



Les réseaux de chercheurs

JDEV 2017

A-Phat LY – Yvan Stroppa

Objectifs

- Travailler sur les réseaux de chercheurs
- Pour connaître les relations entre eux
- Pour essayer de déterminer les types de réseaux qu'ils constituent
- Pour essayer de déterminer comment les chercheurs d'une même structure collaborent (intra) entre eux et avec l'extérieur (extra).
- Pour essayer de savoir comment les entités travaillent ensembles
- Pour essayer de savoir comment les disciplines travaillent ensembles : notion d'interdisciplinarité
- ..
- Construire une plateforme applicative autour de ce thème

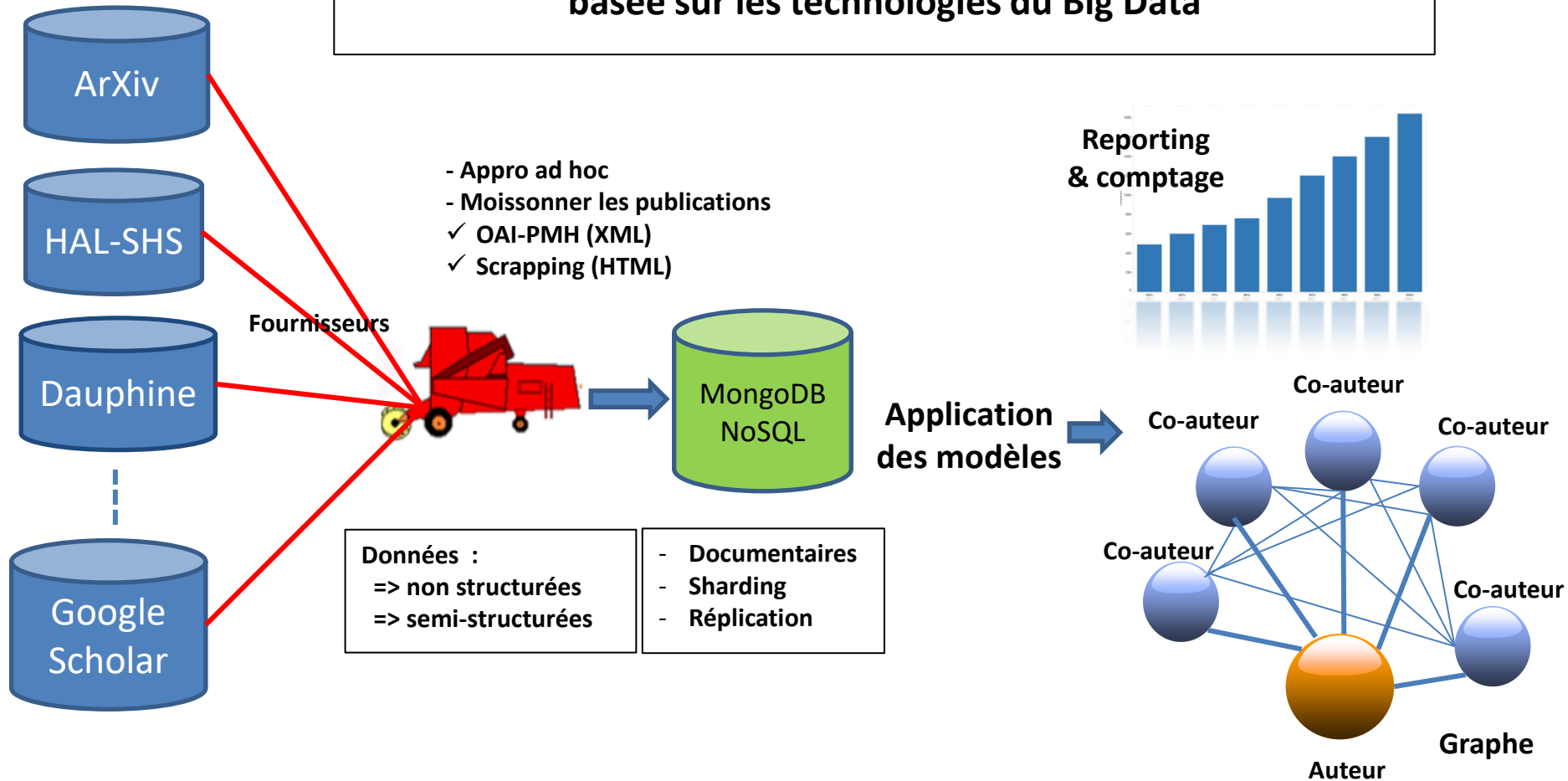
Démarche

- Trouver l'information
 - Hal, arXiv, Google scholar, Dbpl, ...
- Collecter l'information
 - Selon les différentes approches possibles
 - Soit par OAI-PMH
 - Soit par du scraping
- Stocker l'information
 - Structure persistante et scalable
- Structure NoSQL
- Traiter l'information et présenter les résultats
 - Par des outils classiques de type:
 - Javascript, R, python et Neo4j



Représentation du système global

Plateforme d'analyse et de compréhension basée sur les technologies du Big Data



Collecter

Stocker

Traiter

Visualiser

Représentation en couches de l'infrastructure

LES DIFFERENTES COUCHES DE L'APPLICATION

CLIENT
Web

Navigateur
(html, D3.js,
jquery)

Chercheurs
connus
(fichiers json)

APACHE
Couche
intermédiaire

Frontal

Ensemble de
fonctions PHP

TOMCAT
Couches services

Web Services
API-REST (JSON)

INFRASTRUCTURE TRAITEMENTS ET STOCKAGE

APPRO
perl, python
OAI-PMH,
Scrapping
(Hal-SHS,
ArXiv,...)

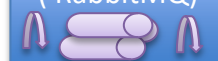


ENTREE ➔

Acces shell
mongo



QUEUE
(RabbitMQ)



PERSISTANCE

Réplication

config

Cluster1

mongos

Cluster2

ClusterN

Visualisation des résultats

- Affichage sous forme de Graphe
- Interaction entre les éléments et possibilité d'étendre les graphes par sélection
- Recherche de similitude entre les différents graphes
- Recherche de relation autour de mot clés
- Calculs intermédiaires des différentes grandeurs associées aux réseaux ainsi reconstruits (avec R et les modules spécialisés graphes : tnet, statnet, igraph)
- Stockage et exploitation de Neo4j

Vue annuelle des collaborations



Status... [2007]--Hurlin C

- Prof
- Resource

Years...

- 2004 ● 2015
- 2005 ● 2016
- 2006 ● 2017
- 2007
- 2008
- 2010
- 2011
- 2012
- 2013
- 2014

Download...

Hurlin_C_1.json

Synthèses...

Articles :6 -->categ A:0 B:0 C:0

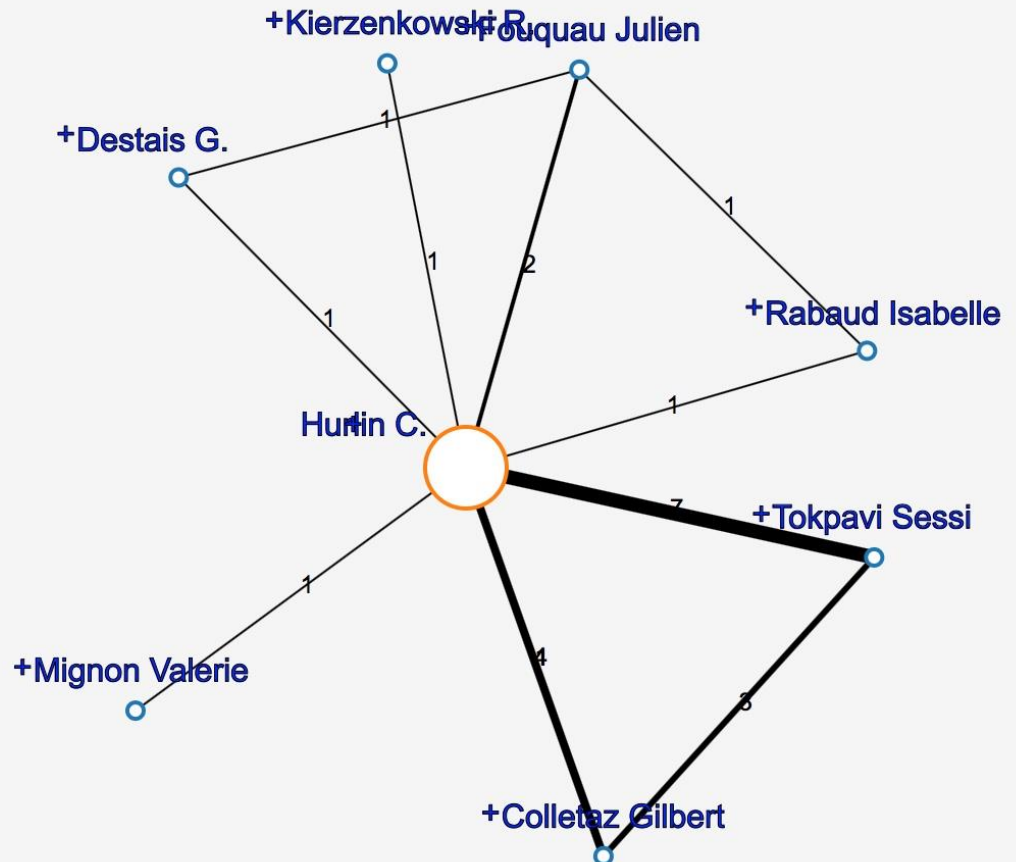
WP :undefined

Comm/conf :2

Chap. Ouvrage :0

Rapport :0

Ouvrage :undefined



Vue globale des collaborations



Status... [Global]--Hurlin C A DEFINIR

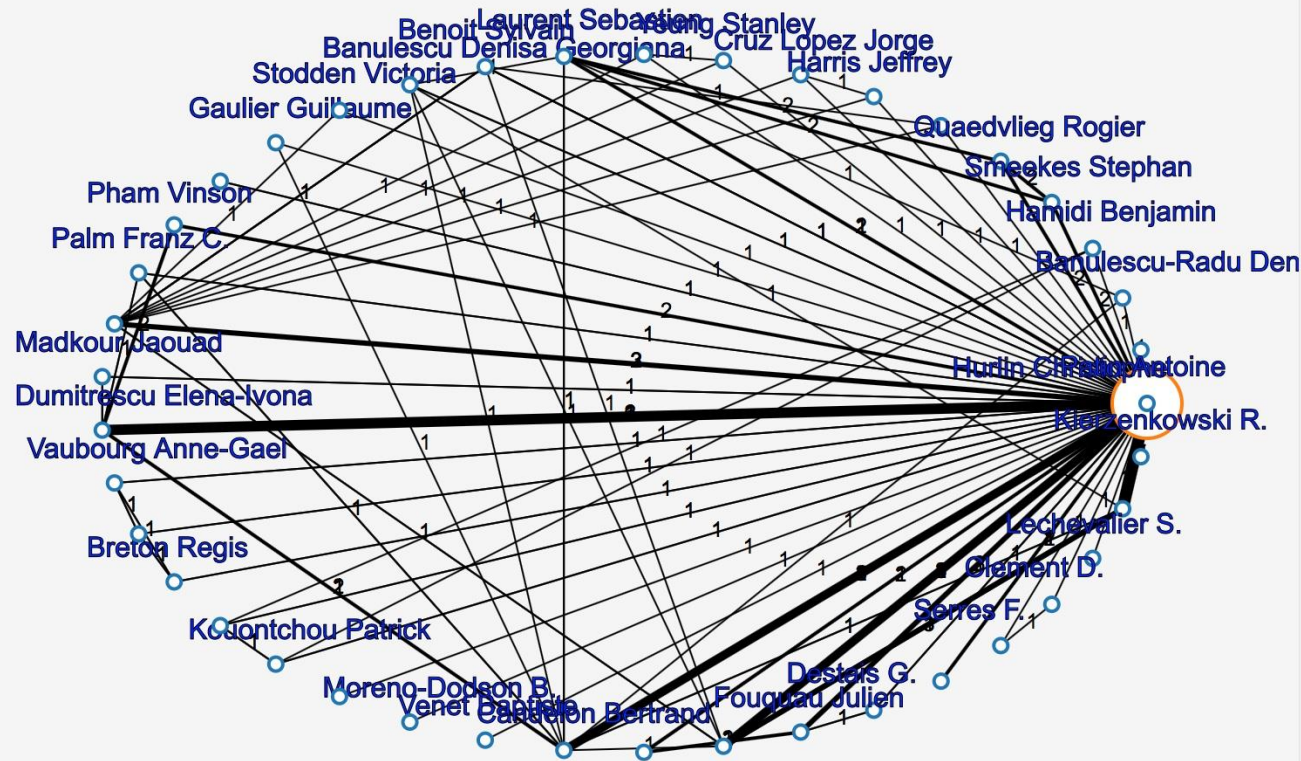
- Prof
- Resource

Years...

- 2004 ● 2015
- 2005 ● 2016
- 2006 ● 2017
- 2007
- 2008
- 2010
- 2011
- 2012
- 2013
- 2014

Download...

Hurlin_C_1.json



Extension du graphe à partir d'un auteur



Status... [2007]--Hurlin C 1;Tokpavi Sessi 1;

- Prof
- Resource

Years...

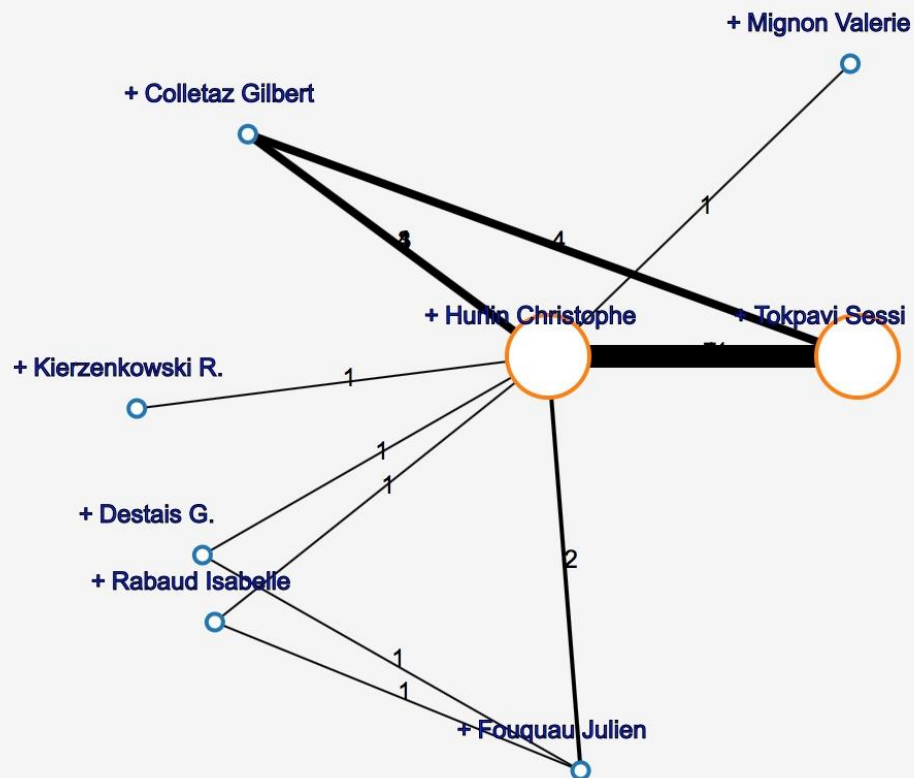
- 2004 ● 2015
- 2005 ● 2016
- 2006 ● 2017
- 2007
- 2008
- 2010
- 2011
- 2012
- 2013
- 2014

Download...

Hurlin_C_1.json

Synthèses...

Articles :135 -->categ A:3 B:3 C:3
WP :undefined
Comm/conf :3
Chap. Ouvrage :undefined
Rapport :undefined
Ouvrage :undefined
Other Publications :3



Visualisation des publications



<div> <div>Show</div> <div>10</div> <div>entries</div> </div>				
Title	Type	Co-authors	Year	
Backtesting Value-at-Risk Accuracy: A New Simple Test	Journal articles	Hurlin Christophe Tokpavi Sessi	2007	
Credit Market Disequilibrium in Poland: Can we find what we expect? Non stationarity and the Short Side Rule	Journal articles	Hurlin Christophe Kierzenkowski R.	2007	
Energy Demand Models: A Threshold Panel Specification of the Kuznets Curve	Journal articles	Destais G. Fouquau Julien Hurlin C.	2007	
How to Estimate Public Capital Productivity?	Preprints, Working Papers, ...	Hurlin Christophe	2007-06-22	
Irregularly Spaced Intraday Value at Risk (ISIVaR) Models : Forecasting and Predictive Abilities	Preprints, Working Papers, ...	Hurlin Christophe Colletaz Gilbert Tokpavi Sessi	2007-07-13	
Irregularly Spaced Intraday Value at Risk (ISIVaR) Models: Forecasting and Predictive Abilities	Other publications	Colletaz Gilbert Hurlin Christophe Tokpavi Sessi	2007	
Irregularly Spaced Intraday Value-at-Risk (ISIVaR) Models: Forecasting and Predictive Abilities	Conference papers	Hurlin Christophe	2007-09-24	
Irregularly Spaces Intraday Value-at-Risk (ISIVaR) Models: Forecasting and Predictive Abilities	Other publications	Colletaz Gilbert Hurlin Christophe Tokpavi Sessi	2007	
Modeles a changement de regimes et macro-economiques	Other publications	Colletaz Gilbert Hurlin Christophe	2007	
Modeles Non Lineaires et Previsions	Preprints, Working Papers, ...	Colletaz Gilbert Hurlin Christophe	2007-06-22	

Choix de la solution de stockage

- Qui doit répondre :
 - Structure de données différentes entre les sources d'approvisionnement
 - Données évolutives (elles se réactualisent de façon indépendante)
 - Données redondantes (entre les sources et dans les sources)
- Solutions :
 - ➔ constituer des clés d'indexation d'une même production
 - ➔ Constituer des clés pour les auteurs

Choix d'une solution de stockage

- MongoDB
 - Flexibilité
 - Simplicité
 - Performance
 - Pour ses dispositifs
 - Map-reduce
 - Cluster
 - Sharding

Mise en oeuvre dans MongoDB

- Une fois les fichiers extraits des sources sous forme XML (OAI-PMH) ou sous forme texte pour Google Scholar
- Construction et utilisation :
 - Des scripts perl de parsing et d'importation dans mongoDB
- Pour produire
 - Databases et collections
 - Collections de documents

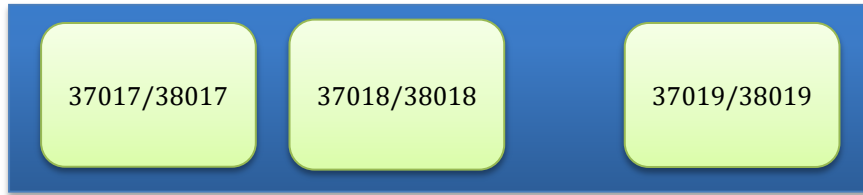
Détails pour Mongo

- Se connecter :
 - `mongo --port XXXX --host NNNNN`
- Ouverture d'un terminal sous mongod
 - Actions
 - `show dbs`
 - `show collections`
 - Exemple de scripts d'extractions :
 - Exemple d'extractions :simples :
 - `db.publications.find({"coauthors":{"$regex":"Hurlin"},"options":"i"},"title":1,"_id":-1})`
 - Regroupement :
 - `db.publications.aggregate({$group:{_id:{date:{$substr:["$date",0,4]}}},nb:{$sum:1}}},{ $sort:{_id: -1}})`
 -

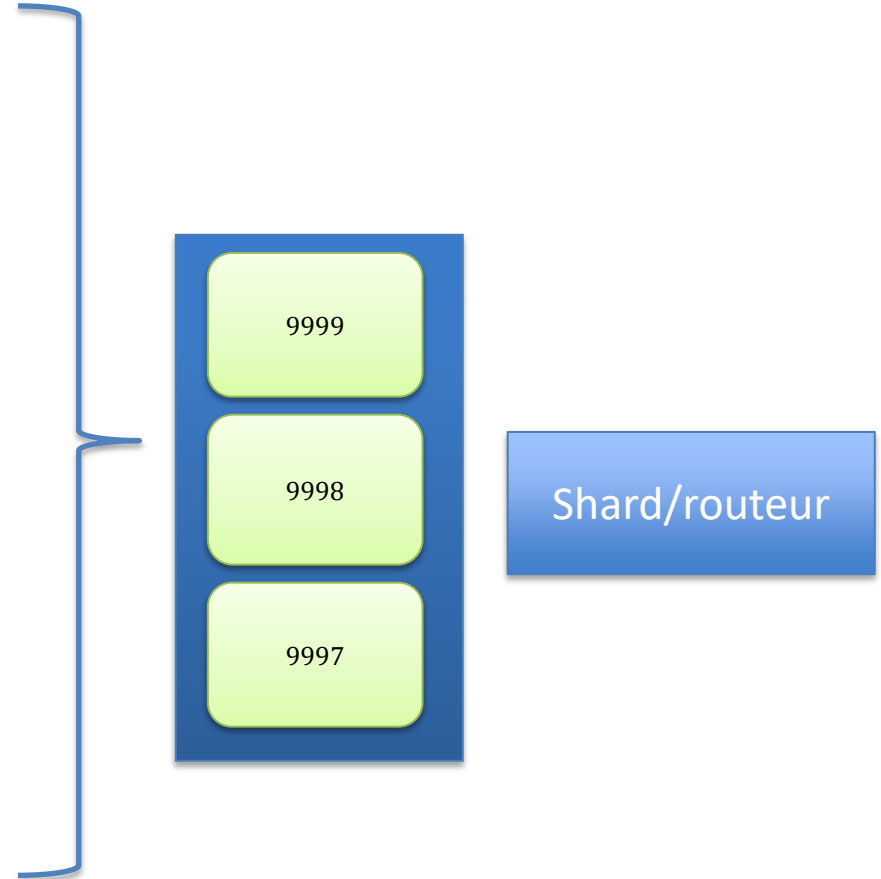
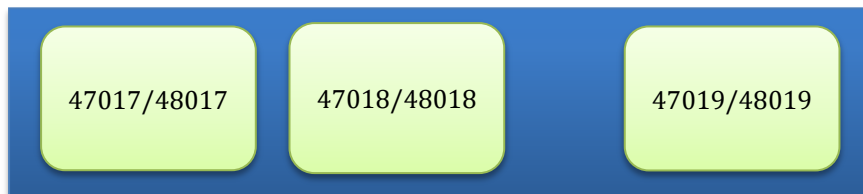
Première étape

Architecture utilisée

Cluster A



Cluster Z



Cluster de configuration

Réplication sur du modèle Maître/esclaves

Le Clustering sous Mongo

- Simple à monter
 - Fichier de configuration
- Peut-être mono machine ou Multi-machines (physiques ou virtuelles)
- Containers (docker)
- Mode Maître-Esclaves
 - Un seul master autorisé à écrire
 - Les autres sont en accessibilité limitée
 - Si le master n'est plus disponible, élection entre les esclaves et réélection d'un nouveau master

Clustering

where to write logging data.

systemLog:

destination: file

logAppend: true

path: /home/jdev/clusterA/instanceA_1/journal/mongod.log

1

Where and how to store data.

storage:

dbPath: /home/jdev/clusterA/instanceA_1/data

journal:

enabled: true

how the process runs

processManagement:

fork: true

pidFilePath: /home/jdev/clusterA/instanceA_1/mongod.pid #

network interfaces

net:

port: 37017

http:

enabled: true

JSONPEnabled: true

RESTInterfaceEnabled: true

replication:

repSetName: "clusterA"

Initialisation du cluster et ajout
des instances participantes

#mongo --port 37017

>rs.initiate()

>rs.add("M1:37018")

>rs.add("M1:37019")

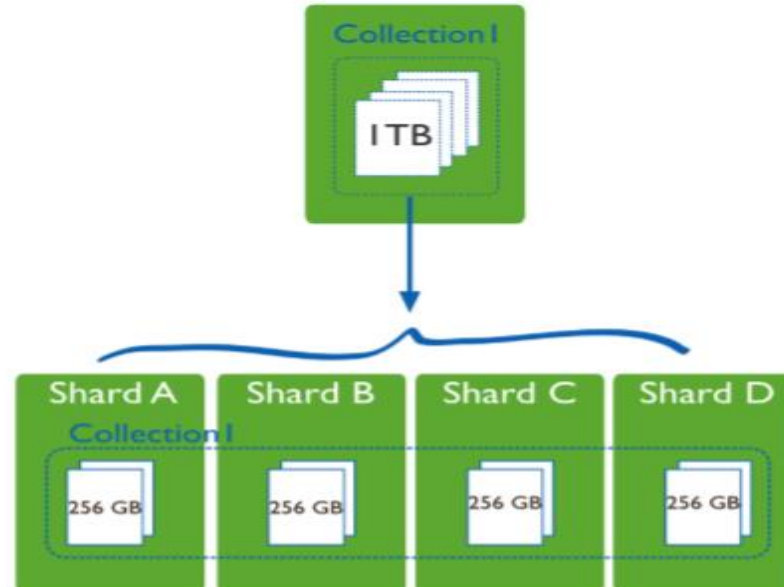
>rs.status()

2

Indication de l'appartenance de
l'instance au cluster

Sharding

- Possibilité de répartir les datas sur l'infrastructure distribuée
 - Règle à fixer et à définir au niveau du shard



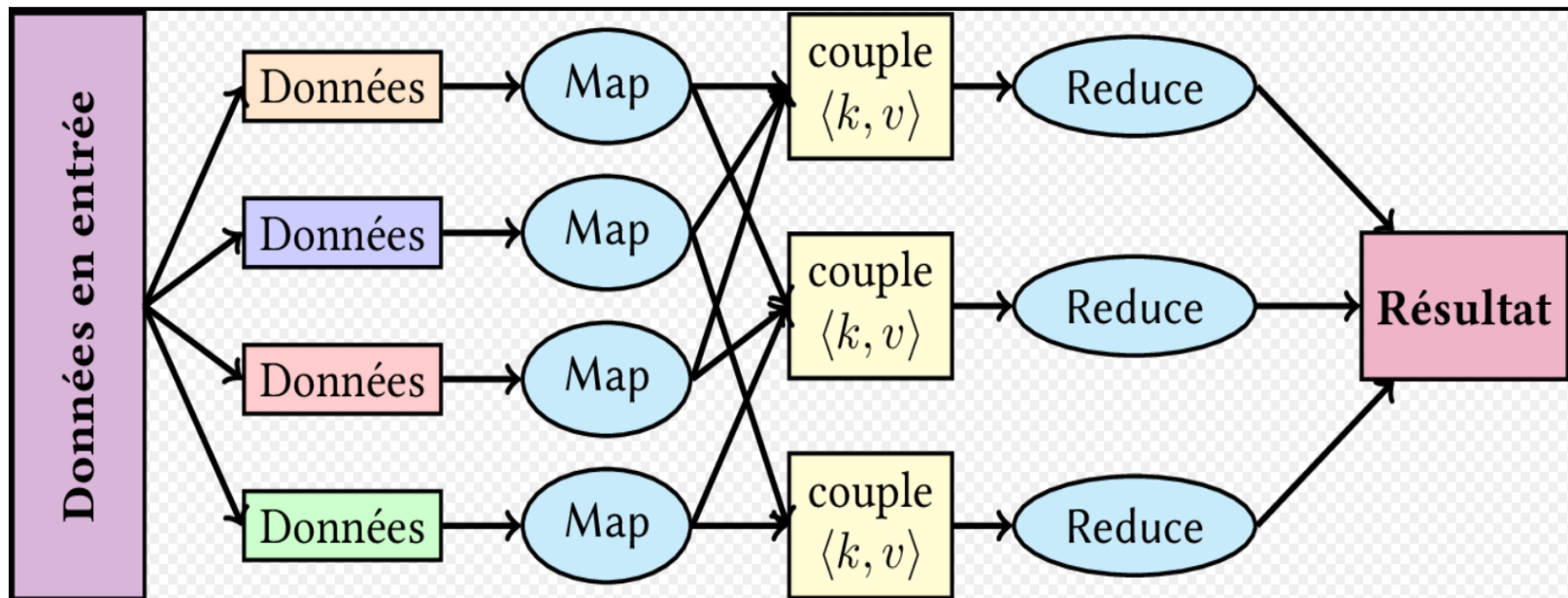
Sharding

- Une fois connecté sous mongos :
- Ajout des clusters au SHARD de Mongo
 - `sh.addShard("clusterA/math2I51:37018, math2I51:37017, math2I51:37019")`
 - `sh.addShard("clusterB/math2I51:47018, math2I51:47017, math2I51:47019")`
- Initialisation de la database cible
 - `use Hal;`
- Création de la collection et de l'index pour la répartition :
 - `db.createCollection("publications");`
 - `db.collection.createIndex({date:1})`
- Activation de la répartition
 - `sh.enableSharding ("Hal")`
 - `sh.shardCollection("Hal.publications",{date:1})`
- On peut importer les données à partir du routeur qui utilisera les règles pour distribuer les datas sur les clusters.

Deuxième étape du traitement distribué

Map-reduce sous mongodb

- Le but est d'utiliser les possibilités de map-reduce pour exécuter des traitements



Une exécution au crible

- Le map est défini par une fonction Javascript construit à partir de la console ou d'un outil graphique (robomongo)
 - n'a pas d'argument
 - est associé à un et un seul document
 - c'est le corps pour le traitement
 - émet une ou plusieurs structures de données {key, values}

Reduce

- Est déclenché après le Map
- Reçoit une entrée {key, values}
- Traite selon le contexte des valeurs reçues
- Ne peut pas accéder aux documents
- Et émet également un couple {key, values}
- La value doit être de la même structure que la sortie de la Map

Finalize

- Permet d'effectuer un traitement de fin pour chaque key.

Merci de votre attention