecen preferir la opción APDO anterior, es posible que no siempre sea posible traducir ciertos tipos de transformaciones a Snowflake. Es posible que no haya equivalentes de SQL para ciertos tipos de procesamiento. En tales casos, además de utilizar el motor nativo de Informatica, también puede configurar la opción de tiempo de ejecución de Informatica como Spark. Informatica CDI Elastic puede usar un motor Spark para transformar datos de gran volumen a escala empresarial.

Caso de uso de almacenamiento de datos de Snowflake #3:
Lectura de datos de Snowflake

Toda la conectividad de Informatica es bidireccional. Puede usarlo para cargar datos en Snowflake, procesarlos dentro de Snowflake y leerlos o descargarlos para compartirlos con aplicaciones externas. Puede hacer esto usando la misma herramienta de Cloud Mapping Designer, en cuyo caso Snowflake se convierte en una "fuente" en el mapeo**.**

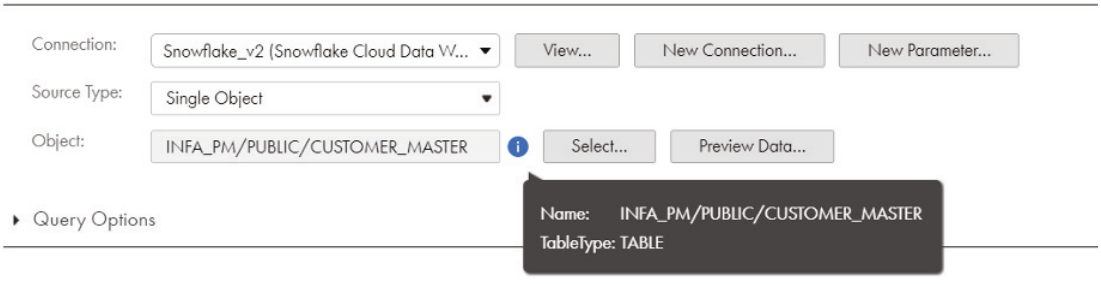
Dicha definición de origen puede hacer referencia a una tabla Snowflake, una tabla externa, una vista o una vista materializada. También puede definir consultas SQL sobre Snowflake y usar el resultado como fuente de la asignación. Esto le permite realizar cualquier procesamiento dentro de Snowflake antes de obtener los datos para cualquier procesamiento o uso compartido externo.

Uso de consultas

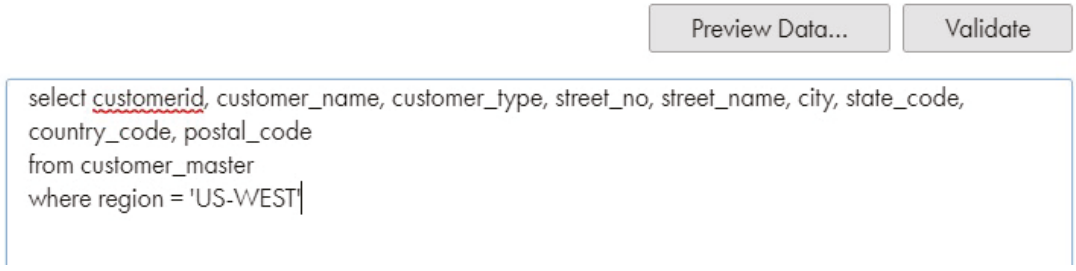
Snowflake es una base de datos en columnas y varía significativamente en términos de estructura de datos en comparación con las bases de datos relacionales clásicas como Oracle. Como resultado, el rendimiento de una operación CRUD puede variar significativamente según la cantidad de columnas involucradas en ella.

Cuando configura una fuente de Snowflake para el mapeo, el tipo de fuente predeterminado que se muestra es "objeto", que se refiere a una tabla o una vista en Snowflake. Le permite navegar y seleccionar una tabla o vista desde la fuente. Cuando utiliza un objeto como fuente, mientras que todos los campos de la fuente están vinculados a la siguiente transformación de forma predeterminada, Informatica lee los datos solo para los campos que realmente se usan en la lógica de asignación o están vinculados al destino. Cuando escribe una consulta que anula este comportamiento, es importante seleccionar solo los campos necesarios, especialmente cuando se utilizan bases de datos orientadas a columnas.

Por ejemplo, digamos que está leyendo datos de una tabla con 300 columnas, pero solo necesita leer y transformar algunas de ellas, por ejemplo, 20. En un escenario de base de datos en columnas, es significativamente eficiente restringir la lectura solo a las columnas específicas. necesitas.



Puede hacerlo creando una vista en Snowflake (si espera que se reutilice mucho) o usando el tipo de objeto "consulta" para una función de configuración de origen en una asignación, que le permite configurar una consulta SQL específica. para ser utilizado como fuente. En este ejemplo, puede seleccionar solo las 20 columnas que necesita de dicha consulta, y solo esas se leerán desde la fuente. Una selección de columnas tan limitada siempre es eficiente, pero especialmente con las bases de datos en columnas.

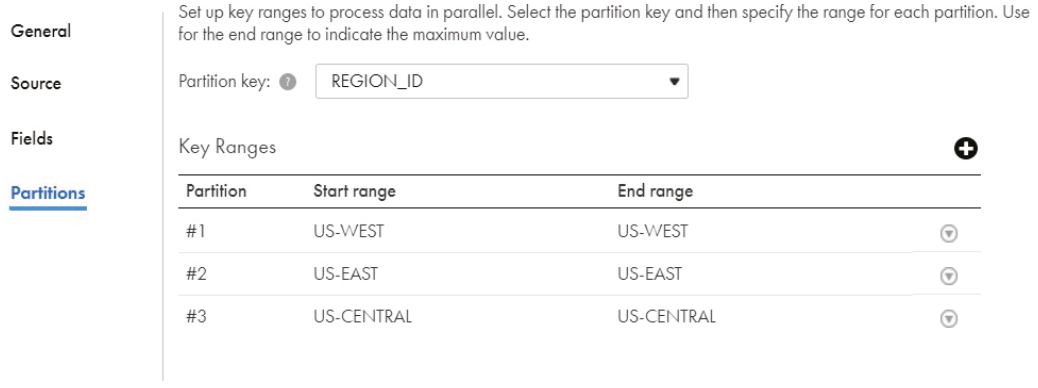


Un comportamiento similar se aplica también en el lado del objetivo. Si está actualizando solo unas pocas columnas de una tabla, o si está insertando datos solo para unas pocas columnas, vincule solo esas columnas en una asignación. Informatica emite los comandos en función de las columnas vinculadas al destino.

Uso de particiones

También puede usar particiones mientras lee los datos de Snowflake para leer y procesar conjuntos de filas en paralelo. En este escenario, puede identificar una columna para usarla como clave de partición y luego especificar dos o más particiones con un rango de valores para dicha clave.

El siguiente ejemplo muestra las particiones de origen utilizando la columna "region_id" con tres rangos diferentes. Cuando se ejecuta una asignación de este tipo, Informatica lee los datos de la tabla Snowflake mediante tres subprocesos paralelos, cada uno para un conjunto lógico de filas. Tenga en cuenta que esto es independiente de la lectura paralela de datos que realiza Snowflake como parte de su implementación de la operación de lectura.



Cómo obtener el mejor rendimiento para su mapeo o tarea

Existen numerosas formas de obtener el mejor rendimiento de su mapeo o una tarea:

1. Uso de la optimización pushdown avanzada
2. Optimización de la ubicación de Informatica Secure Agent
3. Ajuste de las propiedades del agente seguro
4. Ajuste de los atributos de origen/destino de las asignaciones de Informatica
5. Uso de tareas de ingesta masiva
6. Realización de cambios de configuración específicos de Snowflake

Uso de la optimización pushdown avanzada

Si sus datos ya están cargados en Snowflake y desea procesarlos más para cargarlos en otras tablas dentro de Snowflake, primero puede evaluar si la tarea en cuestión puede ejecutarse en modo APDO.

La guía del conector de Informatica Snowflake describe las transformaciones y funciones compatibles con APDO. Cuando utiliza APDO, Informatica convierte la lógica de mapeo en comandos de Snowflake y los ejecuta dentro de Snowflake. Dependiendo de la naturaleza de los datos y el procesamiento, esta podría ser la opción más eficiente.

Ubicación de Informatica Engine/Secure Agent

Si utiliza Informatica Secure Agent, asegúrese de que esté lo más cerca posible de la región Snowflake. Por ejemplo, si su agente está en Amazon EC2, asegúrese de que esté en la misma región que la cuenta de Snowflake. Si utiliza APDO, también puede utilizar el tiempo de ejecución alojado en Informatica.

Ajuste de las propiedades del agente seguro

Por lo general, puede mejorar el rendimiento de las tareas modificando el tamaño del almacenamiento dinámico de Java y el límite de memoria del cliente. **Consulte la Guía de ajuste de rendimiento de Snowflake para obtener más detalles**.

Ajuste de atributos de origen y destino

Hay varios atributos de ajuste de rendimiento para los orígenes y destinos de Snowflake que se pueden configurar en un mapeo, en función de sus datos y preferencias de carga/descarga. Las siguientes secciones describen cada atributo que puede modificar para mejorar el rendimiento, todos los cuales son aplicables cuando se usa el motor de Informatica para procesar los datos.

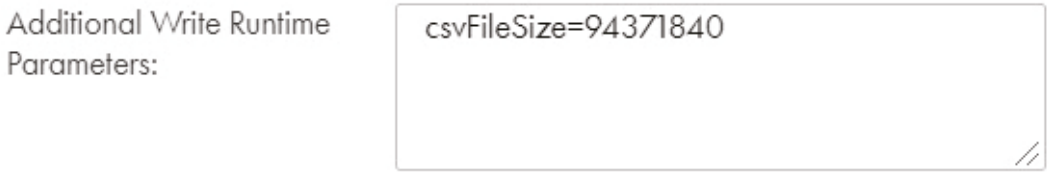
Cuando ejecuta una asignación que escribe en Snowflake, primero crea archivos de preparación locales internamente, en la máquina donde ocurre el procesamiento. Si está utilizando IICS, es la máquina donde se encuentra el entorno de tiempo de ejecución. Luego, estos archivos se cargan en paralelo a la etapa de usuario, configurada con la cuenta de Snowflake. A partir de ahí, los comandos COPY se invocan en paralelo para cargar datos de estos archivos en la tabla de Snowflake.

Al leer datos de Snowflake, los pasos exactos dependen de cómo esté configurada la fuente de Snowflake. En función de si el origen es un objeto o una consulta, primero se invoca un comando COPY INTO correspondiente para descargar datos en el área de preparación. Luego se lee para su posterior procesamiento.

Tamaño del archivo CSV

Snowflake recomienda dividir los datos grandes en archivos con un tamaño de entre 10 MB y 100 MB comprimidos al cargar datos en la plataforma Snowflake. Cuando utiliza una asignación de Informatica con un destino de Snowflake, esto se puede ajustar mediante la propiedad "csvFileSize", que se encuentra entre los parámetros de tiempo de ejecución de escritura adicionales en las propiedades de destino avanzadas.

El valor predeterminado para esta propiedad es 50 MB. La siguiente captura de pantalla muestra un valor de 90 MB (90*1024*1024).



Número de archivos de ensayo locales

A medida que los archivos se cargan en el área de preparación de Snowflake, este atributo funciona como el umbral después del cual los datos se escriben en la tabla de destino. Un valor pequeño para este atributo da como resultado una mayor cantidad de comandos COPY. El valor predeterminado es 64.

Tenga en cuenta que Snowflake ejecuta la operación COPY con todos esos archivos cargados en subprocesos paralelos e independientes. Si hay errores de datos en cualquiera de estos subprocesos, ese subproceso en particular puede cancelarse; pero otros subprocesos pueden continuar cargando los datos si no se encuentran errores en esos subprocesos, dependiendo de su configuración de tiempo de ejecución. Debe tener en cuenta este comportamiento al configurar este valor.

Por ejemplo, considere un escenario donde el tamaño del archivo CSV es de 50 MB. Si el número se establece en 5, se invocará una COPIA después de que se carguen cinco archivos de 50 MB cada uno. La copia cargará los cinco archivos en paralelo. Si el número se establece en 100, se invocará un comando COPIAR después de que se carguen 100 de esos archivos. La copia luego

Recursos adicionalescargará los 100 en paralelo.

Puede determinar el número óptimo según el tamaño de sus datos y el tamaño de su almacén (y recuerde: "almacén" en el caso de Snowflake significa "motor de cómputo"). Recomendamos comenzar con el valor predeterminado y ajustarlo en cualquier dirección para ver si mejora su rendimiento.

Tamaño de registro de lote

No recomendamos cambiar este valor a menos que desee específicamente archivos más pequeños, en función de una cierta cantidad de registros. Esta es una forma alternativa de determinar el tamaño de los archivos locales individuales. La diferencia entre esto y el atributo Tamaño de archivo CSV es que este valor se basa en un registro lógico, no en el tamaño del archivo. Este valor tiene prioridad solo si el tamaño resultante es más pequeño que el tamaño del archivo CSV; como resultado, ajustar el tamaño del registro por lotes solo es efectivo si lo configura cuidadosamente en función del tamaño de su registro y la cantidad de registros que desea escribir en cada archivo.

Uso de tareas de ingesta masiva

Si no está realizando transformaciones sobre los datos, puede evaluar si las tareas de Cloud Mass Ingestion admiten las combinaciones de origen y destino. Consulte las secciones anteriores para obtener más detalles.

Afinación específica de copo de nieve

Además del ajuste relacionado con Informatica, hay varios cambios de configuración en Snowflake que pueden ayudar a mejorar el rendimiento, como usar un almacén más grande, usar almacenes separados para cargar datos en lugar de consultarlos, etc. Para obtener más detalles, consulte la **documentación de ajuste del rendimiento de Snowflake.**

Seguridad

De forma predeterminada, Informatica admite la autenticación basada en las credenciales de inicio de sesión asociadas con una cuenta de Snowflake. Los clientes que utilizan Okta para el inicio de sesión único (SSO) en su organización pueden configurarlo con Informatica. Consulte la guía del conector para obtener más detalles.

Los clientes que hayan configurado Snowflake con una configuración de red específica del proveedor de la nube, como un enlace privado de AWS, también pueden trabajar con el servicio de atención al cliente global de Informatica para ayudarlos a abordar cualquier problema de configuración de este tipo.

Recursos adicionales

Juntos, Informatica y Snowflake ofrecen una arquitectura de datos unificada que ofrece las mejores capacidades de almacén de datos en la nube, gestión de datos en la nube y lago de datos, y se ejecuta en cualquier nube.

Para obtener orientación adicional y mejores prácticas sobre cómo aprovechar al máximo Snowflake con Informatica, aquí hay algunos recursos adicionales:

1. Visite el **portal de documentación de Informatica** para explorar una amplia biblioteca de documentación para las versiones de productos actuales y recientes.
2. Utilice la **base de conocimientos de Informatica** para encontrar recursos de productos, como artículos de procedimientos, prácticas recomendadas, tutoriales en vídeo y respuestas a las preguntas más frecuentes.
3. Obtenga más información sobre la **asociación de Informatica y Snowflake.**
4. Más información sobre **Informatica Intelligent Data Management Cloud**

Esta guía es una traducción al español de la versión original Best Practices for Data Warehousing With Snowflake publicado por Informatica

PowerData | Informatica PLATINUM PARTNER

PowerData, es una compañía multinacional de origen español con destacada presencia regional, como especialistas en gestión de datos en la nube, está preparada para ayudar a las organizaciones a acelerar su camino hacia la transformación digital y brindar la previsión necesaria para que sean más ágiles y aprovechen nuevas oportunidades de crecimiento.

PowerData e Informatica se convirtieron en socios desde el año 2000, desde entonces se ha estrechado la relación con más de 100 organizaciones de diversos sectores, categorías y nichos, quienes han depositado su confianza en las soluciones de Informatica, y en el expertise y Know-how de PowerData.

Te invitamos a explorar los proyectos donde aportamos valor con la gestión de datos.
powerdata.es