

# Intégration de données :

## Mise en pratique sur Talend Open Studio

Mourad Ouziri

[Mourad.Ouziri@ParisDescartes.fr](mailto:Mourad.Ouziri@ParisDescartes.fr)

Maître de conférences à l'Université Paris Descartes  
Formateur indépendant

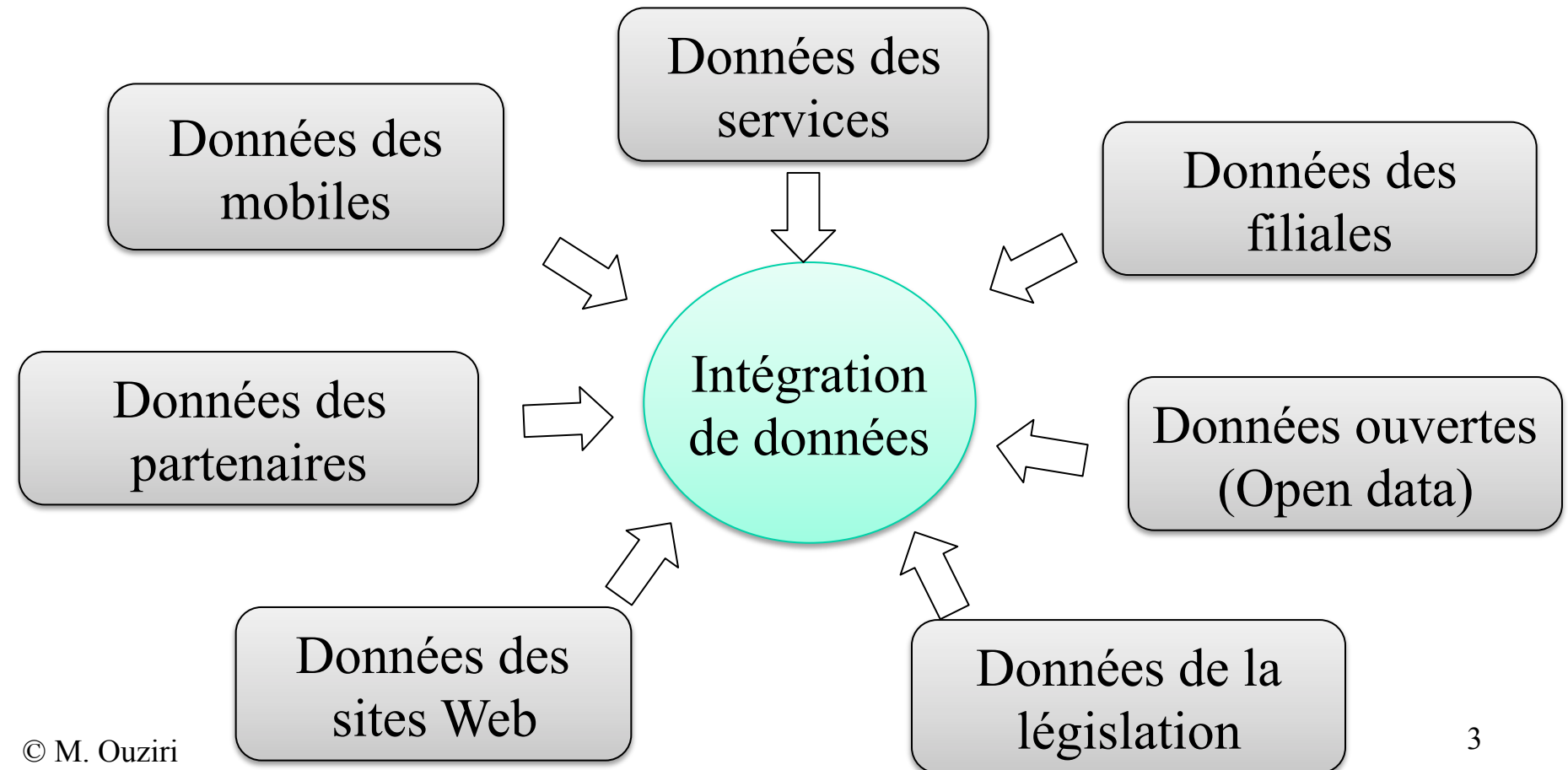
# Programme

1. Intégration de données : qu'est ce que c'est ? pourquoi ?
2. Architectures d'intégration de données
3. Mises en pratique sur Talend Open Studio en modes :
  1. ETL : Extract-Transform-Load (1 jour)
  2. ELT sur un SGBD et la programmation de procédures stockées (1/2 jour)
  3. ELT sur Hadoop : mise en œuvre sur HDFS, Hive, Pig et HBase (2,5 jour)

# Intégration de données

## Définition

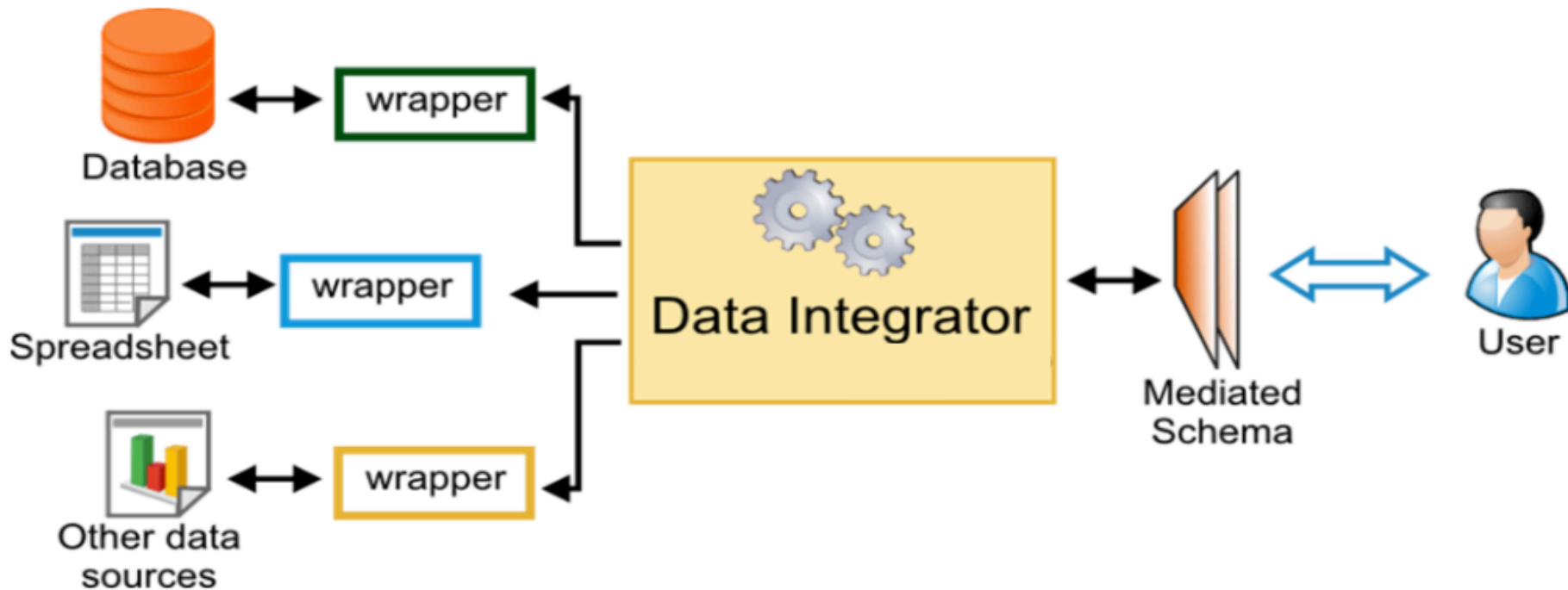
➡ Processus de fusion, de rapprochement, de croisement et de consolidation de données issues de multiples sources de données



# Intégration de données

## Objectif

- 👉 Objectif technique : interroger plusieurs sources de données hétérogènes via une interface unique



# Intégration de données

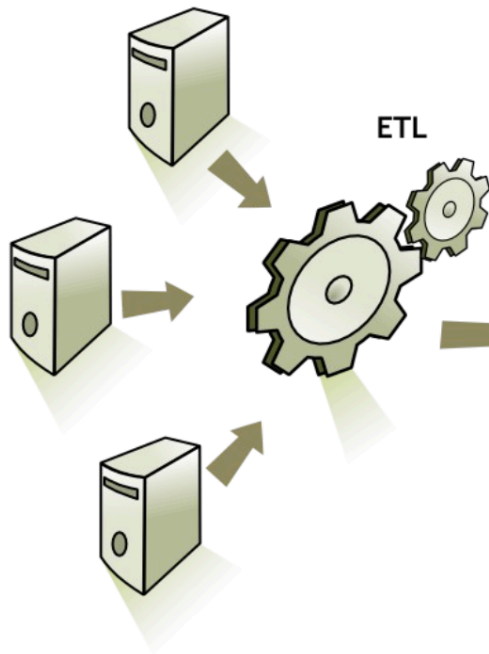
## Objectif qualité

- ☞ Objectif métier : collecter des données riches/complètes et fiables !
- ☞ Comment ? Exploiter tous les canaux (sources de données) possibles
  - Données internes à l'entreprise : bases commerciales, sites Web, centres d'appels
  - Données externes : données des partenaires, des collaborateurs, de l'*Open Data*
- ☞ Caractéristiques des données
  - Proviennent de multiples sources hétérogènes
  - Duplication des entités du monde réel (selon différentes facettes) sous différents identifiants
  - Décrites avec différentes terminologies : Client/Customer, France/FR
  - Conflits/divergence de données : possibles violations de règles de gestion

# Intégration de données

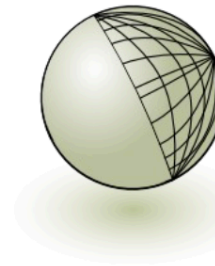
Un étape importante de la chaine décisionnelle

Bases de production



Entrepôt de données  
(datawarehouse, datamarts)

Portail



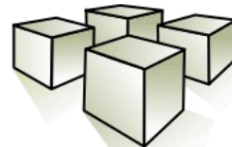
Reporting



Datamining



OLAP



Tableaux de Bord



# Intégration de données

## Meilleure prise de décision

☞ En assurance habitation : meilleure connaissance du client (et de ses biens) pour une tarification plus juste, personnalisation des services, etc.

*Avec peu d'informations*

Adresse : Paris

Nb pièces : 5



L1



L2



Adresse : Paris

Nb pièces : 5

**cotisation L1 = cotisation L2**

*Avec plus d'informations*

Adresse : Paris

Nb pièces : 5

AnnéeConst : 1890

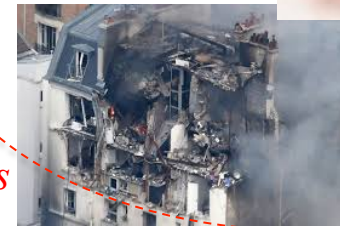
Nb enfants : 3

Risque camb : 0.3

L2



**cotisation L2 >> cotisation L1**



L1



Adresse : Paris

Nb pièces : 5

AnnéeConst : 1985

Nb enfants : 0

Risque camb : 0.01

# Intégration de données

## Meilleure prise de décision

### 👉 Pour la santé :

- Un meilleur diagnostic médical : par l'obtention des antécédents médicaux du patient, des informations sociétales, environnementales, etc.
- Préserver la santé du patient : ne pas ré-administrer les mêmes médicaments (à effets secondaires), ne pas refaire les radiologies/scanners (à rayons X !), etc.
- Economies des examens médicaux déjà réalisés

### 👉 En marketing :

- Meilleure connaissance du client et de ses besoins : mieux cibler, mieux fidéliser, réduire les coûts des opérations de marketing, etc.



# Intégration de données

## Enrichissement de données

☞ Incomplétude des données internes

Source externe		
Appartements de Paris		
Id	Aire m2	nb-occupants
a1	70	3
a5	20	1

Base interne		
Appartements		
Num	Adresse	Type
a1	Paris	T2
a2	Paris	T3



Evaluation de  
requêtes

Info sur les  
appartements  
de Paris ?



Résultats

Id	Adresse	Type
a1	Paris	T2

**Incomplet !**

# Intégration de données

## Enrichissement de données

➔ Hétérogénéité structurelle des données

### Source externe

```
<appartements>
  <appart num="a1">
    <aire>70</aire>
    <nb-occ>3</nb-occ>
    <cp>75011</cp>
  </appart>
  <appart num="a2">
    <aire>20</aire>
    <nb-occ>1</nb-occ>
  </appart>
</appartements>
```

### Base interne

#### Appartements

Num	Adresse	Type
a1	Paris	T2
a2	Paris	T2

Evaluation de  
requêtes

Info sur les  
appartements  
de Paris ?



### Résultats

Id	Adresse	Type
a1	Paris	T2

**Incomplet !**

# Intégration de données

## Enrichissement de données

👉 Duplication des entités du monde réel

datasource1		
Clients		
Id	tél	a-voiture
c1	06.20	v1
c2	07.40	v2

Datasource 2		
Clients		
Id	tél	a-maison
cc10	06.20	m1
cc20	01.50	



Evaluation de  
requêtes

Résultats

Id	tél	a-voiture
c1	06.20	v1



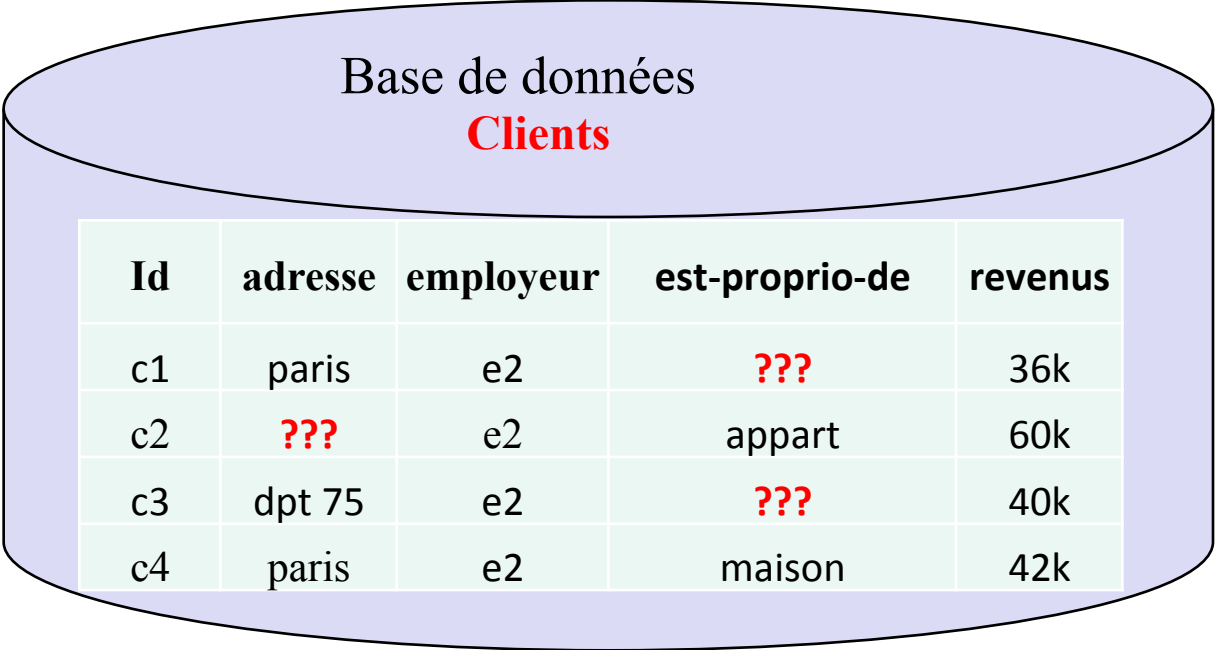
⁉️ Tout savoir sur le  
client **c1** ?

**Incomplet !**

# Qualité de données

Enrichissement par les méthodes statistiques (non traitées !)

☞ Imputation de données manquantes



Base de données  
**Clients**

Id	adresse	employeur	est-proprio-de	revenus
c1	paris	e2	???	36k
c2	???	e2	appart	60k
c3	dpt 75	e2	???	40k
c4	paris	e2	maison	42k

☞ Quelques méthodes :

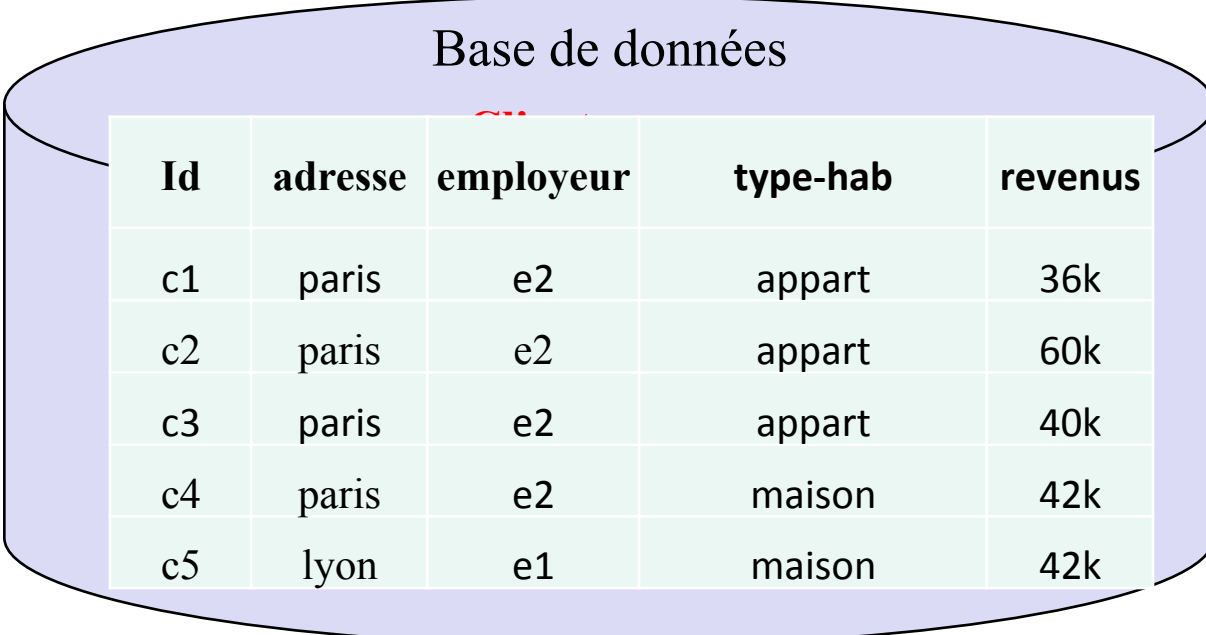
Knn (kppv), Régression, Inférence bayésienne, Gradient, etc.

☞ La qualité de l'approximation dépend de la richesse des données !

# Qualité de données

Enrichissement par les méthodes statistiques (non traitées !)

➡ Améliorer la fiabilité des modèles prédictifs



Base de données

Id	adresse	employeur	type-hab	revenus
c1	paris	e2	appart	36k
c2	paris	e2	appart	60k
c3	paris	e2	appart	40k
c4	paris	e2	maison	42k
c5	lyon	e1	maison	42k

Règle d'association :

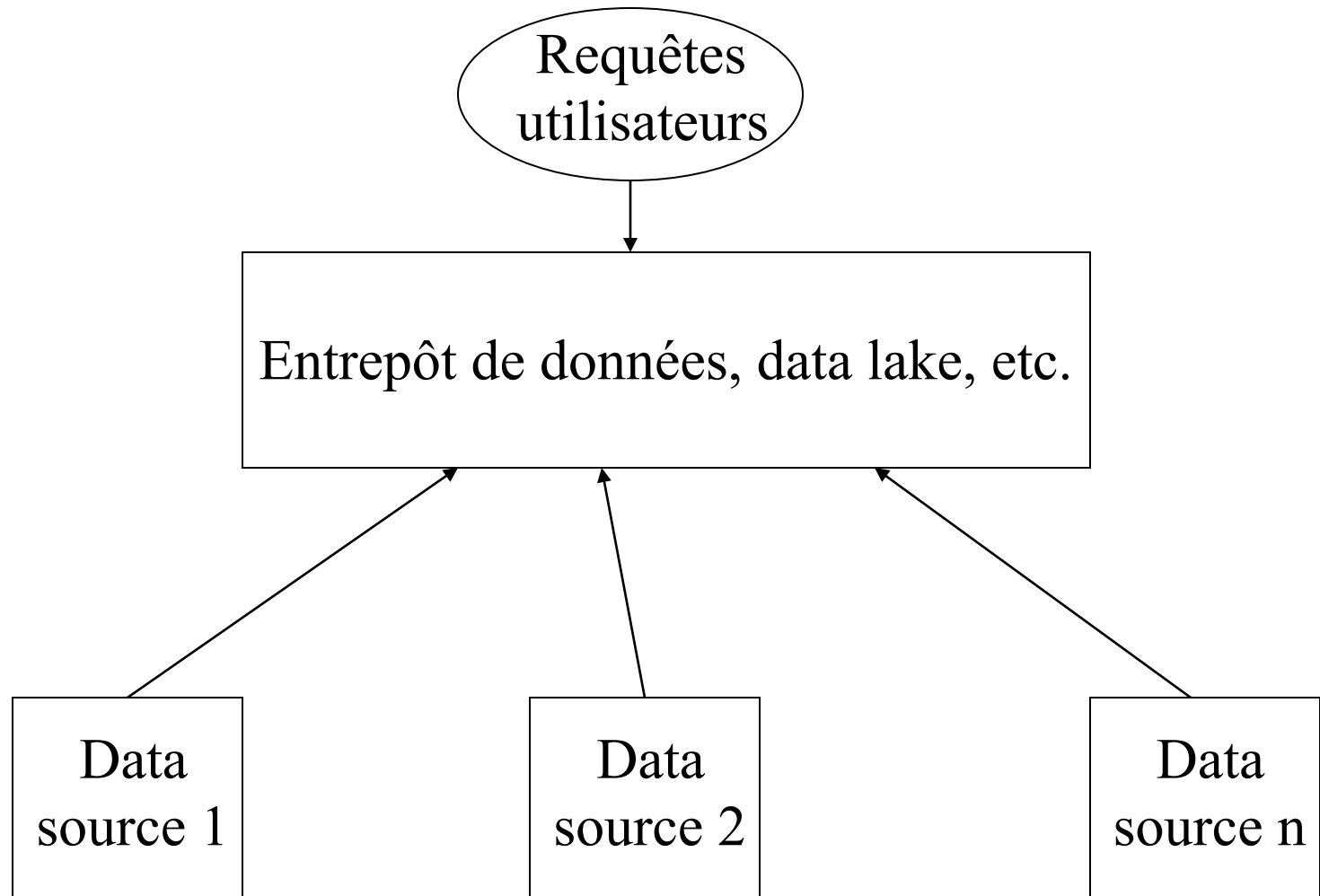
**(adresse="paris", employeur="e2") → type-hab="appart"**

support (indicateur de fiabilité) = 4 (ou 80% en relatif)

© M. Ouziri confiance (indicateur de précision) = 75%

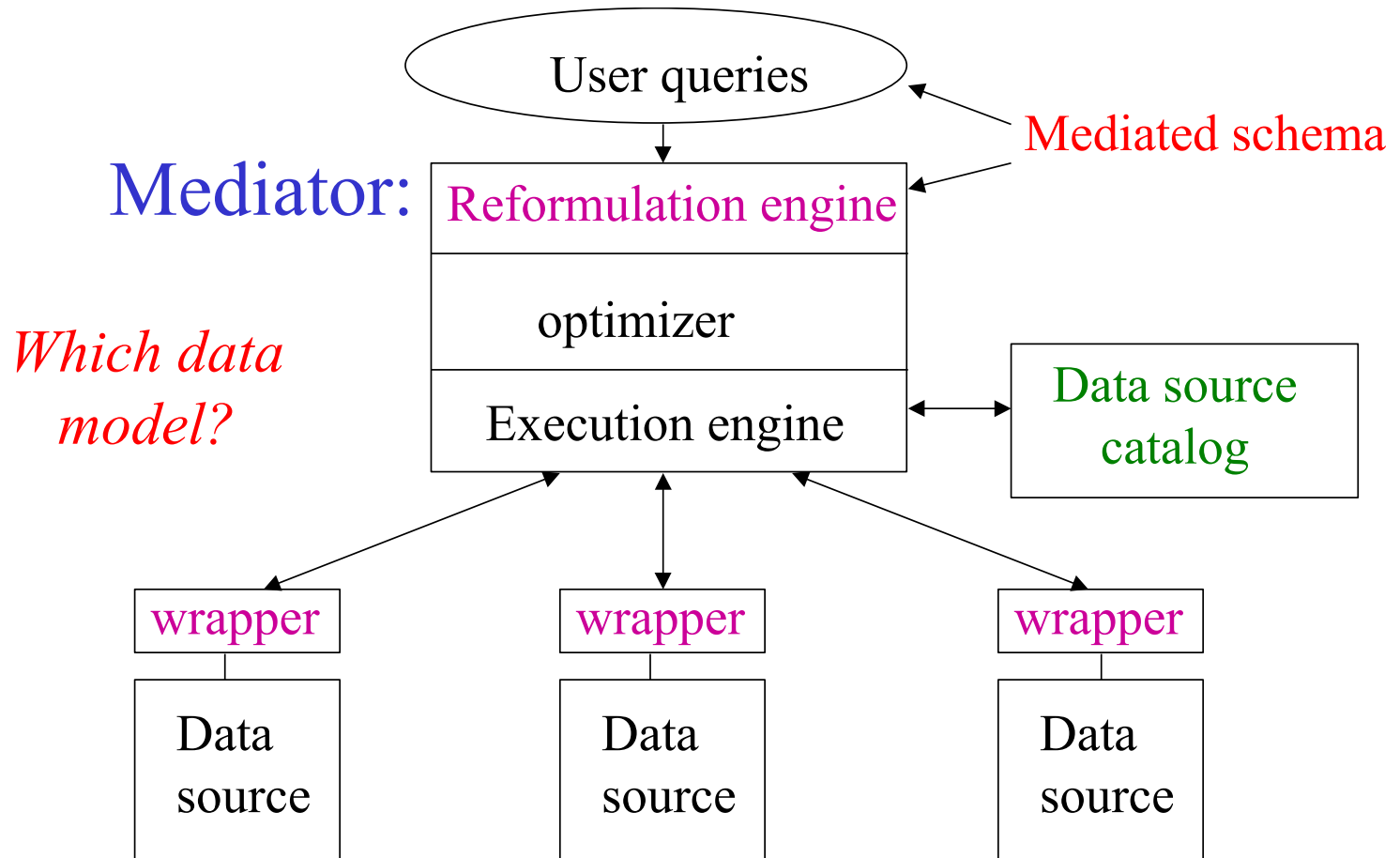
# Architectures d'intégration de données

## Intégration matérialisée



# Architectures d'intégration de données

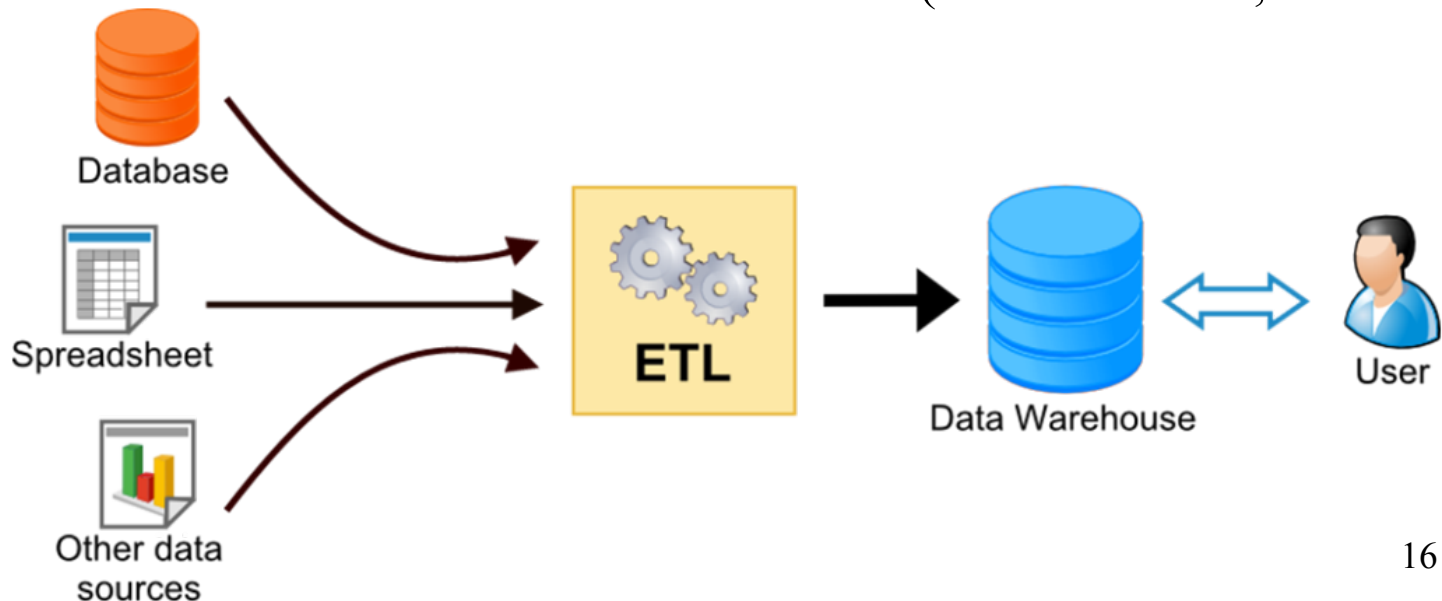
## Intégration virtuelle



# Outils d'intégration de données : ETL

👉 ETL pour Extract, Transform, Load

- Extract : récupération de données à partir de sources hétérogènes (fichiers textes, bases de données, services web, etc.)
- Transform : traitements de données (conversion de types et formats, de valeurs, nettoyage de données, agrégations, jointures, etc.)
- Load : insérer les données dans les sources cibles (datawarehouse, data lake, etc.)

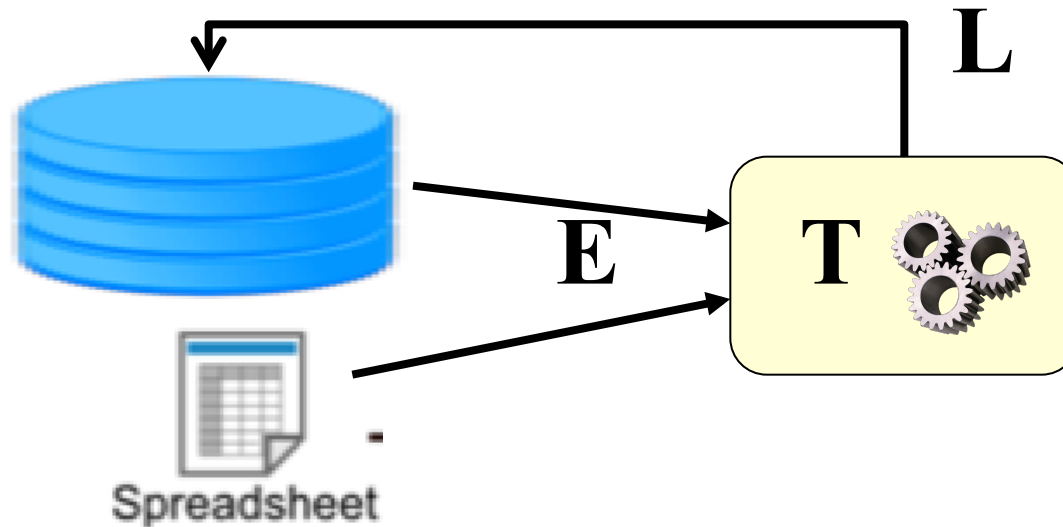




# Intégration de données dans le Big Data

## Limites du processus ETL

- ☞ Limites du mode ETL : problèmes de performances pour le traitement de gros volumes de données



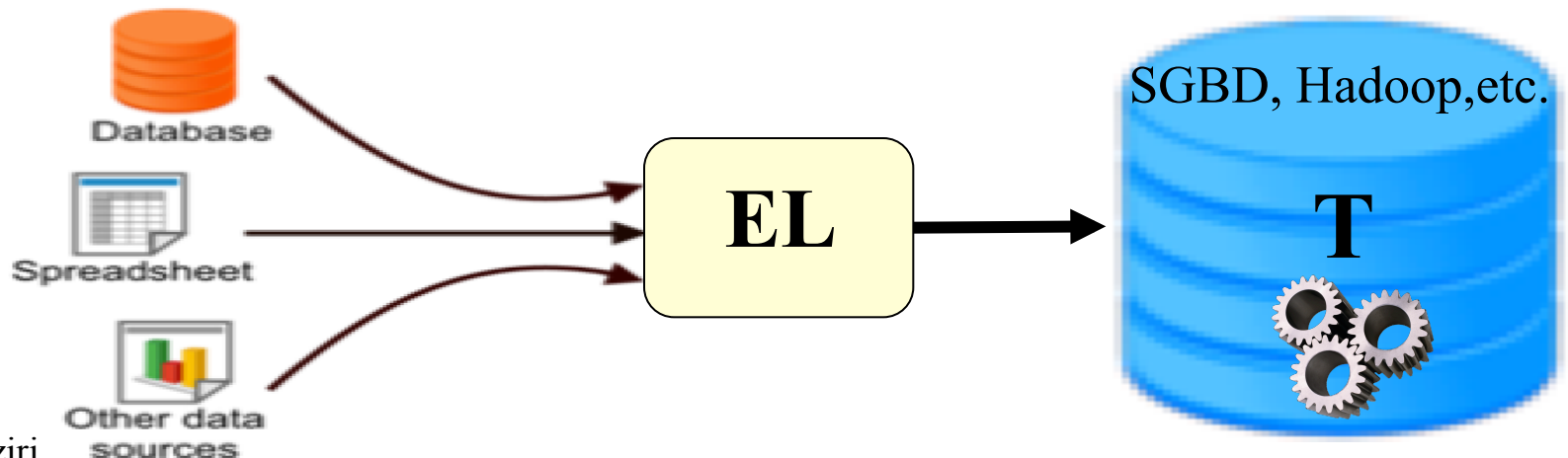
- ☞ Exemple

- Mise à jour d'une base de données avec des données issues d'un fichier externe !

# Intégration de données dans le Big Data : le processus ELT

👉 ELT : Extract, Load, Transform

- Mêmes objectifs que l'ETL, mais la manière diffère !
- Pousse les traitements (T de ELT) vers les données (philosophie de Hadoop !)
- Les traitements sont ainsi réalisés par le système cible hébergeant les données (SGBD, ERP, Hadoop, CRM, etc.)
- Par conséquent, les traitements sont écrits dans le langage cible (SQL, PL/SQL, Hive, Pig, MapReduce, Spark, etc.)



# Talend

<https://www.talend.com>

- ☞ Talend est un ETL/ELT open source de type générateur de code (Java)
- ☞ Produit de la société Telend créée en 2006 à Suresnes
- ☞ Talend offre une palette de composants graphiques de lecture/extraction, traitement/transformation et écriture/chargement de données
- ☞ Il dispose d'un éditeur graphique (basé sur Eclipse RCP) permettant de concevoir des processus (Job) d'intégration de données complets
- ☞ Il offre des composants permettant d'écrire du code Java et d'autres pour le chargement de package Java externes

# Talend

Talend Open Studio (3.2.0.M1\_r26328)

File Edit View Window Help

Repository Selezione

Job demo03\_tMap 0.1 Job beforeRunJobs 0.1 \*Job ComponentRow 0.1

Business Models  
Job Designs  
t01\_Compo  
CustomCode  
Databases  
Bulk  
InOut  
SCD  
SP  
ComponentRow 0.1  
Connection 0.1  
DataQuality  
ELT  
File  
Internet  
LogError  
Misc  
Orchestration  
Processing  
System\_

Find component...

Business  
Business Intelligence  
Custom Code  
Data Quality  
Databases  
AS400  
Access  
DB Generic  
DB JDBC  
DB2  
FireBird  
Greenplum  
HSQLDb  
Informix  
Ingres  
Interbase  
JavaDB  
ELT  
File  
Internet  
Logs & Errors  
Misc  
MultiSchema  
Orchestration  
Processing  
System  
XML

mysqlRow is the specific component for this database query.  
It executes the SQL query stated onto the specified database.  
The row suffix means the component implements a flow in the job design although it doesn't provide out..

Create table  
tCreateTable\_1 "demotable"

Insert the data  
tRowGenerator\_1  
tRowGenerator\_1 (Main)  
tMysqlOutput\_1

Drop the primary key  
tMysqlRow\_1

Set the primary key  
tMysqlRow\_2

OnSubJobOk

Designer Code

Job(Compo) Contexts(J) Component Run (Job C) Problems Modules Talend Exc Scheduler Job Hierarc

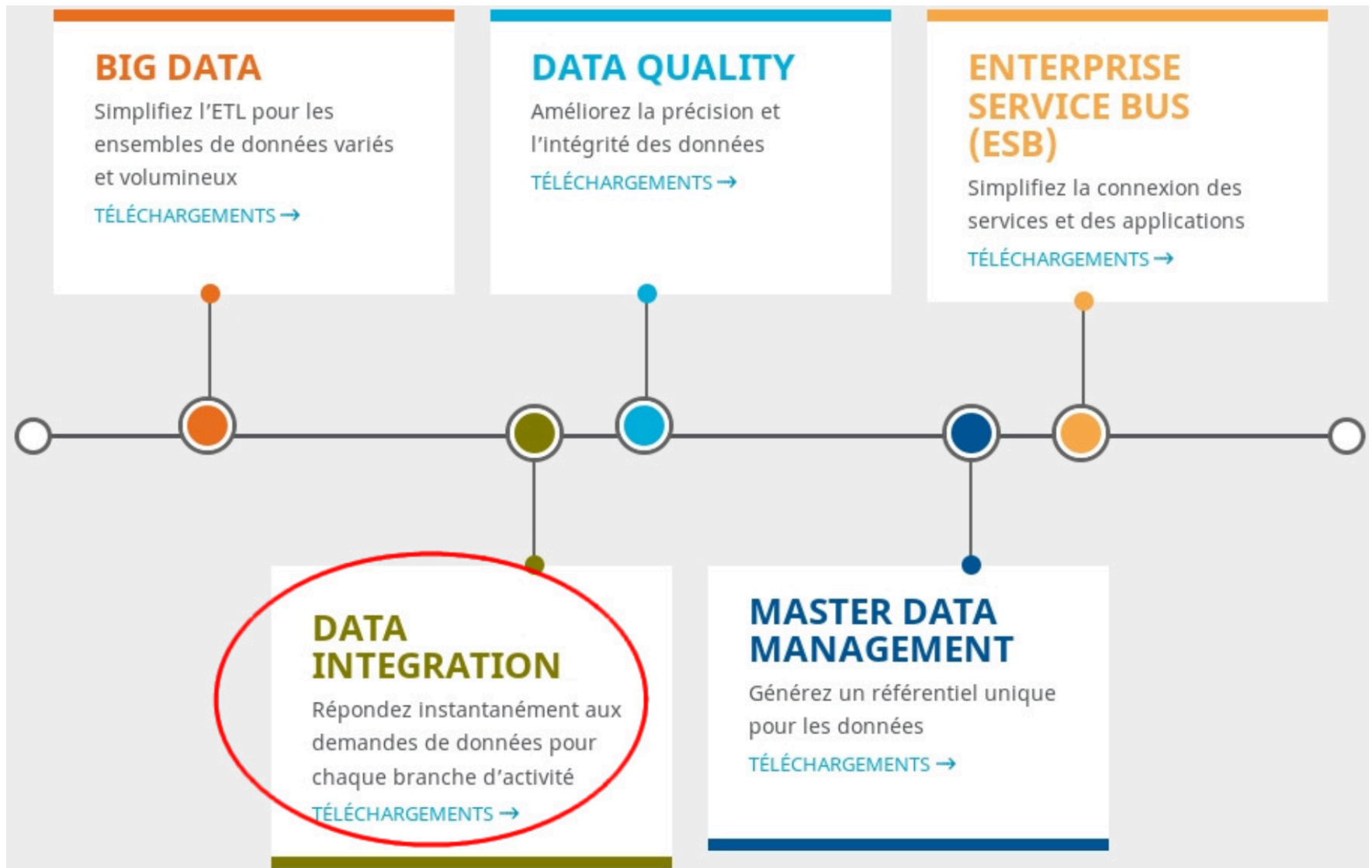
0 errors, 0 warnings, 1 info

Description	Resource
Errors (0 items)	
Warnings (0 items)	
Infos (1 item)	

tCreateTable\_1(tCreateTable\_1 "demotable")  
tMysqlOutput\_1  
tMysqlRow\_1(tMysqlRow\_1)  
tMysqlRow\_2(tMysqlRow\_2)  
tRowGenerator\_1(tRowGenerator\_1)

# Talend

Plusieurs éditions adaptées



# Talend

## Composants graphiques

### ☞ Lecture de données (E)

- Fichiers : tFileInputDelimited, tFileInputExcel, tFileInputXML,...
- Bases de données : tMySQLInput, tMyOracleInput, tMongoDBInput,...
- Services Web : tREST, tSOAP, etc.

### ☞ Traitement et transformation de données (T)

- tMap : transformation, jointure, filtre.
- tSortRow, tFilterRow, tUniqRow : tri, filtre, déduplication
- tNormalize et tDenormalize : normaliser les données tabulaires

### ☞ Ecriture de données (L)

- Fichiers : tFileOutputDelimited, tFileOutputExcel...
- Bases de données : tMySQLOutput, tMongoDBOutput,...

# Talend

## Composants graphiques

- ☞ Visualiser des résultats temporaires pour débogage
  - tLogRow : afficher les résultats sur la sortie standard de Talend
- ☞ Assemblage des composants : deux types de liens
  - Par transmissions de données (row)
  - Par évènement (trigger) : composant ok/error, Iterate
- ☞ Les matrices de données sont traitées sous forme de flux
  - Chaque ligne est traitée dans une itération
  - tAggregateRow permet de faire l'agrégation de plusieurs lignes

# Mise en pratique sur Talend Open Studio for Big Data !