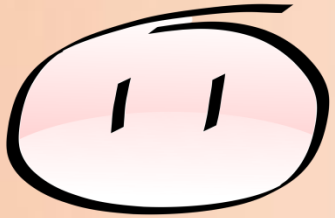


France Labs
Open Source Enterprise Search

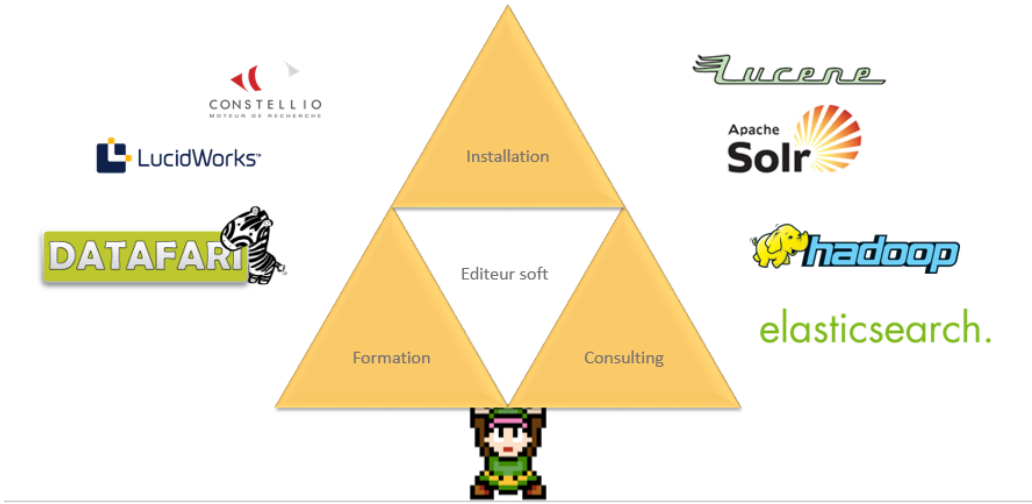
Fouillez facilement dans votre système Big Data

Olivier TAVARD

Introduction

- A propos de moi :
  - Cofondateur de la société France Labs
  - Développeur (principalement Java)
  - Formateur en technologies de moteurs de recherche (Solr, Elasticsearch)

- A propos de France Labs :
  - Startup créée en 2011
  - Experts en technos de search (consulting)
  - Partenaire officiel de LucidWorks
  - Editeur de la solution Datafari



Plan

- **Rappels sur Hadoop**
- Présentation (rapide) de Solr
- Intégration d'Hadoop et de Solr
- Démo

Présentation d'Hadoop

Hadoop

- Créé par Doug Cutting en 2004
- A l'origine pour Lucene : Projet Nutch
- Framework open source
- Inspiré par les papiers sur Google Map Reduce et Google File System

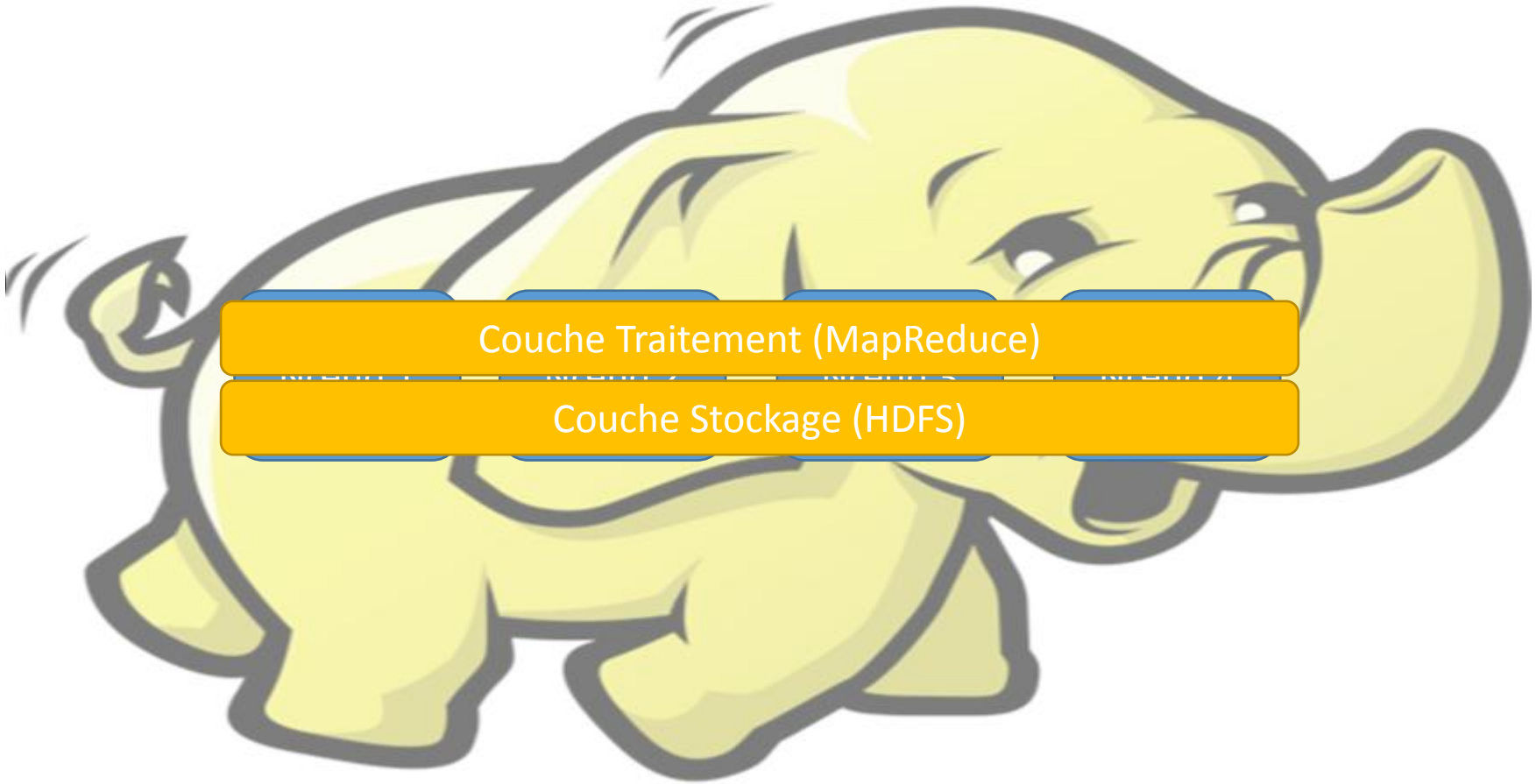
Présentation d'Hadoop

Hadoop

- Traiter des grands volumes de données en un minimum de temps
- Stocker des immenses volumes de données
- Fonctionne sur machines de configuration faible et peu coûteuses

Présentation d'Hadoop

- Architecture (très) simplifiée de Hadoop v1



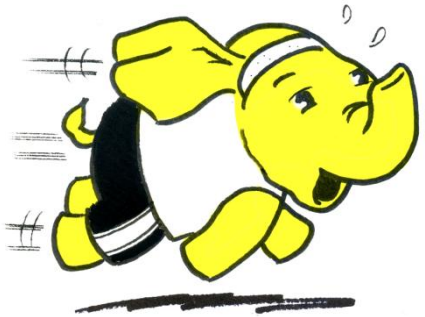
Présentation d'Hadoop

- Couche stockage (HDFS)
  - Système de fichier distribué
  - Utilise les disques « normaux » de chacun des nœuds…

    …pour donner l'impression de n'utiliser qu'un seul énorme disque
  - Fiable (données répliquées 3 fois)

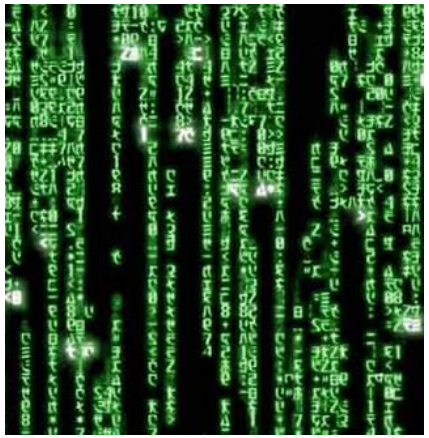
Présentation d'Hadoop

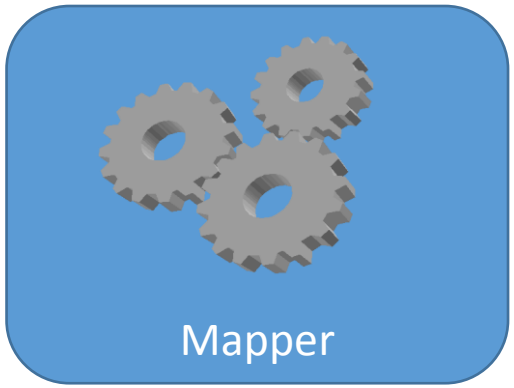
- Couche traitement (MapReduce)
  - Hadoop exécute des jobs Map Reduce
  - Un algorithme doit donc implémenter:
    - Un mapper
    - Un reducer
  - Pour être executé de manière distribuée par Hadoop en tant que job



Présentation d'Hadoop

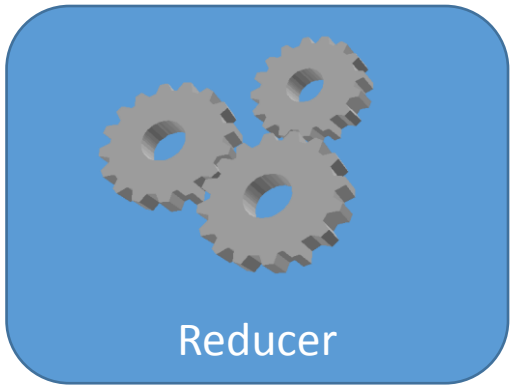
- Mapper
  - Prend des données en entrée
  - Et les transforme en paire de clé valeur





- Reducer
  - Combine les valeurs ayant la même clé

Clé 1 : Val2
Clé 2 : Val2
Clé 1 : Val3
Clé 2 : Val4



Présentation d'Hadoop

Hadoop

- Un exemple simple?
  - On a en entrée plusieurs textes
  - On veut compter le nombre d'occurrences des mots
  - De manière distribuée
  - En utilisant un job Map Reduce

Solr

Apache Solr est une couche web open source basée sur la librairie de recherche Apache Lucene. Elle ajoute des fonctionnalités de serveur web, du faceting.

Datafari

Datafari est notre solution open source de recherche d'entreprises, basée sur Lucene, Solr, et le framework de connecteurs Apache ManifoldCF.

Lucene

Apache Lucene est un moteur de recherche et d'indexation, en open source.

Présentation d'Hadoop

Hadoop

Solr

Apache Solr est une couche web open source basée sur la librairie de recherche Apache Lucene. Elle ajoute des fonctionnalités de serveur web, du faceting.

Datafari

Datafari est notre solution open source de recherche d'entreprises, basée sur Lucene, Solr, et le framework de connecteurs Apache ManifoldCF.

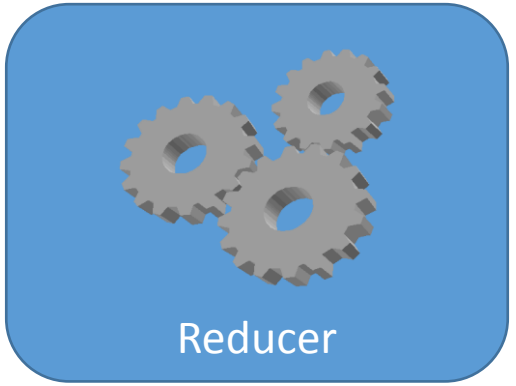
Lucene

Apache Lucene est un moteur de recherche et d'indexation, en open source.

Apache : 1
Apache : 1
Solr : 1
Lucene : 1
…

Apache : 1
Solr : 1
Lucene : 1
…

Apache : 1
Lucene : 1
…

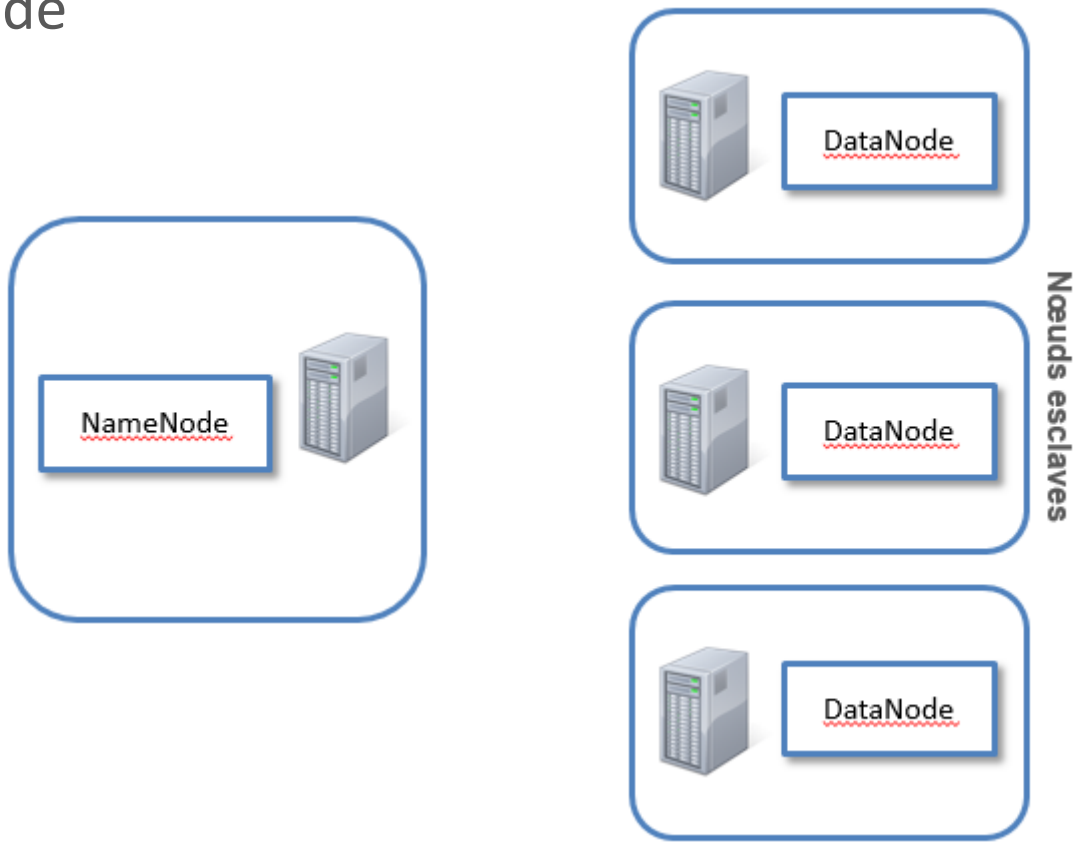


Apache : 4
Solr : 2
Lucene : 3
…

Présentation d'Hadoop

Hadoop 2

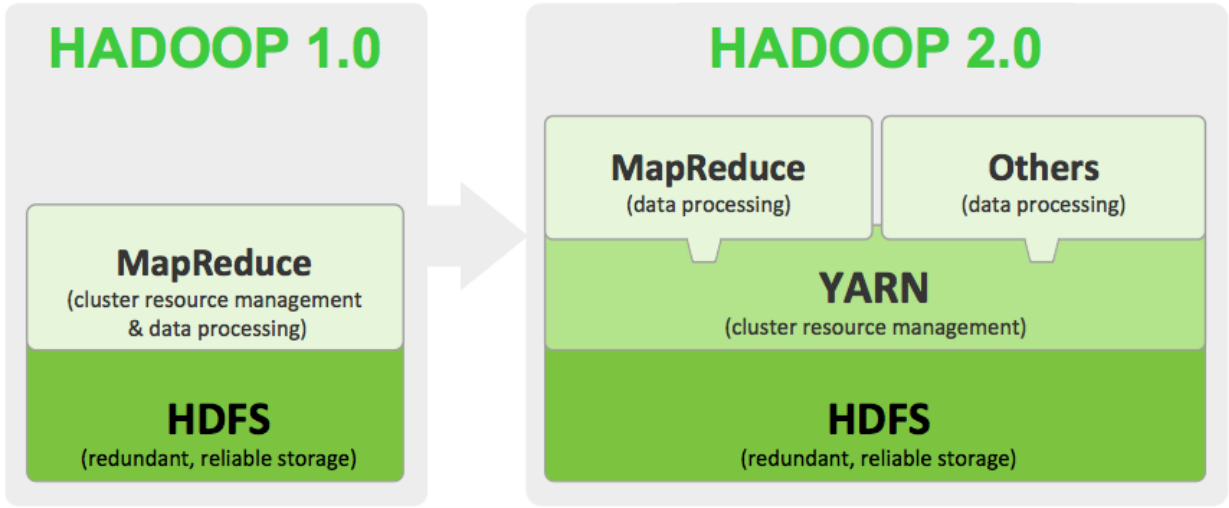
- HDFS 2 :
  - Haute disponibilité NameNode
  - Support de NFS



Présentation d'Hadoop

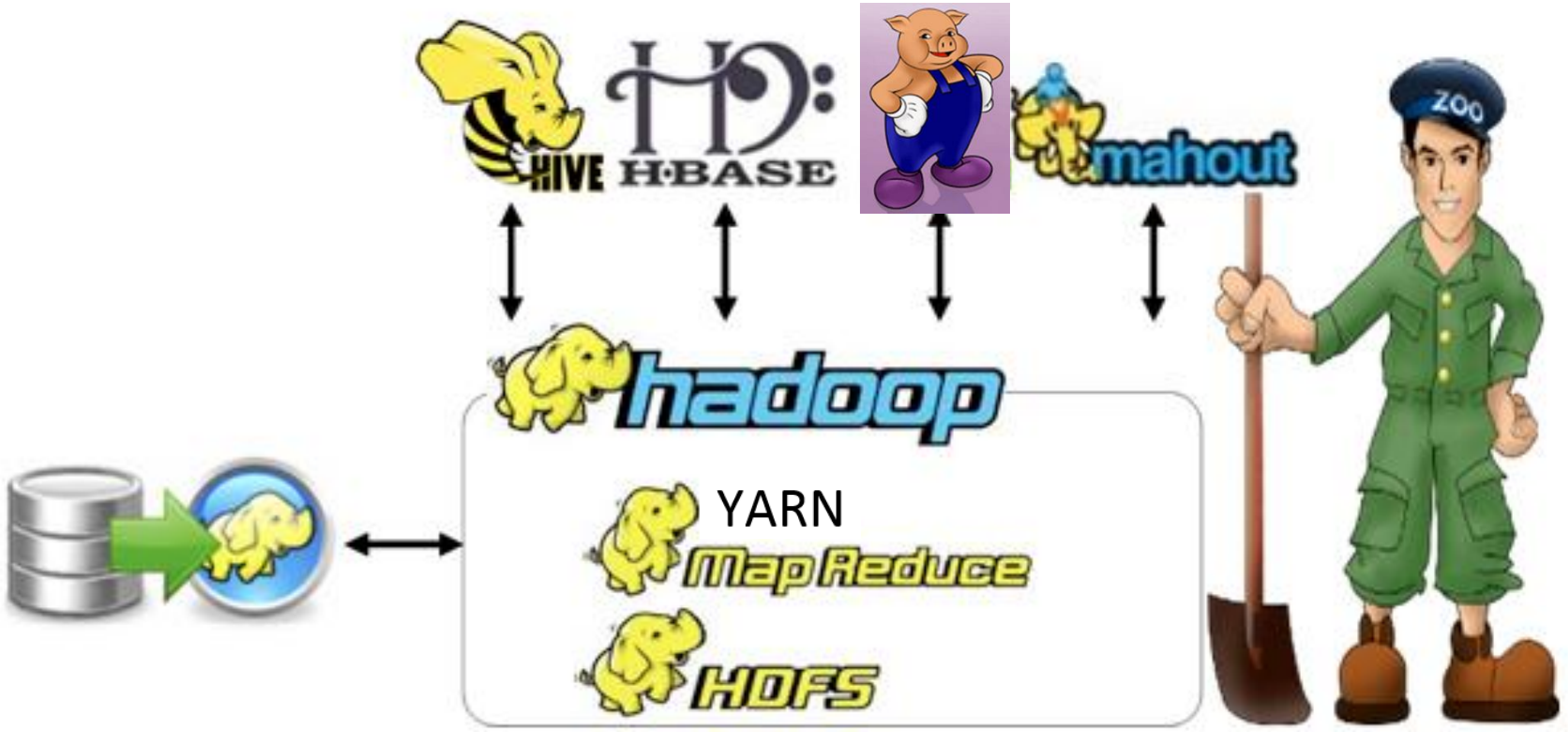
Hadoop 2

- YARN
  - Gestion jobs et ressources



Présentation d'Hadoop

Ecosytème



Plan

- Rappels sur Hadoop
- **Présentation (rapide) de Solr**
- Intégration d'Hadoop et de Solr
- Démo

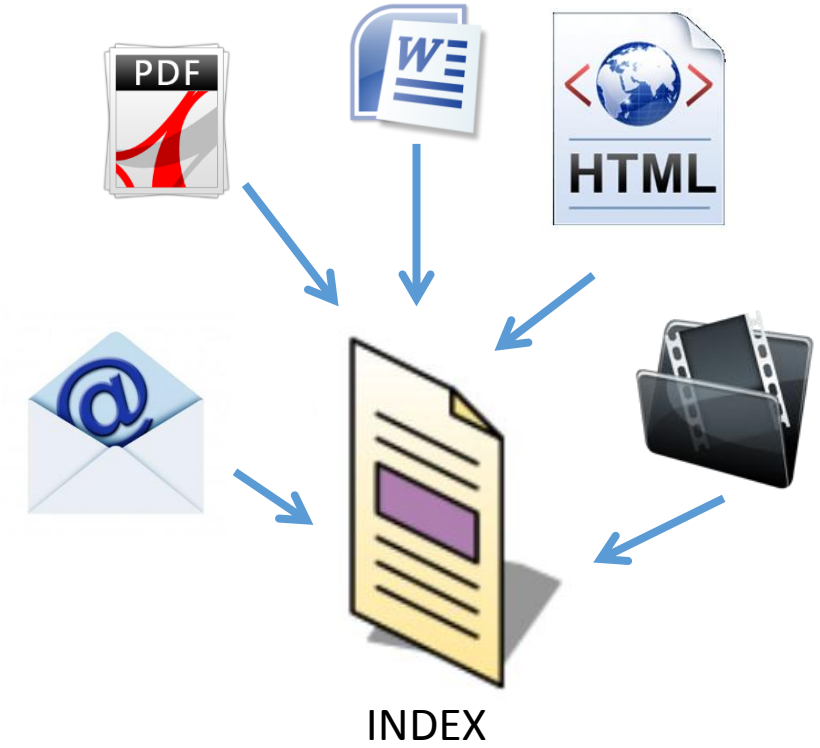
Présentation de Solr

La recherche est un oignon matriciel!

France Labs
Open Source Enterprise Search

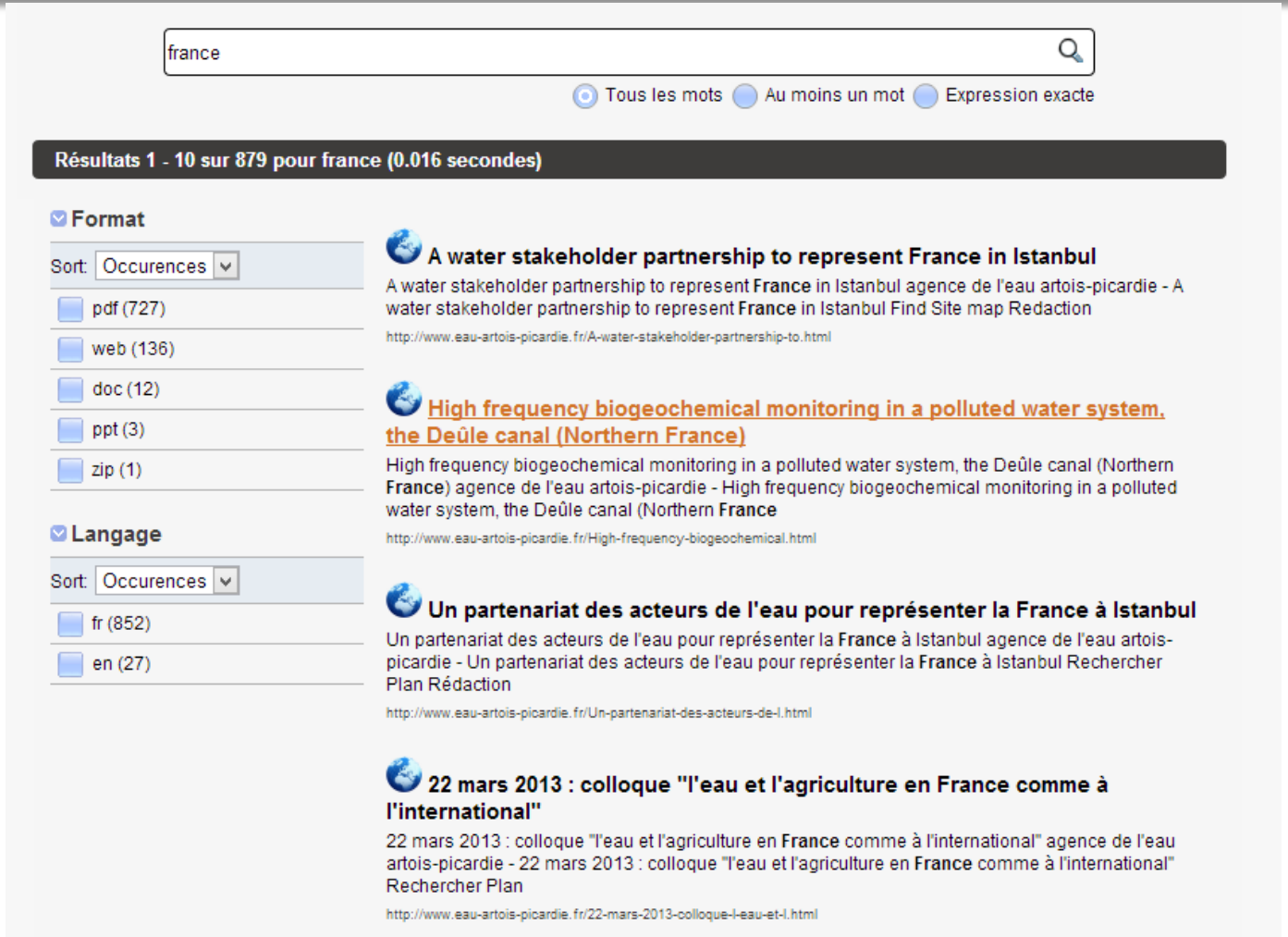
Présentation de Solr

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



Présentation de Solr

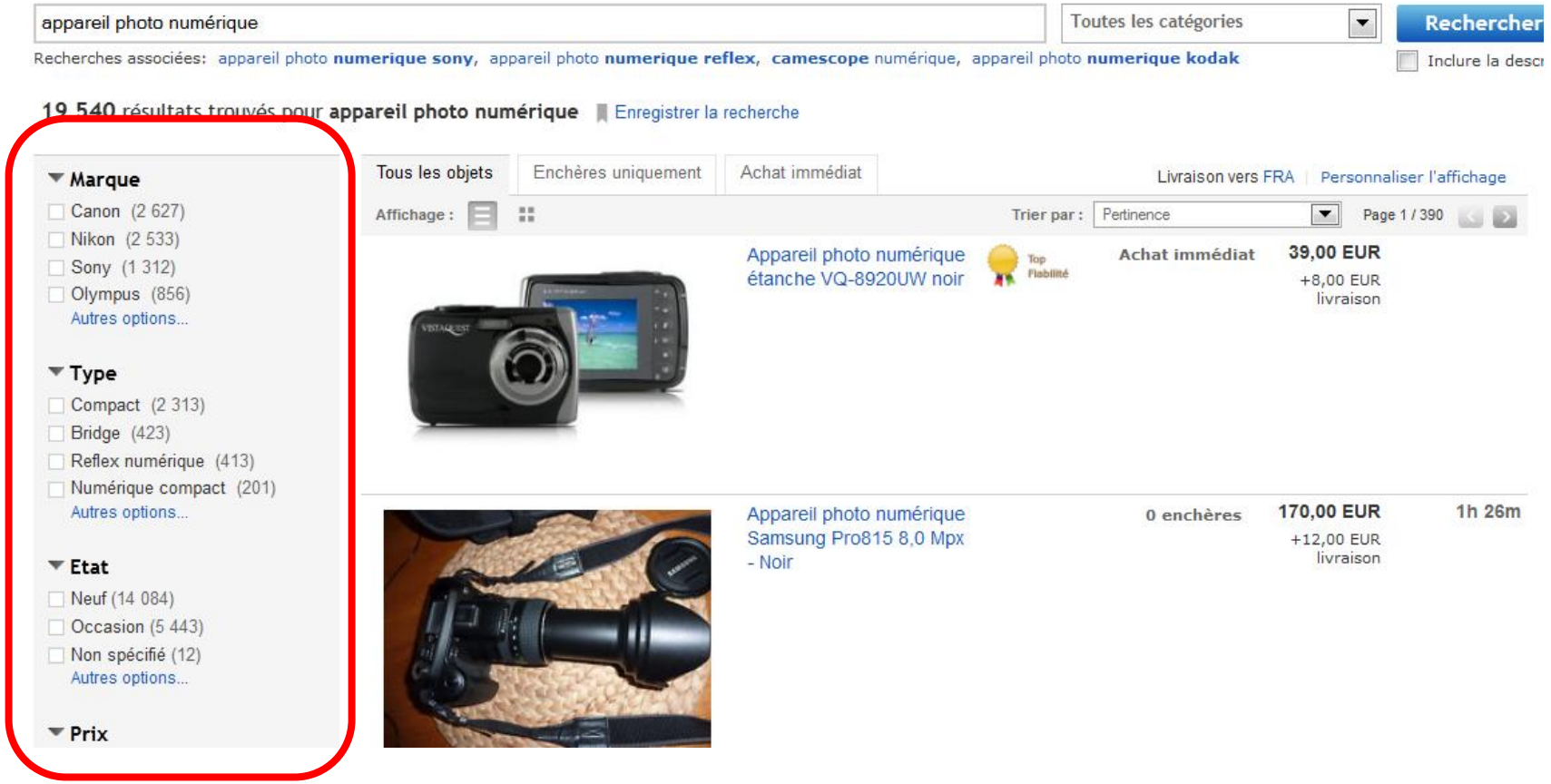
A quoi ça ressemble ?



France Labs
Open Source Enterprise Search

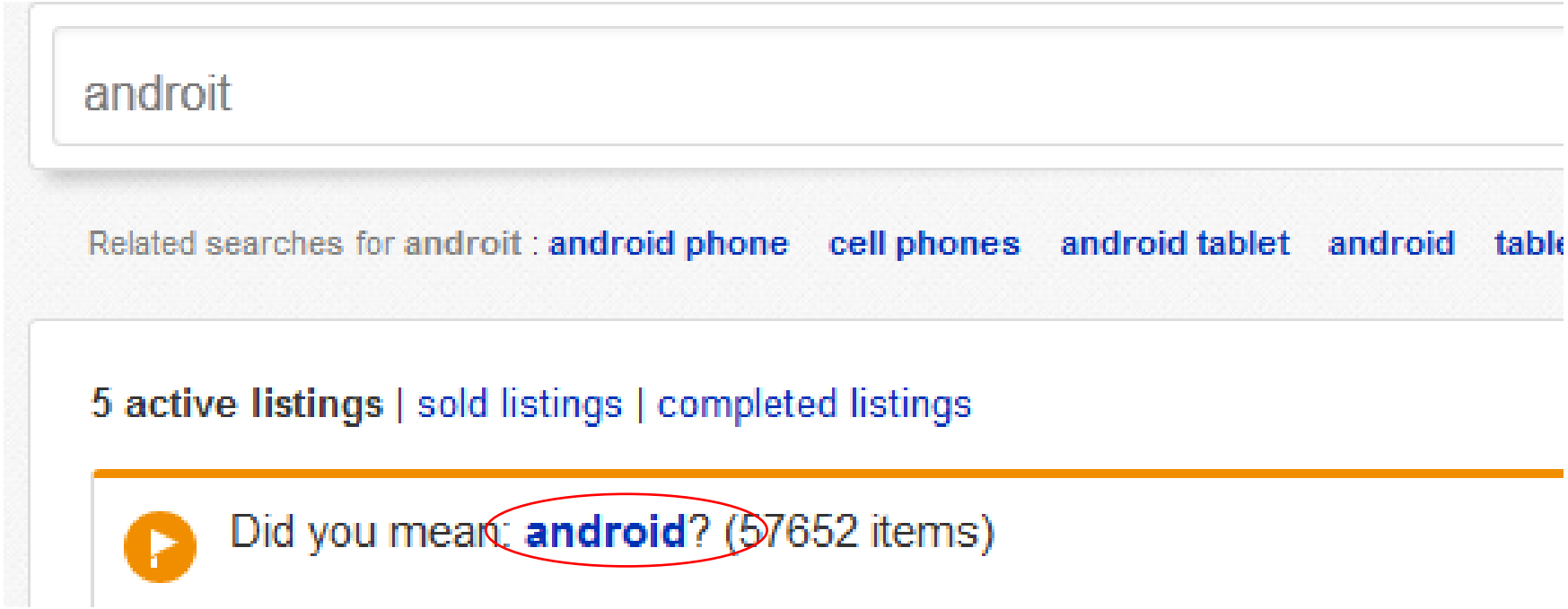
Présentation de Solr

Facettes



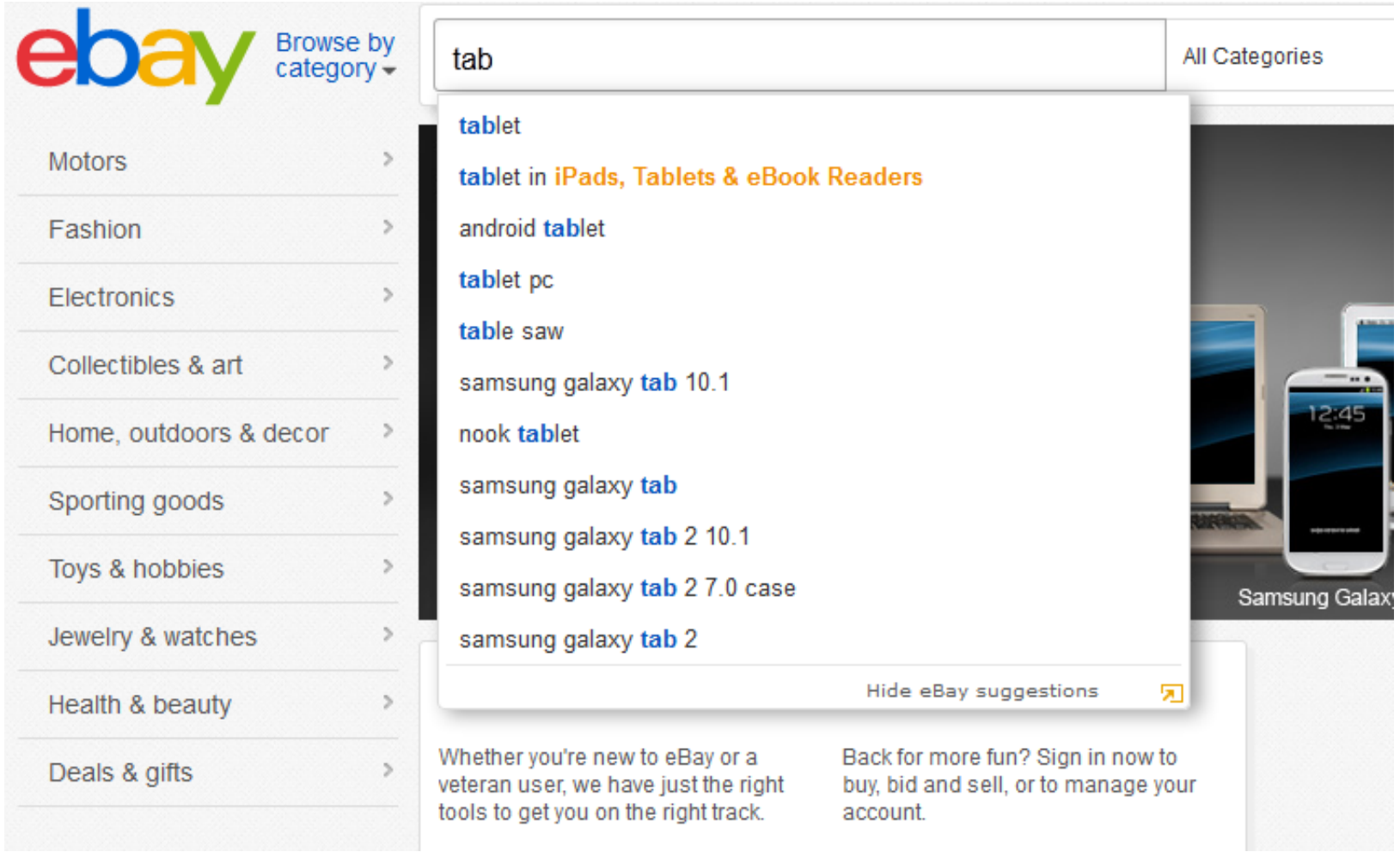
Présentation de Solr

Spellcheck



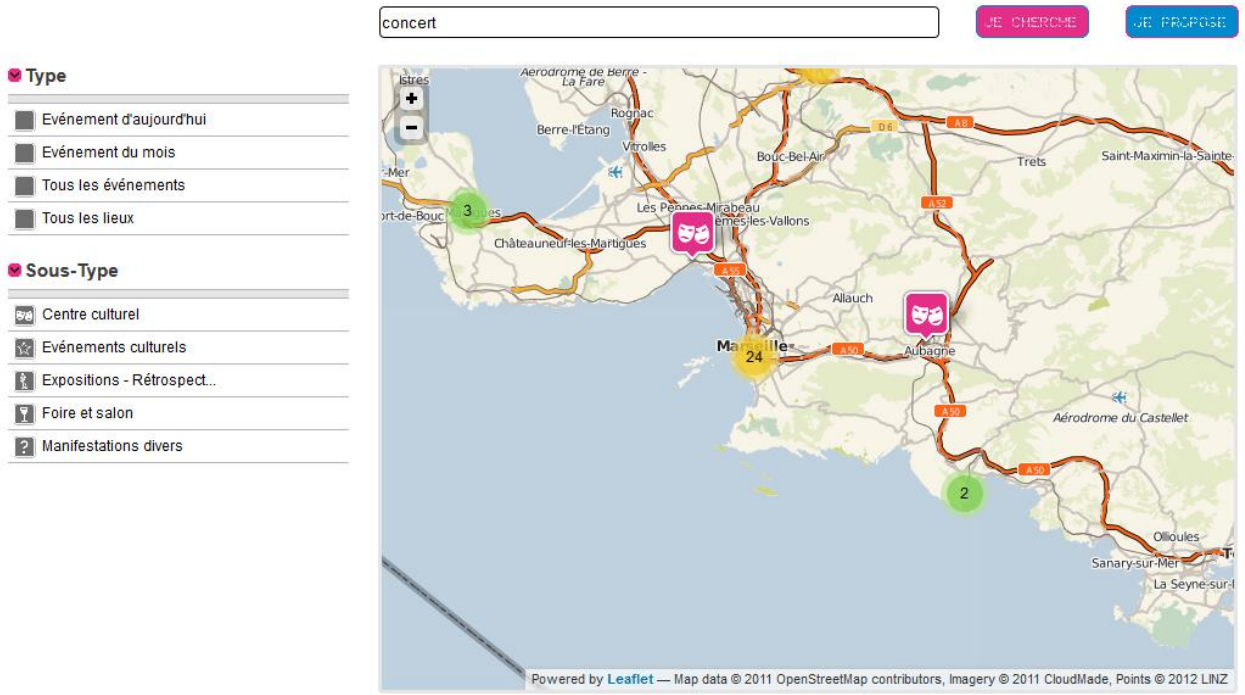
Présentation de Solr

Autocomplete



Présentation de Solr

Géolocalisation



France Labs
Open Source Enterprise Search

Autres fonctionnalités

- Highlighting

15 janv. 2012 – **Solr** provides a collection of **highlighting** utilities which can be reused by various Request Handlers to include "**highlighted**" matches and **...**

- More Like This
  - Obtenir des documents similaires à un document
  - Similarité textuelle

Présentation de Solr

Business Intelligence (BI)/Marketing

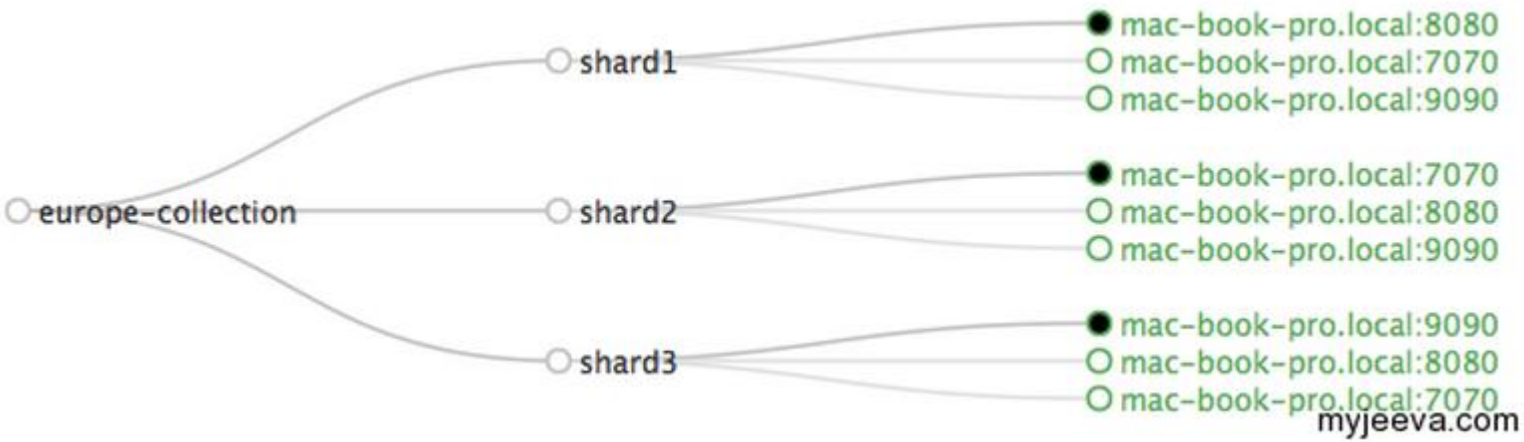
Présentation de Solr

Analyse de logs (for devops)

Présentation de Solr

Solr Cloud

- Solr Cloud permet la mise à l'échelle:
  - Architecture flexible en cluster de serveurs
  - Volumétrie élevée
  - Montée en charge
  - Tolérance aux pannes



Démo Solr

Tester Solr

Démo !

Apache Solr

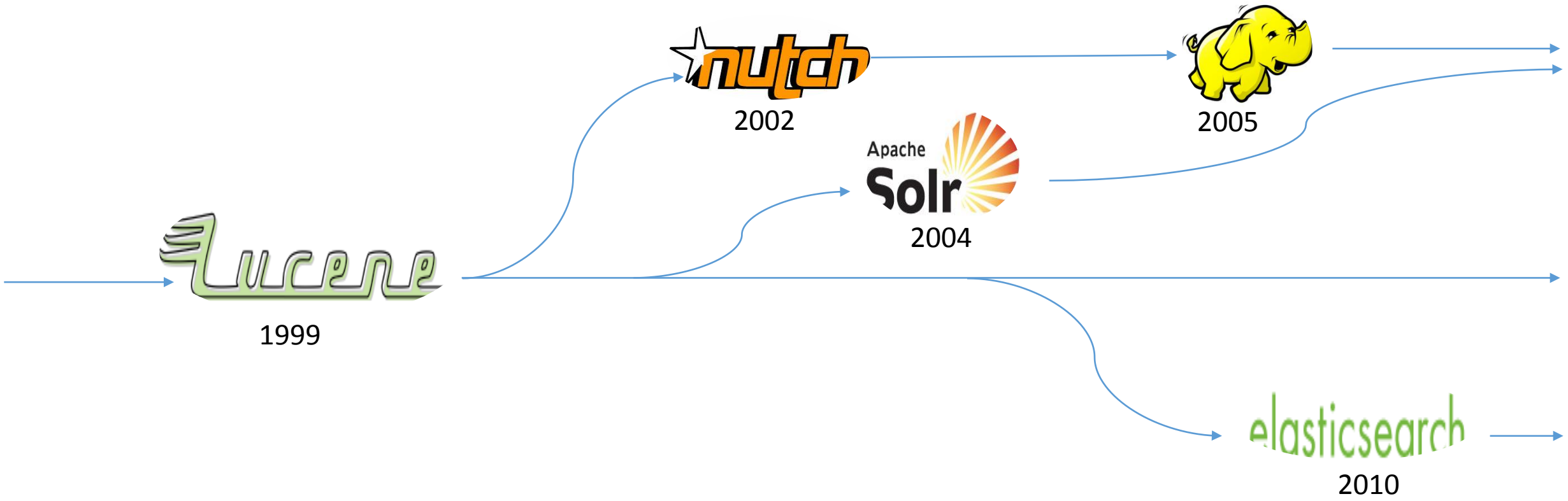
France Labs
Open Source Enterprise Search

Plan

- Rappels sur Hadoop
- Présentation (rapide) de Solr
- **Intégration d'Hadoop et de Solr**
- Démo

Présentation d'Hadoop

Projets liés



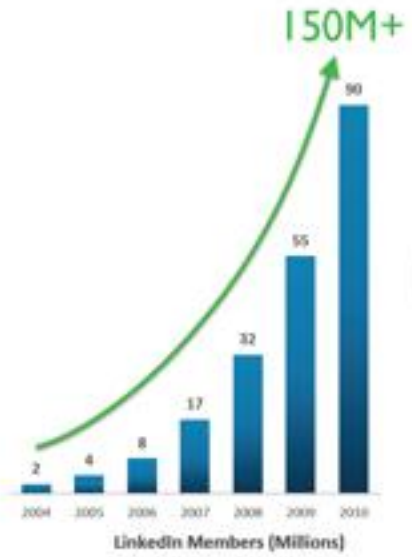
Intégration d'Hadoop et Solr

Exemples de scénarios Big Data

- Text mining
- Création et analyse de graphes
- Reconnaissance de patterns
- Moteurs de recommandations, ciblage publicité
- Analyse de menaces, surveillances de marchés

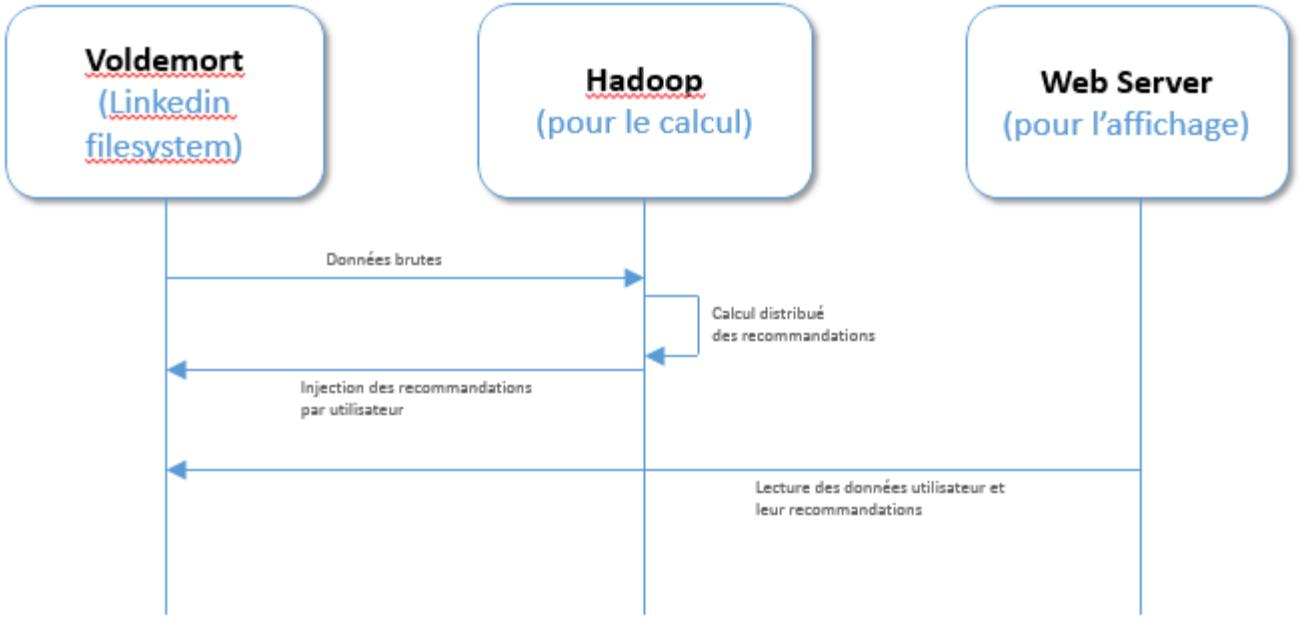
Intégration d'Hadoop et Solr

Hadoop pour les recommandations



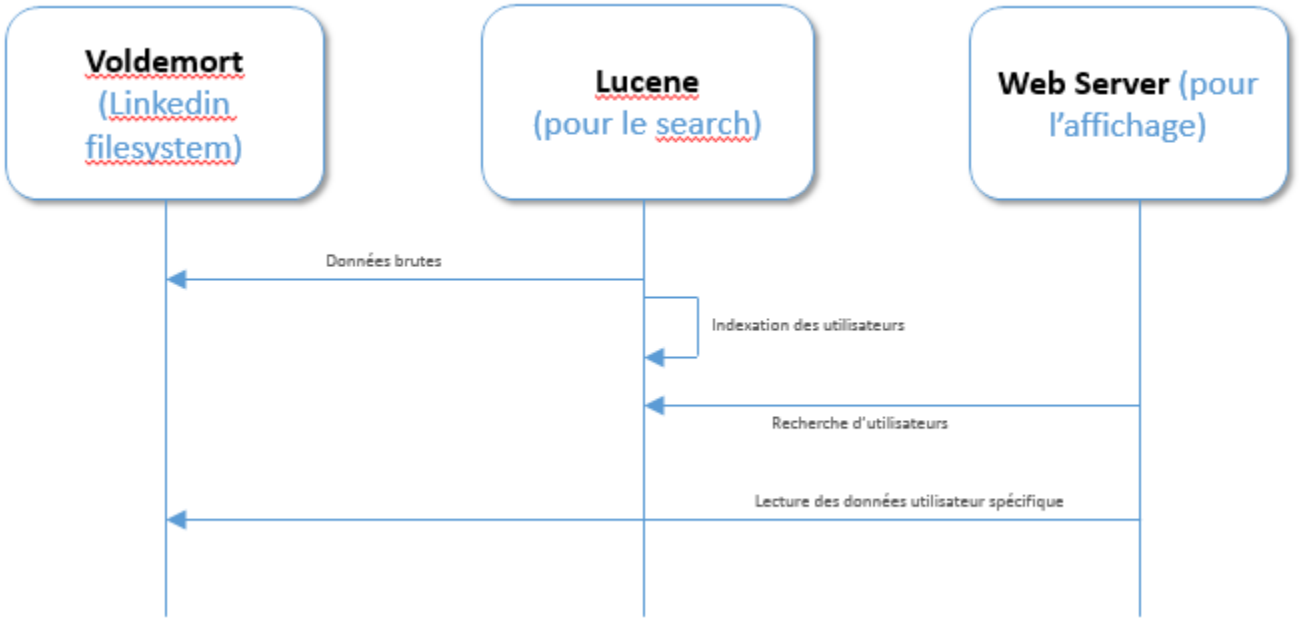
Intégration d'Hadoop et Solr

Hadoop pour les recommandations



Intégration d'Hadoop et Solr

Lucene pour la recherche d'utilisateurs



Intégration d'Hadoop et Solr

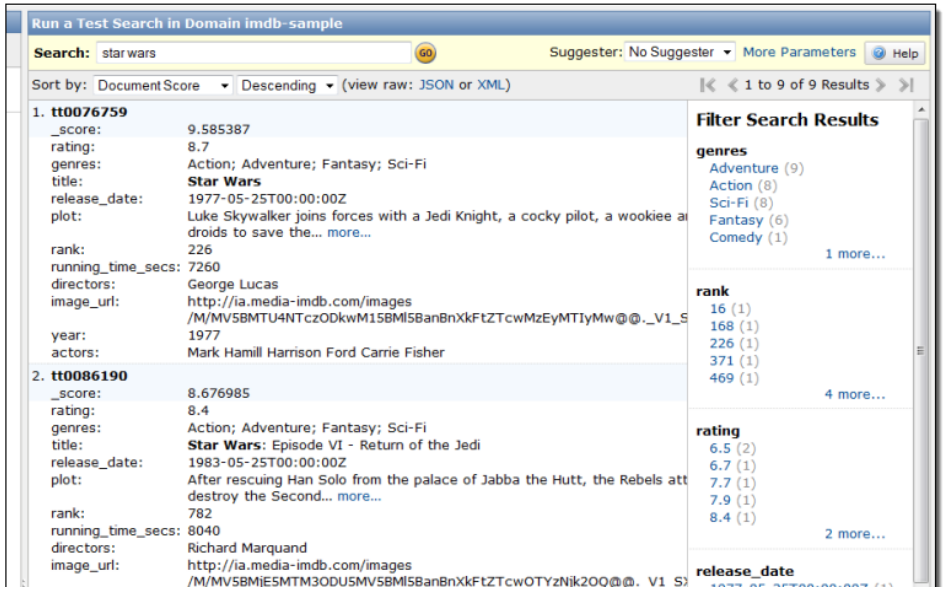
Exemples d'utilisation d'Hadoop

- Netflix
  - Site de streaming légal de vidéos
  - Avant Hadoop :
    - Logs étaient traités pendant la nuit
    - Importés dans une BDD
    - Analyse/BI
    - => Il fallait plus de 24h pour traiter une journée de logs
  - Avec Hadoop :
    - En seulement une heure de jobs Hadoop les données sont traitées
    - Traite quotidiennement 1 To de données par jour approximativement
  - Solr
    - Recherche de contenus par l'utilisateur

Intégration d'Hadoop et Solr

Exemples d'utilisation Big Data

- Amazon CloudSearch
  - Utilisation de Solr depuis Mars 2014
  - Scalable
  - Performant



Intégration d'Hadoop et Solr

- Scénario : analyse de tickets de caisse
  - Que faire des données brutes ?
  - Comment les analyser, les traiter ?
  - Comment rechercher dedans ?
  - => Solution : Hadoop et Solr !

BOUTON OR CHIPS SCHT	Eur1,17 A
FRUIT ET LEGUME	Eur0,80 A
TOP BUDGET BISCOTTES	Eur1,68 A
TB SUCRE POUDRE KG 1	Eur0,72 A
CHABRIOR EXTRA MOEL.	Eur0,52 A
TB SUCRE POUDRE KG 1	Eur0,72 A
TB SUCRE POUDRE KG 1	Eur0,72 A
PAT MUNSTER 200G	Eur1,95 A
T.BUDGET SALAMI 20TR	Eur0,85 A
PAT PTITS FONDANTS 2	Eur2,55 A
T.BUDGET RIZ LONG CE	Eur0,74 A
PAT YT A BOIR.FRAMBO	Eur1,47 A
T.BUDGET COQUILLETTE	Eur0,72 A
TOP BUDGET SPAGHETTI	Eur0,69 A
PAT C.FRAICHE ISIGNY	Eur2,15 A
EMMENTAL 45% PORTION	Eur2,94 A
LIMES EMERI PM	Eur2,70 B
BOUCHERIE LS	Eur3,00 A
BOUCHERIE LS	Eur1,00 A
CHARCUTERIE LS	Eur0,50 A
*ELODIE CUBO OURSON	Eur3,49 B
MONTANT DU	Eur31,08
CHEQUE AUTO.	Eur31,08

Nombre d'articles vendus= 21

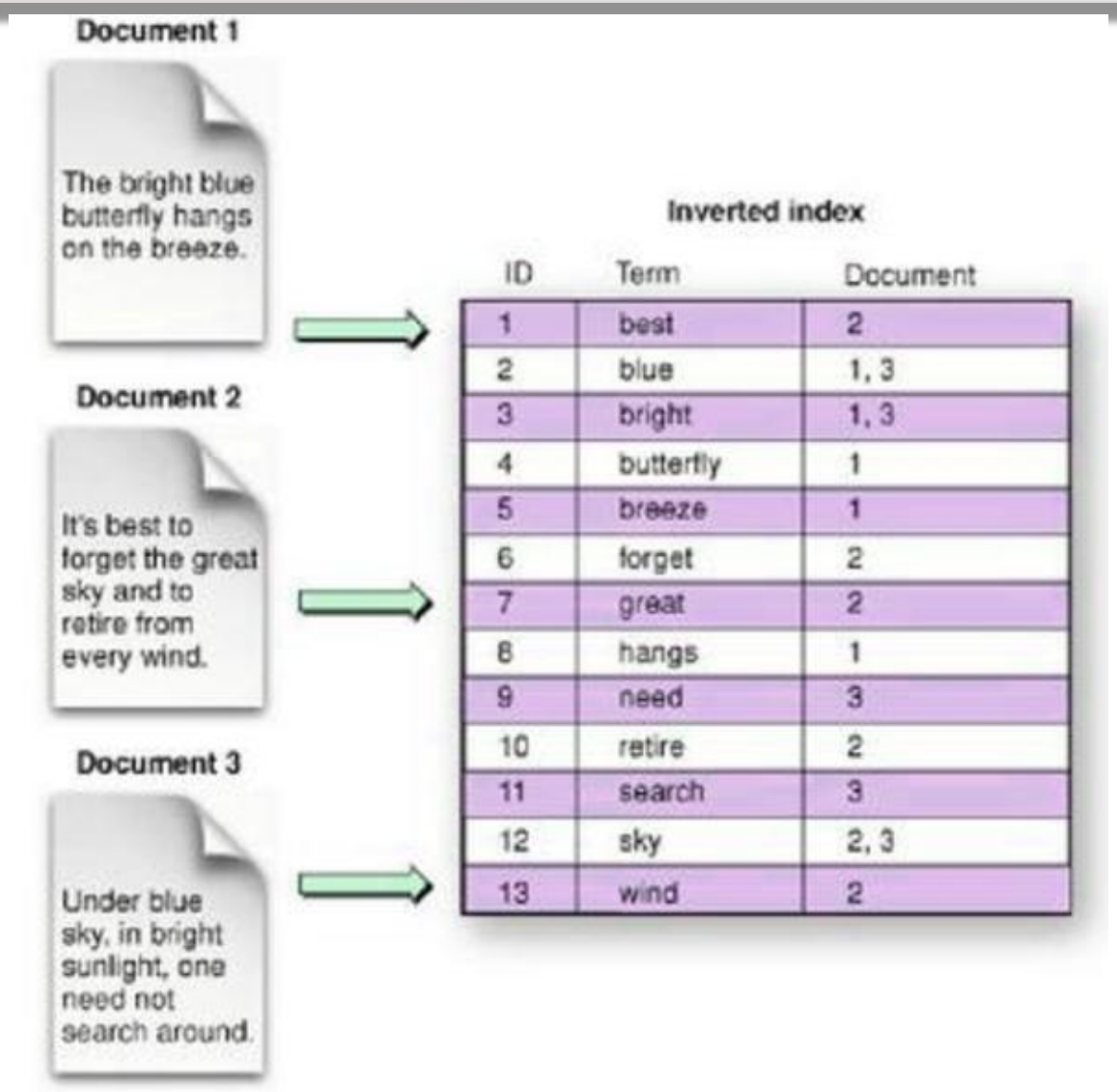
Intégration d'Hadoop et Solr

1er niveau d'intégration

- Hadoop et Solr indépendants
  - Traitements et calculs sur Hadoop (ex : Pig)
    - Import d'un CSV avec Pig
    - Opérations : analyse, calculs panier moyen, TVA etc…
  - Export des données d'Hadoop
    - Génération d'un fichier XML avec Pig
  - **Indexation** dans un Solr standalone (ex : DIH)
    - Utilisation d'une contribution de Solr : DIH

Intégration d'Hadoop et Solr

Index Solr



Intégration d'Hadoop et Solr

1er niveau d'intégration

- Avantages :
  - Si cluster Solr déjà présent, ajout d'un index Solr dédié
  - Mature
- Inconvénients:
  - Réplication de HDFS vers système de fichiers du Solr

Intégration d'Hadoop et Solr

2e niveau niveau d'intégration

- Stockage de l'index de Solr sur HDFS
  - Lancement de Solr avec utilisation de HDFS

```
java -Dsolr.directoryFactory=HdfsDirectoryFactory
     -Dsolr.lock.type=hdfs
     -Dsolr.hdfs.home=hdfs://host:port/path
```

Intégration d'Hadoop et Solr

2e niveau niveau d'intégration

- Avantages :
  - Tolérance aux pannes
  - Gain espace disque (cluster Hadoop déjà existant avec HDFS)

Intégration d'Hadoop et Solr

Différents niveau d'intégration

- Création de l'index Solr à l'aide de jobs Map Reduce
  - Travail en cours (https://issues.apache.org/jira/browse/SOLR-1045)
  - Contrib expérimentale fournie avec Solr
  - Avantages :
    - Augmentation des performance d'indexation sur de gros volumes de données
    - Très utile si indexes déjà construits par HDFS

Intégration d'Hadoop et Solr

Solr

- Fonctionnalité de Search basée sur Solr intégrée par plusieurs distributions
  - MapR
  - Cloudera Search
  - Hortonworks
  - Lucidworks

Plan

- Rappels sur Hadoop
- Présentation (rapide) de Solr
- Intégration d'Hadoop et de Solr
- **Démo**

Démo Hadoop et Solr

Work in progress

- Scénario des tickets de caisse
- Etape 1 : import dans Pig des tickets en CSV
- Etape 2 : opérations sur les données avec Pig
- Etape 3 : indexation des données en Solr
- Etape 4 : recherche sur ces données avec Solr

Démo Hadoop et Solr

Work in progress

```
<field name="id" type="uuid" indexed="true" stored="true" required="true"/>

<field name="gtin_cd" type="string" indexed="true" stored="true"  multiValued="false" />
<field name="gtin_nm" type="string" indexed="true" stored="true"  multiValued="false" />
<field name="gtin_img" type="string" indexed="true" stored="true"  multiValued="false" />
<field name="brand_nm" type="string" indexed="true" stored="true"  multiValued="false" />
<field name="price" type="float" indexed="true" stored="true"  multiValued="false" />
<field name="id-show" type="string" indexed="true" stored="true"  multiValued="false" />
<field name="shop-name" type="string" indexed="true" stored="true"  multiValued="false" />
 <field name="categorie" type="string" indexed="true" stored="true"  multiValued="false" />

<field name="position" type="location" indexed="true" stored="true"/>
```


```
"docs": [
  {
    "gtin_cd": "0071249812044",
    "gtin_nm": "Hydra Perfecte Concealer Medium Deep",
    "gtin_img": "http://product.okfn.org.s3.amazonaws.com/images/gtin/gtin-007/0071249812044.jpg",
    "brand_nm": "L'Oréal Paris",
    "price": 6.21,
    "id-show": "shop-48888",
    "shop-name": "Champs sur Marne",
    "position": "48.85436,2.58183",
    "categorie": "article",
    "id": "f3bf1047-8b0c-42f7-b77a-1c7d3d48fcd1",
    "_version_": 1466082742653419500
  },
```

Démo Hadoop et Solr

Solr directement couplé avec Hadoop

Démo !

Contact

Site web : www.francelabs.com

Email : contact@francelabs.com

Twitter : Francelabs