

目录结构说明

- `mapreduce` 为项目源代码
- `result` 目录为最后生成的在HDFS系统下的截图与倒排文件
- `data.txt` 为20,000条记录文件
- `HBase` 为设计的 HBase 数据库。`paper` 表记录 `paper` 的所有信息，`row`为 Java 系统的纳秒时间，`author` 表的存储设计为 `row` `papername: paperId = paperName`，其中`row`为 `authorName`，而 `papername: paperId` 中的 `papername` 为列族，`paperId` 为相应 `paper`在`paper`表中的`rowKey`，`paperName` 为`paper`的`title`。这样的设计是为了考虑两点：
 1. 一个作者可以写几个`paper`;
 2. 一次查询就可以获的查询`paper`的入口。