

L'informatique décisionnelle

Cours 2 : Les ETL

- **Les outils d'Extraction, Transformation et Load
(ETL)**
- **Pentaho Data Integration (PDI)**

Les outils ETL (Extract, Transform and Load)

✓ Un outil ETL:

- Extrait des données sources,
- Les transforme
- Les charge dans des données cibles.

✓ Exemple :

Extraire et collecter les données de plusieurs sources hétérogènes pour alimenter un entrepôt de données servant à analyser les données

Extraction

- Extraire des données hétérogènes provenant de diverses sources: SGBD (oracle, MySql, SqlServer...), fichiers (txt, Excel, XML, Csv...), NoSQL...
- Mise en place des primitives de connexion, déconnexions aux différentes sources
- Prise en compte des propriétés des données: domaine, type et spécificités (clé primaire / étrangère)....
- Extraction des données mises à jour ou insérées depuis la dernière phase d'extraction.

Transformation

- Retraitement des données extraites pour qu'elles soient utilisables dans un processus décisionnel
- Filtrage des données selon des critères (ex: produits dont le prix > 1000)
- Traitements et calculs : aggrégation(sum, count, avg, max...)
- Génération de clé primaire

Chargement

- Chargement de ces données dans des cibles différentes (généralement un entrepôt de données)
- Gestion de l'ajout et de la mise à jour des données
- Gestion des erreurs (incompatibilité des données....)

Autres fonctionnalités

- Planification des exécutions: lancer une des phases d'une manière automatique, périodique ou selon un critère donné
- Une interface d'administration
- Accès aux fonctionnalités moyennant des privilèges spécifiques aux utilisateurs
- Rapports d'erreurs, méthodes de reprise, affichage des statistiques relatives aux exécutions des traitements....

Cas d'utilisation

- La business intelligence
- La migration de données: passage d'une version d'une BD à une autre
- Changement de système: transfert des données d'un environnement vers un autre
- Synchronisation des données: les données sont gérées séparément par de multiples applications.

Exemples d'outils ETL open source

- Talend Open Studio for Data Integration
- Pentaho Data Integration (Kettle)
- GeoKettle
- Coudera
- Birt

Talend Open Studio for Data Integration



- Talend est un éditeur de solutions de gestion de données
- Cet ETL est développé sous Java
- Elle propose une interface de modélisation graphique basée sur l'environnement IDE Eclipse.
- Sa gestion de la performance permet de manier du Big Data avec l'approche ELT.
- L'éditeur est connu pour sa bibliothèque collaborative de plus de 900 composants et connecteurs aux sources de données.
- Ses outils sont aujourd'hui leaders du marché.

Pentaho Data Integration

Pentaho est une plate-forme décisionnelle open source complète possédant une couverture globale des fonctionnalités de la Business Intelligence :

- ETL (intégration de données)
- Reporting
- Tableaux de bord ("*Dashboards*")
- Analyse *ad hoc* (requêtes à la demande)
- Analyse multidimensionnelle (OLAP)

- Projet étudiant québécois crée entre 2006 et 2009,
- Développé à partir de **Kettle** (Pentaho Data Integration).
- En 2009, la compagnie Spatialytics a été créée pour poursuivre le développement des projets GeoKettle et GeoMondrian
- Une version de Pentaho Data Integration spécialisée dans **le traitement des données géospatiales**.
- Il s'installe sur Windows, Mac OS, Linux et Solaris et peut extraire des données issues de plus de 35 database (Oracle, MySQL, etc..)
- GeoKettle constitue un ETL complet et entièrement gratuit rivalisant avec produits propriétaire.

Cloudera

- Ce fournisseur développe CDH, une distribution d'Hadoop, comme Apache.
- C'est un pure player soutenu par Intel.
- Il propose des fonctions de sécurité et d'intégration, et délivre des formations et certifications aux développeurs, administrateurs et analystes.

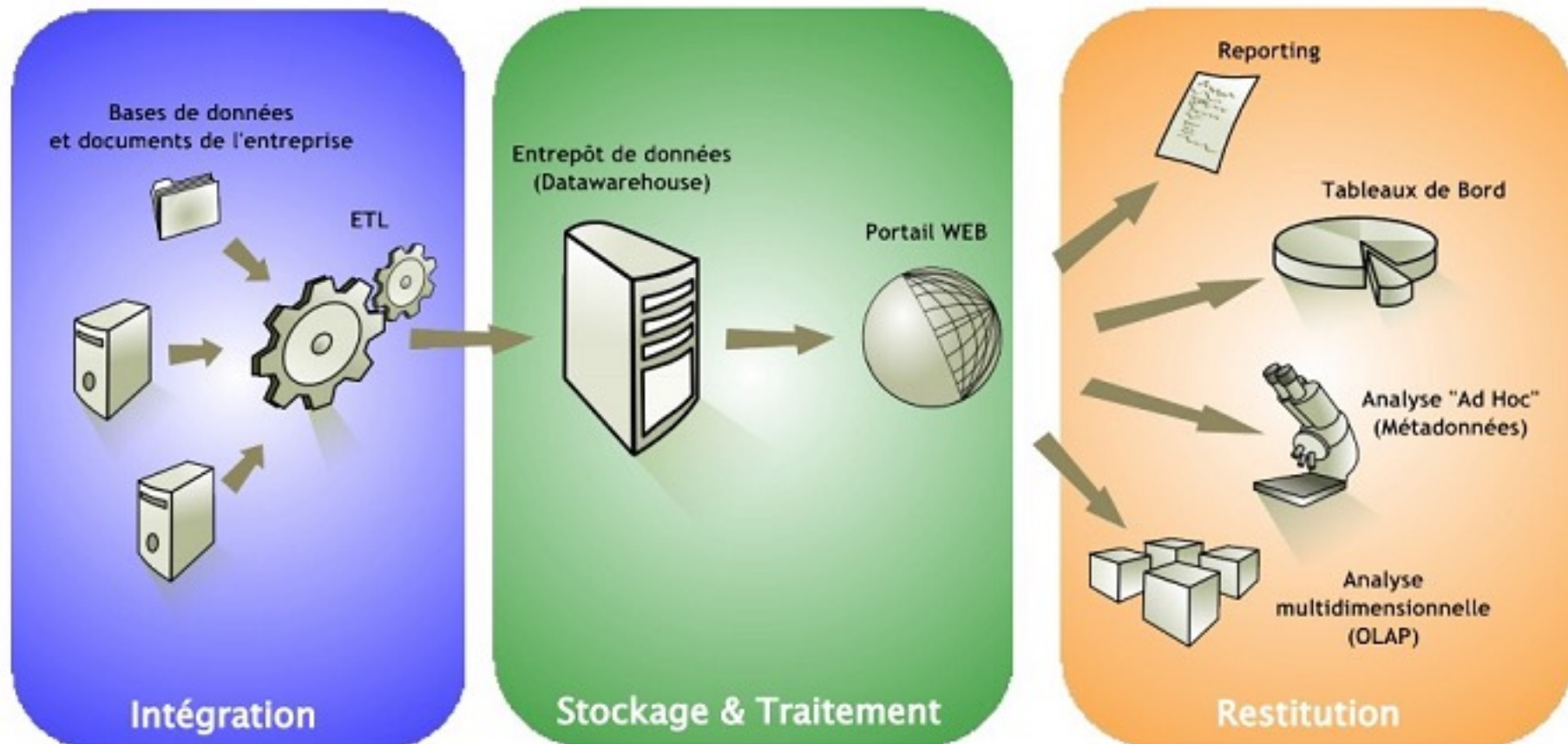
BIRT

- **BIRT** signifie **Business Intelligence and Reporting Tools**.
- Un projet d'Actuate élaboré dans un environnement Java / J2EE.
- Centré sur les reportings,
- Propose une plateforme client avec visualisation des données automatisée.

PENTHO : les composants

Fonctionnalité	Module Pentaho
Extraction, Transformation, Load (ETL)	Pentaho Data Integrator (anciennement Kettle)
Reporting Standard	Pentaho Reporting (Jfree Report), Jasper Report, BIRT (Business Intelligence Reporting Tools)
Reporting Ad'hoc	Pentaho BI's Metadata, Pentaho Reporting (Jfree)
Analyse OLAP	Pentaho Analysis (Mondrian + Jpivot)
Tableau de bord	Pentaho Dashboard
Data Mining	Weka

PENTHO : Open source de Business Intelligence



Pentaho Data Integration (PDI)

- Créé en 2001 par Matt Casters pour ses besoins personnels
- Une solution d'informatique décisionnelle open source entièrement développée en Java
- Fournit une interface graphique pour la manipulation des données

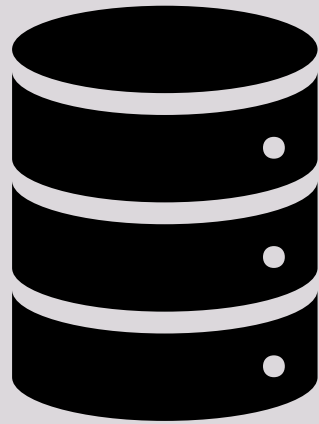
Fonctionnalités

PDI est un environnement qui permet :

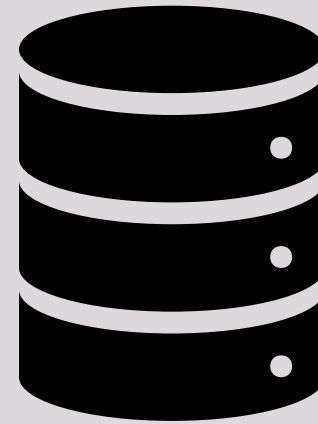
- la connexion à plusieurs bases de données
- la définition des transformations sur les données
- L'exécution de ces transformations
- La sauvegarde des transformations dans des fichiers ou dans un référentiel bases de données
- Permet de faire du:
 - Reporting simple,
 - OLAP (OnLine Analytical Processing)
 - Data mining Report

Exemples Dans PDI

Objectif : créer une table Enseignant à partir des deux sources de données



Source 1 : base de données Oracle



Source 2 : fichier xls

Source 1 : base de données Oracle

Lignes de l'étape: Extraction depuis table (4 lignes)

▲ #	ID_ENS	NOM	PRENOM	TELEPHONE
1	1	Dubois	Jean	012345678
2	2	LEGRAND	EMILIE	012345678
3	10	Nom10	Prenom10	012345678
4	11	Nom10	Prenom11	012345678

Source 2 : fichier xls

	A	B	C	D	E	F	G	H	I
	ID	Nom	Prenom	Statut	Provenance	FormationPrecedente	NiveauInsertion	ID_Cours	Libelle_Cours
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	5	ID
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	5	ID
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	6	ID
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	6	ID
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	2	SGBDA
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	2	SGBDA
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	4	SGBDA
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	4	SGBDA
0	3	DUBAIN	HELENE	etudiant	France	Bac+4	M1	1	SGBD
L	4	HANZ	SOPHIE	etudiant	Italie	Bac+3	M1	1	SGBD
2	3	DUBAIN	HELENE	etudiant	France	Bac+4	M1	3	SGBD
3	4	HANZ	SOPHIE	etudiant	Italie	Bac+3	M1	3	SGBD
4	1	Dubois	Jean	enseignant				2	SGBDA
5	1	Dubois	Jean	enseignant				4	SGBDA
5	2	LEGRAND	EMILIE	enseignant				5	ID
7	2	LEGRAND	EMILIE	enseignant				6	ID
3	3	MARTIN	Eric	enseignant				1	SGBD
9	3	MARTIN	Eric	enseignant				3	SGBD
0									
L									
2									
3									

Extraction de la source 1 : extraction depuis table



(41) OBM - Webmail - Mozilla Firefox

Database Connection

General
Advanced
Options
Pooling
Clustering

Connection Name:
connexion 1

Connection Type:
MS Access
MS SQL Server
MS SQL Server (Native)
MaxDB (SAP DB)
MonetDB
MySQL
Native Mondrian
Neoview
Netezza
OpenERP Server
Oracle
Oracle RDB
Palo MOLAP Server
PostgreSQL

Access:
Native (JDBC)
ODBC
OCI
JNDI

Settings
Host Name:
172.19.1.71
Database Name:
MIAGE
Tablespace for Data
Tablespace for Indices
Port Number:
1521
User Name:
marukoz
Password:

Tester Spécifica Explorer

OK Cancel

Explorer la base de données

Actions

- DEALERS
- DEP
- DEPARTEMENT
- DEPARTEMENTO
- ENSEIGNANT
- ENSEIGNANTS**
- ENSEIGNE
- ENSEIGNEMENT
- ETUDIANT
- ETUDIANTS
- INSCRIPTION

Valider

Annuler

Extraction Table

Extraction depuis table

Connexion connexion 1

Editer...

Nouvelle...

Assistant...

Obtenir script SQL select ...

Données prévisualisées

Lignes de l'étape: Extraction depuis table (4 lignes)

#	ID_ENS	NOM	PRENOM	TELEPHONE
1	1	Dubois	Jean	012345678
2	2	LEGRAND	EMILIE	012345678
3	10	Nom10	Prenom10	012345678
4	11	Nom10	Prenom11	012345678

String(45)

Fermer

Afficher Trace

Remplacer les variables dans le script SQL

Insérer données à partir de

Exécuter pour chaque ligne

Limite

0

Help

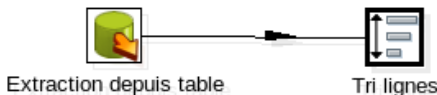
OK

Prévisualiser

Annuler

L'étape Tri lignes : trier les lignes selon un ou plusieurs champs

Note : certaines étapes nécessitent une étape de Tri lignes au préalable



Extraction depuis table → Tri lignes

Tri lignes

Nom étape: Tri lignes

Répertoire de tri: %%java.io.tmpdir%% Parcourir...

Préfixe fichiers TMP: out

Nbr lignes en mémoire: 1000000

Conserver min mémoire libre (%JVM):

Compresser fichiers: ☐

Transmettre valeurs distinctes: ☐

Champs :

#	Nom champ	Ascendant	Respecter la casse	Presorted?
1	ID_ENS	O		
2	NOM	O		
3	PRENOM	O		
4	TELEPHONE	O		

Help OK Annuler Récupérer Champs

Extraction source 2

- Récupérer les champs de toutes les feuilles du fichier
- Récupérer les lignes dont le statuts = enseignants
- Enlever les doublants
- Retirer les champs qui ne sont pas nécessaires

	A	B	C	D	E	F	G	H	I
	ID	Nom	Prenom	Statut	Provenance	FormationPrecedente	NiveauInsertion	ID_Cours	Libelle_Cours
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	5	ID
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	5	ID
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	6	ID
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	6	ID
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	2	SGBDA
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	2	SGBDA
	1	HOMMS	CECILE	etudiant	France	Ingenieur	M1	4	SGBDA
	2	MURIEL	RICHARD	etudiant	Allemagne	Bac+4	M2	4	SGBDA
)	3	DUBAIN	HELENE	etudiant	France	Bac+4	M1	1	SGBD
L	4	HANZ	SOPHIE	etudiant	Italie	Bac+3	M1	1	SGBD
2	3	DUBAIN	HELENE	etudiant	France	Bac+4	M1	3	SGBD
3	4	HANZ	SOPHIE	etudiant	Italie	Bac+3	M1	3	SGBD
4	1	Dubois	Jean	enseignant				2	SGBDA
5	1	Dubois	Jean	enseignant				4	SGBDA
5	2	LEGRAND	EMILIE	enseignant				5	ID
7	2	LEGRAND	EMILIE	enseignant				6	ID
3	3	MARTIN	Eric	enseignant				1	SGBD
3	3	MARTIN	Eric	enseignant				3	SGBD
)									



Input

Fichier Excel en entrée

Nom étape:

[Ajouter champ\(s\)](#)

Fichiers | Feuilles | Contenu | Gestion erreurs | Champs | Champs additionnels

Type de tableau:

Fichier ou répertoire: [Ajouter](#) [Parcourir...](#)

Caractères joker:

Exclure caractères joker:

Fichiers sélectionnés:

#	Fichier/Répertoire	Caractères joker (RegExp)	Exclure caractères joker	Exigé	Parcourir sous-répertoire
1					

[Supprimer](#) [Editer](#)

Accepter noms de fichier crée dans les étapes précédentes

Accepter noms de fichier crée dans les étapes ☐

Étape origine des noms de fichiers:

Champ contenant le nom de fichier:

[Afficher la ou les fichier\(s\)...](#)

[OK](#) [Prévisualiser lignes](#) [Annuler](#)

Récupérer tous les champs du fichier xls

Extraction

Fichier Excel en entrée

Nom étape
Ajouter feuille(s) Extraction depuis fichier MS Excel

Fichiers Feuilles Contenu Gestion erreurs **Champs** Champs additionnels

#	Nom	Type	Longueur	Précision	Trim type	Répéter	Format	Devise	Décimal	Groupement
1	ID	Number			none	N				
2	Nom	String			none	N				
3	Prenom	String			none	N				
4	Statut	String			none	N				
5	Provenance	String			none	N				
6	FormationPrecedente	String			none	N				
7	NiveauInsertion	String			none	N				
8	ID_Cours	Number			none	N				
9	Libelle_Cours	String			none	N				
10	Type_Cours	String			none	N				
11	Niveau_Cours	String			none	N				
12	Note	Number			none	N				

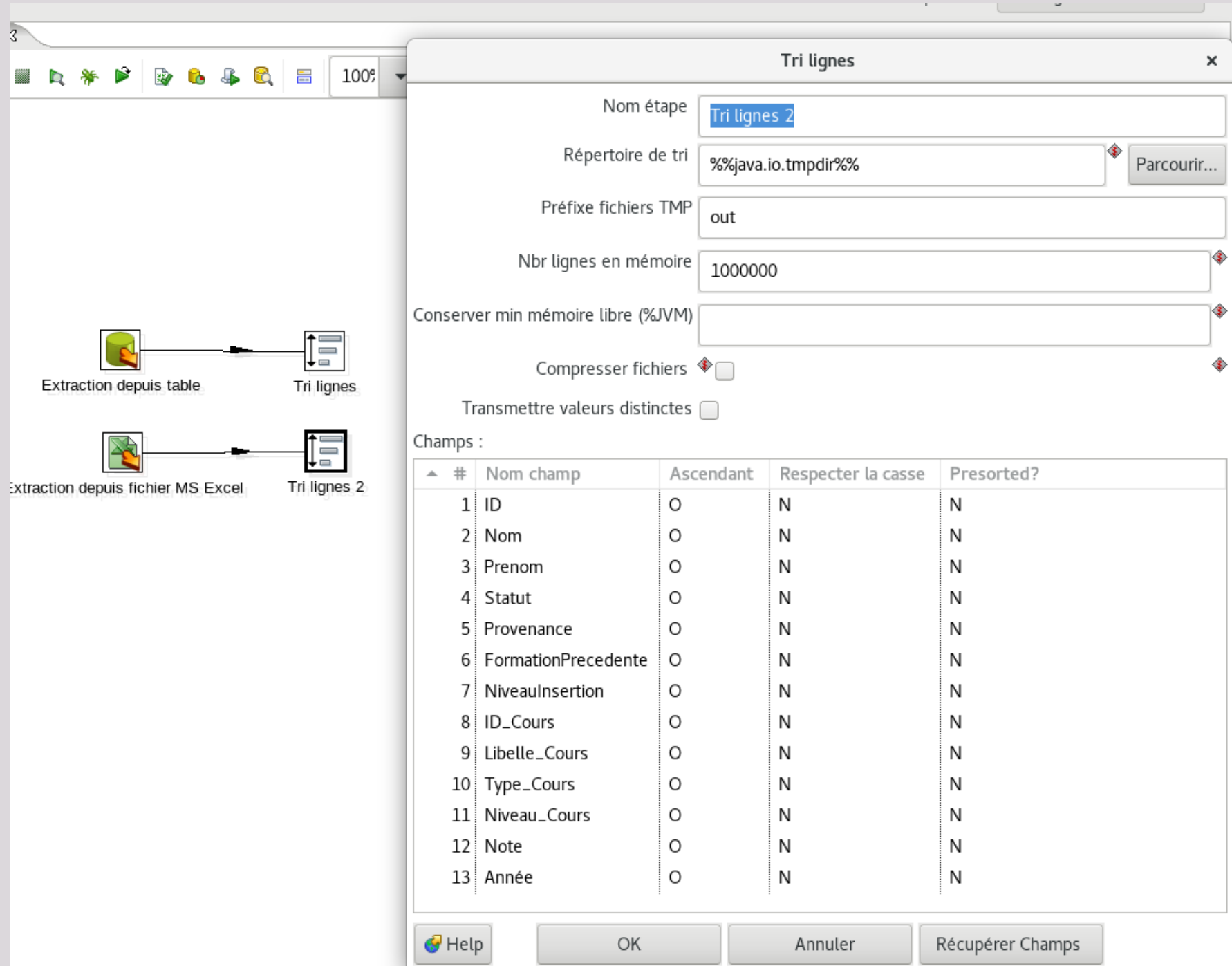
Récupérer les champs depuis la ligne d'en tête...

OK Prévisualiser lignes Annuler

Help

Generation de numeros de carte

- Rajouter une étape de tri pour la sortie de l'extraction du fichier xls



The screenshot shows a data extraction tool interface. On the left, a workflow diagram illustrates two extraction paths: 'Extraction depuis table' and 'Extraction depuis fichier MS Excel', both leading to a 'Tri lignes' (Sort lines) step. The 'Tri lignes' step is highlighted with a red box.

On the right, the 'Tri lignes' dialog box is open, showing the configuration for the sorting step. The dialog box has a title bar 'Tri lignes' and a close button 'X'. The configuration fields are as follows:

- Nom étape: Tri lignes 2
- Répertoire de tri: %%java.io.tmpdir%% (with a 'Parcourir...' button)
- Préfixe fichiers TMP: out
- Nbr lignes en mémoire: 1000000
- Conserver min mémoire libre (%JVM):
- Compresser fichiers: ☐
- Transmettre valeurs distinctes: ☐

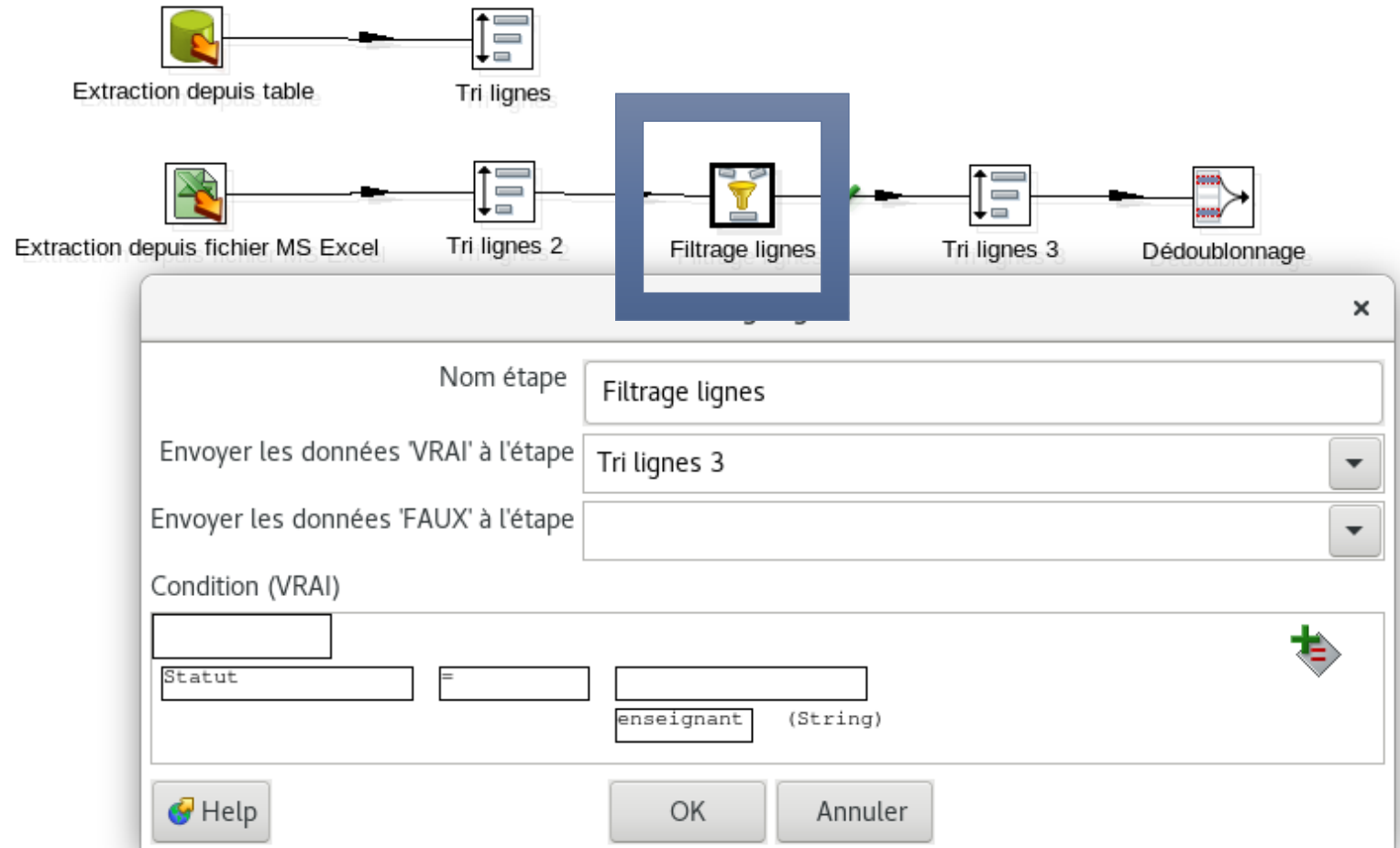
Below the configuration fields, there is a section titled 'Champs :' (Fields) containing a table with 5 columns: '#', 'Nom champ', 'Ascendant', 'Respecter la casse', and 'Presorted?'. The table lists 13 fields:

#	Nom champ	Ascendant	Respecter la casse	Presorted?
1	ID	O	N	N
2	Nom	O	N	N
3	Prenom	O	N	N
4	Statut	O	N	N
5	Provenance	O	N	N
6	FormationPrecedente	O	N	N
7	NiveauInsertion	O	N	N
8	ID_Cours	O	N	N
9	Libelle_Cours	O	N	N
10	Type_Cours	O	N	N
11	Niveau_Cours	O	N	N
12	Note	O	N	N
13	Année	O	N	N

At the bottom of the dialog box, there are four buttons: 'Help', 'OK', 'Annuler', and 'Récupérer Champs'.

- **Filtrages lignes :**
filtre les lignes en
entrée selon une
ou plusieurs
conditions

*Note : ici, on
s'intéresse aux
enseignants*



- **Dédoublonnage:** enlever les doublants selon un ou plusieurs champs

Note : plusieurs lignes détiennent le même nom, prénom et ID

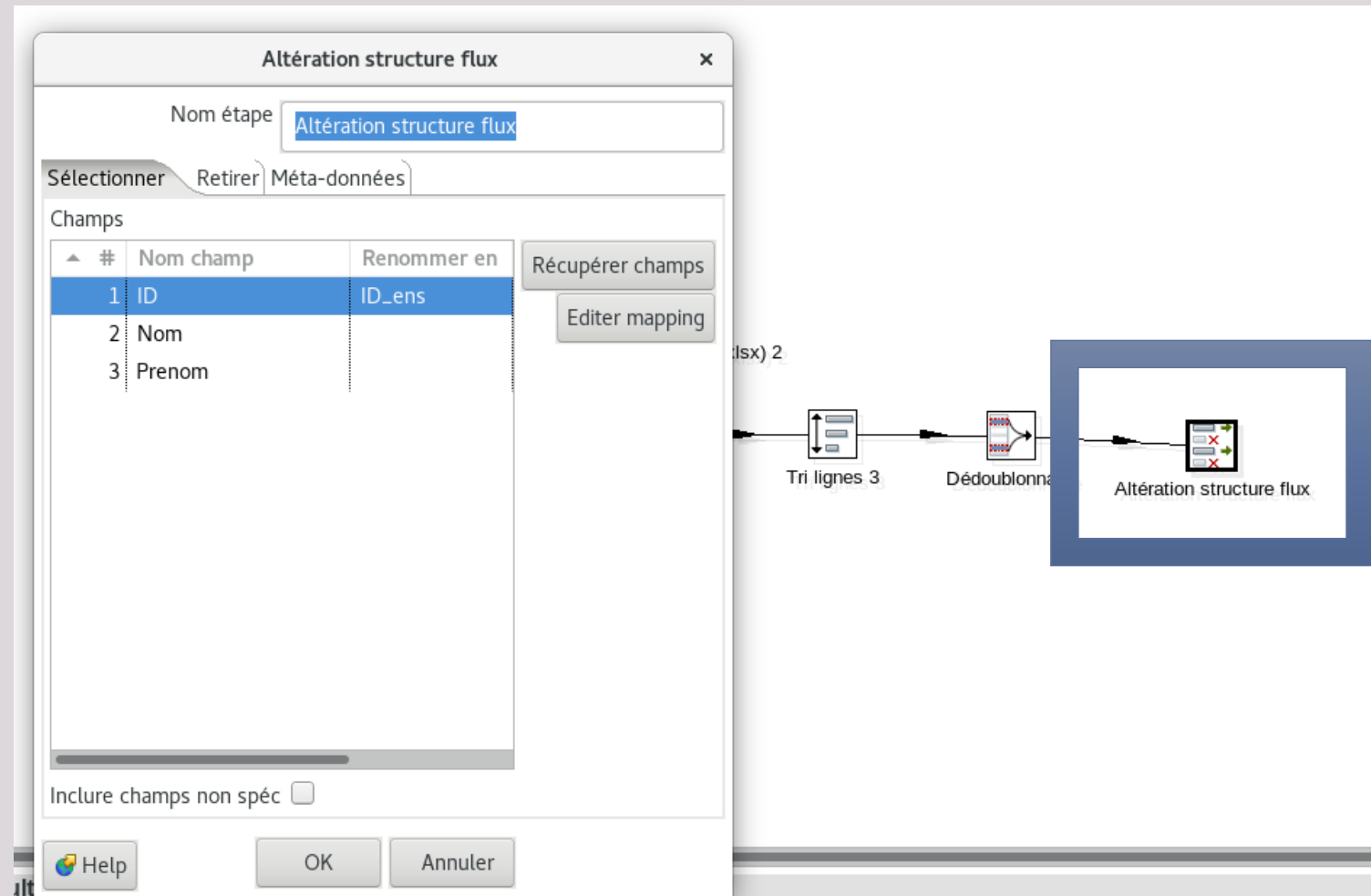
The screenshot shows a data workflow with two parallel paths. The top path consists of 'Extraction depuis table' followed by 'Tri lignes'. The bottom path consists of 'Extraction depuis fichier MS Excel', 'Tri lignes 2', 'Filtrage lignes' (marked with a green check), and 'Tri lignes 3'. Both paths converge into a 'Dédoublonnage' (Deduplication) step, which is highlighted with a blue square. Below the workflow, the 'Dédoublonnage' configuration window is open. It has a title bar 'Dédoublonnage' and a close button. The 'Nom étape' field contains 'Dédoublonnage'. Under 'Paramètres', there are two options: 'Ajouter un compteur à la sortie' with an unchecked checkbox and a 'Champ compteur' text box, and 'Rejeter lignes doublons' with an unchecked checkbox and a 'Description erreur' text box. Below this is a section 'Champs de comparaison (aucun champ saisi : comparaison de toute la ligne)' containing a table:

▲ #	Nom champ	Ignorer la casse
1	ID	N
2	Nom	N
3	Prenom	N

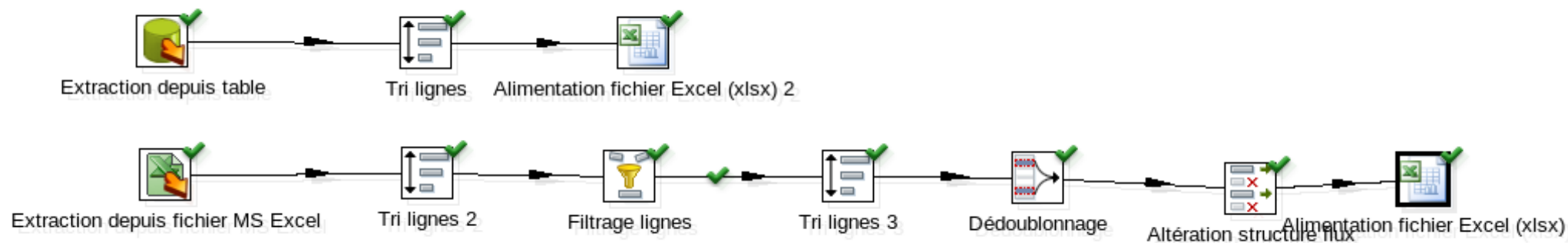
At the bottom of the window are buttons for 'Help', 'OK', 'Annuler', and 'Récupérer champs'.

- **Altération structure flux :** modifier (ajouter, supprimer, renommer) les champs des données en entrée

Note : plusieurs champs ne sont pas nécessaires à l'entité Enseignant (comme provenance, niveau insertion , etc.)



Vérification des résultats : une idée est de charger les résultats dans un fichier excel (alimentation fichier Excel) pour visionner les résultats des transformations



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID_ens	Nom	PreNom										
2	1	Dubois	Jean										
3	2	LEGRAND	EMILIE										
4	3	MARTIN	Eric										
5													
6													
7													

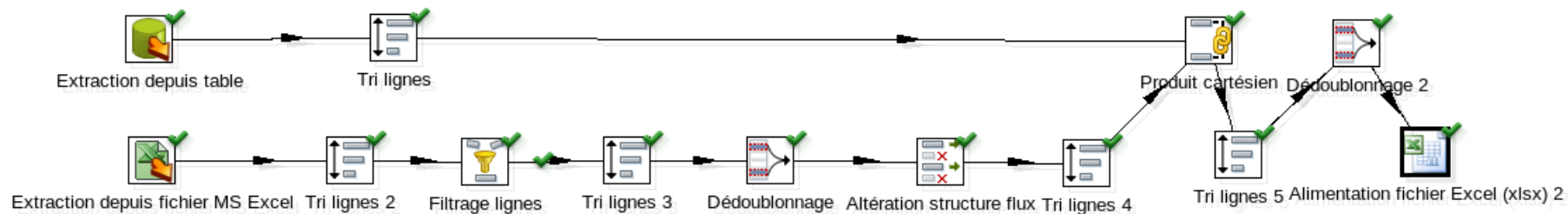
Joindre les données récoltées et transformées des deux sources par un **produit cartésien** sur le Nom et le Prénom

The image shows a data integration workflow with two parallel paths. The top path starts with 'Extraction depuis table', followed by 'Tri lignes', and then a 'Produit cartésien' connector (highlighted with a blue box). The bottom path starts with 'Extraction depuis fichier MS Excel', followed by 'Tri lignes 2', 'Filtrage lignes', 'Tri lignes 3', 'Dédoublonnage', 'Altération structure flux', 'Tri lignes 4', and 'Tri lignes 5'. Both paths converge into 'Alimentation fichier Excel (xlsx) 2'. A 'Produit cartésien Dédoublonnage 2' connector is also shown. A dialog box titled 'Produit cartésien' is open, showing configuration options: 'Nom étape' (Produit cartésien), 'Répertoire temp' (%%java.io.tmpdir%%), 'Préfixe fichier TMP' (out), and 'Mémoire cache Max.(en lignes)' (500). The 'Etape source' dropdown is empty. The 'Condition' section is highlighted with a blue box and contains the following logic:

```
Condition:  
  
NOM = NOM  
  
AND  
  
PRENOM = PRENOM
```

At the bottom left, a 'stats exécution' table is partially visible:

Nom étape	Lignes maj	Lignes rejetées	Lignes en erreur	S
Extraction depuis tab	0	0	0	T
Extraction depuis fich	0	0	0	T
Tri lignes 2	0	0	0	T



- Faire un dernier tri
- Nettoyer les doublants
- Charger dans un fichier Excel

bddxls.xls

Fichier Édition Affichage Insertion Format Feuille Données

Arial 10

A1 Σ = ID_ENS

	A	B	C	D	E	F
1	ID ENS	NOM	PRENOM	TELEPHONE		
2	1	Dubois	Jean	012345678		
3	2	LEGRAND	EMILIE	012345678		
4	10	Nom10	Prenom10	012345678		
5	11	Nom10	Prenom11	012345678		
6						
7						
8						
9						

Chargement dans une table

- **Exécution Script** : permet d'écrire un script de création de tables dans une base de données
- **Insertion dans Table** : permet d'insérer les résultats d'une transformation dans une table d'une base de données