Nutch Lucene Solr

2014年4月5日 13:24

"nutch 爬虫生成索引导入到solr中进行查询" nutch是hadoop的,solr本身不是hadoop的,而是solrcloud.

hadoop提供的这个和lucene结合的实质，是将索引写到本地文件，再传到hdfs上。

6， 我在搜索资料的过程中，了解到几种开源产品，诸如blur ， nutch ， katta等用于**整合** lucene和hadoop，虽然我没有详细看过他们源码，但是从网友的讨论中得知，nutch也是基于先写本地再写hdfs，其他 两个被使用的比较少，具体实现还不知道。

来自 <http://www.verydemo.com/demo_c441_i10123.html>

7	Lucene in action中文版	(美) Otis Gospodnetic, Erik Hatcher著；谭鸿[等]译	电子工业出版社	2007	TP393.09/G27	8	5	
8	征服Ajax + Lucene构建搜索引擎	李刚,宋伟,邱哲编著	人民邮电出版社	2006	TP393.09/L31	6	5	

来自 <http://opac.lib.szu.edu.cn/opac/searchresult.aspx?anywords=lucene&dt=ALL&cl=ALL&dept=ALL&sf=M_PUB_YEAR&ob=DESC&page=1&dp=20&sm=table>

Building a Search Enginewith Nutch and Solr in 10minutes

来自 <http://www.building-blocks.com/thinking/building-a-search-engine-with-nutch-and-solr-in-10-minutes/>

How SolrCloud Works

来自 <https://cwiki.apache.org/confluence/display/solr/How+SolrCloud+Works>

lucene索引数据放在哪里？

放在hbase，hbase是典型的key-value存储，不适合跨行事务。虽然在随机读取方面借助缓存有优势，但如果一次事务读取key太多，延迟也挺低的。我在hbase上面存储lucene index以及借助hbase机制重新设计lucene index，性能始终不是特别好。

如果lucene+hadoop，索引数据存储在hdfs，而hdfs目前并不适合大量小文件存储。lucene索引过程涉及大量索引小文件合并，而且hdfs设计目标也并不是一个实时流读写系统，因此对于lucene核心的near time index也是个挑战。

生搬硬凑在一起玩玩可以，产线恐怕还真要慎重。

目前最成熟的都是通过hadoop mapreduce生成索引先存储在hdfs,然后从hdfs将索引下载到本地的solr索引目录，整个过程在于加快全量索引生成效率。

来自 <http://www.oschina.net/question/220491_117079>

Just rephrasing it once, nutch just crawls and stores data in db and it does not do indexing by itself. Solr is needed to index. Am I right?

来自 <http://stackoverflow.com/questions/10844792/nutch-vs-solr-indexing>

but I can make Nutch to call Solr to index the data Nutch has crawled if I have Solr installed too. This can be done running a single command which tell Nutch to both crawl and index.

summary:

目前的Hadoop与Lucene的结合方法都是这样的:

由于Lucene需要随机读/写索引,而Hadoop只支持随机读,不支持随机写,只能顺序写,所以要使用Hadoop,需要一些折衷的方法.

①contrib/index: 这个是Hadoop自己提供的方法. 这个方法中,因为hdfs只支持序列化的写所以索引文件无法被直接存在hdfs上，所以会被存在lucene的本地目录上。这个方法先在本地建一个tmpdir,把索引写在上面,然后再上传到hadoop上的perm dir,删除tmpdir. hadoop提供的这个和lucene结合的实质，是将索引写到本地文件，再传到hdfs上。在这个方法下,MapReduce的只是用作建立索引的过程,而不是搜索过程.

②Nut: 索引不放在HDFS上了,索引放在一个名叫"索引服务器"或"搜索服务器"的地方(可以是一个集群). 另外,使用Hadoop Mapper/Reducer 建立索引。再将索引从HDFS分发到各个索引服务器。

③solr: 使用"SolrCloud"项目来提供实现分布式特性. 也是就是用SolrCould来管理这些分布式的索引数据和进行分布式的搜索. 简单地说,Solr与MapReduce无关.

However,

Solr has support for writing and reading its index and transaction log files to the HDFS distributed filesystem. This does not use Hadoop Map-Reduce to process Solr data, rather it only uses the HDFS filesystem for index and transaction log file storage.

来自 <https://cwiki.apache.org/confluence/display/solr/Running+Solr+on+HDFS?focusedCommentId=33297067#comment-33297067>