

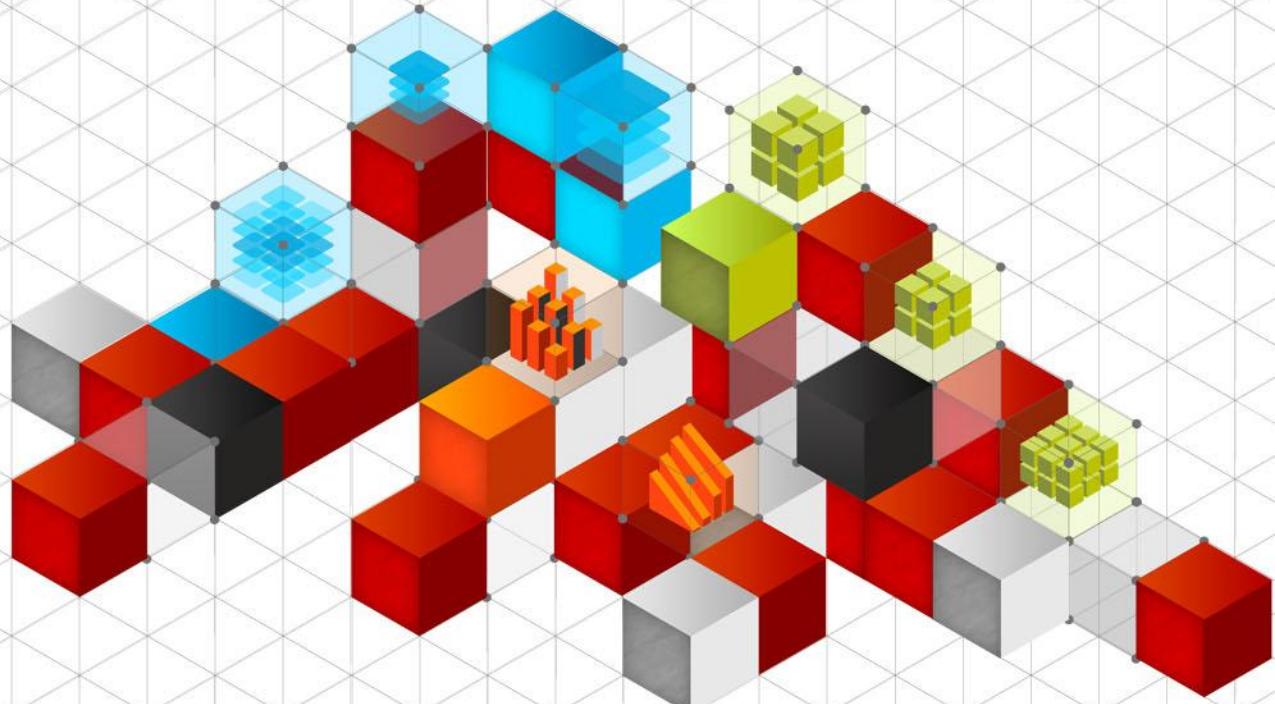
Les journées SQL Server

Paris, les 12 & 13 Décembre 2011

Rejoignez la Communauté

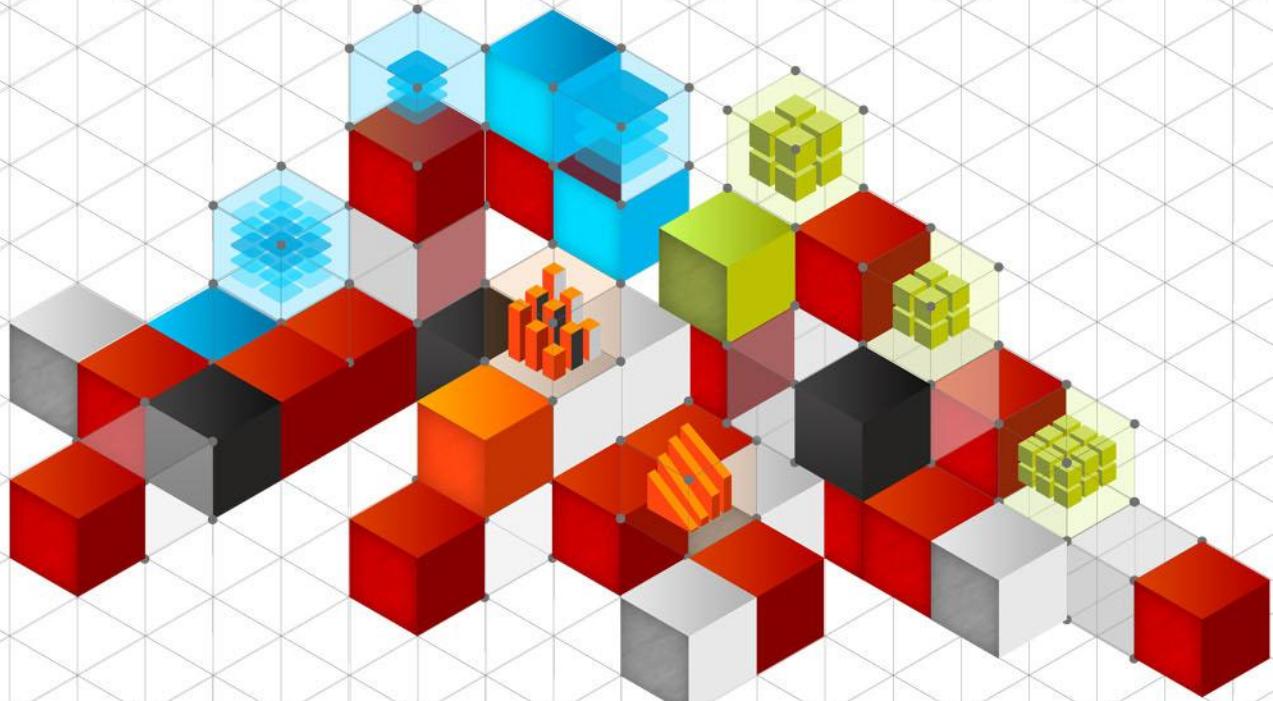
MODÉLISATION DIMENSIONNELLE

Le fondement du datawarehouse



MODÉLISATION DIMENSIONNELLE

Le fondement du datawarehouse



QUI SOMMES NOUS?

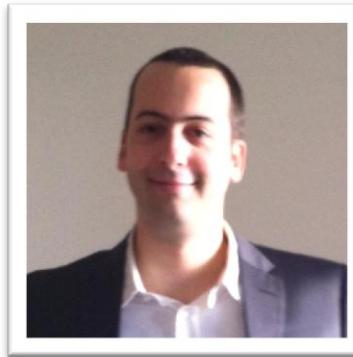
- Jean-Pierre Riehl

- Pratice Manager SQL – Azeo
- MVP SQL Server
- 5+ années sur le décisionnel Microsoft
- Blog : <http://blog.djeepy1.net>



- Florian Eiden

- Architecte Décisionnel - MCNEXT
- 5+ années sur le décisionnel Microsoft
- Blog : <http://www.fleid.fr>



MCNEXT (ESPACE PARTENAIRE)

- **En Bref**

- 100% Microsoft
- 150 collaborateurs / 40% de croissance
- 4 pôles : **BI, SharePoint, .NET, BizTalk**
- Présent à Paris, Lyon et Genève



- **Pôle décisionnel reconnu par Microsoft**

- Certifié « Gold Décisionnel »
- Sélectionné dans le programme METRO - SQL Server 2012

- **Expertise technique forte sur la Suite décisionnelle**
(SQL Server : SSIS, SSAS, SSRS; PowerPivot, MDS...)

- **Accompagnement global**

- Maîtrise d'ouvrage / Maitrise d'œuvre

Jean-Pierre Riehl – Responsable Practice SQL

<http://blog.djeepy1.net>



- **MVP** SQL Server
- **MCITP** : Business Intelligence Developer 2008
- **MCITP** : Database Administrator 2008
- **MCPD** : Enterprise Application



Pure-Player sur l'expertise Microsoft

- Practice Collaboration
- Practice SQL (Business Intelligence-Data Management)
- Practice Infrastructure
- Practice Développement

<http://www.azeo.com>

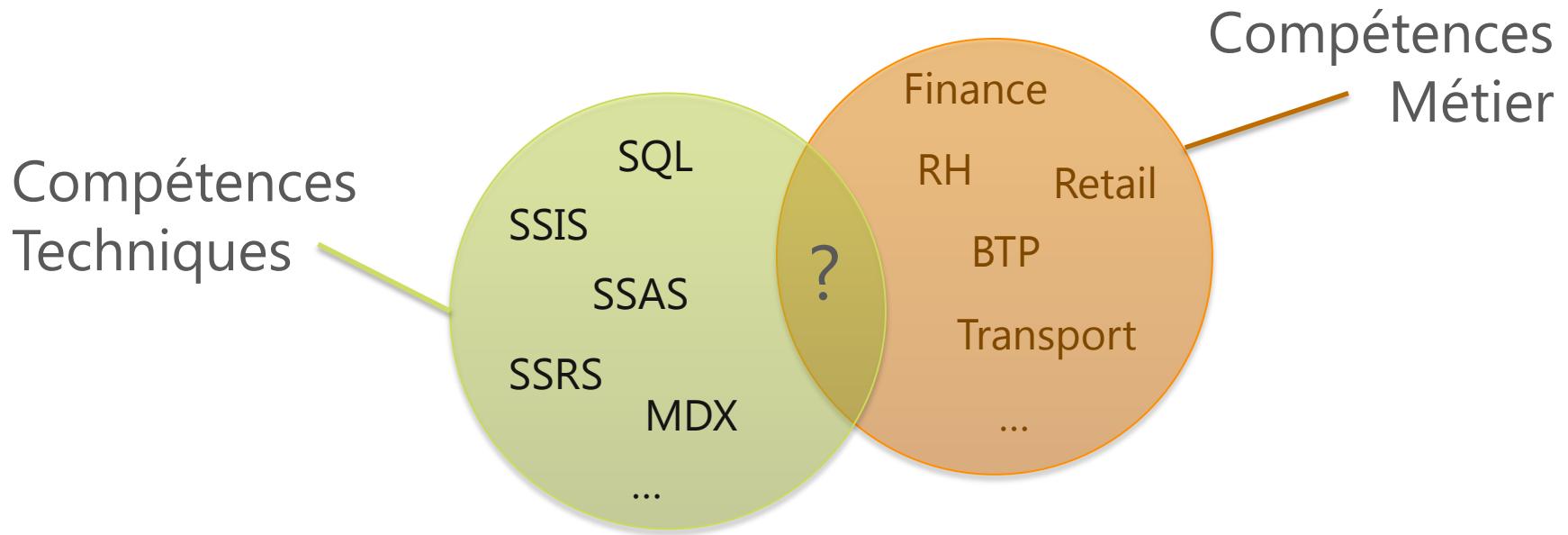


Microsoft® Partner

Silver Server Platform
Silver Application Integration



POURQUOI CETTE PRÉSENTATION



AGENDA

**1. Modélisation
dimensionnelle**

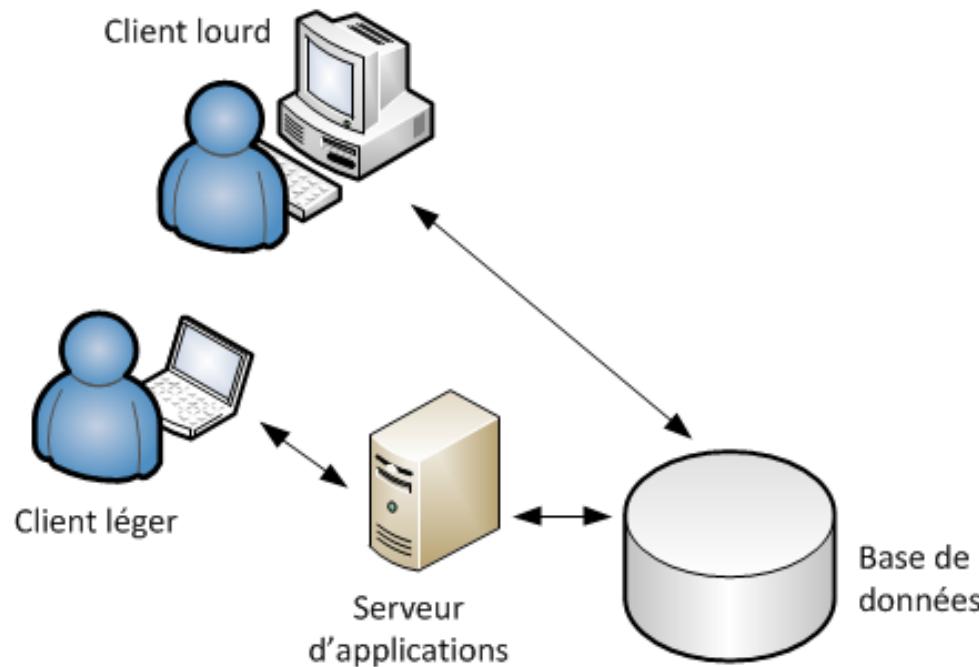
01.1
5

**2. Concepts
avancés**

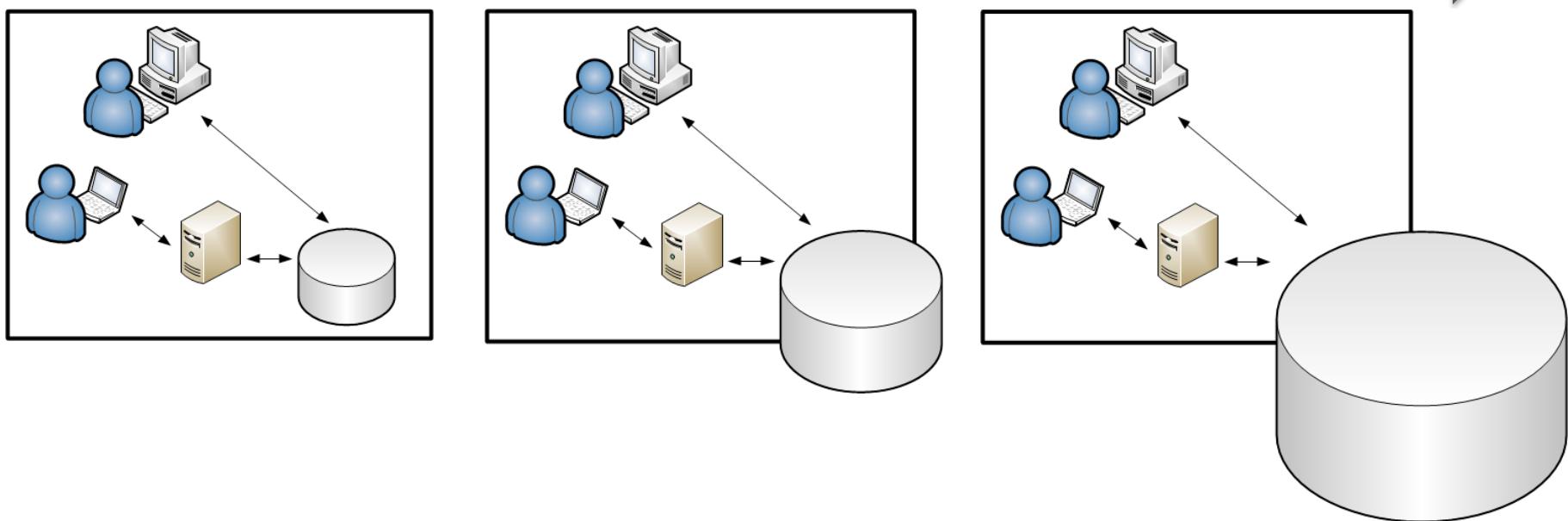


OBJECTIFS DU DÉCISIONNEL

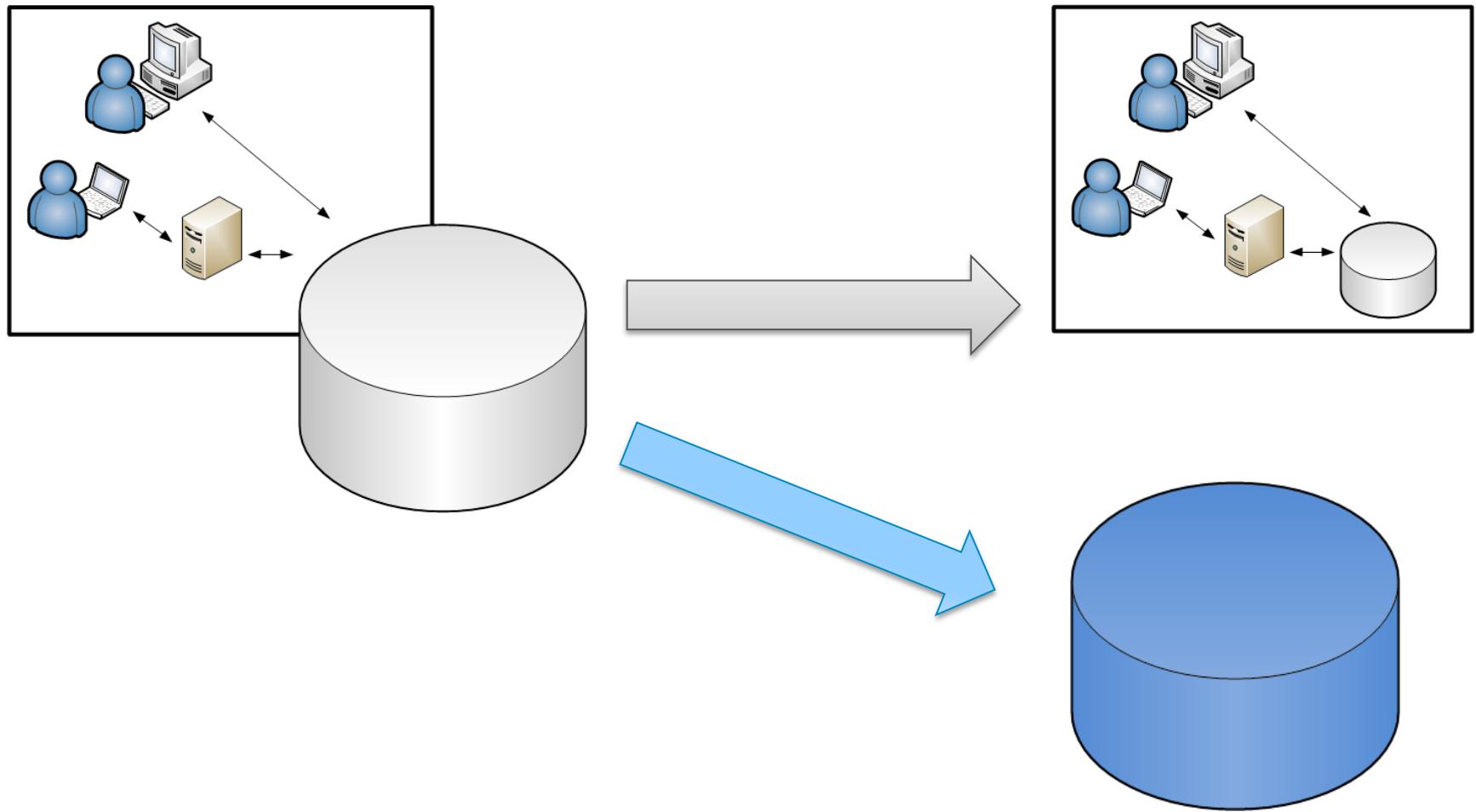
- Tout commence dans le monde applicatif...



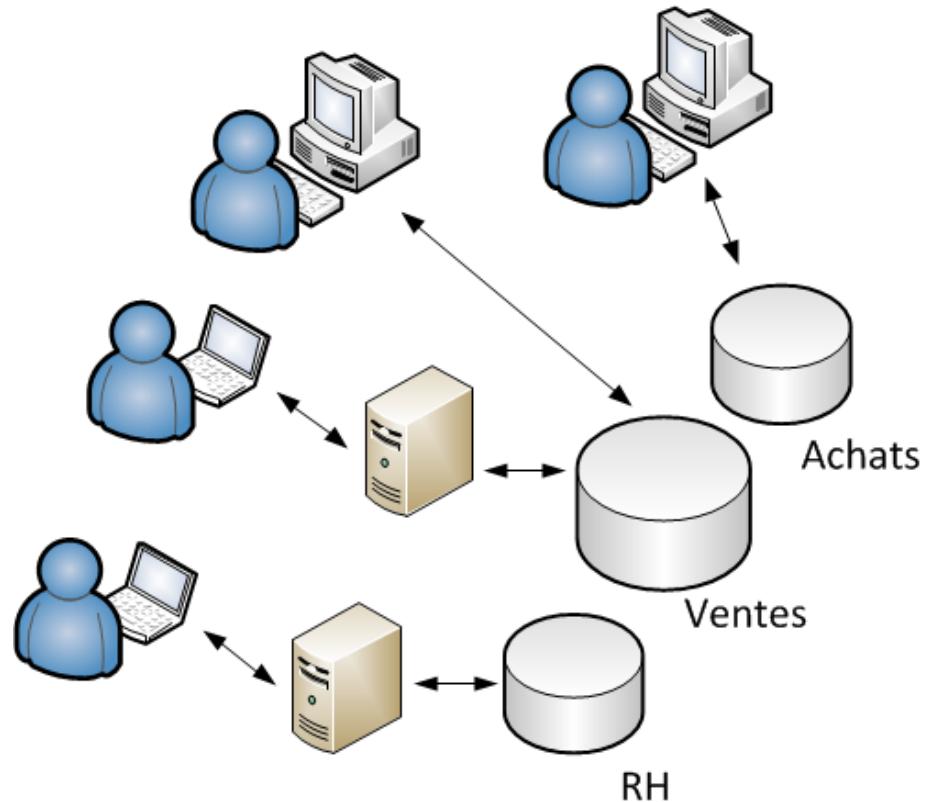
1^{ER} BESOIN MÉTIER



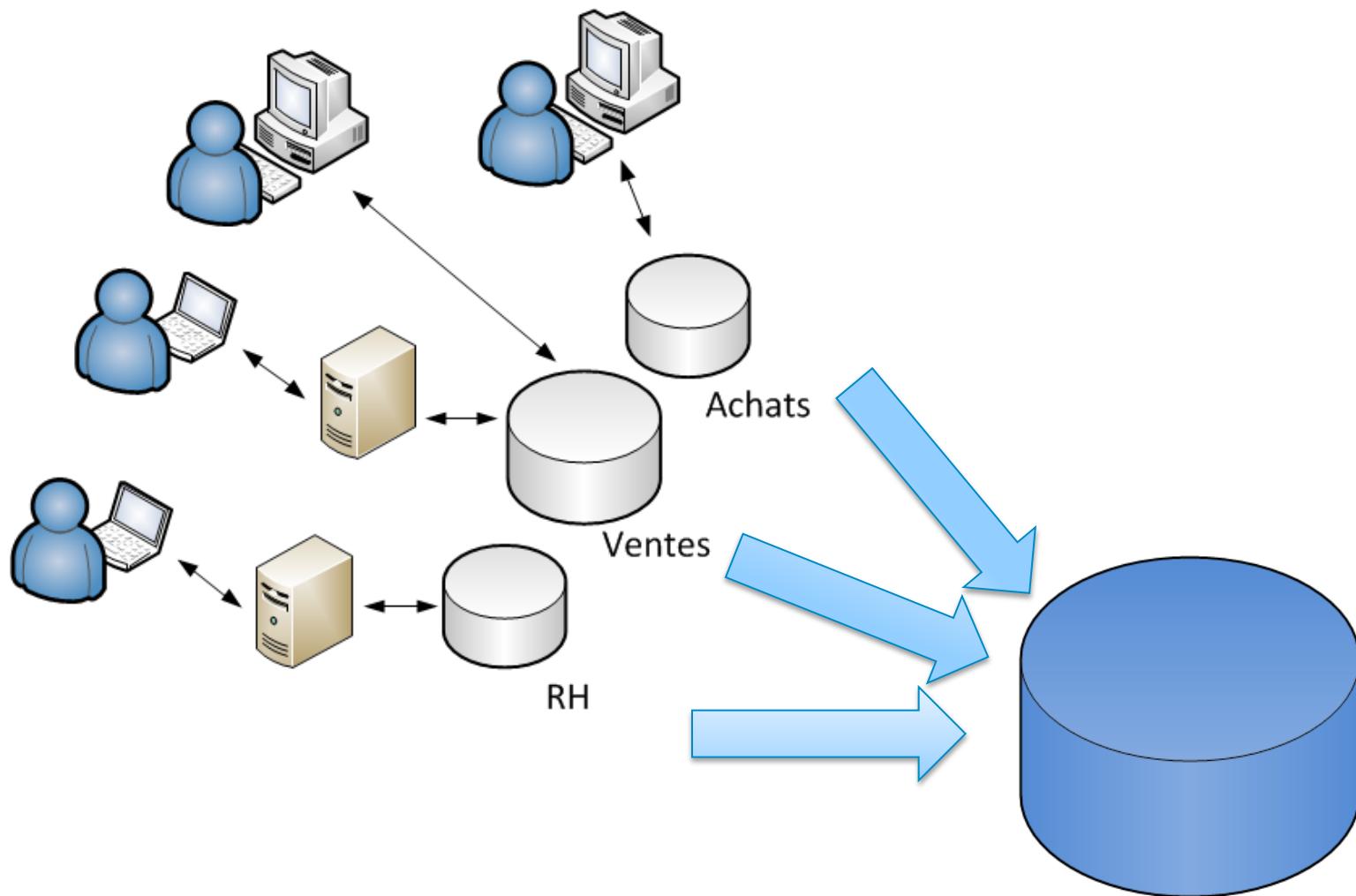
BESOIN : HISTORISATION



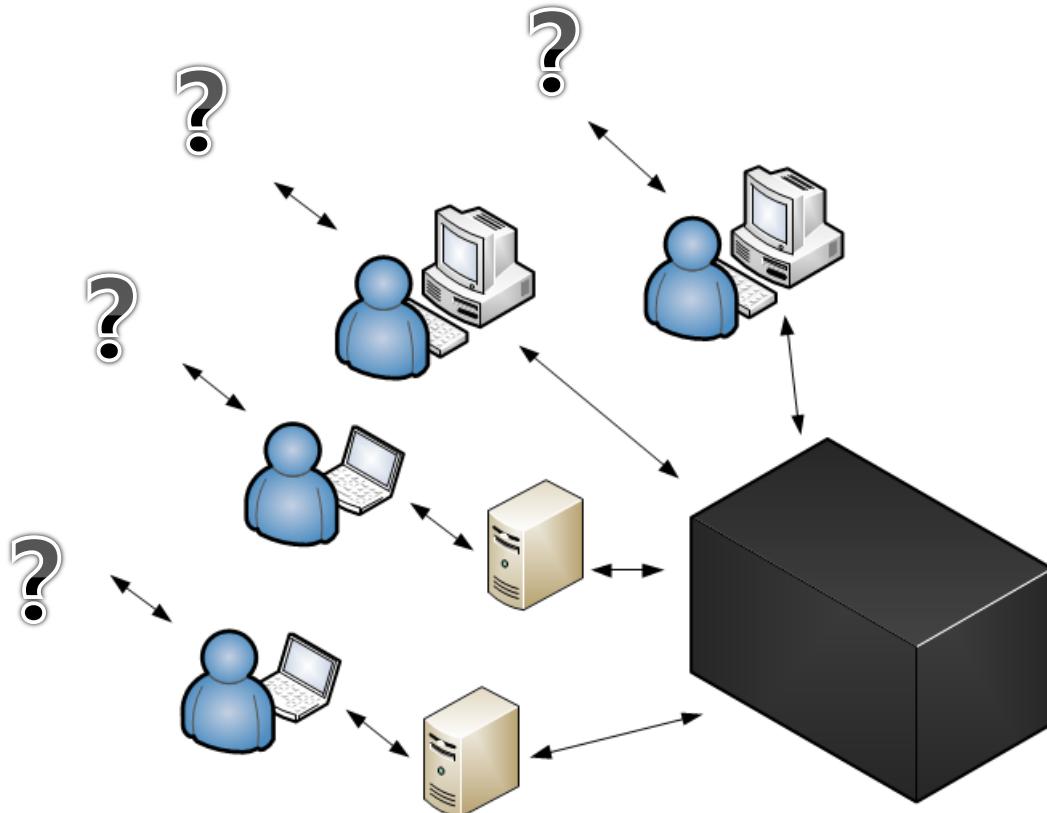
2ÈME BESOIN MÉTIER



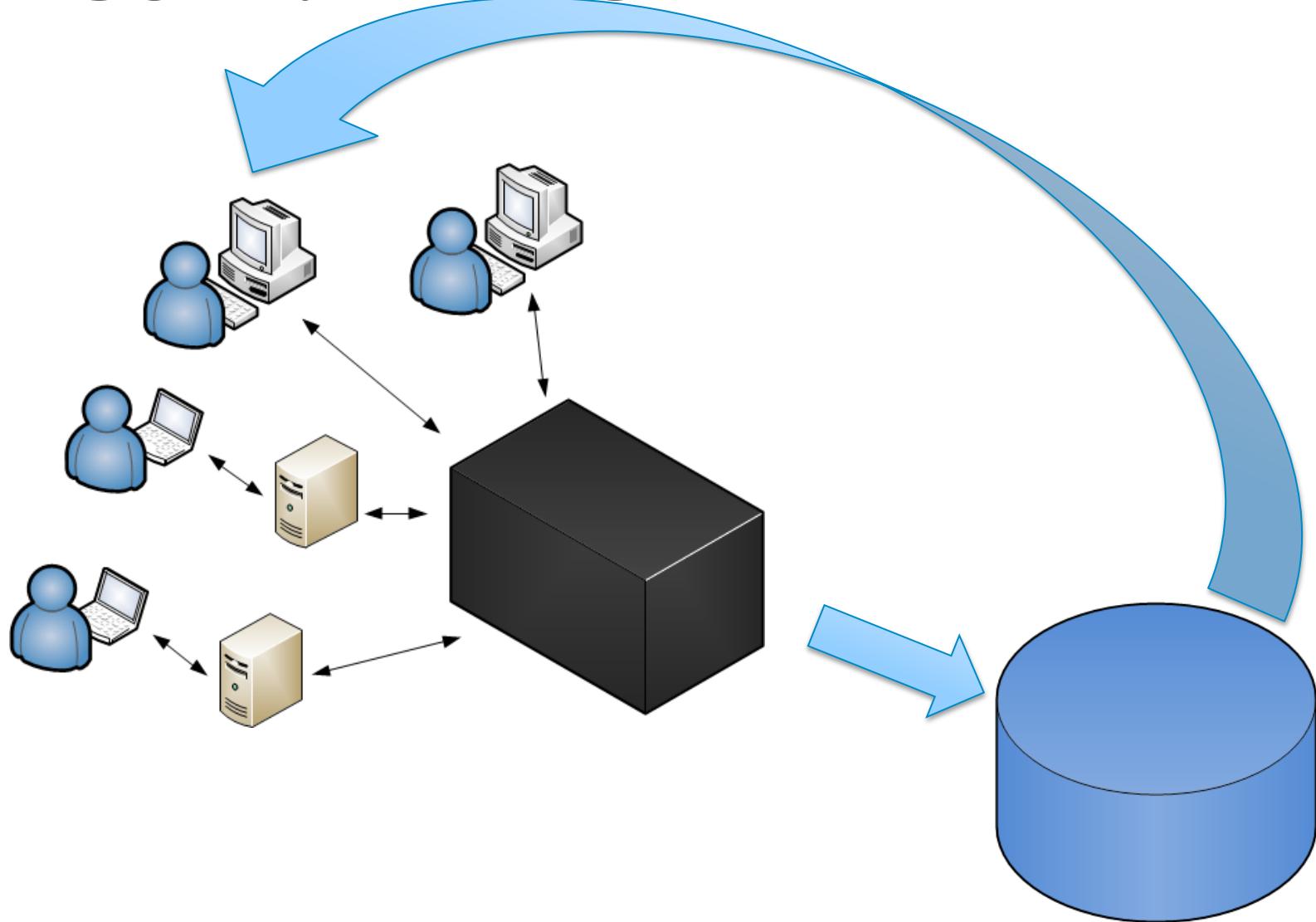
BESOIN : CENTRALISATION



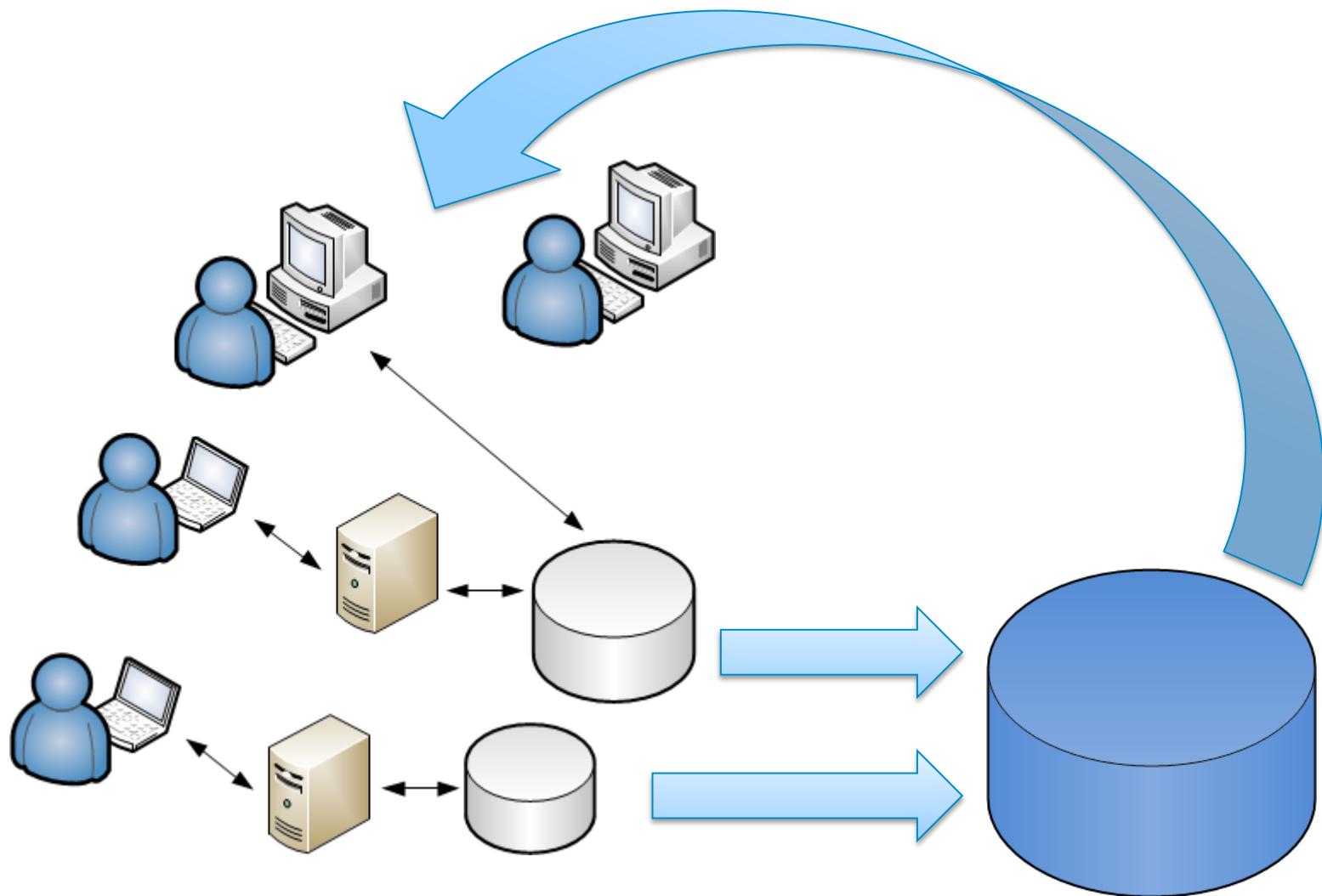
3ÈME BESOIN MÉTIER



BESOIN : ANALYSER



LE DATAWAREHOUSE

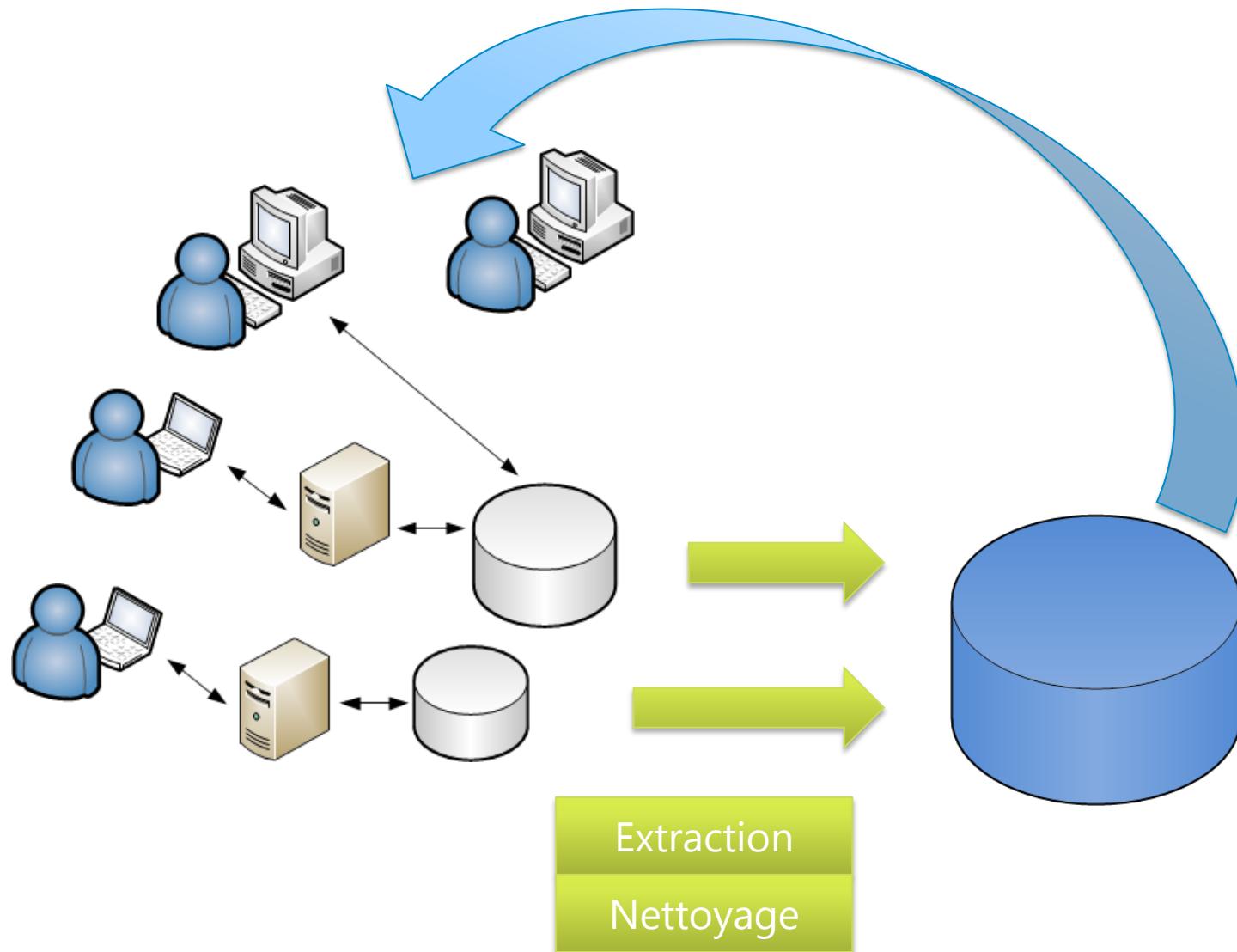


5 ÉTAPES POUR 3 BESOINS

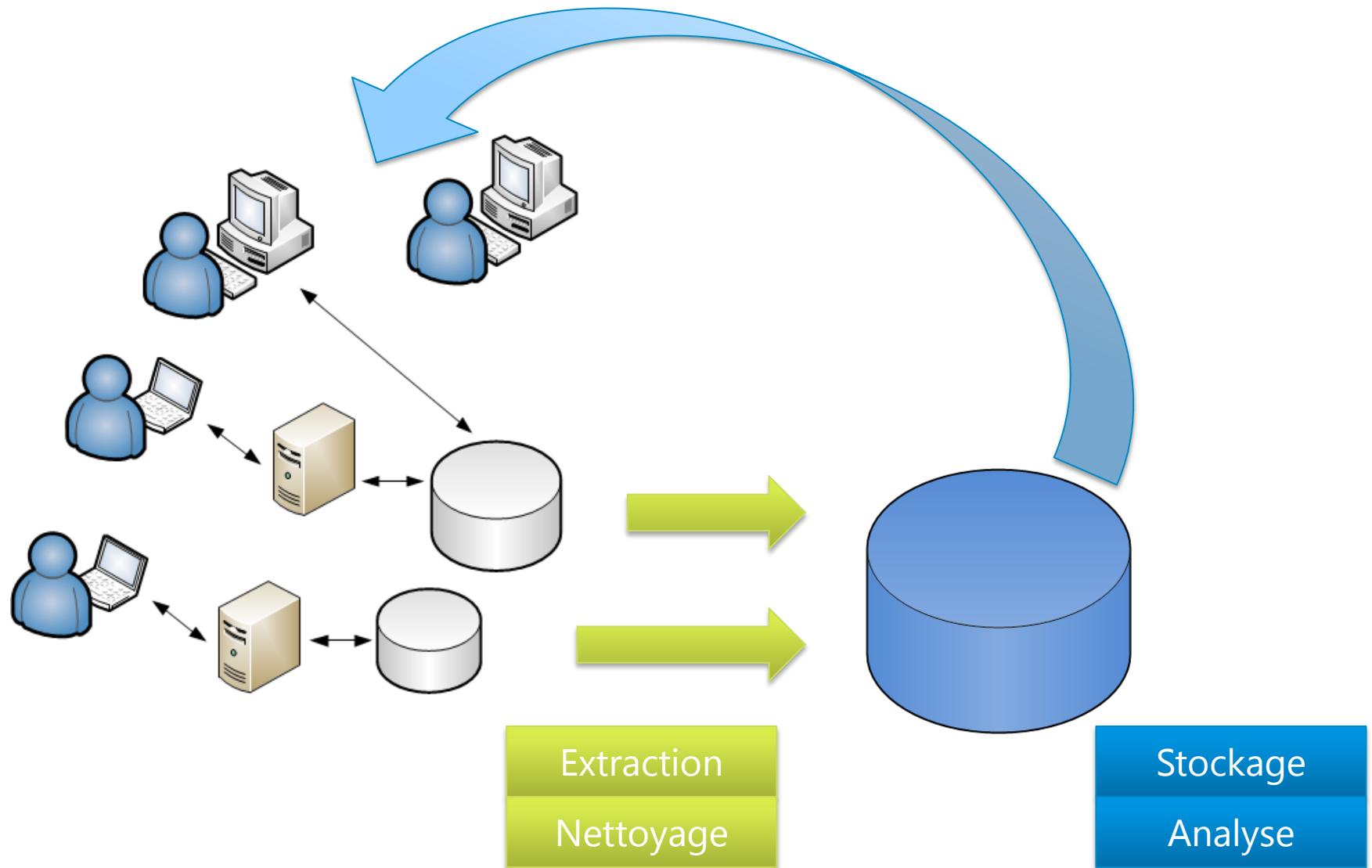
- Préparation
 - Extraction
 - Nettoyage
 - Stockage
 - Archivage
 - Historisation
- Présentation
 - Analyse
 - Reporting



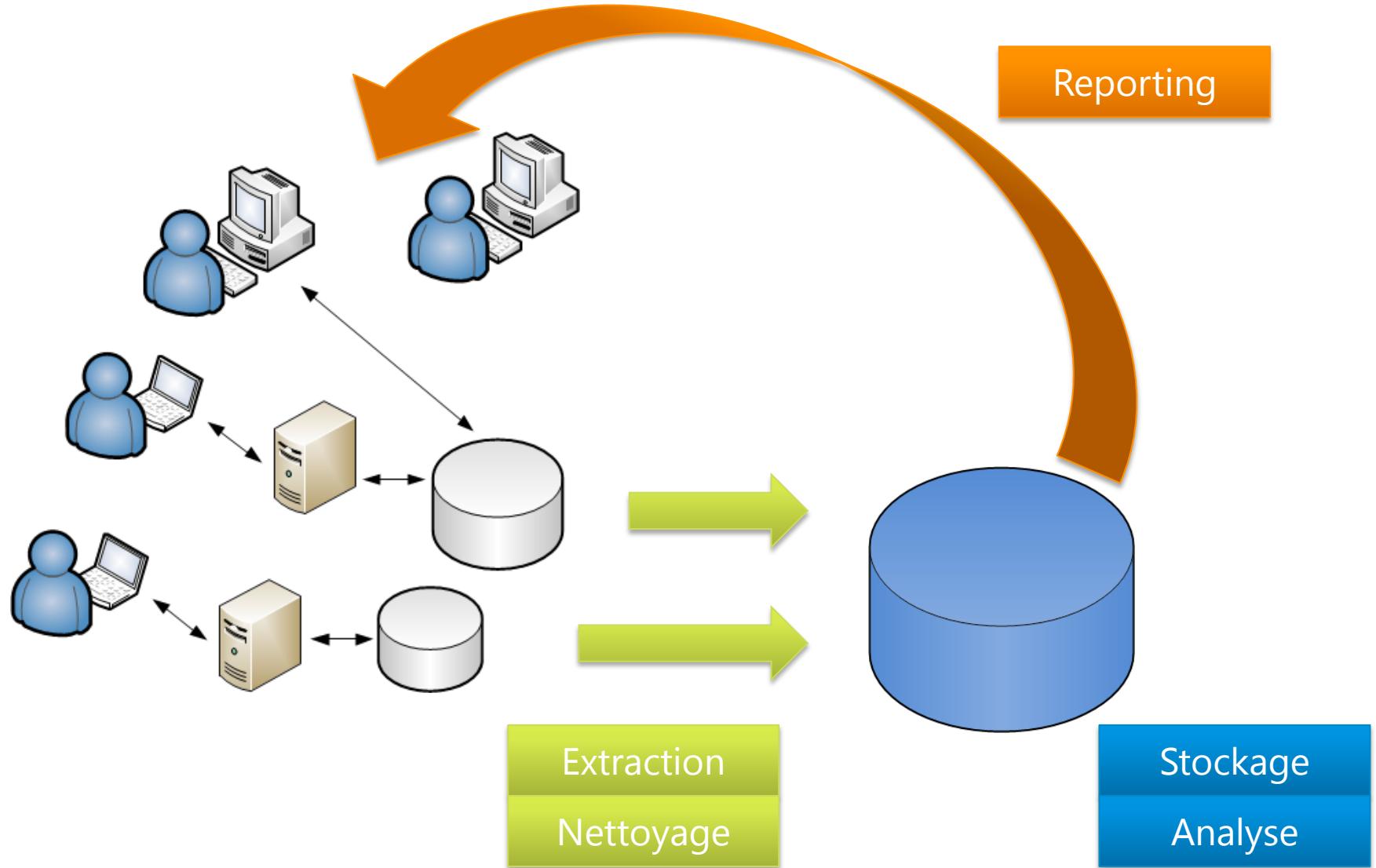
LES ÉTAPES DU DÉCISIONNEL



LES ÉTAPES DU DÉCISIONNEL



LES ÉTAPES DU DÉCISIONNEL



MÉTAPHORE DU RESTAURANT

Préparation : Cuisine

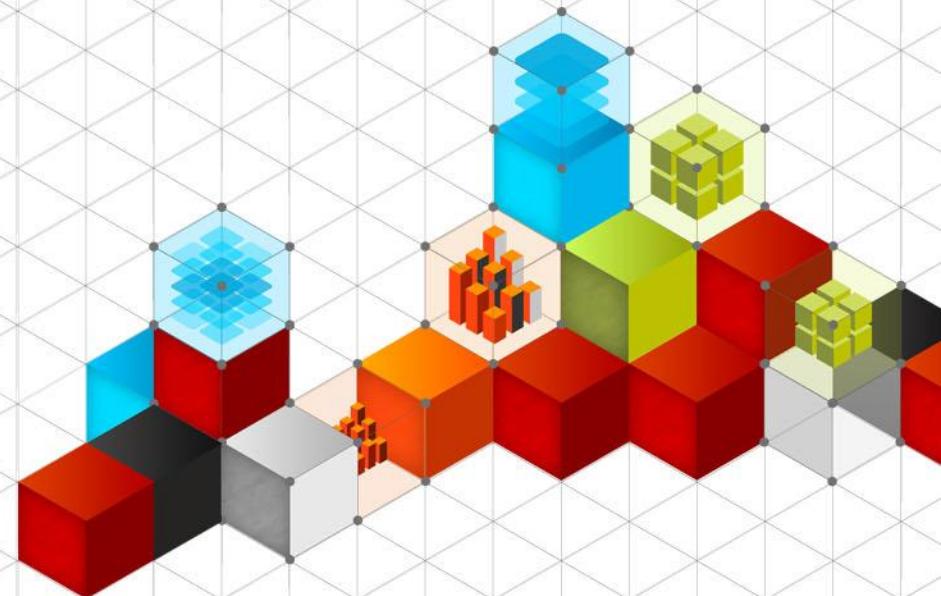
- Objectifs : préparer, transformer
- Caractéristiques : rigueur, constance



Présentation : Salle à manger

- Objectif : consommer
- Caractéristiques : beau et bon

APPROCHES ACADÉMIQUES



Les journées
SQL Server

GUSS
GROUPE DES UTILISATEURS
FRANÇAIS DE
MICROSOFT SQL SERVER

Microsoft®

LES APPROCHES ACADEMIQUES

- Les 2 grandes philosophies actuelles



	Kimball	Inmon
Processus	Bottom-Up	Top-Down
Organisation	Datamarts	Datawarehouse
Schématisation	Etoile	Flocon

- Une méthode alternative : Data Vault

PROCESSUS

Que chacun construise ce qu'il veut, on intégrera ce qu'il faudra quand il faudra!

On ne fait rien tant que tout n'est pas désigné, le datawarehouse doit être exhaustif!



Bill Inmon

Corporate Information Factory

www.inmoncif.com

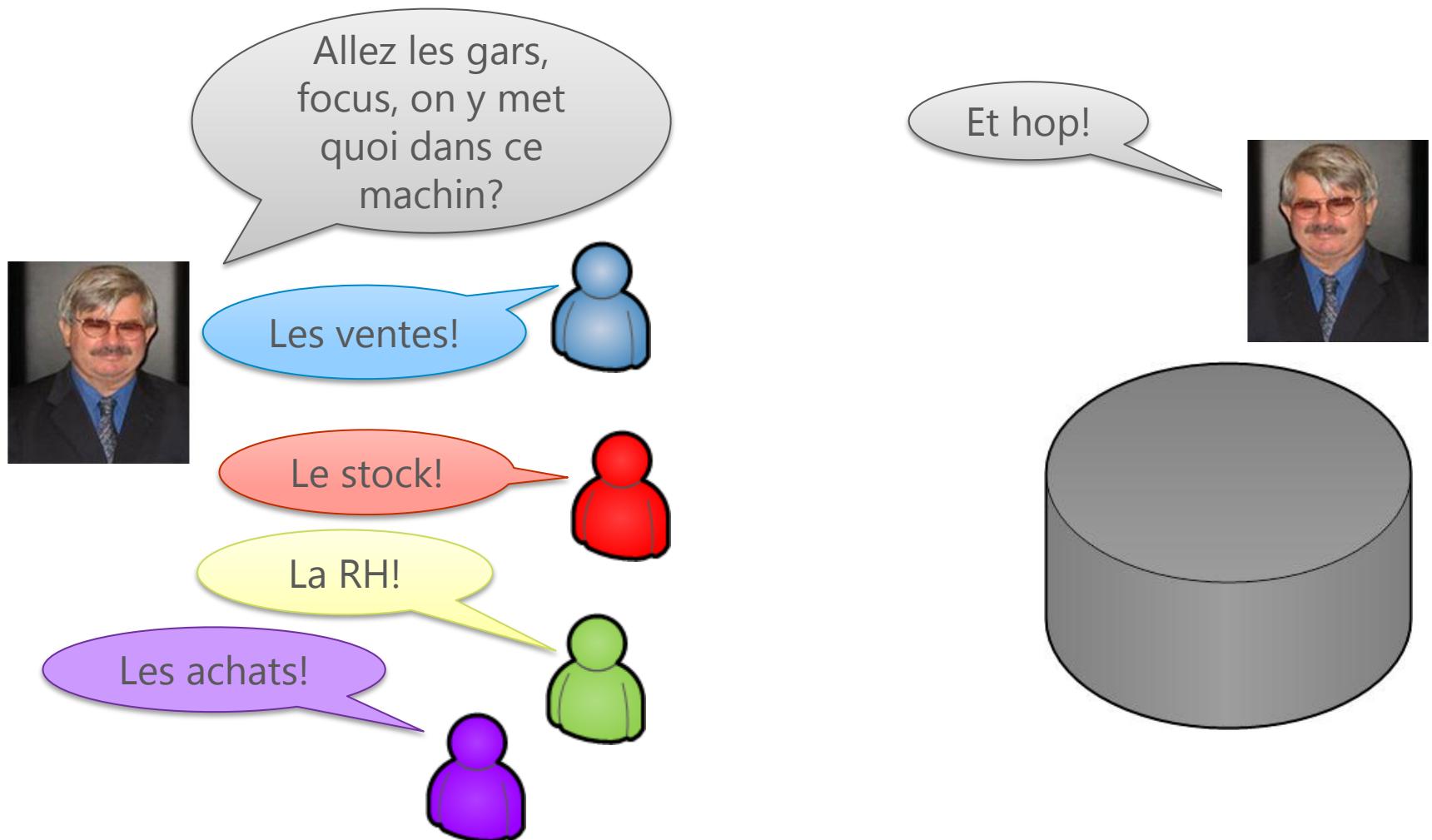


Ralph Kimball

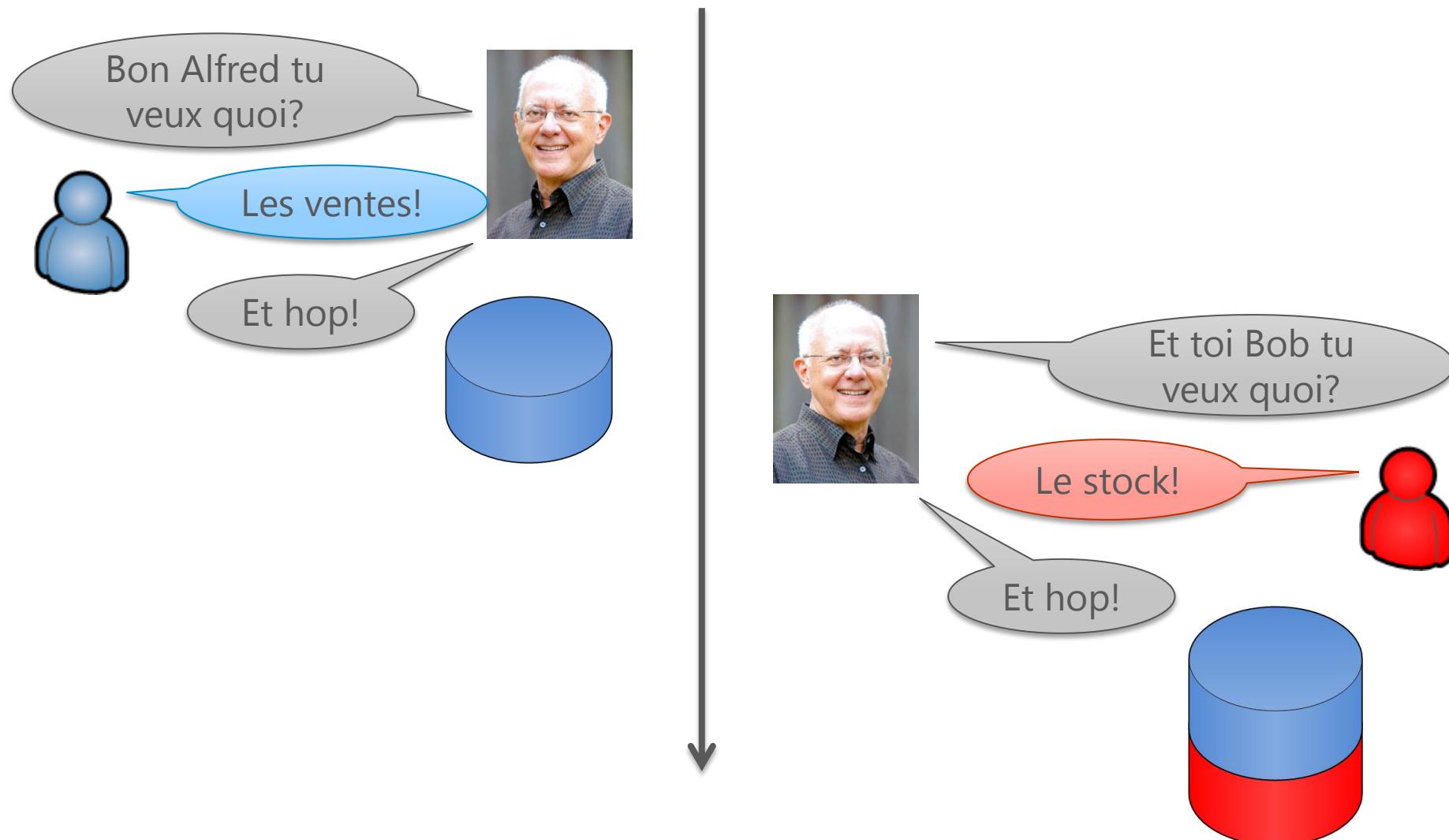
Kimball Group

www.kimballgroup.com

INMON : TOP DOWN



KIMBALL: BOTTOM UP



LE POINT COMMUN?



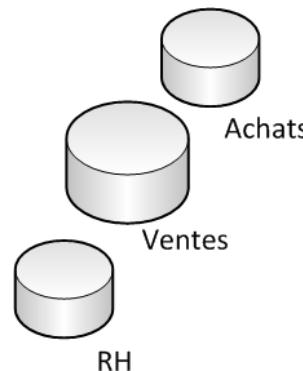
Allez **les gars**,
focus, on y met
quoi dans ce
machin?

Bon **Alfred** tu
veux quoi?

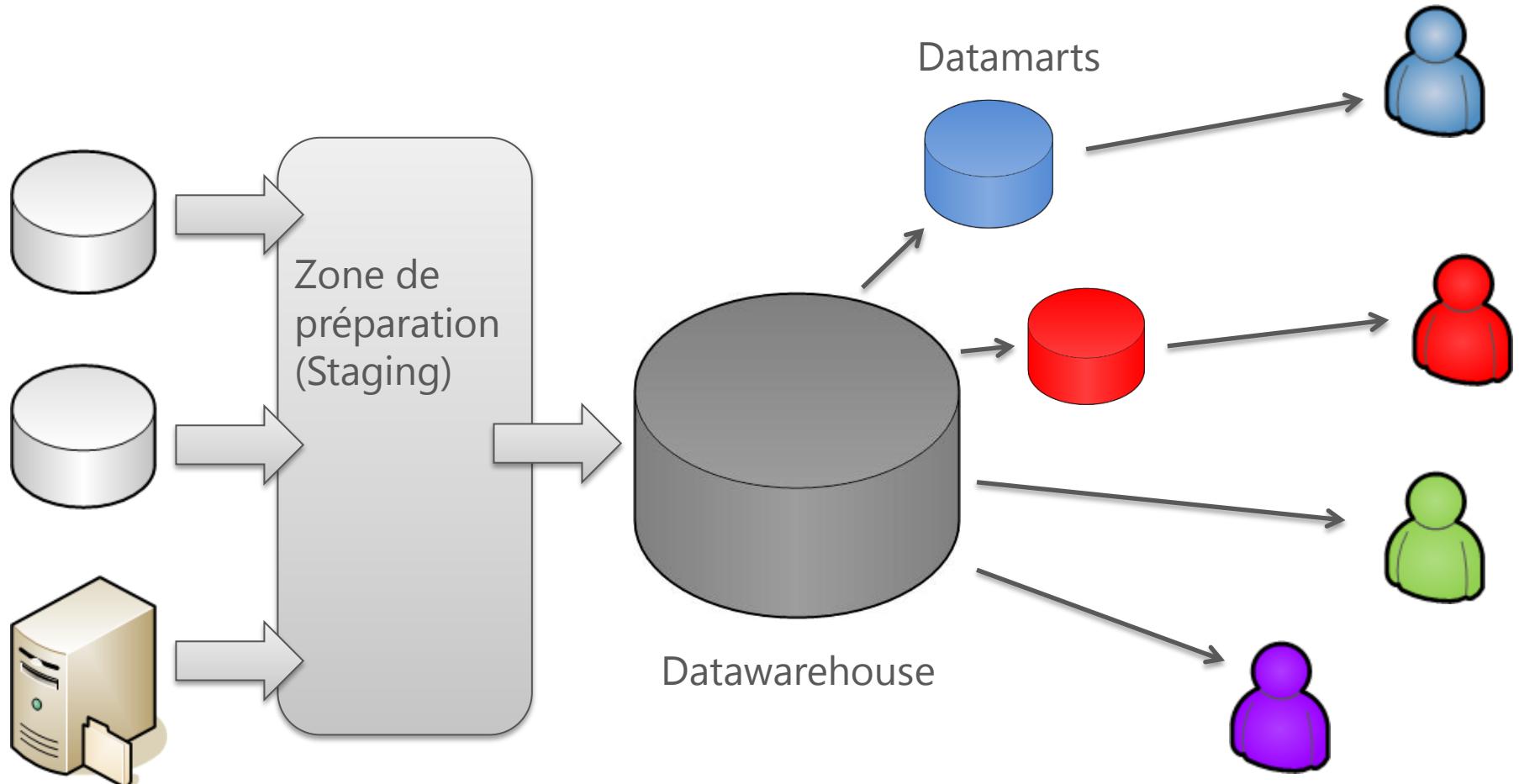


Et pas :

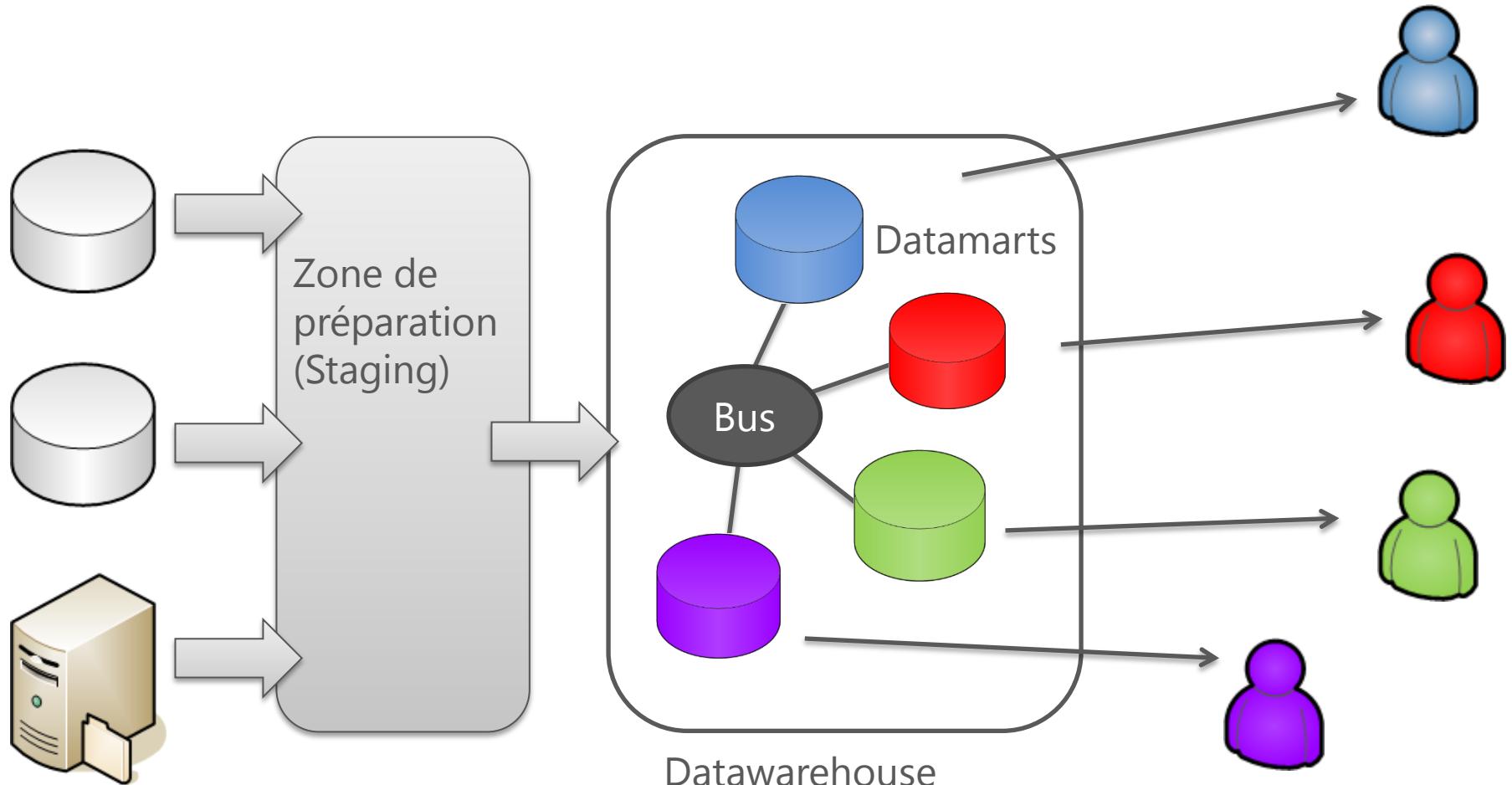
Hum, voyons ce
qu'il y a là dedans



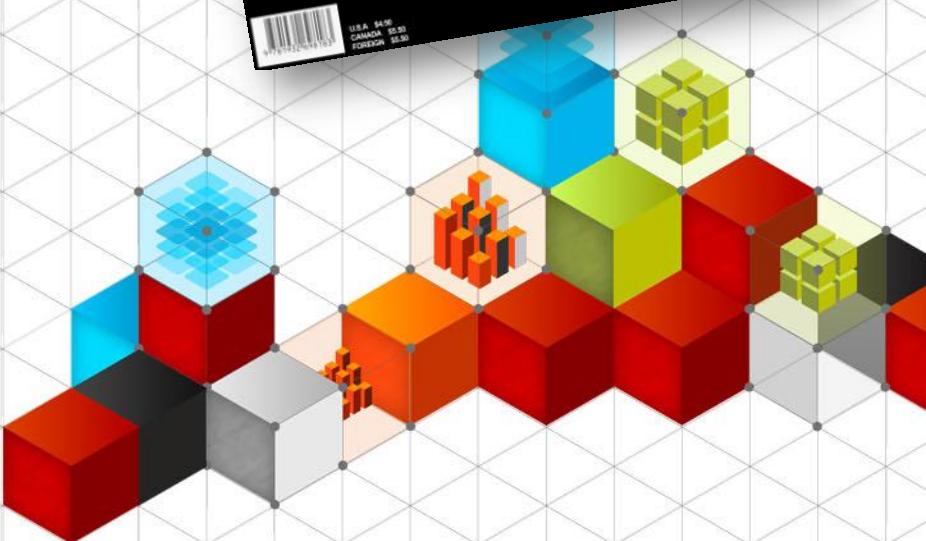
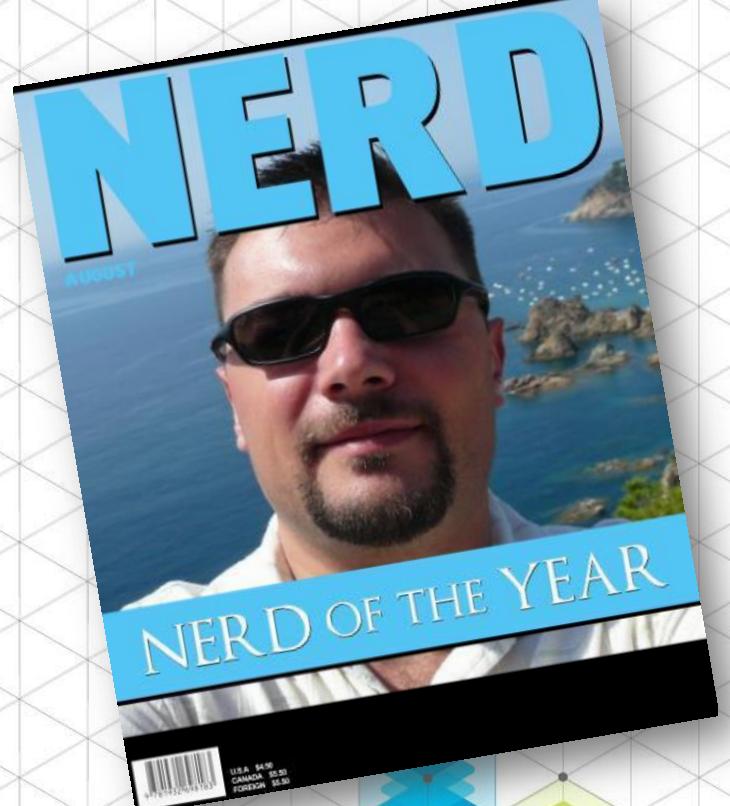
ORGANISATION : INMON



ORGANISATION : KIMBALL



LA MINUTE HYPE



Les journées
SQL Server

GUSS
GROUPE DES UTILISATEURS
FRANÇAIS DE
MICROSOFT SQL SERVER

Microsoft®

LES TENDANCES

Big Data

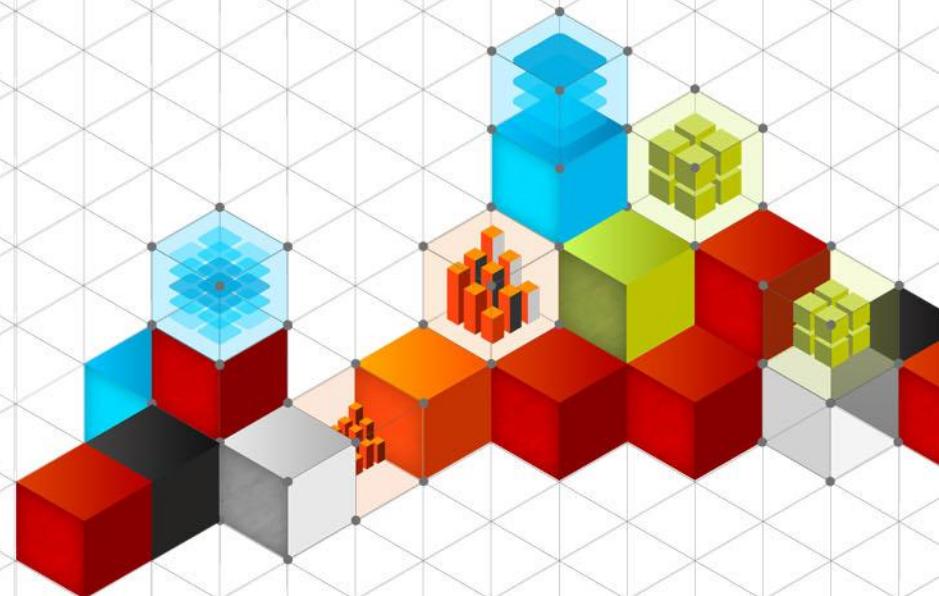
- **HADOOP** : Framework de stockage parallèle de données massives et pas forcément structurées
- **HIVE** : Framework de requêtage SQL sur HADOOP
- Limité, Design to Fail, peu d'outils...
- ...Mais des données à prendre en compte

NoSQL

- « **Not Only SQL** »
- Corollaire du Big Data

Microsoft est dans la course et des solutions arrivent

LA SCHÉMATISATION



Les journées
SQL Server

GUSS
GROUPE DES UTILISATEURS
FRANÇAIS DE
MICROSOFT SQL SERVER

Microsoft®

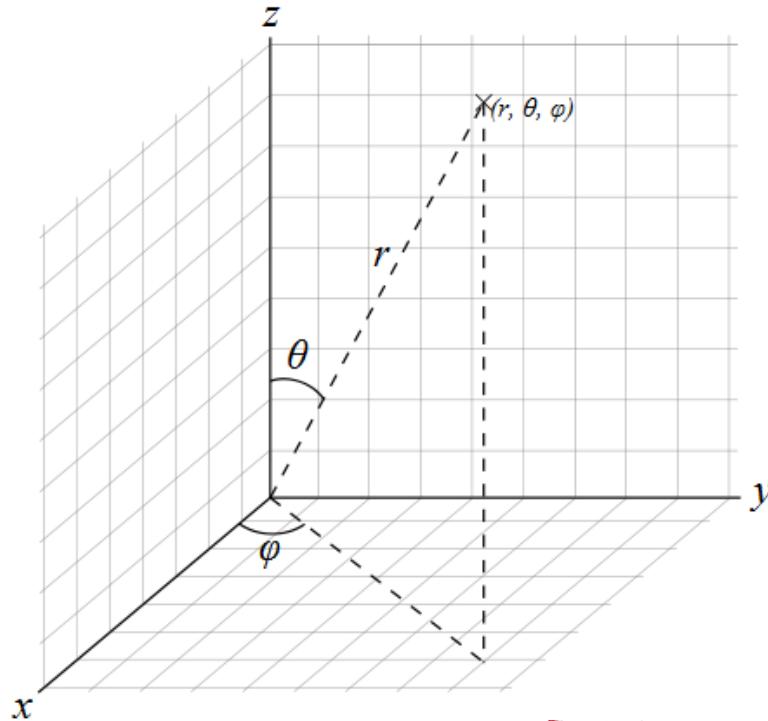
LE DÉCISIONNEL EN 3 MOTS

1 : La mesure



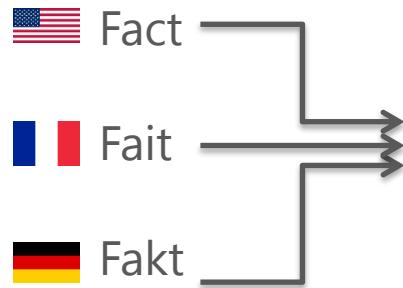
LE DÉCISIONNEL EN 3 MOTS

2 : Les dimensions



LE DÉCISIONNEL EN 3 MOTS

3 : Les faits



Factum : Acte, événement

Il s'est passé quelque chose, et on l'a mesuré selon notre référentiel, nos dimensions

ON RANGE

- **Un fait**, c'est une ligne, dans une table de faits

Table de Faits

Il s'est passé quelque chose

Il s'est passé autre chose

Il s'est passé quelque chose d'autre

ON ORDONNE

- **Les dimensions** donnent le contexte du fait

Table de Faits		
Quand	Où	
Hier	Ici	Il s'est passé quelque chose
Hier	Là bas	Il s'est passé autre chose
Aujourd'hui	Ici	Il s'est passé quelque chose d'autre

ON COMPTE

- **Les mesures** donnent les valeurs numériques du fait

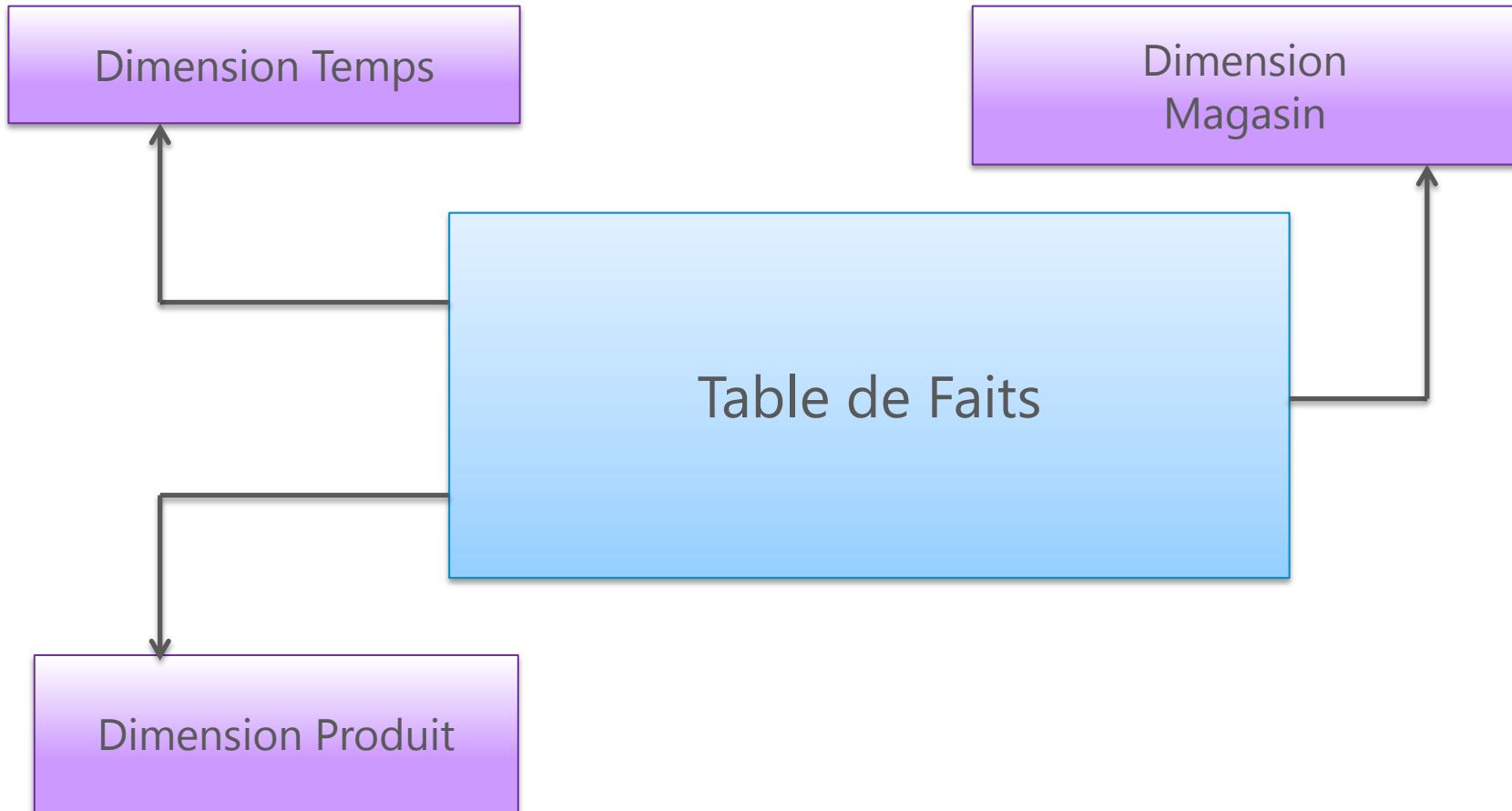
Table de Faits : Ventes				
Date	Magasin	Produit	Combien	Prix Unitaire (€)
Hier	AGY	Jouet	3	10
Hier	BGH	Saucisson	2	2,5
Aujourd'hui	AAZ	Parapluie	5	5

PROBLÉMATIQUE

- Mes dimensions ont des attributs
 - Hier
 - Date : 12/12/2011
 - Jour : Lundi
 - Jour de la semaine : 1er
 - Mois : Décembre
 - Trimestre : 4ème
 - Année : 2011
 - ...
- Mes dimensions sont réutilisables
 - Mêmes produits à l'achat, la vente et l'inventaire!

SOLUTION

- Chaque dimension devient une table



PREMIER ÉLÉMENT DE DESIGN

Dimension Temps			
Id Temps	Jour	Mois	Année
20111212	12/12/2011	Décembre	2012
20111213	13/12/2011	Décembre	2012

Dimension Magasin		
Id Magasin	Code Magasin	Région
22	AAZ	Ile de France
35	AGY	Ile de France
56	BGH	Lorraine

Table de Faits : Ventes

Id Temps	Id Magasin	Id Produit	Quantité	Prix Unitaire (€)
20111212	35	5	3	10
20111212	56	8	2	25
20111212	22	12	5	5
20111212	35	5	1	10
20111213	56	5	6	15
20111213	56	8	7	20
20111213	22	12	2	5

Dimension Magasin			
Id Produit	Code Produit	Nom	Catégorie
5	Z947342983	Lapin Malin	Jouet
8	Z238943135	Saucisson sec	Alimentaire
12	E238953123	Parapluie	Accessoire

PROBLÉMATIQUE

- Un magasin a des attributs uniques
 - Un code
 - Un nom
 - Une adresse
 - ...
- Il a aussi des attributs communs à d'autres
 - Région
 - Pays
 - Catégorie
 - Enseigne
 - ...

ATTRIBUTS PARTAGÉS

- Premier réflexe :

Dimension Magasin					
Id Magasin	Code Magasin	...	Région	Pays	...
22	AAZ	...	Ile de France	France	...
35	AGY	...	Ile de France	France	...
56	BGH	...	Lorraine	France	...
67	BHJ	...	Lorraine	France	...
78	BKJ	...	Lorraine	France	...
79	BKL	...	Alsace	France	...
95	ZDE	...	Hérault	France	...
96	ZRT	...	Hérault	France	...
105	ZYU	...	Hérault	France	...

- Oui mais mon dba n'aime pas quand je répète:
 - Standard des bases de données : Normalisation 3NF!

NORMALISATION

- Dimension normalisée

Dimension Magasin N1			
Id Magasin N1	Code Magasin	...	Id Magasin N2
22	AAZ	...	1
35	AGY	...	1
56	BGH	...	5
67	BHJ	...	5
78	BKJ	...	5
79	BKL	...	6
95	ZDE	...	21
96	ZRT	...	21
105	ZYU	...	21

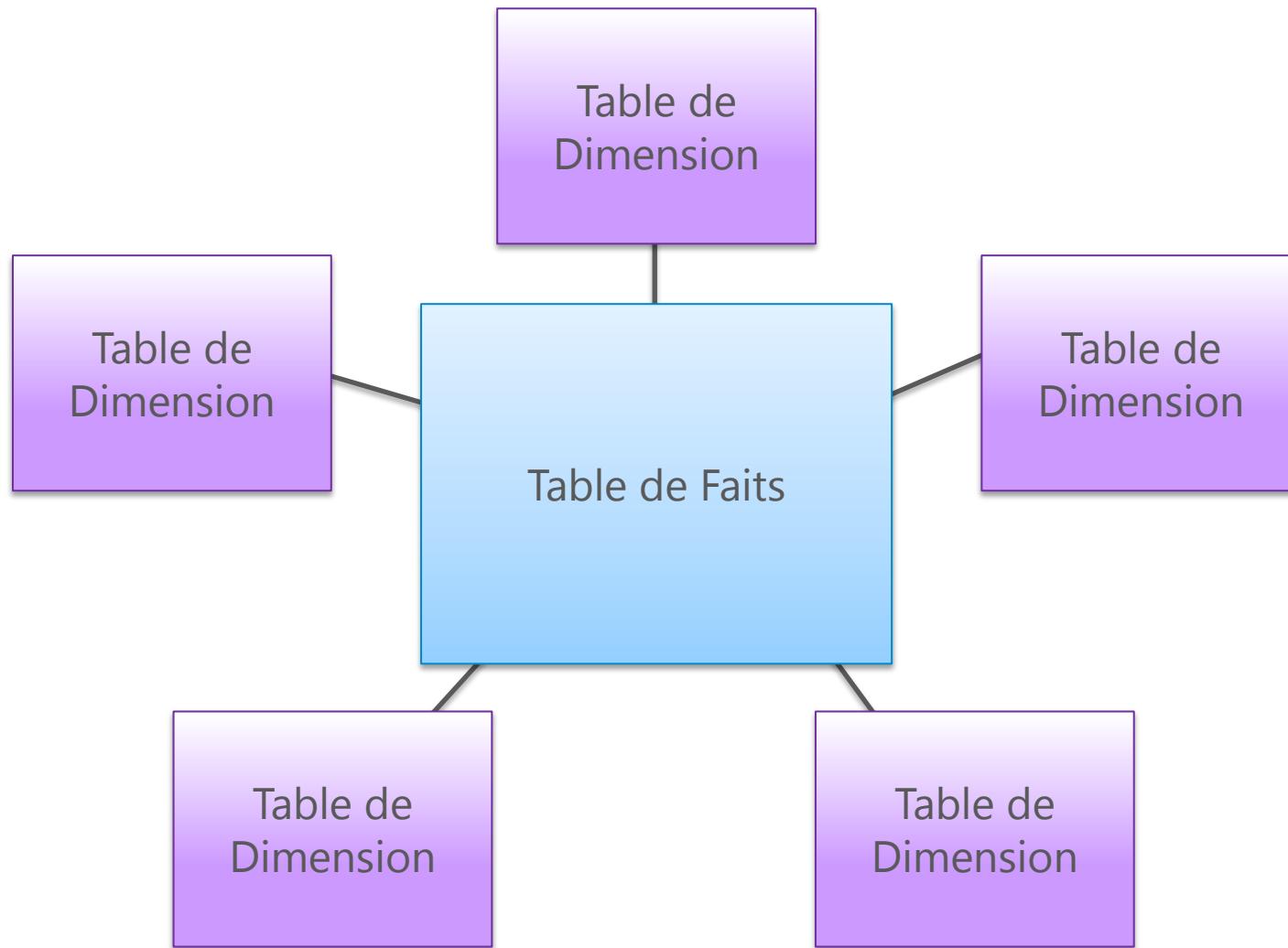
Dimension Magasin N2			
Id Magasin N2	Région	...	Id Magasin N3
1	Ile de France	...	1
5	Lorraine	...	1
6	Alsace	...	1
21	Hérault	...	1

Dimension Magasin N3		
Id Magasin N3	Pays	...
1	France	...
2	Allemagne	...

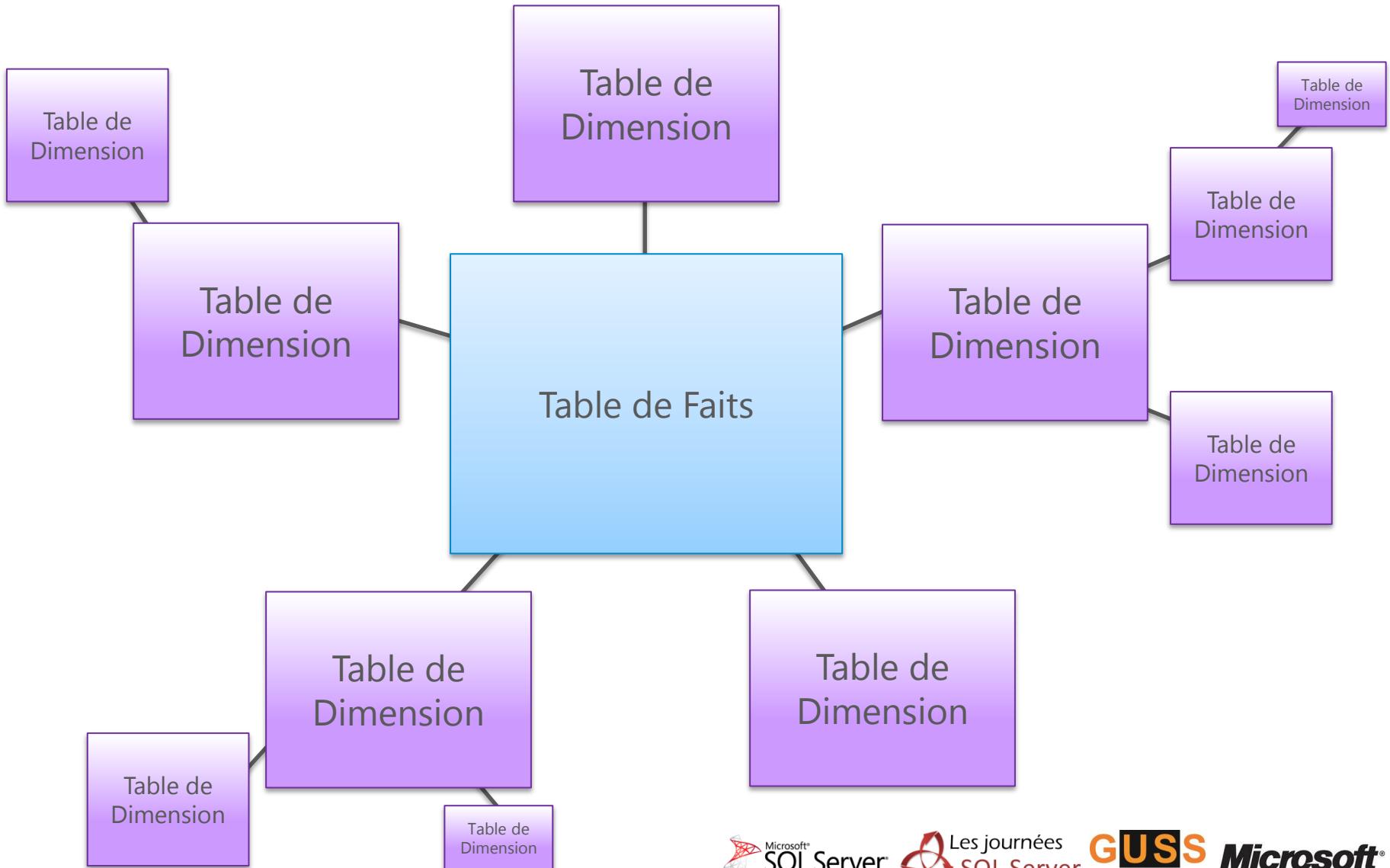


- Cette question, c'est Flocon vs Etoile!
 - 1 table par niveau de hiérarchie?
 - 1 table par dimension?

KIMBALL : SCHÉMA EN ÉTOILE



INMON : SCHÉMA EN FLOCON



SCHÉMATISATION

Etoile

DimOrganisation			
ID	Orga Niveau 3	Orga Niveau 2	Orga Niveau 1
1	Paris	France	Europe
2	Lyon	France	Europe
3	Berlin	Allemagne	Europe

- Avantage : lisibilité et performance des requêtes

```
SELECT  
    [Orga Niveau 3]  
    ,[Orga Niveau 2]  
    ,[Orga Niveau 1]  
FROM DimOrganisation
```

```
Query 1: Query cost (relative to the batch): 100%  
SELECT * FROM TRW_DWH_Dim_Categorie
```

```
SELECT  
Cost: 0 %  
Clustered Index Scan (Clustered)  
[TRW_DWH_Dim_Categorie].[PK_TRW_DWH..  
Cost: 100 %
```

Flocon

DimOrganisationN3		
ID	Orga Niveau 3	ID Orga N2
1	Paris	26
2	Lyon	26
3	Berlin	87

DimOrganisationN2		
ID	Orga Niveau 2	ID Orga N1
26	France	2
87	Allemagne	2

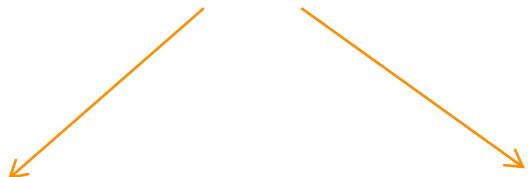
DimOrganisationN1	
ID	Orga Niveau 1
2	Europe

- Avantage : Espace disque
- Avantage : Performance des updates

ETOILE OU FLOCON?

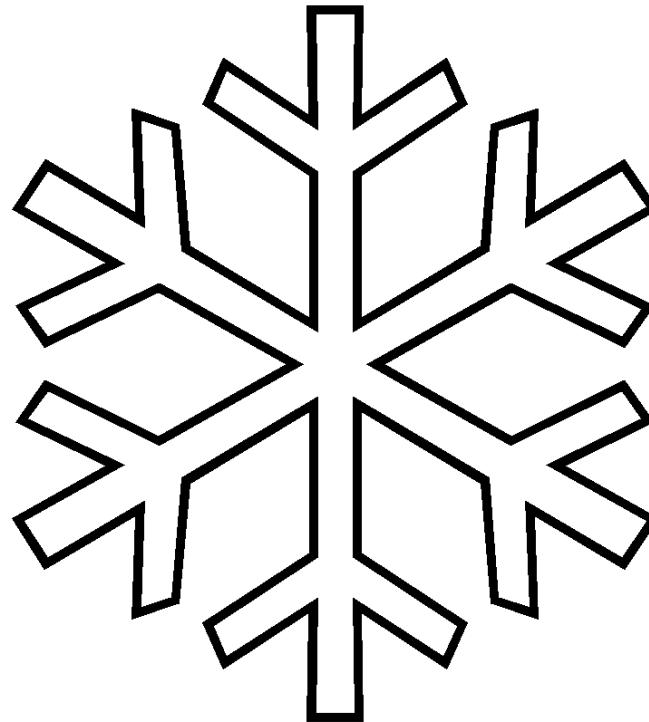
- Etoile !
 - Plus simple à alimenter et maintenir
 - Les outils Microsoft ont une sensibilité Kimball

Notez la ressemblance



QUAND FLOCONNER?

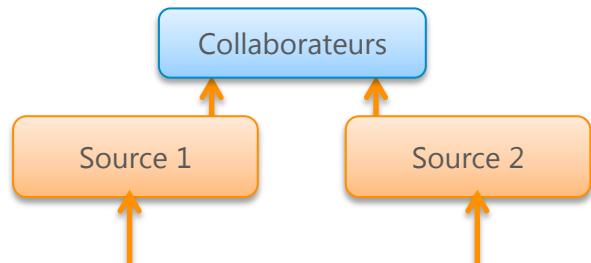
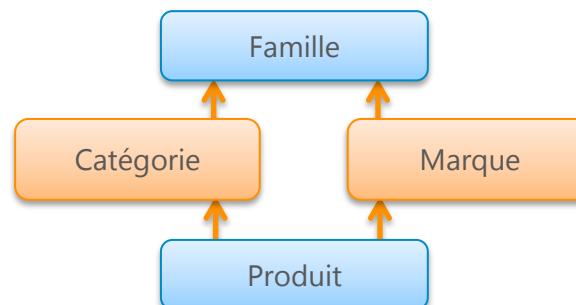
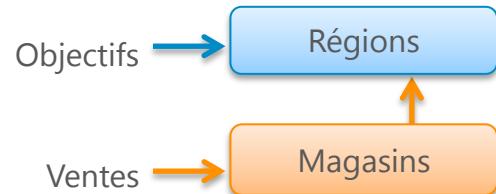
- Si on préfère les flocons aux étoiles
 - Mieux vaut jouer avec des outils qu'on aime!



QUAND FLOCONNER?

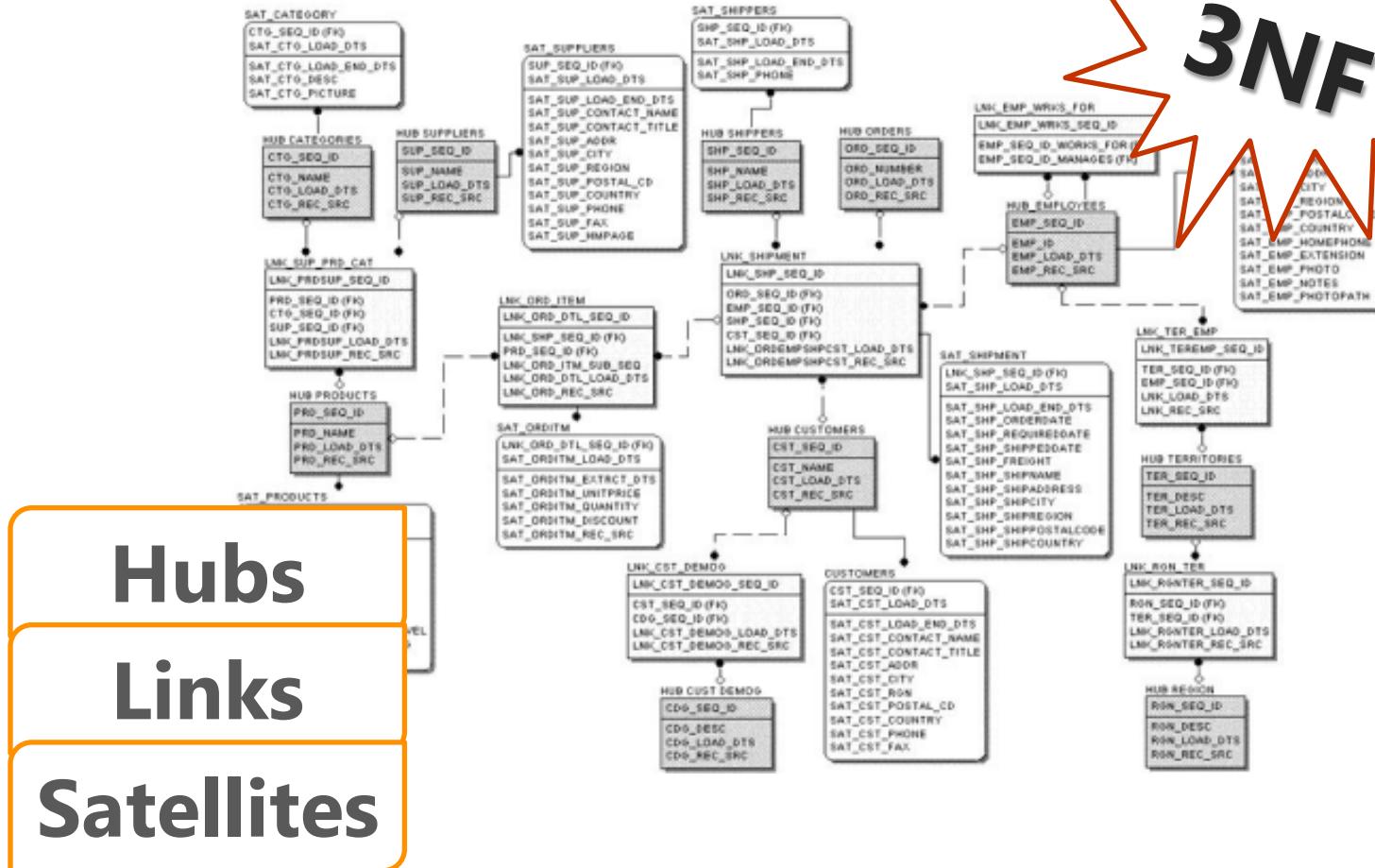
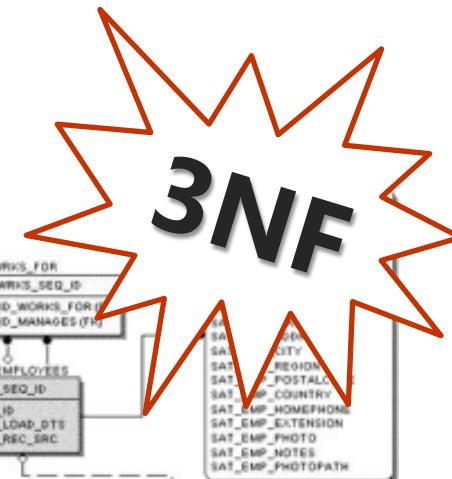
- Si on est en étoile :

- Des faits arrivent à différents niveaux de la même dimension
- Hiérarchies de la même dimension qui peuvent agréger sur des membres supérieurs différents
- Niveaux fins différents pour une même dimension

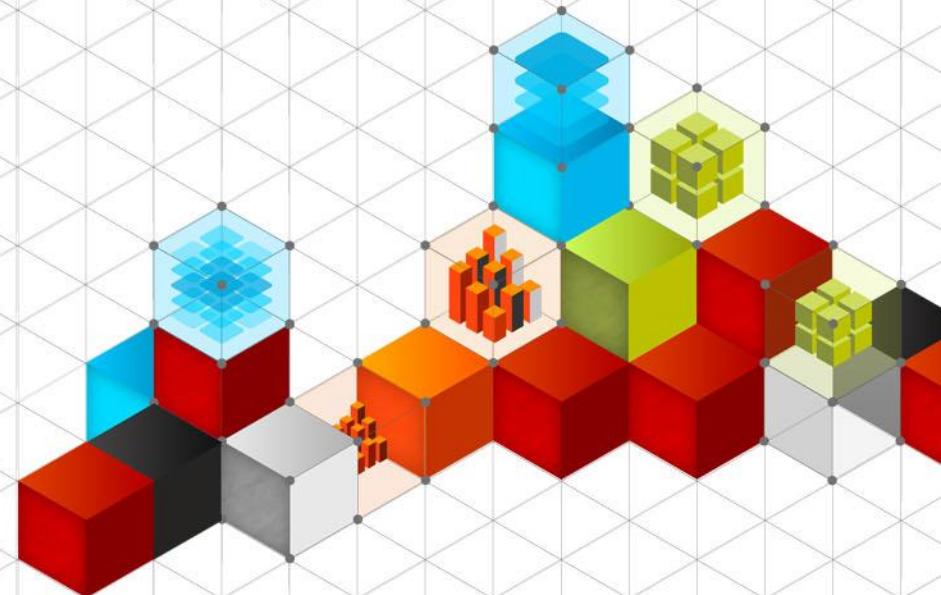


SCHÉMATISATION

Data Vault



LE PROCESSUS DE MODÉLISATION



Les journées
SQL Server

GUSS
GROUPE DES UTILISATEURS
FRANÇAIS DE
MICROSOFT SQL SERVER

Microsoft®

PROCESSUS DE MODÉLISATION

Sponsor

1

2

3

4

- 0 : Le sponsor

Identifier l'utilisateur, celui qui sait, celui a le dernier mot, sinon ça ne sert à rien



- 1 : On modélise quoi ? Quel processus métier?

> **Quel acte est mesuré?**

- Une transaction?
 - Une vente, une commande, un voyage...
- Un stock?
 - **Un inventaire**, une position financière...



• 2 : Déclarer le grain

- > Que représente la ligne dans la table de fait?
- 1 ligne d'un ticket de caisse?
- Le total du ticket de caisse?
- Le stock en fin de semaine du magasin?

- Dans notre cas : comment considère t'on le « où »:
- Dans un rack?
- Dans une rangée?
- **Dans l'entrepôt de quel magasin?**





BURGERVILLE
Better taste comes from better food.
NUTRIMATE receipt (nu-tri-KAT) v. To nutritionally educate.

Qty	Item	Price	Calories	Fiber(g)	Fat(g)	Carbs(g)
1	HALIBUT SANDWICH BASKET	\$8.48	-	-	-	-
	Halibut Sandwich		488	2	27	43
	No Tartar Sauce		-130	0	-14	-1
	Reg. Sweet Potato Frie		530	11	24	75
	Reg Self-Serve Bev*		-	-	-	-
	NUTRITION TOTALS	888	13	37	117	
	% DAILY VALUE - 2000 CALORIES	44%	528	578	39%	
	% DAILY VALUE - 2500 CALORIES	36%	428	458	318	
1	TURKEY BURGER BASKET	\$7.09	-	-	-	-
	Turkey Burger		519	2	22	47
	No Mayo		-100	0	-11	0
	No Tomato		-4	0	0	-1
	Regular French Fries		360	3	15	52
	Reg Self-Serve Bev*		-	-	-	-
	NUTRITION TOTALS	775	5	26	98	
	% DAILY VALUE - 2000 CALORIES	39%	198	408	33%	
	% DAILY VALUE - 2500 CALORIES	31%	158	328	268	
	1 Reg Choc Monk Smoothie	\$3.99	470	4	2	104
	* Visit our website for nutrition info on this item.					
	Sub Total	\$19.56				
	TOTAL	\$19.56				
	VISA	\$19.56				
	PAID	\$19.56				
	Did You Know ?					
	For great-tasting desserts under 200 calories, try our Triple Berry Frozen Yogurt Sundae, a dish of Frozen Yogurt, or a Vanilla Frozen Yogurt Cone.					

- 3 : Choisir les dimensions qui s'appliquent
 - Réutiliser les dimensions disponibles

	Date	Produit	Magasin	Entrepôt	Fournisseur	Promotion
Forecasting	x		x			x
Achats	x	x		x	x	
Commandes	x	x	x			x
Livraisons	x		x	x		

- Inventaire en Magasin :
 - Date
 - Produit
 - Magasin

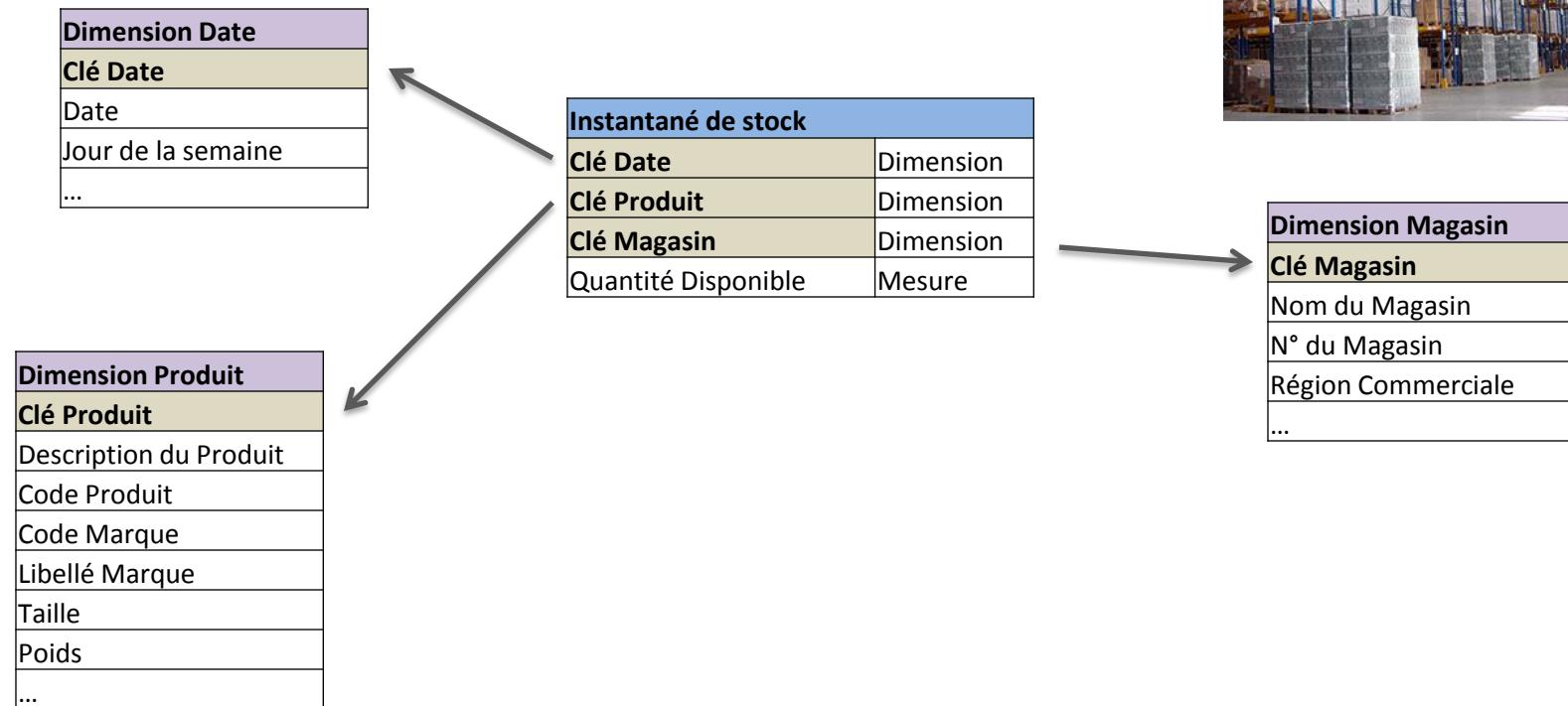
• 4 : Identifier les mesures

- Attention aux unités
 - Devise
 - Métrique
 - Conteneur
 - Palette?
 - Carton?
 - **Unité?**



PREMIER DATAMART

- Inventaire en magasin



DEUXIÈME DATAMART

- Ventes dans le magasin



Dimension Date
Clé Date
Date
Jour de la semaine
...

Dimension Produit
Clé Produit
Description du Produit
Code Produit
Code Marque
Libellé Marque
Taille
Poids
...

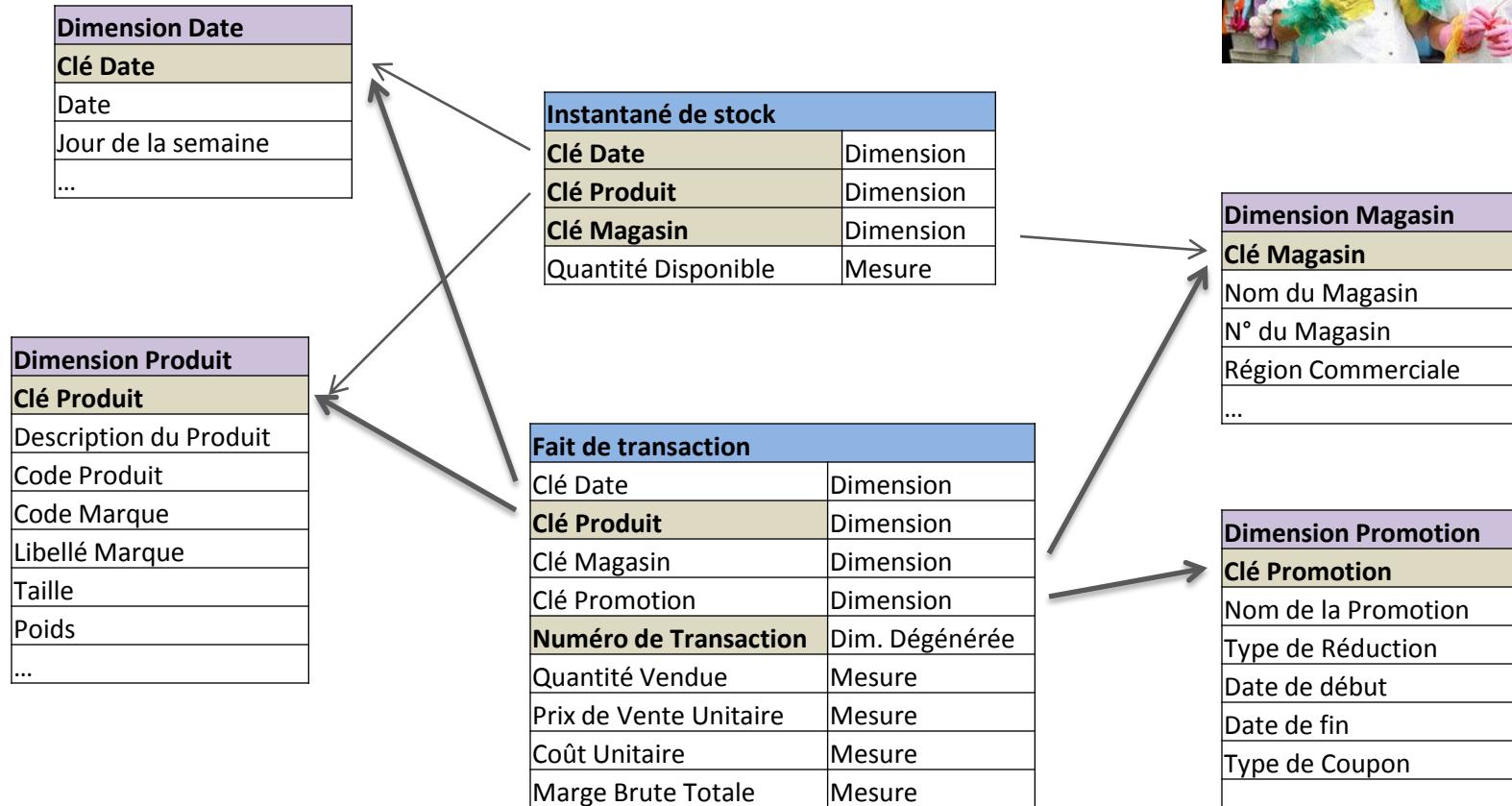
Fait de transaction	
Clé Date	Dimension
Clé Produit	Dimension
Clé Magasin	Dimension
Clé Promotion	Dimension
Numéro de Transaction	Dim. Dégénérée
Quantité Vendue	Mesure
Prix de Vente Unitaire	Mesure
Coût Unitaire	Mesure
Marge Brute Totale	Mesure

Dimension Magasin
Clé Magasin
Nom du Magasin
N° du Magasin
Région Commerciale
...

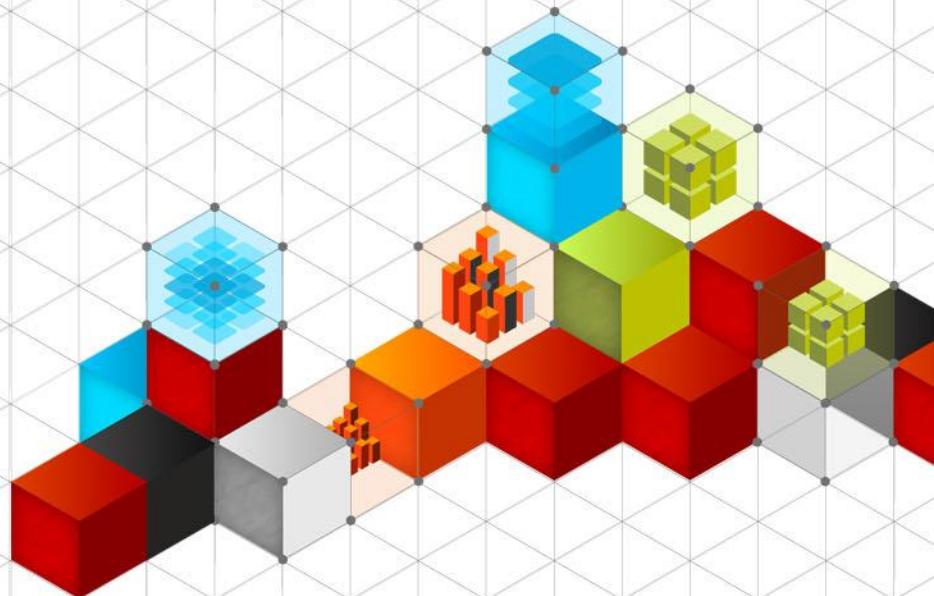
Dimension Promotion
Clé Promotion
Nom de la Promotion
Type de Réduction
Date de début
Date de fin
Type de Coupon
...

EXEMPLES FONCTIONNELS

- Et hop, un datawarehouse!



IMPLÉMENTATION DANS SQL SERVER



Les journées
SQL Server

GUSS
GROUPE DES UTILISATEURS
FRANÇAIS DE
MICROSOFT SQL SERVER

Microsoft®

SQL SERVER 2012

- Trois moteurs pour modéliser

Capacité de modélisation des technologies Microsoft	Etoile	Flocon	Data Vault
SQL Server Moteur Relationnel	5/5	5/5	5/5
BISM Multidimensionnel (SSAS)	5/5	4/5	0/5
BISM Tabular (PowerPivot)	5/5	4/5	2/5

AGENDA

**1. Modélisation
dimensionnelle**

**2. Concepts
avancés**

01.15

CONCEPTS AVANCÉS

Semi-additivité

Gestion des NULL

Pondération

Distinct Count

Gestion des devises

SCD

Temporal Snapshot

Surrogate Keys

Factless Tables

Junk Dimension

Ragged Hierarchies

Many-to-Many

Sécurité dynamique

SCD : SLOWLY CHANGING DIMENSIONS

- Besoin
 - Les dimensions ne sont pas immuables : les attributs peuvent changer
- Dimensions à évolution lente
 - 3 Techniques
 - **Type 1** : On écrase

Clé Produit	Code Produit	Nom	Rayon
29	AA-56	Lapin Malin	Jouet

UPDATE DimProduit SET...



Clé Produit	Code Produit	Nom	Rayon
29	AA-56	Lapin Malin	Boucherie

- 2 types de clefs
 - Naturelle / Business, c'est le code que comprend le métier
 - Technique /« Surrogate », c'est une clef propre au DWH qui ne doit jamais apparaître sur un rapport
- Le type de SCD est un attribut de colonne, pas de dimension, ici c'est le rayon

SCD : SLOWLY CHANGING DIMENSIONS

- Dimensions à évolution lente
 - 3 Techniques...
 - **Type 2** : Nouvelle ligne

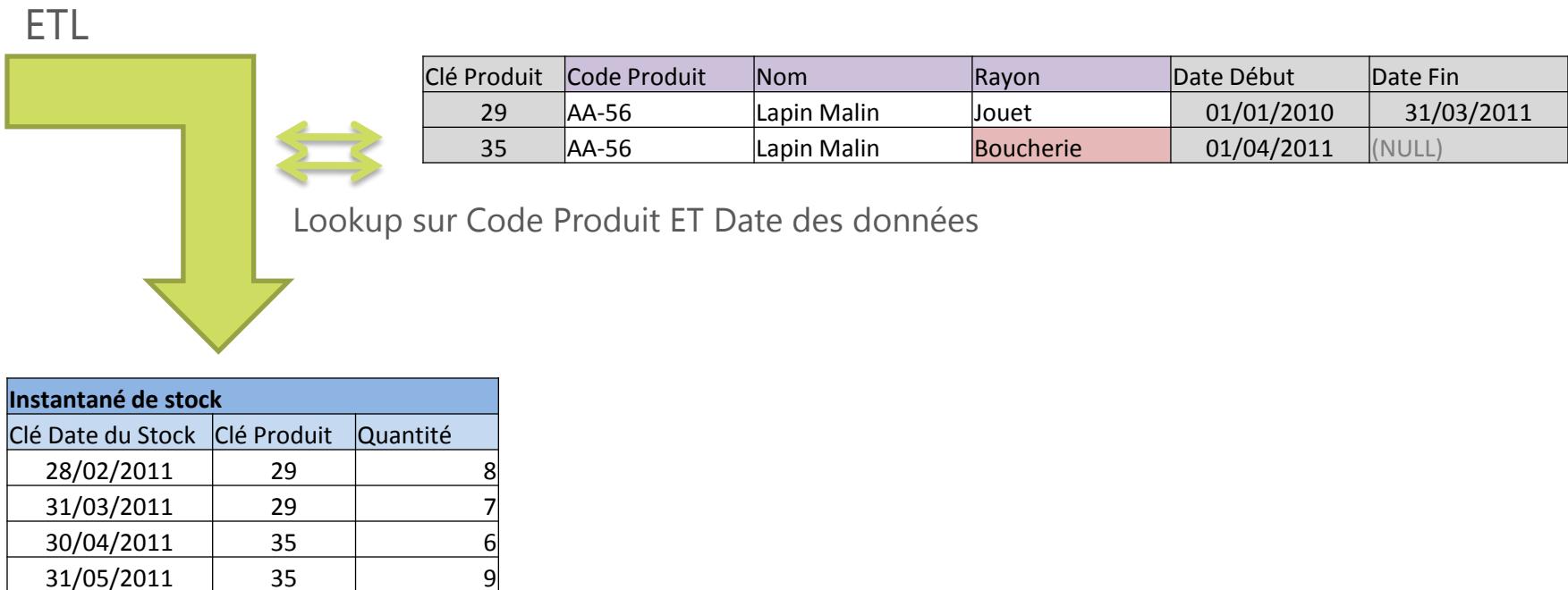
Clé Produit	Code Produit	Nom	Rayon	Date Début	Date Fin
29	AA-56	Lapin Malin	Jouet	01/01/2010	(NULL)

UPDATE DimProduit SET...
INSERT INTO DimProduit...

Clé Produit	Code Produit	Nom	Rayon	Date Début	Date Fin
29	AA-56	Lapin Malin	Jouet	01/01/2010	31/03/2011
35	AA-56	Lapin Malin	Boucherie	01/04/2011	(NULL)

SCD : SLOWLY CHANGING DIMENSIONS

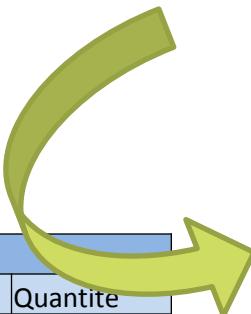
- Dimensions à évolution lente
 - 3 Techniques...
 - **Type 2** : Nouvelle ligne



SCD : SLOWLY CHANGING DIMENSIONS

- Dimensions à évolution lente
 - 3 Techniques...
 - **Type 2** : Nouvelle ligne

Clé Produit	Code Produit	Nom	Rayon	Date Début	Date Fin
29	AA-56	Lapin Malin	Jouet	01/01/2010	31/03/2011
35	AA-56	Lapin Malin	Boucherie	01/04/2011	(NULL)



Instantané de stock		
Clé Date du Stock	Clé Produit	Quantité
28/02/2011	29	8
31/03/2011	29	7
30/04/2011	35	6
31/05/2011	35	9

Somme de Quantité Étiquettes de lignes	Étiquettes de colonnes				
	02 Février	03 Mars	04 Avril	05 Mai	Total général
Boucherie			6	9	15
Lapin Malin			6	9	15
Jouet	8	7			15
Lapin Malin	8	7			15
Total général	8	7	6	9	30

SCD : SLOWLY CHANGING DIMENSIONS

- Dimensions à évolution lente
 - 3 Techniques...
 - **Type 3** : Nouvelle colonne

Clé Produit	Code Produit	Nom	Rayon Actuel	Ancien Rayon
29	AA-56	Lapin Malin	Jouet	(NULL)

UPDATE DimProduit SET...

Clé Produit	Code Produit	Nom	Rayon Actuel	Ancien Rayon
29	AA-56	Lapin Malin	Boucherie	Jouet

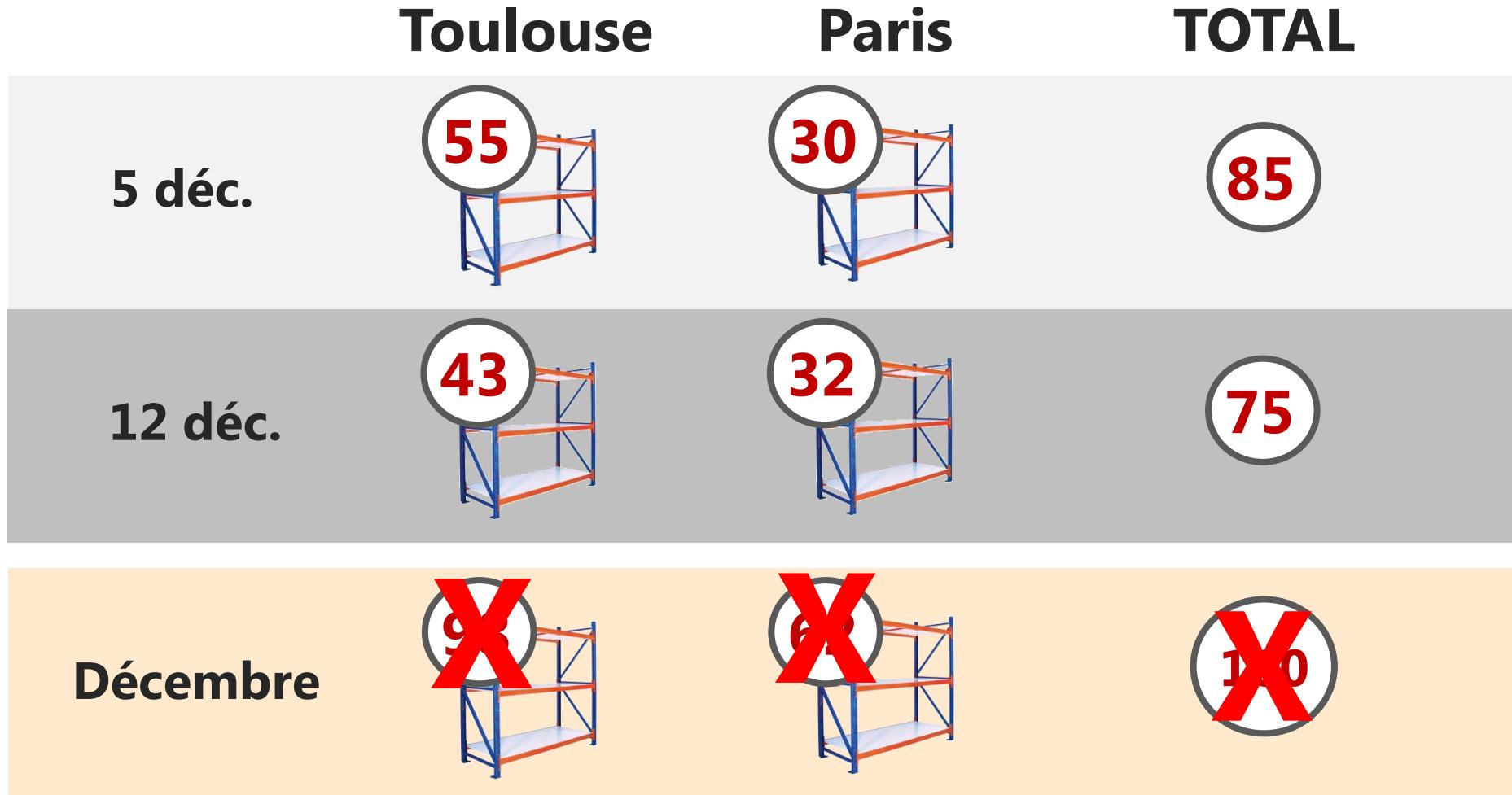
- Et les **hybrides**

Clé Produit	Code Produit	Nom	Rayon	Ancien Rayon	Date Début	Date Fin
29	AA-56	Lapin Malin	Jouet	(NULL)	01/01/2010	31/03/2011
35	AA-56	Lapin Malin	Boucherie	Jouet	01/04/2011	(NULL)

- Et l'évolution rapide ? Si la granularité temporelle des changements rejoint celle des tables de fait. Il faut créer une nouvelle dimension contenant les attributs concernés: tout rentre dans l'ordre!

SEMI-ADDITIVITÉ

Problématique : analyse de stock



SEMI-ADDITIVITÉ

Solution :

Changer la formule d'agrégation

- Faire le calcul dans les outils de Reporting ?
- Faire un membre calculé en MDX ?
- **Directement dans la mesure dans SSAS**

Attention : Edition Enterprise uniquement

GESTION DES NULL

« gestion de l'absence de valeur »

Problématique :

- Un fait référence un membre de la dimension qui n'existe pas
 - Ex : une vente remonte sur un nouveau produit
 - dans la base : NULL ou FK invalide

Par défaut : **erreur** au traitement du cube

GESTION DES NULL

Solution :

- Mise en place de l'**Unknown Member**

Alternatives :

- Création de membre « spéciaux »
- Traitement dans l'ETL

Problématique de **Qualité des données**

MANY-TO-MANY

Problématique :

- Un fait est rattaché à plusieurs membres d'une dimension
 - Ex : promotion sur une vente (fidélité, remise, etc.)

Pistes :

- En relationnel, on passe par une table d'association

MANY-TO-MANY

Solution :

- On met une table pour faire le pont entre les ventes et les promotions
- On la transforme en table de faits
 - Bridge-factless-fact-table
- SSAS fait le reste : **Many-To-Many relation**

Attention : pas de ventilation de la mesure !

JUNK DIMENSIONS

Problématique :

- J'ai de nombreux axes d'analyse à faible cardinalité
 - Ex : genre, situation matrimoniale, EstDecideur, etc.
- Ajouter des dimensions pour chaque axe entraîne :
 - Une complexité pour le moteur d'agrégation
 - Une complexité pour l'utilisateur

JUNK DIMENSIONS

Solutions :

- Ajouter les valeurs dans la table de faits
→ **dimension factuelle** (ou dimension dégénérée)
- Créer une dimension qui contient toutes les combinaisons existantes
→ **Junk dimension**

Attention à la gestion des NULL

CONCEPTS AVANCÉS

Semi-additivité

Gestion des NULL

Pondération

Distinct Count

Gestion des devises

SCD

Temporal Snapshot

Surrogate Keys

Factless Tables

Junk Dimension

Ragged Hierarchies

Many-to-Many

Sécurité dynamique

MERCI !

Des questions ?

