[raw\_note 1](#_Toc497196909)

[raw\_user 2](#_Toc529579086)

[raw\_comments 3](#_Toc1463465234)

[raw\_answers 4](#_Toc1329978747)

[raw\_snapshot 4](#_Toc1943002853)

[raw\_trend 5](#_Toc1412614089)

# raw\_note

普通数据(存储到raw\_note)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type”,⼀般有blog，video，tweet等

url: ⽹⻚URL，默认配置为ref=”\_url”， 有时会调整

fet\_time：⽹⻚爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

title：标题

content：内容

pub\_time：发布时间

image：图⽚，⽀持单个或者多个

uid：⽤户id,对应raw\_user中的idstr,string类型，⼀般要添加action=”tostring”

uname：⽤户名称,对应raw\_user中的name

uurl：⽤户主⻚URL,对应raw\_user中的url

head：⽤户头像, 对应raw\_user中的head

video：视频

vstct：观看数或者阅读数

cmtct：评论数

retct：转发数

likct：点赞数

ulkct：点踩数

aswct：回答数

bmkct：收藏数

thkct: 答谢数

edtct: 编辑数

tag:标签

app：发布设备

# raw\_user

⽤户数据(存储到raw\_user)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型

id: 存储到hbase中rowkey，是我们数据仓库中的唯⼀标识

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type”，得出的结果⼀定是” user ”

url：⽤户URL，⼀般默认配置为ref=”\_url”，有时会调整

fet\_time：爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

head：⽤户头像url

g: ⽤户性别，取值是f,m,n，其中f是⼥，m是男, n是未知

age：⽤户年龄

name：⽤户姓名

birth：⽤户⽣⽇

loc：⽤户地理位置

desc: ⽤户描述

country:国家

province: 省份

city：城市

geo：⽤户geo信息

scho：毕业院校

comp：所属公司

folct：粉丝数

frdct：朋友数

stact：发帖数

likct:被点赞数

retct：被转发或者转帖数

bmkct：收藏数

thkct: 答谢数

edtct: 编辑数

reg\_time:⽤户注册时间

tag：标签

app：发布设备

verify\_type：认证类型。1-个⼈认证, 2-机构认证 3-达⼈ 0-普通⼈

# raw\_comments

评论数据(存储到raw\_comments)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型，⼀般要添加

action=”tostring”

pid\_str:指该评论所属⽬标的idstr，对应raw\_note⾥的idstr，必须保证是string类型，⼀

般要添加action=”tostring”

pid ：指该评论所属⽬标的id,对应raw\_note⾥的id，必须是byte数组类型，⼀般是需要

handler="com.sixestates.template.custom.handler.common.IdHandler"，参考普通数据的

id字段。

id: 存储到hbase中rowkey，是我们数据仓库中的唯⼀标识

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type” ,得出的结果⼀定是” comments ”

url: ⽹⻚URL，⼀般默认配置为ref=”\_url”，有时会调整

fet\_time：⽹⻚爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

title：标题

content：内容

pub\_time：发布时间

image：图⽚，⽀持单个或者多个

uid：⽤户id,对应raw\_user中的idstr,string类型，⼀般要添加action=”tostring”

uname：⽤户名称,对应raw\_user中的name

uurl：⽤户主⻚URL,对应raw\_user中的url

head：⽤户头像, 对应raw\_user中的head

video：视频

vstct：观看数或者阅读数

cmtct：评论数

retct：转发数

likct：点赞数

ulkct：点踩数

aswct：回答数

bmkct：收藏数

thkct: 答谢数

edtct: 编辑数

tag：标签

app：发布设备

# raw\_answers

回答数据(存储到raw\_answers)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型，⼀般要添加

action=”tostring”

pid\_str:指该评论所属⽬标的idstr，对应raw\_note⾥的idstr，必须保证是string类型，⼀

般要添加action=”tostring”

pid：产品或者⽬标id在我们数据库中的唯⼀标识，对应raw\_note⾥的id

id: 存储到hbase中rowkey，是我们数据仓库中的唯⼀标识，必须是byte数组类型。⼀般

由pid + toBytes(idstr)⽣成

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type” ,得出的结果⼀定是” answer ”

url: ⽹⻚URL，⼀般默认配置为ref=”\_url”，有时会调整

fet\_time：⽹⻚爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

title：标题

content：内容

pub\_time：发布时间

image：图⽚，⽀持单个或者多个

uid：⽤户id,对应raw\_user中的idstr,string类型，⼀般要添加action=”tostring”

uname：⽤户名称,对应raw\_user中的name

uurl：⽤户主⻚URL,对应raw\_user中的url

head：⽤户头像, 对应raw\_user中的head

video：视频

vstct：观看数或者阅读数

cmtct：评论数

retct：转发数

likct：点赞数

ulkct：点踩数

aswct：回答数

tag：标签

app：发布设备

# raw\_snapshot

snapshot数据(存储到raw\_snapshot)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型，⼀般要添加

action=”tostring”

id: 存储到hbase中rowkey，是我们数据仓库中的唯⼀标识

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type”,⼀般有blog，video，tweet等

url: ⽹⻚URL，默认配置为ref=”\_url”， 有时会调整

fet\_time：⽹⻚爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

title：标题

content：内容

pub\_time：发布时间

image：图⽚，⽀持单个或者多个

uid：⽤户id,对应raw\_user中的idstr,string类型，⼀般要添加action=”tostring”

uname：⽤户名称,对应raw\_user中的name

uurl：⽤户主⻚URL,对应raw\_user中的url

head：⽤户头像, 对应raw\_user中的head

video：视频

vstct：观看数或者阅读数

cmtct：评论数

retct：转发数

likct：点赞数

ulkct：点踩数

aswct：回答数

bmkct：收藏数

thkct: 答谢数

edtct: 编辑数

tag:标签

app：发布设备

# raw\_trend

trend数据(存储到raw\_trend)

idstr：业务id，数据源内部使⽤的唯⼀标识，必须保证是string类型

id: 存储到hbase中rowkey，是我们数据仓库中的唯⼀标识

source: 数据源名称

type: 数据类型名称，默认配置为ref=”\_type”

url: ⽹⻚URL，默认配置为ref=”\_url”， 有时会调整

fet\_time：⽹⻚爬取时间, 默认配置为ref=”\_fet\_time”

lang：语⾔zh或者en，⼀般配置为value=”zh”或者value=”en”

extra：外部标识字段，默认配置为ref=”\_extra”

title：标题

content：内容

pub\_time：发布时间

image：图⽚，⽀持单个或者多个

uid：⽤户id,对应raw\_user中的idstr,string类型，⼀般要添加action=”tostring”

uname：⽤户名称,对应raw\_user中的name

uurl：⽤户主⻚URL,对应raw\_user中的url

head：⽤户头像, 对应raw\_user中的head

video：视频

vstct：观看数或者阅读数

cmtct：评论数

retct：转发数

likct：点赞数

ulkct：点踩数

aswct：回答数

bmkct：收藏数

thkct: 答谢数

edtct: 编辑数

tag:标签

app：发布设备