Vers un développement de systèmes d'information orientés données distribuées :
requêtes distribuées dans un environnement NoSQL

M. El Malki

H. Ben Hamadou , A. Laadhar, O.Teste

IRIT - Equipe SIG

(Systèmes d'Informations Généralisés)

Journée InfraMed / Cogiter 2017

Plan

- **Big Data & noSQL**
- **Hétérogénéité des schémas**
- **Interrogation multi-structures**
- **Quelques solutions**

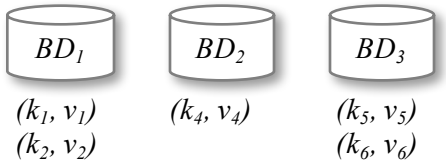
Contexte

- **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*

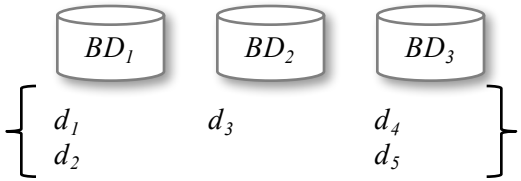
- **Nouveaux systèmes de stockage**
  - NoSQL « *Not-Only-SQL* »
  - Principes
    - Distribution des données et des traitements (volume)
    - Extensibilité et Flexibilité des données (variété, vélocité)

Contexte

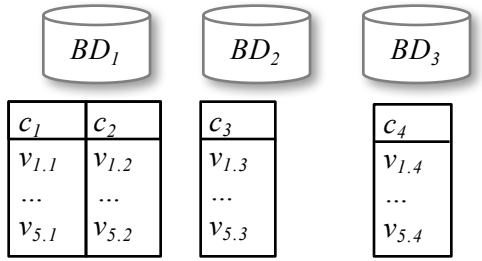
- **Plusieurs paradigmes NoSQL**
  - orientés clé/valeur
    - Données = { (clé, valeur) }
      - Clé : identifiant
      - Valeur : pas de structure
    - Stockage des couples



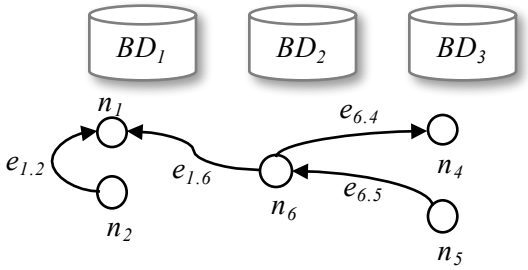
  - orientés documents
    - Données = { documents }
      - Identifiant de document
      - Structures variables & Imbrication
    - Stockage horizontal des documents



  - 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

- **Points communs (1/2)**
  - **Imbrication/dénormalisation des données**

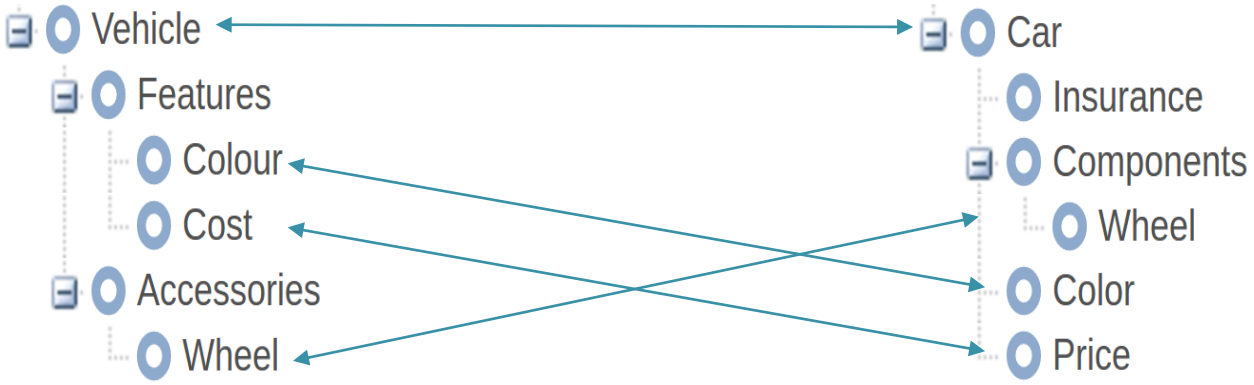
    - Comment modéliser les données (imbrication des données) ?
      - Systèmes NoSQL remettent en cause l'indépendance données/traitements
      - Placement des données dépendant des traitements

    - Plusieurs modèles de données possibles (pas de modèles génériques)
      - *Modèle plat*
      - *Imbriqué*
      - *Hybride*

    - Plusieurs systèmes de stockage
      - Mono-store
      - Multi-store

Contexte

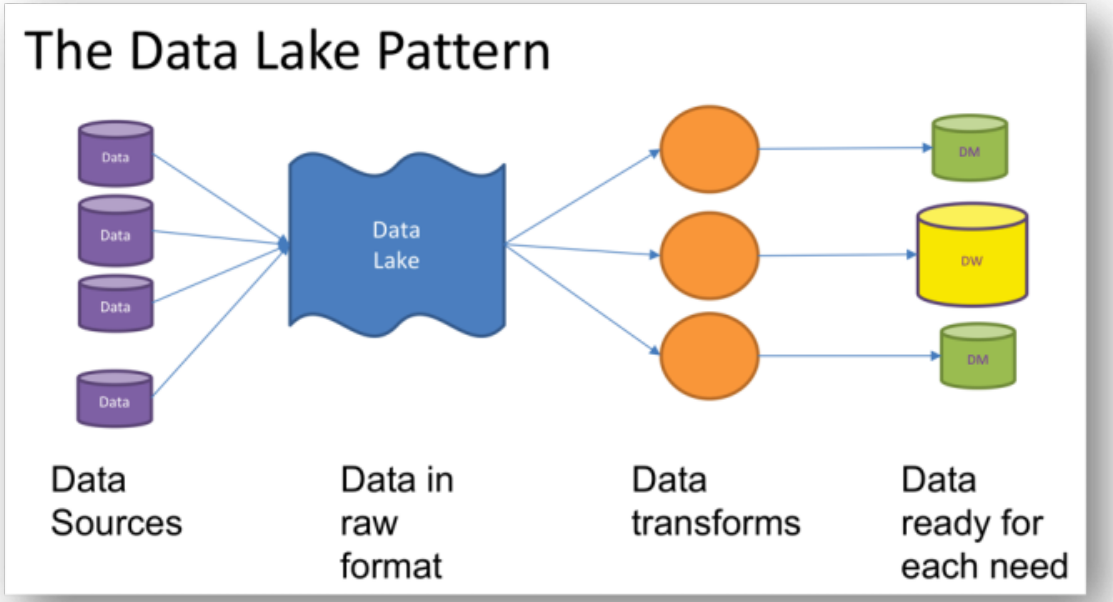
- **Points communs (1/2)**
  - Flexibilité de schémas
    - Chaque enregistrement peut avoir son propre schémas
      - **Problème d'hétérogénéité**
        - ***Problème d'alignement***
          - Structure
          - Syntaxique
          - sémantique



Comment interroger ces schémas multi-structures?

Mono-store

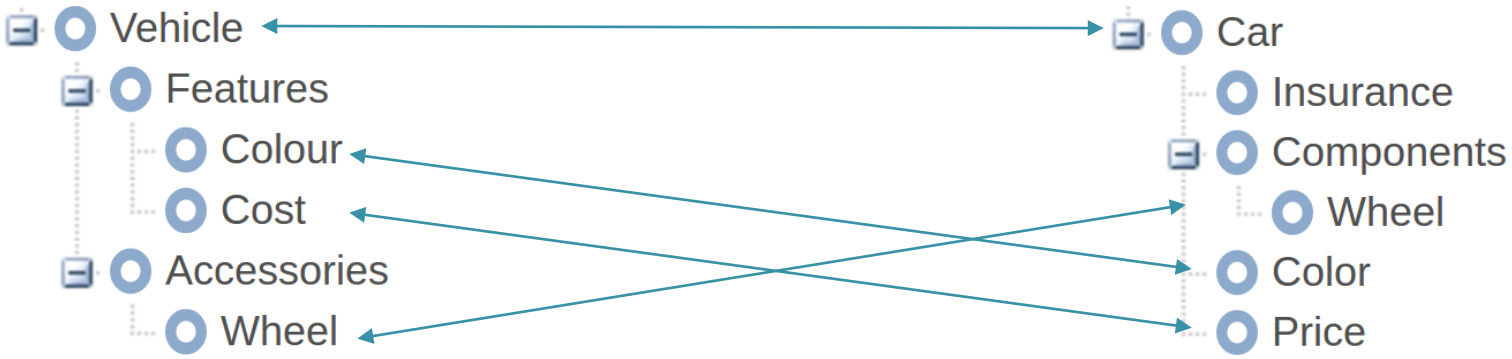
Multi-store



https://medium.com/data-ops/the-data-lake-is-a-design-pattern-888323323c66

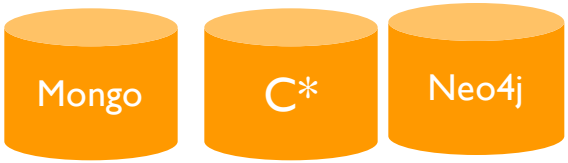
Problématique

- **Multi-structures – mono-store (noSQL - Mongodb)**



db.collection.find({ Vehicule.Colour : 'red' })

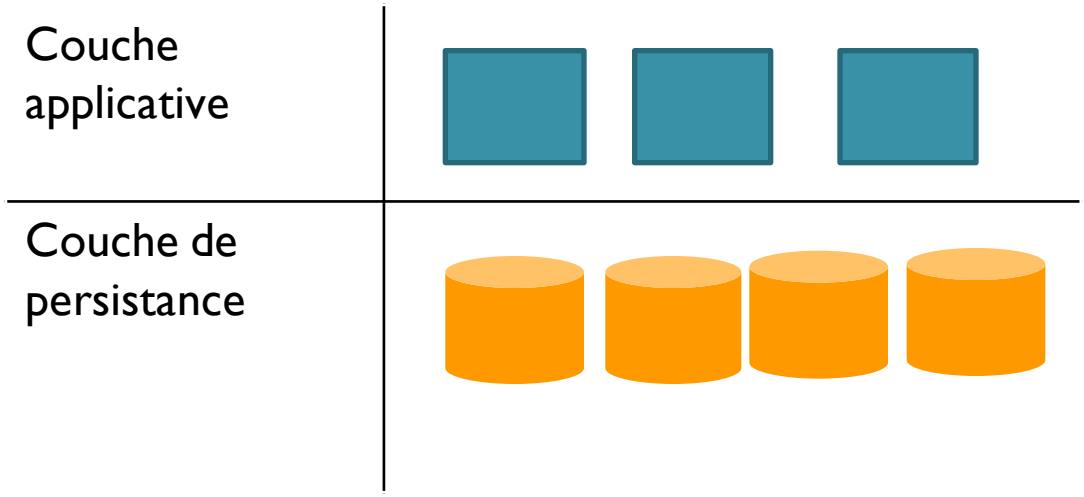
Requête n'est pas suffisamment complète pour obtenir les résultats attendus



Proposition

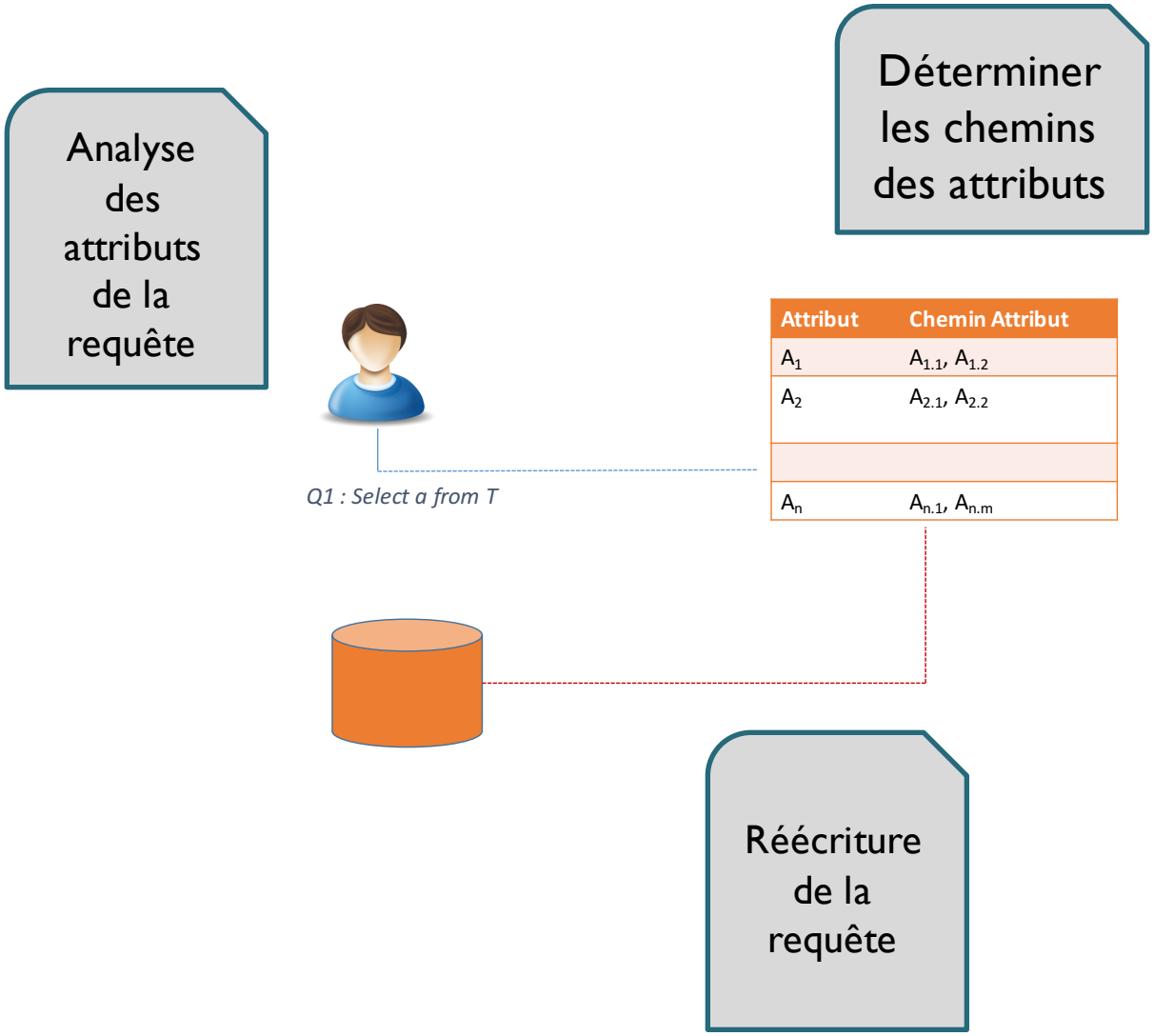
- **Uniformisation des schémas**
  - Métadonnées
  - Contre la philosophie noSQL

- **Multi-structures**
  - Flexibilités des schémas et des modèles
  - Travail consistant au niveau applicatif



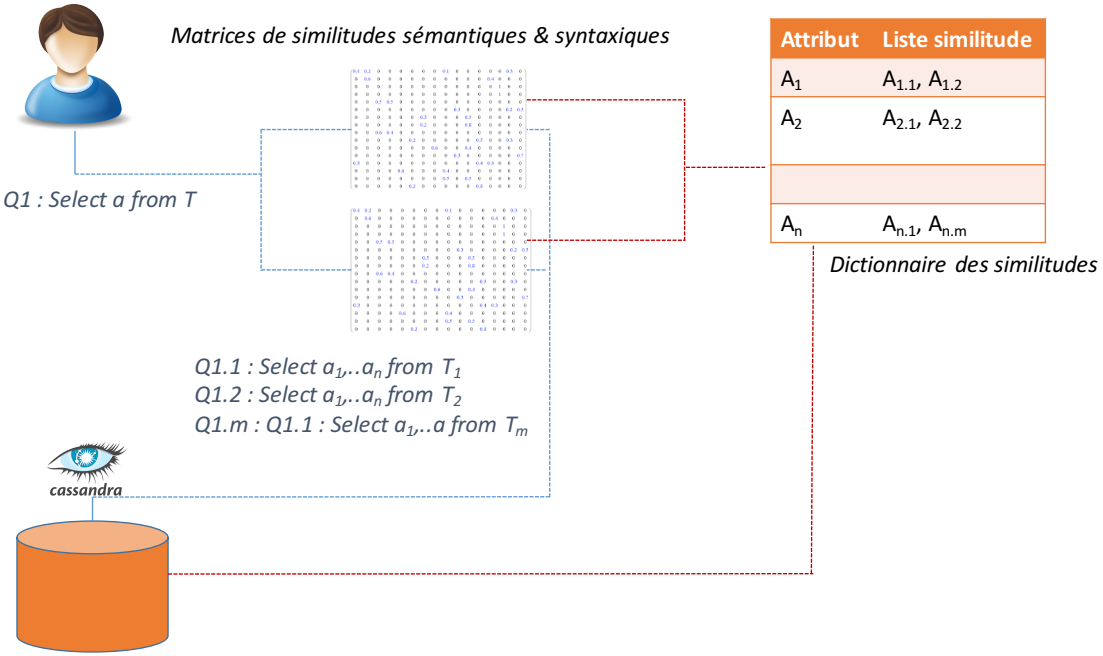
Proposition

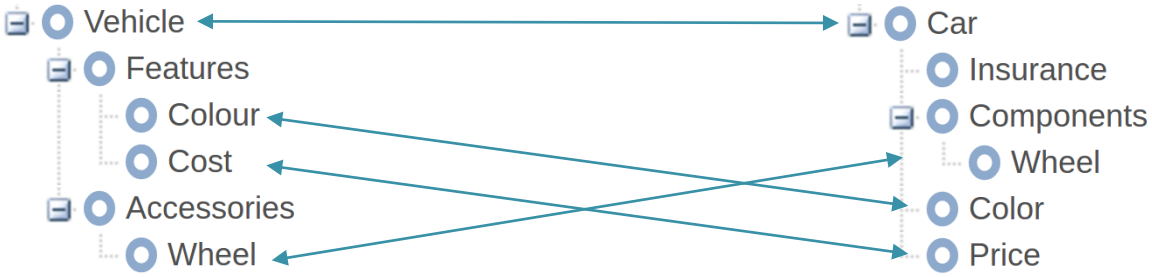
- **Multi-structures – mono modèle (noSQL - Mongodb)**

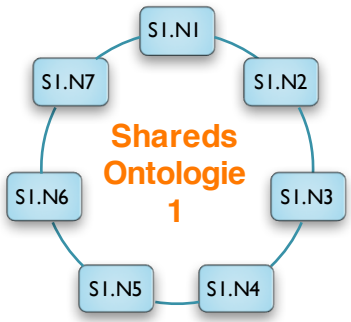


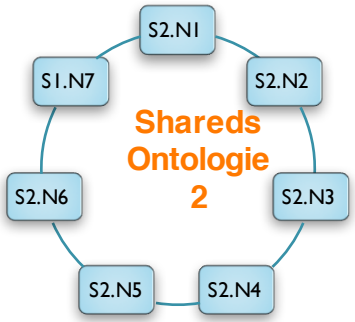
Proposition

- **Multi-structures**



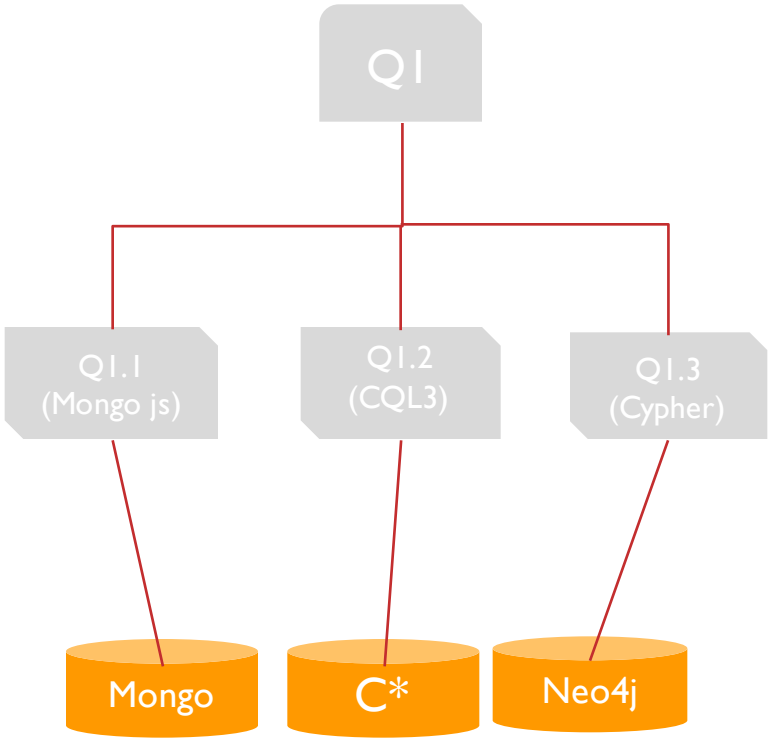






Proposition

- **Récriture de la requête**
  - Déterminer attributs similaires (*select* - *where*);
  - Découpage requêtes (BD reparties)
  - Réécriture requêtes – langages spécifiques



Limites/perspectives

- **Pertinence des résultats (100% ??)**
- **Algorithmes d'alignements conçus pour un environnement mono-machine**
- **A quel niveau de l'architecture doit-on gérer le calcul de similarité**
  - Au niveau de la couche de médiation, toujours?
  - Au niveau de chaque nœud?
  - Au niveau du maitre ?

Merci de votre attention

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.