



Módulo Profesional: Big Data Aplicado

Ingestión de datos con NiFi

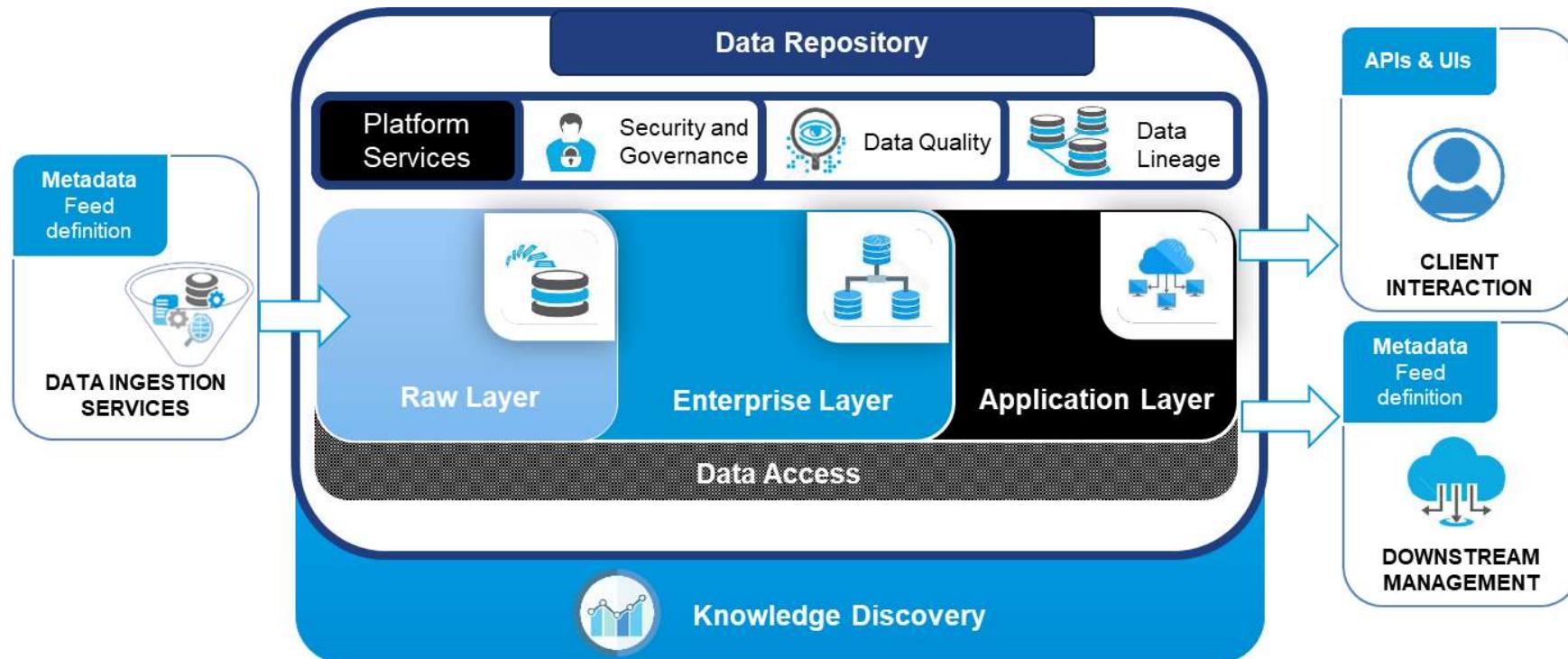
Ingestión de Datos

- **Introducción**
 - ¿Qué es la Ingesta de Datos?
 - Errores comunes
 - Herramientas de Ingestión de Datos
- **Apache NiFi**
 - Introducción y características
 - Conceptos básicos e Interfaz de usuario
 - Construyendo un Data Flow
 - Revisión en profundidad y más: procedencia de los datos, responsables de tratamientos,...
 - Subproyectos: MiNiFi y Registro
- **Historias de éxito**
- **StreamSets:**
 - Introducción y componentes
- **Airbyte: Introducción**

¿Qué es la Ingestión de Datos?

La Ingestión de Datos es el proceso de adquirir datos de los sistemas de origen y enviarlos (y/o almacenarlos) en un sistema destino, siendo clave para obtener la información mejor y más rápida, reduciendo tiempos.

Introducción



Introducción: Errores comunes



1. Formatos:

El formato de los datos puede variar como datos estructurados (JSON, XML, Bases de Datos) hasta datos no estructurados (texto, imágenes, sonido o vídeo) y también cualquier formato intermedio (CSV, registros, HTML,...).



2. Delivery:

Los datos se pueden entregar utilizando todo tipo de mecanismos, como SFTP, intermediarios de mensajes, interfaces REST, carga directa, bases de datos, sockets,...

Introducción: Errores comunes



3. Frecuencia:

Los proveedores de datos envían los datos en diferentes frecuencias, desde tiempo real, hasta por lotes, incluido casi tiempo real.



4. Cambios:

Las fuentes de datos no son estáticas y están sujetas a cambios. De hecho, a lo largo del tiempo cambiarán.

Características de una herramienta de Ingestión de Datos

Teniendo en cuenta los obstáculos mencionados anteriormente que debemos superar, ¿qué características debería tener una herramienta de Ingestión de Datos?



Fácil de añadir
nuevas fuentes y destinos



Interfaz gráfica



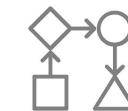
Catálogo de componentes



Escalabilidad



Monitorización



Trazabilidad

Herramientas de Ingestión de Datos



Herramientas de Ingestión de Datos



Apache NiFi es un sistema distribuido dedicado a extraer, transformar y cargar datos (ETL). Es Open Source y está desarrollado y mantenido por Apache Software Foundation.



Es una plataforma de gestión de datos en tiempo real que ofrece la capacidad de recopilar, mover, transformar y analizar datos en tiempo real. Es una solución completa que simplifica el proceso de gestión de datos al proporcionar una interfaz intuitiva y fácil de usar.

Herramientas de Ingestión de Datos



Es una aplicación con interfaz de línea de comando para transferir datos entre bases de datos relacionales y Hadoop.



Es una herramienta distribuida y Open Source. Se encarga de recopilar, agregar y mover datos desde diversas fuentes hasta almacenamientos de datos.

Herramientas de Ingestión de Datos



Open Source, herramienta para recolectar, almacenar, buscar, analizar y visualizar grandes volúmenes de datos en tiempo real, escalando horizontalmente y ofreciendo soluciones rápidas y efectivas para problemas comunes en entornos de Big Data, como la gestión de logs, la monitorización y el análisis de datos no estructurados.



Es una plataforma de análisis de datos en tiempo real que permite supervisar y analizar toda la infraestructura de TI. Esto incluye desde servidores y aplicaciones hasta dispositivos de red, bases de datos y servicios en la nube.

Herramientas de Ingestión de Datos



Es un recopilador de datos de código abierto para una capa registrada unificada.



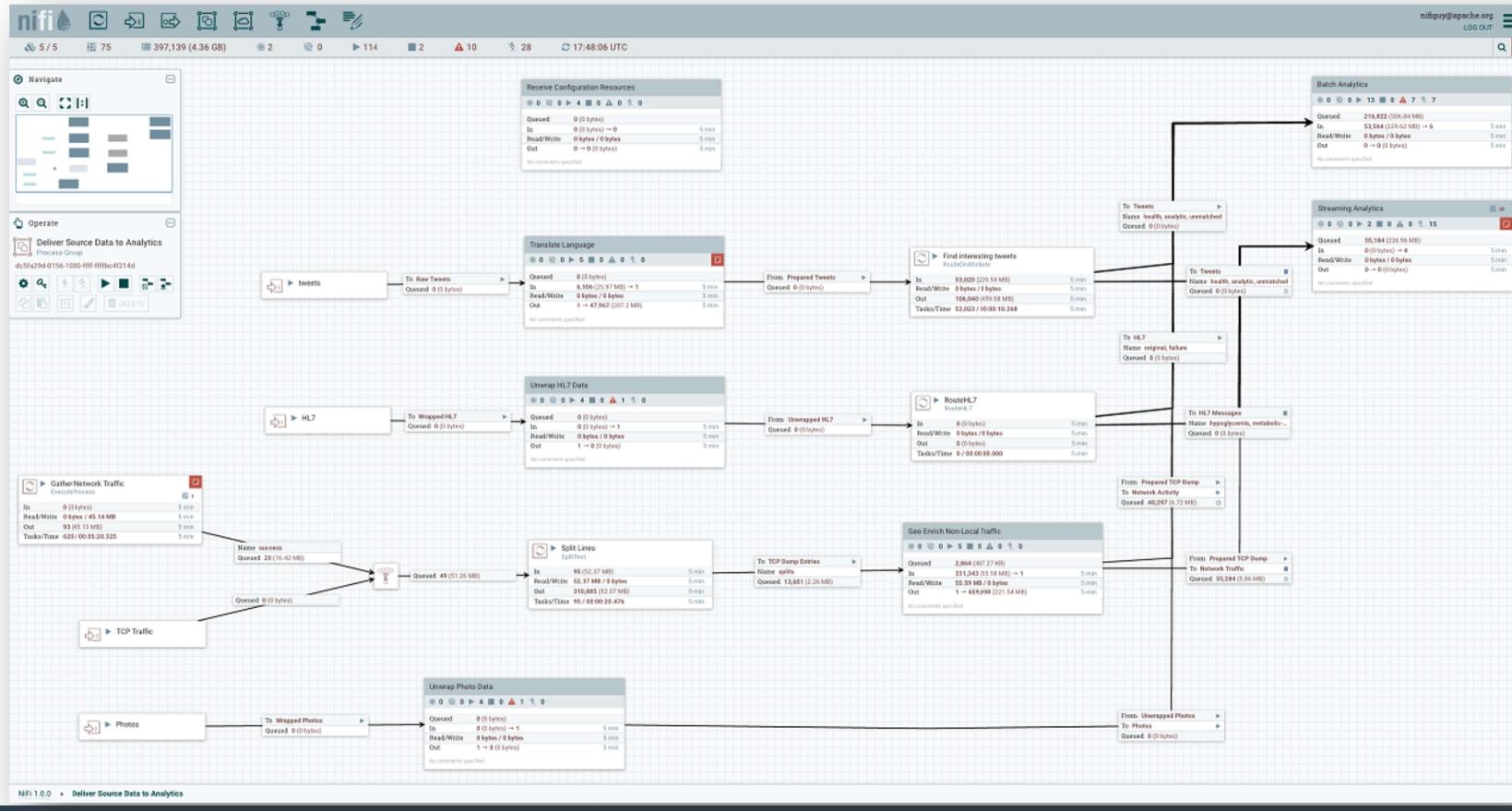
Es una infraestructura de movimiento de datos de código abierto para crear canalizaciones de datos de extracción y carga (EL).

Apache NiFi

Es un proyecto de software de **Apache Software Foundation** diseñado para automatizar el flujo de datos entre sistemas de software. *Definición de Wikipedia.*

Un sistema fácil de usar, potente y confiable para procesar y distribuir datos. Admite gráficos dirigidos potentes y escalables de enrutamiento de datos, transformación y lógica de mediación del sistema. *Sitio NiFi*

Apache NiFi: Ejemplo



Apache NiFi: Características

- **Interfaz de usuario basada en web.** Experiencia perfecta entre diseño, control, retroalimentación y monitorización.

- **Altamente configurable.**

Entrega garantizada.

Alto rendimiento

Priorización dinámica

El flujo se puede modificar en tiempo de ejecución.

Contrapresión.

- **Procedencia de los datos.**

Seguimiento del flujo de datos de principio a fin.

- **Desarrollo rápido y pruebas efectivas.**

Apache NiFi: Características

- Seguro SSL, SSH, HTTPS, contenido cifrado, etc.

Gestión de políticas/autorización interna.



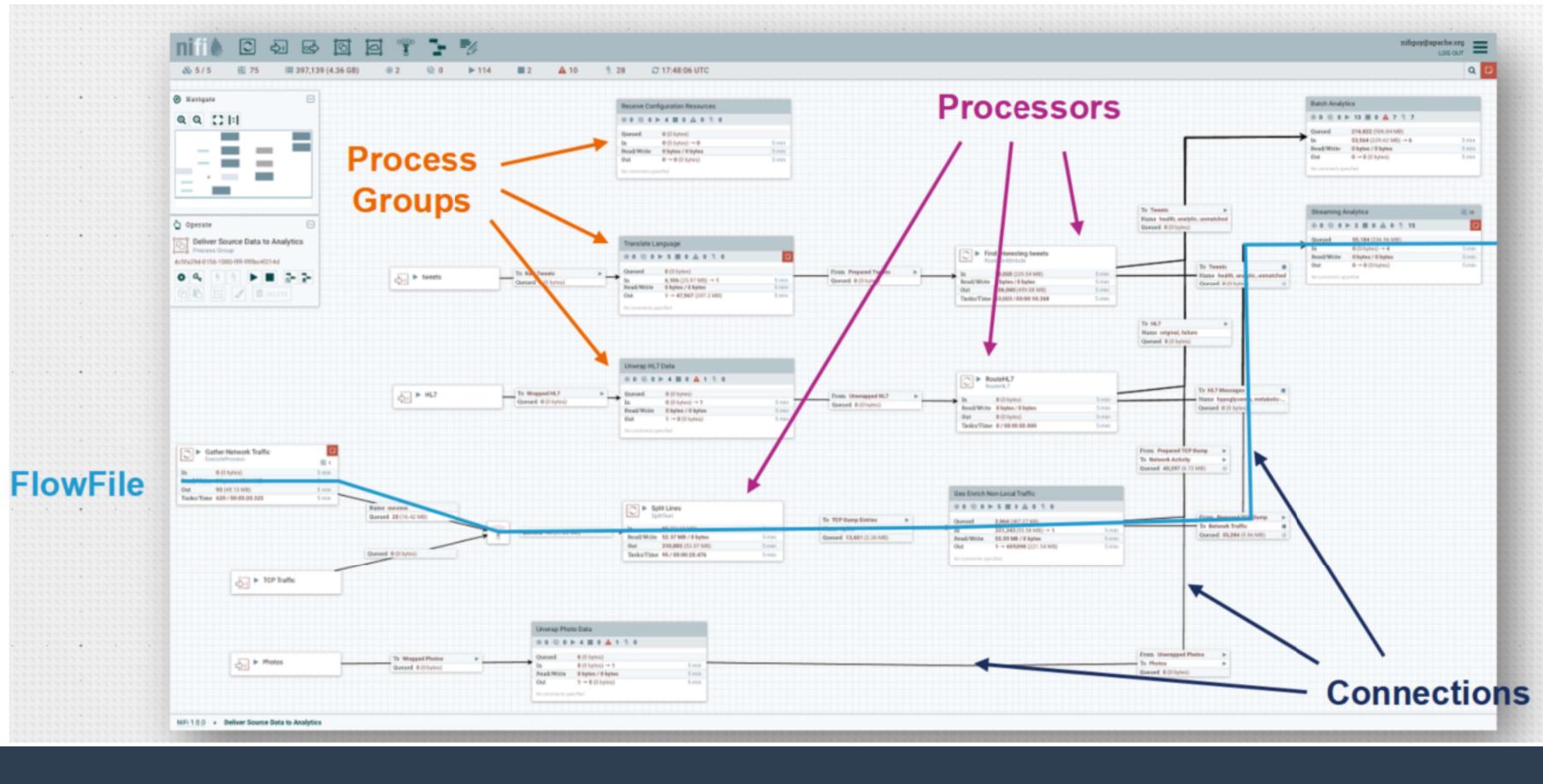
Apache NiFi: Conceptos básicos

	Términos NiFi	FBP Term	Descripción
	FlowFile	Information Packet	Un FlowFile representa cada objeto que se mueve a través del sistema y, para cada uno, NiFi realiza un seguimiento de mapa de atributos de pares clave/valor y su contenido asociado de cero o más bytes.
	FlowFile Processor	Black Box	Los procesadores realizan trabajo. Un procesador realiza alguna combinación de enrutamiento, transformación o mediación de datos entre sistemas. Los procesadores tienen acceso a los atributos de un FlowFile determinado y su flujo de contenido. Los procesadores pueden operar en cero o más FlowFiles en una unidad de trabajo determinada y confirmar ese trabajo o revertirlo.

Apache NiFi: Conceptos básicos

	Términos NiFi	FBP Term	Descripción
	Connection	Bounded Buffer	Las conexiones proporcionan el vínculo real entre los procesadores. Estos actúan como colas y permiten que varios procesos interactúen a diferentes velocidades. Estas colas se pueden priorizar dinámicamente y pueden tener límites superiores de carga, lo que permite contrapresión.
	Flow Controller	Scheduler	El Flow Controller mantiene el conocimiento de cómo se conectan los procesos y gestiona los subprocesos y su asignaciones que utilizan todos los procesos, actuando como intermediario de FlowFiles entre procesadores.
	Process Group		Un Process Group es un conjunto específico de procesadores y sus conexiones, que pueden recibir datos a través de puertos y enviar datos a través de puertos de salida.

Apache NiFi: Ejemplo



Apache NiFi: FlowFile

Attributes

```
<file name: ... >  
<file path: ... >  
<id: ... >
```

Content

* bytes

Atributos comunes:

- **uuid**: único identificador FlowFile generado
- **filename**: nombre de FlowFile
- **path**: ruta relativa del cluster
- **entrydate**: fecha y hora de cuándo fue creado FlowFile.
- **lineageStartDate**: Fecha y hora de cuando su antecesor fue creado.
- **fileSize**: tamaño del FlowFile.

Atributos específicos del tipo de procesador Definir atributos como usuarios

Formatos:

- **JSON, Avro, Text, Proprietary**