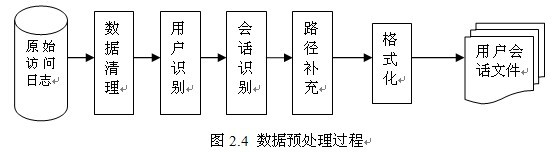
# 模块开发----数据预处理

## 主要目的

过滤“不合规”数据，清洗无意义的数据

格式转换和规整

根据后续的统计需求，过滤分离出各种不同主题(不同栏目path)的基础数据。



## 实现方式

开发一个mr程序WeblogPreProcess(***内容太长，见工程代码***)

|  |
| --- |
| *public class WeblogPreProcess {*  *static class WeblogPreProcessMapper extends Mapper<LongWritable, Text, Text, NullWritable> {*  *Text k = new Text();*  *NullWritable v = NullWritable.get();*  *@Override*  *protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {*  *String line = value.toString();*  *WebLogBean webLogBean = WebLogParser.parser(line);*  *// WebLogBean productWebLog = WebLogParser.parser2(line);*  *// WebLogBean bbsWebLog = WebLogParser.parser3(line);*  *// WebLogBean cuxiaoBean = WebLogParser.parser4(line);*  *if (!webLogBean.isValid())*  *return;*  *k.set(webLogBean.toString());*  *context.write(k, v);*  *// k.set(productWebLog);*  *// context.write(k, v);*  *}*  *}*  *public static void main(String[] args) throws Exception {*    *Configuration conf = new Configuration();*  *Job job = Job.getInstance(conf);*  *job.setJarByClass(WeblogPreProcess.class);*  *job.setMapperClass(WeblogPreProcessMapper.class);*  *job.setOutputKeyClass(Text.class);*  *job.setOutputValueClass(NullWritable.class);*  *FileInputFormat.setInputPaths(job, new Path(args[0]));*  *FileOutputFormat.setOutputPath(job, new Path(args[1]));*  *job.waitForCompletion(true);*    *}*  *}* |

* 运行mr对数据进行预处理

|  |
| --- |
| hadoop jar weblog.jar com.gec.bigdata.hive.mr.WeblogPreProcess /weblog/input /weblog/preout |

## 点击流模型数据梳理

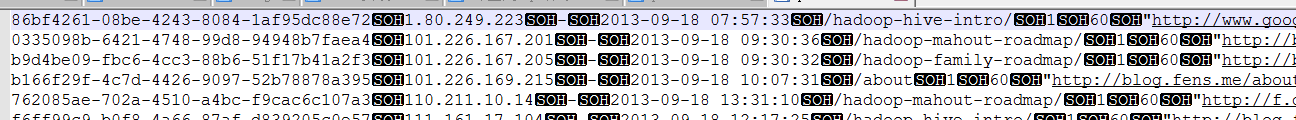
由于大量的指标统计从点击流模型中更容易得出，所以在预处理阶段，可以使用mr程序来生成点击流模型的数据。

### 点击流模型pageviews表

Pageviews表模型数据生成, 详细见：ClickStreamPageView.java



此时程序的输入数据源就是上一步骤我们预处理完的数据。经过此不处理完成之后的数据格式为：



### 点击流模型visit信息表

*注：“一次访问”=“N次连续请求”*

*直接从原始数据中用hql语法得出每个人的“次”访问信息比较困难，可先用mapreduce程序分析原始数据得出“次”信息数据，然后再用hql进行更多维度统计*

用MR程序从pageviews数据中，梳理出每一次visit的起止时间、页面信息

详细代码见工程：ClickStreamVisit.java

