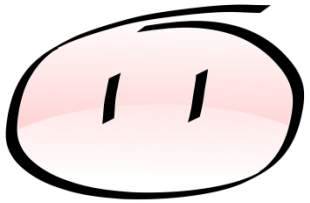


France Labs
Open Source Enterprise Search

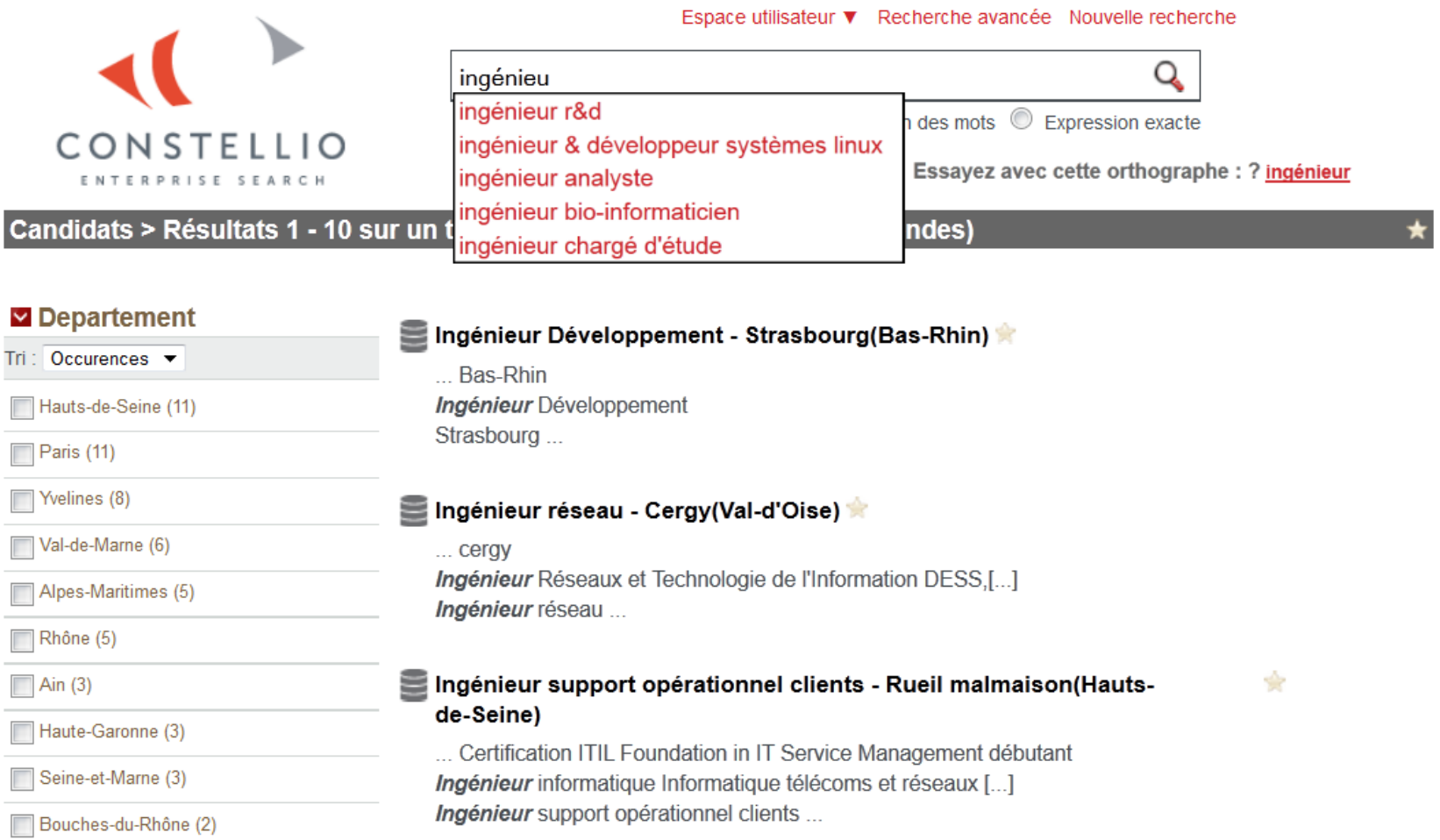
Big Search

Aurélien MAZOYER
Olivier TAVARD

Plan

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

A quoi ressemble un moteur de recherche aujourdhui?



Qu'y a-t-il dans Constellio?

Un oignon

Qu'y a-t-il dans Constellio?

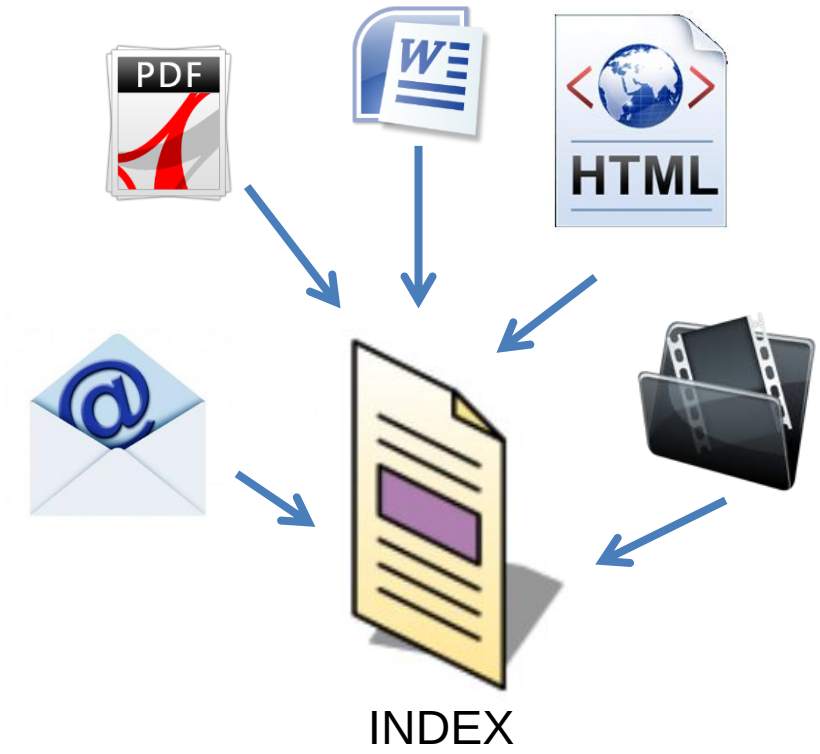
Lucene

- 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

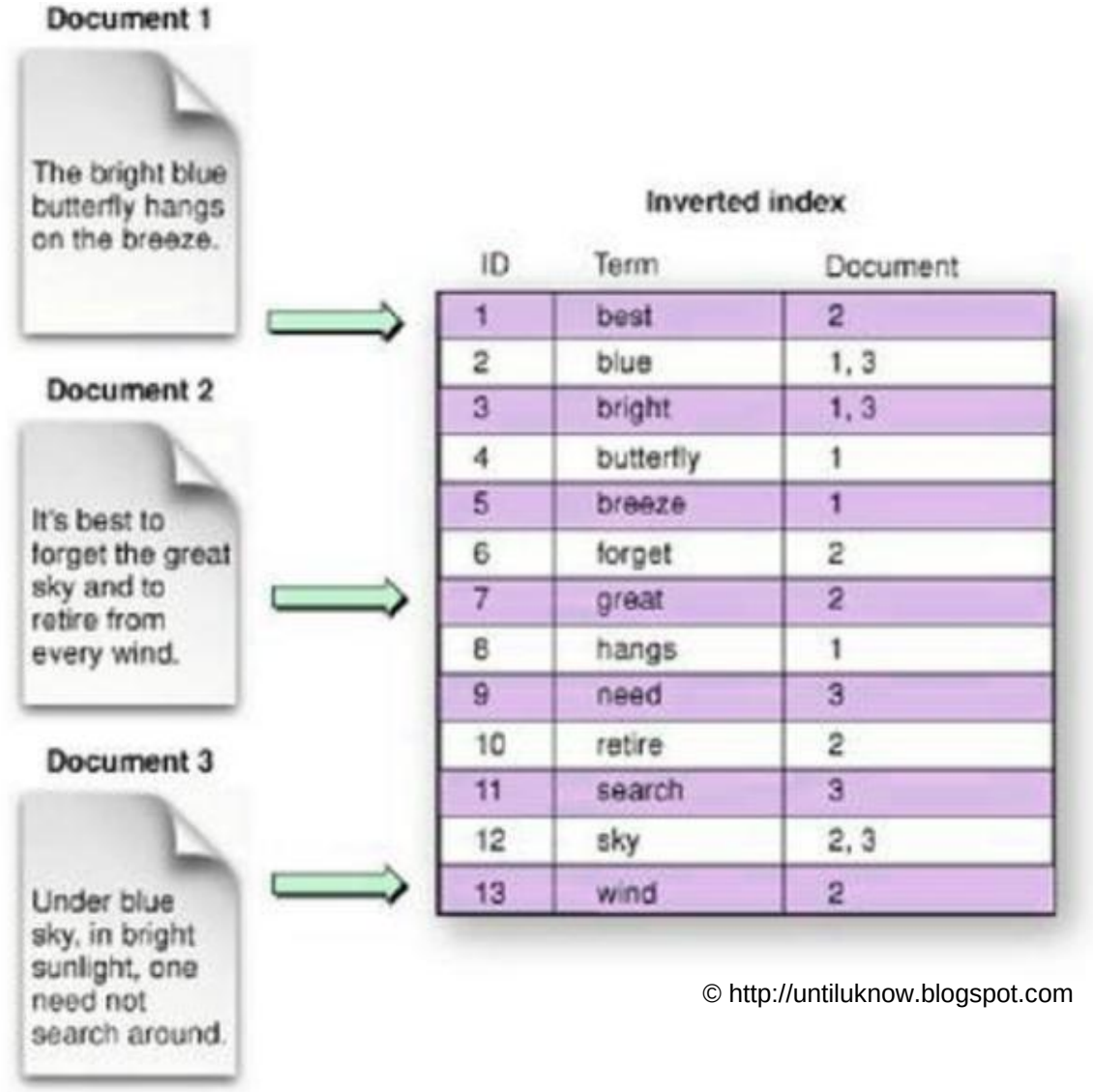
Indexation

- Un outil qui permet:
  - De créer un index à partir de documents



Lucene

Index inversé



Lucene

Recherche

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

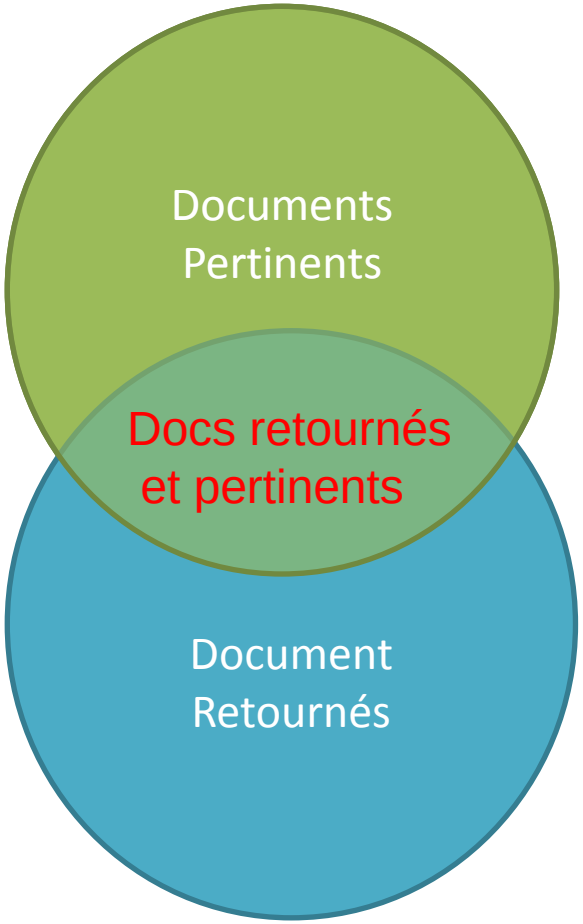
INDEX

Lucene

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



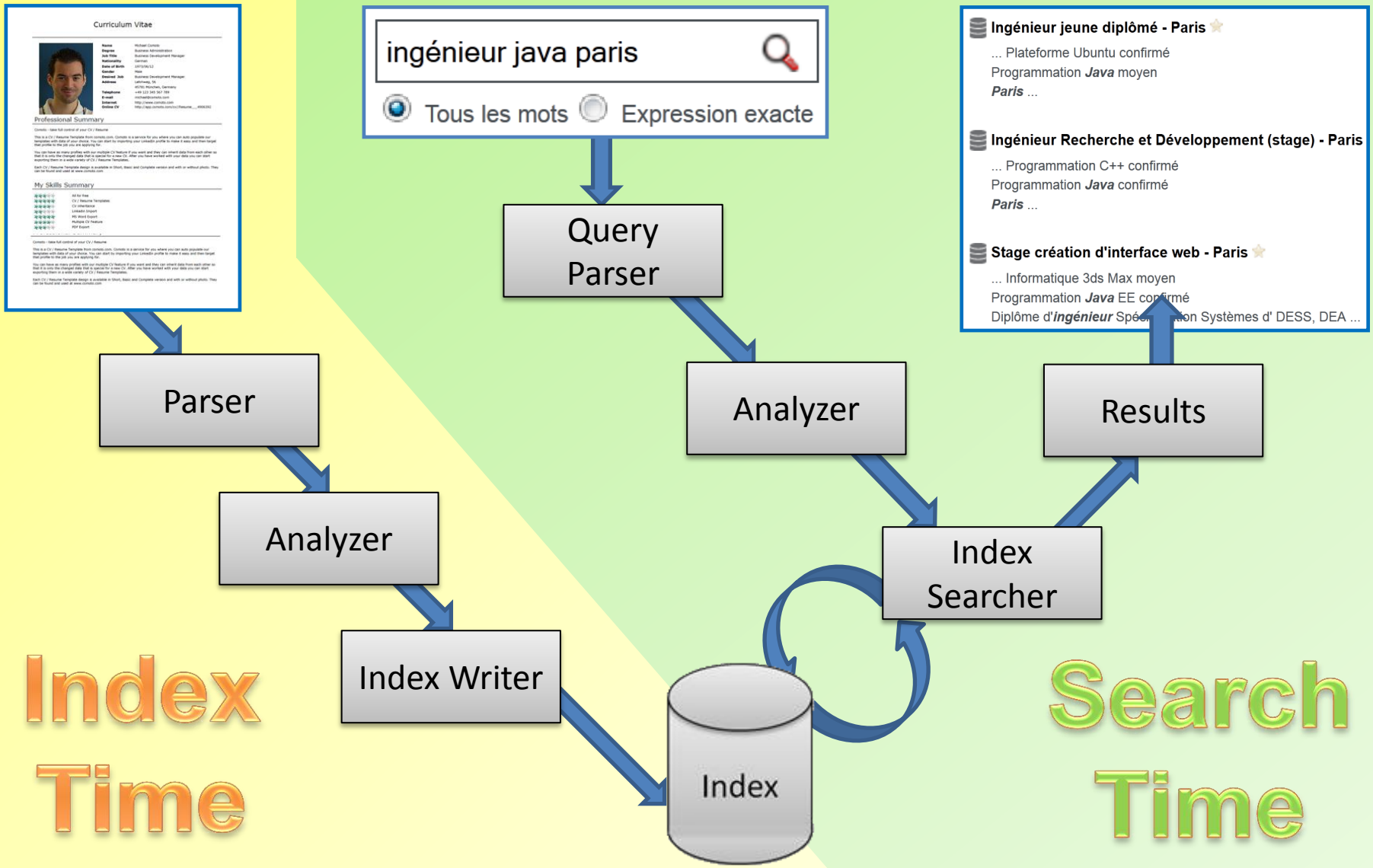
Trouver un bon compromis…

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

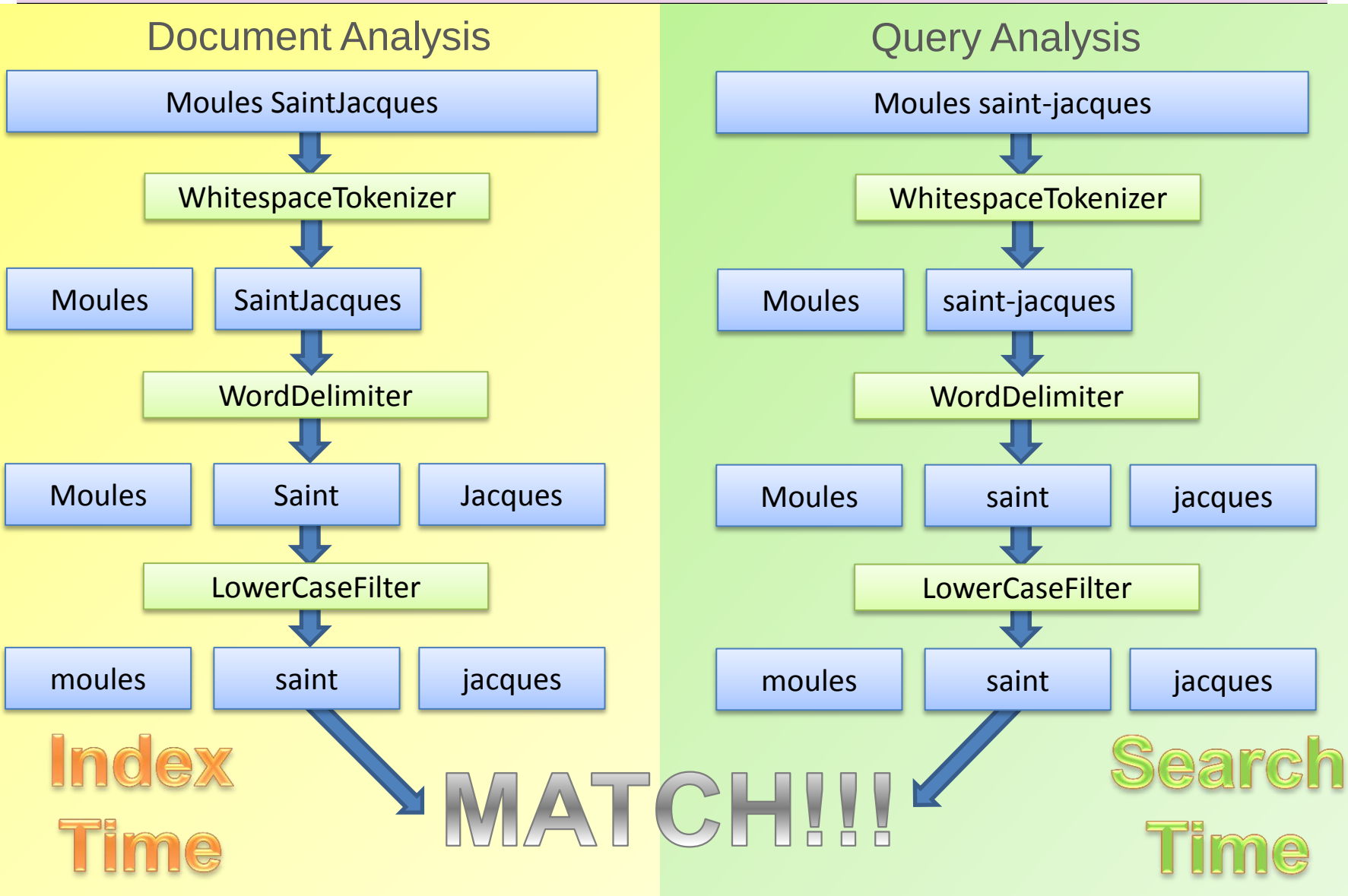
Lucene

Indexation et Recherche



Lucene

Analyzers



Lucene

Scoring

$$\Sigma_{t \text{ in } q} \left(tf(t \text{ in } d) \times idf(t)^2 \times boost(t.field \text{ in } d) \times lengthNorm(t.field \text{ in } d) \right) \times coord(q,d) \times queryNorm(q)$$

- Formule paramètrable
- Combinaison de
  - Boolean Model
  - Vector Space Model
    - Term Frequency
    - Inverse Document Frequency
    - …

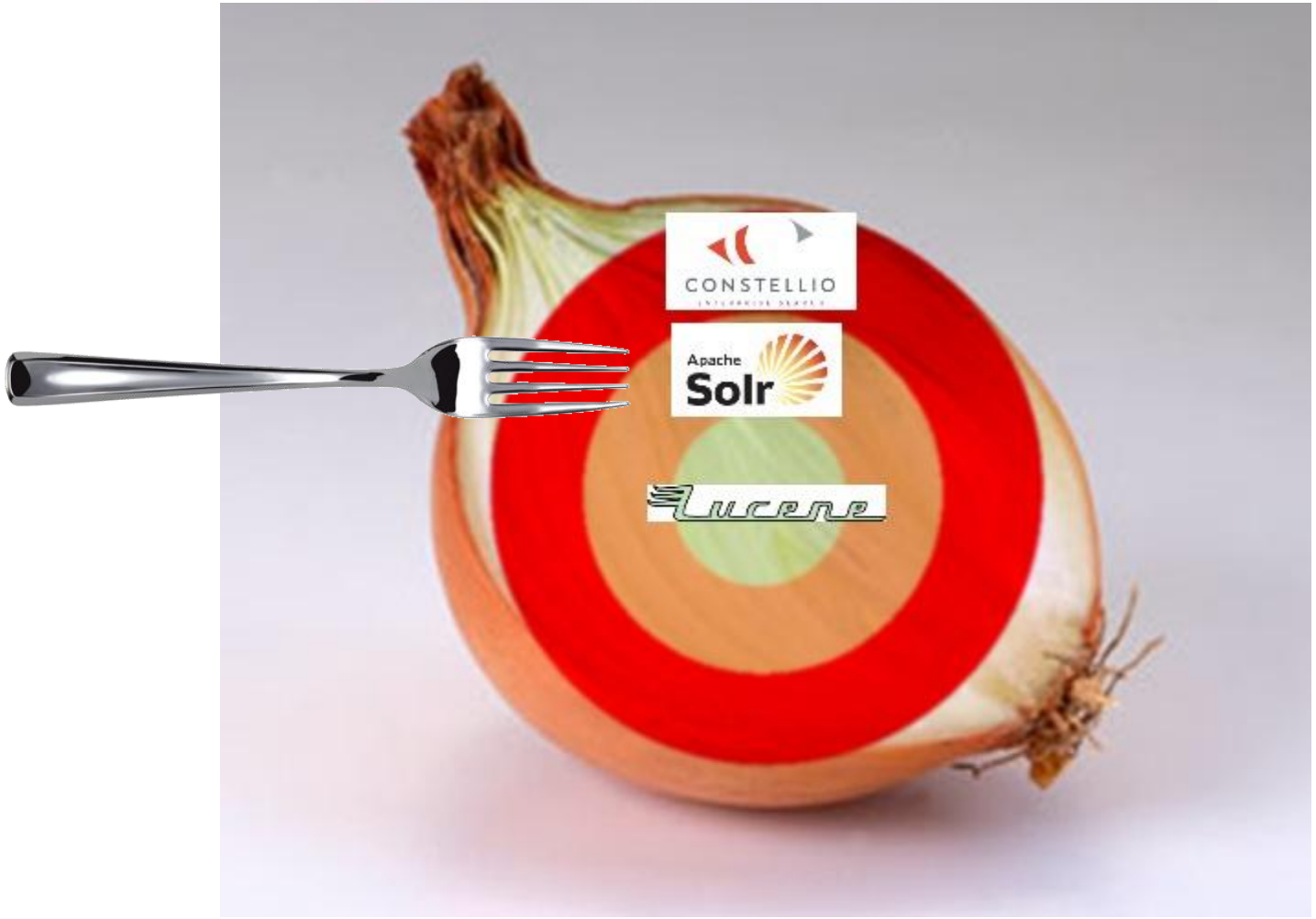
Lucene

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 : Sorl 3.6 (Avril 2012)

Apache Solr™

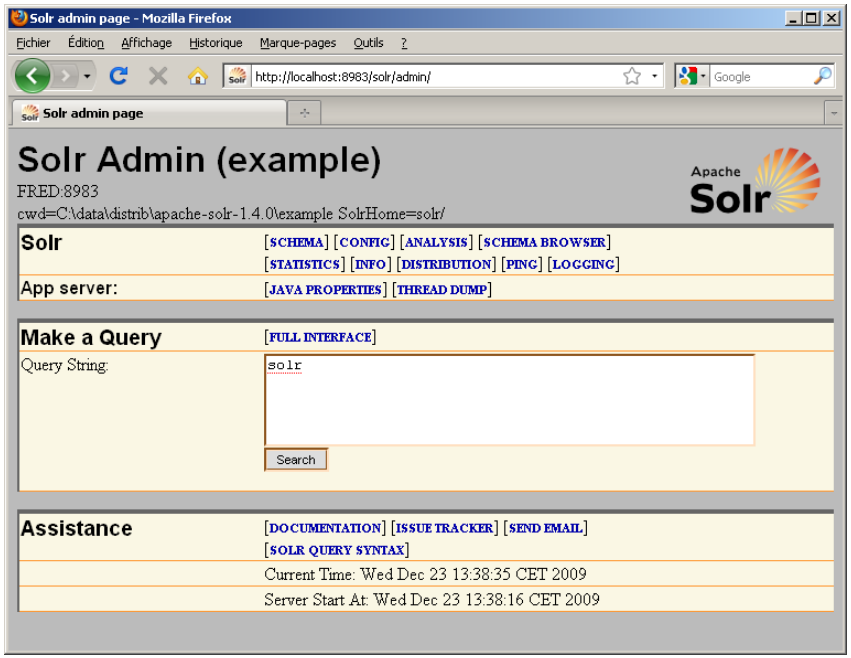
Solr

- 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

Fonctionnalités supplémentaires

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

Constellio

Connectors

- 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

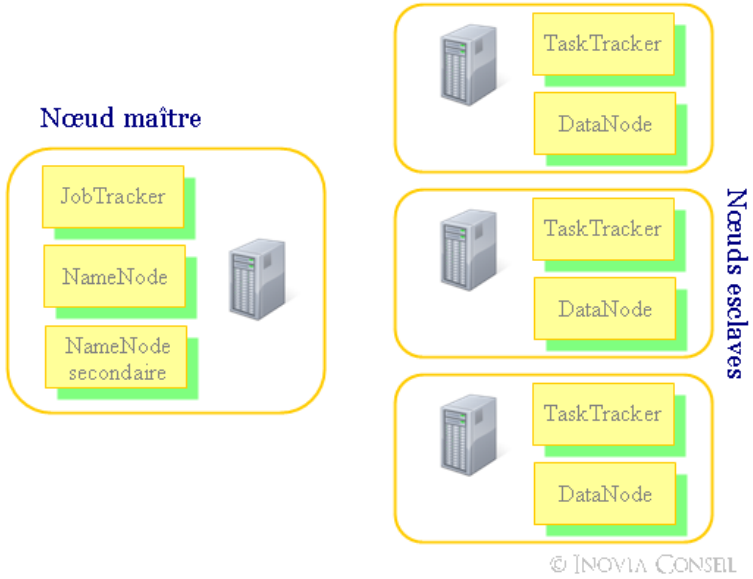
Vue d'ensemble

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

Hadoop

HDFS

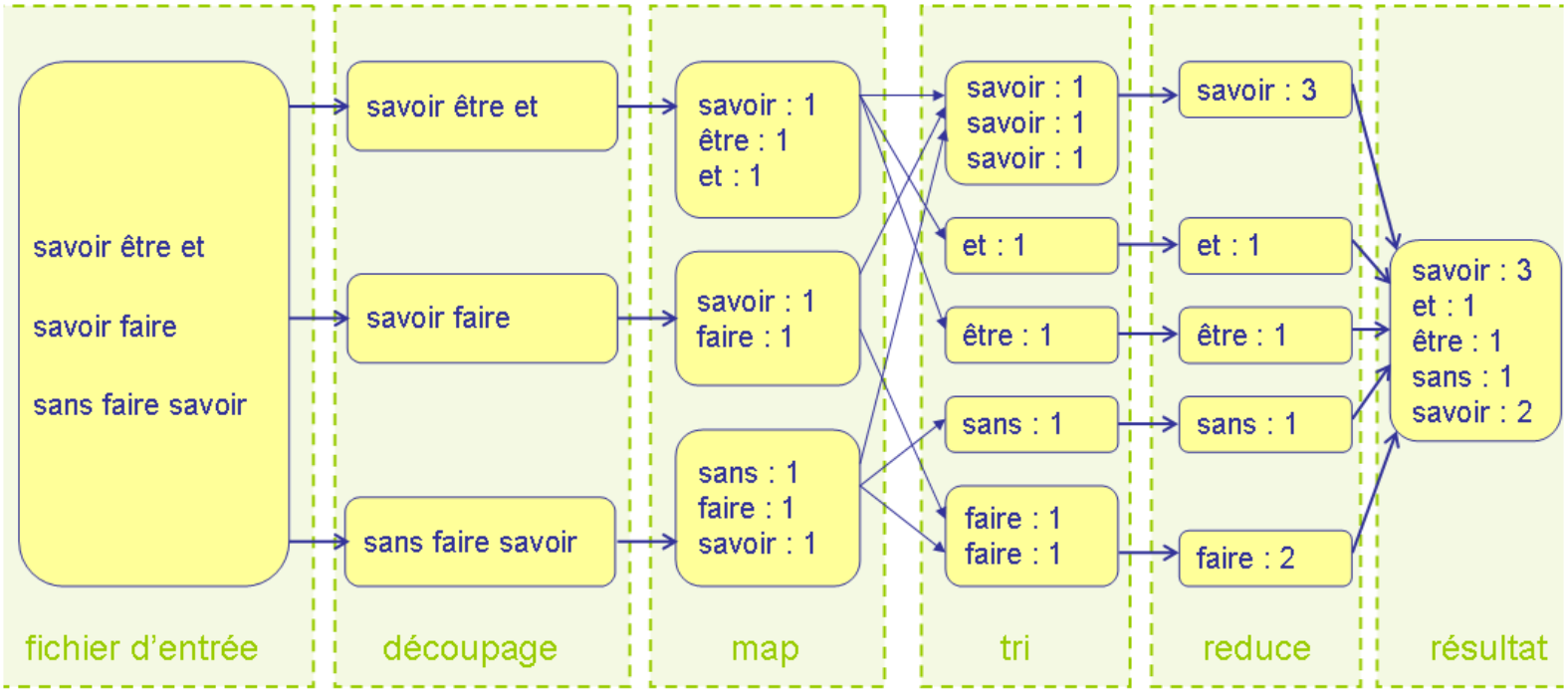
- Données converties en blocs et distribuées sur des nœuds
- Chaque bloc est répliqué



Hadoop

Map/Reduce

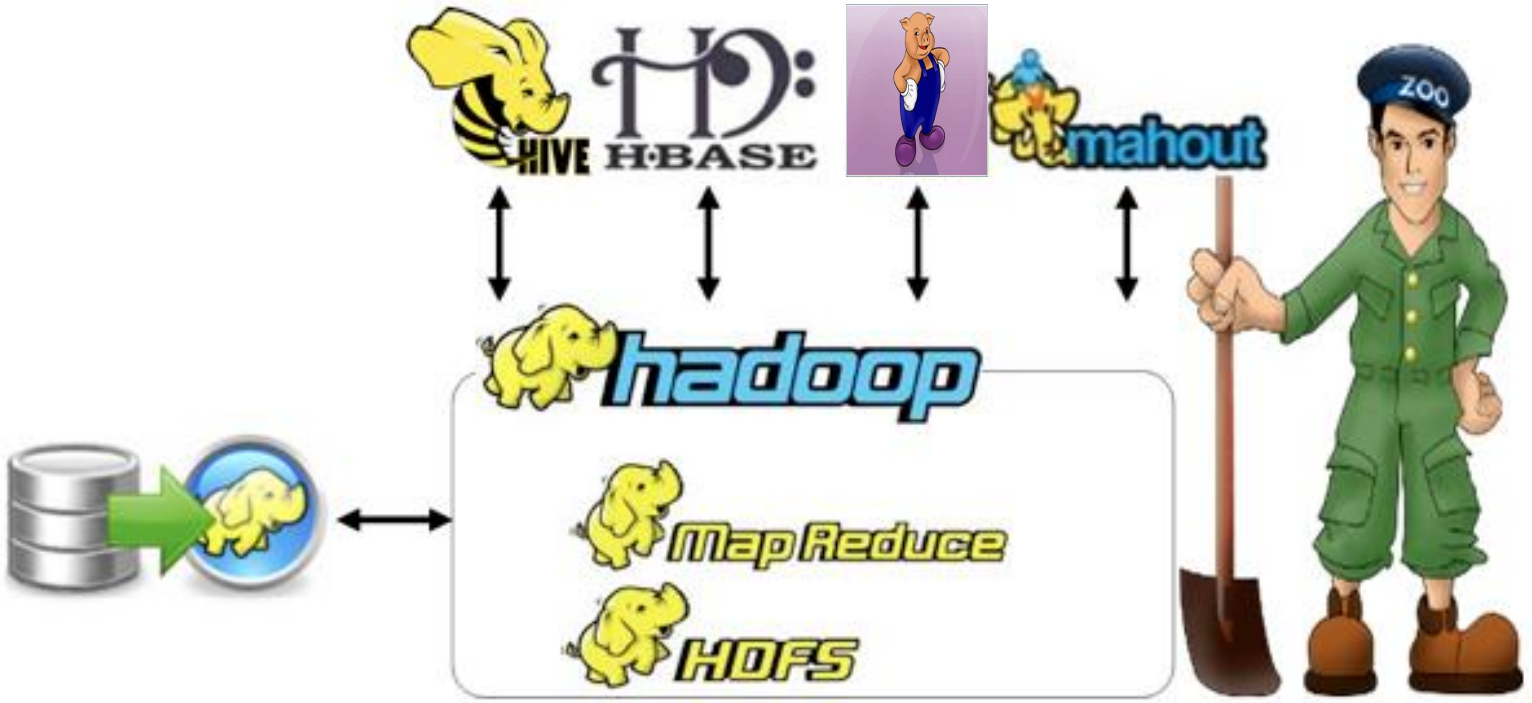
- Map : données sous forme clés/valeurs
- Reduce : fusion par clé pour former résultat



http://blog.inovia-conseil.fr/?p=46

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 BIIIIG
  - Hadoop
  - Hadoop / Solr
  - MapReduce / Search
  - Solr

Démonstration par l'exemple

Google

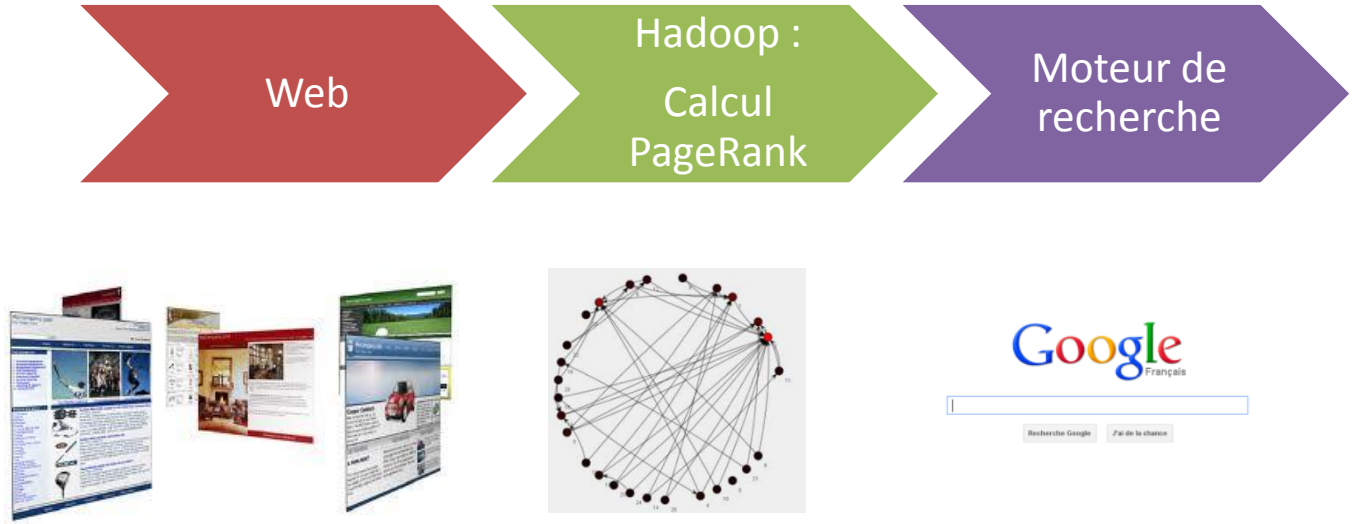
- 1 000 000 000 000 d'URLS uniques (2008)
- Pagerank : le ranking d'une page est estimé par sa popularité plutôt que par son contenu

Google

Démonstration par l'exemple

Google

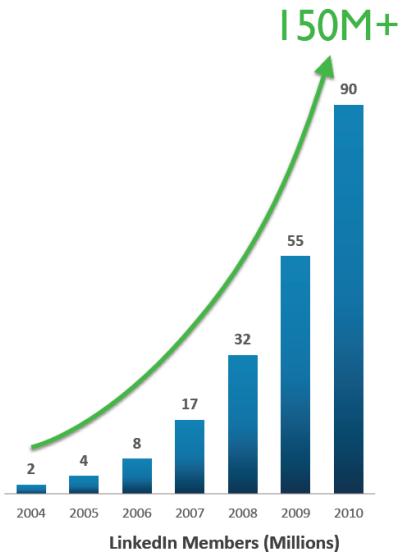
- Construire PageRank grâce à Map/reduce



Google

Démonstration par l'exemple

Linkedin



75%
Fortune 100 Companies use LinkedIn to hire

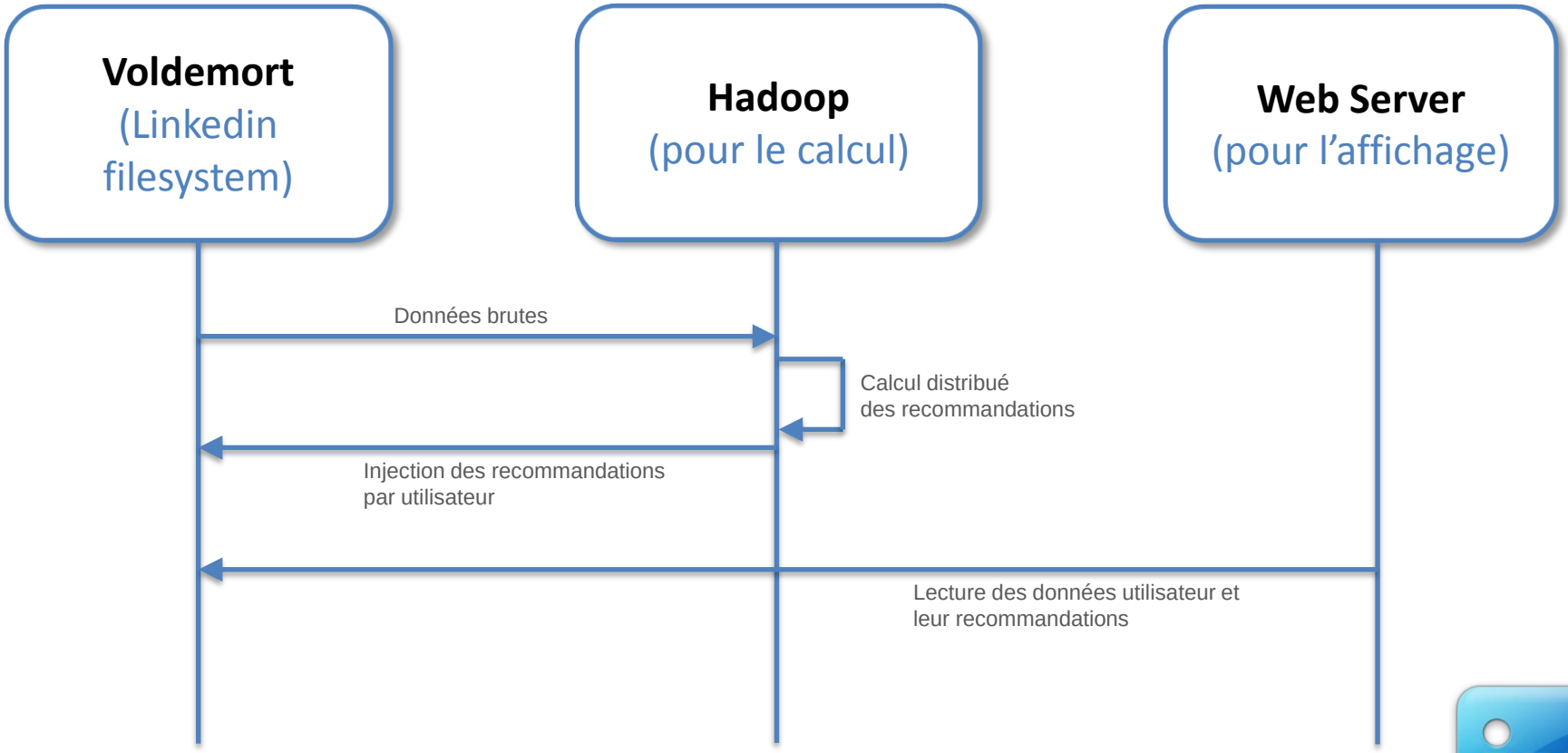
>2M
Company Pages

~4B
Searches in 2011

People You May Know

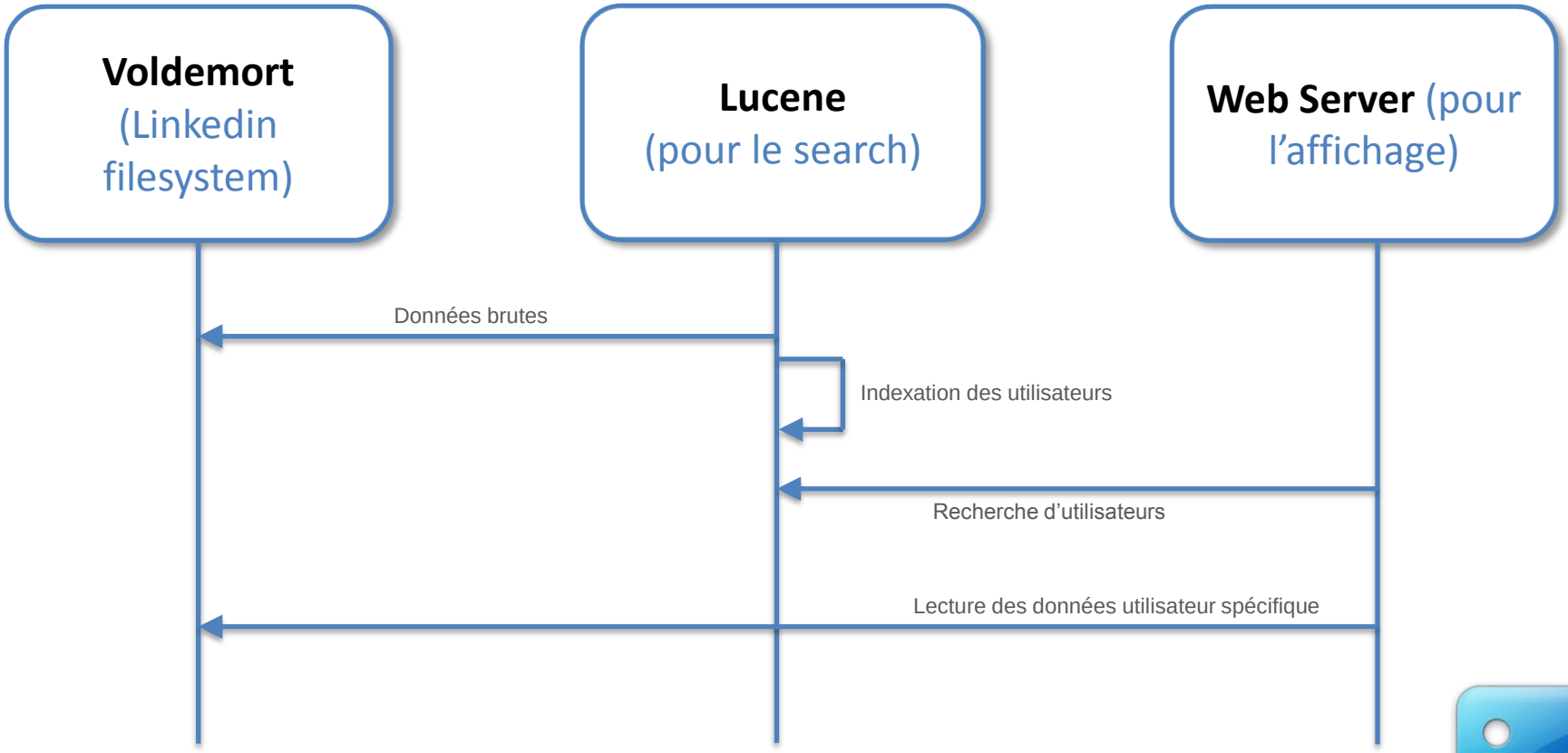
Démonstration par l'exemple

1er cas : Hadoop pur pour les recommandations



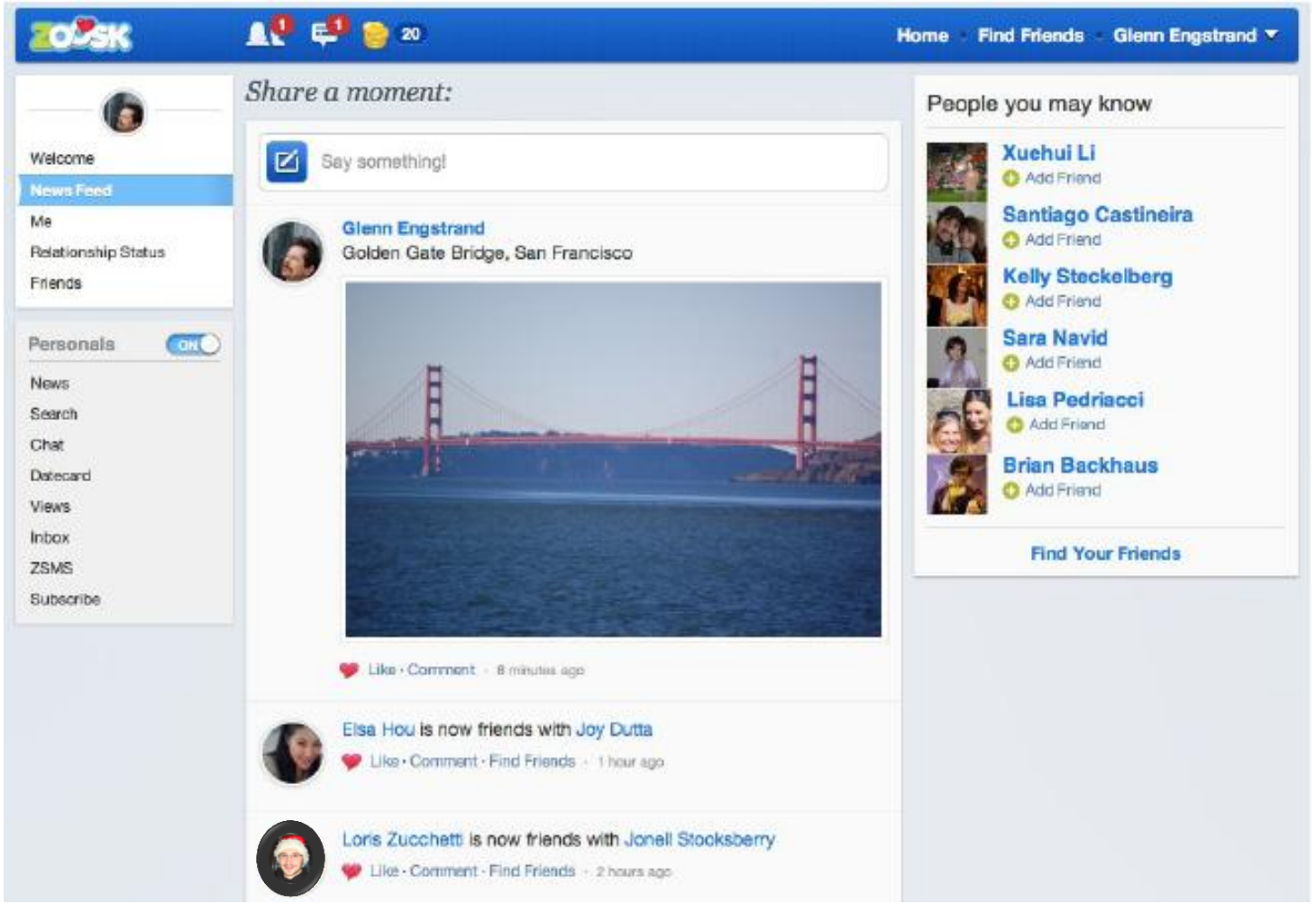
Démonstration par l'exemple

2e cas : Lucene pur pour la recherche d'utilisateurs



Démonstration par l'exemple

Zoosk



Démonstration par l'exemple

Zoosk

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

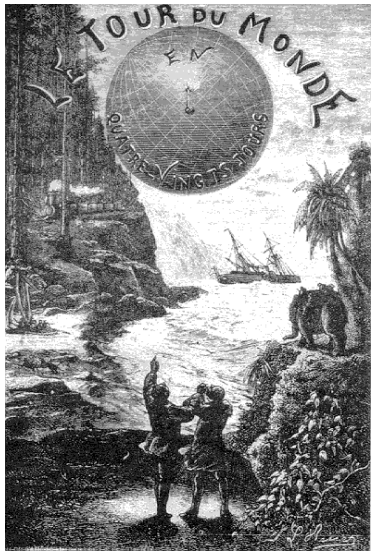
Plan

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

Démonstration Hadoop/Solr

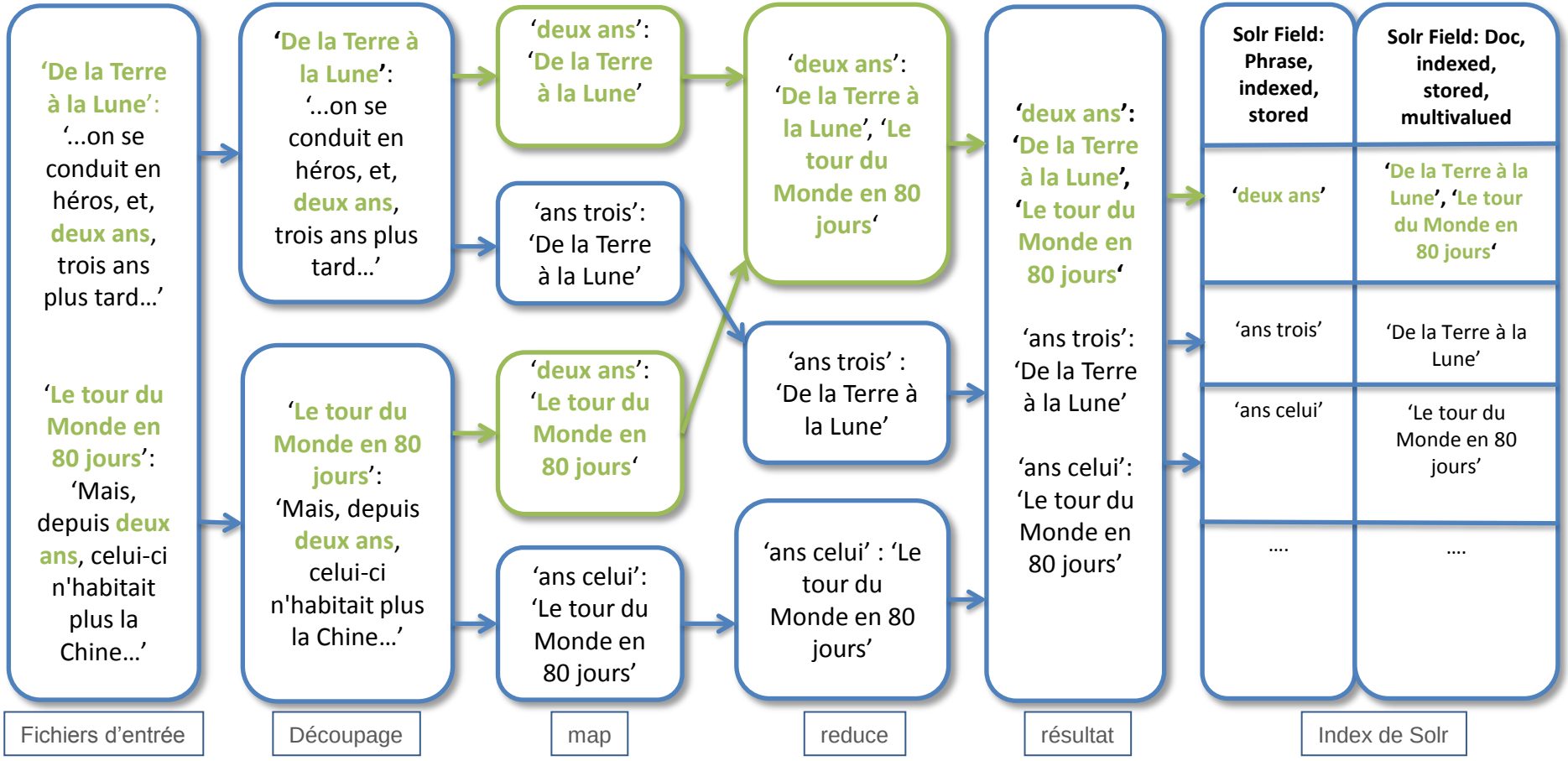
Objectif

- 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



hadoop

Apache Solr

France Labs
Open Source Enterprise Search

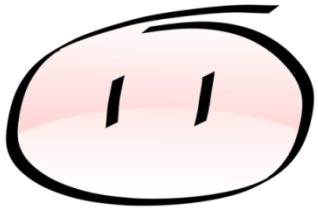
Contacts

Site web : www.francelabs.com

Email : contact@francelabs.com

Twitter : @Francelabs

Incubateur Paca-Est
Innover, Entreprendre



TELECOM ParisTech

France Labs
Open Source Enterprise Search