netmind
a bts company

Módulo 2 - Data Engineering & Migration

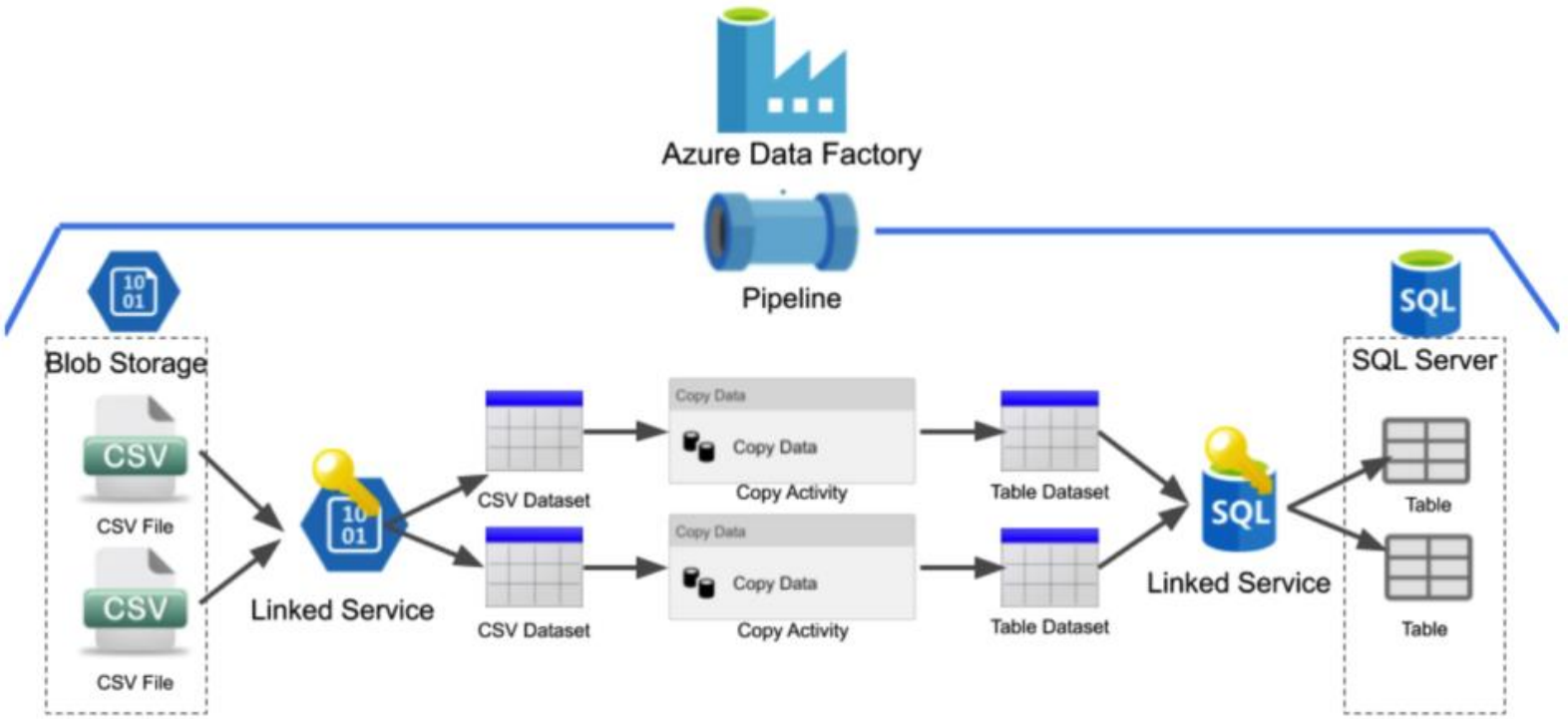
Azure Data Factory

Índice

01

Introducción. Aprovisionamiento.

Qué es Azure Data Factory



Source: medium.com

Azure Data Factory es el servicio **ETL/ELT en la nube** de Azure para la integración y transformación de datos sin servidor de escalabilidad horizontal.
Ofrece una **interfaz de usuario sin código** que favorece la creación intuitiva y una supervisión y administración desde un único panel.

https://docs.microsoft.com/es-es/azure/data-factory/

Cómo funciona

Ingest

- Multi-cloud and on-premise hybrid copy data
- 100+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

Control Flow

- Design code-free data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, variables, parameters, …

Data Flow

- Code-free data transformations that execute in Spark
- Scale-out with Azure Integration Runtimes
- Generate data flows via SDK
- Designers for data engineers and data analysts

Schedule

- Build and maintain operational schedules for your data pipelines
- Wall clock, event-based, tumbling windows, chained

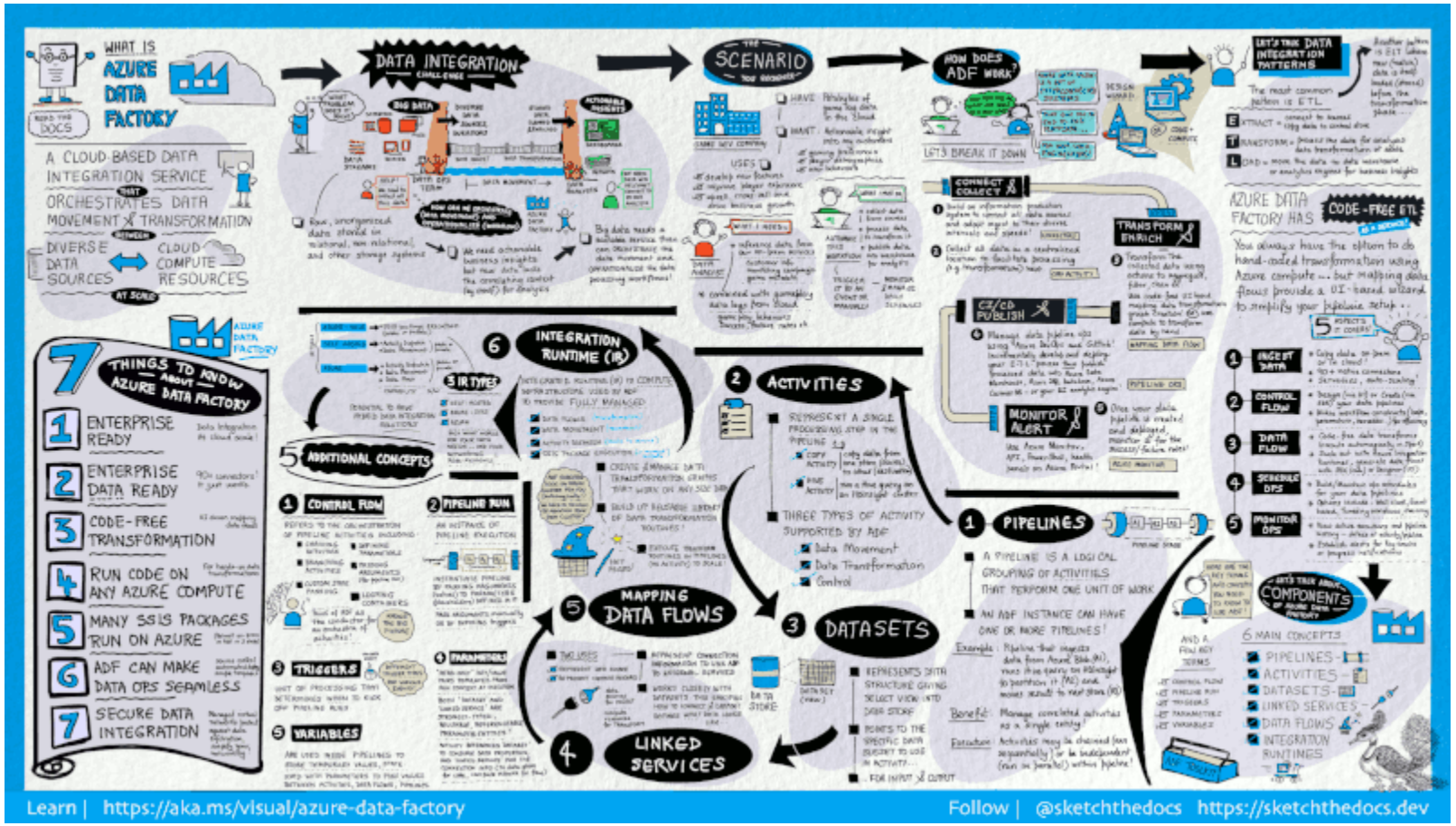
Monitor

- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

Source: docs.microsoft.com

Data Factory contiene una serie de sistemas interconectados que proporcionan una plataforma completa de un extremo a otro para los ingenieros de datos.

Guía visual



Source: docs.microsoft.com

https://docs.microsoft.com/es-es/azure/data-factory/media/introduction/data-factory-visual-guide.png#lightbox

Resource Group

- Contenedor que contiene recursos relacionados
- Puede contener todos los recursos para la solución o recursos selectivos
- Despliégua, actualiza y elimina recursos como un grupo
- Almacena metadatos sobre los recursos.

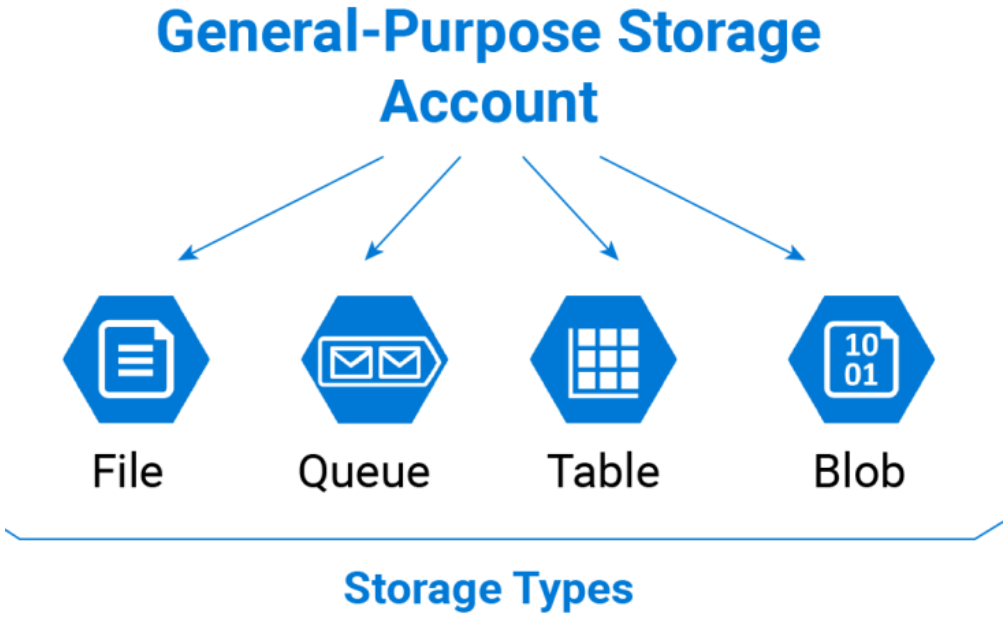
Azure Storage account

¿Qué es un atorage account?

- Cuenta de almacenamiento de propósito general
- Proporciona almacenamiento en la nube altamente disponible, seguro, duradero, escalable y redundante.

Servicios disponibles

- Tablas
- Colas
- Archivos
- Blobs
- Azure VM Disks



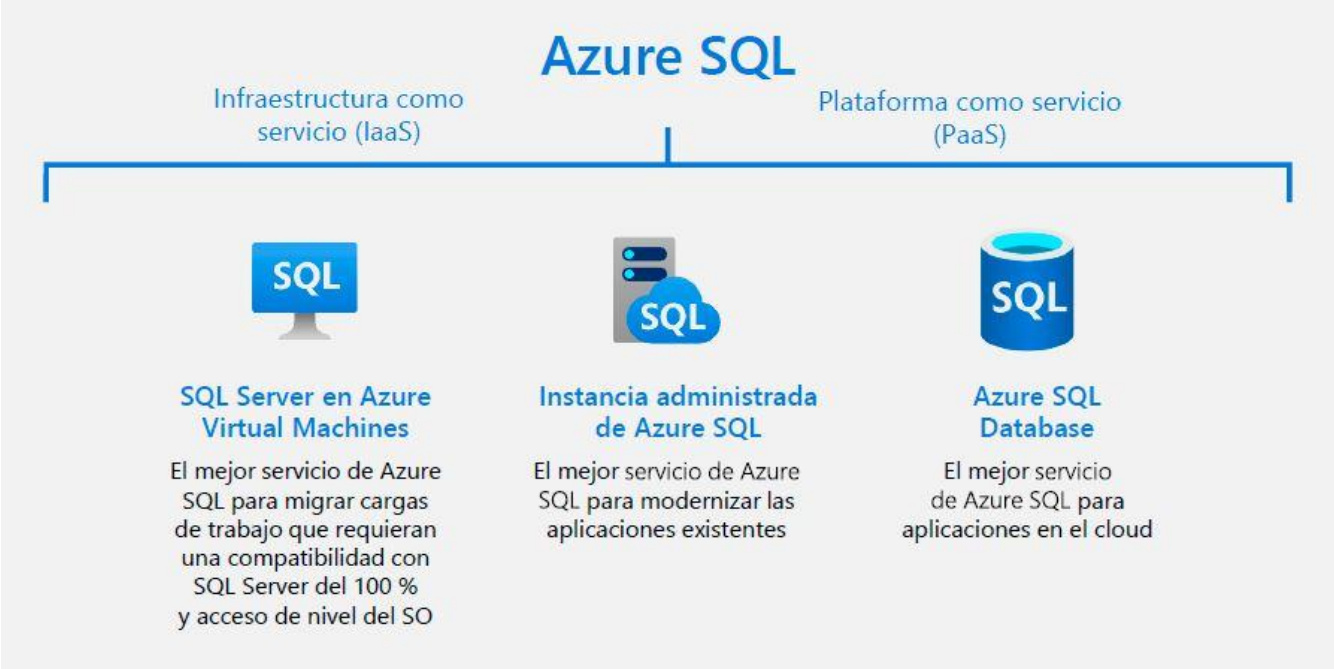
Azure SQL DB

¿Qué es un atorage account?

- Base de datos relacional de propósito general

Estructuras soportadas

- Datos relacionales
- JSON
- Espacial
- XML



Aprovisionamiento de Azure Data Factory

Data Factory

- El nombre debe ser globalmente único.
- Suscripción
- Grupo de recursos
- Versión (V1 frente a V2)
- Ubicación
- Control de versiones

ACTIVIDAD

Crear recursos

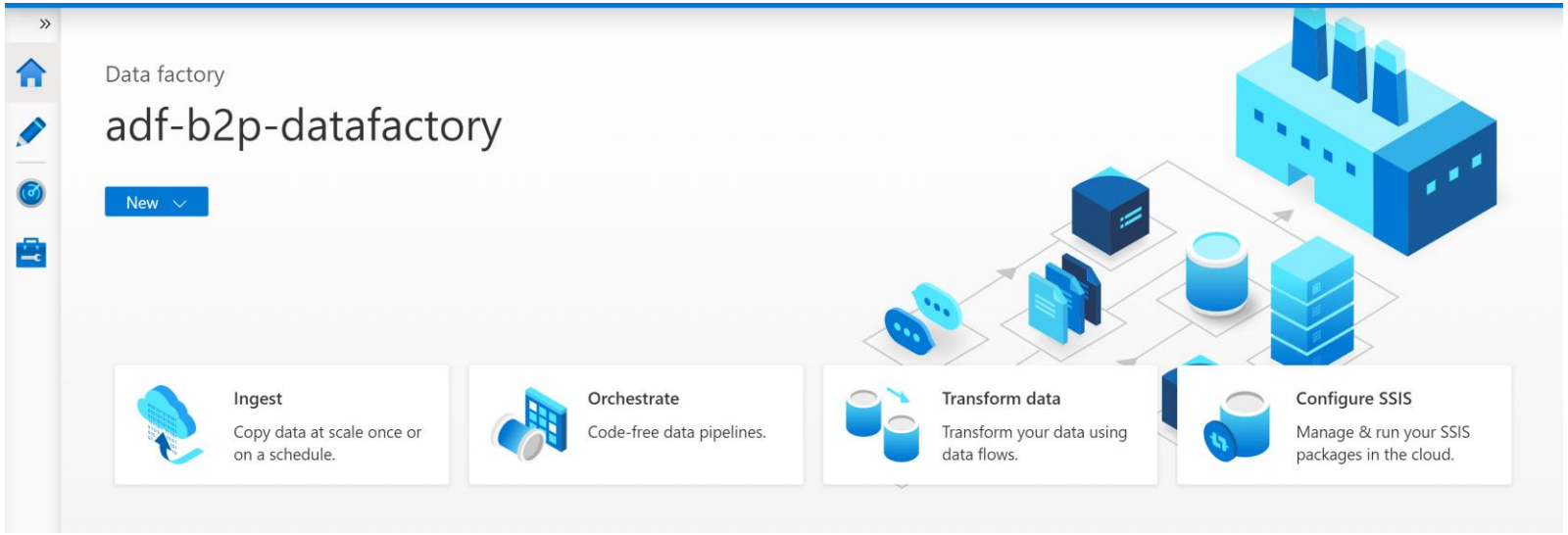
02

Entorno de Data Factory

Home

Actions

- Ingest (Copy Data Activity Wizard)
- Orchestrate (Create Pipeline)
- Transform Data (Create Data flow)
- Configure SSIS Runtime



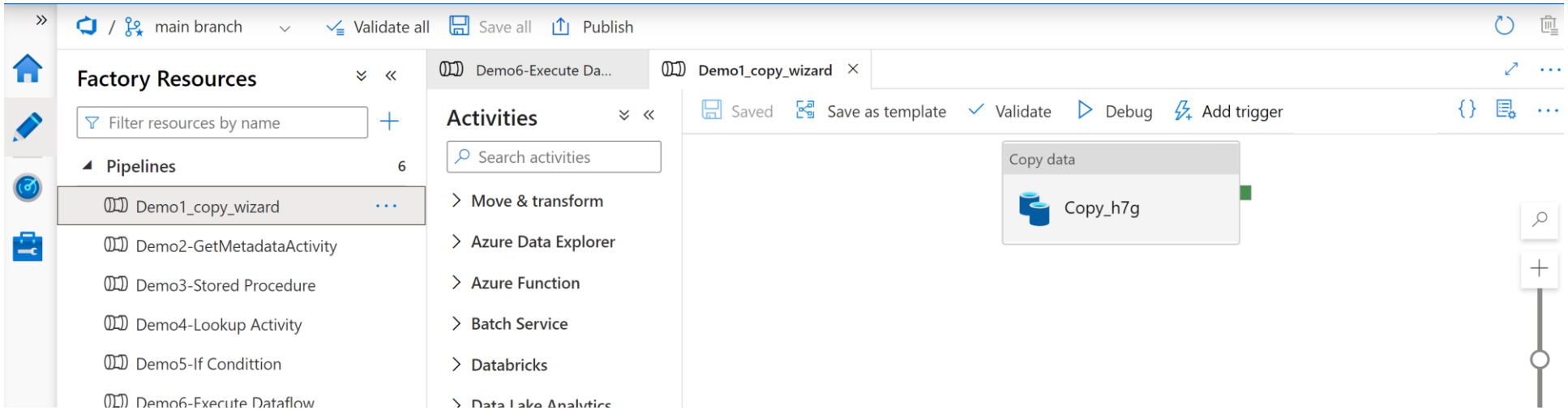
Otras áreas

- Discover More
- Recent Resources
- Feature showcase
- Resources

Centro de autor

Área de Diseño

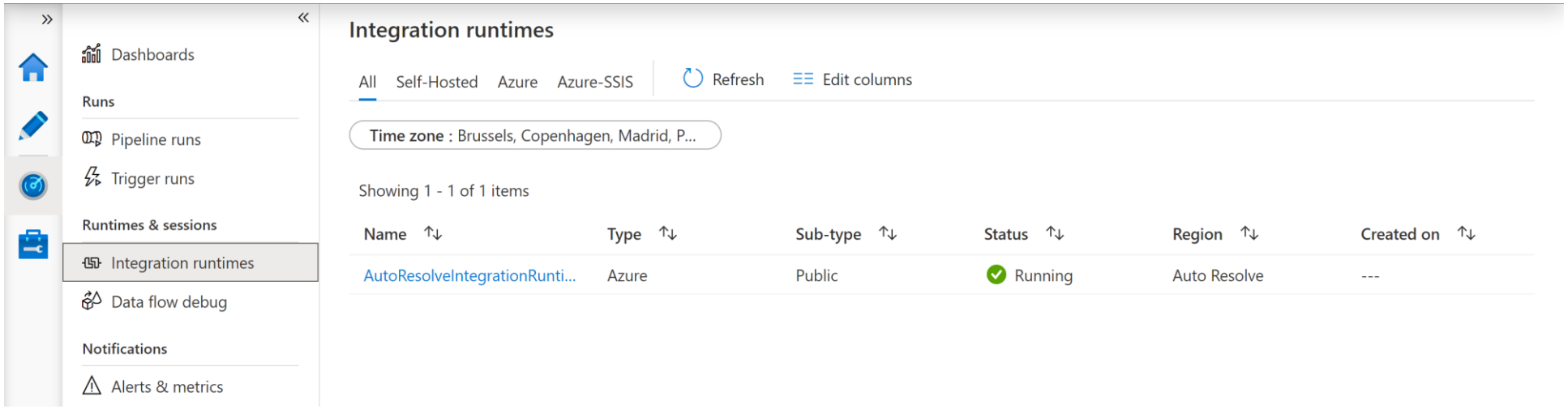
- Pipelines
- Datasets
- Data flows
- Power Query



Centro de monitoreo

Opciones de monitoreo

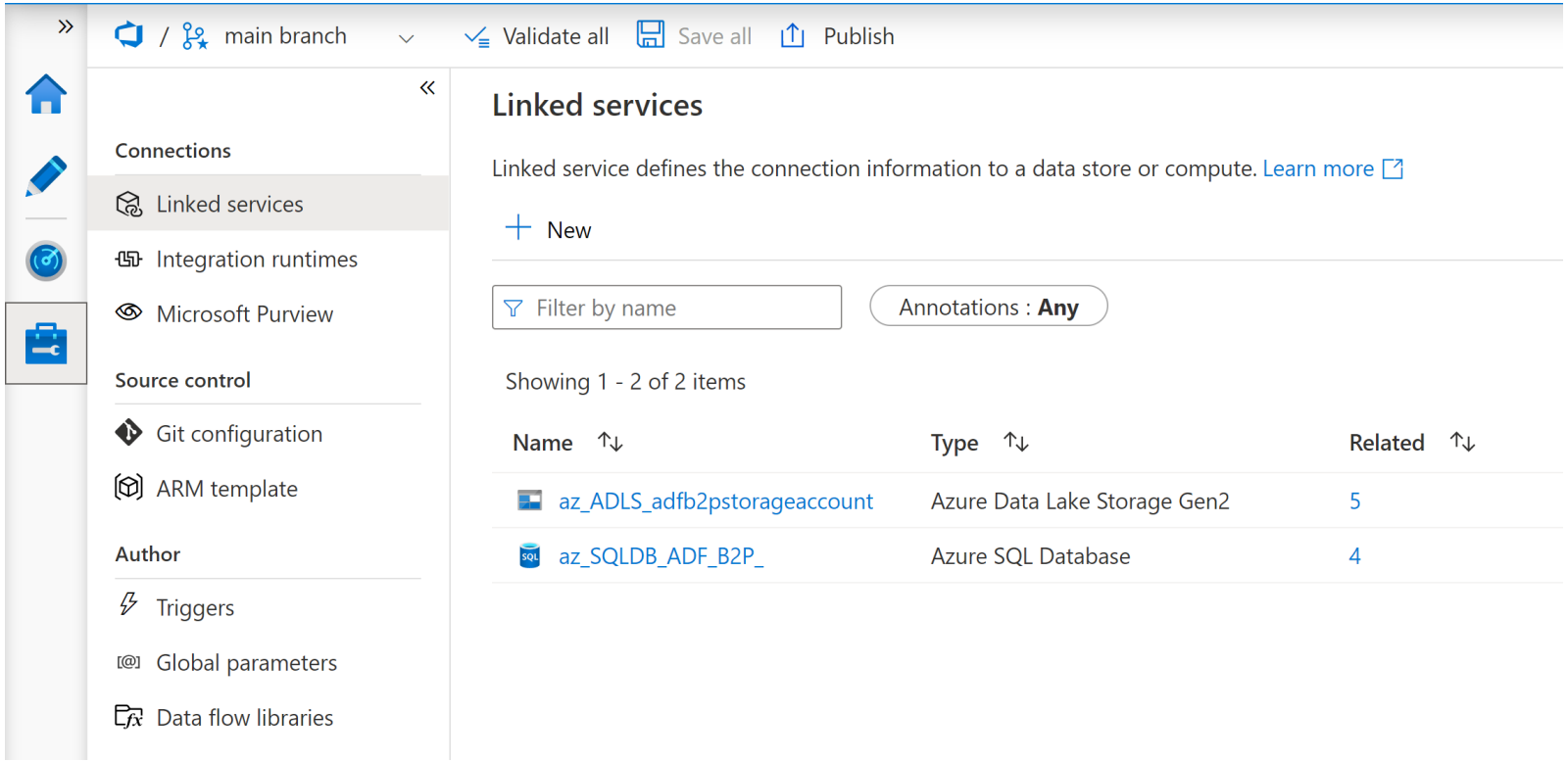
- Dashboards
- Pipeline Runs
- Trigger Runs
- Integration Runtimes
- Data flow debug



Centro de gestión

Opciones de administrador

- Connections
- Source Control
- Author
- Security



03

Recursos de Data Factory

Integration Runtimes

Integration Runtime (centro de gestión)

- Es la infraestructura informática utilizada por ADF para proporcionar las siguientes capacidades de integración de datos:
    1. Movimiento de datos (Azure IR, IR autohospedado)
    2. Ejecución del paquete SSIS (Azure-SSIS IR)

Integration Runtime autohospedado

- Capaz de ejecutar actividades de copia entre almacenes de datos en la nube y almacenes de datos privados

Linked Services y Datasets

Linked Services (centro de gestión)

- Define la información de conexión para que Data Factory pueda conectarse al origen de datos.
- Se puede reutilizar entre pipelines en una factoría de datos

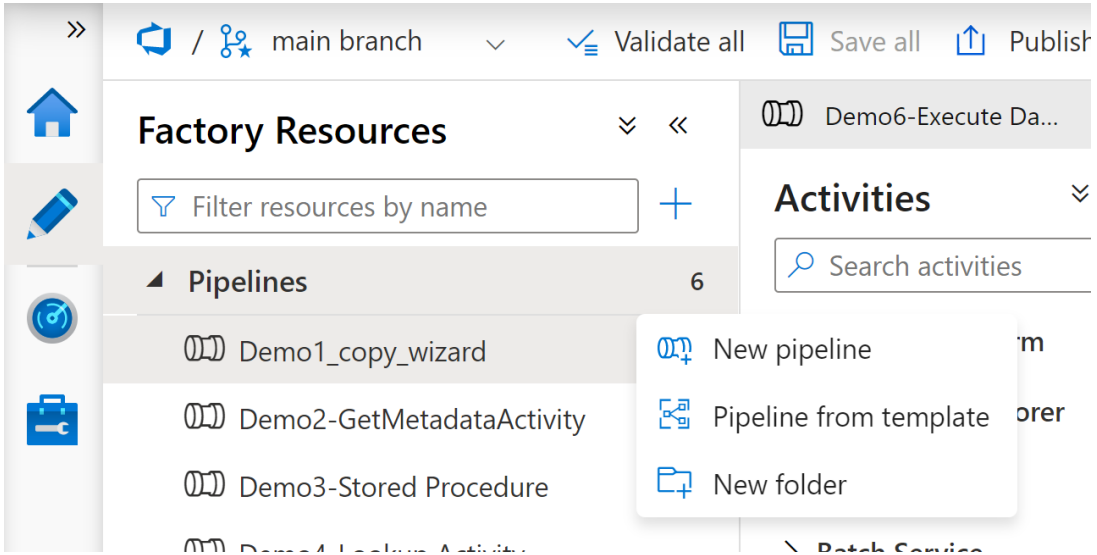
Conjuntos de datos/Datasets (centro de autor)

- Vista con nombre de datos que señala o hace referencia a los datos
- Almacenes de datos: tablas, archivos, carpetas y documentos

Organización de recursos

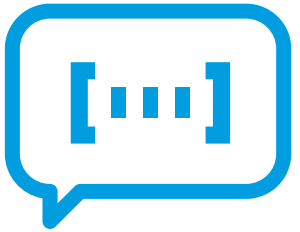
Carpetas

- Se utiliza para agrupar recursos de canalización
- Se utiliza para agrupar recursos de conjuntos de datos
- Se utiliza para agrupar flujos de datos



ACTIVIDAD

Crear un Linked Service



Actividades y pipelines

04

Copy Activity Wizard

Cadencia de tareas o progrmación

- Ejecutar una vez ahora
- Ejecutar regularmente según lo programado (Crear trigger)

Almacén de origen

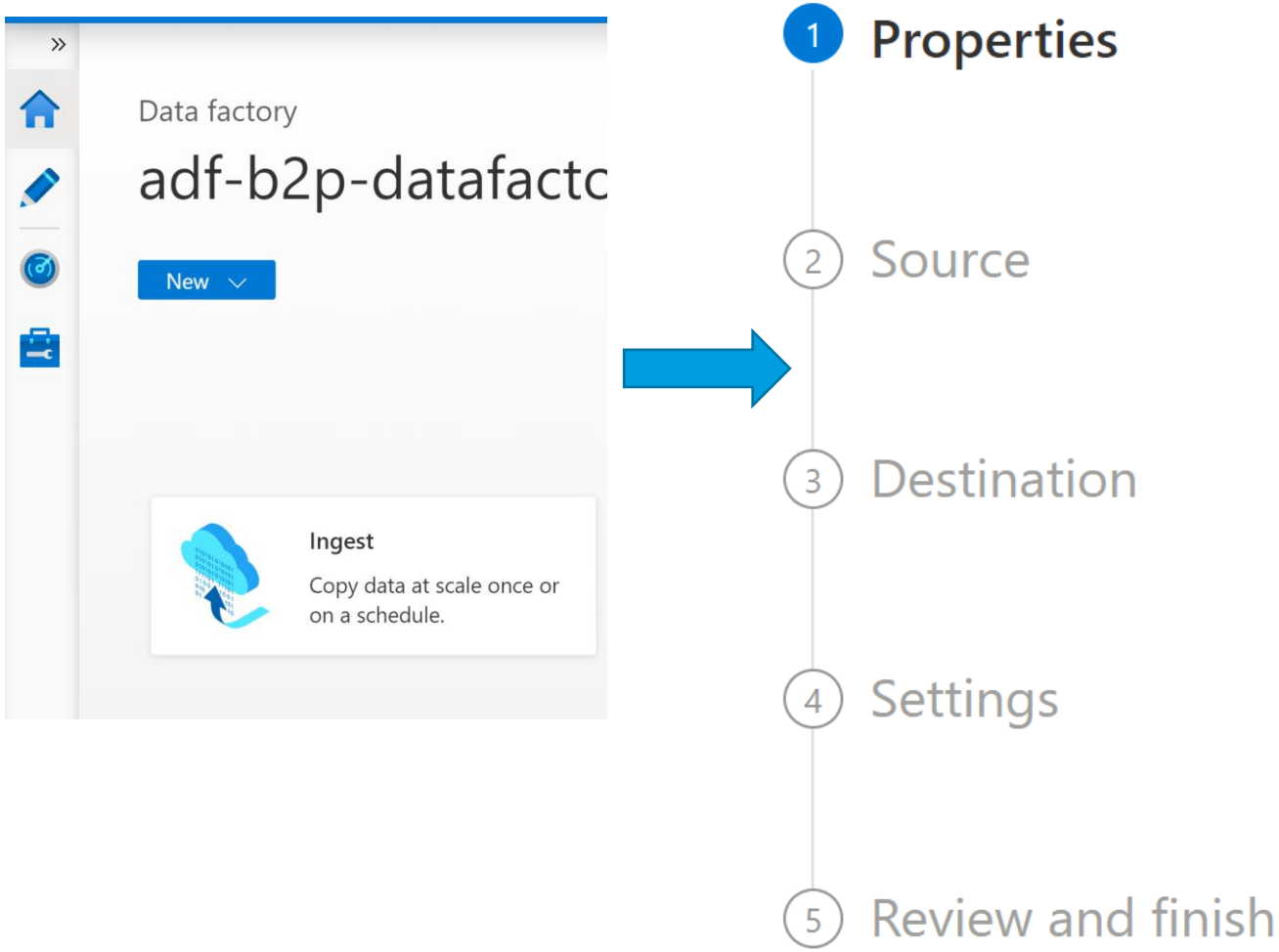
- Elijir un conjunto de datos existente
- Crear nuevo conjunto de datos

Almacén de destino

- Elijir un conjunto de datos existente
- Crear nuevo conjunto de datos

Ajustes

- Unidad de Integración de Datos
- Grado de paralelismo de copia



ACTIVIDAD

Demo Copy Activity Wizard and Pipeline

ACTIVIDAD

Demo Copy Pipeline

OBJETIVO

Pipeline de copia de datos de datalake a DB

INSTRUCCIONES

1. Crea un pipeline para copiar los contenidos del archivo Emp2.csv (carpeta Exercises) a la tabla dbo.emp.
2. Modifica el proceso anterior para que se copien a una nueva tabla dbo.emp2.
3. Crea otro bloque de copia para copiar un subconjunto de datos del archivo iris.csv a una nueva tabla dbo.iris. Concretamente solo se deben copiar las columnas: "petal.length","petal.width","variety" en "petal_length", "petal_width", "variety"

20 min

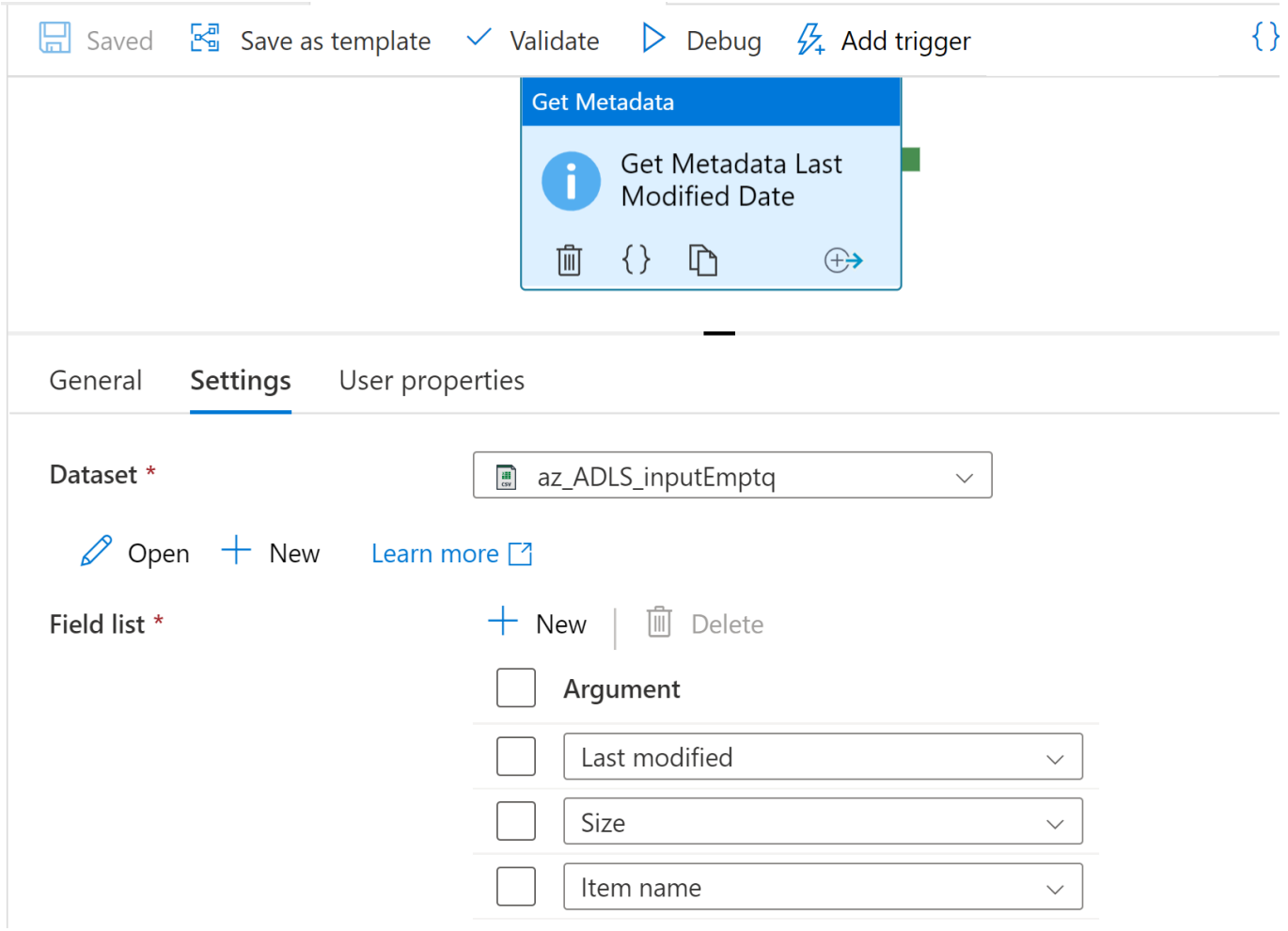
Get Metadata activity

Objetivo

- Recuperar información de metadatos de datos

Opciones de metadatos

- Nombre del item
- Tipo de item
- Tamaño
- Creado
- Última modificación
- Artículos secundarios
- Contenido MD5
- Estructura
- Recuento de columnas
- Existe



Get Metadata activity – Parámetros de salida

Parámetros de salida

- La salida se pueden utilizar en otras actividades.

Nombres de parámetros de salida

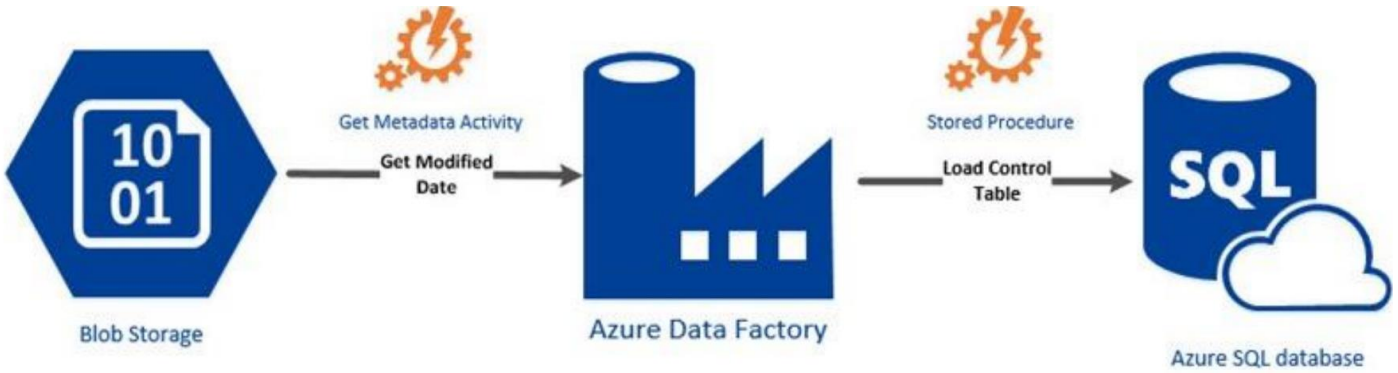
- Agregar contenido dinámico
- Resultados de depuración (salida de actividad)


```
{
    "lastModified": "2022-09-06T08:29:47Z",
    "size": 52,
    "itemName": "inputEmp_tq.txt",
    "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (East US)",
    "executionDuration": 0,
    "durationInQueue": {
        "integrationRuntimeQueue": 0
    },
    "billingReference": {
        "activityType": "PipelineActivity",
        "billableDuration": [
            {
                "meterType": "AzureIR",
                "duration": 0.016666666666666666,
                "unit": "Hours"
            }
        ]
    }
}
```

ACTIVIDAD

Demo Get Metadata Activity

Metadata Activity → Stored Procedure Activity



OBJETIVO

Metadata de varios archivos

INSTRUCCIONES

1. Crea un nuevo dataset que incluya toda la carpeta Source
2. Revisa la actividad "ForEach":
   - https://learn.microsoft.com/es-es/azure/data-factory/control-flow-for-each-activity
3. Crea un pipeline que extraiga los metadatos (lastmodified, size, name) de todos los archivos

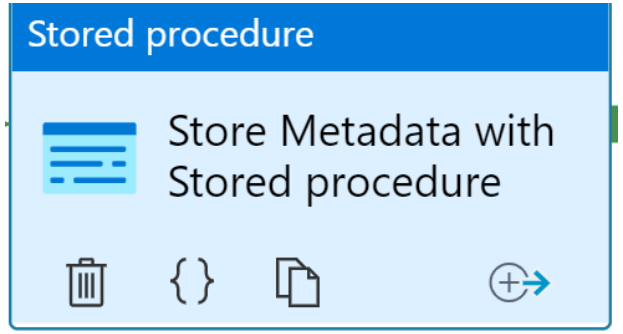
- Tip: https://stackoverflow.com/questions/64504159/get-all-files-names-in-subfolders-azure-data-factory

20 min

Stored Procedure Activity

Objetivo

- Invocar un procedimiento almacenado
- Utilizar productos de otras actividades



Soporta

- Base de datos Azure SQL
- Synapse Analytics (Azure SQL DW)
- Base de datos del servidor SQL

Limitaciones

- No hay parámetros de salida a ADF

ACTIVIDAD

Demo Stored Procedure Activity

OBJETIVO

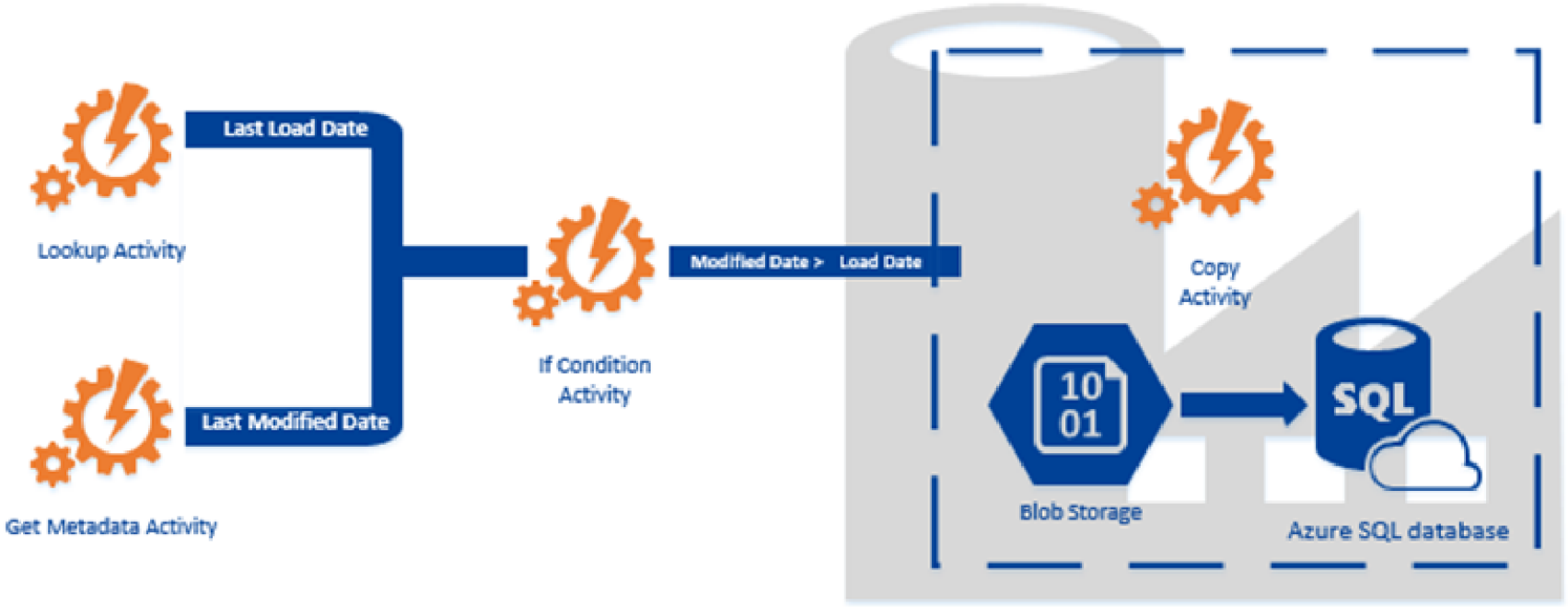
Stored procedure para varios archivos

INSTRUCCIONES

1. Crea un nuevo pipeline que use el ejercicio “Metadata de varios archivos” para almacenar la meta-información de todos los archivos de la carpeta Source.

20 min

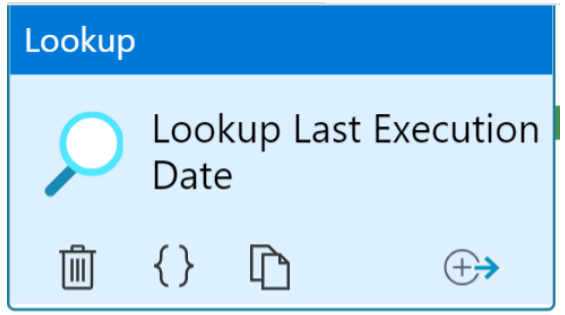
Diseño del Pipeline



Lookup Activity

Objetivo

- Recuperar un conjunto de datos



Soporta

- Cualquier origen de datos de Azure Data Factory
- Ejecución de procedimientos almacenados
- Ejecución de secuencias de comandos SQL
- Parámetros de salida

Salidas

- Valor único
- Matriz / Objeto

ACTIVIDAD

Demo Lookup Activity

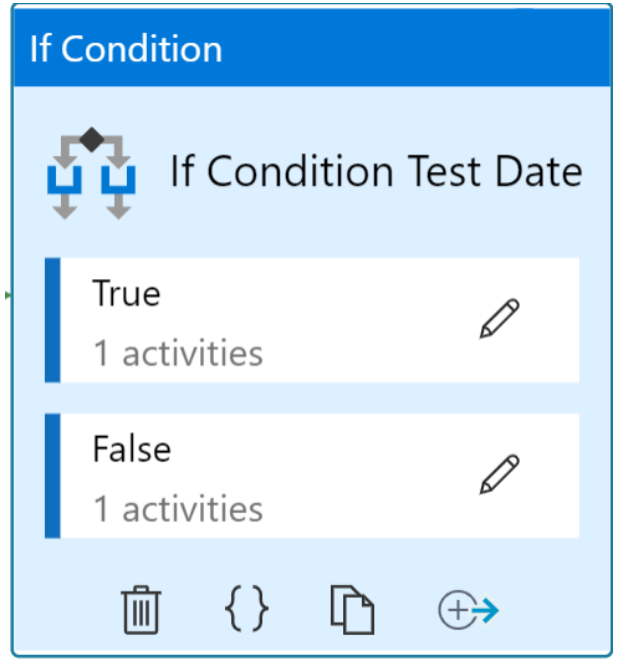
If Condition Activity

Objetivo

- Funcionalidad de sentencia if
- Expresión booleana (Verdadero/Falso)

Soporta

- Expresiones y funciones ADF
- Si las actividades verdaderas
- Si las actividades son falsas



ACTIVIDAD

Demo If Condition Activity

OBJETIVO

Más Actividades de Flujo de Control

INSTRUCCIONES

1. En la documentación de Azure Data Factory (https://learn.microsoft.com/es-es/azure/data-factory/), ve a la sección:
   - Guías paso a paso > Flujo de control
2. Revisa cada una de las actividades que pueden usarse para generar un pipeline.
   - Experimenta con ellas en tu Azure Data Factory

OBJETIVO

Actividad ForEach

INSTRUCCIONES

1. Modifica el pipeline para actualizar la tabla de control para más de un fichero según la última ejecución que ocurrió.

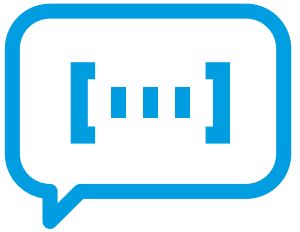
OBJETIVO

Pipeline de ventas

INSTRUCCIONES

1. Considera los archivos de la carpeta Demo5.
2. Crea un pipeline que cargue los datos de cada uno de los ficheros de la carpeta internetSales en la tabla TaskQueue
    1. Para crear la tabla usa: Create_TaskQueue.sql
    2. Para actualizar los datos usa el procedimiento definido en: usp_InsertFileNames.sql

netmind
a bts company



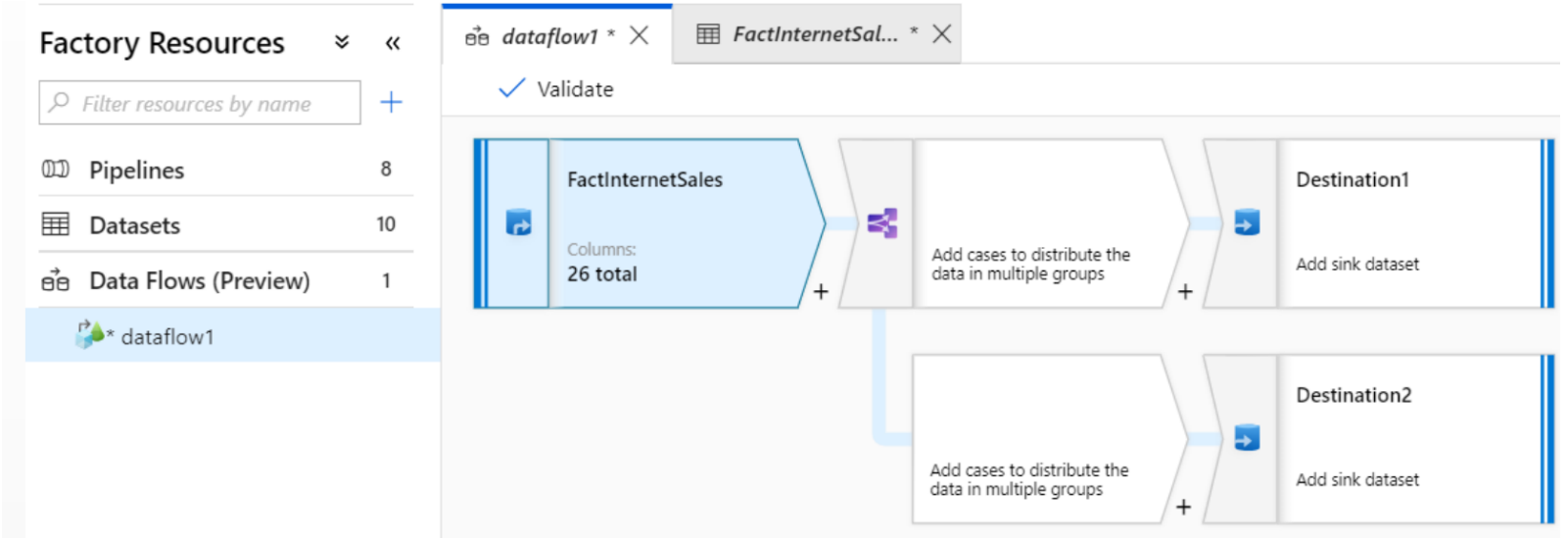
Data Flows

05

Data Flows

Objetivo

- Permite transformaciones de datos.



Elementos

- Fuente
- Transformaciones
- Sink

Cómo ejecutar

- Depurar
- Actividad de flujo de datos

Código ADF se convierte a Scala

- Los flujos de datos se ejecutan en Azure Databricks.
- Escalado automático según sea necesario

Más en: https://learn.microsoft.com/es-es/azure/data-factory/concepts-data-flow-overview

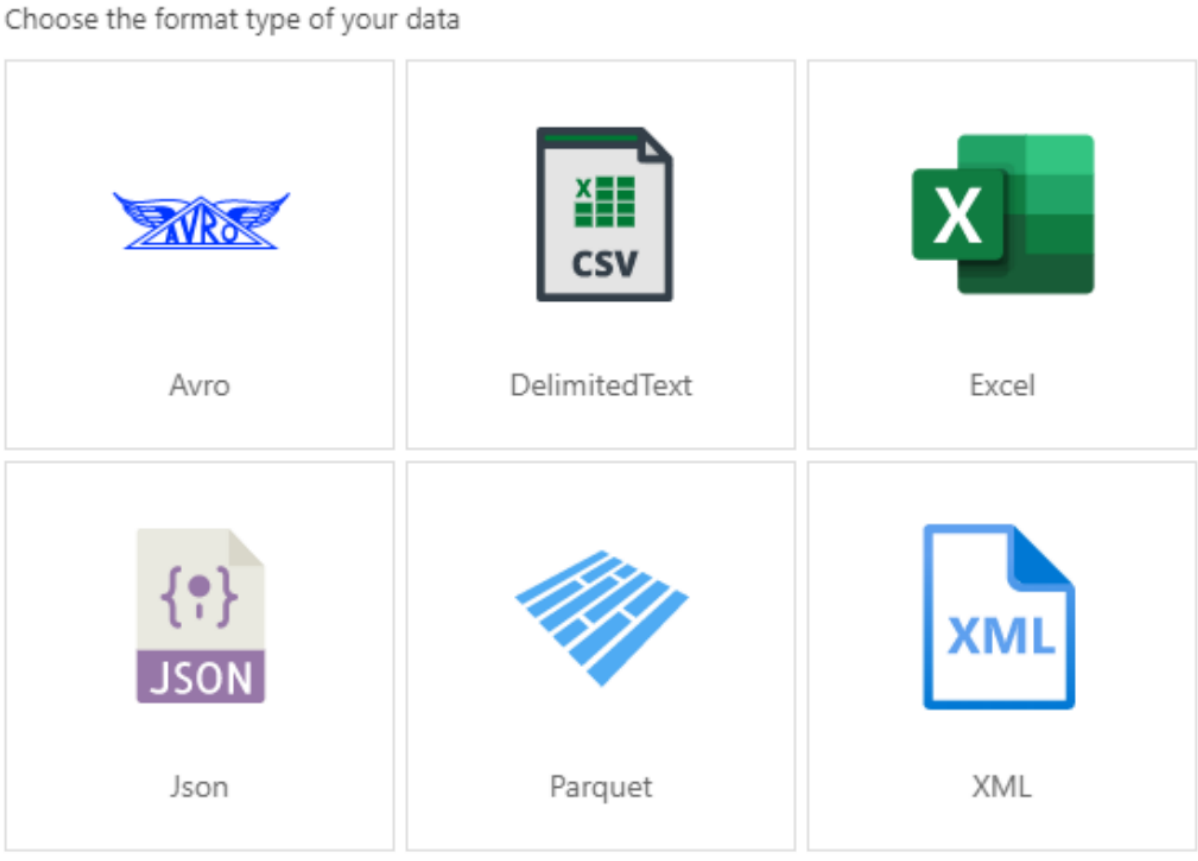
Parquet

Formato de archivo

- Almacenamiento de datos orientado a columnas
- formato vs orientado a filas

Beneficios

- Almacenamiento
- Actuación



OBJETIVO

CSV a Parquet

INSTRUCCIONES

1. Crea un pipeline para convertir iris.csv a un fichero con formato parquet iris.parquet
2. Revisa el archivo resultante en el datalake.
3. Usa un viewer para parquet para revisar el resultado (https://www.parquet-viewer.com/)

OBJETIVO

SQL a Parquet

INSTRUCCIONES

1. Crea un pipeline para convertir dbo.emp a un fichero con formato parquet emp.parquet

Tip: https://www.tech-findings.com/2021/09/i-getting-started-with-adf-creating-and.html

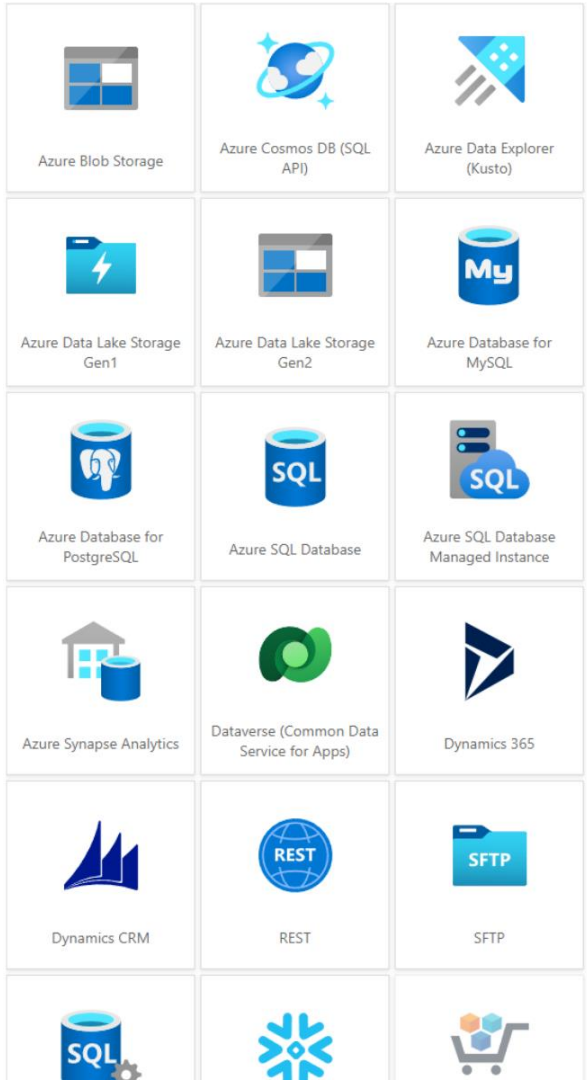
Fuente de datos

Opciones Disponibles

- Azure SQL Data Warehouse
- Azure SQL Database
- Cosmos DB
- Azure Blob
- ADLS Gen1/2
- Synapse Analytics

Items

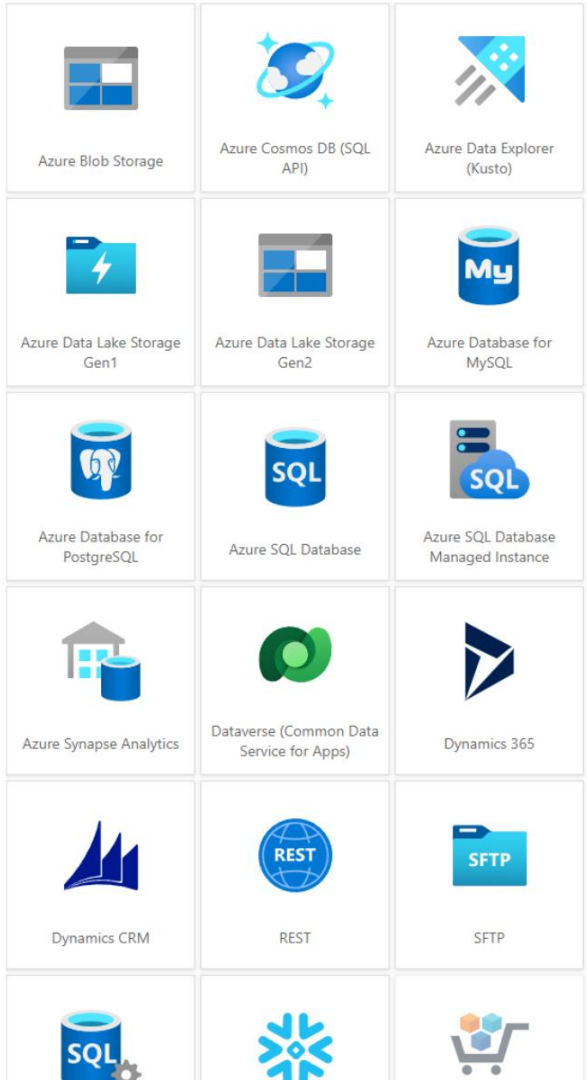
- Mínimo de 1 fuente



Transformaciones

Opciones Disponibles

- New Branch
- Join
- Conditional Split
- Derived Column
- Lookup
- Select
- Sort
- Filter
- Etc…



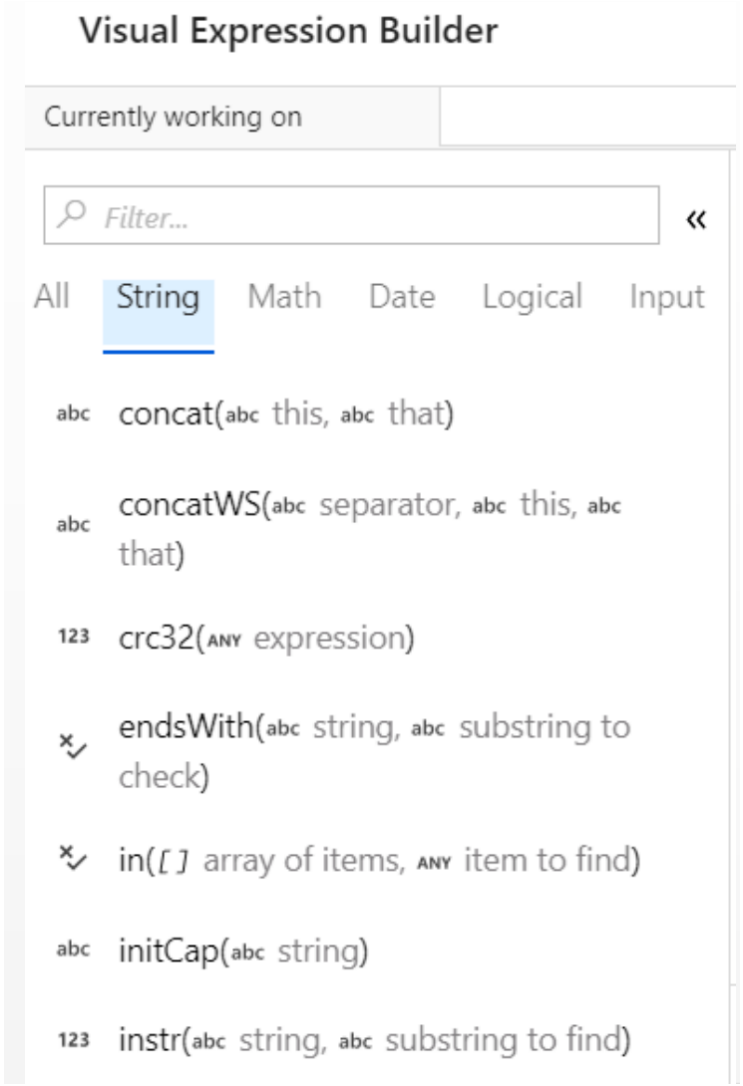
Expresiones

Generador de expresiones visuales

- Ciertas transformaciones requieren el uso del lenguaje de expresión ADF

Debug

- Permite tener una vista previa en vivo y en directo de los resultados de la expresión que se está creando.



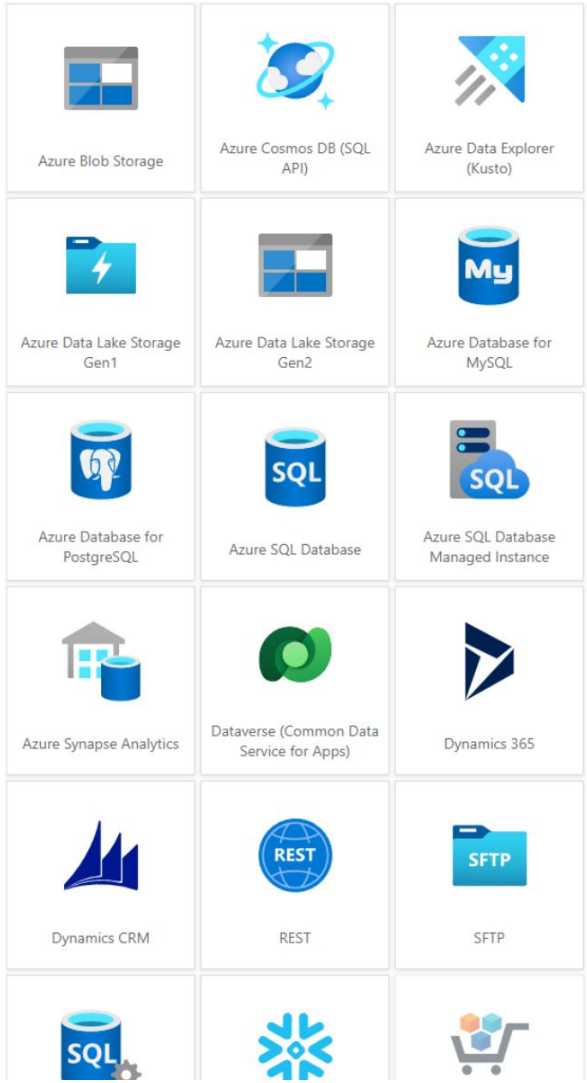
Sink (destino)

Opciones disponibles

- Azure SQL Data Warehouse
- Azure SQL Database
- Cosmos DB
- Azure Blob
- ADLS Gen1/2
- Synapse Analytics

Items

- Mínimo de 1 sink

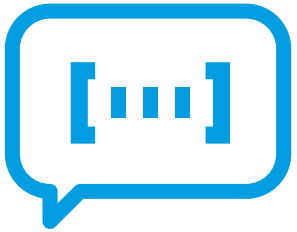


ACTIVIDAD

Demo Data Flow

Escenario de caso

- Mi organización ha solicitado obtener un archivo que enumere todos los productos que se venden. (Fuente)
- También quieren la descripción del modelo del producto que proviene de una tabla diferente. (Buscar y seleccionar)
- Debe incluirse el peso del envío, pero debe calcularse añadiendo al peso real un 10% para tener en cuenta el embalaje (columna derivada)
- No necesitamos productos que tengan un precio de lista de $0.00 (Filtro)
- Finalmente, necesitamos ordenar los datos de salida en un archivo por precio descendente (Sort & Sink)



Triggers

06

Triggers

Schedule trigger

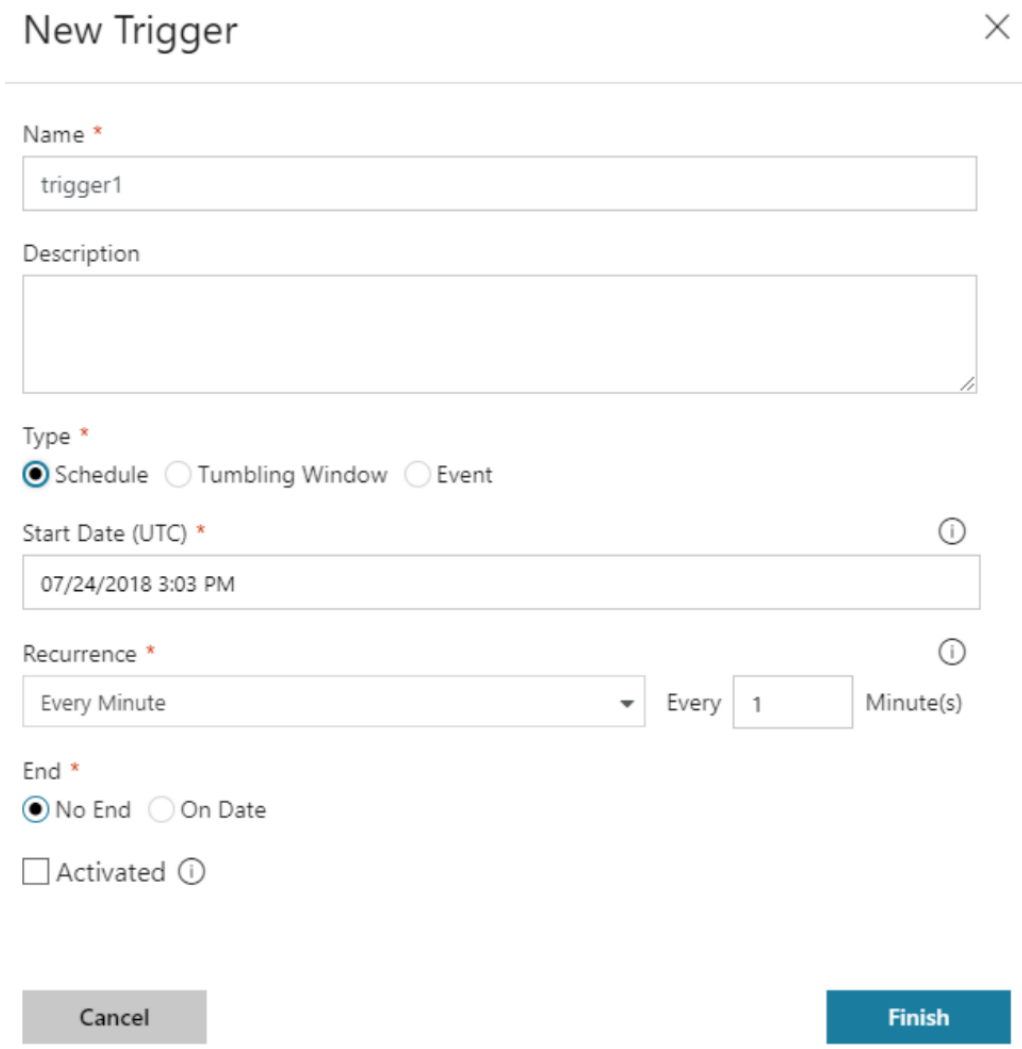
- Invoca un pipeline en un horario concreto.

Tumbling window trigger

- Opera en un intervalo periódico, al mismo tiempo que retiene el estado

Event-based trigger

- Responde a un evento



Más en: https://learn.microsoft.com/es-es/azure/data-factory/concepts-pipeline-execution-triggers

ACTIVIDAD

Demo Trigger

OBJETIVO

Diferentes triggers

INSTRUCCIONES

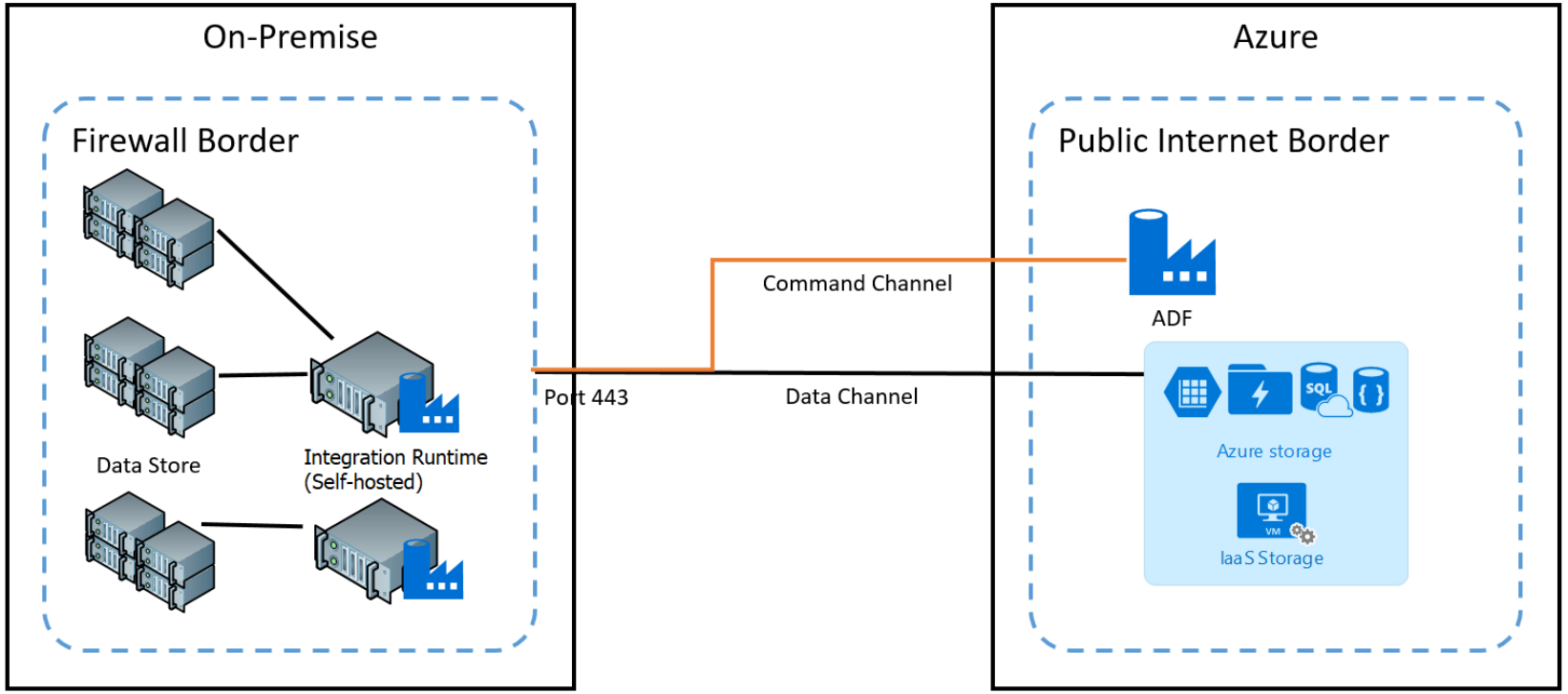
1. Modifica el pipeline para que se dispare la ejecución del mismo cuando alguno de los archivos de fuente se modifique.
    1. Caso 1: sobreescribir salida
    2. Caso2: rotar salida a nombre con fecha: yyyy_mm_dd
2. Añade otro trigger para que se ejecute 2 veces al día (8:00 y 20:00 horas).

07

Seguridad

Seguridad

- Datos
- Networking
- Integration Runtime
- Azure Key Vault
- Credenciales



Más en: https://learn.microsoft.com/es-es/azure/data-factory > security

Azure Key Vault

- Azure Key Vault es un servicio en la nube que se utiliza para guardar y acceder a secretos de forma segura.
- El secreto podría ser cualquier cosa que queramos proteger, como la clave API, las credenciales, etc.
- Proporciona cifrado de datos cuando se mueve de un almacén de claves a una aplicación cliente, lo que lo hace más seguro.

Más en: https://learn.microsoft.com/es-es/azure/data-factory/store-credentials-in-key-vault

ACTIVIDAD

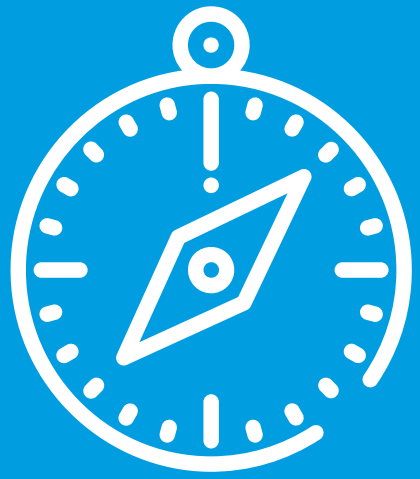
Demo Key Vault

OBJETIVO

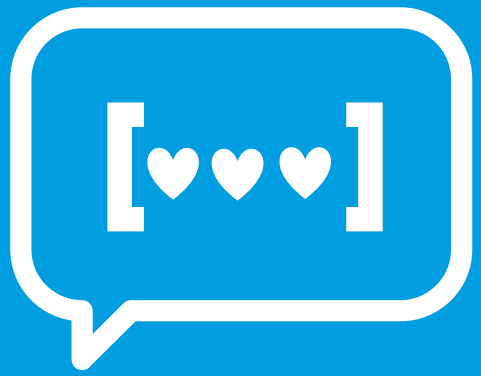
Securiza todos los datasets

INSTRUCCIONES

1. Modifica los datasets y actividades para securizarlos usando Azure Key Vault



Next steps



We would like to know your opinion!

Please, let us know what you think about the content.
From Netmind we want to say thank you, we appreciate time and effort you have taking in answering all of that is important in order to improve our training plans so that you will always be satisfied with having chosen us
quality@netmind.es

netmind
a bts company

Thanks!

Follow us: