

Procesos ETL

Tema 7: Diseño de flujos de trabajo a través de jobs

Adquisición y preparación de datos

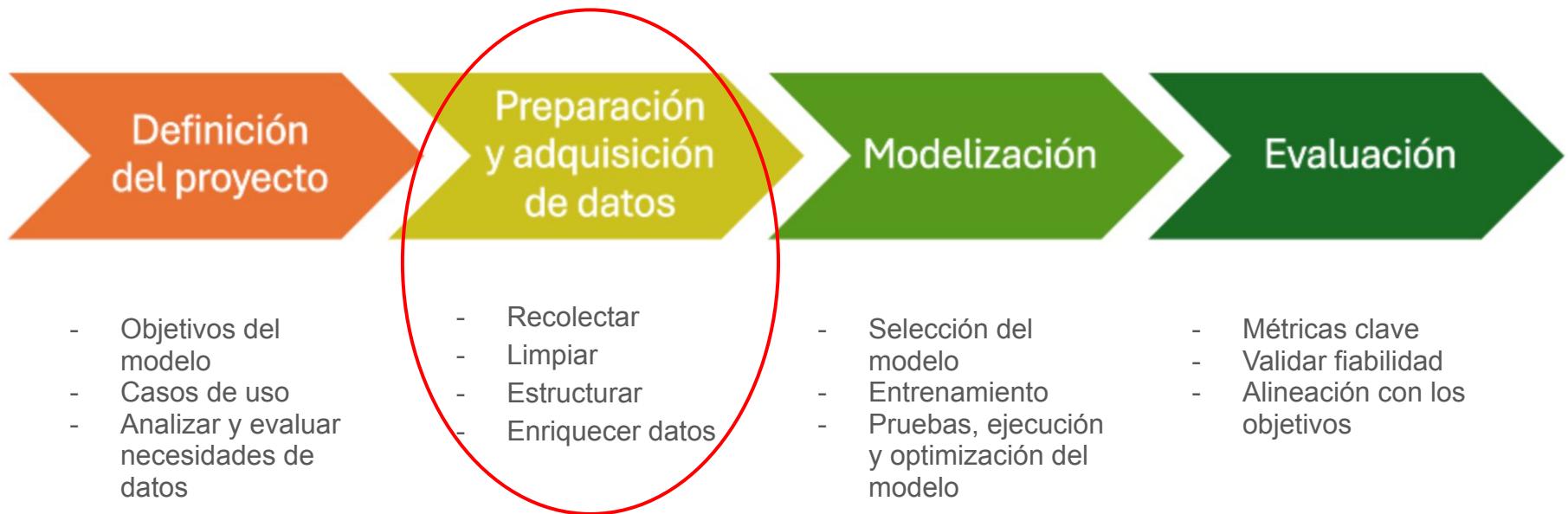


Universitat d'Alacant
Universidad de Alicante

Tema 7: Diseño de flujos de trabajo a través de jobs

- Diseño de jobs
- Variables y parámetros

Ciclo de vida del proyecto IA



Flujos de trabajo

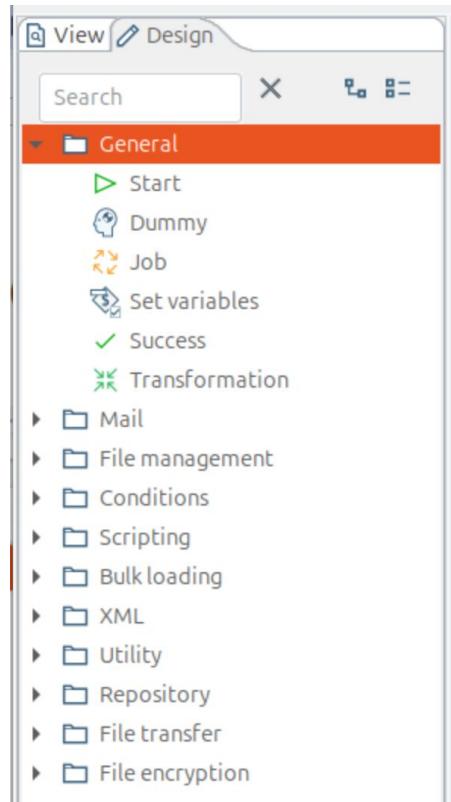
Existen diferentes herramientas que permiten trabajar con grandes volúmenes de datos y facilitan su análisis.

Para ello se diseñan y desarrollan flujos de trabajo que se pueden crear utilizando pasos o entradas unidas por saltos que pasan datos de un elemento al siguiente.

Los flujos de trabajo utilizan dos tipos de archivos:

1. Transformaciones que realizan tareas ETL.
2. Trabajos que organizan actividades de ETL, como la definición del flujo, dependencias y ejecución.

Pentaho Data Integration



Pasos para el diseño de jobs

General, pasos de inicio de trabajo, ejecución transformaciones o trabajos entre otras operaciones

Mail, envío de correos, recuperación y validación de cuentas

File management, gestión de operaciones sobre ficheros y carpetas, HTTP, etc

Conditions, comprobaciones sobre conexiones, BD, ficheros, etc

Scripting, scripts en JavaScript, Shell y SQL

Bulk loading, cargas a MySQL y MSSQL

XML, validaciones XML, DTD, XSD y XSL

Utility, pasos para ejecución de transformaciones como ping, enviar información a syslogs, Nagios, etc

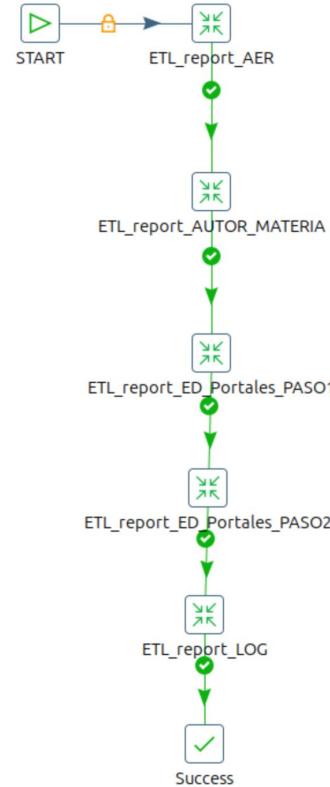
Repository, operaciones con el repositorio de transformaciones y trabajos.

File transfer, gestión ficheros FTP, FTPS y SFTP

File encryption : sistema PGP (Pretty Good Privacy) para envío y recepción de ficheros

Flujos de trabajo

Job



Flujos de trabajo

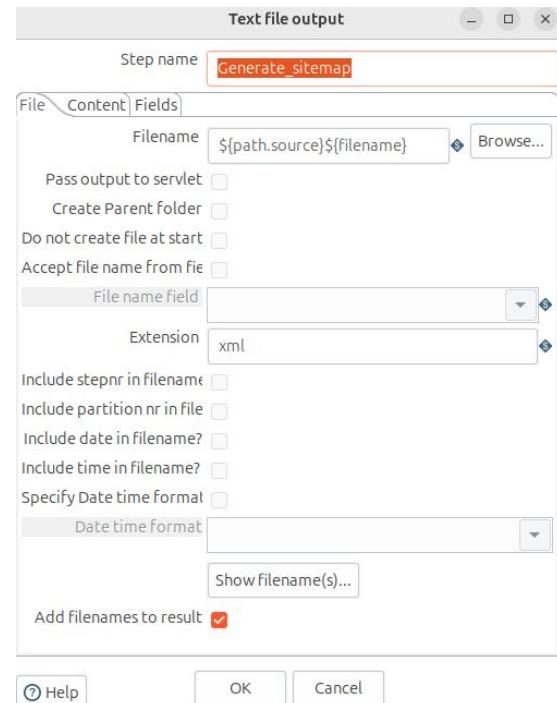
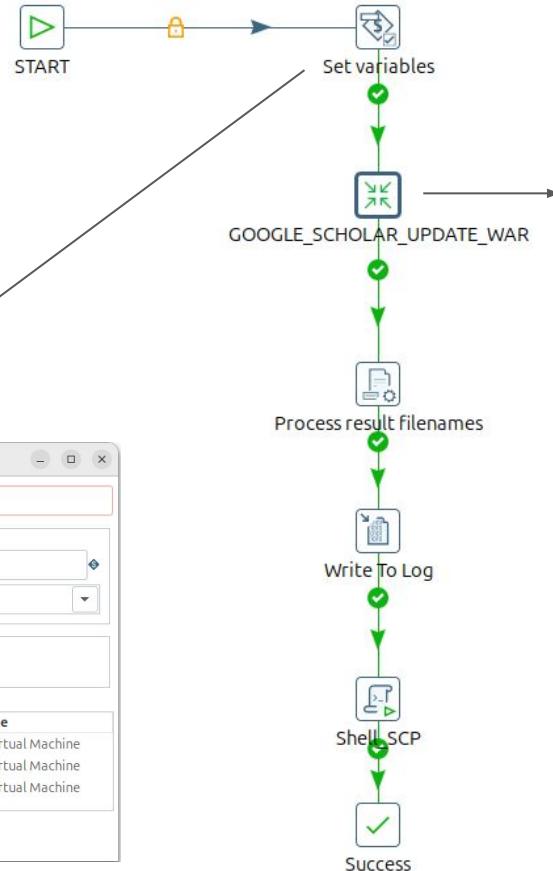
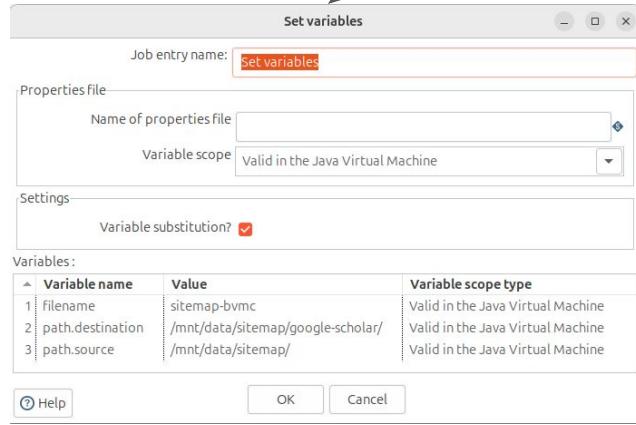
Job



Flujos de trabajo

Job

Establecer variables del flujo de trabajo, paso



Flujos de trabajo

Job

Process result filenames

Job entry name: **Process result filename:**

Action: Copy

Destination folder: \${path.destination}

Create destination folder

Replace existing file

Remove source filenames from result

Add destination filename to result

Add date to filename

Add time to filename

Specify Date time format

Date time format:

Add date before extension

Limit to

Limit action to

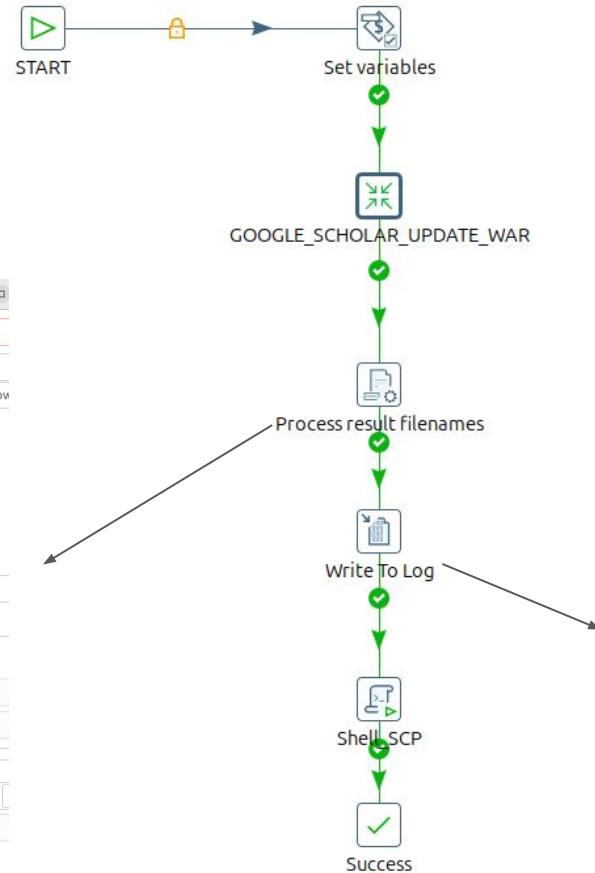
Wildcard (RegEx)

Exclude wildcard (RegEx)

Success On

Success condition: Success when all works fine

Nr errors lesser than: 10



Write to log

Job entry name: **Write To Log**

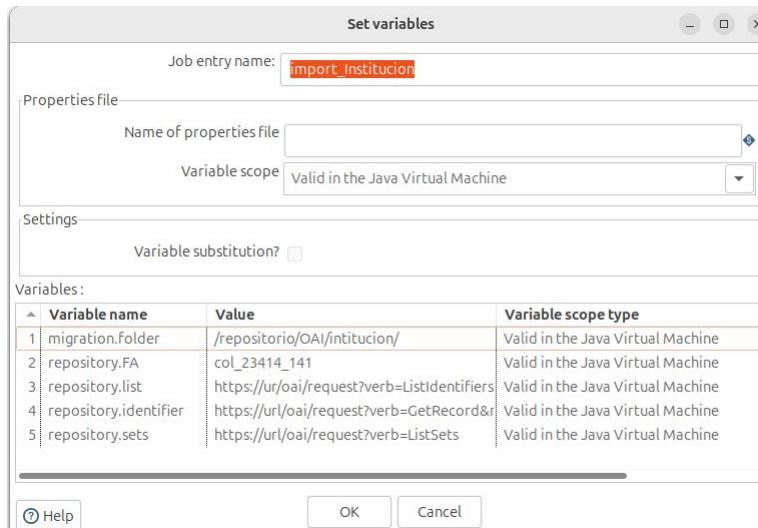
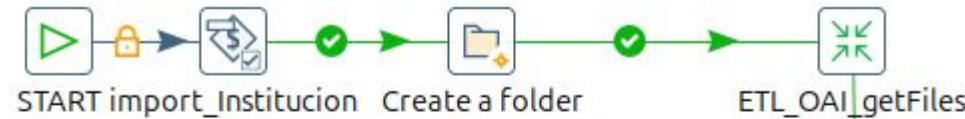
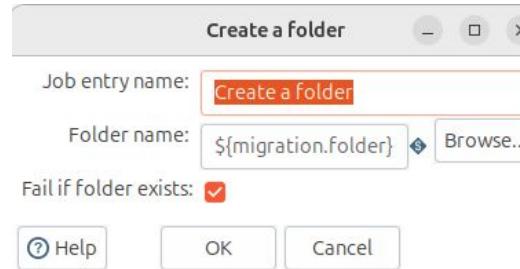
Log level: Basic

Log subject:

Log message:
.....sitemap is ok \${filename}
RESULT scp \${path.source}\${filename}.xml root

OK Cancel

Flujos de trabajo



Flujos de trabajo

Uso de variables en procesos ETL

- Las variables se configuran a nivel de sistema o dinámicamente en un trabajo
- Ámbito específico que permite ejecutar en paralelo el mismo trabajo o transformación con diferentes variables configuradas en el mismo sistema a nivel de JVM, de job actual o sesión
- Tipos de variables: sistema o definida por el usuario

Transformation properties

Transformation Parameters Logging Dates Dependencies Miscellaneous Monitoring

Parameters:

Parameter	Default Value	Description
1 FILE_IDS	ETL_inputs/set_authority.csv	csv file with manifestation identifiers
2 UPDATE_DATE	2025-05-23	update publication date

