Implementación de Ingesta Batch con Apache NiFi

Apache NiFi es una herramienta de integración y automatización de flujos de datos que facilita la ingesta, transformación y distribución de datos entre diferentes sistemas de manera eficiente. Es una plataforma **open-source** (código abierto) diseñada para **automizar el flujo de datos** entre sistemas y **simplificar** tareas complejas como la **ingesta batch**.

por Kibernum Capacitación S.A.

¿QUÉ ES APACHE NIFI?

Apache NiFi es una plataforma robusta para la automatización de flujos de datos entre sistemas. Permite capturar, transformar, enrutar, y almacenar datos provenientes de diversas fuentes como bases de datos, archivos, API, y sistemas de mensajería. A través de una interfaz gráfica fácil de usar, NiFi facilita el diseño, monitoreo y gestión de flujos de datos sin necesidad de escribir código.

¿PARA QUÉ SIRVE APACHE NIFI EN LA INGESTA BATCH?

Apache NiFi es muy útil en la ingesta de datos batch debido a las siguientes funcionalidades:

- Automatización de Flujos de Datos: NiFi puede orquestar la ingesta batch de datos desde fuentes a un sistema de destino en intervalos programados, asegurando que los datos sean procesados de manera eficiente y en el momento adecuado.
- Gestión de Grandes Volúmenes de Datos: Es capaz de procesar grandes cantidades de datos, lo que es ideal para escenarios de Big Data. NiFi gestiona eficientemente tanto datos estructurados como no estructurados.
- Transformación de Datos: NiFi proporciona un conjunto de procesadores que permiten realizar transformaciones simples o complejas, como la conversión de formatos (CSV a JSON, por ejemplo), limpieza de datos, y la agregación de información.
- Integración con Diversas Fuentes: NiFi puede conectarse a bases de datos, sistemas de archivos, servicios en la nube y APIs, lo que lo convierte en una solución versátil para la ingesta de datos desde diversas fuentes.

PRINCIPALES CARACTERÍSTICAS DE APACHE NIFI

Interfaz Gráfica Intuitiva

NiFi ofrece una interfaz basada en la web que permite diseñar flujos de datos sin necesidad de escribir código. Esto facilita a los usuarios no técnicos crear y gestionar flujos de datos de manera eficiente.

Procesadores para Conectar Fuentes y Destinos

NiFi incluye más de 300 procesadores que permiten leer desde, escribir a, y transformar datos entre diversas fuentes y destinos, como bases de datos, servicios web, sistemas de mensajería, y más.

Enrutamiento Condicional de Datos

Los datos pueden ser dirigidos dinámicamente a diferentes destinos según condiciones específicas (por ejemplo, datos de diferentes categorías pueden ser enviados a distintos sistemas de almacenamiento).

Escalabilidad y Flexibilidad

NiFi está diseñado para escalar horizontalmente, lo que significa que puede manejar grandes volúmenes de datos distribuidos a través de múltiples nodos en un clúster.

Monitoreo en Tiempo Real

NiFi proporciona herramientas para monitorear el rendimiento de los flujos de datos en tiempo real, lo que facilita la gestión y resolución de problemas rápidamente.

Garantía de Entrega de Datos

NiFi asegura que los datos sean entregados de forma confiable a su destino, utilizando características como al menos una vez, exactamente una vez, o a lo más una vez para el procesamiento de datos.

MÉTODOS DE IMPLEMENTACIÓN DE LA INGESTA BATCH CON APACHE NIFI

Ejemplo de Implementación de Ingesta Batch con Apache NiFi

Caso de Uso: Ingesta de Datos de Ventas de Archivos CSV a una Base de Datos

Objetivo: La empresa necesita procesar los registros de ventas diarios en formato CSV y cargarlos en una base de datos PostgreSQL para su análisis.

Extract (Extraer)

Utilizar el procesador GetFile para leer los archivos CSV almacenados en un directorio específico.El procesador ListFile se puede configurar para leer continuamente nuevos archivos CSV conforme se vayan generando.

Transform (Transformar)

Utilizar el procesador ConvertRecord para convertir los archivos CSV en formato JSON para que sean compatibles con el sistema de bases de datos.El procesador UpdateRecord puede aplicarse para limpiar y normalizar los datos (por ejemplo, asegurando que las fechas estén en el formato correcto).

Load (Cargar)

Utilizar el procesador PutDatabaseRecord para insertar los datos transformados en una base de datos PostgreSQL. Configurar las credenciales y la conexión a la base de datos en NiFi para que los datos se carguen en la tabla correspondiente.

Programación y Monitoreo

Configurar el procesador para que el flujo de datos se ejecute a intervalos regulares (por ejemplo, todos los días a las 2 a.m.) para cargar los datos de ventas al sistema de análisis.

Resultado: El flujo de datos se ejecutará de manera automática todos los días, transformando y cargando los datos de ventas a la base de datos sin intervención manual.

La implementación de ingesta de datos batch con Apache NiFi es una opción potente y flexible que permite gestionar grandes volúmenes de datos de manera eficiente. NiFi facilita la automatización del flujo de datos, la transformación, y la integración de diversas fuentes, lo que permite a las empresas consolidar y analizar datos a gran escala. Su interfaz visual y su capacidad para integrarse con diferentes sistemas lo convierten en una herramienta ideal para gestionar tareas de ingesta batch complejas.

LICENCIAS DE APACHE NIFI

Apache NiFi es una plataforma de integración de datos que es completamente de código abierto bajo la Licencia Apache 2.0. Esta licencia es muy permisiva y permite a los usuarios usar, modificar y distribuir el software tanto en proyectos comerciales como no comerciales.

A continuación, se detallan los aspectos clave relacionados con las licencias de Apache NiFi, sus versiones y capacidades.

Licencia de Apache NiFi

Apache NiFi se distribuye bajo la Licencia Apache 2.0, que es una licencia de código abierto. Esta licencia permite:

- Uso libre: Puedes usar el software en proyectos personales o comerciales sin tener que pagar por licencias.
- Modificación: Puedes modificar el código para ajustarlo a tus necesidades.
- Distribución: Puedes redistribuir el código modificado o no modificado, siempre y cuando se respeten las condiciones de la licencia.
- Patentes: La licencia Apache 2.0 también proporciona una licencia de patente que asegura que el uso del software no infrinja las patentes de los desarrolladores.

Versiones y Capacidades de Apache NiFi

Apache NiFi no tiene versiones comerciales o premium. Toda la funcionalidad está disponible en la versión de código abierto, lo que significa que no hay versiones de pago ni restricciones por nivel de acceso. La única diferencia en las capacidades viene del entorno y la infraestructura en la que se ejecuta NiFi.

Características Clave de Apache NiFi (Versión Open Source):

La versión de código abierto de Apache NiFi incluye:

Interfaz Gráfica de Usuario (GUI)

NiFi ofrece una interfaz gráfica de usuario basada en navegador para diseñar, gestionar y monitorear flujos de datos. Puedes crear flujos de trabajo con arrastrar y soltar para automatizar la ingesta, transformación y carga de datos sin necesidad de escribir código.

Conectividad con Múltiples Fuentes de Datos

NiFi tiene más de 300 procesadores que permiten la integración con diversas fuentes de datos como bases de datos, sistemas de archivos, servicios web y más.

Escalabilidad

Apache NiFi se puede escalar tanto vertical como horizontalmente. Los flujos de datos pueden ser gestionados de manera distribuida en un cluster de NiFi, permitiendo el procesamiento de grandes volúmenes de datos.

Procesamiento Distribuido

Soporta la ejecución de flujos de datos en múltiples nodos en un clúster, lo que mejora el rendimiento y la resiliencia del sistema.

Enrutamiento Condicional de Datos

Los datos pueden ser procesados de forma condicional y dirigidos a diferentes destinos según criterios definidos en el flujo de trabajo.

Seguridad y Control

NiFi incluye características de autenticación y autorización (soporte para LDAP, Kerberos, etc.), cifrado de datos en tránsito y en reposo, y auditoría de accesos.

Integración con otras herramientas de Big Data

NiFi se integra fácilmente con tecnologías como Hadoop, HBase, Kafka, Spark, y Hive, lo que lo hace adecuado para proyectos de Big Data.

Apache NiFi en la Nube y Versiones Adicionales:

- Si bien la versión de código abierto de NiFi es suficiente para muchas empresas, algunas organizaciones optan por ejecutarlo en la nube (por ejemplo, en AWS, Azure, o Google Cloud) para aprovechar la escalabilidad y la infraestructura gestionada que estas plataformas ofrecen.
- Las versiones adicionales de Apache NiFi como NiFi Registry proporcionan características para la gestión de versiones y control de flujos a nivel organizacional, permitiendo mantener una evolución controlada de los flujos de trabajo.

¿Existen Versiones de Pago de Apache NiFi?

No, Apache NiFi es completamente de código abierto bajo la Licencia Apache 2.0, y no existen versiones de pago o comerciales directamente ofrecidas por la Fundación Apache. Sin embargo, algunas empresas pueden ofrecer servicios de soporte comercial, como Hortonworks (ahora parte de Cloudera), que proporciona soporte y servicios adicionales para NiFi en entornos empresariales.

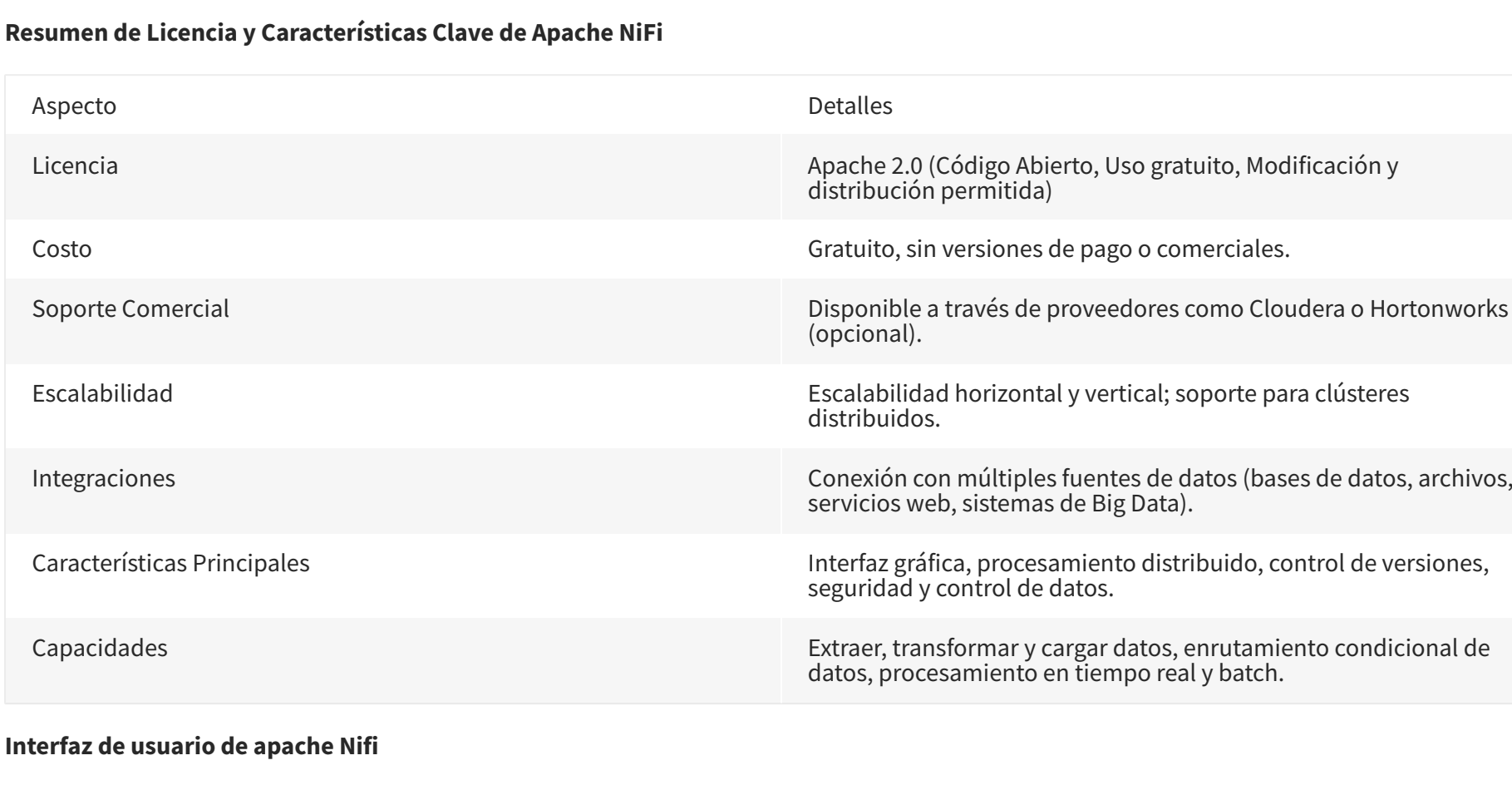
Servicios de Soporte Comercial (opcional):

Cloudera y otras compañías ofrecen soporte empresarial para NiFi, que incluye capacitación, soporte técnico, y herramientas adicionales. Estas ofertas son opcionales y no son necesarias para usar NiFi en su versión de código abierto.

Resumen de Licencia y Características Clave de Apache NiFi

Aspecto	Detalles
Licencia	Apache 2.0 (Código Abierto, Uso gratuito, Modificación y distribución permitida)
Costo	Gratuito, sin versiones de pago o comerciales.
Soporte Comercial	Disponible a través de proveedores como Cloudera o Hortonworks (opcional).
Escalabilidad	Escalabilidad horizontal y vertical; soporte para clústeres distribuidos.
Integraciones	Conexión con múltiples fuentes de datos (bases de datos, archivos, servicios web, sistemas de Big Data).
Características Principales	Interfaz gráfica, procesamiento distribuido, control de versiones, seguridad y control de datos.
Capacidades	Extraer, transformar y cargar datos, enrutamiento condicional de datos, procesamiento en tiempo real y batch.

Interfaz de usuario de apache Nifi



ACTIVIDAD PRÁCTICA GUIADA

Título: Ejecutar Apache NiFi en Cloudera y Acceder a la Interfaz de Usuario (UI)

Ojetivo : El objetivo de esta actividad es guiar a los estudiantes en el uso de Apache NiFi dentro de la plataforma Cloudera. En este ejercicio, se enseñará cómo ejecutar NiFi en un entorno administrado y acceder a la interfaz de usuario (UI) de NiFi para gestionar flujos de datos.

Iniciar sesión en Cloudera Manager

El primer paso consiste en acceder a la plataforma Cloudera Manager utilizando las credenciales proporcionadas por el administrador del sistema.

Verificar que Apache NiFi está disponible

- Acceder al Panel de Cloudera Manager
- Buscar el servicio Apache NiFi

Acceder a la Interfaz de Usuario de NiFi

Una vez identificado el servicio, hacer clic en el enlace correspondiente para abrir la interfaz de usuario de Apache NiFi.

Navegar por la Interfaz de Usuario de NiFi

Explorar las diferentes secciones y funcionalidades disponibles en la interfaz gráfica de NiFi.

Detener NiFi cuando hayas terminado

Al finalizar la práctica, asegurarse de detener correctamente el servicio para liberar recursos del sistema.