

France Labs  
Open Source Enterprise Search

## *Big Search*

Aurélien MAZOYER

Olivier TAVARD

# Plan

- ❑ **Solr/Constellio**
- ❑ Hadoop
- ❑ Démonstration par l'exemple
- ❑ Démonstration Hadoop/Solr



# A quoi ressemble un moteur de recherche aujourd'hui?

Espace utilisateur ▼ Recherche avancée Nouvelle recherche

ingénieur

ingénieur r&d  
ingénieur & développeur systèmes linux  
ingénieur analyste  
ingénieur bio-informaticien  
ingénieur chargé d'étude

Expression exacte

Essayez avec cette orthographe : ? ingénieur

Candidats > Résultats 1 - 10 sur un total de 100 (10 pages)

## ☑ Département

Tri : Occurences ▼

Hauts-de-Seine (11)

Paris (11)

Yvelines (8)

Val-de-Marne (6)

Alpes-Maritimes (5)

Rhône (5)

Ain (3)

Haute-Garonne (3)

Seine-et-Marne (3)

Bouches-du-Rhône (2)

### 🗄 Ingénieur Développement - Strasbourg(Bas-Rhin) ★

... Bas-Rhin  
**Ingénieur** Développement  
Strasbourg ...

### 🗄 Ingénieur réseau - Cergy(Val-d'Oise) ★

... cergy  
**Ingénieur** Réseaux et Technologie de l'Information DESS,[...]  
**Ingénieur** réseau ...

### 🗄 Ingénieur support opérationnel clients - Rueil malmaison(Hauts-de-Seine) ★

... Certification ITIL Foundation in IT Service Management débutant  
**Ingénieur** informatique Informatique télécoms et réseaux [...]  
**Ingénieur** support opérationnel clients ...

# Qu'y a-t-il dans Constellio?

## Un oignon



# Qu'y a-t-il dans Constellio?

## Lucene

- ❑ Créé en 2000 par Doug Cutting. Version Actuelle : Lucene v. 3.6 (Avril 2012)
- ❑ Projet Open Source, Apache depuis 2001
- ❑ Librairie de recherche “full-text”
- ❑ Rapide, stable, performant, modulable
- ❑ 100% Java (pas de dépendances)



# Lucene

## Indexation

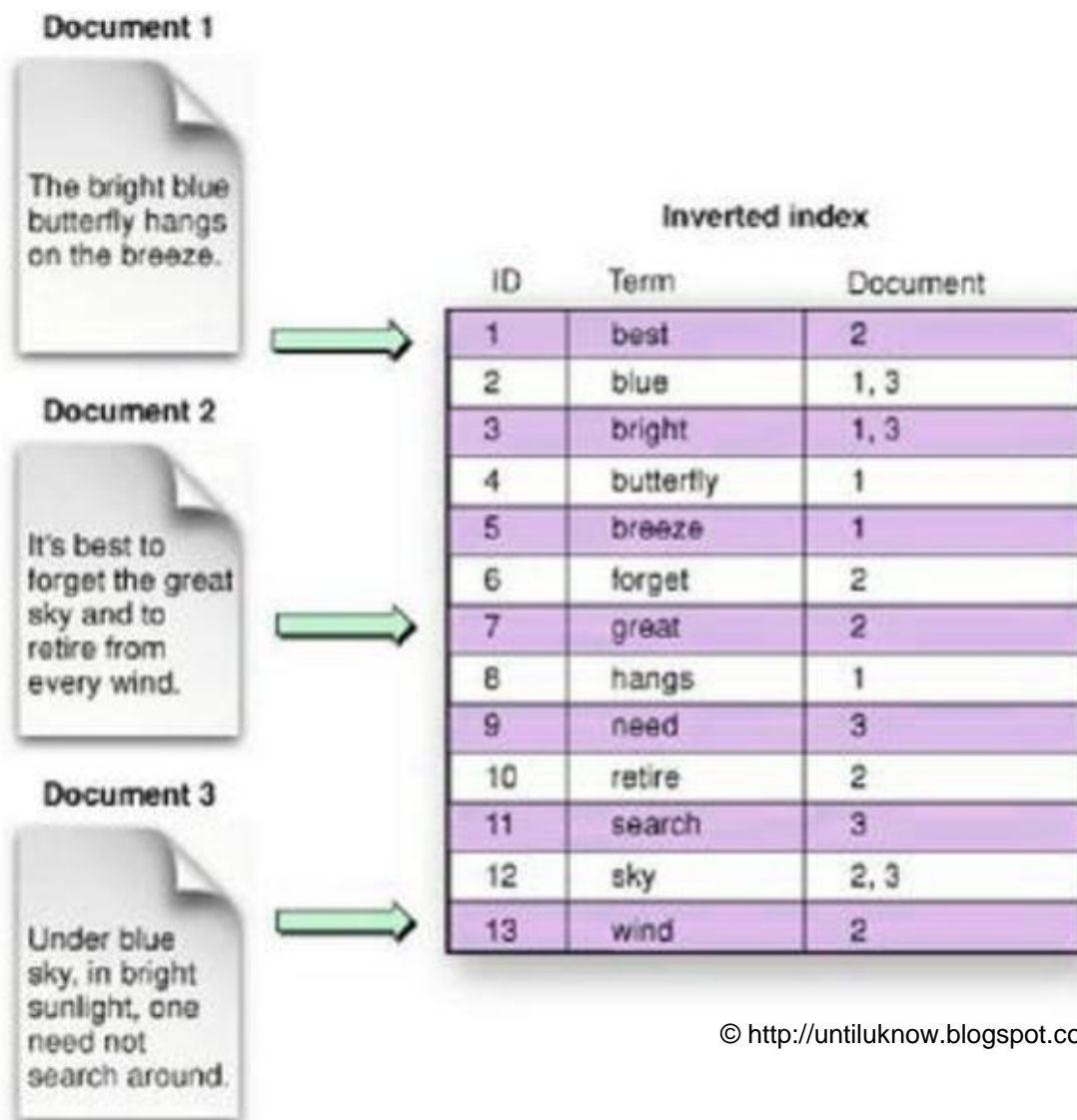
### □ Un outil qui permet:

- De créer un index à partir de documents



# Lucene

## Index inversé



- Un outil qui permet:
  - De créer un index à partir de documents
  - D'effectuer des recherches dans cet index

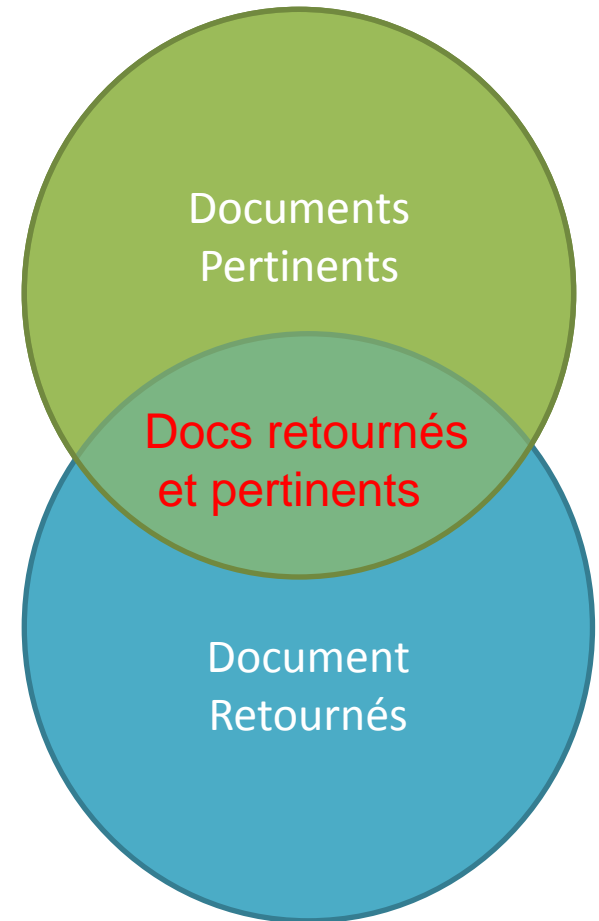


INDEX



## Recherche pertinente

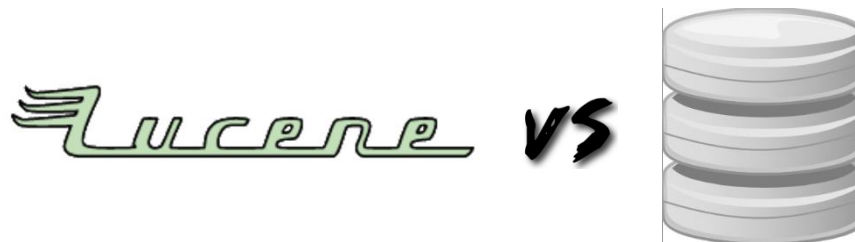
- ❑ Récupérer les bons résultats...  
... et seulement ceux là
- ❑ Precision
  - Pourcentage de docs pertinents sur les docs retournés
- ❑ Recall
  - Pourcentage de docs pertinents retournés sur le total des docs pertinents



***Trouver un bon compromis...***

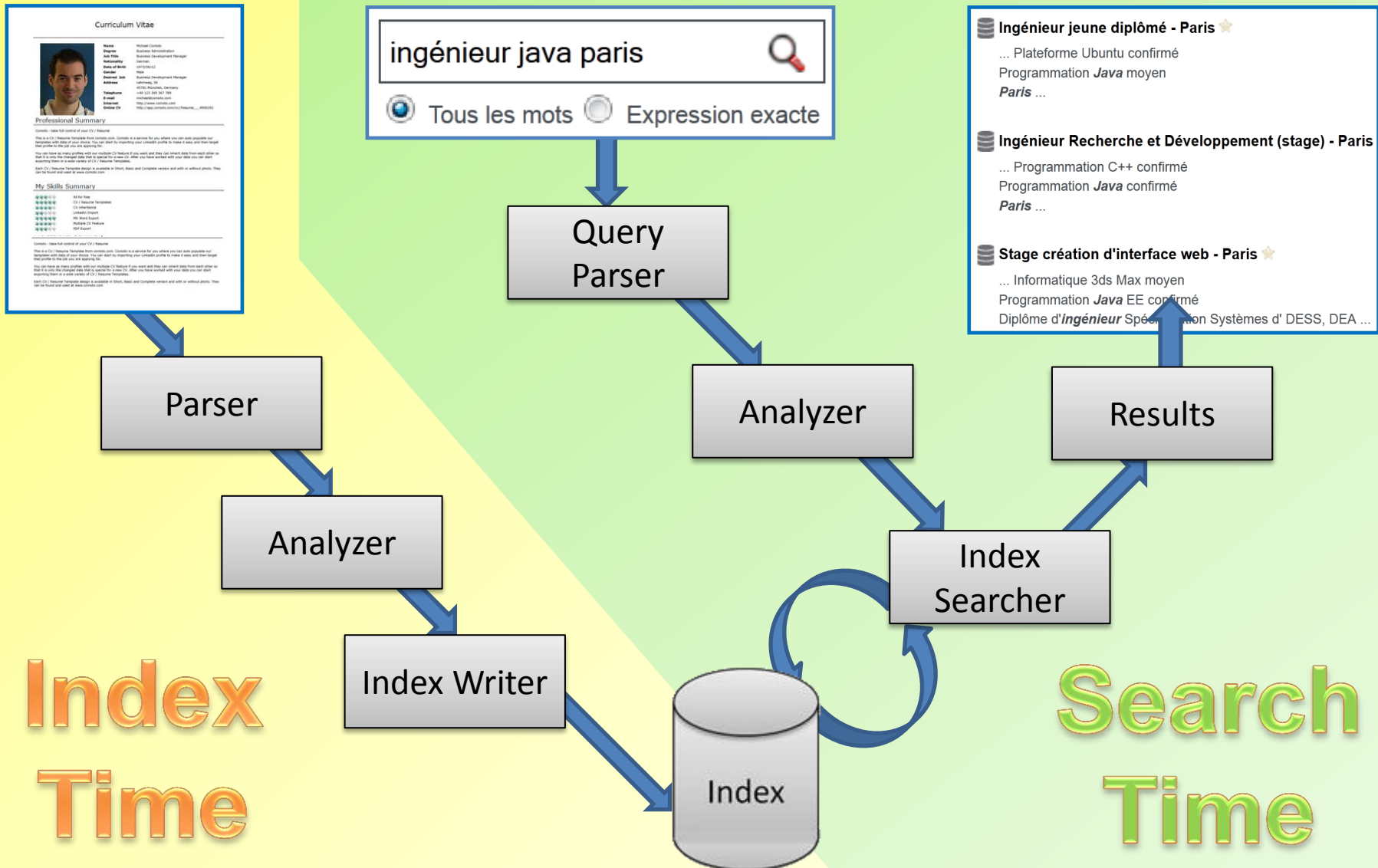
## Différence avec une base de données

- ❑ Plus rapide pour récupérer un doc à partir de son contenu
- ❑ Résultats scorés
- ❑ Non relationnelle, structure non fixe
- ❑ Champs qui peuvent contenir plusieurs valeurs



# Lucene

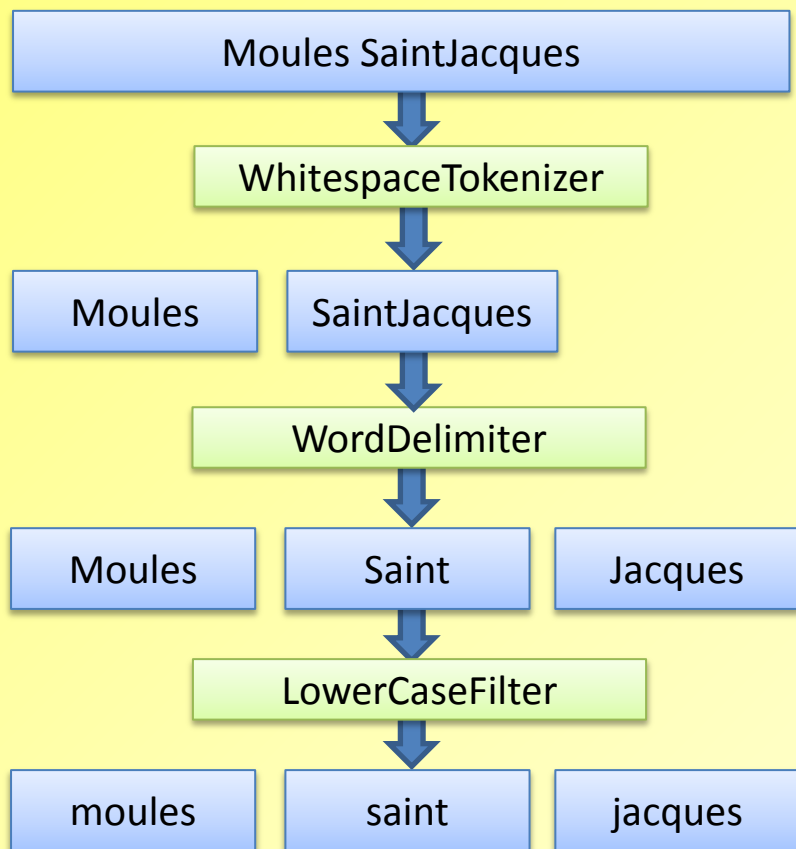
## Indexation et Recherche



# Lucene

## Analyzers

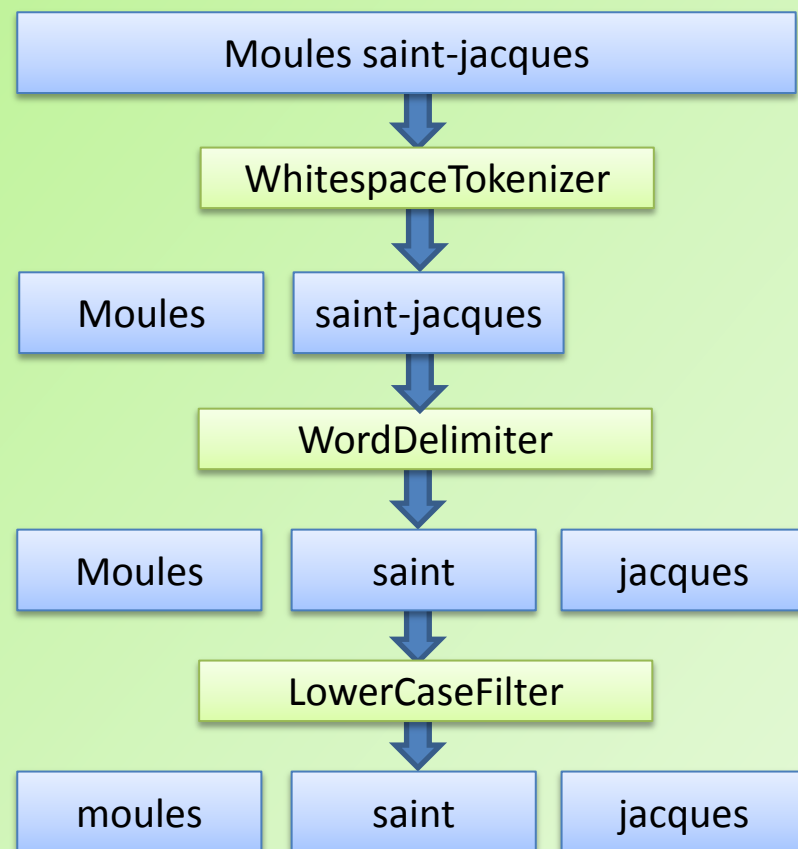
### Document Analysis



**Index  
Time**

**MATCH!!!**

### Query Analysis



**Search  
Time**

# Lucene

## Scoring

$$\sum_{t \text{ in } q} (tf(t \text{ in } d) \times idf(t) \times boost(t, \text{field in } d) \times lengthNorm(t, \text{field in } d)) \times coord(q, d) \times queryNorm(q)$$

- ❑ Formule paramétrable
- ❑ Combinaison de
  - Boolean Model
  - Vector Space Model
    - Term Frequency
    - Inverse Document Frequency
    - ...

## Pourquoi n'est pas suffisant?

- ❑ Simple bibliothèque
- ❑ Besoin d'une couche serveur

# Qu'y a-t-il dans Constellio?

## Solr



# Solr

- ❑ Lucene « embarqué » dans une webapp
- ❑ Créé en 2004 par Yonik Seeley à CENT Networks
- ❑ In 2006, Solr devient open-source et été cédé à la Apache Software Foundation
- ❑ En 2010, fusion des projets Lucene et Solr
- ❑ Version Actuelle : Solr 3.6 (Avril 2012)



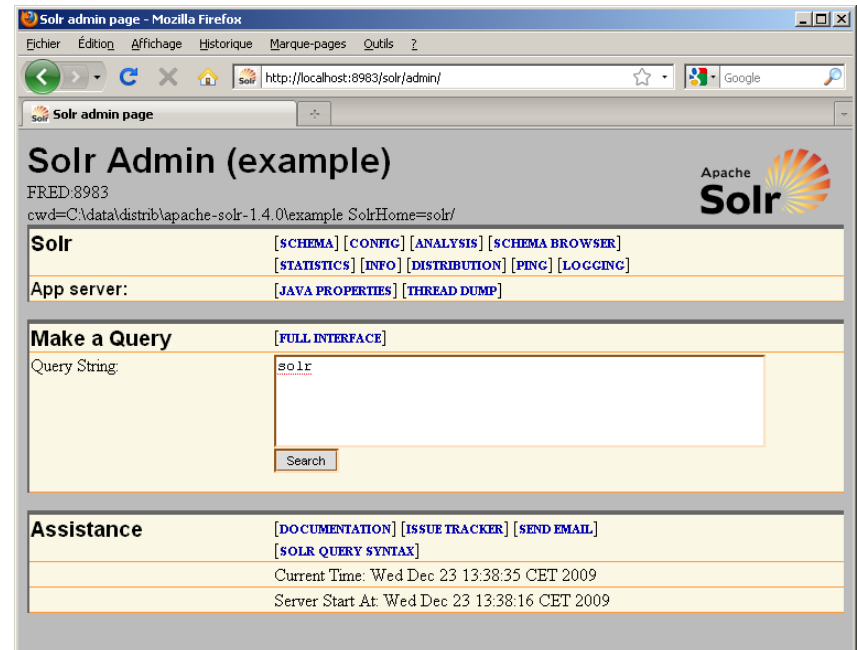


- ❑ APIs XML/HTTP de type REST
  - ajouter des documents (POST)
    - <http://localhost:8983/solr/update>
  - effectuer des recherches (GET)
    - <http://localhost:8983/solr/select>
- ❑ Configuration par fichiers XML
- ❑ Mécanisme de Cache, Réplication
- ❑ Interface admin web

# Pas suffisant?

## Solr

- ❑ Pas de gestion de la sécurité
- ❑ Pas de connecteurs
- ❑ Interface web « limitée »



# Qu'y a-t-il dans Constellio?

## Constellio



- ❑ Interface Web 2.0
- ❑ Sécurité
  - Gestions des ACLs sur les fichiers
  - Connexion avec un LDAP
  - Gestion du SSO (Kerberos, SAML)

- ❑ Compatible avec les Google Connectors (open source)
  - Http
  - File
  - Database (MySQL, Oracle)
  - GED (Alfresco, Nuxeo)
  - Mail
  - XML



# Plan

- ❑ Solr/Constellio
- ❑ **Hadoop**
- ❑ Démonstration par l'exemple
- ❑ Démonstration Hadoop/Solr



# Hadoop

## Vue d'ensemble

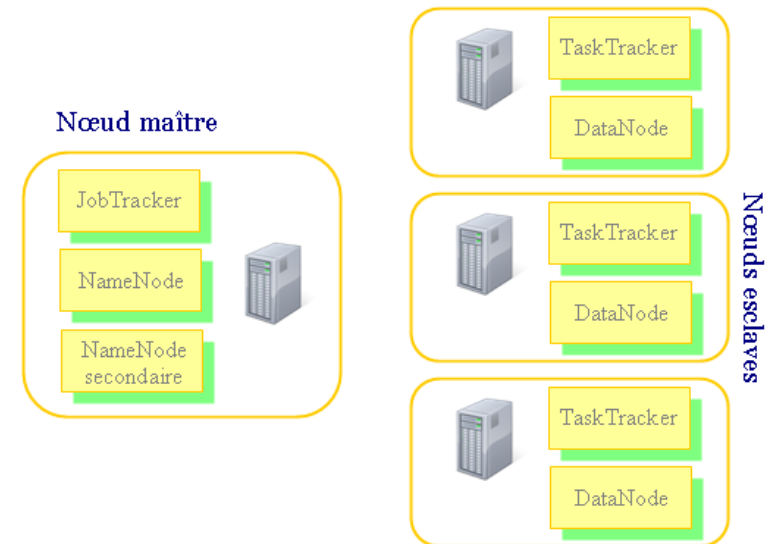
- ❑ Créé par Doug Cutting
- ❑ Framework open source
- ❑ Inspiré par les papiers sur Google Map Reduce et Google File System



# Hadoop

## HDFS

- ❑ Données converties en blocs et distribuées sur des nœuds
- ❑ Chaque bloc est répliqué



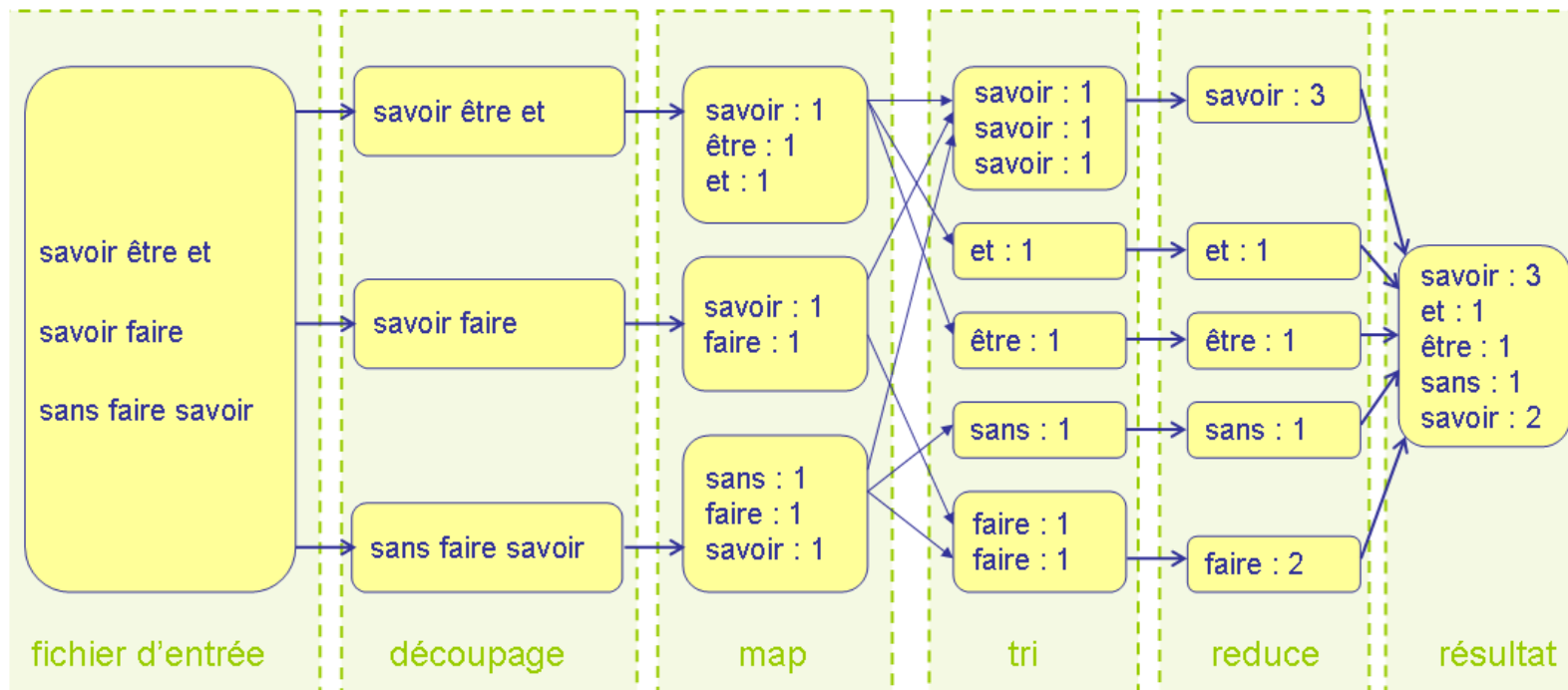
© INOVIA CONSEIL



# Hadoop

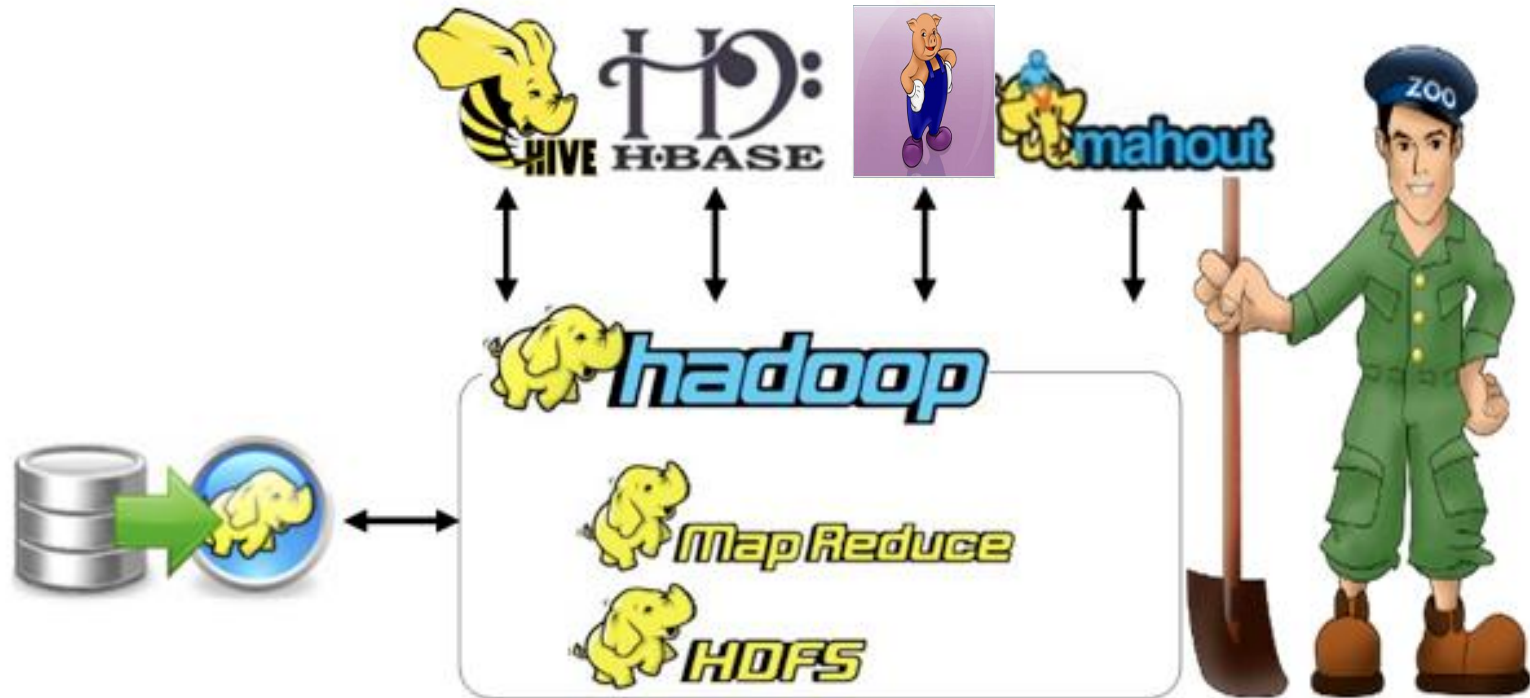
## Map/Reduce

- ❑ Map : données sous forme clés/valeurs
- ❑ Reduce : fusion par clé pour former résultat



# Hadoop

## Ecosystème Hadoop



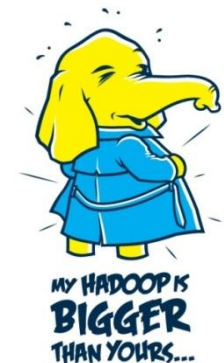
<http://cloudstory.in/2012/04/introduction-to-big-data-hadoop-ecosystem-part-1/>



# Hadoop

## Avantages

- ❑ Traiter des grands volumes de données en un minimum de temps
- ❑ Stocker des immenses volumes de données : plusieurs To ou même Po
- ❑ Fonctionne sur machines de configuration faible et peu coûteuses



# Plan

- ❑ Solr/Constellio
- ❑ Hadoop
- ❑ **Démonstration par l'exemple**
- ❑ Démonstration Hadoop/Solr



# Démonstration par l'exemple

## Big Search dans la vraie vie

- Exemples d'entreprises utilisant différentes technologies pour différents scénarios BIIIG
  - Hadoop
  - Hadoop / Solr
  - MapReduce / Search
  - Solr

The Google logo, featuring the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red) with a trademark symbol.

# Démonstration par l'exemple

## Google

- ❑ 1 000 000 000 000 d'URLS uniques (2008)
- ❑ Pagerank : le ranking d'une page est estimé par sa popularité plutôt que par son contenu

Google™  
PageRank

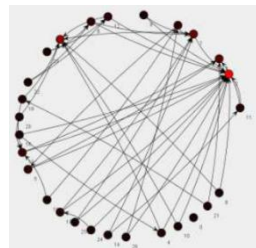
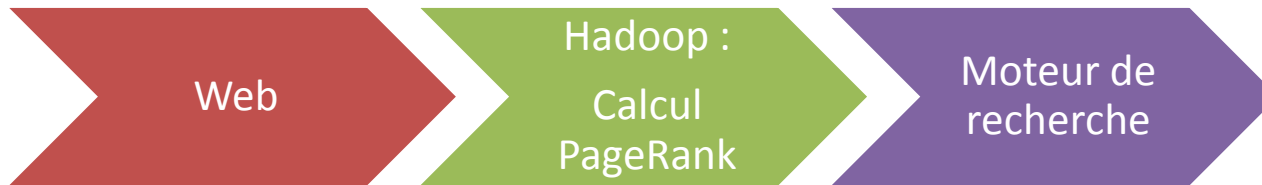


Google

# Démonstration par l'exemple

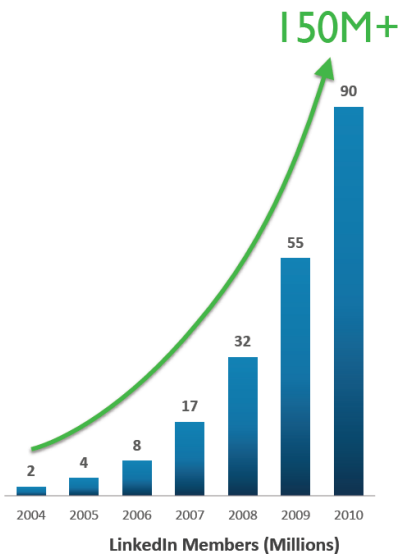
## Google

- Construire PageRank grâce à Map/reduce



# Démonstration par l'exemple

## LinkedIn



**75%**

Fortune 100 Companies use LinkedIn to hire



**>2M**

Company Pages





**~4B**

Searches in 2011

## People You May Know

**People You May Know**

-  **Cedric Ulmer**  
Président de France Labs  
Région de Nice , France · Technologies et services de l'information
-  **Aurélien MAZOYER**  
Search Technologies Expert - Co-founder of France Labs  
Région de Nice , France · Logiciels informatiques
-  **Olivier Tavard**  
Search Technologies Expert, Co-Founder at France Labs  
Région de Nice , France · Technologies et services de l'information

[See more »](#)

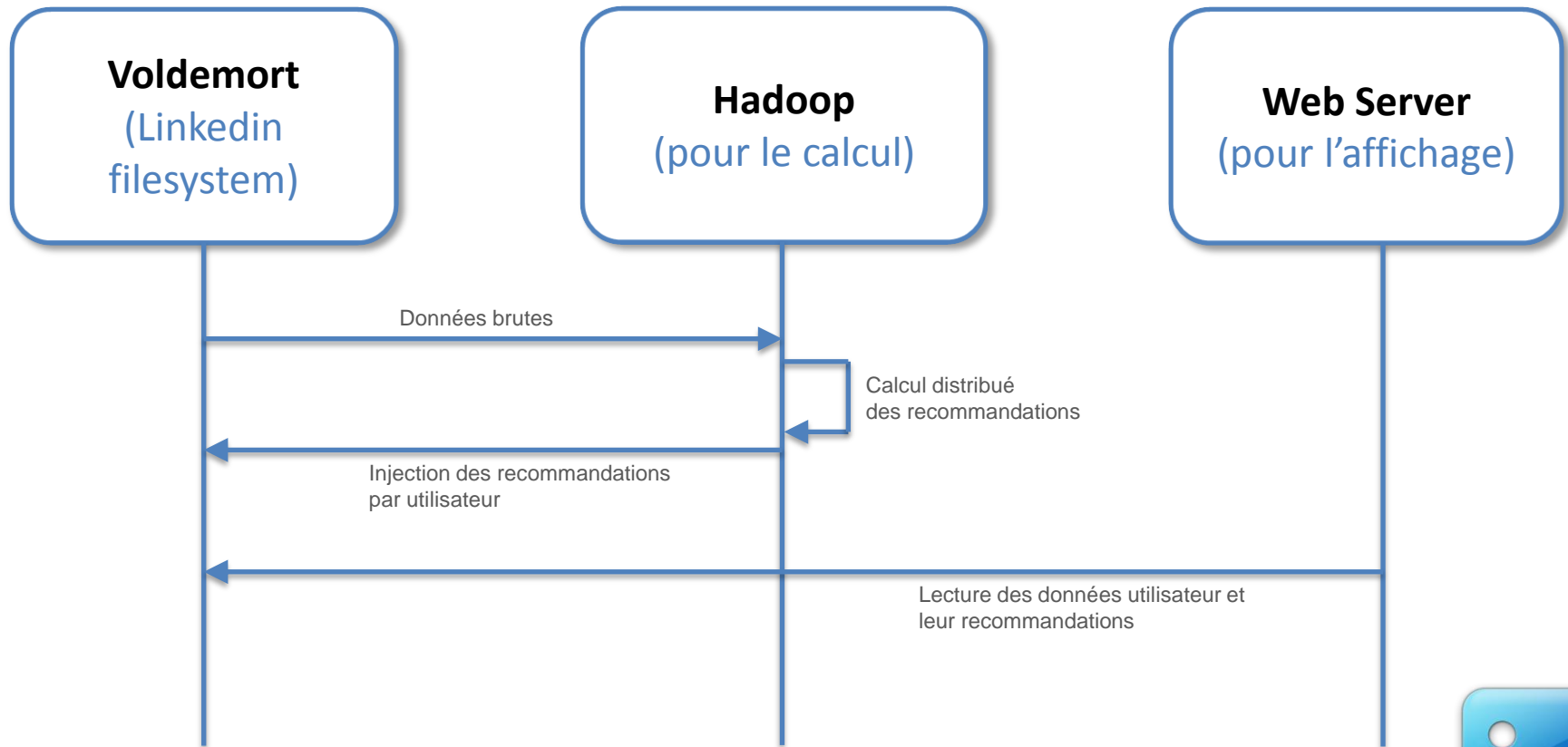


France Labs  
Open Source Enterprise Search



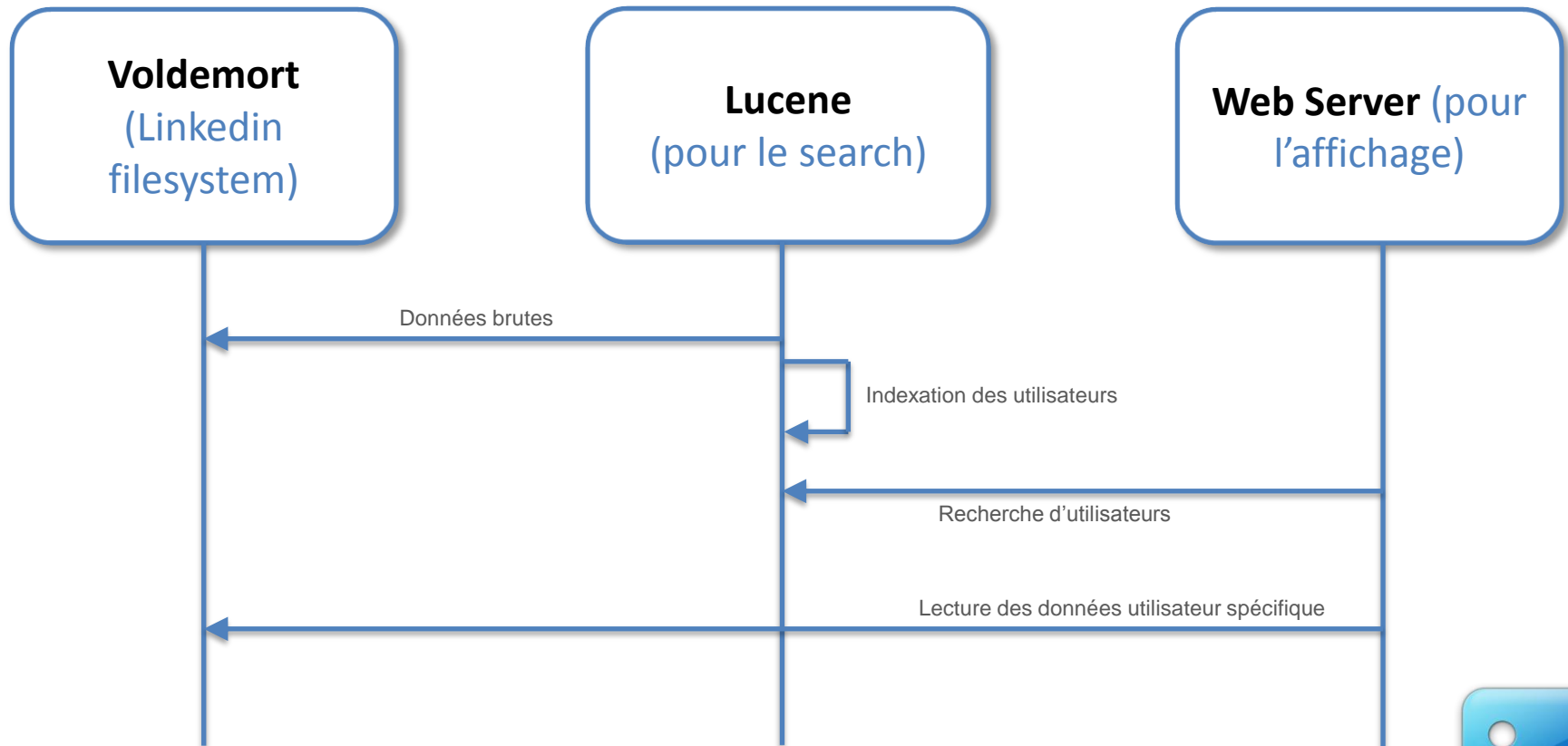
# Démonstration par l'exemple

## 1<sup>er</sup> cas : Hadoop pur pour les recommandations



# Démonstration par l'exemple

## 2<sup>e</sup> cas : Lucene pur pour la recherche d'utilisateurs



# Démonstration par l'exemple

## Zoosk

The screenshot displays the Zoosk website interface. At the top, the Zoosk logo is on the left, and navigation links for Home, Find Friends, and the user's name (Glenn Engstrand) are on the right. The main content area is titled "Share a moment:" and features a text input field with a checkmark icon and the placeholder text "Say something!". Below this is a post by Glenn Engstrand, who is located at the Golden Gate Bridge in San Francisco. The post includes a photograph of the bridge and the text "Like · Comment · 8 minutes ago". Below the post are two notifications: "Elsa Hou is now friends with Joy Dutta" and "Loris Zucchetti is now friends with Jonell Stocksberry". On the right side, there is a section titled "People you may know" which lists several users with their profile pictures and "Add Friend" buttons. At the bottom of this section is a "Find Your Friends" button. The left sidebar contains a navigation menu with options like News Feed, Me, Relationship Status, Friends, Personals, News, Search, Chat, Datecard, Views, Inbox, ZSMS, and Subscribe.



# Démonstration par l'exemple

## Zoosk

- Big Search avec Solr
  - Recherche de profil
  - Flux d'actualités
  - Trouver un partenaire



# Plan

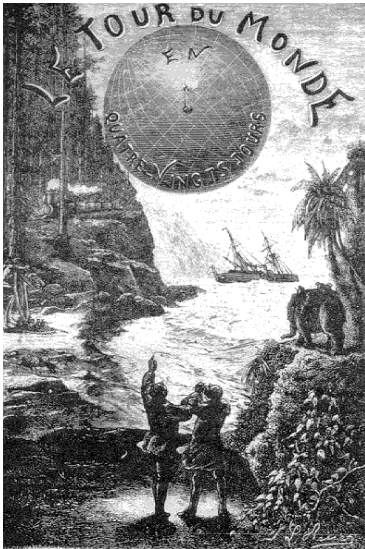
- ❑ Solr/Constellio
- ❑ Hadoop
- ❑ Démonstration par l'exemple
- ❑ **Démonstration Hadoop/Solr**



# Démonstration Hadoop/Solr

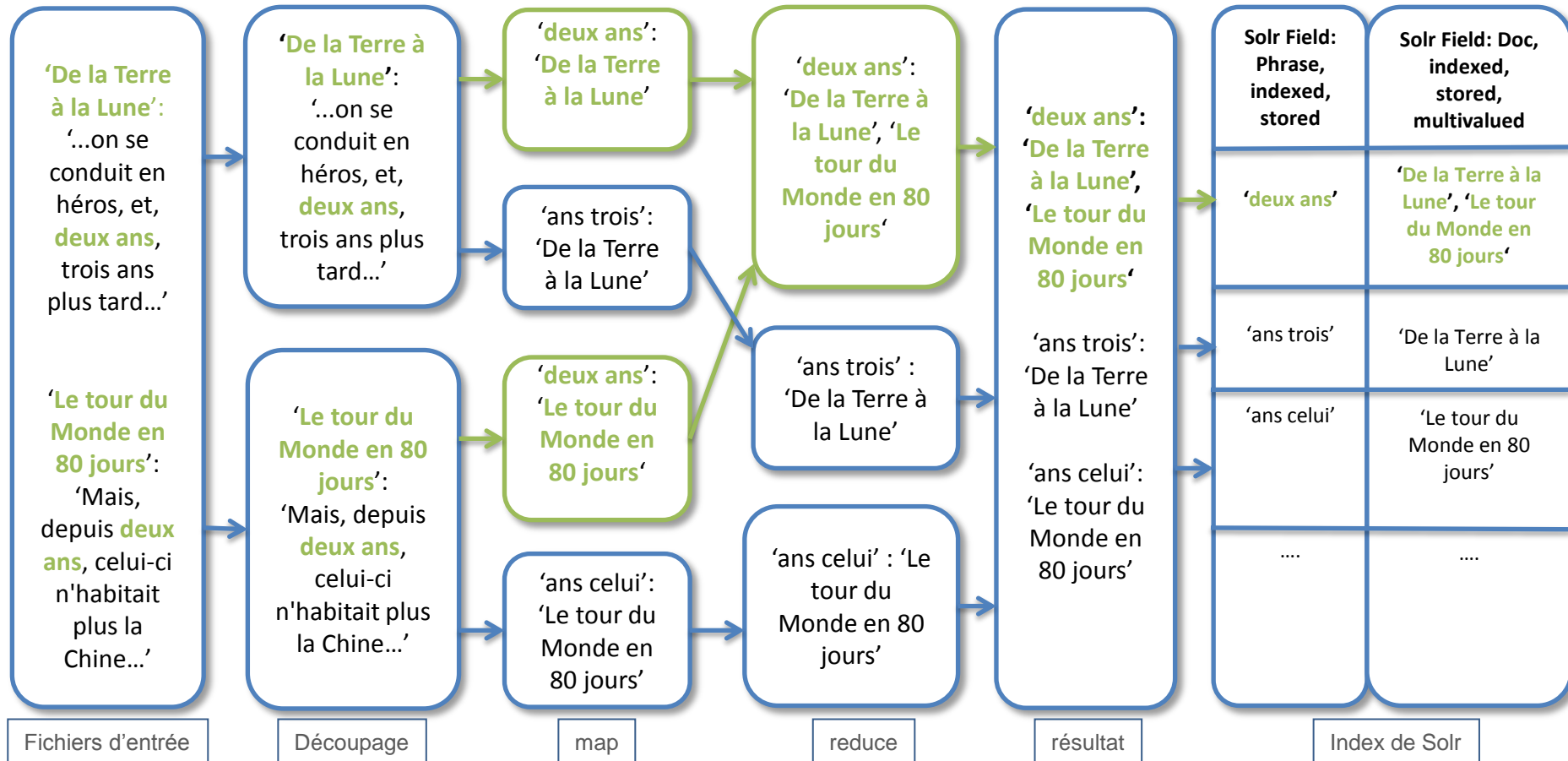
## Objectif

- Dans un texte, trouver toutes les phrases qui contiennent un mot cherché et dans quel document il se trouve (basé sur Lucid Imagination)



# Démonstration Hadoop/Solr

## Etapes



# Contacts

***Site web : [www.francelabs.com](http://www.francelabs.com)***

***Email : [contact@francelabs.com](mailto:contact@francelabs.com)***

***Twitter : [@Francelabs](https://twitter.com/Francelabs)***

