net.ai需求

通过搜索关键字抓取网站内容

技术架构

1数据库nosql，

2 http模拟okhttp,

3 日志打算主要用kafka试探集成flume参考（第11章 场景实战：Kafka 与日志采集高性能服务系统构建与实战）

4 分布式的定时任务

5 图片分布式文件服务

6 maven架构，多个项目

7 尝试搜索引擎

需求

网站分类 1技术 2网盘 3 新闻 等

Csdn

1 根据关键字抓取csdn文章，存在nosql数据库中

2 页面需要列表展示，可以通过标题关键字搜索，文章关键字搜索，点击标题可以以（word

，pdf） 形式看到整个文章。

注：需要特殊抓取的：<http://blog.csdn.net/lifetragedy?t=1>;csdn上袁鸣凯的原创文章

特别是 ：通向架构师的道路系列和亲密接触Redis系列单点登录之CAS SSO从入门到精通系列

pan.baidu.com

根据关键字抓取pan.baidu.com的信息，存储

展示，搜索