**ElasticSearch大数据搜索技术概述**

**摘要**

随着大数据时代的到来，数据量以极快的速度增长，形成了海量数字资源。在此情况下，数据检索面临着检索不全面和检索时间长的问题。（概述本文所讲内容）

**1 引言**

（写一下背景）

* 1. 国内外搜索技术现状

目前的搜索系统主要采用两种技术模式：1)以数据库为基础，利用成熟的商业检索系统，如Oracle, Mysql ; 2)在全文检索基础上开发的搜索系统。

* 1. 介绍

**2 ElasticSearch介绍**

EasticSearch(简称ES)是一个基于Lucene构建的开源、分布式、RESTful搜索引擎。在云计算中，ES能够达到实时搜索，稳定，可靠，快速，支持通过HTTP使用JSON进行数据索引。它致力于使开发者能够使用尽量少的开发成本快速实现一个功能完善的检索系统。主要特性有：

1. 集群（cluster）。集群中有多个节点，其中一个为主节点，可以通过选举产生，主从节点是相对于集群内部来定义的。ES中一个概念是去中心化，是想对集群外部来定义的。因为从外部看ES集群，从逻辑上是个整体，与任何一个节点的通信和与整个集群通讯是等价的。
2. 索引分片（shards）。ES把一个完整的索引分成多个分片，把一个大的索引拆分成多个，分布到不同节点上，构成分布式检索。
3. 索引副本（replicas）。ES可以设置多个索引副本，主要作用有两个，一个是提高系统的容错性，当某个节点的某个分片损坏或丢失时，可以从副本中恢复；另一个是提高ES的查询效率，ES会对检索请求进行负载均衡。
4. 数据恢复（recovery）。ES在有节点加入或退出时会根据机器的负载对索引分片进行重新分配，挂掉的节点重新启动时也会进行数据恢复。
5. ES索引的持久化存储方式（gateway）。ES默认先把索引放在内存中，当内存满了再持久化到硬盘。当这个ES集群重新启动时就会从gateway中读取索引数据。
6. Discovery.zen. ES的自动发现节点机制。ES是一个基于P2P的系统，先通过广播寻找存在的节点，再通过多播协议进行节点之间的通信，同时也支持点对点的交互。  
   2.2 索引数据结构

ES索引采用倒排索引机制构建索引。

1. Item: 最小的索引单位，直接代表一个关键词以及其在源文档中的出现位置和出现次数等信息。
2. Field: 一个关联的元祖，包括域名和域值。域名是一个字符串，域值是一个项。
3. Document: 包括所有的域信息。
4. Segment: 包含若干个文档，若干段组成子索引或索引。

ES索引的数据结构如下图所示：

2.3 分词技术

分词技术是查询串进行初步处理后，用各种匹配方法进行匹配的一种技术。分为3类：

1. 机械分词法：
2. 语义分词法：
3. 统计分词法：  
   目前主要的分词器如表一所示：

**5 结束语**

本文在深入研究分布式检索的原理和全面剖析ES相关技术的基础上...