

Procesos ETL

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

Adquisición y preparación de datos

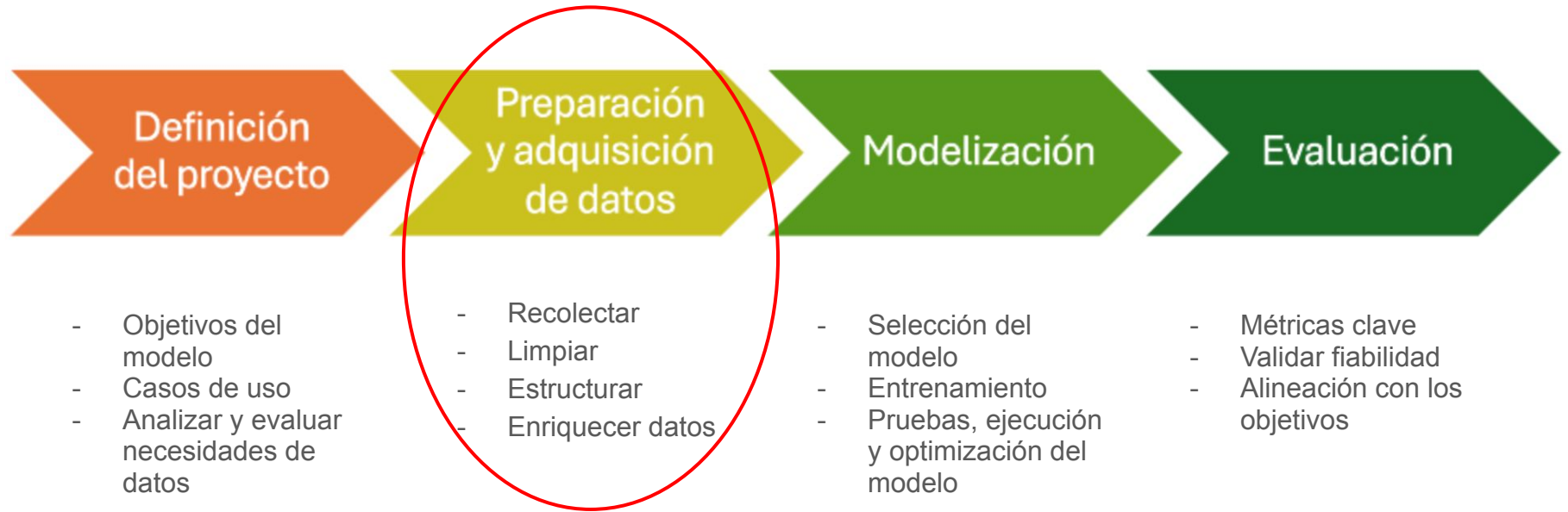


Universitat d'Alacant
Universidad de Alicante

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

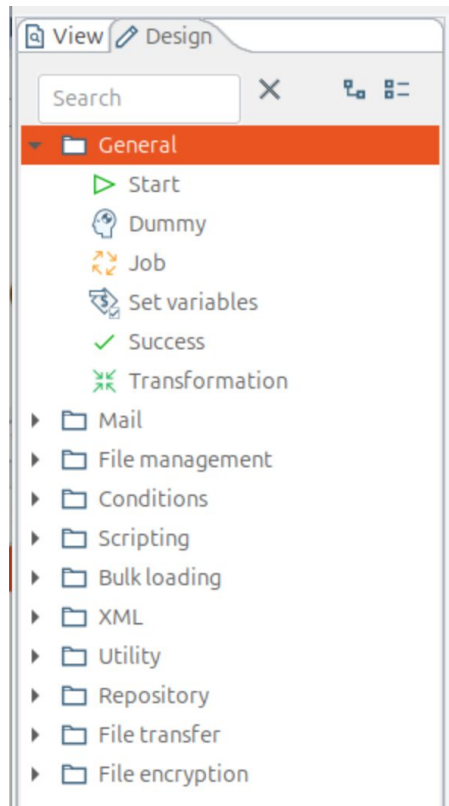
XML, validaciones XML, DTD, XSD y XSL

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

Repository, operaciones con el repositorio de transformaciones y trabajos.

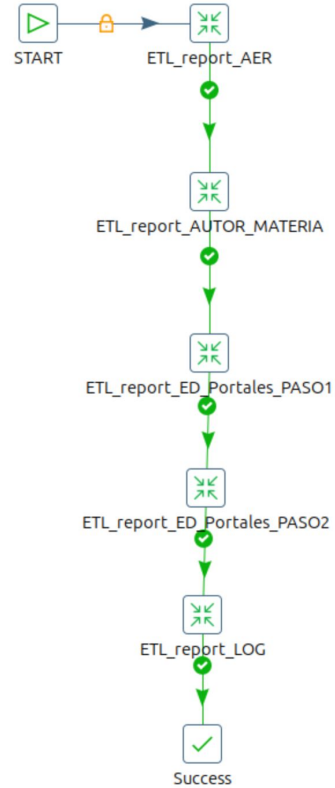
File transfer, gestión ficheros FTP, FTPS y SFTP

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



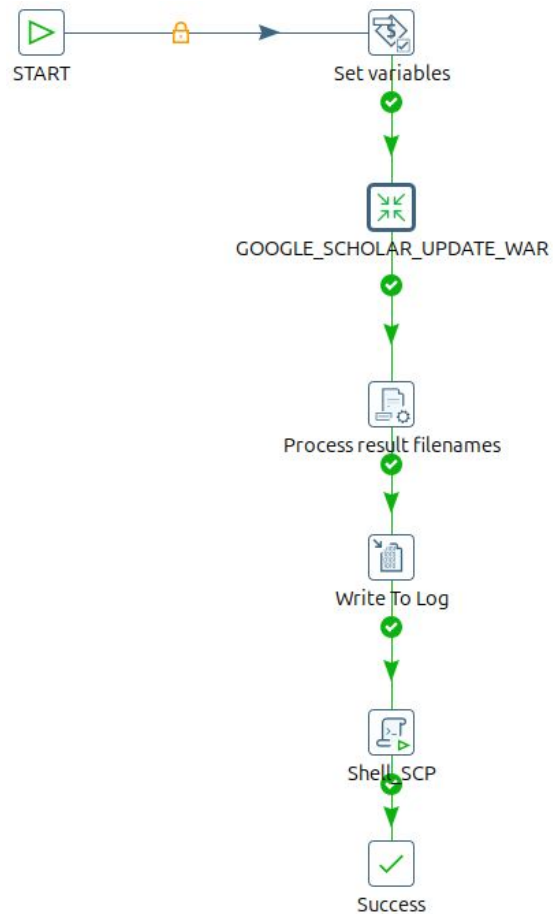
Flujos de trabajo

Job



Flujos de trabajo

Job



Flujos de trabajo

Job

Establecer variables del
flujo de trabajo, paso



Set variables

Job entry name:

Properties file

Name of properties file

Variable scope

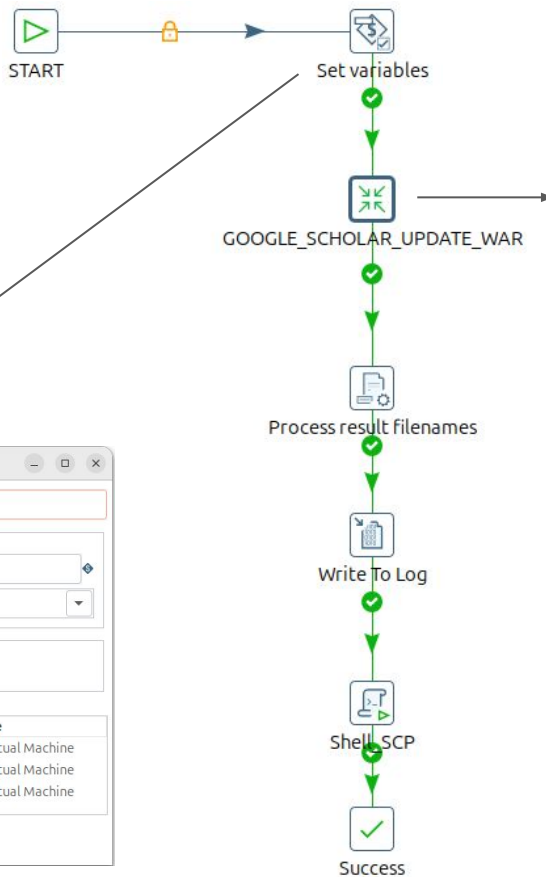
Settings

Variable substitution? ☒

Variables:

Variable name	Value	Variable scope type
1 filename	sitemap-bvmc	Valid in the Java Virtual Machine
2 path.destination	/mnt/data/sitemap/google-scholar/	Valid in the Java Virtual Machine
3 path.source	/mnt/data/sitemap/	Valid in the Java Virtual Machine

Help OK Cancel



Text file output

Step name

File Content Fields

Filename Browse...

Pass output to servlet ☐

Create Parent Folder ☐

Do not create file at start ☐

Accept file name from file ☐

File name field

Extension

Include stepnr in filename ☐

Include partition nr in file ☐

Include date in filename? ☐

Include time in filename? ☐

Specify Date time format ☐

Date time format

Show filename(s)...

Add filenames to result ☒

Help OK Cancel

Flujos de trabajo

Job

Process result filenames

Job entry name:

Action:

Destination folder:

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 ☐

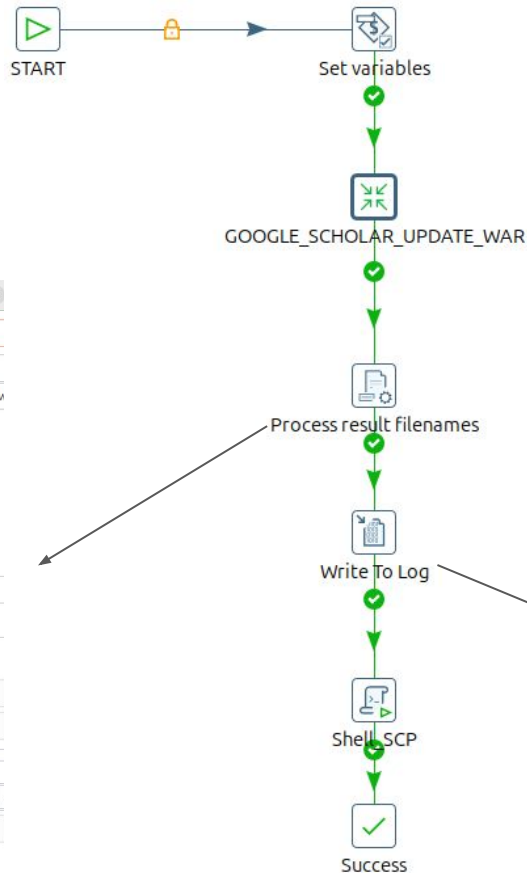
Wildcard (RegExp)

Exclude wildcard (RegExp)

Success On

Success condition

Nr errors lesser than



Write to log

Job entry name:

Log level:

Log subject:

Log message:

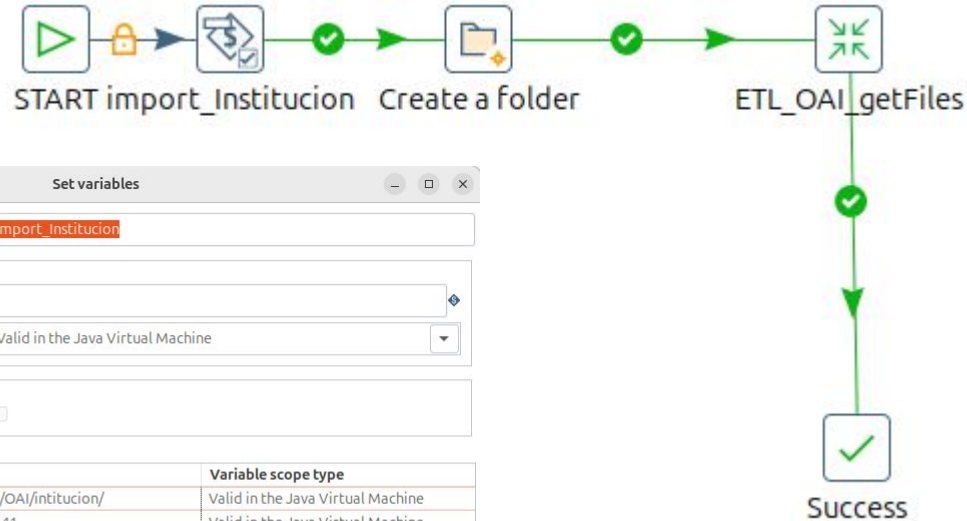
Flujos de trabajo

Create a folder

Job entry name:

Folder name:

Fail if folder exists: ☒



Set variables

Job entry name:

Properties file

Name of properties file

Variable scope:

Settings

Variable substitution? ☐

Variables:

Variable name	Value	Variable scope type
1 migration.folder	/repositorio/OAI/intitucion/	Valid in the Java Virtual Machine
2 repository.FA	col_23414_141	Valid in the Java Virtual Machine
3 repository.list	https://ur/oai/request?verb=ListIdentifiers	Valid in the Java Virtual Machine
4 repository.identifier	https://ur/oai/request?verb=GetRecord&r	Valid in the Java Virtual Machine
5 repository.sets	https://ur/oai/request?verb=ListSets	Valid in the Java Virtual Machine

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

