QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2018

《爱奇艺十亿全网视频仓库建设》

帅伟良

主办方 Geekbang. 极客邦科技 InfoQ

目录

- 项目背景
- 视频仓库实战
  - 长视频仓库建设
  - 短视频仓库建设
- 未来的路

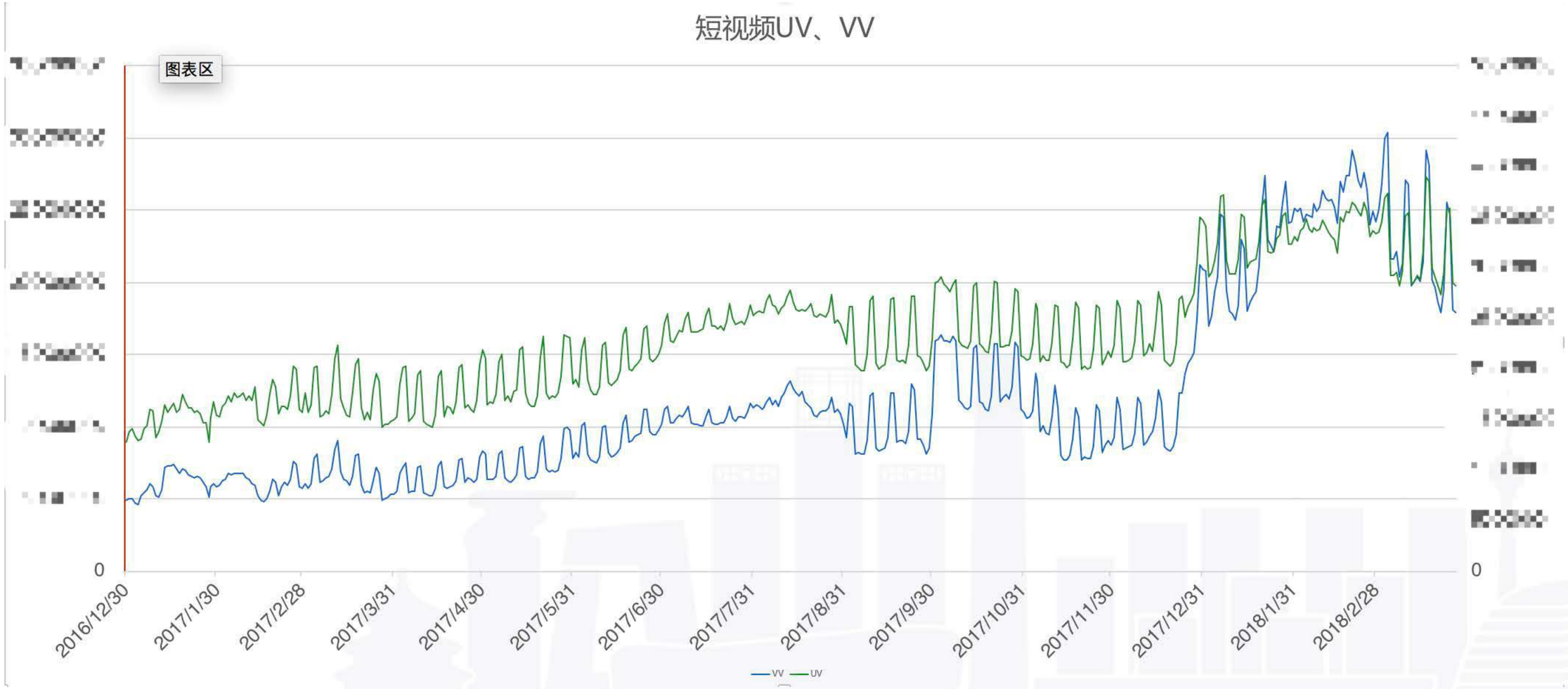
项目背景

长视频版权大战


搜索" 南方有乔木 ",共找到约 5.6万 个视频
相关 最新 最热 高级筛选
VIP
南方有乔木
更新至2集(共40集)
南方有乔木 2018
7.6
周六至周四每天两集，周五一集
导演: 林妍 主演: 陈伟霆 白百何
简介: 清秀却寡淡的南家三小姐南乔在儿时好友常剑雄的设计下撞见未婚夫周然出轨，毅然与其分手。周然威胁撤资，为使自己钟爱的科技事业平稳推进，南乔去了酒吧面见投资人，却误打误...
详细>
1 2 3 4 5 6 7 8 9 10 11 12
爱奇艺
腾讯
优酷
搜狐
VIP
南方有乔木
更新至2集(共40集)
南方有乔木DVD版 2018
周日至周五每天两集，周六一集
导演: 林妍 主演: 陈伟霆 白百何
简介: 清秀却寡淡的南家三小姐南乔在儿时好友常剑雄的设计下撞见未婚夫周然出轨，毅
然威胁撤资，为使自己钟爱的科技事业平稳推进，南乔去了酒吧面见投资人，却误打误...
1 2 3 4
爱奇艺


搜索" 这就是街舞 ",共找到约 37.0万 个视频
相关 最新 最热 高级筛选
这！就是街舞第1季 2018
每周六20:00
独播
别名: 这就是街舞
主持: 易烊千玺
简介: 《这！就是街舞》是一档大型街舞真人秀节目。节目选用"明星导师+专业舞者真人秀"的全新赛制颠覆所有传统舞蹈节目模式，通过"大海选"吸纳优秀街舞舞者，设置"舞者近身斗舞...
详细>
2018-03-24: 第5期：疯狂24小时齐舞燃炸天
2018-03-17: 第4期：49强突围，队长花式"护...
2018-03-10: 第3期：百强"互杀"队长不忍直视
2018-03-03: 第2期：队长"变卦"选手情绪失控
2018-02-28: 顶配版 易烊千玺严中选优 黄子...
2018-02-24: 第1期：队长崩溃！这海选太难了
优酷
热血街舞团 2018
别名: 街舞
简介: 《热血街舞团》是由爱奇艺倾力打造的一档热血系街舞竞技剧情真人秀。全国顶尖舞团踏上厂牌荣誉，在两组引领潮流与话题的顶级明星带领下，打破舞林格局，于生死未卜的激烈赛制...
详细>
2018-03-24: 鹿晗王嘉尔寻最饿舞者 陈伟霆...
2018-03-17: 召集人公演炸屏来袭 热血之城...
安装爱奇艺客户端，立即下载此片
爱奇艺
这！就是街舞会员顶配版 第一季 2018
简介: 《这！就是街舞》是一档大型街舞真人秀节目。节目选用"明星导师+专业舞者真人秀"的全新赛制颠覆所有传统舞蹈节目模式，通过"大海选"吸纳优秀街舞舞者，设置"舞者近身斗舞...
详细>
2018-03-20: 黄子韬补刀选手年龄 罗志祥自...
2018-03-14: 易烊千玺哥哥力爆棚 黄子韬赞1...
2018-03-07: 黄子韬加赛复活太纠结 易烊千...
2018-02-28: 易烊千玺严中选优 黄子韬激将...
优酷
独播

短视频消费量爆发


短视频UV、VV
图表区
0
2016/12/30
2017/1/30
2017/2/28
2017/3/31
2017/4/30
2017/5/31
2017/6/30
2017/7/31
2017/8/31
2017/9/30
2017/10/31
2017/11/30
2017/12/31
2018/1/31
2018/2/28
VV
UV

视频仓库主要任务

- 全网版权剧仓库构建
  - **覆盖主流版权剧网站**
  - 聚合相同版权内容
- 全网短视频仓库构建
  - **覆盖主流短视频网站**
  - **数据量大**

视频仓库实战

全网长视频采集

南方有乔木 TV版_高清视频在线观看_爱奇艺

更新至 15集/共40集　网络更新
2018 | 内地 | 国语　VIP每日24
主演：陈伟霆 白百何 李现 白冰 张若
类型：言情 | 偶像
简介：该剧述出身优越，研发无人飞行器的
与她身份背景相差巨大的男人时樾，
改变的故事[。 更多>>

爱奇艺 搜狐 腾讯 优酷

15集 新 | 14集 | 13集 | 12集 | 11集 | 10集
7集 | 6集 | 5集 | 4集 | 3集 | 2集

www.iqiyi.com -

腾讯视频 首页 频道　我是大侦探 热搜榜 全网搜　看过 看单 上传 下载客户端

YOUKU 导视　首页 发现 订阅 会员 我的　登录 注册　搜库

搜狐视频 首页 导航　南方有乔木 女超人 哥谭镇　动物系恋人啊　登录 注册 记录

南方有乔木　7.1 | 56人评分　播放数：872万

主演：白百何 /陈伟霆 /秦海璐 /李现 /白冰 /张宥浩　导演：林妍
类型：言情剧 /都市剧　年份：2018

简介：清秀却寡淡的南家三小姐南乔在儿时好友常剑雄的设计下撞见未婚夫周然出轨，毅然与其分手。周然威胁撤资，为使自己钟爱的科技事业平稳推进，南乔去了酒吧面见投资人，却误打误撞与高大冷峻的酒吧老板时樾相识，时樾意外发现南乔似乎与自己早已尘封的一段过往有着扑朔迷离的关系，他有意接近南乔，原想布下情感陷阱彻查当年那段往事，不想自... 展开

现在观看　追剧提醒　分享到：

概览 | 片花& 预告 | 分集剧情 | 系列 | 评论

共40集，更新到15集，会员24点同步卫视，非会员次日24点观看

1-15

1	2	3	4	5	6	7	8	9	10
11	12	13	14 预	15 预	14 会员	15 会员			

片花& 预告　更多>>

演职员表


白百何
南乔


陈伟霆
时樾


秦海璐
安宁


李现
常剑雄


白冰
温笛


张宥浩
郝杰

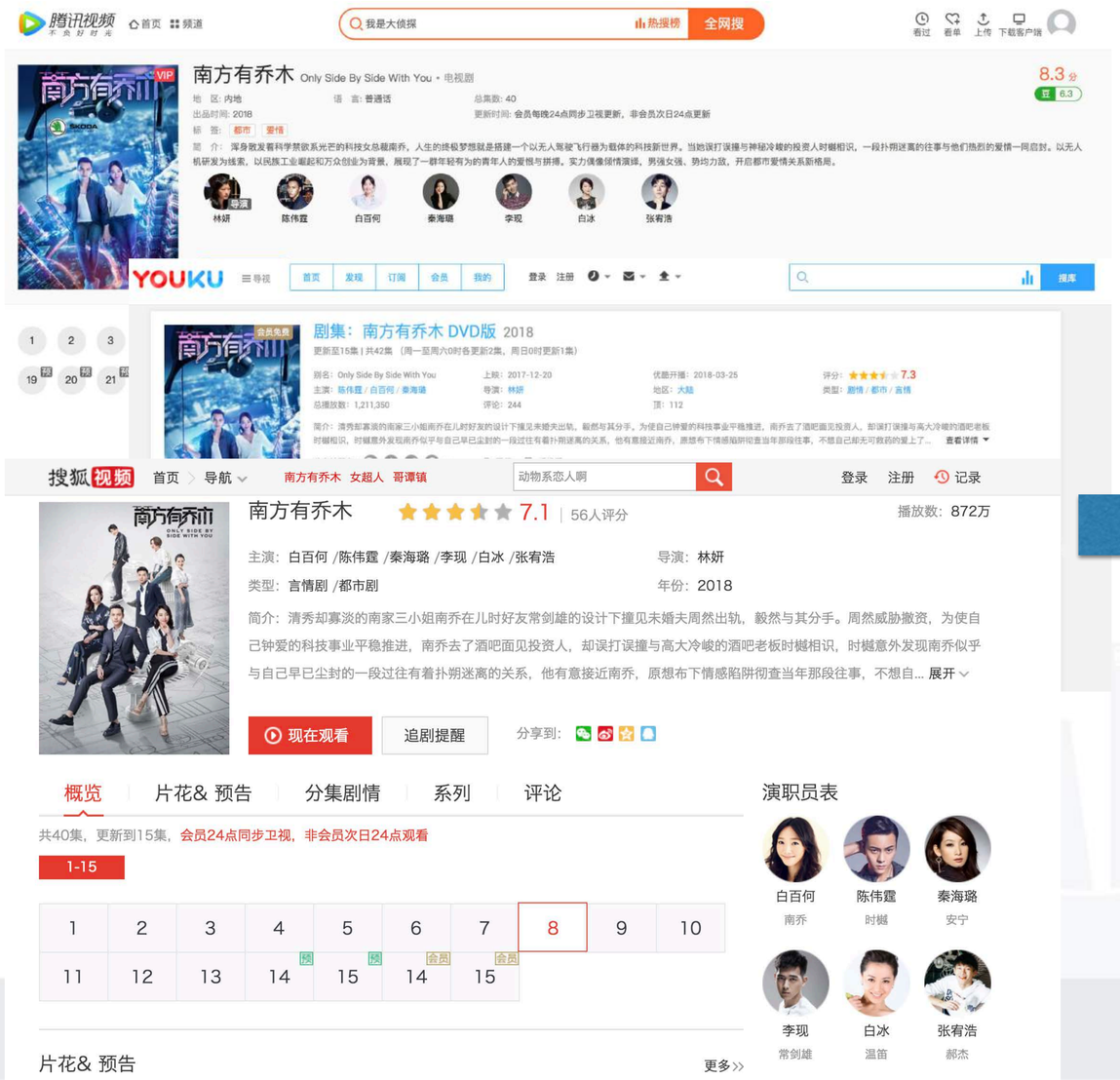
全网长视频采集

完成方式	优点	缺点	上线时间
阶段1： 抓**搜索聚合页**	1. 开发**简单** 2. 去重**简单**	1. 更新频率依赖聚合站点 2. 站点源依赖聚合站点 **3. 永远都是跟随者**	到2015年中，我们一直通过抓取聚合页完成
阶段2： 抓**各网站详情页**	1. 可**灵活配置** 2. 站点源**不受限制**	**1. 工作量大** **2. 维护成本高**	2015年中，我们完成了第一版**云爬虫**的开发，对版权剧的采集进行迁移

全网长视频采集

我们的站点效果已经超过友商

全网长视频去重



全网长视频去重

找特征 → 定公式 → 升效率

标题
描述
主演
导演
主持人
频道
发行日期

相似度计算公式：
启发式
机器学习

快速收敛候选集大小

找特征

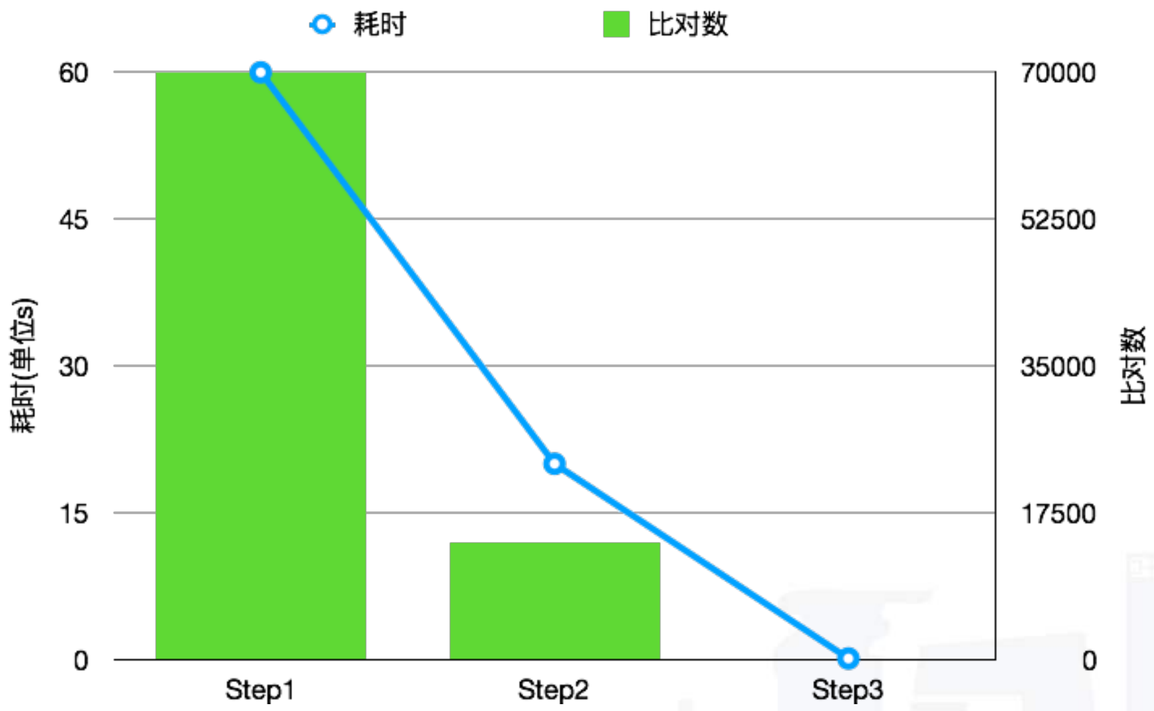
标题	英文名	别名	描述	频道	地区	语言	总集数	发行时间	导演	演员
南方有乔木	Only side by side with you		浑身散发着科学禁欲…	电视剧	内地	普通话	40	2018	林妍	陈伟霆 白百合 秦海璐 李现

定公式

特征类型	特征	样例	
决定性特征	季、期	第一季、第二季	$f(a,b)=\begin{cases}1 & a=b\\0 & a\neq b\end{cases}$
	语言	粤语、普通话、英语	
	版本	OVA、剧情版	
接受一定偏差特征	发布时间	2018	$e^{e(-0.1 * distanceYear) - 1}$
一般特征	地区	内地、香港、台湾、美国	$Jaccard(A,B)=\frac{A\cap B}{A\cup B}$
	导演	王晶、周星驰	
	演员	刘德华	
	总集数	40	
	描述		
	标签	搞笑、经典	

$$Sim(a, b) = Dete(a, b) * Perti(a, b) * \sum Wj*P(a, b)$$

升效率



70,000

- 加载全部meta到内存，线性比对全部专辑，最坏比对140000次，平均耗时 1 min

14,000

- Meta按照频道分桶（电影、电视剧、综艺、动漫、其它），最坏需要比对桶中最大数，耗时平均降低到 20s

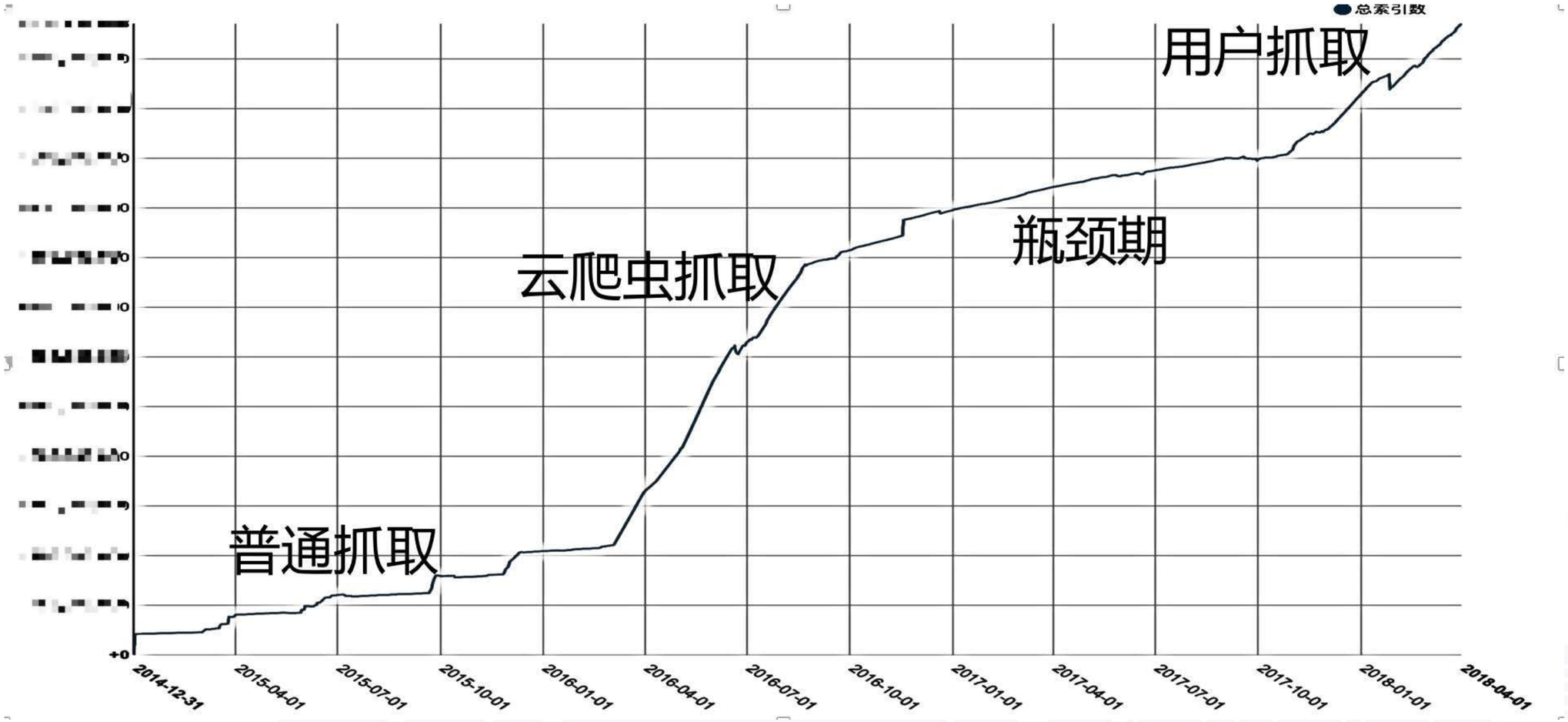
10

- 利用meta关键词建立倒排，按照相关性初排序，选举出相关性最高的10个meta，从而达到近实时的效果

全网短视频采集


全网影视
V.iQIYI.COM
初音未来的视频
搜全网
00:26
简介：《王者荣耀》是腾讯天美工作室历时3年推出的东方英雄即时对战手游大作,抗塔强杀...
05:39
马里奥大叔VS初音未来，马里奥尽然攻击公主
发布时间：2017-09-25
上传者：空梦_1918
简介：转载优酷http://v.youku.com/v_show/id_XMTM4...
01:32
初音未来歌姬计划X op
发布时间：2017-01-10
上传者：pudashoes_2633
04:08
【百合向MMD】charActer feat. 初音ミク、巡音ルカ
发布时间：2018-01-26
来源：bilibili
简介：简介：那个...你们的小公主，被女王吃掉了（这不挺好的嘛？）姬姬的感觉真棒...
04:55
VOCALOID初音ミク《BLACK ★ ROCK SHOOTER》
发布时间：2017-03-31
来源：腾讯
02:30
初音未来 - Ievan Polkka (甩葱歌)[高清版]
发布时间：2013-10-08
上传者：木力
简介：闲了累了，听听看看

全网短视频采集


总索引数
用户抓取
瓶颈期
云爬虫抓取
普通抓取
+0
2014-12-31
2015-04-01
2015-07-01
2015-10-01
2016-01-01
2016-04-01
2016-07-01
2016-10-01
2017-01-01
2017-04-01
2017-07-01
2017-10-01
2018-01-01
2018-04-01

挑战

- 下载
  - 站点多
  - 防爬虫问题
  - 下载量大
  - 分布式QPS控制
- 数据全覆盖
- 数据量大

解决下载问题

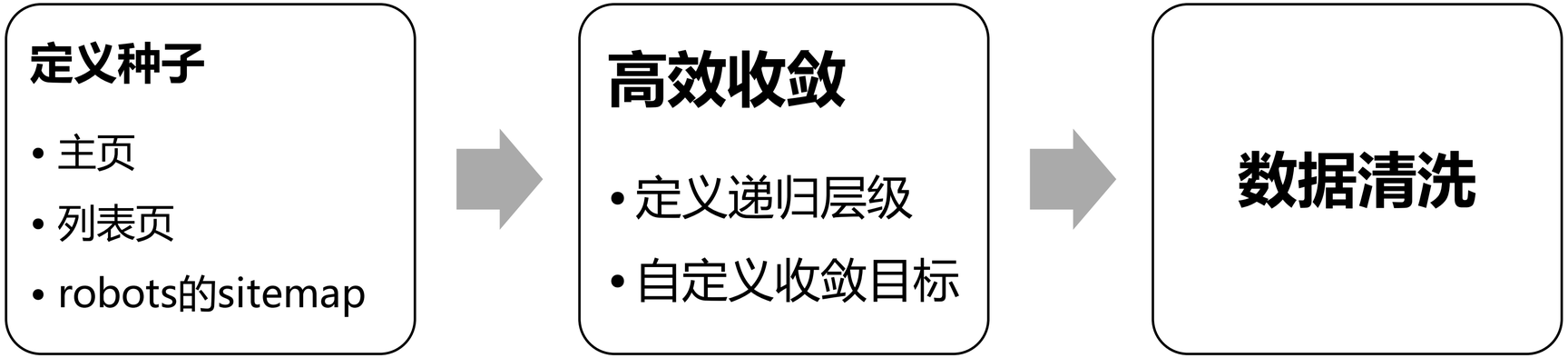
站点多

防爬虫问题

下载量大

分布式QPS控制

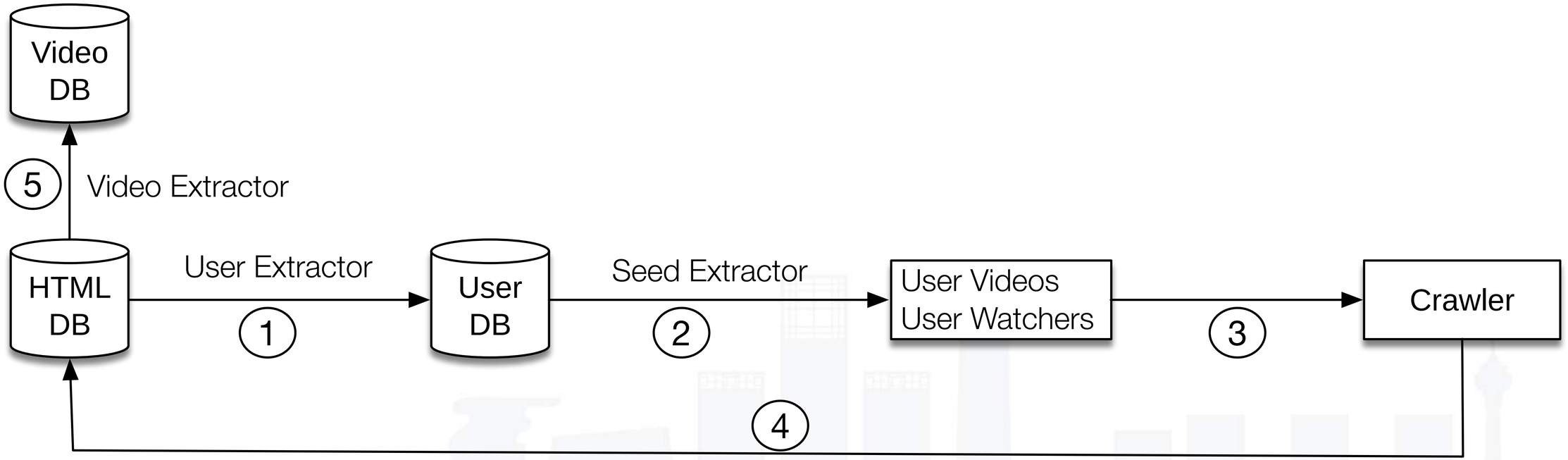
解决全覆盖问题



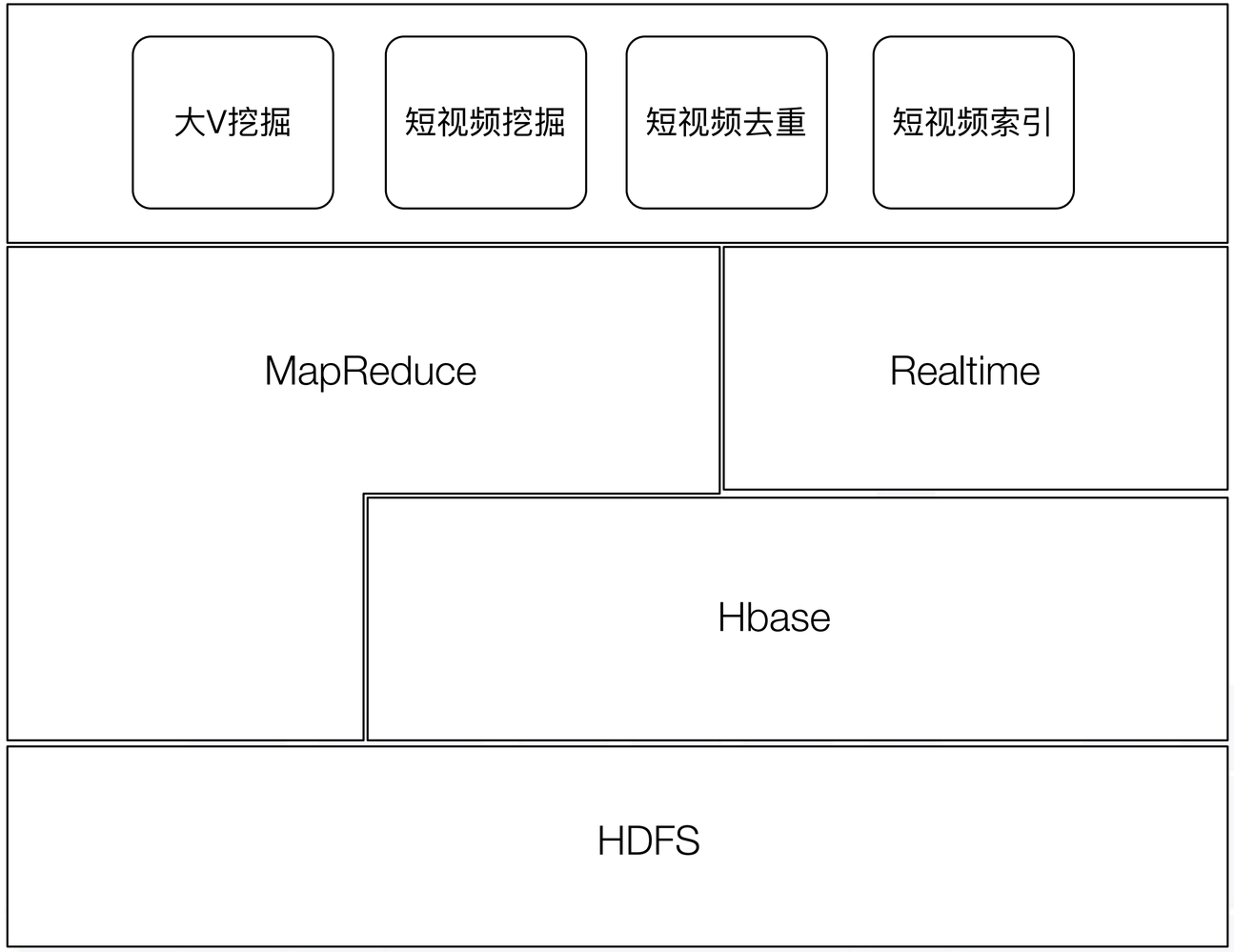
传统爬虫问题：

1. 整站抓取，收敛太慢
2. 自定义收敛，数据抓取不全
3. 信息流推荐场景无法整站抓取

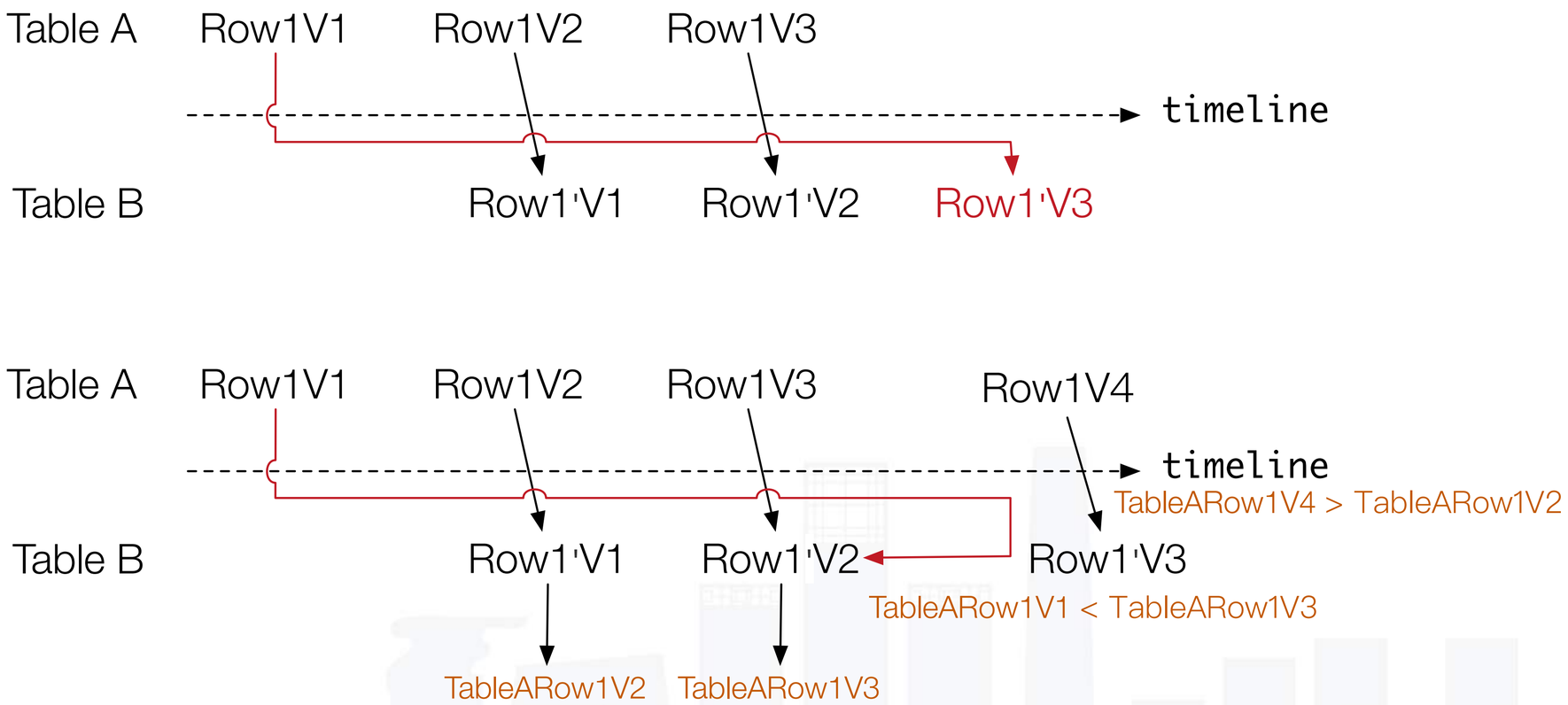
基于up主的抓取


Video
DB
5
Video Extractor
HTML
DB
User Extractor
1
User
DB
Seed Extractor
2
User Videos
User Watchers
3
Crawler
4

解决数据量大问题



全量覆盖实时问题解决



- Version选择Cell的**timestamp**
- 比较使用CAS **checkAndPut**

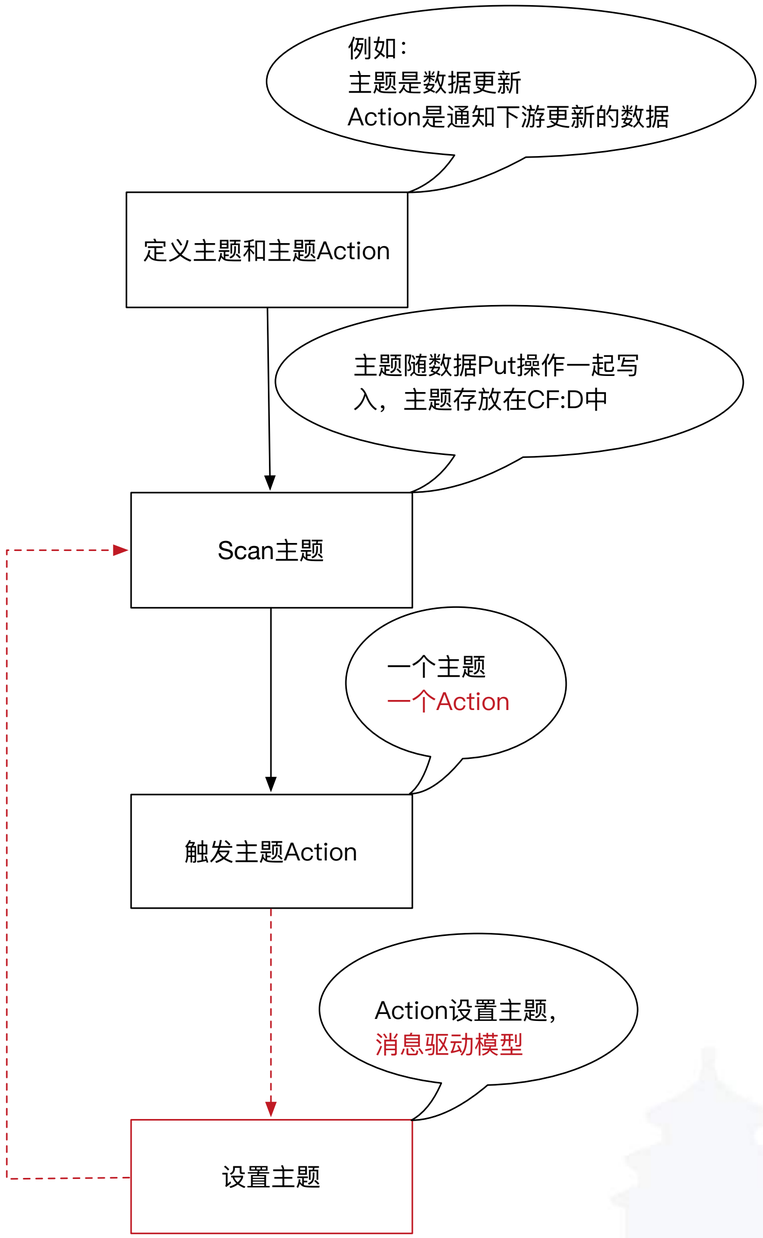
实时更新原理

Row Key	Column family : C	Column family : **D**
youku|url1	column **d**, column **s**	column update
youku|url2	column **d**, column **s**	column update
qq|url1	column **d**, column **s**	column update
qq|url2	column **d**, column **s**	column update

Site prefix for site scan

d for normal data
s for statistic data

Column Family D is **in memory**, and the data is very **small** to mark the state of the row for good performance



实时更新原理

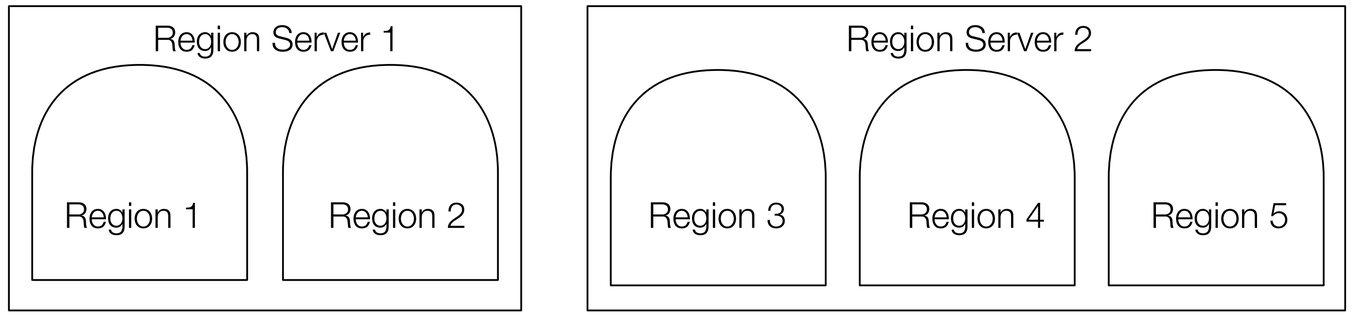
Row Key	Column family：C	Column family：**D**
youku|url1	column **d**, column **s**	column update
youku|url2	column **d**, column **s**	column update
qq|url1	column **d**, column **s**	column update
qq|url2	column **d**, column **s**	column update

- 相比于写入数据后发送一个消息到MQ

此方案充分利用Hbase put操作在某行的完整性，解决单行事务的问题

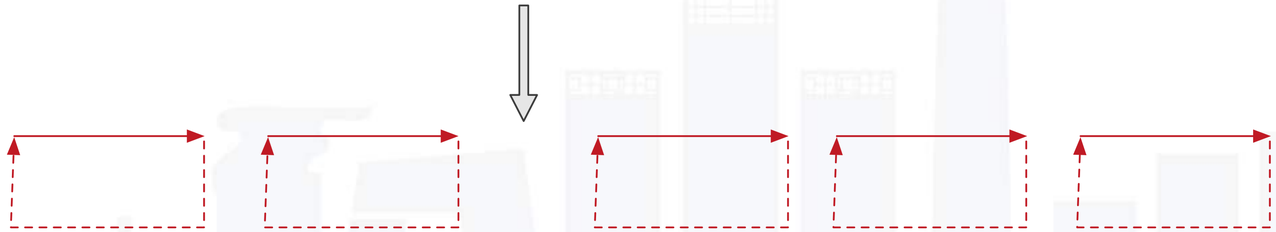
问题：数据量大，Region变得非常多，Region server的不稳定，如何减少这种不稳定对系统的影响

实时更新原理



顺序Scan

Region Scan



假设region的异常概率是p，前者的概率$n * p$，后者p^n

实时更新-后续的路

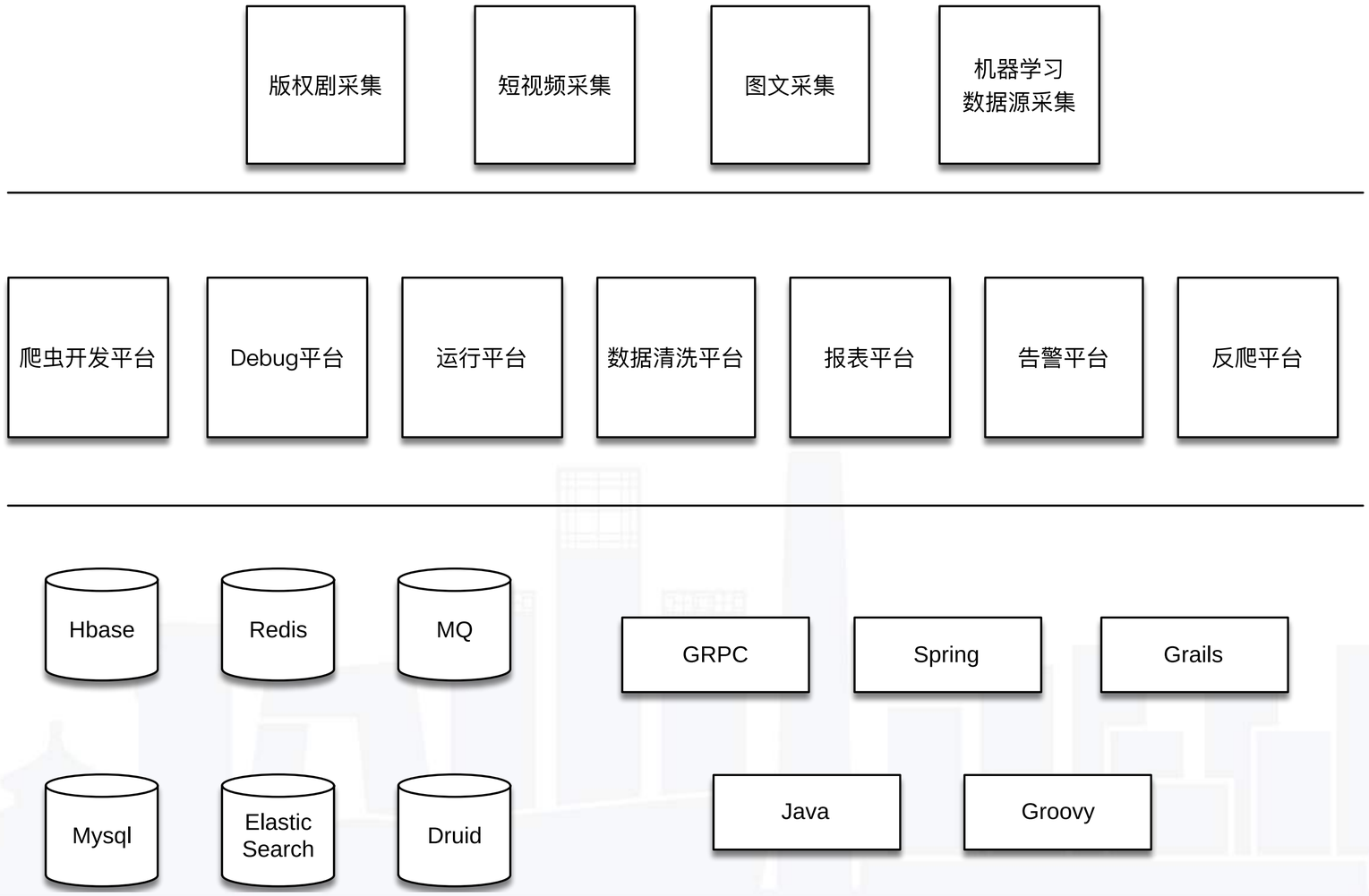
- 高可用建设
- 平台化，业务只需要关心定义主题和主题绑定的Action，分布式执行和性能调优交给平台

视频仓库实战总结

通过中间件快速实现业务，业务发展又不断推进中间件功能升级

通过开源，快速构建新的中间件

感谢和致敬开源



未来的路

数据挖掘

- 竞品分析
- 优质内容挖掘
- 辅助内容生态构建
- 机器学习、深度学习特征库

关注QCon微信公众号，
获得更多干货！

Thanks!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

主办方 Geekbang. 极客邦科技 InfoQ