MOBIDIK

nosql MOdeling of BIg Data, Information and Knowledge

M. Chevalier, A. Kopliku, M. El Malki, O. Teste, R. Tournier

IRIT - Equipe SIG

(Systèmes d'Informations Généralisés)

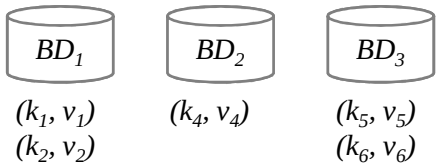
Contexte

- **Mégadonnées ou « *Big Data* »**
  - Des systèmes de gestion de données pour faire face
    - au *v*olume
    - à la *v*ariété
    - à la *v*élocité
  - Par exemple
    - Collections de données du Web (Google, Facebook, Twiter…)
      - 2003 « *The Google file system* » [SOSP03]
      - 2004 « *MapReduce: Simplified Data Processing on Large Clusters* » [OSDI04]
    - Autres collections
      - Astronomie, Biologie, Météorologie, *etc*

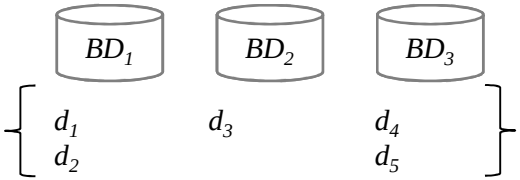
- **Nouveaux systèmes de stockage**
  - NoSQL « *Not-Only-SQL* »
  - Principes
    - Distribution des données et des traitements (volume)
    - Extensibilité et Flexibilité des données (variété, vélocité)

Contexte

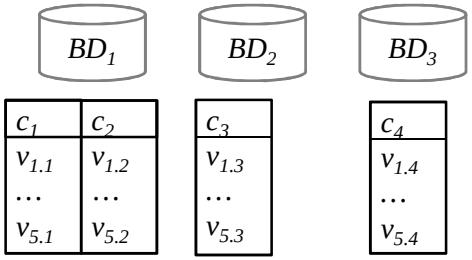
- **Plusieurs paradigmes NoSQL**
  - orientés clé/valeur
    - Données = { (clé, valeur) }
      - Clé : identifiant
      - Valeur : pas de structure
    - Stockage des couples



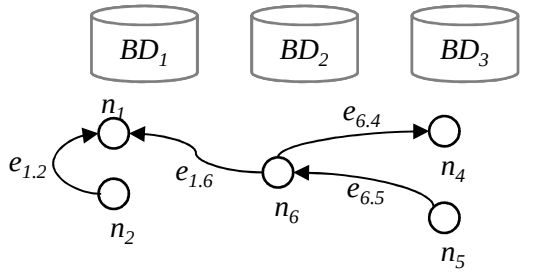
  - orientés documents
    - Données = { documents }
      - Identifiant de document
      - Structures variables & Imbrication
    - Stockage horizontal des documents



  - orientés colonnes
    - Données = Table { tuples }
      - Identifiant de tuple
      - Structures variables
    - Stockage vertical des colonnes



  - orientés graphes
    - Données = { nœuds } { arcs }
      - Graphe étiqueté orienté
      - Structures variables des nœuds & des arcs
    - Stockage distribué du graphe



Contexte

- **Nombreux logiciels qui apparaissent et disparaissent…**

Problématique

- **Projets *versus* Solutions NoSQL**
  - Quel paradigme NoSQL ? Quel logiciel ?
    - Différentes manières de distribuer
  - Comment modéliser les données ?
    - Systèmes NoSQL remettent en cause l'indépendance données/traitements
    - Placement des données dépendant des traitements
  - Comment réutiliser les données ?
    - Migrer d'un système à un autre
    - Transformer d'un modèle à un autre

Problématique

- **Comment modéliser les données ?**

NoSQL => Dépendance Données / Traitements

- **Illustration du problème**
  - Données

PRODUIT	
NP	**DES**
P1	D1
P2	D2
P3	D3

SITE	
NS	URL
S1	w1
S2	w2
S3	w3
S4	w4

FOURNIR		
NS	**NP**	**PU**
S1	P1	10
S1	P3	20
S2	P1	15
S2	P2	20
S2	P3	30
S3	P3	15
S3	P2	15

  - Traitements
    - Le dilemme de l'imbrication
    - q_1 « *Restituer par site web les produits les plus chers* »

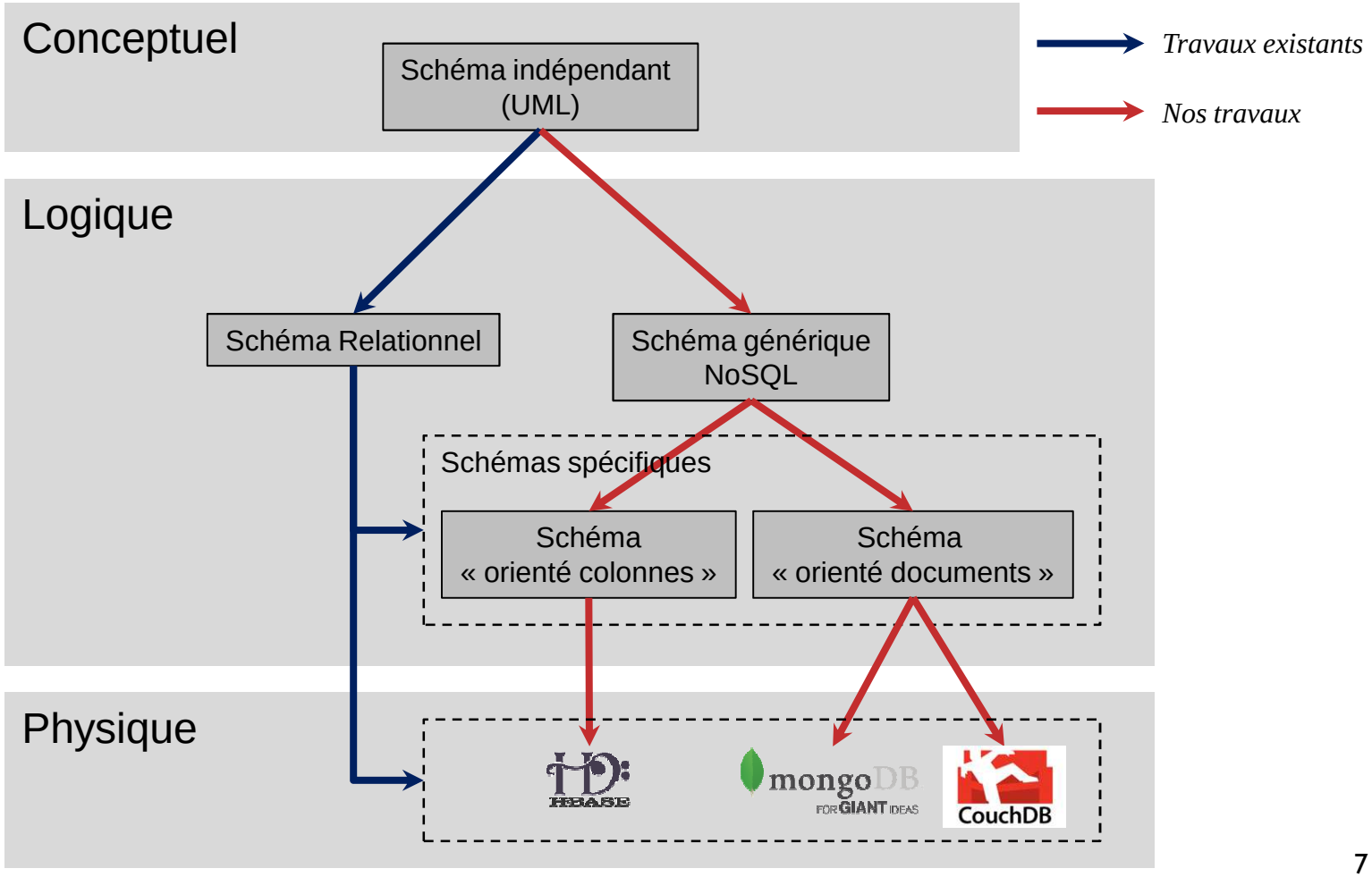
```
SITE                                                FOURNIR                                          PRODUIT
{[NS, URL, FOURNIR {[NP, DES, PU]}]} ≺ {[NS, URL, NP, DES , PU]} ≺ {[NP, DES, FOURNIR {[NS, URL, PU]}]}
```

    - q_2 « *Restituer par désignation de produit le numéro du site web qui le propose au prix le plus bas* »

```
PRODUIT                                             FOURNIR                                          SITE
{[NP, DES, FOURNIR {[NS, URL, PU]}]} ≺ {[NS, URL, NP, DES , PU]} ≺ {[NS, URL, FOURNIR {[NP, DES, PU]}]}
```

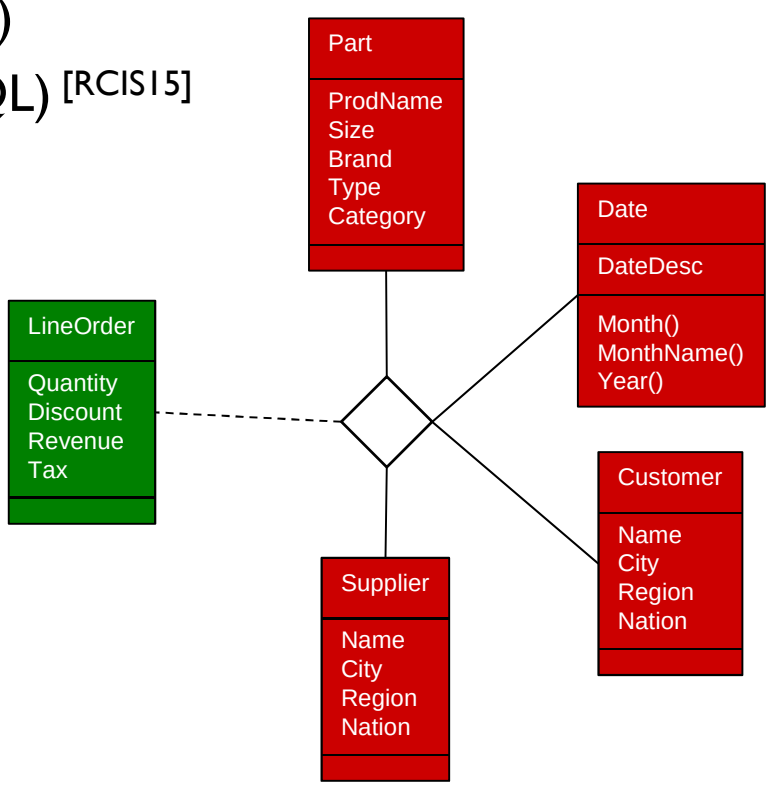
Proposition

- **Différents niveaux d'abstraction** [ICEIS15]
- **Processus de transformation de modèles** [ADBIS15, DAWAK15]
- **Modèles distinguant Schéma / Valeur** [EDA15] [VSST15]



Proposition

- **Etude de cas : entrepôts de données multidimensionnelles**
  - Schéma en étoile
    - Fait : « *sujet de l'analyse* » (indicateurs numériques)
    - Dimension : « *axe de l'analyse* » (paramètres)
- **Benchmark**
  - SSB (ROLAP)
  - SSB+ (NoSQL) [RCIS15]



Proposition

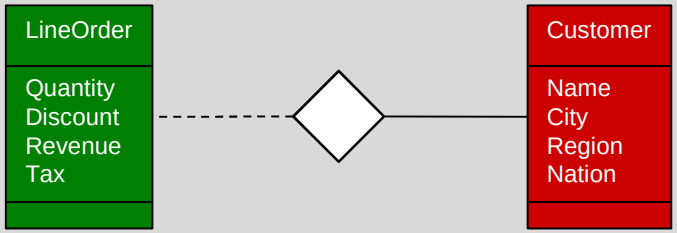
- **Modèle générique assurant séparation structure/valeur**
  - Notations
    - [ ] structure
    - { } ensemble

- **Plusieurs processus générique de transformations**
  - MLD_0 « Flat »
  - MLD_1 « Deco »
  - MLD_2 « Shattered »
  - MLD_3 « Hybrid »

Proposition

- **Exemple MLD_1 « Deco »**

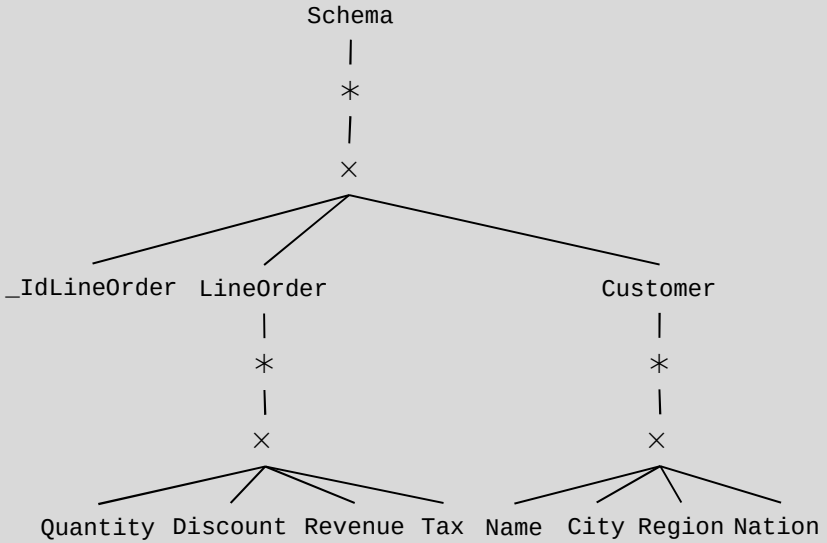
Conceptuel



MLD_1

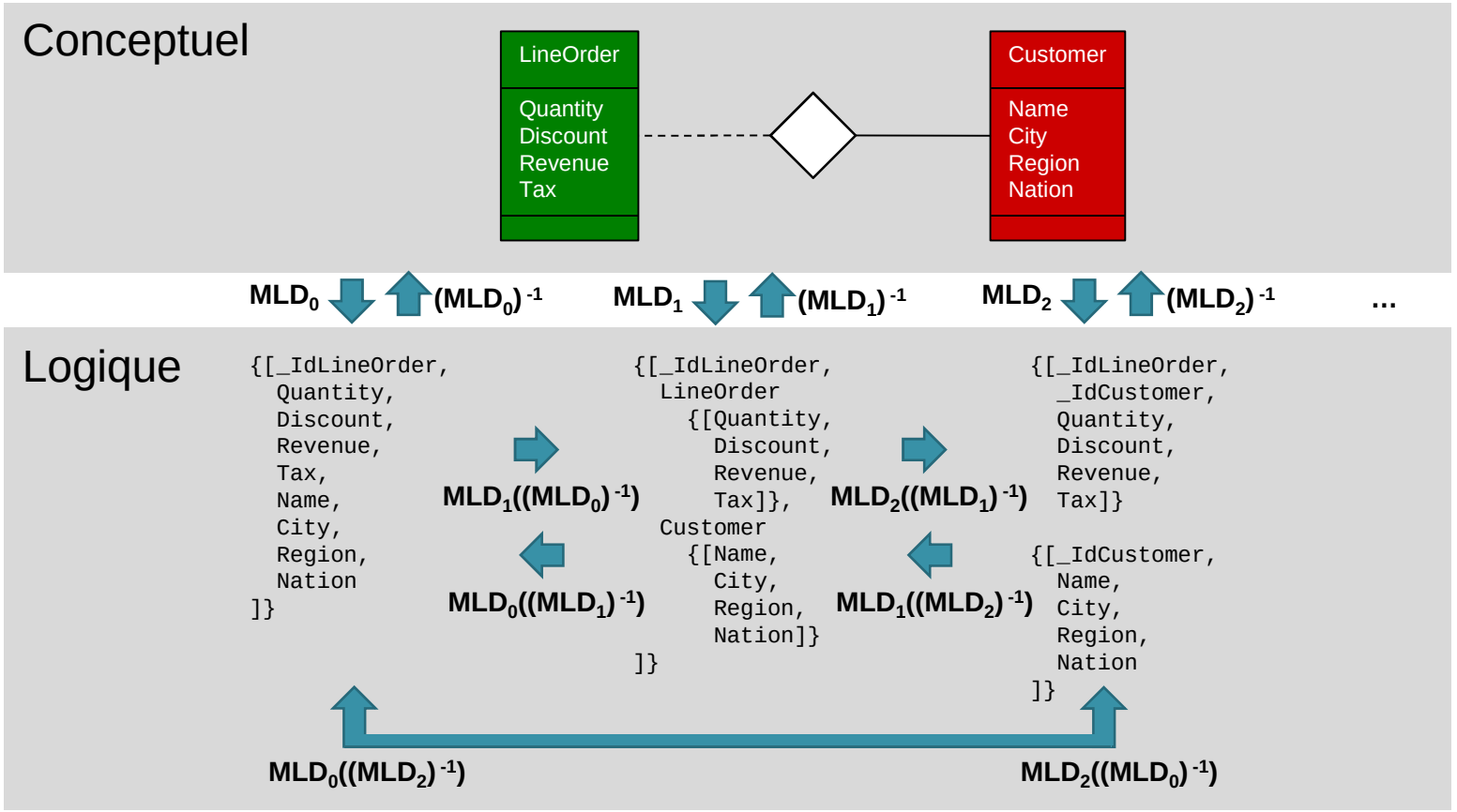
Logique

```
Schema
{[_IdLineOrder,
  LineOrder
    {[Quantity,
      Discount,
      Revenue,
      Tax]},
  Customer
    {[Name,
      City,
      Region,
      Nation]}
]}
```



Proposition

- **Processus de transformation**
  - Automatique
  - Processus réversible inter-modèle



Proposition

- **Processus de transformation**
  - Exemple

```
{
 _id    : "22021972",
 Name   : "Olivier",
 City   : "Toulouse",
 Region : "Midi-Pyrénées",
 Nation : "France"
}
```

$MLD_1((MLD_0)^{-1})$

```
{
 _id      : "22021972",
 Customer : { Name   : "Olivier",
              City   : "Toulouse",
              Region : "Midi-Pyrénées",
              Nation : "France"
            }
}
```

$MLD_1((MLD_1)^{-1})_{HBase}$

$MLD_1((MLD_0)^{-1})_{HBase}$

RowKey	Customer			
22021972	Name	City	Region	Nation
	Olivier	Toulouse	Midi-Pyrénées	France

Proposition

- **Expérimentations**
  - Systèmes NoSQL
    - Stockage horizontal mongoDB. FOR GIANT IDEAS
    - Stockage vertical HBASE
  - Benchmark SSB+
    - Opérationnel
    - Scripts de génération par MDL_i paramétrable *Facteur d'échelle*
      - sf1 $\Rightarrow$ 10^7 données LineOrder
      - sf10 $\Rightarrow$ 10×10^7 données LineOrder
      - sf200 $\Rightarrow$ 200×10^7 données LineOrder (~1.5 To avec mongoDB. FOR GIANT IDEAS)
      - …
    - Jeux de requêtes
      - LMD (~INSERT, ~UPDATE)
      - LID (~SELECT)

Conclusion

- **Autres travaux**
  - Contourner la dépendance données/traitements
    - Règles de choix des modèles en fonction des traitements
    - Tenir compte de la distribution et des transferts de données
  - Prendre en compte la variabilité des données

Références

[ADBIS15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Implementation of multidimensional databases in column-oriented NoSQL systems*. East-European Conference on Advances in Databases and Information Systems (ADBIS'15), Poitiers, France, p.79-91. *doi: 10.1007/978-3-319-23135-8_6*

[DAWAK15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Not Only SQL Implementation of multidimensional database*. International Conference on Big Data Analytics and Knowledge Discovery (DAWAK'15), Valencia, Spain, p.379-390. *doi: 10.1007/978-3-319-22729-0_29*

[EDA15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Entrepôts de données multidimensionnelles NoSQL*. 11ème Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA'15), vol. RNTI-B-11, Bruxelles, Belgique, p.161-176.

[ICEIS15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Implementing Multidimensional Data Warehouses into NoSQL*. International Conference on Enterprise Information Systems (ICEIS'15), Barcelona, Spain, p.172-183. *doi: 10.5220/0005379801720183*

[OSDI04]

J. Dean, S. Ghemawat (2004). *MapReduce: Simplified Data Processing on Large Clusters*. 6th Symposium on operating system design and implementation (OSDI'04), San Francisco, California, p.137-150.

[RCIS15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Benchmark for OLAP on NoSQL Technologies*. International Conference on Research Challenges in Information Science (RCIS'15), IEEE, Athens, Greece, p. 480-485. *doi: 10.1109/RCIS.2015.7128909*

[SOSP03]

S. Ghemawat, H. Gobioff, S-T. Leung (2003). *The Google file system*. ACM Symposium on operating systems principles (SOSP '03), New York, NY, USA, p.29-43. *doi:10.1145/945445.945450*

[VSST15]

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier (2015). *Implantation « Not-Only-SQL » des bases de données multidimensionnelles*. Colloque Veille Stratégique Scientifique et Technologique (VSST'15), Grenade, Espagne.