Journées INRAE Clermont-Ferrand
Les systèmes d'information agro-environnementaux à l'ère du Big Data
Vendredi 29 Janvier 2021

Nouveaux paradigmes de modélisation – NoSQL orienté-document

Olivier TESTE

olivier.teste@irit.fr

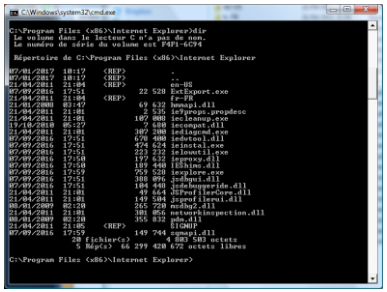
https://www.irit.fr/~Olivier.Teste/

La genèse

- **Technologie des Bases de données**

  - 1969 C.W. Bachman
    Data Structure Diagrams. DATA BASE 1(2): 4-10

  - 1973 Prix Turing



Fichiers

Abstraction

La genèse

- **Bases de données relationnelles**

  - 1970 E.F. Codd
    A Relational Model of Data for Large Shared Data Banks. Commun. ACM 13(6), p.377-387

- **Théorie de la normalisation**
  - 1971 E. F. Codd
    Further Normalization of the Data Base Relational Model. IBM Research Report, San Jose, California RJ909
  - 1981 Prix Turing

BD = { Relation **}**

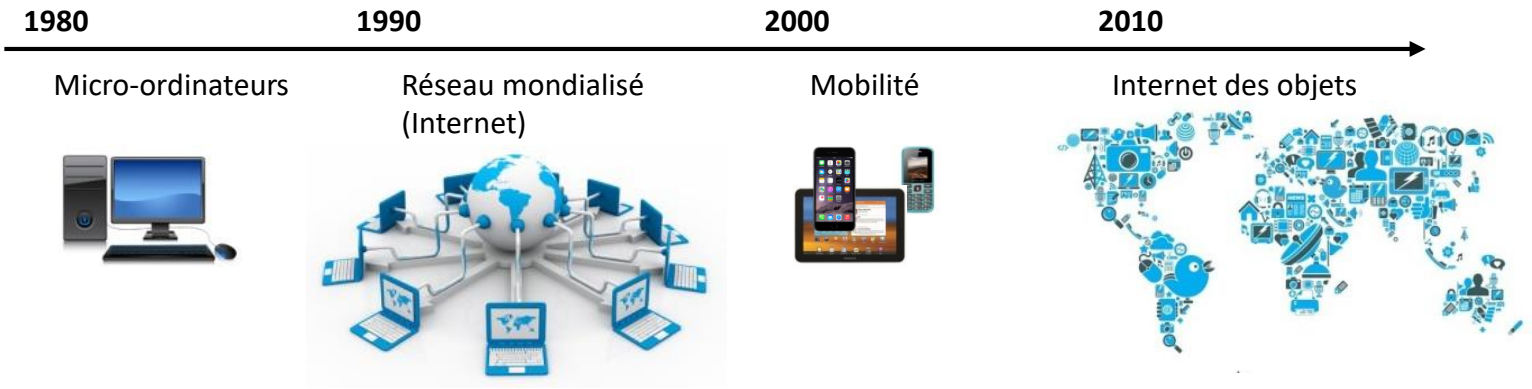
Ri = < Schéma**,** Extension **>**
Schéma **= <a1,…, an>**
Extension **= <t1,…, tm>**

Séparation Données/Traitements
→ Concept de Relations (1NF : atomique)
→ Algèbre Relationnelle

Non redondance
→ Normalisation des schémas (1NF, 2NF, 3NF)

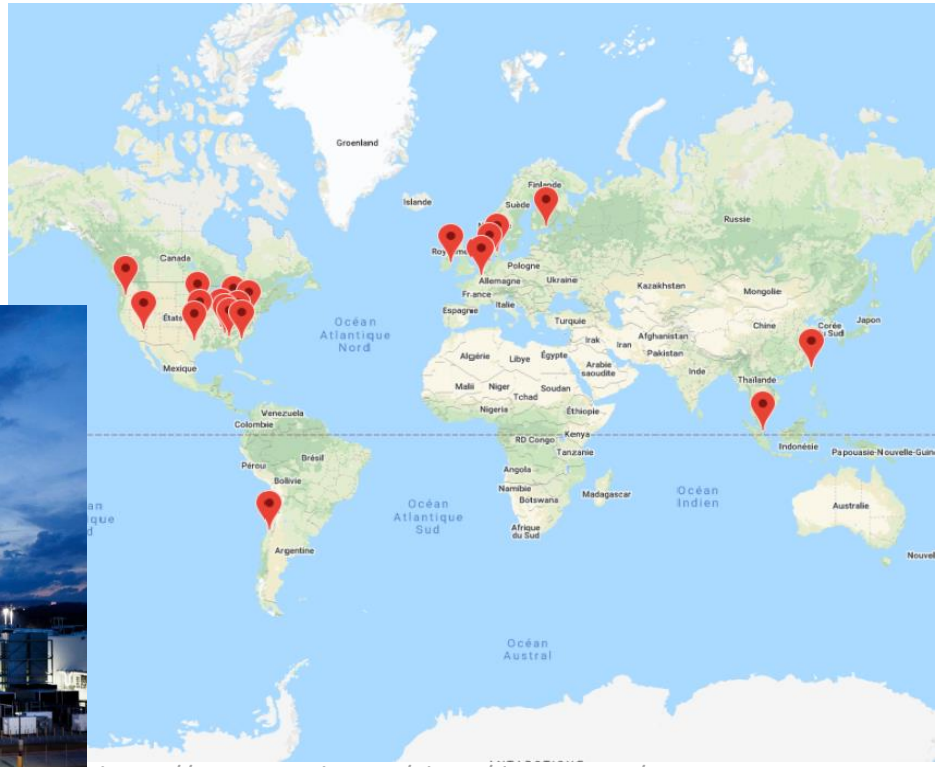
La genèse

- Depuis 1990
  - Un déferlement de données
  - Une inversion de paradigmes
- Pourquoi ?



La genèse

- Google un cloud mondial
  - Des « data centers » répartis dans le monde
    - Amérique du Nord (13)
    - Amérique du Sud (1)
    - Europe (5)
    - Asie (2)



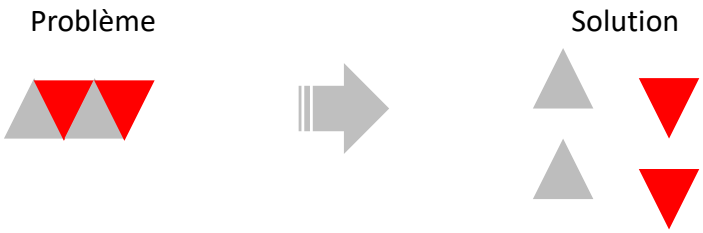
https://www.google.com/about/datacenters/

https://www.google.com/about/datacenters/

La genèse

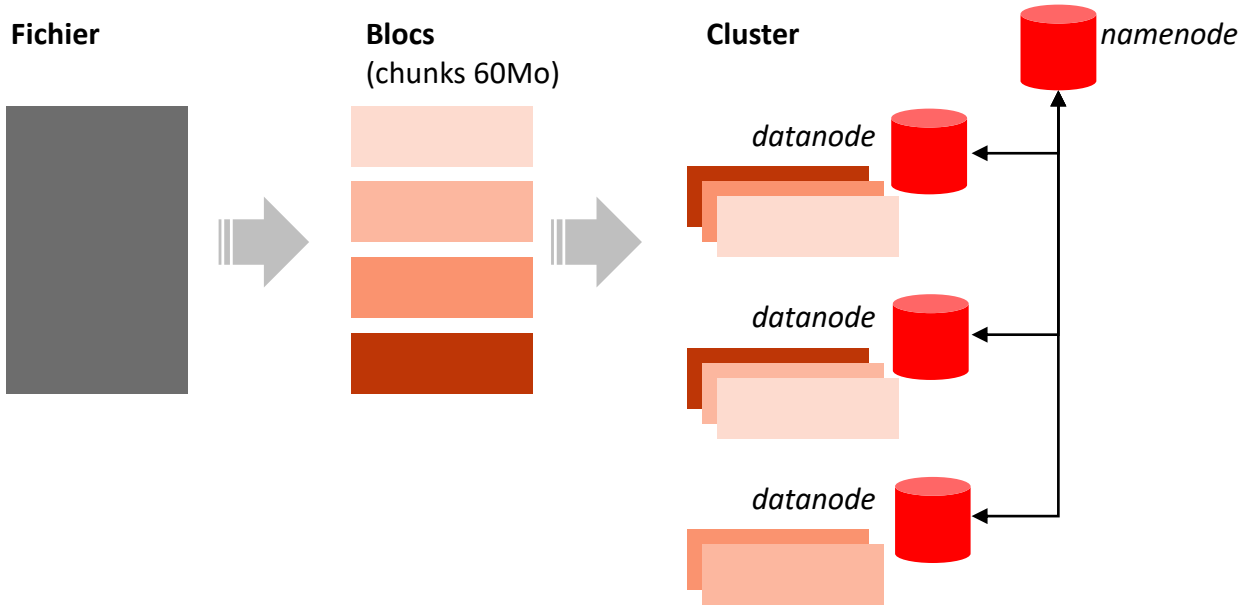
- Comment gérer les Big Data ?
  - A l'initiative des grands opérateurs du Web
  - Remise en cause des approches classiques par de nouveaux systèmes
    - Centralisées => Distribution des données
    - Structurées => Déstructurées
    - Normalisées => Dénormalisées (Redondances)

« divide and conquer »



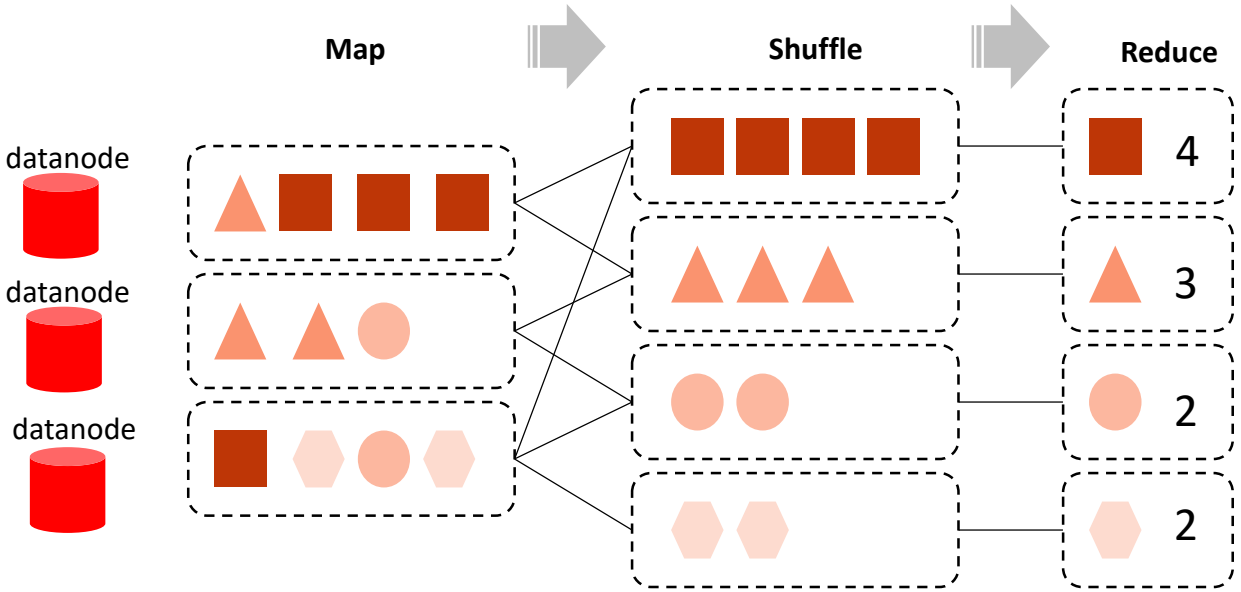
La genèse

- Données
  - 2003
    S. Ghemawat, H. Gobioff, S-T. Leung
    The Google file system. SOSP, p.29-43



La genèse

- Traitements
  - 2004
    J. Dean, S. Ghemawat
    MapReduce: Simplified Data Processing on Large Clusters. OSDI, p.137-150



La genèse

- Avènement des systèmes « Not-Only SQL »
  - Sans schéma
    - clé/valeur
  - Schémas dynamiques
    - orienté-colonne
    - orienté document
    - orienté-graphe

NoSQL orienté-documents

- Schéma dynamique (« schemaless »)

En SQL

BD = { (Si; Vi) }

```
[{_id:10,prenom:"Charles",nom:"Bachman",annee:1969},
 {_id:20,prenom:NULL,     nom:"Codd",    annee:1970},
 {_id:30,prenom":"Peter", nom:"Chen",    annee:NULL}]
```

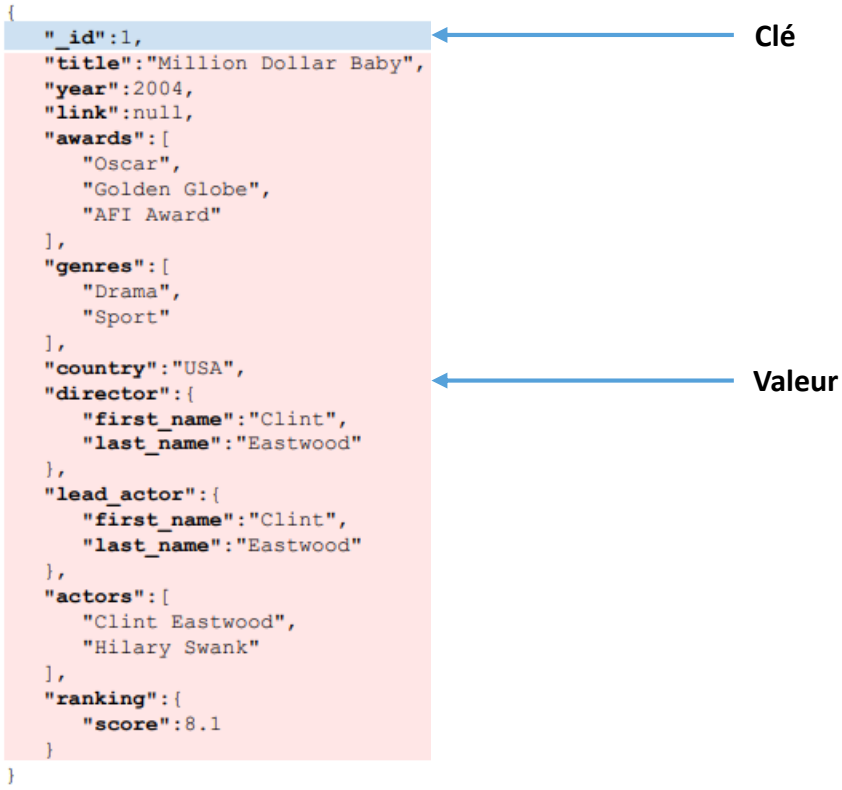
En NoSQL

BD = { (Vi) }

```
[{_id:10,prenom:"Charles",nom:"Bachman",annee:1969},
 {_id:20,                 name:"Codd",   year:1970},
 {_id:30,prenom":"Peter", nom:"Chen"}]
```

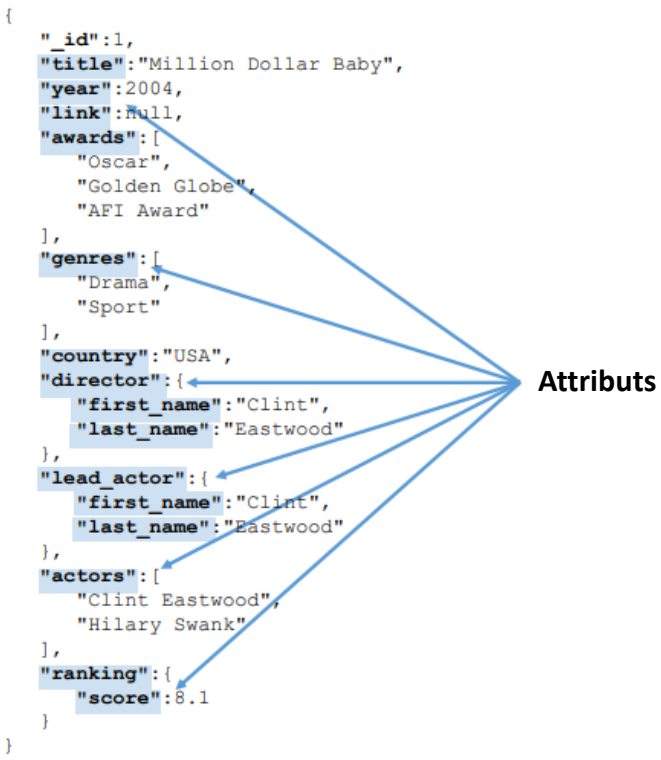
NoSQL orienté-documents

- Représentation JSON des documents



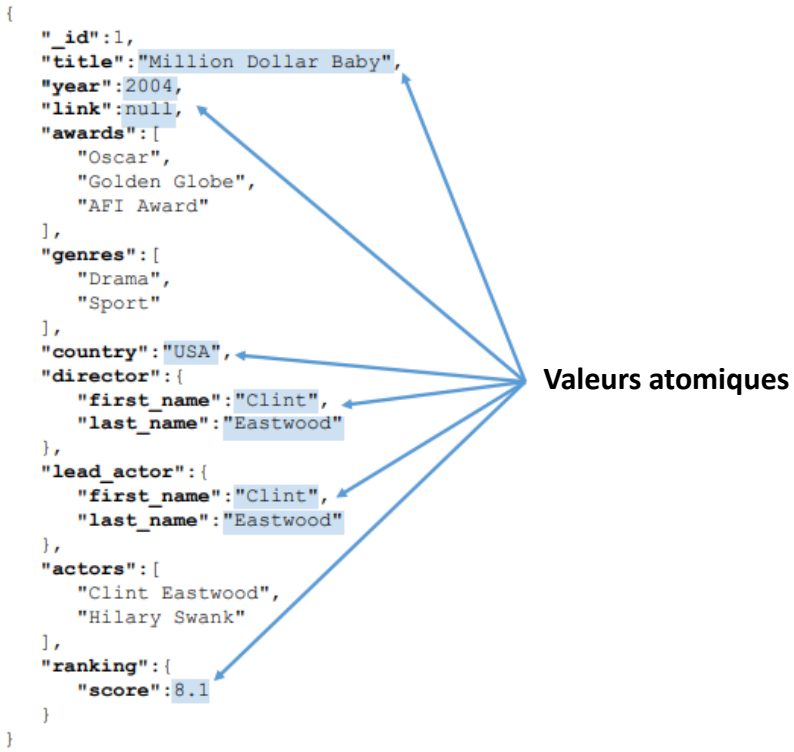
NoSQL orienté-documents

- Représentation JSON des documents



NoSQL orienté-documents

- Représentation JSON des documents



NoSQL orienté-documents

- Représentation JSON des documents

```
{
   "_id":1,
   "title":"Million Dollar Baby",
   "year":2004,
   "link":null,
   "awards":[
      "Oscar",
      "Golden Globe",
      "AFI Award"
   ],
   "genres":[
      "Drama",
      "Sport"
   ],
   "country":"USA",
   "director":{
      "first_name":"Clint",
      "last_name":"Eastwood"
   },
   "lead_actor":{
      "first_name":"Clint",
      "last_name":"Eastwood"
   },
   "actors":[
      "Clint Eastwood",
      "Hilary Swank"
   ],
   "ranking":{
      "score":8.1
   }
}
```

Ensemble ordonné de valeurs

Valeurs imbriquées

NoSQL orienté-documents

- Représentation JSON des collections de documents


```
{
   "_id":1,
   "title":"Million Dollar Baby",
   "year":2004,
   "link":null,
   "awards":[
      "Oscar",
      "Golden Globe",
      "AFI Award"
   ],
   "genres":[
      "Drama",
      "Sport"
   ],
   "country":"USA",
   "director":{
      "first_name":"Clint",
      "last_name":"Eastwood"
   },
   "lead_actor":{
      "first_name":"Clint",
      "last_name":"Eastwood"
   },
   "actors":[
      "Clint Eastwood",
      "Hilary Swank"
   ],
   "ranking":{
      "score":8.1
   }
}
```

```
{
   "_id":2,
   "title":"In the Line of Fire",
   "info":{
      "year":1993,
      "country":"USA",
      "link":"https://goo.gl/2A253A",
      "genres":[
         "Drama",
         "Action",
         "Crime"
      ],
      "people":{
         "director":{
            "first_name":"Clint",
            "last_name":"Eastwood"
         },
         "lead_actor":{
            "first_name":"Clint",
            "last_name":"Eastwood"
         },
         "actors":[
            "Clint Eastwood",
            "John Malkovich"
         ]
      },
      "ranking":{
         "score":7.2
      }
   }
}
```

NoSQL orienté-documents

- Interrogation des collections de documents mongoDB
  - Sans tenir compte de la variabilité (hétérogénéité)

```
aggregate([
   { $match:{
       "country":"USA"}}
   { $project:{
       "title":1}}
])
```

```
[
 {"_id":1, "title":"Million Dollar Baby" }
]
```

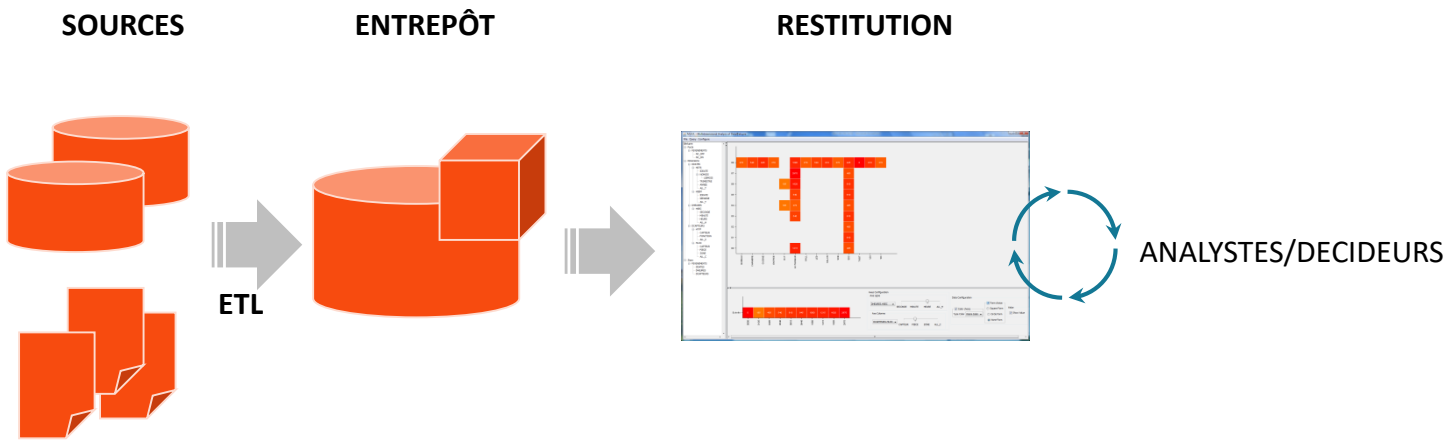
  - Avec prise en compte de la variabilité

```
aggregate([
   { $match:{
       $or:[
           {"country":"USA"},
           {"info.country":"USA"}]}}
   { $project:{
       "title":1}}
])
```

```
[
 {"_id":1, "title":"Million Dollar Baby" }
 {"_id":2, "title":"In the Line of Fire" }
]
```

OLAP en NoSQL orienté-documents

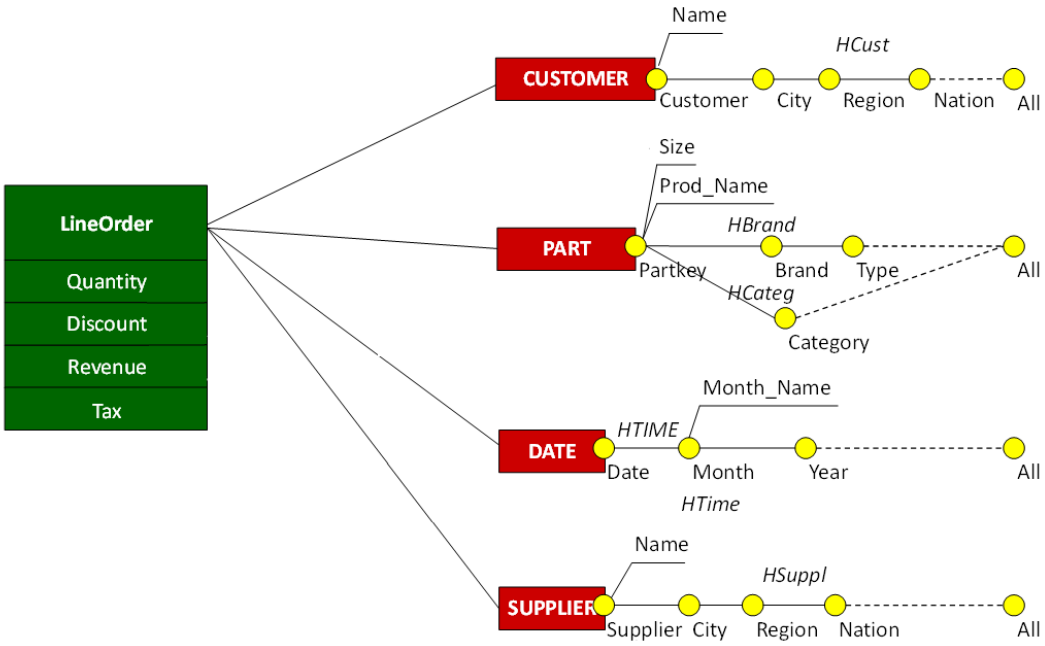
- Principes des entrepôts R-OLAP



- Dans le contexte NoSQL, attention au processus ETL
  - Schéma dynamique => variabilité des données entreposées

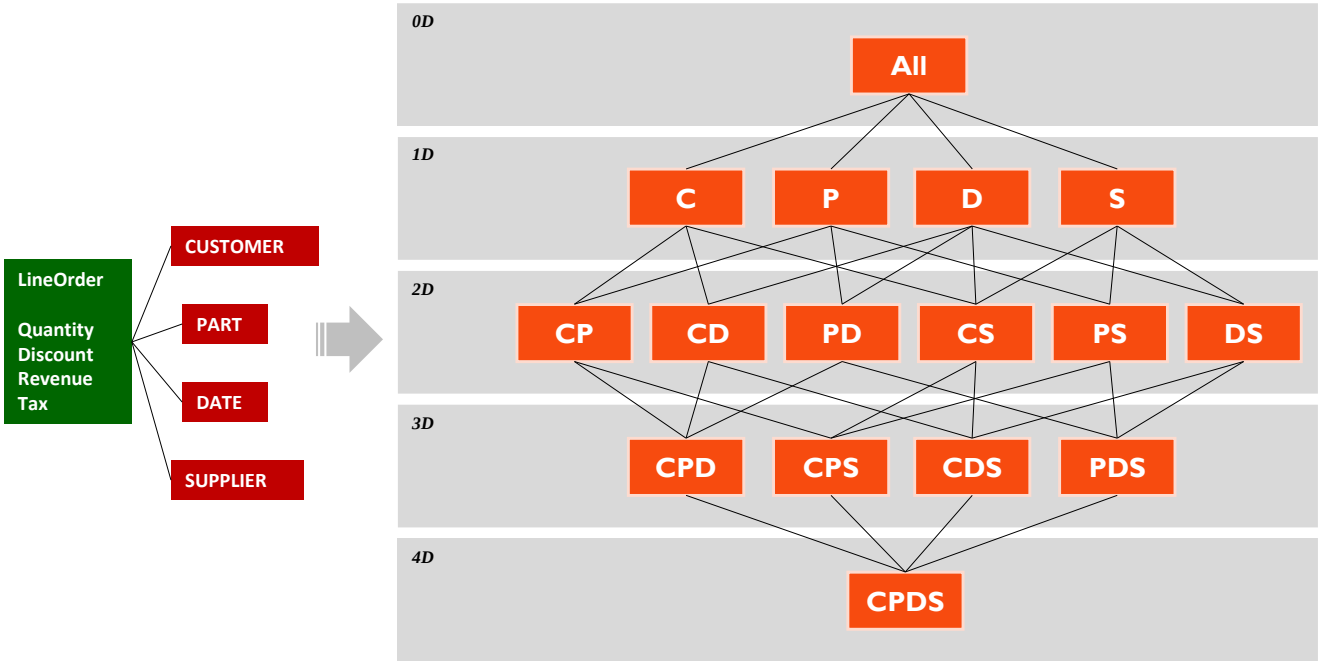
OLAP en NoSQL orienté-documents

- Modélisation multidimensionnelle des entrepôts R-OLAP
  - Schémas en étoile



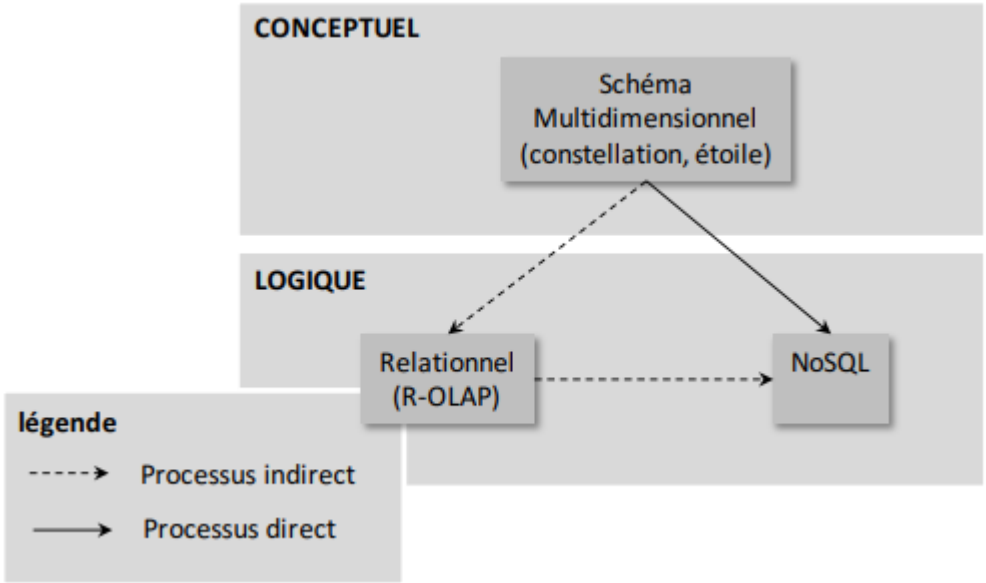
OLAP en NoSQL orienté-documents

- Modélisation multidimensionnelle des entrepôts R-OLAP
  - Optimisation par pré-calculs d'agrégats



OLAP en NoSQL orienté-documents

- Comment gérer en entrepôt de données en NoSQL ?

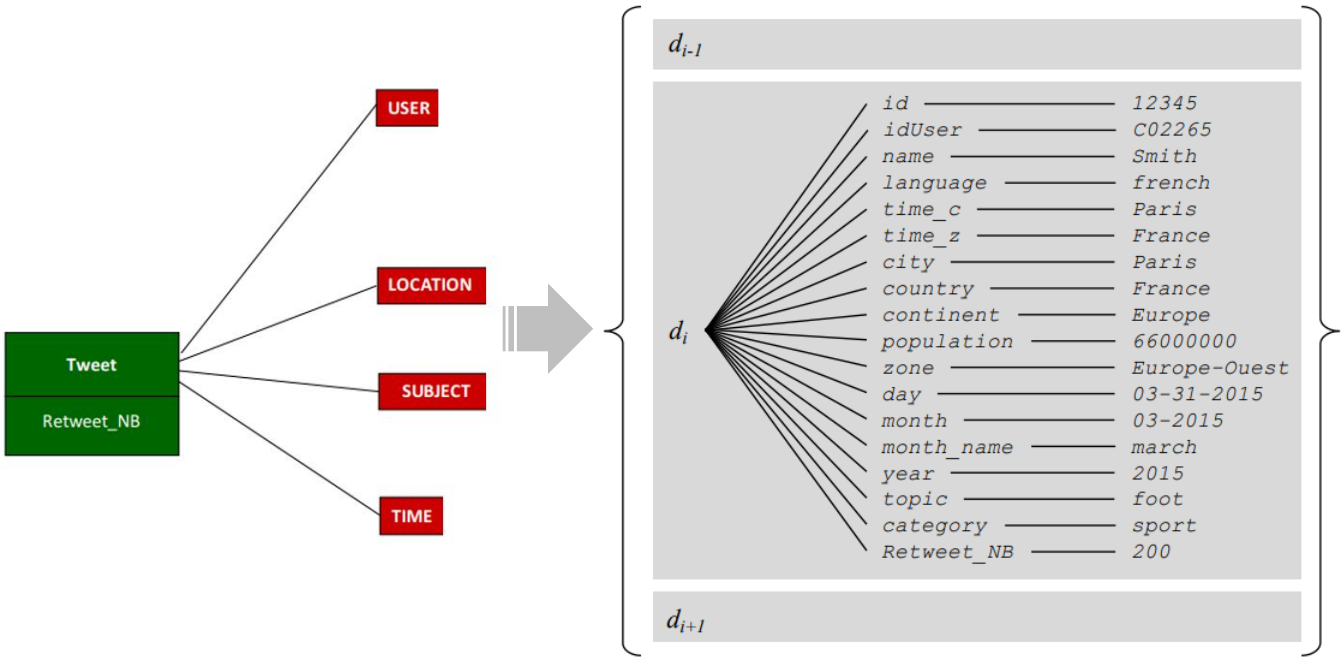


OLAP en NoSQL orienté-documents

- Approche directe
  - Exploiter les avantages du NoSQL
    - Schéma dynamique
    - Imbrication
  - 4 modèles développés
    - À plat – DFL (Document Flat Logic)
    - Imbriqué – DNL (Document Nested Logic)
    - Hybride – DHL (Document Hybrid Logic)
    - Éclaté – DSL (Document Shattered Logic)

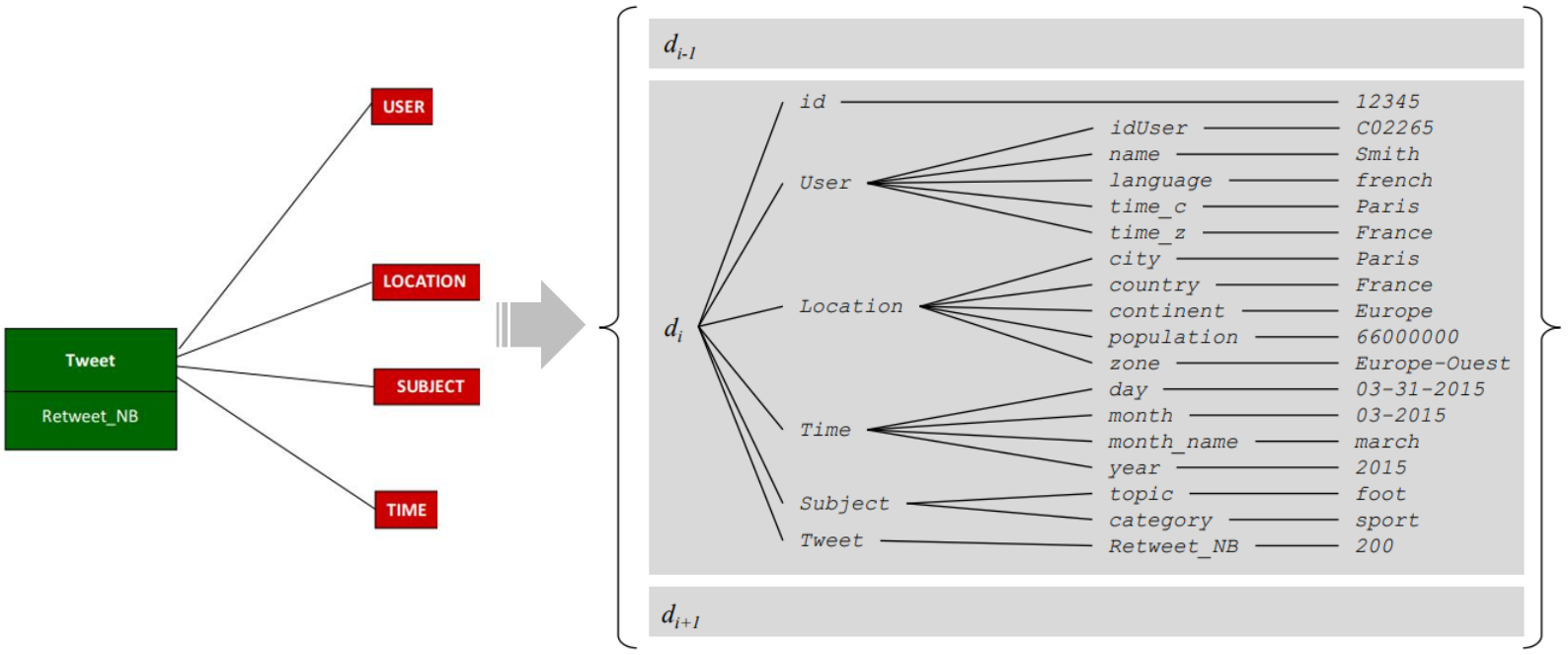
OLAP en NoSQL orienté-documents

- Approche directe
  - À plat – DFL



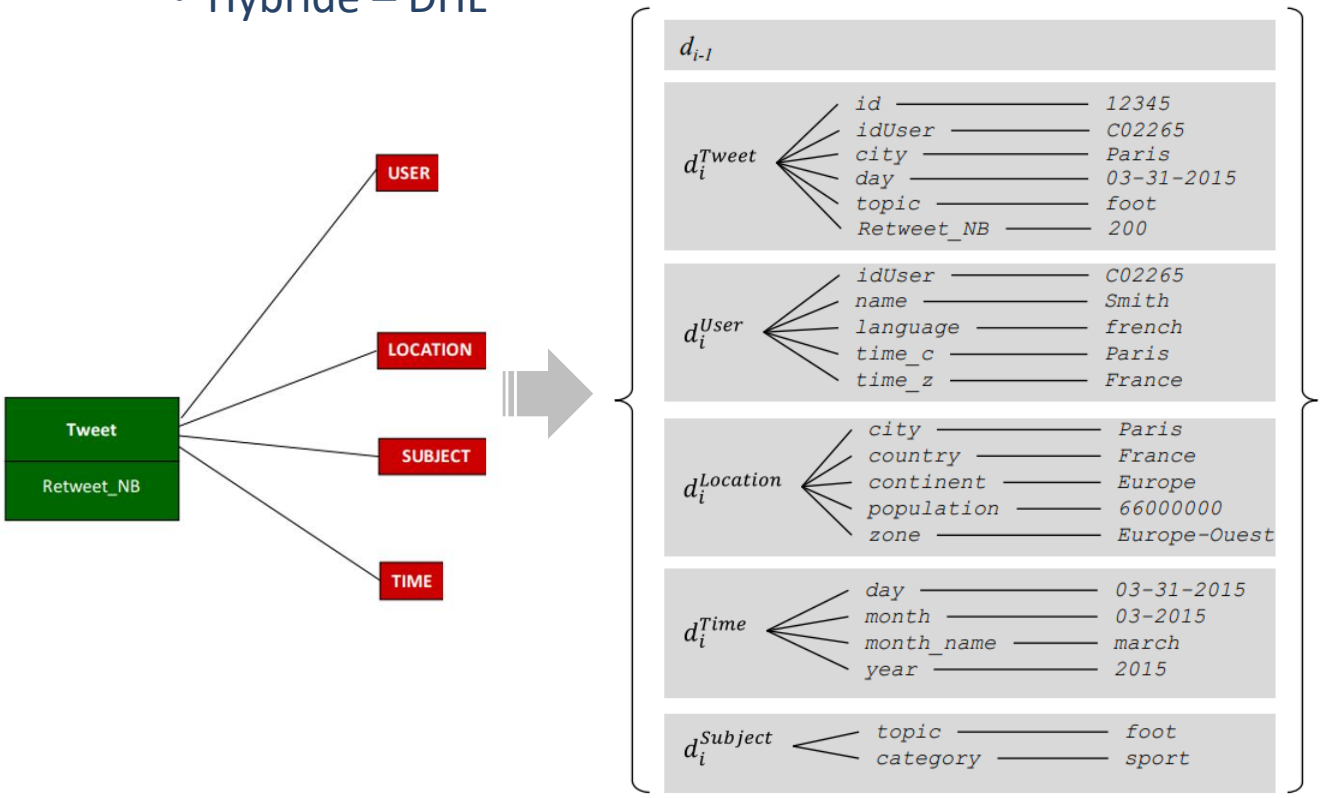
OLAP en NoSQL orienté-documents

- Approche directe
  - Imbriqué – DNL



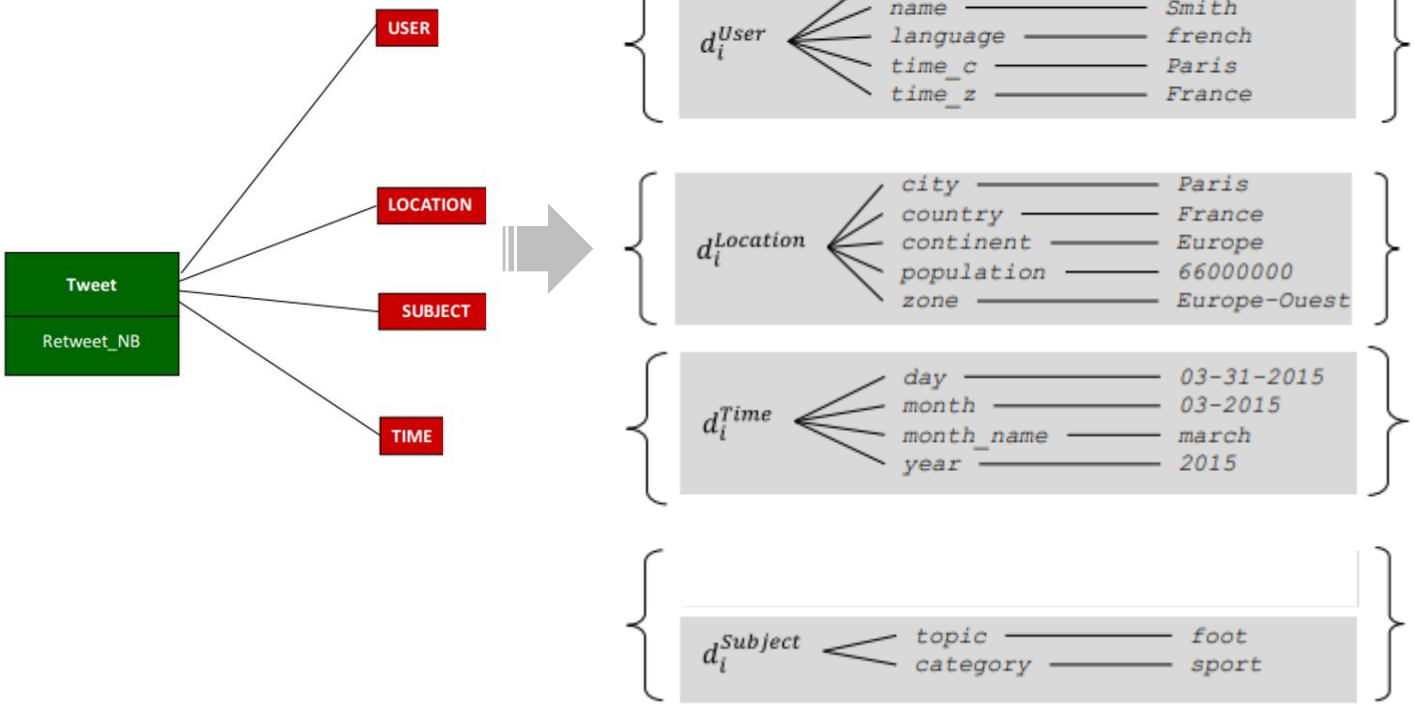
OLAP en NoSQL orienté-documents

- Approche directe
  - Hybride – DHL



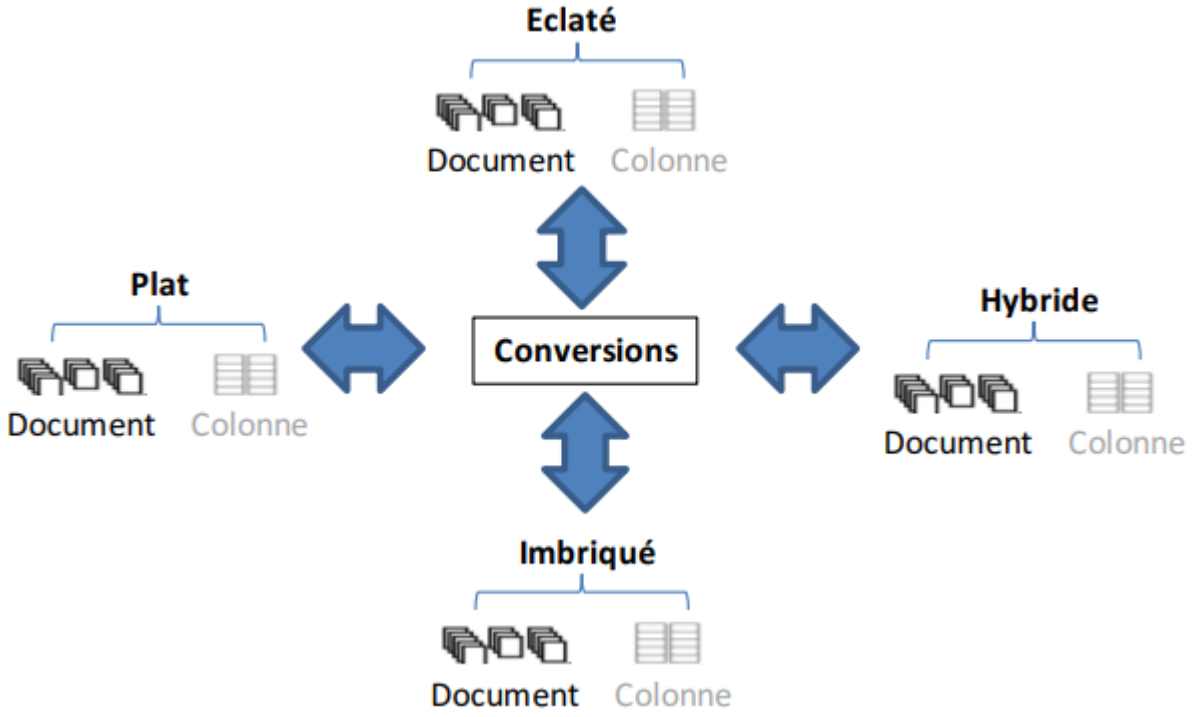
OLAP en NoSQL orienté-documents

- Approche directe
  - Éclaté – DSL



OLAP en NoSQL orienté-documents

- Conversions intra-modèles



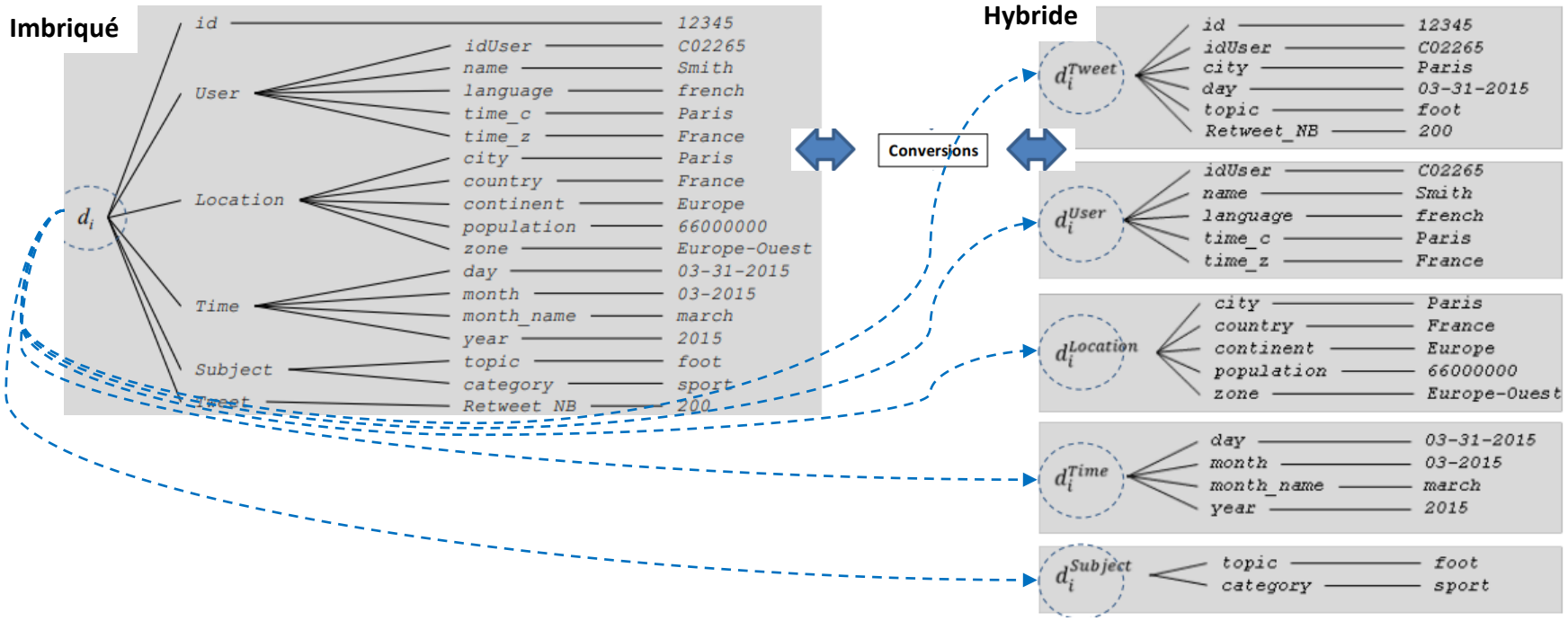
OLAP en NoSQL orienté-documents

- Conversions intra-modèles
  - Processus formalisés / automatisables

			Schéma cible orienté documents			
			Plat	**Imbriqué**	**Hybride**	**Eclaté**
Schéma origine orienté documents	**Plat**	$C^F.d_i.m$ $C^F.d_i.a$		$C^F.d_i.\mathbf{N^F.m}$ $C^F\ d_i.\mathbf{N^D.a}$	$C^F.d_i.\mathbf{m}$ $C^F.\mathbf{d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$	$C^F.d_i.\mathbf{m}$ $\mathbf{C^D.d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$
	Imbriqué	$C^F.d_i.N^F.m$ $C^F\ d_i.N^D.a$	$C^F.d_i.\mathbf{m}$ $C^F.d_i.\mathbf{a}$		$C^F.d_i.\mathbf{m}$ $C^F.\mathbf{d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$	$C^F.d_i.\mathbf{m}$ $\mathbf{C^D.d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$
	Hybride	$C^F.d_i.m$ $C^F.d_j.a$ $a \equiv id^D \Rightarrow$ $C^F.d_i.a$	$C^F.d_i.\mathbf{m}$ $C^F.d_i.\mathbf{a}$	$C^F.d_i.\mathbf{N^F.m}$ $C^F\ d_i.\mathbf{N^D.a}$		$C^F.d_i.\mathbf{m}$ $\mathbf{C^D.d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$
	Eclaté	$C^F.d_i.m$ $C^D.d_j.a$ $a \equiv id^D \Rightarrow$ $C^F.d_i.a$	$C^F.d_i.\mathbf{m}$ $C^F.d_i.\mathbf{a}$	$C^F.d_i.\mathbf{N^F.m}$ $C^F\ d_i.\mathbf{N^D.a}$	$C^F.d_i.\mathbf{m}$ $C^F.\mathbf{d_j.a}$ $a \equiv id^D \Rightarrow$ $C^F.d_i.\mathbf{a}$	

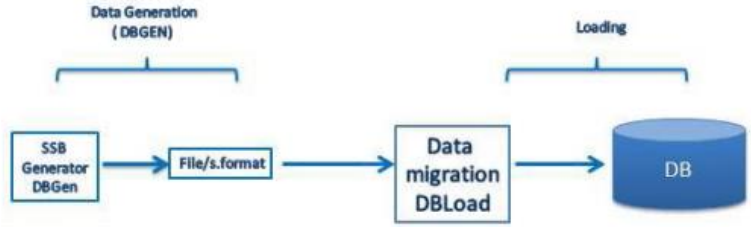
OLAP en NoSQL orienté-documents

- Conversions intra-modèles

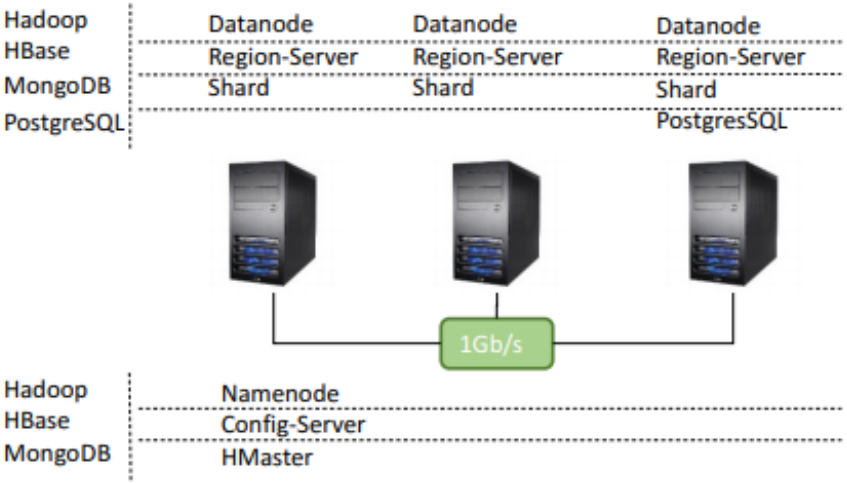


Banc d'essai SSB+ (Star Schema Benchmark)

- Benchmark SSB+ [Chevalier et al. 2015]



- Infrastructure



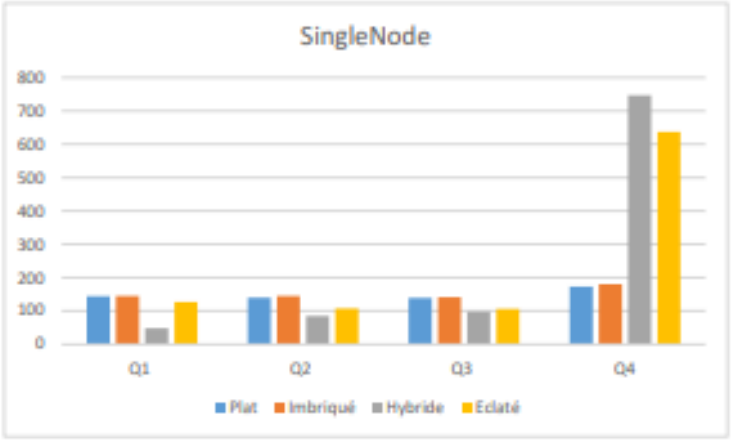
Banc d'essai SSB+ (Star Schema Benchmark)

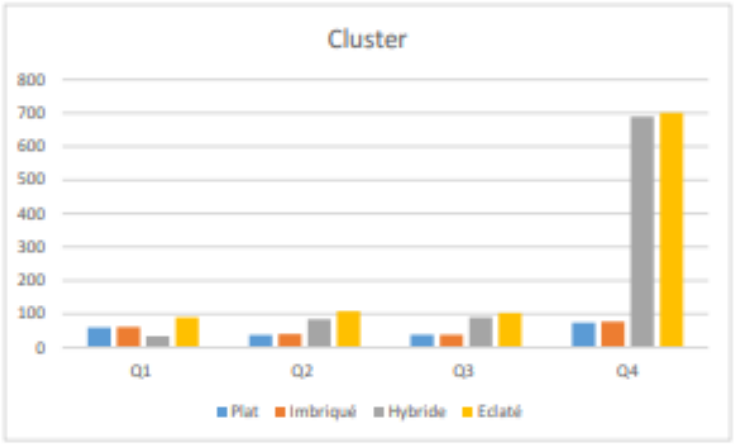
- Volume
  - 1 nœud: DFL/ DNL $^{(x3)}$**>>** DHL/DSL
  - 3 nœuds: idem
- Temps de construction (alimentation)
  - 1 nœud: DFL/ DNL $^{(x1.5/4)}$**>>** DHL $^{(x2.5/4)}$**>>** DSL
  - 3 nœuds: peu d'écarts (« Balancer » qui équilibre les données)

Implantations / Configuration	**Plate** (DFL)	**Imbriquée** (DNL)	**Eclatée** (DSL)	**Hybride** (DHL)
sf=1, une seule machine	1306s/15Go	1235s/15Go	1005s/4.2Go	501s/4.2Go
sf=10, une seule machine	16680s/150Go	16080s/150Go	4320s/42Go	4407s/42Go
sf=25, une seule machine	46704s/375Go	44220s/375Go	10980s/105Go	11020s/105Go
sf=1, cluster	4246s/15Go	4304s/15Go	3767s/4.2Go	3737s/4.2Go

Banc d'essai SSB+ (Star Schema Benchmark)

- Interrogation
  - Q1/Q2/Q3 – sélections –
    - DFL/ DNL/DHL/DSL les approches sont à peu près comparables (différences fonction du taux de sélectivité dans les Qi)
  - Q4 – jointures –
    - DFL/ DNL $^{(x7)}$<< DHL/DSL les approches Hybrides et Eclatés (normalisées) sont très couteuses





Conclusion

- NoSQL orienté-documents
  - Interrogation avec prise en charge de la variabilité (schéma dynamique) [Ben Hamadou, 2019]
  - Construction d'Entrepôts multidimensionnels [El Malki, 2016]

Références

- Interrogation NoSQL
  - [Ben Hamadou, 2019] Hamdi Ben Hamadou, Querying heterogeneous data in NoSQL document stores. Paul Sabatier University, Toulouse, France, 2019
  - [Ben Hamadou, 2019] Hamdi Ben Hamadou, Faiza Ghozzi, André Péninou, Olivier Teste, Schema-independent querying for heterogeneous collections in NoSQL document stores. Inf. Syst. 85: 48-67 (2019)
- Entrepôts de données NoSQL
  - [El Malki, 2016] Mohammed El Malki, Modélisation NoSQL des entrepôts de données multidimensionnelles massives. University of Toulouse-Jean Jaurès, France, 2016
  - [Chevalier et al. 2015] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, Ronan Tournier, How Can We Implement a Multidimensional Data Warehouse Using NoSQL? ICEIS (Revised Selected Papers) 2015: 108-130
  - [Chevalier et al. 2015] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, Ronan Tournier, Benchmark for OLAP on NoSQL technologies comparing NoSQL multidimensional data warehousing solutions. RCIS 2015: 480-485