Solr&Lucene

搜索引擎技术

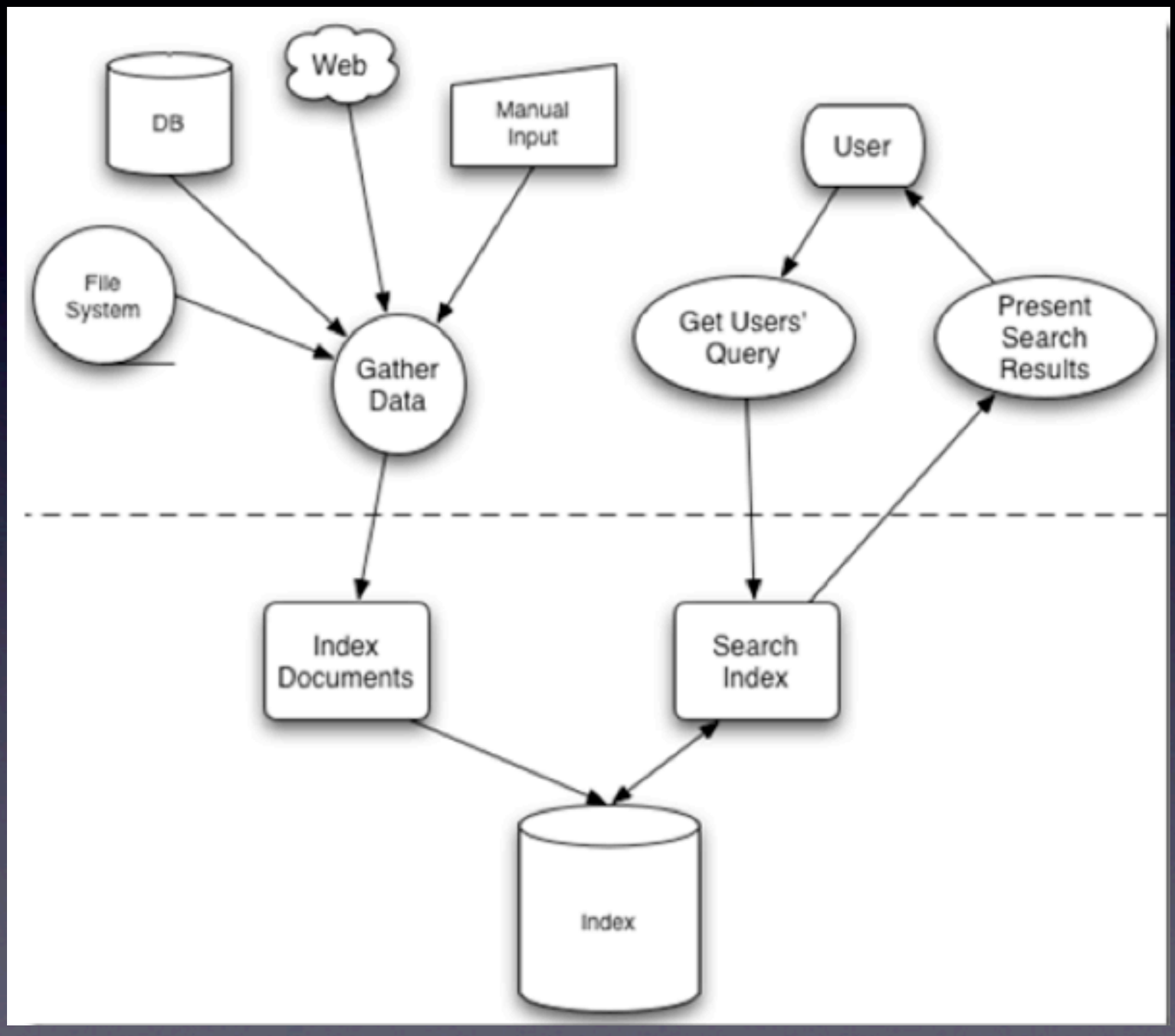
如何检索?

- search算法　（同类的东西）
- DB 对多个维度进行查询（Index）
- 顺序搜索不规则的文档
- 倒序索引

倒序索引

- DB 正序索引 key -> column -> rows
- 倒序索引 keywords -> docs

lucene解决方法


Web
DB
Manual Input
File System
Gather Data
User
Get Users' Query
Present Search Results
Index Documents
Search Index
Index

索引存储

- 文件方式存储
- lucene定义的结构，完成文档的存储

文档

- 文档定义 （solr中schema.xml）
- 文档生成（文本，DB，image...）

搜索

- QueryParser分析
- 结果呈现
- 查询结果相关统计

分词

- 将文档切分成许多的关键词
- 英文中，按空格分，另外一些做单独处理
- 中文。（ik-analyzer,mmseg4j,paoding ...）

Solr是什么？

- 基于lucene
- 简化索引过程，文档结构定义等
- 提供http形式的搜索接口，便于使用
- 充分的定制性

Step 1

- 定义Schema.xml
  - 哪些内容要索引，索引的内容是什么类型，要不要做分词？（field，fieldType，analysier...）
- 定制自己需要的fieldType （分词器...）

Step II

- 编写建索引程序
- solr支持json，xml，csv，javabean方式建索引
- 通过http接口发送需要索引的数据给Solr服务器

Step III

- 选择需要的组件，通过solrconfig.xml配置
- 选择自己需要的SearchHandler
  - edismax 提供N多有用的特性
- 确定打分算法 (boost)
  - boostQuery，functionQuery...

Step IX

- 编写搜索程序，基于HTTP接口
- lucene语法
- 组件支持的功能 eg：facet,highlight,edismax,wt,

Step X

- 定制自己的分词器
  - ik-analyzer
  - multiCore

Step XI

- replication
- HA
- realtime search...

Solr & Lucene Rocks!

Q&A