# 新闻类采集的扩展需求

## 1.1采集来源

1、环球网（http://www.huanqiu.com/）：国际国内实时新闻类数据，非评论类，非图片和视频新闻。重点采集的频道（排名分先后）：海外看中国、国际（独家）、国内头条（反腐正风）、台海（滚动新闻）、军事、科技；

2、凤凰网（https://www.ifeng.com/）：国际国内实时新闻类、评论类数据，非图片和视频新闻。重点采集的频道（排名分先后）：资讯中的大陆、国际、台湾、港澳；

3、台湾联合新闻网（<https://udn.com/news/index>，需翻墙）：中国大陆相关的的新闻、评论类数据，非图片和视频新闻。重点采集的频道（排名分先后）：即时中的两岸、要闻、国际，要闻中的总统会六都市长、民进党主席补选，政治，军事；

4、新加坡联合早报（http://www.zaobao.com/），中国大陆相关的的新闻、评论类数据，非图片和视频新闻。重点采集的频道（排名分先后）：即时报道中的中港台即时、中国、国际，新闻中的中港台、国际。

## 1.2采集方式

1、新闻网站的频道实时监控和采集；

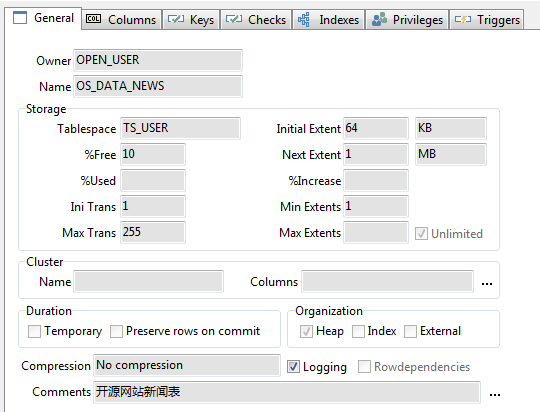
2、新闻文章的浏览量、转载量、评论量实时监控和采集；

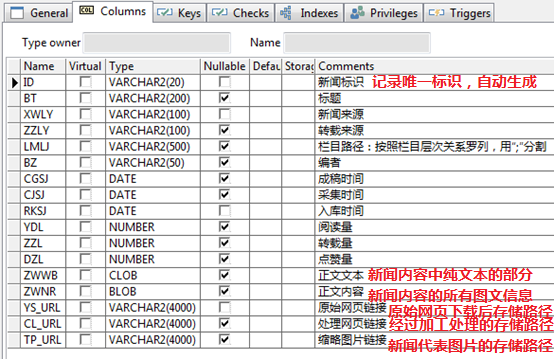
3、2018年6月份以来的历史新闻数据一次性采集，总量不少于3W条。

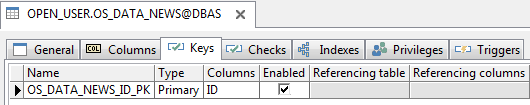
## 1.3加工方式

1、加工后数据持久化存储在Oracle 11g数据库（属性、文本）和FTP服务器（原始网页、处理后的网页和图片）中同时存储在HBase和HDFS中，由于时间要求急迫，先满足Oracle和FTP存储，HBase和HDFS存储要求暂缓。

2、Oracle数据库新闻表结构如下：







3、文章加工内容（以环球网为例）：





# 百科类采集的扩展需求

## 2.1采集来源

1、维基百科（https://www.wikipedia.org/）：政治、经济、外交、军事、社会等领域相关的重要人物介绍，重要法律法规条文介绍，国家介绍，主题事件介绍，专有名词解释等，百科数据内容集中在中美南海问题、朝韩半岛问题、中印洞朗事件相关（规则由十所确定）；

2、百度百科（https://baike.baidu.com/）：政治、经济、外交、军事、社会等领域相关的重要人物介绍，重要法律法规条文介绍，国家介绍，主题事件介绍，专有名词解释等，百科数据内容集中在中美南海问题、朝韩半岛问题、中印洞朗事件相关（规则由十所确定）。

## 2.2采集方式

1、现有的维基百科、百度百科数据按照采集来源要求进行筛选（规则由十所确定）；

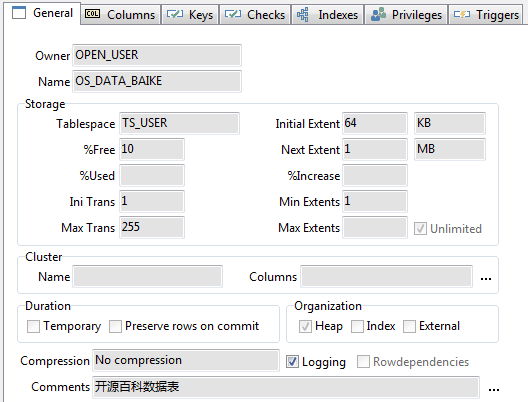
2、按照中美南海问题、朝韩半岛问题、中印洞朗事件需要的百科进行增补（规则由十所确定）；

3、一次性采集，总量不少于20W条。

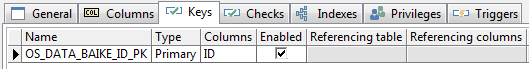
## 2.3加工方式

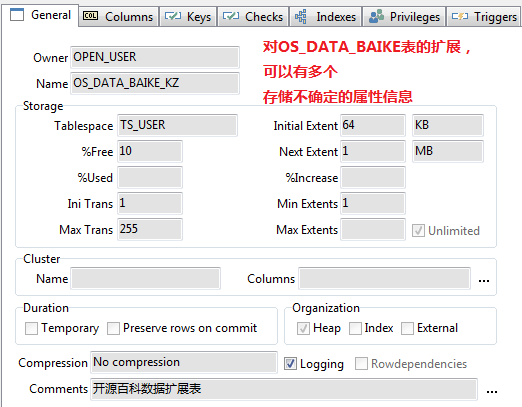
1、加工后数据持久化存储在Oracle 11g数据库（属性、文本）和FTP服务器（原始网页、处理后的网页和图片）中同时存储在HBase和HDFS中，由于时间要求急迫，先满足Oracle和FTP存储，HBase和HDFS存储要求暂缓。

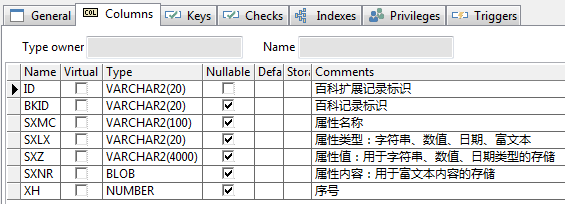
2、Oracle数据库新闻表结构如下：

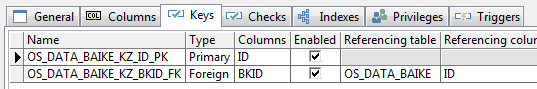












3、百科加工内容（百度百科）：







4、百科加工内容（维基百科）：

