

# Méthodologie : Introduction


Giuseppe Berio

[giuseppe.berio@univ-ubs.fr](mailto:giuseppe.berio@univ-ubs.fr)

2023



# Introduction aux méthodologies de conceptions des entrepôts/data-marts



- Typologies
- Approche « supply driven »
- Difficultés

# Typologies

3

	Supply-Driven	User-Driven	Goal-Driven
Basic approach	Bottom-up	Bottom-up	Top-Down
Users involvement	Low: DB Administrators	High: Business users	High: Top management
Constraints	Existence of a reconciled data level	Business users must have a good knowledge of the processes and organization of the company	Willingness of top management to participate in the design process
Strengths	The availability of data is ensured	Ensure the acceptance of the system.	Maximize the probability of a correct identification of the relevant KPIs.
Risks	The multidimensional schemata do not fit business user requirements.	Quick obsolescence of the multidimensional schemata due to changes of the business users.	Difficulties in being supported by top management and in translating the business strategy into quantifiable KPIs.
Targeting organizational level	Operational and tactical	Depends on the level of the interviewed users, typically tactical	Strategic and tactical
Skills of project staff	DW designers	Moderators; DW designers	Moderators; Economist; DW designers
Risk of obsolescence	Low	High	Low
Number of source systems	Low	Moderate	High
Cost	Low	High	High

*Méthodologies hybrides*

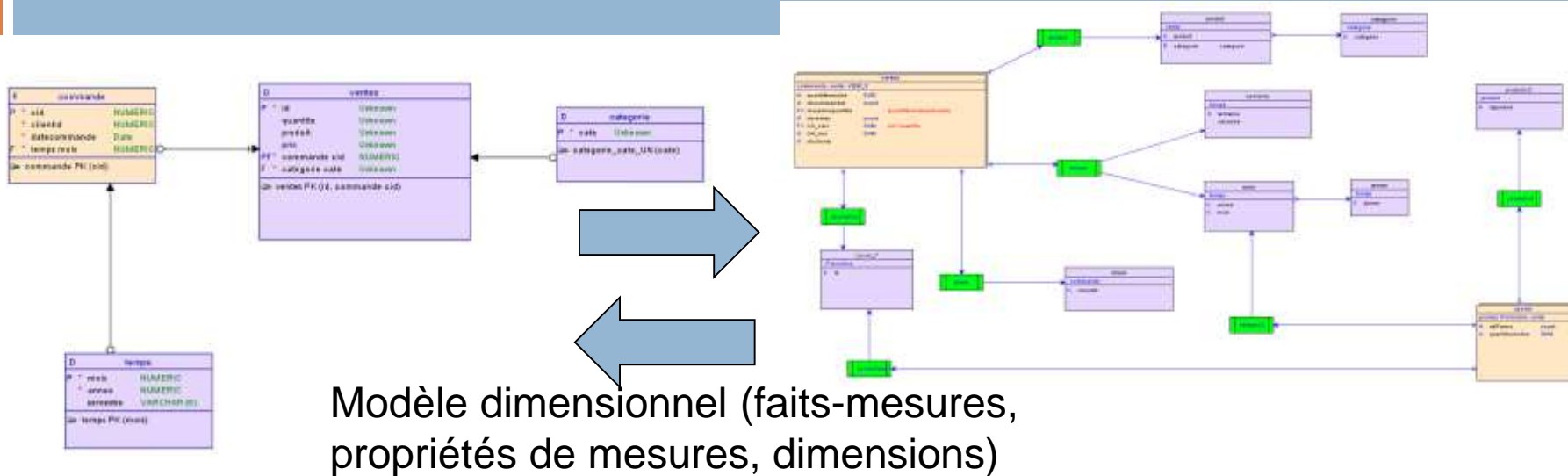
Source: DEXA2002

# Approche « supply-driven » : étapes

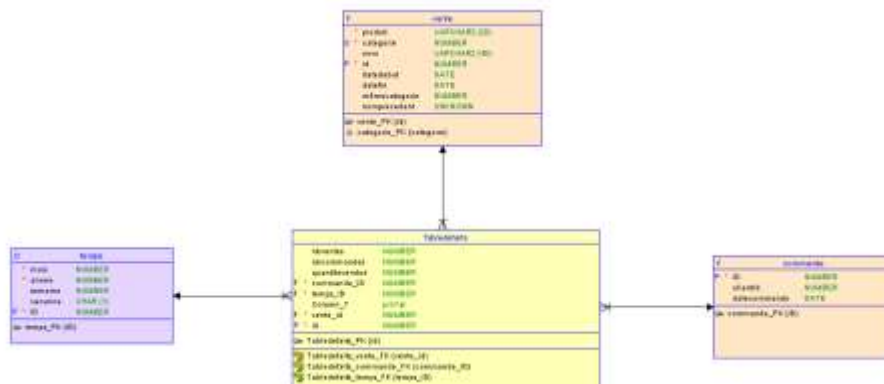
- **Identification des sources** (avec une analyse préalable de la **qualité de données** dans les sources) et **réimplantation des sources**
  - Transformation de **schémas**
- **Conception d'un schéma intégré (réconciliation des conflits)** et mise en œuvre
  - Transformation et Intégration de schémas
- **Conception d'un schéma multidimensionnel (faits et dimensions)** et mise en œuvre
  - Transformation de schémas
- **Conception et programmation des flux ETL (via le schéma intégré)**
  - **Flux d'extraction**
  - **Flux d'intégration**
  - **Flux de chargement**
  - Échange de **données**
  - Intégration de données
- **Chargement de données (sans fin)**
  - Rappel : Données ≠ Schéma; les données sont décrites par un schéma

# Rappel : Conception d'un schéma multidimensionnel (faits et dimensions) et mise en œuvre (ROLAP)

5



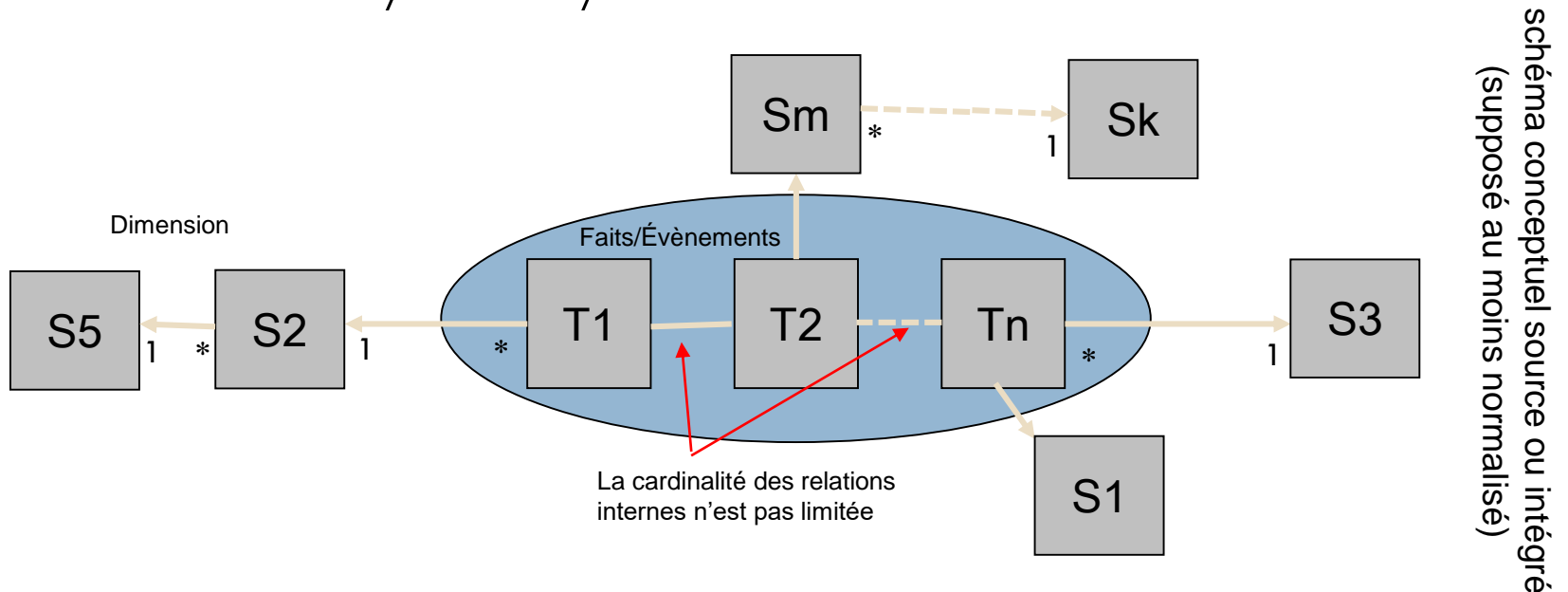
Modèle logique ROLAP ( star vs snowflake, TdF, clés, SCD, RCD)



Modèle physique (vues matérialisées, index, partitionnement, stockage par colonne, parallélisation)

# Conception d'un schéma multidimensionnel conceptuel à partir d'un schéma conceptuel source ou intégré

- Les informations qui sont « autour » des  $T1, \dots, Tn$  peuvent être utilisées comme dimensions/niveaux/hierarchies

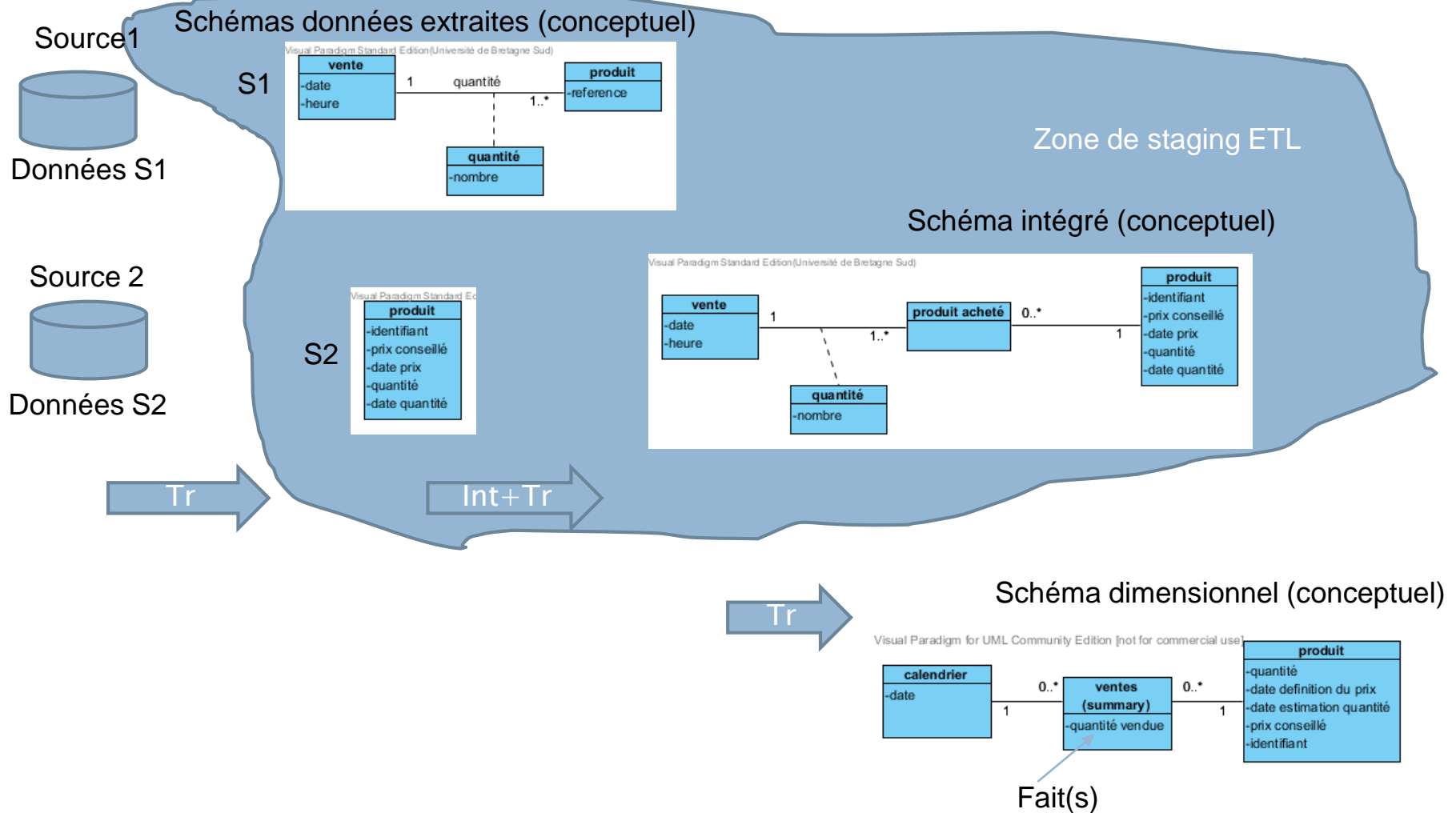


- Moins convenable pour définir des dimensions conformes car dépendent des sources ; des transformations de schéma sont possibles pour redéfinir les dimensions et grâce à la transitivité des relations 1..\*

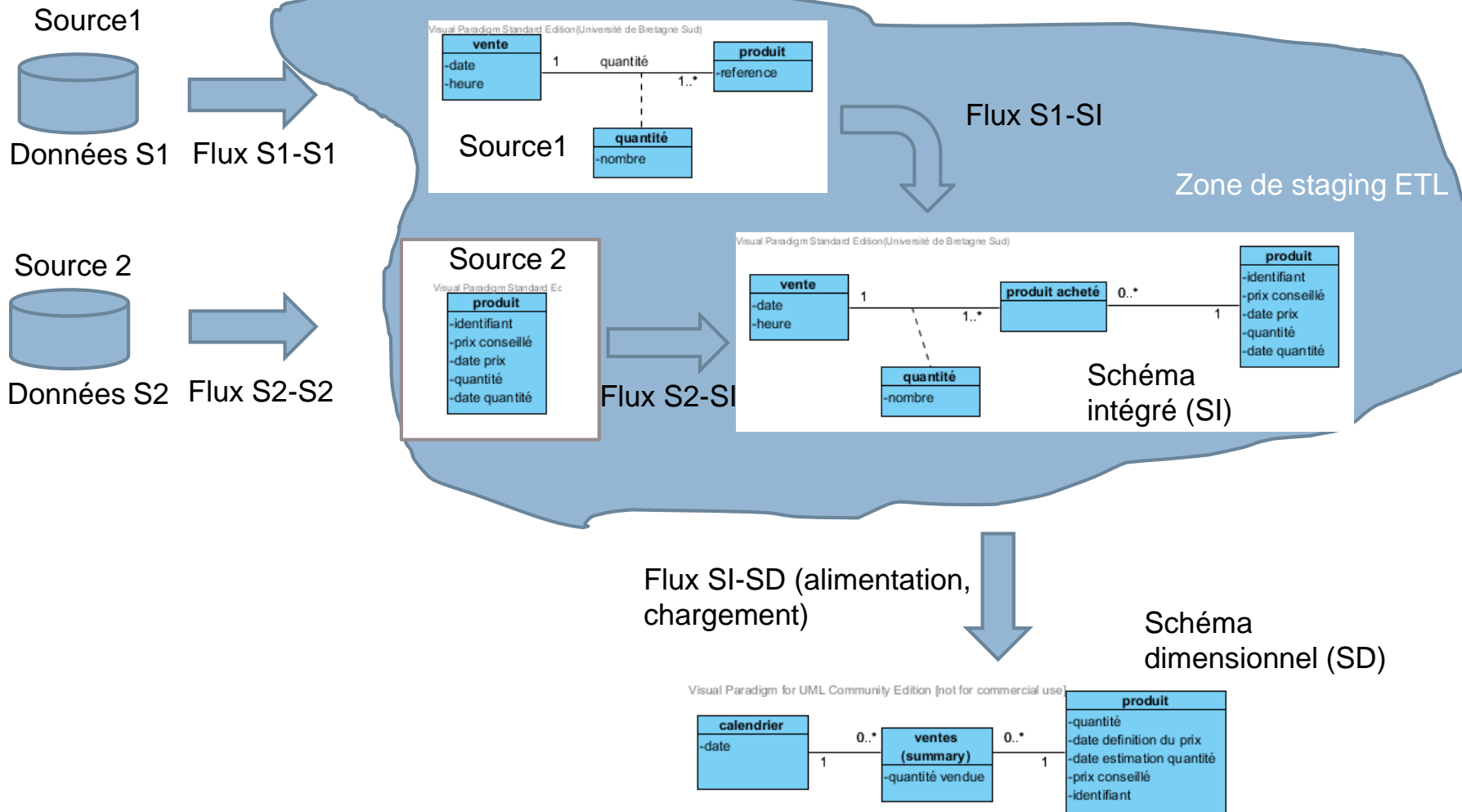
Exemple détaillé des étapes de conception :

[https://www.ibm.com/support/knowledgecenter/en/SS9UM9\\_9.1.1/com.ibm.datatools.dimensionai.ui.doc/topics/c\\_dm\\_design\\_phase\\_cont.html](https://www.ibm.com/support/knowledgecenter/en/SS9UM9_9.1.1/com.ibm.datatools.dimensionai.ui.doc/topics/c_dm_design_phase_cont.html)

# ETL : Zone de staging et schémas



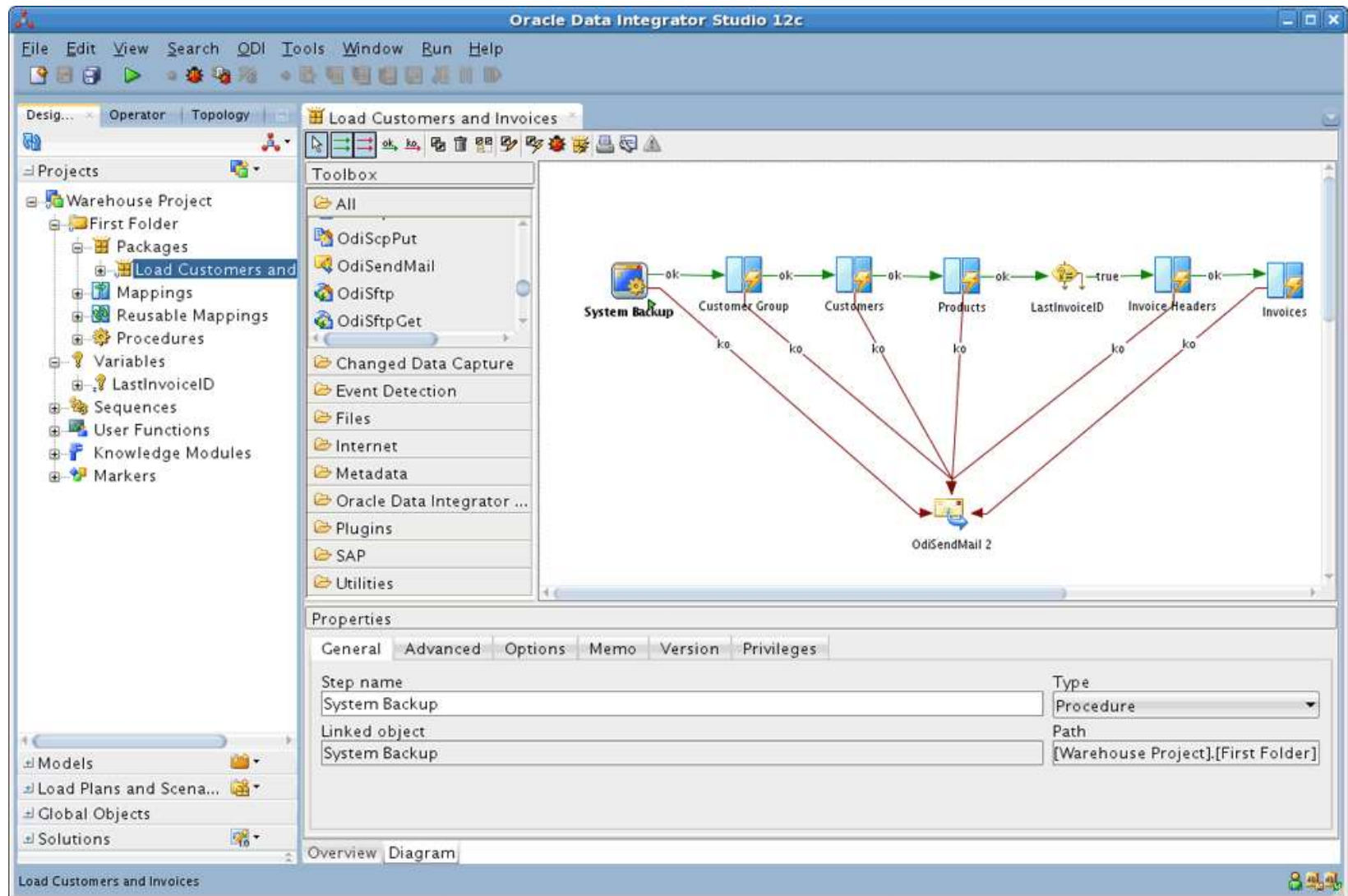
# ETL : Flux





# Interfaces programmation flux ETL (ORACLE DI)

9



# Interfaces programmation flux ETL (SAS DI)

10

**Prepare Customer Target TRAIN data**

Up Run Stop Continue Step Run From Selected Transformation Run Selected Transformations

WEB\_CUSTO... (WEB\_CUST...)  
Target Summe... (WEB\_SUMM...)  
WEBPATH (WEBPATH)

1 SQL Join  
2 SQL Join  
3 Sort  
4 Frequency  
5 Transpose  
Transpose\_O... (WS1JZNNP)

Diagram Code Log Output

**Details**

Status Warnings and Errors Statistics Control Flow

Node	Name	Status
0	Precode	Completed successfully
1	Sort	Completed successfully
2	Frequency	Completed successfully
3	Transpose	Completed successfully
4	Postcode	Completed successfully
	Prepare Customer Target TRAIN ...	Completed successfully

**View Data: Transpose\_OUTPUT2 (290 rows) (Browse)**

#	customer_id	_NAME_	COL1	COL2	COL3	
19	0x115	requested_file	/Cookie_C...	/Departme...	/Email.jsp ...	/Home
20	0x116	requested_file	/Billing.jsp ...	/Cart.jsp ...	/Confirm.js...	/Cooki
21	0x117	requested_file	/Product.js...			
22	0x118	requested_file	/CDMA/Err...			
23	0x119	requested_file	/Home.jsp ...	/Product.js...	/Search.js...	/Site_5
24	0x12	requested_file	/Billing.jsp ...	/Cart.jsp ...	/Confirm.js...	/Depai
25	0x120	requested_file	/Catalog_...	/Catalog_...	/Cookie_C...	/Depai
26	0x121	requested_file	/Home.jsp ...			
27	0x122	requested_file	/MiniDisc/E...	/Product.js...	/Retail_Sto...	
28	0x123	requested_file	/Cookie_C...	/Departme...	/Product.js...	/Subca

Completed successfully

# Matrice méthodologique

11

Identification sources	Conception SI	Conception SD	Conception Flux	Conception
But : Limiter les erreurs, expliciter les données, standardiser les données disponibles dans les sources ; Transformation de schémas	But : Disposer de données sans redondance, plus complètes, plus fraîches ; Intégration de schémas ET Transformation de schémas	But : Choisir des mesures, les regrouper et les décliner selon des dimensions d'intérêt ; Transformation de schémas	Conception flux d'extraction	
			Conception flux d'intégration	
			Conception flux de chargement vers l'entrepôt	
Réimplantation sources	Mise en œuvre SI	Mise en œuvre SD	Programmation flux	Mise en œuvre
Transformation de schémas + déploiement	Transformation de schémas + déploiement	Transformation de schémas + déploiement	Programmation flux d'extraction (y compris parallélisation, extractions successives)	
			Programmation flux d'intégration (y compris parallélisation, intégrations successives)	
			Programmation flux de chargement vers l'entrepôt (y compris parallélisation, chargements successifs)	

# Exemple (Conception schéma dimensionnel)

12

- FLIGHTS (flightNumber, airline, fromAirport:AIRPORTS, toAirport:AIRPORTS, departureTime, arrivalTime, carrier)
- FLIGHT\_INSTANCES (FlightNumber:FLIGHTS, date)
- AIRPORTS (IATAcode, name, city, country)
- TICKETS (ticketNumber, flight:FLIGHT\_INSTANCES, seat, fare, passengersFirstName, passengersSurname, passengersGender)
- CHECK-IN (ticketNumber:TICKETS, CheckInTime, numberOfBags)

A noter : les relations 1..\* ne sont pas explicites au niveau du modèle logique relationnel de la source

# Exemple (mesures/formalisation)

13

## □ nbBaggage →

- Fait (observé) : Check-in.nbBaggages
- $M(\text{Checkin.nbBaggages}, Nht, \dots, Nij) = \text{Select Sum}(\text{Checkin.nbBaggages}), Nht, \dots, Nij \text{ from } Nht, \dots, Nij \text{ (jointure) group by } Nht, \dots, Nij$ 
  - $Nht = T.c \text{ } Nij = T'.c' \text{ and } c \neq c'$
- $Fl(Nij, \dots, Nkj) = M(\text{Checkin.nbBaggages}, Nij, \dots, Nkj)$
- M est additive car définie par Sum
- M est agregable car elle est additive

## □ NbVols →

- Fait (observé) : Vol.num
- $M(\text{Vol.num}, Nht, \dots, Nij) = \text{select Count}(\text{Distinct Vol.num}), Nht, \dots, Nij \text{ from } Nht, \dots, Nij \text{ (jointure) group by } Nht, \dots, Nij$ 
  - $Nht = T.c, \dots, Nij = T'.c' \text{ and } c \neq c'$
- M est additive par rapport à Trajet
- M n'est pas additive par rapport à Passager

# Formalisation de la notion de mesure

14

- Hypothèses :
  - **Fait** est l'observation comme valeur numérique (ou \* est l'évènement sans fait)
  - **FI** est un fait inféré
  - $N_{ji}$  est un niveau hiérarchique  $i$  (par exemple le nom d'une colonne d'une table) d'une dimension  $j$
  - **M** est la mesure associée à un Fait, calculable sur **les données disponibles** permettant d'obtenir un FI à savoir  $FI(Nht, \dots, N_{ij}) = M(Fait, Nht, \dots, N_{ij})$  ; **M** réalise l'inférence (il s'agit d'une convention – donc il n'est pas l'observable/observé mais FI est le mesurable/mesuré)
  - $N_{ji}$  et  $N_{ji+1}$  sont 2 niveaux hiérarchiques directement liés au sein de la même dimension  $j$ , étant  $N_{ji}$  le détail et  $N_{ji+1}$  l'agrégé
- **M est additive** ssi pour chaque dimension ( $j$ ) et tout niveau ( $i$ )
  - $M(Fait, N_{ji+1}) = \text{Sum}(M(Fait, N_{ji}), N_{ji+1})$
  - S'il existe au moins 1 dimension ne satisfaisant pas la condition ci-dessus, alors **M est semi-additive**
- **M est agrégable** ssi il existe un *opérateur d'agrégation* **OP** tel que
  - $M(Fait, N_{ji+1}) = OP(M(Fait, N_{ji}), N_{ji+1})$
- **M est calculé** (ou calculable) ssi il existe une formule **F** telle que
  - $M(Fait, Nht, \dots, N_{ij}) = F(M_1(Fait_1, Nht, \dots, N_{ij}), \dots, M_n(Fait_n, Nht, \dots, N_{ij}))$
- **M est dérivée** (ou dérivable) ssi il existe une formule **F** telle que
  - $Fait_{(n+1)} = F(Fait_1, \dots, Fait_n, Nht, \dots, N_{ij})$  et
  - $M(Fait_{(n+1)}, N_{ji+1}) = OP(M(Fait_{(n+1)}, N_{ji}), N_{ji+1})$

# Implantation SD : Table de faits

15

- La table de faits TdF est générée/schématisée sur la base de la schéma dimensionnel établi au niveau conceptuel par transformation de schéma
- La table de faits peut stocker (en fonction de la clé choisie) :
  - ▣ Les faits observés  $\rightarrow$  granularité transactionnelle
  - ▣ Des faits inférés  $FI(N_{ht}, \dots, N_{ij}) = M(\text{Fait}, N_{ht}, \dots, N_{ij}) \rightarrow$  granularité temporelle
- Une mesure calculée n'a pas besoin d'être stockée dans la table de faits
- Une mesure dérivée peut être stockée dans la table de faits en fonction de la granularité de la table
- Le schéma dimensionnel logique peut ensuite être transformé pour obtenir un schéma étoile si nécessaire
- Une fois que la table de faits est alimentée :
  - ▣  $TdF.\langle \text{Fait} \rangle = M(\text{Fait}, N_{11}, \dots, N_{k1})$  ou
  - ▣  $TdF.\langle FI(N_{11}, \dots, N_{k1}) \rangle = M(\text{Fait}, N_{11}, \dots, N_{k1})$s'il y a K dimensions, étant  $N_{j1}$  les 1<sup>ers</sup> niveaux de toutes les K dimensions

# Difficultés méthodologiques

16

- ❑ Analyse de la qualité de données issues des sources, sélection des sources
- ❑ Conception du schéma intégré
- ❑ Conception des flux ETL entre sources et schéma intégré
- ❑ Gestion des exceptions dans les flux ETL
- ❑ Gestion de l'évolution des sources
- ❑ Maîtrise de la réactivité aux changements dans les données issues des sources
- ❑ Maîtrise du compromis coûts/bénéfices