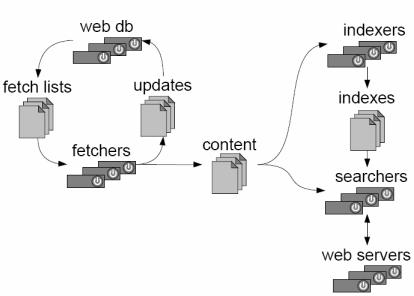
Nutch从总体上看来，分为三个主要的部分：爬行、索引和搜索，各部分之间的关系图所示。Web db是Nutch初始运行的URL集合；Fetcher是用来抓取网页的爬行器，也就是平时常说的Crawler；indexer是用来建立索引的部分，它将会生成的索引文件并存放在系统之中；searcher是查询器，用来完成对某一词条的搜索并返回结果。



**爬虫，**Crawler**：**

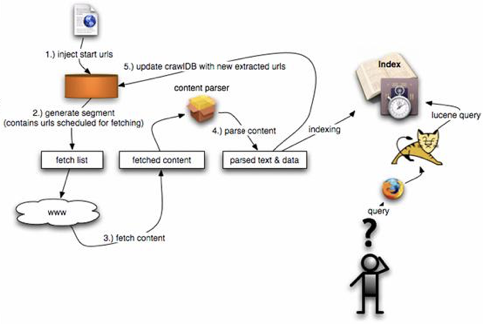
Crawler的重点在两个方面，Crawler的工作流程和涉及的数据文件的格式和含义。数据文件主要包括三类，分别是web database，一系列的segment加上index，三者的物理文件分别存储在爬行结果目录下的db目录下webdb子文件夹内，segments文件夹和index文件夹。那么三者分别存储的信息是什么呢？

Web database，也叫WebDB，其中存储的是爬虫所抓取网页之间的链接结构信息，它只在爬虫Crawler工作中使用而和Searcher的工作没有任何关系。WebDB内存储了两种实体的信息：page和link。Page实体通过描述网络上一个网页的特征信息来表征一个实际的网页，因为网页有很多个需要描述，WebDB中通过网页的URL和网页内容的MD5两种索引方法对这些网页实体进行了索引。Page实体描述的网页特征主要包括网页内的link数目，抓取此网页的时间等相关抓取信息，对此网页的重要度评分等。同样的，Link实体描述的是两个page实体之间的链接关系。WebDB构成了一个所抓取网页的链接结构图，这个图中Page实体是图的结点，而Link实体则代表图的边。

一次爬行会产生很多个segment，每个segment内存储的是爬虫Crawler在单独一次抓取循环中抓到的网页以及这些网页的索引。Crawler爬行时会根据WebDB中的link关系按照一定的爬行策略生成每次抓取循环所需的fetchlist，然后Fetcher通过fetchlist中的URLs抓取这些网页并索引，然后将其存入segment。Segment是有时限的，当这些网页被Crawler重新抓取后，先前抓取产生的segment就作废了。在存储中。Segment文件夹是以产生时间命名的，方便我们删除作废的segments以节省存储空间。

Index是Crawler抓取的所有网页的索引，它是通过对所有单个segment中的索引进行合并处理所得的。Nutch利用Lucene技术进行索引，所以Lucene中对索引进行操作的接口对Nutch中的index同样有效。但是需要注意的是，Lucene中的segment和Nutch中的不同，Lucene中的segment是索引index的一部分，但是Nutch中的segment只是WebDB中各个部分网页的内容和索引，最后通过其生成的index跟这些segment已经毫无关系了。

Nutch 的运行流程如下图所示。

1. 将起始 URL 集合注入到 Nutch 系统之中。

2. 生成片段文件，其中包含了将要抓取的 URL 地址。

3. 根据URL地址在互联网上抓取相应的内容。

4. 解析所抓取到的网页，并分析其中的文本和数据。

5. 根据新抓取的网页中的URL集合来更新起始URL集合，并再次进行抓取。

6. 同时，对抓取到的网页内容建立索引，生成索引文件存放在系统之中。