ECナビ 商品検索でのSolrの利用

株式会社ECナビ
システム本部ECナビラボ
春山 征吾

アジェンダ

- ECナビでのSolrの利用
- ECナビ商品検索の概要
- システム構成
- ECNaviTokenizer
- 問題点

ECナビでのSolrの利用

- 商品検索
- ソーシャルブックマークサイトBuzzurlでの検索
- …

商品検索でのSolrの利用

以下でSolrを利用しています.

- 通常の全文検索 (ECナビとすべてのOEM)
- アイテム一覧ページ (一部のOEMを除く)
- NavicSearch API

ECナビ商品検索の概要(続き)

件数 約2300万件

クエリ数/日 最大約100万

元データサイズ 約10GB

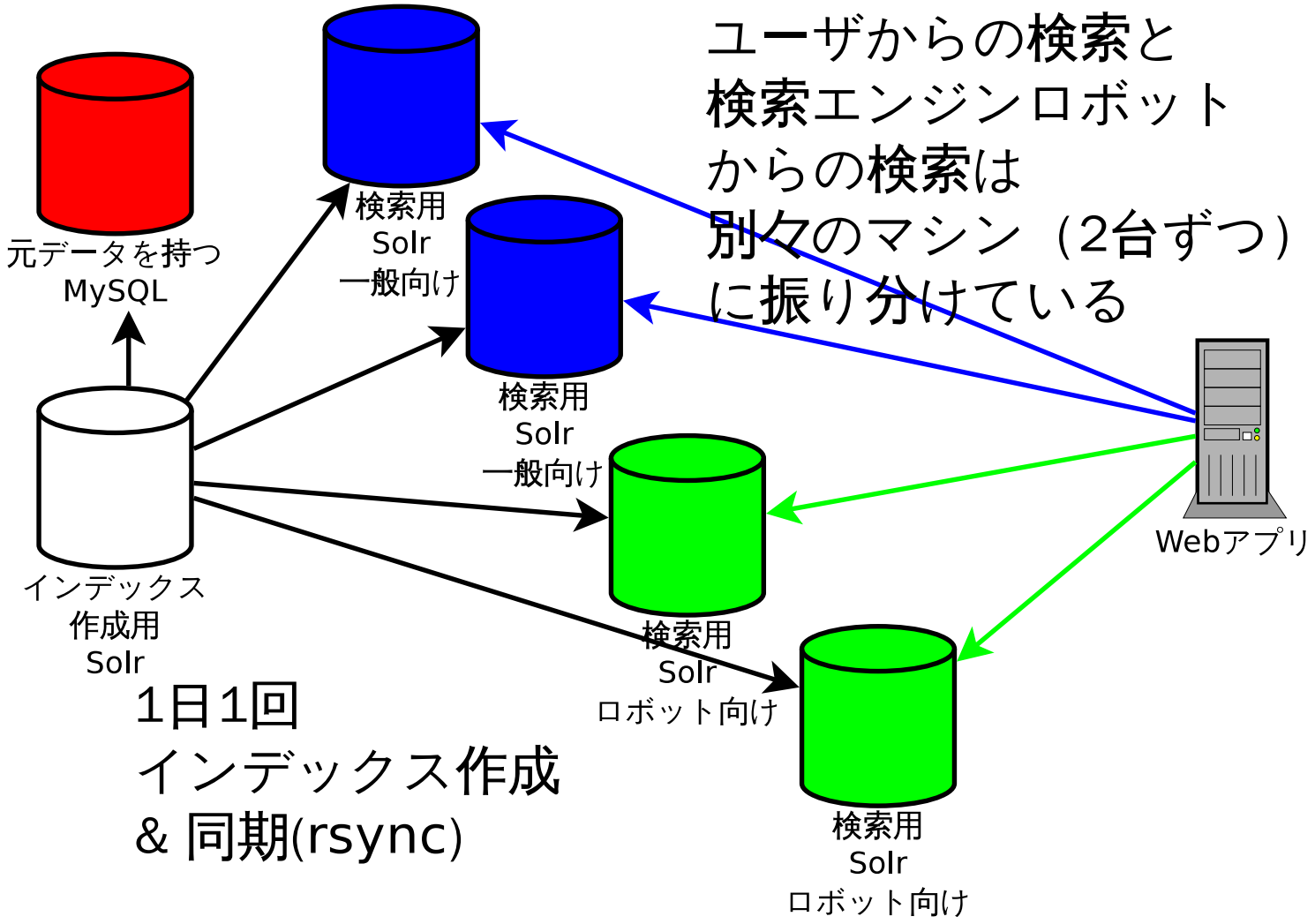
インデックスサイズ 約18GB

利用しているSolrの機能

- DistributedSearch(shards)
- rsyncでのレプリケーション
- ファセット
- 独自Tokenizerの組込み

デモ

システム構成


元データを持つ
MySQL
検索用
Solr
一般向け
検索用
Solr
一般向け
ユーザからの検索と
検索エンジンロボット
からの検索は
別々のマシン（2台ずつ）
に振り分けている
Webアプリ
インデックス
作成用
Solr
検索用
Solr
ロボット向け
検索用
Solr
ロボット向け
1日1回
インデックス作成
& 同期(rsync)

マシンスペック

インデックス作成機

メモリ 4GB

HDD 120GB RAID なし

検索機

メモリ 18GB

HDD 120GB RAID1+0

ソフトウェア構成

- Solr 1.3
- Tomcat 6
- Sun JDK 1.6
- CentOS 4.6

インデックス作成

- １日１回インデックス作成
  → rsyncでレプリケーション
- shardを３つにわけて並列に更新
- インデックス更新にかかる時間は,
  ２時間

検索

- 検索エンジンのロボットからのクエリが多数
- そのため，４台のマシン中２台がロボット用２台が人間用

検索 (続き)

どれくらいロボットからのクエリが多いかというと

ロボット 50 万〜80 万/日

それ以外 20 万弱/日

ECNaviTokenizer

武田さんのCJKTokenizerを改造したもの

- 「つのだ☆ひろ」「漫★画太郎」といった語句の検索のために「☆」などの文字があってもなくても同じTokenが出力されるようにした

ECナビ独自のTokenizer(続き)

- 商品の型番によくある 「cyber-shot」「cybershot」などといったハイフンの表記揺れに対応するため，「cyber-shot」から「cyber」「shot」「cybershot」というTokenを生成するようにした．(検索時には，すべてのハイフンは取り除いてクエリが送られる)
- いわゆる半角カナを全角カナに正規化するようにした

問題点

- インデックスがメモリに乗らない大きさ
- 検索エンジンロボット由来のクエリにはキャッシュの効果がない
- Disk Accessが発生 → HDDがボトルネック

というわけでＬＴに続く