# Nutch分析

## Architecture



图1 Nutch Architecture

1. **Injector：**takes a flat file of URLs and adds them to the of pages to be crawled (CrawlDB)

The URL files contain one URL per line, optionally followed by custom metadata separated by tabs

* 1. **CrawlDB:** is Mapfile <Text, CrawlDatum>

1. **Generator:** 
   1. Select, sort by score, limit by URLs/host
   2. Partition by host, sort randomly
2. **Fetcher：**
   1. content：is Mapfile <Text, Content>
   2. crawl\_fetch
3. **Parse segment**
   1. **crawl\_parse**
   2. **parse\_data：**is Mapfile <Text, ParseText>
   3. **parse\_text：**is Mapfile <Text, ParseData>
4. **CrawlDB Update**
5. **InvertLinks**
   1. **LinkDB：is Mapfile <Text, Inlinks>**
6. **Indexer**
   1. **Index Dedup**

## 

### Injector