FRANCELABS
EXPERTS DU SEARCH

Apache SolrCloud pour la fouille de données Big Data

AURÉLIEN MAZOYER, SEARCH EXPERT, FRANCE LABS
SOPHIACONF WORLDWIDE EVENT
JULY 2015

France Labs – Qui sommes-nous ?

LES EXPERTS DU SEARCH:

- Consulting sur les meilleures technos open source de search
  - Apache Solr
  - Elasticsearch
  - Apache ManifoldCF
  - Formations, audit techniques, AMOA, MOE, MEP, LOL, MDR
- Editeur de Datafari:
  - Solution clé en main de recherche
  - Gestion de la sécurité
  - Big data
  - Multisource, multi-types
  - Sémantique

Agenda

- Solr:
  - Kezako ?
  - Place dans les Data Science !
  - C’est fastoche !
- SolrCloud:
  - C’est le Big Data
  - Kezako ?
  - C’est fastoche !
- Solr et Data Science:
  - Cas d’usages

Solr – Kezako ?

- Moteur de recherche et d'analyse
- Fondation Apache
- Basé sur Lucene
- Fonctionnalités : autocomplétion, faceting, suggestions, highlighting…

Solr et les Data Science

- L’axe de la fouille de données
- Pas d’algorithmes complexes
- Priorité à l’instantané
- Sommes, min, max, pourcentages, moyenne

Solr c’est fastoche

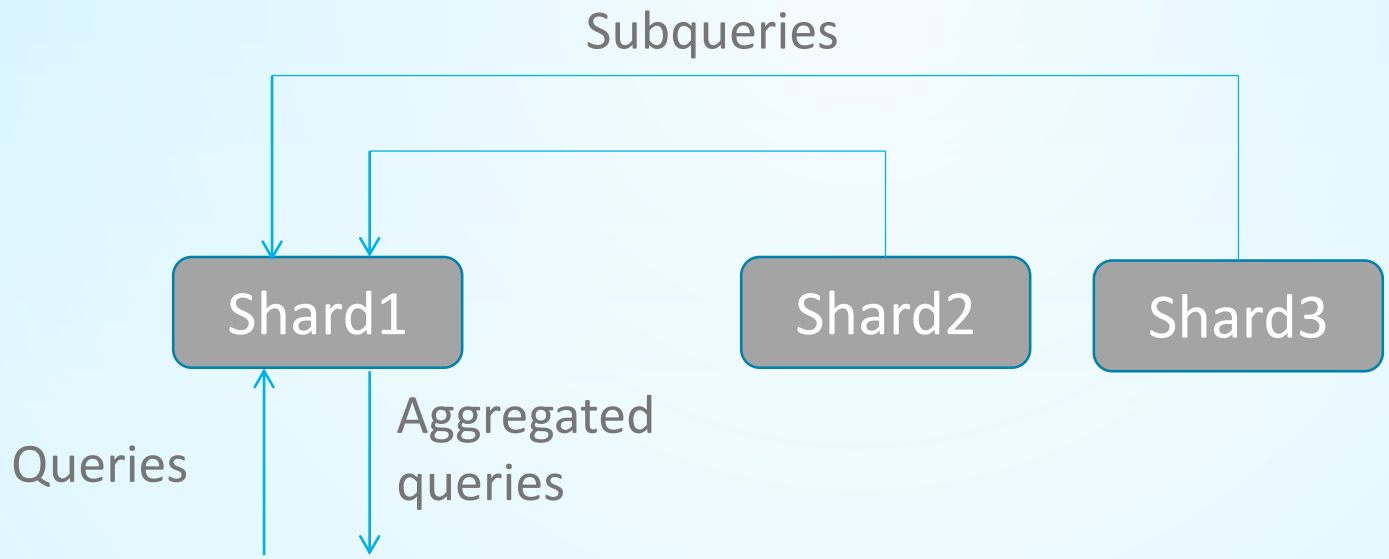
- La preuve par l’exemple

SolrCloud – C'est le Big Data

- Scalabilité horizontale pour:
  - Gérer un index énorme
  - Gérer un nombre massif de requêtes
  - Gérer la haute disponibilité
- Même philosophie qu'Hadoop – Utilise Zookeeper
  - Gère la configuration des systèmes distribués
  - Conserve et gère les configs d'index du SolrCloud
  - Conserve et gère le statut des shards enregistrés
  - Les Solr s'abonnent et se désabonnent auprès du Zookeeper

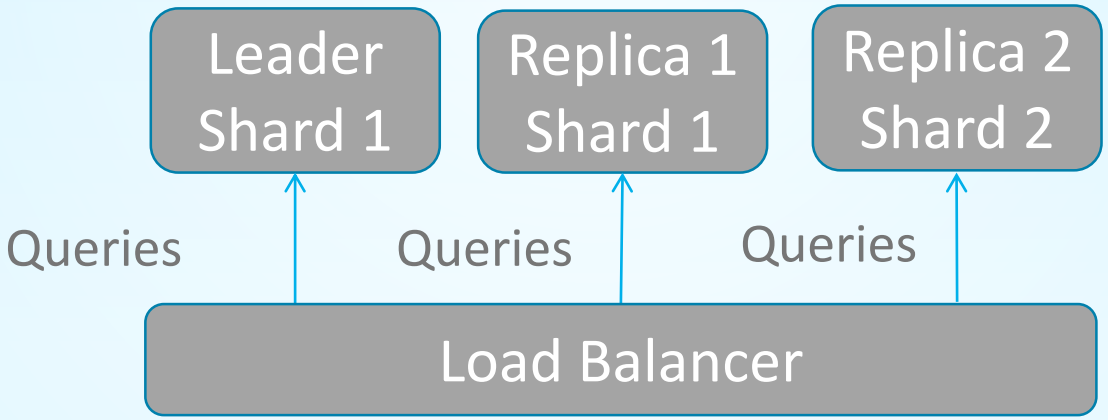
SolrCloud – Les shards

- Sharding :
  - Un shard est un morceau d'index
  - Une recherche distribuée se fait sur tous les shards (donc l'index complet)
  - Utile pour gérer un gros index



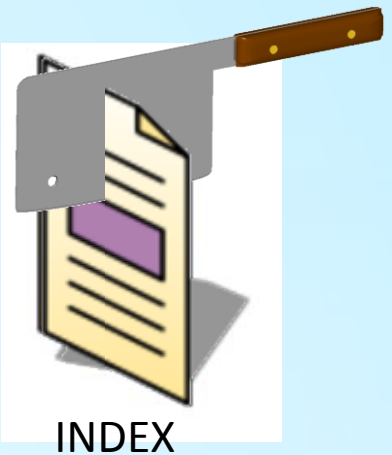
SolrCloud: Leaders/Replicas

- Leaders et replicas pour la réplication:
  - Pas de « maître/esclave » mais des « leaders/replicas » (un replica peut devenir un leader)
  - Un leader et ses replicas contiennent le même shard
  - Utile pour gérer une grosse charge de requêtes et pour la haute disponibilité



SolrCloud - Illustration

- On veut répartir l’index sur 2 shards
- On veut un réplica par shard
- D’où 4 machines



SolrCloud – Démo locale

- Démarrage d’un cluster local, check de la config et arrêt brutal d’un serveur

SolrCloud – Démo AWS

- Indexation de 2.300.000 pages web pour jouer un peu
- Quasi instantané

Solr et Big Data – Cas d'usages

- BOX:
  - Index de 10 To
  - 10 Mds docs
  - 100 M requêtes / jour

- Bloomberg
  - 80 Mds docs
  - Doc ~ 50 Ko, variance 1ko -> 100 Mo
  - Centaines de champs
  - <10 requêtes / seconde

FRANCELABS

EXPERTS DU SEARCH

QUESTIONS ? (Que représente le logo France Labs ?!)

A VOTRE DISPOSITION AU COCKTAIL

AURELIEN.MAZOYER@FRANCELABS.COM
@FRANCELABS