



Módulo Profesional: Big Data Aplicado

Ingestión de datos con NiFi

Procesadores - Otros

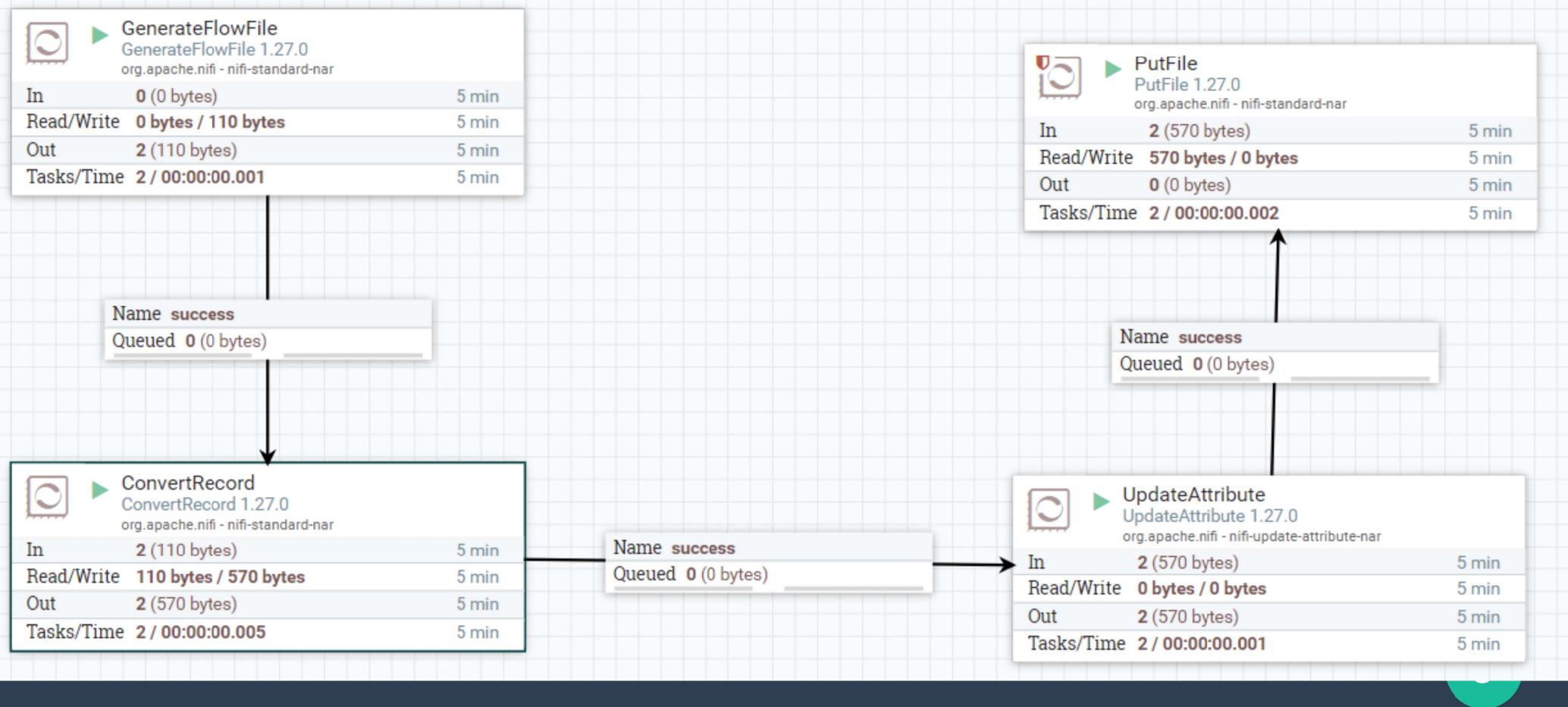
- **Conversions:** Permite realizar conversiones de formato de datos entre diferentes tipos, facilitando la interoperabilidad entre sistemas.
- **JSON to Avro:** Convierte datos en formato JSON a formato Avro, que es un formato de serialización de datos compacto y eficiente.

¿Qué es Apache Avro?

- Es un formato de datos binario muy popular en arquitecturas de streaming de eventos. Los datos se almacenan por filas, cada una de forma independiente.
- Las estructura de datos que soporta Avro son muy completas, con tipos de datos simples y complejos y se integra en su mayoría de lenguajes como Java, C++, JavaScript, Python y Ruby.

Más información: https://www.tutorialspoint.com/avro/avro_schemas.htm

Procesadores - Ejemplo de Convertir JSON a Avro



Procesadores - Otros

- **JSON to SQL:** Transforma datos en formato JSON en instrucciones SQL para insertar o actualizar registros en base de datos.
- **Avro to Parquet.** Convierte datos del formato Avro al formato Parquet, que es un formato de almacenamiento columnar optimizado para consultas analítica.

¿Qué es Apache Parquet?

- Formato de archivo de almacenamiento de datos columnar y abierto que se utiliza para almacenar datos en HDFS (*) y otros sistemas de almacenamiento distribuidos como almacenamiento de objetos.
- Este formato es compatible con muchos lenguajes de programación y utilizado por varios proyectos de Big Data como Apache Spark, Apache Hive,...

Más Información: <https://aprenderbigdata.com/apache-parquet/>

(*) Hadoop Distributed File System, componente principal del ecosistema Hadoop.

Procesadores - Otros

CSV to Avro: Convierte datos en formato CSV a formato Avro, permitiendo una mejora integración y almacenamiento eficiente de datos.

Split (JSON, XML, text...): Este procesador divide datos en múltiples FlowFiles basados en estructuras específicas, como JSON, XML o texto, facilitando el procesamiento granular de los datos.

CompressContent: Comprime los datos para reducir su tamaño y optimizar el almacenamiento o la transferencia.

Procesadores - Otros

- **Evaluate** (JSONPath, Xpath...): Evalúa y extrae información de los FlowFiles utilizando expresiones de consulta como JSONPath o facilitando la manipulación de datos estructurados.
- **Execution:** Ejecuta una script o comando externo.

Procesadores - Otros

- **Merge content:** Combina múltiples FlowFiles en un solo, permitiendo la consolidación de datos para su procesamiento o almacenamiento.
- **Parse (syslog, netflow, CEF, ...):** Analiza y transforma datos de formatos específicos como syslog,...facilitando su manipulación posterior.
- **Wait:** Pausa el procesamiento de FlowFiles hasta que se cumplan ciertas condiciones, permitiendo el control del flujo de datos basado en eventos externos.

DATA PROVENANCE

NiFi Data Provenance



NiFi Data Provenance

Displaying 1,000 of 1,000

Oldest event available: 09/30/2024 16:51:27 UTC

Showing the most recent 1,000 of 1,000+ events, please refine the search.

Filter	by component name	▼	Search			
Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type	Actions
10/14/2024 13:19:49.261 UTC	ATTRIBUTES_MODIFIED	9b82be48-b1b6-4bc8-bcf1-740a86cd...	6 bytes	ExtractText	ExtractText	🔗 ➔
10/14/2024 13:19:49.250 UTC	CONTENT_MODIFIED	9b82be48-b1b6-4bc8-bcf1-740a86cd...	6 bytes	ReplaceText	ReplaceText	🔗 ➔
10/14/2024 13:19:49.239 UTC	CREATE	9b82be48-b1b6-4bc8-bcf1-740a86cd...	10 bytes	GenerateFlowFile	GenerateFlowFile	🔗 ➔
10/14/2024 13:19:46.390 UTC	DROP	86ab85cb-59e9-411f-8854-19054b02...	277 bytes	Guardar en /tmp/out/n2yo	PutFile	🔗 ➔
10/14/2024 13:19:46.389 UTC	SEND	86ab85cb-59e9-411f-8854-19054b02...	277 bytes	Guardar en /tmp/out/n2yo	PutFile	🔗 ➔
10/14/2024 13:19:46.379 UTC	ATTRIBUTES_MODIFIED	86ab85cb-59e9-411f-8854-19054b02...	277 bytes	UpdateAttribute	UpdateAttribute	🔗 ➔
10/14/2024 13:19:46.370 UTC	RECEIVE	86ab85cb-59e9-411f-8854-19054b02...	277 bytes	International Space Station Position	InvokeHTTP	🔗 ➔
10/14/2024 13:19:46.256 UTC	ATTRIBUTES_MODIFIED	d95a87fb-189f-47c5-98d7-c03f26966...	6 bytes	ExtractText	ExtractText	🔗 ➔
10/14/2024 13:19:46.245 UTC	CONTENT_MODIFIED	d95a87fb-189f-47c5-98d7-c03f26966...	6 bytes	ReplaceText	ReplaceText	🔗 ➔
10/14/2024 13:19:46.237 UTC	CREATE	d95a87fb-189f-47c5-98d7-c03f26966...	10 bytes	GenerateFlowFile	GenerateFlowFile	🔗 ➔
10/14/2024 13:19:43.254 UTC	ATTRIBUTES_MODIFIED	148bb6ef-5af2-44ea-80f1-6a8e3fea4...	6 bytes	ExtractText	ExtractText	🔗 ➔
10/14/2024 13:19:43.244 UTC	CONTENT_MODIFIED	148bb6ef-5af2-44ea-80f1-6a8e3fea4...	6 bytes	ReplaceText	ReplaceText	🔗 ➔
10/14/2024 13:19:43.235 UTC	CREATE	148bb6ef-5af2-44ea-80f1-6a8e3fea4...	10 bytes	GenerateFlowFile	GenerateFlowFile	🔗 ➔
10/14/2024 13:19:40.262 UTC	ATTRIBUTES_MODIFIED	c76304ad-b9b4-4067-bdd9-4cbf2717...	6 bytes	ExtractText	ExtractText	🔗 ➔
10/14/2024 13:19:40.247 UTC	CONTENT_MODIFIED	c76304ad-b9b4-4067-bdd9-4cbf2717...	6 bytes	ReplaceText	ReplaceText	🔗 ➔
10/14/2024 13:19:40.231 UTC	CREATE	c76304ad-b9b4-4067-bdd9-4cbf2717...	10 bytes	GenerateFlowFile	GenerateFlowFile	🔗 ➔
10/14/2024 13:19:37.239 UTC	ATTRIBUTES_MODIFIED	873ca98d-cdbe-4675-9204-000b51c7...	6 bytes	ExtractText	ExtractText	🔗 ➔

9

NiFi Data Provenance

Ejemplo práctico de uso NiFi Data Provenance:

Supongamos que tienes un flujo en NiFi que extrae archivos CSV desde un servidor FTP, transforma los datos para ajustarse a un formato JSON y finalmente envía esos datos a una base de datos. Si en algún momento descubres que los datos en la base de datos no tienen el formato correcto o faltan registros, puedes usar la funcionalidad de Data Provenance para investigar:

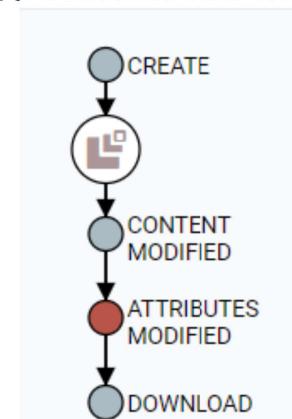
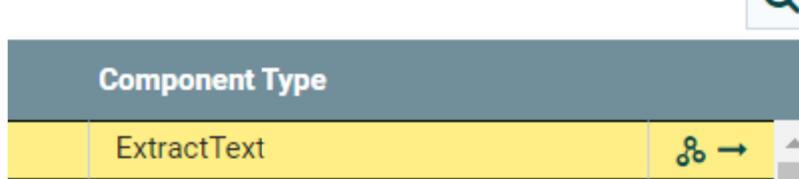
- Ver si el archivo CSV fue recibido correctamente.
- Revisar qué transformaciones se aplicaron y si ocurrieron errores durante la conversión a JSON.
- Comprobar cuándo y cómo se enviaron los datos a la base de datos.

Esto te permitirá no solo identificar el problema, sino también determinar su origen exacto y corregirlo con mayor rapidez.

En resumen, Data Provenance en NiFi proporciona una visibilidad completa y detallada del ciclo de vida de los datos, lo cual es esencial para auditoría, depuración y cumplimiento normativo.

NiFi Data Provenance

La Data Provenance (procedencia de datos) en Apache NiFi es una funcionalidad clave que permite rastrear el origen, la transformación y el movimiento de los datos a través de un flujo de procesamiento. En NiFi, cada vez que los datos se mueven o son modificados, se genera un registro detallado de esas acciones, lo que permite obtener una "historia" completa del ciclo de vida de los datos dentro del sistema. Esta funcionalidad es muy útil en escenarios donde es importante entender de dónde vienen los datos, cómo han sido procesados, y a dónde han sido enviados.



NiFi Data Provenance

NiFi Data Provenance

Displaying 1,000 of 1,000
Oldest event available: 09/30/2024 16:51:27 UTC

Filter	Date/Time	Type
	10/14/2024 13:28:19.639 UTC	ATTRIBUTE_MODIFIED
	10/14/2024 13:28:19.628 UTC	CONTENT_MODIFIED
	10/14/2024 13:28:19.627 UTC	CREATE
	10/14/2024 13:28:17.035 UTC	DROP
	10/14/2024 13:28:17.034 UTC	SEND
	10/14/2024 13:28:17.023 UTC	ATTRIBUTE_MODIFIED
	10/14/2024 13:28:17.015 UTC	RECEIVE
	10/14/2024 13:28:16.642 UTC	ATTRIBUTE_MODIFIED
	10/14/2024 13:28:16.632 UTC	CONTENT_MODIFIED
	10/14/2024 13:28:16.624 UTC	CREATE
	10/14/2024 13:28:13.645 UTC	ATTRIBUTE_MODIFIED
	10/14/2024 13:28:13.635 UTC	CONTENT_MODIFIED
	10/14/2024 13:28:13.619 UTC	CREATE
	10/14/2024 13:28:10.636 UTC	ATTRIBUTE_MODIFIED
	10/14/2024 13:28:10.626 UTC	CONTENT_MODIFIED
	10/14/2024 13:28:10.617 UTC	CREATE
	10/14/2024 13:28:07.635 UTC	ATTRIBUTE_MODIFIED

Provenance Event

DETAILS ATTRIBUTES CONTENT

Time: 10/14/2024 13:28:19.639 UTC

Event Duration: < 1ms

Lineage Duration: 00:00:00.012

Type: ATTRIBUTE_MODIFIED

FlowFile Uuid: cbcd45dd-0d82-4dd6-b574-4b117f7c5584

File Size: 6 bytes

Component Id: 8a5ce62a-0192-1000-9dd5-cd7605da9046

Component Name: ExtractText

Component Type:

Parent FlowFiles (0): No parents

Child FlowFiles (0): No children

OK

Event Duration: Duración del evento
Lineage Duration: Duración Linaje

Type: tipo

FlowFile Uuid: Identificador único Universal

File Size: Tamaño del fichero

Component Id: Identificador del componente

Component Name: Nombre del componente

Component Type: Tipo de componente.

NiFi Data Provenance

Provenance Event

DETAILS ATTRIBUTES CONTENT

Show modified attributes only

Attribute Values

contenido.0	prueba
No value set	
filename	cbed45dd-0d82-4dd6-b574-4b117f7c5584
path	./
uuid	cbed45dd-0d82-4dd6-b574-4b117f7c5584

OK

Attribute Values: son pares clave-valor que se asocian con un FlowFile

filename: Nombre del fichero

path: representa la ubicación o ruta de origen de un FlowFile en el sistema de archivos o en algún otro contexto de almacenamiento. Este atributo es parte de los metadatos asociados con el FlowFile

Uuid: Identificador único Universal

NiFi Data Provenance

Provenance Event

DETAILS ATTRIBUTES CONTENT

Input Claim		Output Claim	
Container	default	Container	default
Section	1	Section	1
Identifier	1728911911059-1	Identifier	1728911911059-1
Offset	19116	Offset	19116
Size	6 bytes	Size	6 bytes

[DOWNLOAD](#) [VIEW](#) [DOWNLOAD](#) [VIEW](#)

Replay

Connection Id
8a5eae74-0192-1000-bdc2-5a7b997e5208

Input Claim: En NiFi es una referencia a los datos reales de un FlowFile que están almacenados en el **Content Repository**. Este mecanismo permite un manejo eficiente de grandes volúmenes de datos, ya que evita la duplicación y copia innecesaria de contenido, optimizando el rendimiento y el almacenamiento dentro de NiFi.

Output Claim se refiere a la referencia a los datos generados por un FlowFile después de que han sido procesados y escritos en el **Content Repository** (repositorio de contenido). Al igual que el Input Claim, el Output Claim señala la ubicación en la que se almacenan los nuevos datos, optimizando así el manejo y el almacenamiento del contenido.

NiFi Data Provenance

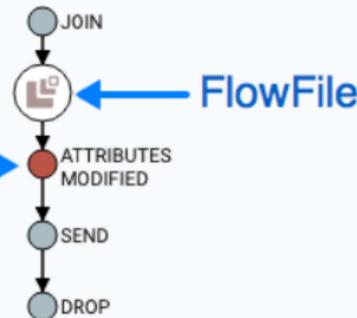
NiFi Data Provenance

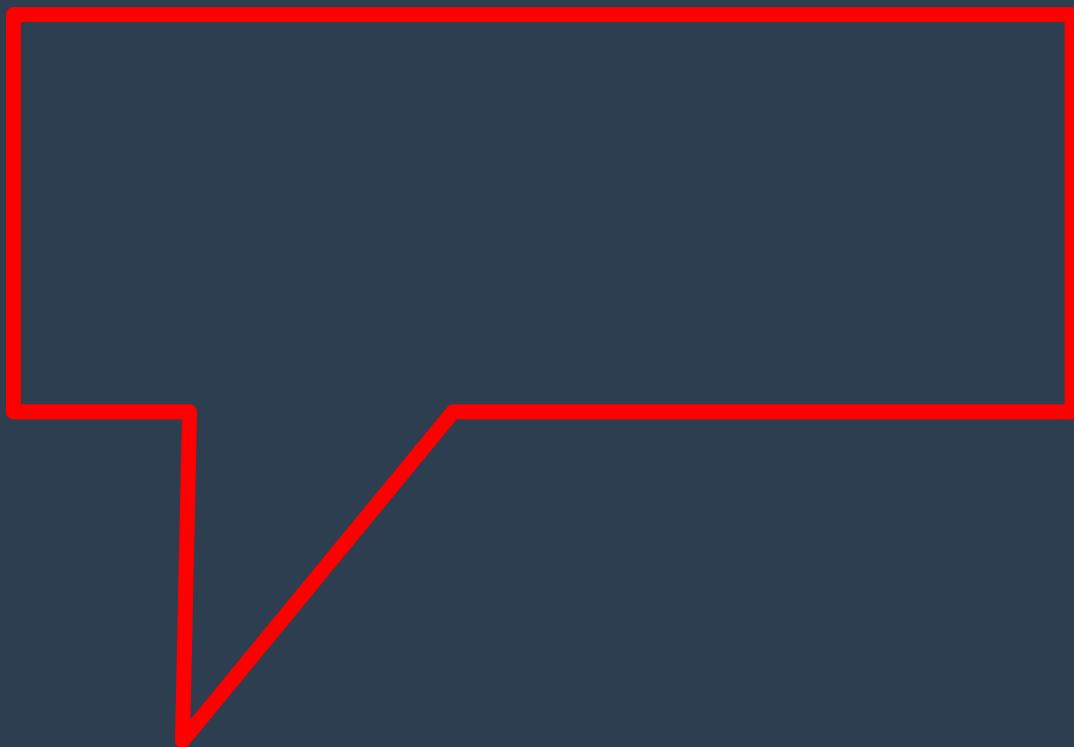
Pop out →

Download an
image of the graph

Return to
Event List

Event whose graph
was selected





16