

SDCC 2017 | 上海

互联网应用架构实战峰会

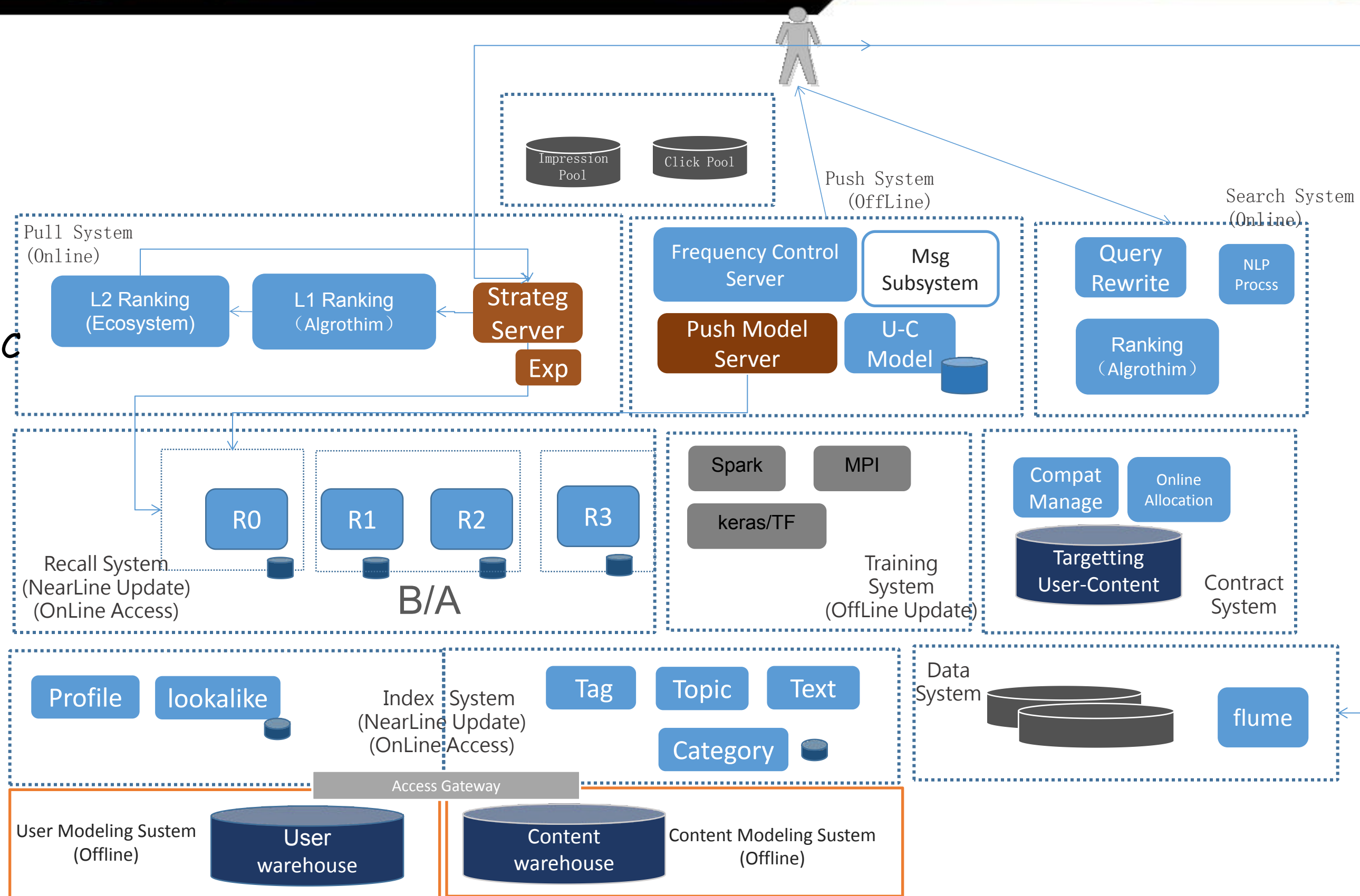
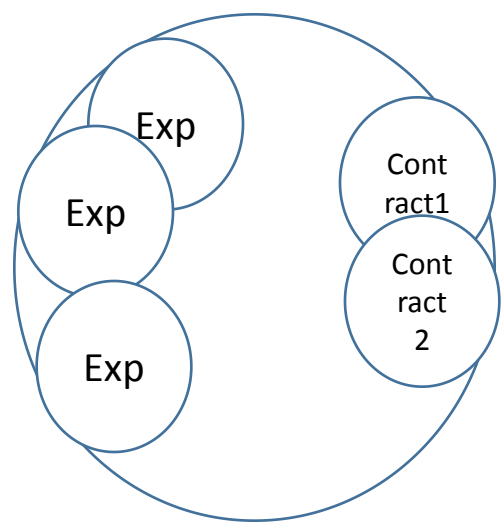
CSDN

Building Content Recommendation System

based on Content modeling

**Tencent Zixunsun
腾讯 孙子荀**

- **Architecture Overview**
 - *Content Side*
 - *User Side*
- **User Modeling**
- **Content Modeling**
- **Content Processing Architecture**



1. $U \rightarrow \text{Model}(U) ? \text{Model}(C) \rightarrow C$
Silarity

2. $U \rightarrow C \rightarrow \text{Model}(C) \rightarrow C$
Cluster

3. $U \Rightarrow U \rightarrow C$
UserCF
 $U \rightarrow C \Rightarrow C$
ItemCF

4. $U \rightarrow \text{Model}(U-C) \rightarrow R(C)$
:FM

Content Description

Data type: Text , URL, Video Refer

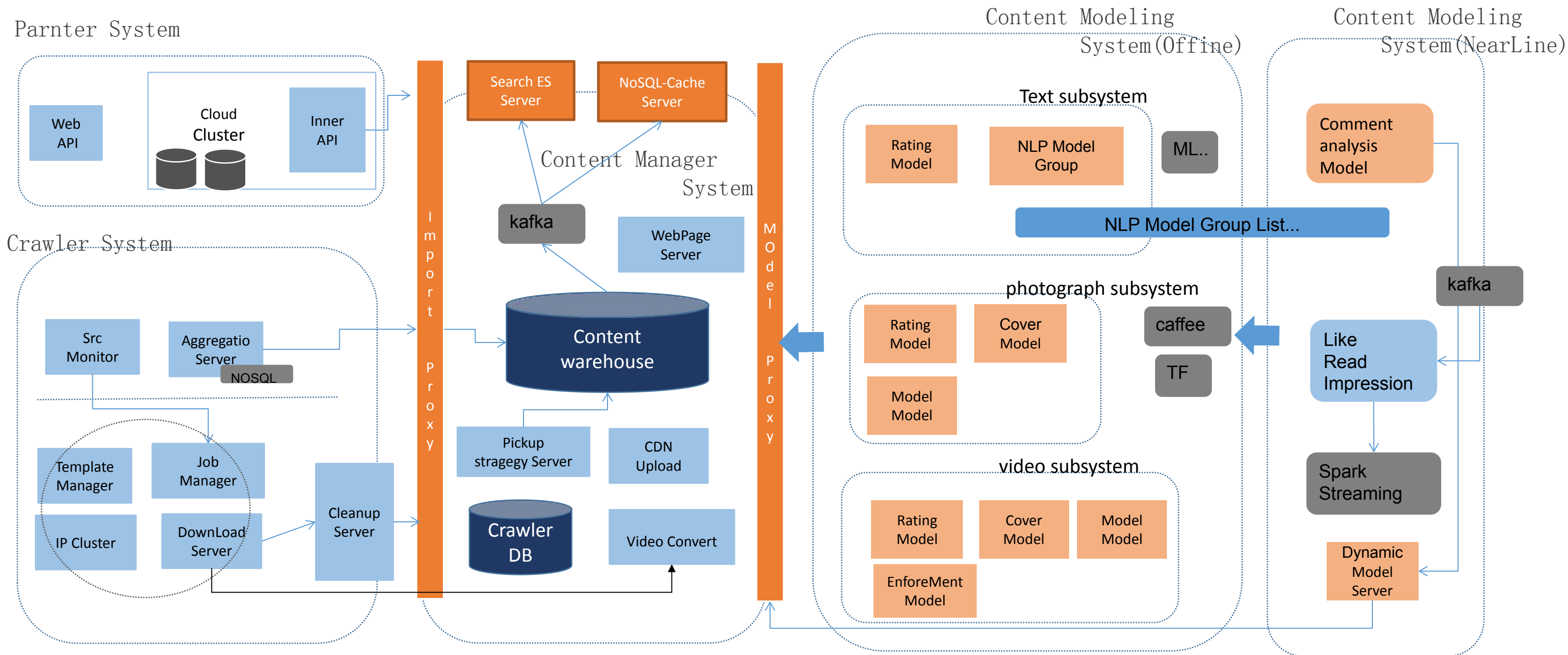
Control Type : Life Cycle Manager

Model Type: Semantic analytics ,Mutile-Classification

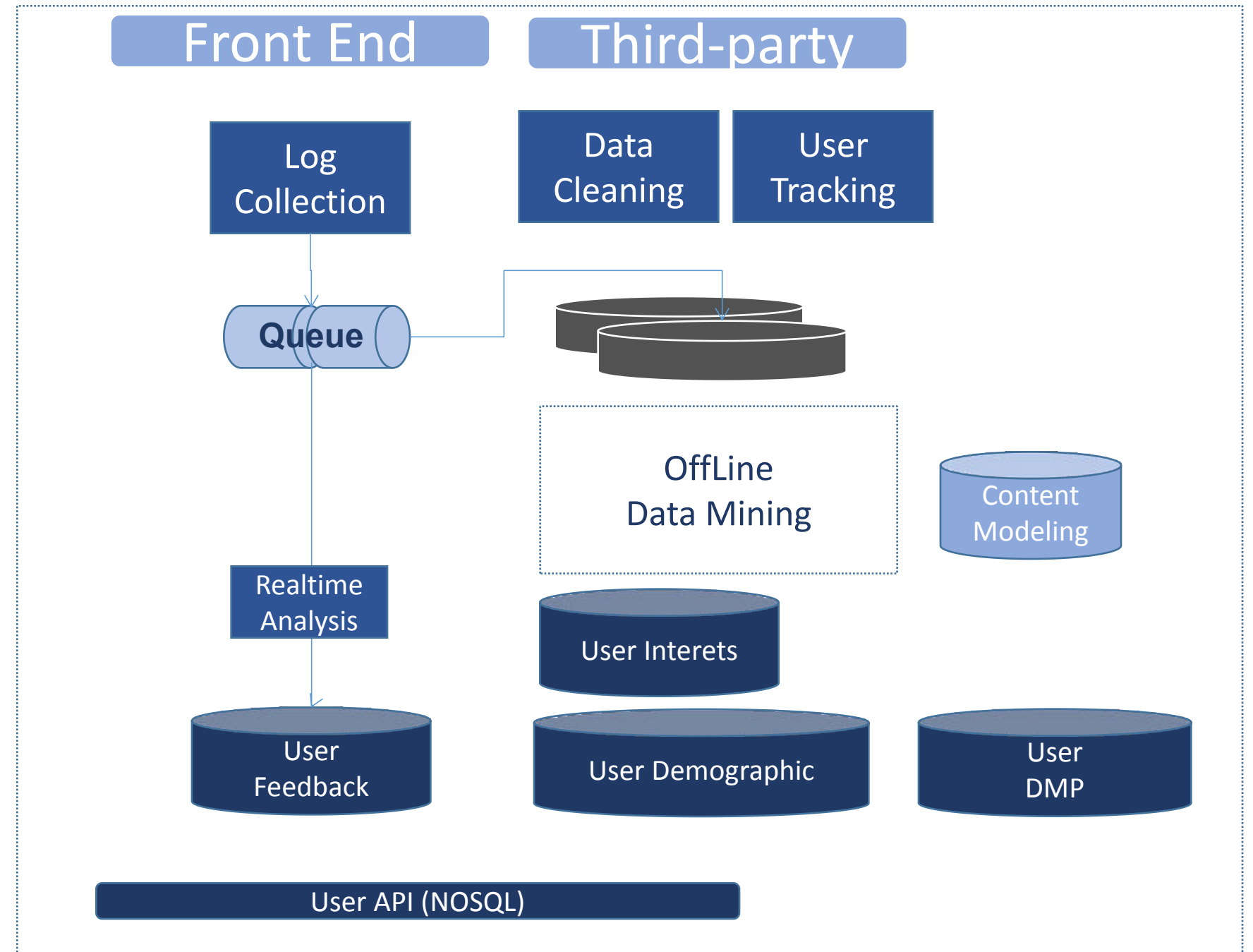
Photo Type : OCR , Enhance, Obj Tag , Rating, Cut...

Evaluation Type : Content Rating ,Media Rating

hot_pic_url
hot_source
hot_topic
insearch_push
is_novel
kv_storage
low_quality_type
manual_chann
manual_sec_chann
manual_tag
media_id
media_name
mid
multi_cover_score
multi_cover_sz
multi_cover_url
need_push_kd
npic
ocr_msg
org_url
pic
pic_file_sz
pic_illegal
pic_max_hot
pic_max_porn
pic_qrcode
pic_score
pic_sz
pic_type
priority
purl
purl_name
raw_cover_score
raw_pic
raw_pic_score
score
sexy_score
share_pic_url
simkey
single_cover_score
single_cover_sz
single_cover_url
src_url
st_Complex
st_TagHot
st_TagNormal
st_TagPron
st_article
st_auth
st_disu
st_edit_flag
st_edit_st
st_insearch
st_is_novel
st_market
st_ocr
st_page_url
st_pdfif
st_pic
st_politic
st_push
st_quality
st_report
st_score
st_sexy
st_share_url
st_tag
st_top
st_txt
st_txt_cover
st_use
st_xinan
summary
t_sh
t_st
teg_chann
teg_disu
teg_mingxing_fm
teg_sec_chann
teg_sec_topic
teg_sex
teg_tag
teg_tag_flag
teg_tag_score
teg_term
teg_term_flag
teg_term_score
teg_title_tag
teg_title_tag_flag
teg_title_tag_score
teg_topic
tips
ts
vid



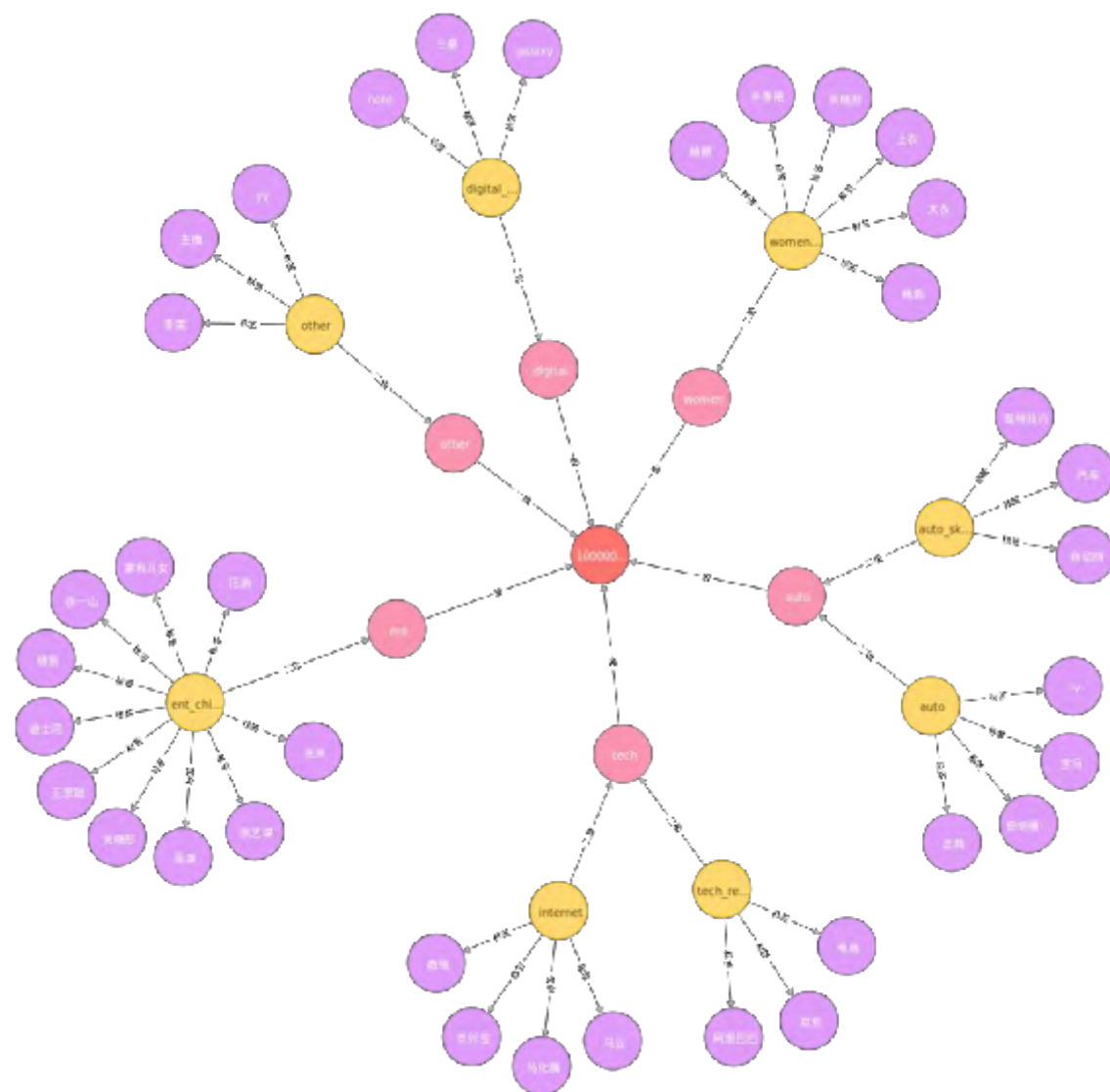
	TimeLiness	Storage	latency	Data Src
Feedback	Real-Time	rockdb	local (2ms)	behavior
Interets	Near Real-Time	redis	remote (10ms)	behavior mining
Demography	Non-Real Time	redis	remote (10ms)	mutil src mining



Traditional user **Interests** profile building from content tag

"As is the Content, As is the User"

Tag is never Enough .
User TAG - > User Knowledge



User KnowLedge Building From
1 Content Modeling
2 User Behavier

"How deeply know the Content,
How strongly know the User"

Content Feature :

Entity: Singer Wangxxx
Category: Entertainment
Desc : Divorce
Emotion Analyses : Negative

Reading completion rate
20%
Comment : Negative



UserA
<Girl, 16>

Content Feature :

Entity: Singer Wangxxx
Category: Entertainment
Desc : Marriage
Emotion Analyses : Neutral

Reading completion rate
80%

Content Feature :

Entity: NBA Player
Category: Sports
Desc : Game Report
Emotion Analyses : Neutral

Nerver Click

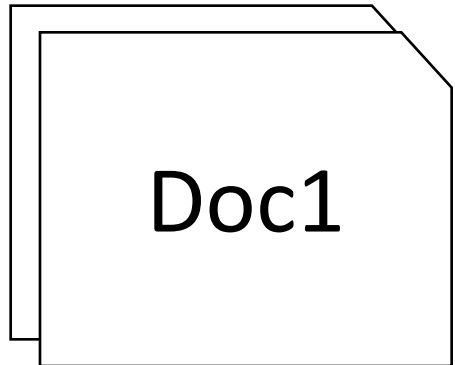


UserB
<Man, 45>

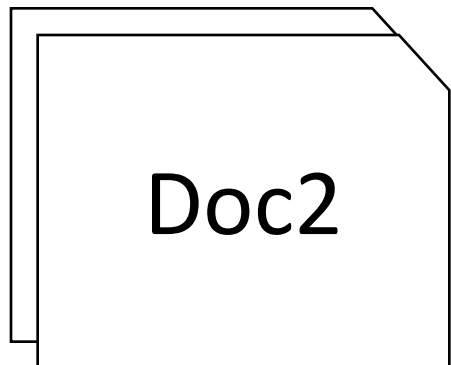
Content Feature :

Entity: NBA Player
Category: Sports
Desc : Gossip
Emotion Analyses : Neutral

Reading completion rate
90%



Doc1

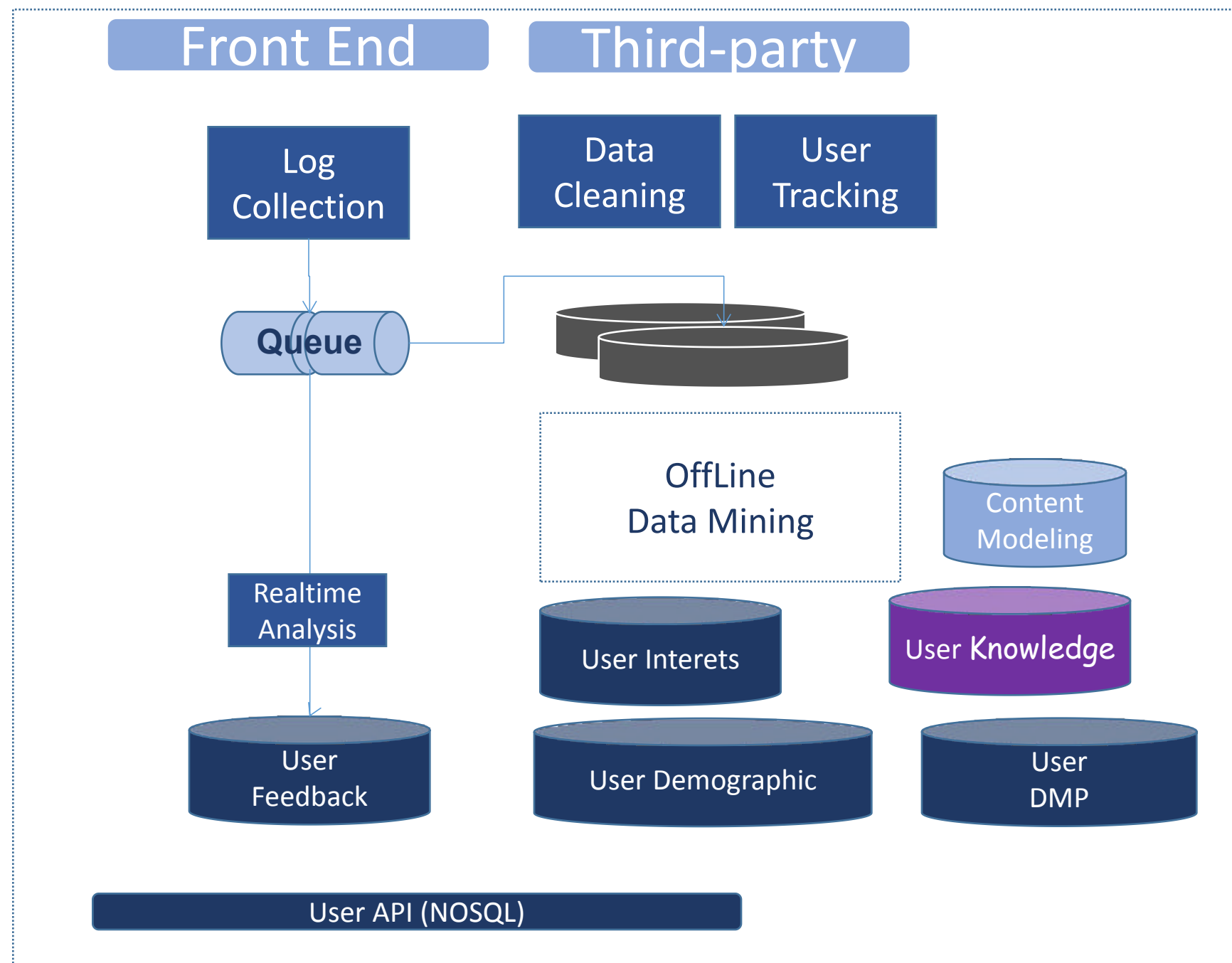


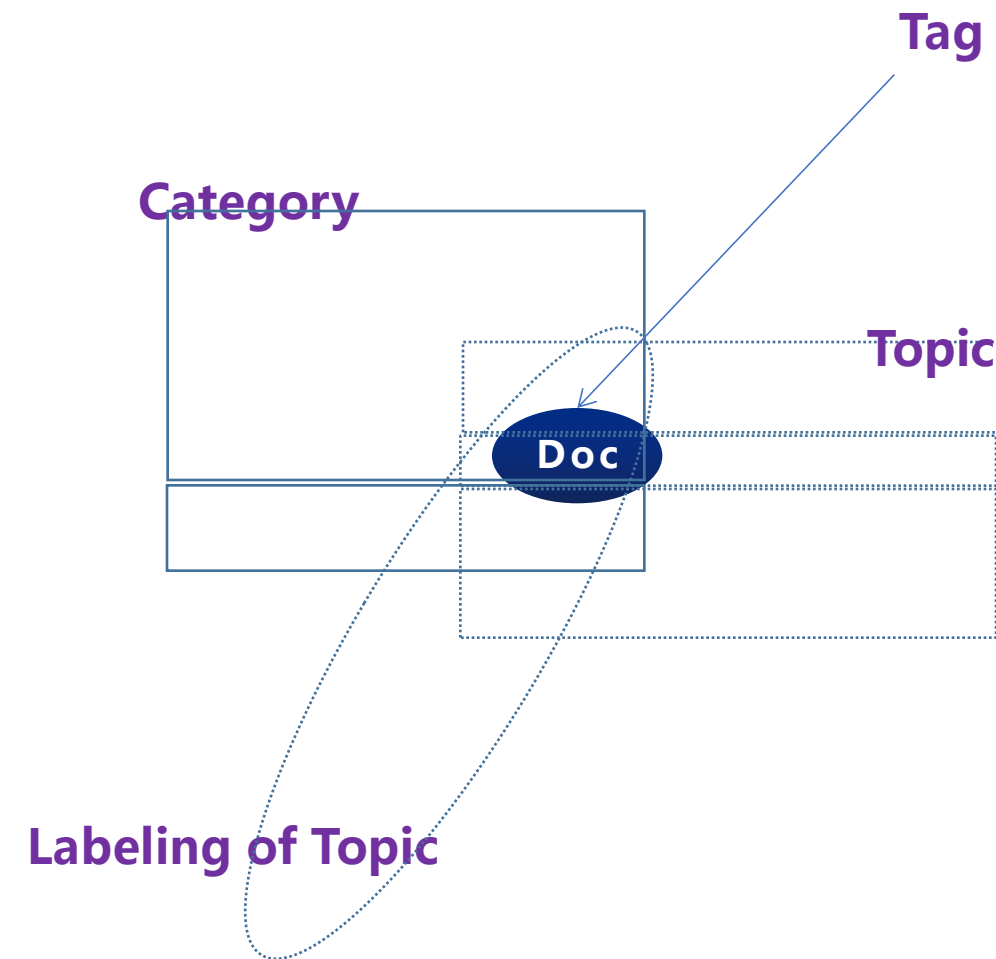
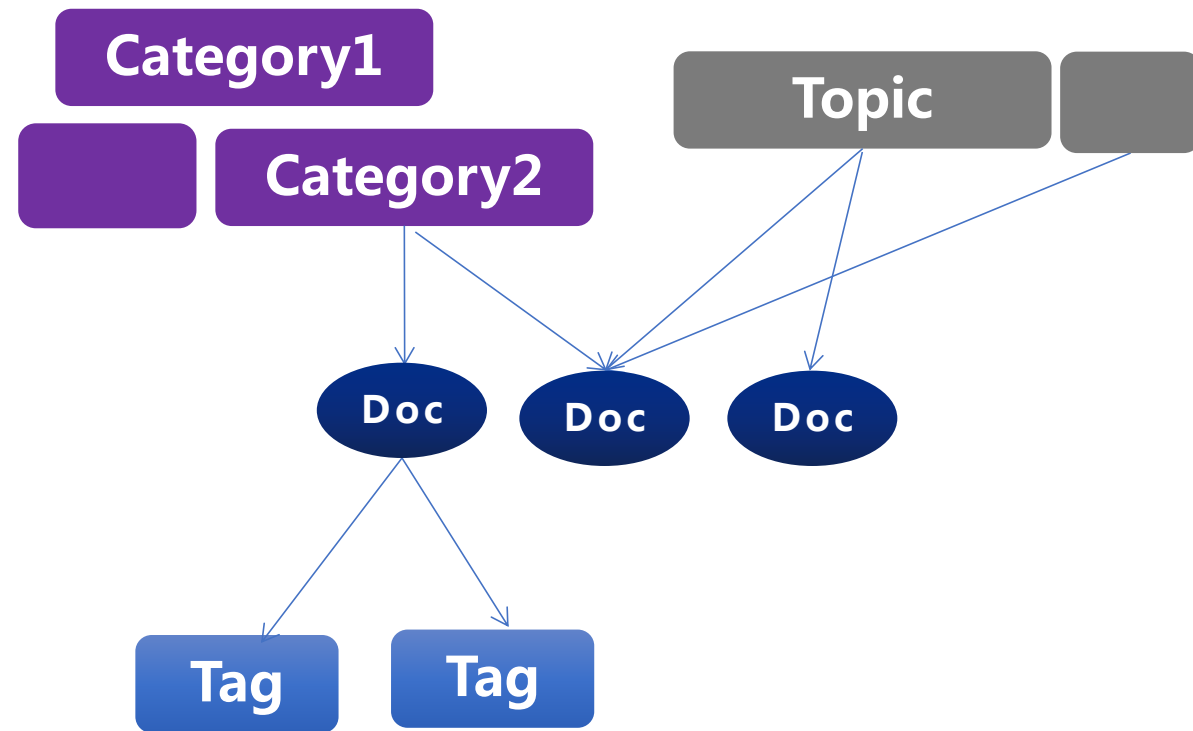
Doc2


```

{
  {"_id": "xxx"},
  {"M": "StarGossip", "S": "1"},
  {"Cond": [
    {"tag": "Wangxxx/"},
    {"category": "ent/sports"},
    {"topic": "123/221"}
  ]}
}
    
```

	TimeLiness	Storage	latency	Data Src
Feedback	Real-Time	rockdb	local (2ms)	behavier
Interets	Near Real-Time	redis	remote (10ms)	behavier mining
Demography	Non-Real Time	redis	remote (10ms)	mutil src mining
Knowledge	Near Real-Time	Mongodb	remote (20ms)	Content + User Modeling





Category : Client eg"navigation menu" , Feature

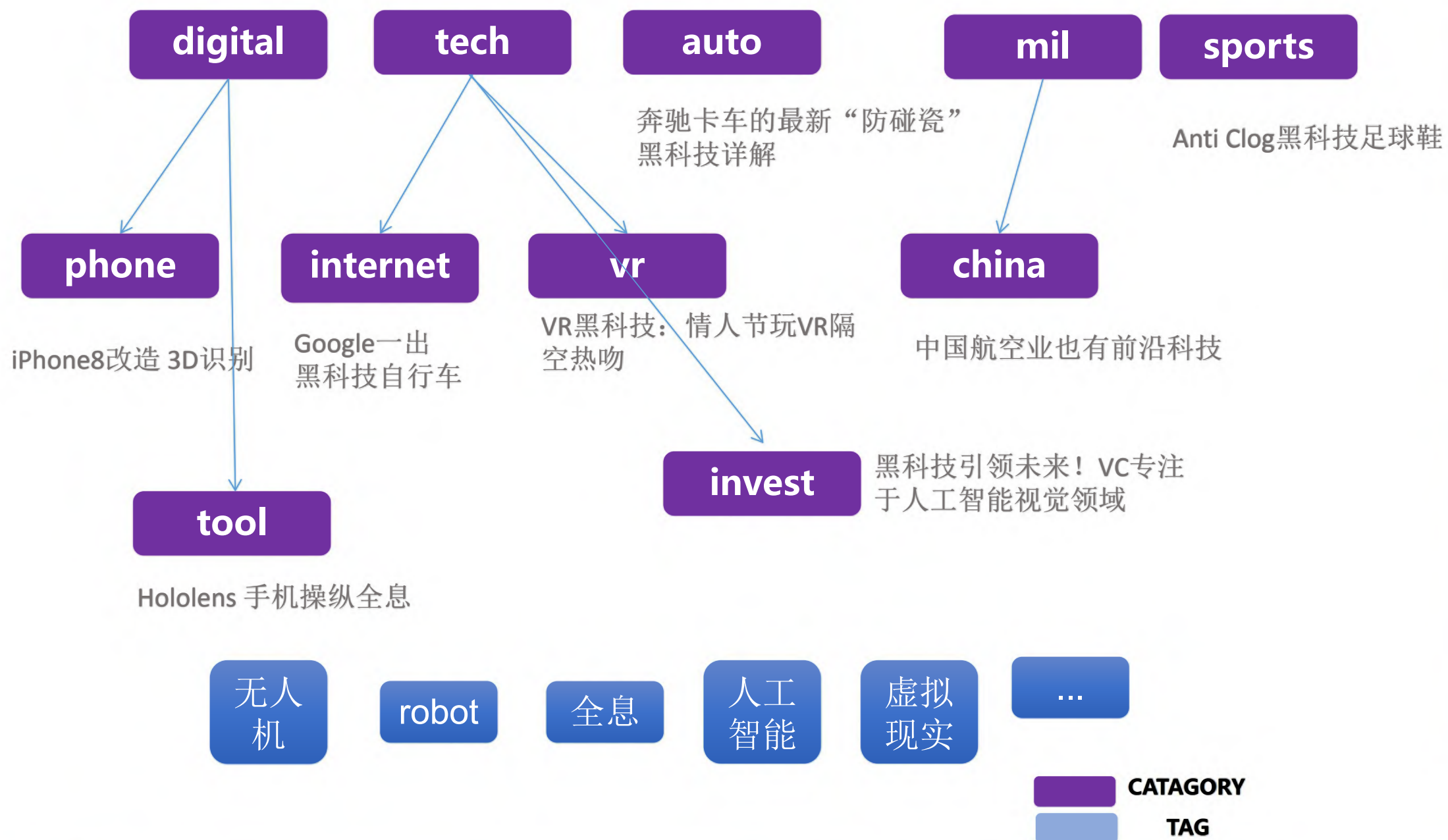
Topic : Feature , User Profile, Retrieval

Tag : Retrieval , User Profile

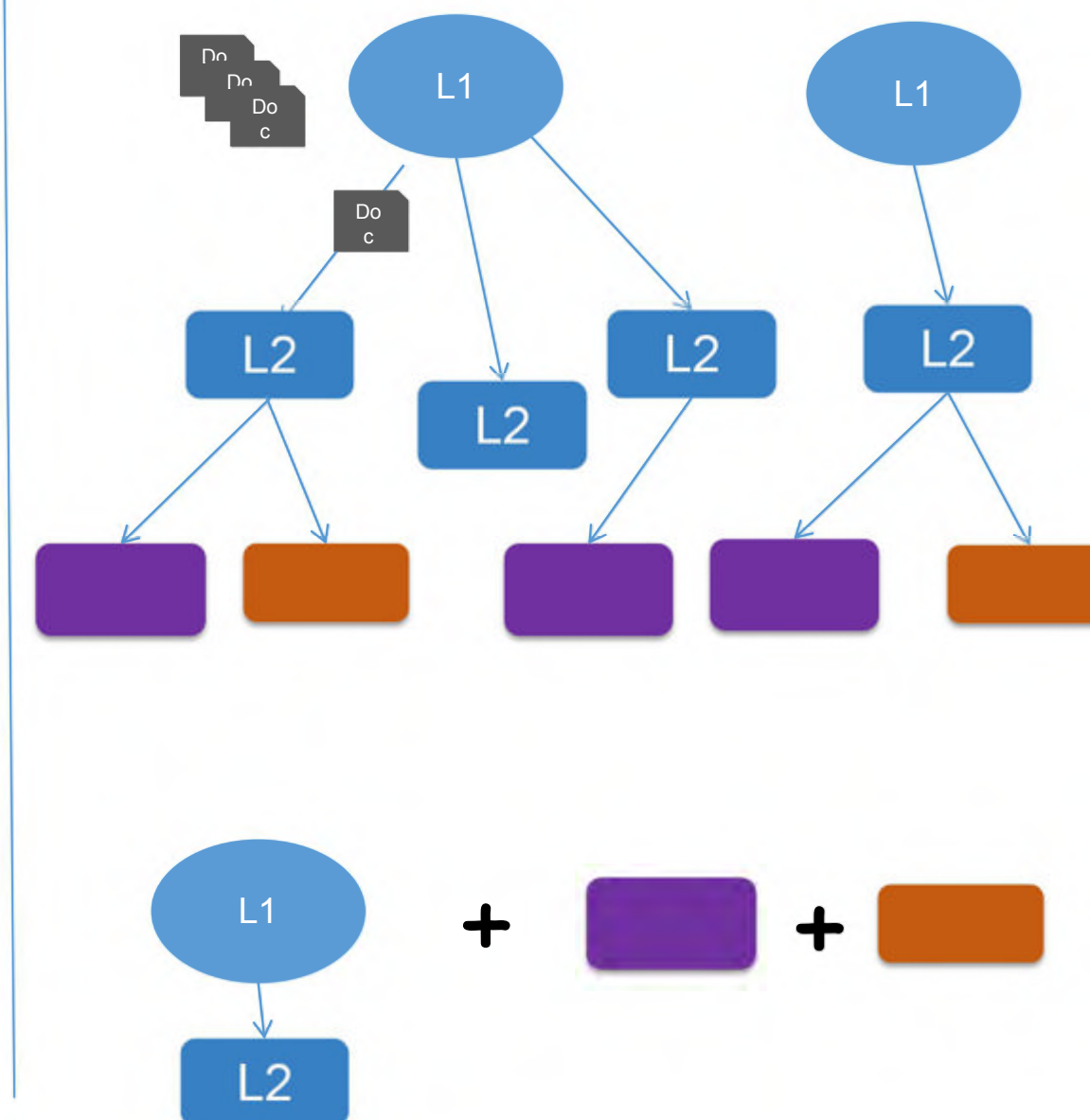
Labeling of Topic : Client , Feature , User Profile, Retrieval

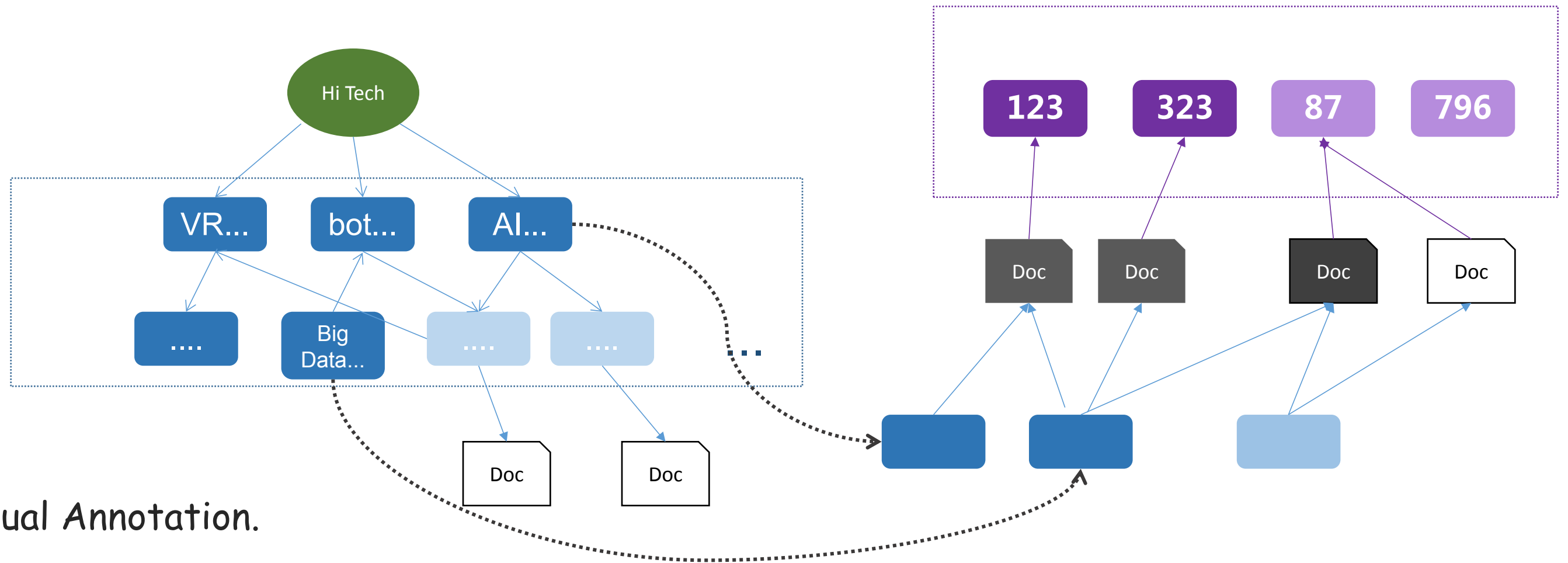
eg:Future Technologies

Labeling of Topic : high tech

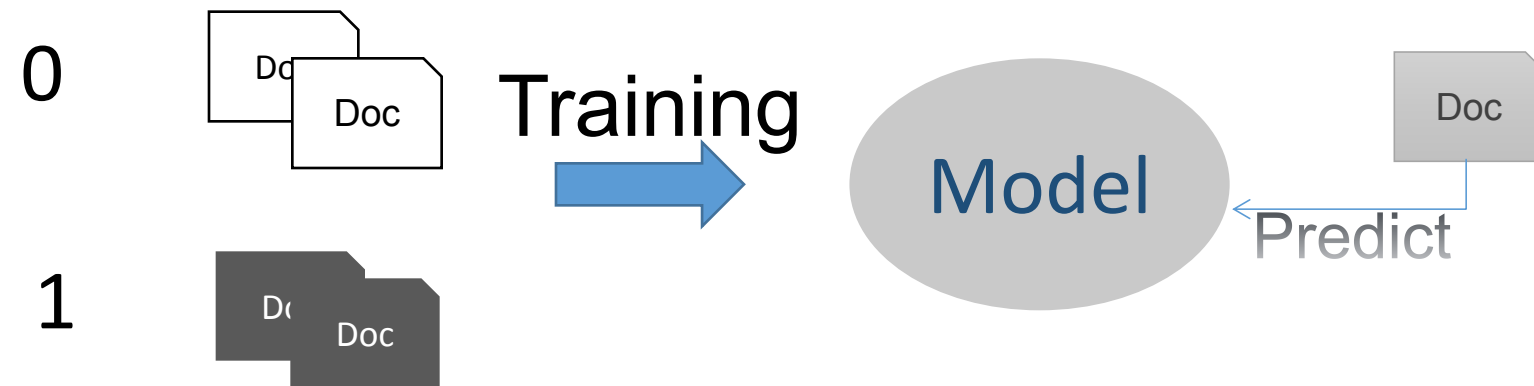


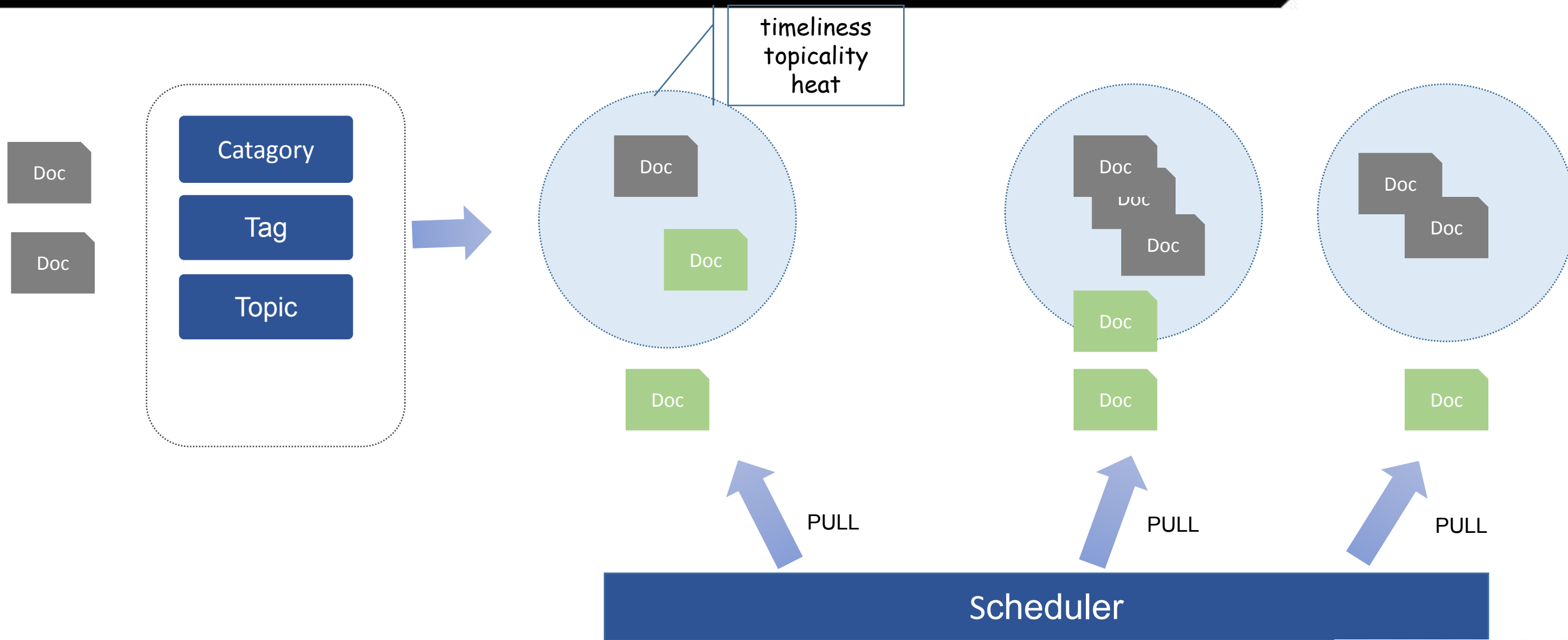
Labeling of Topic: Another dimension to describe content



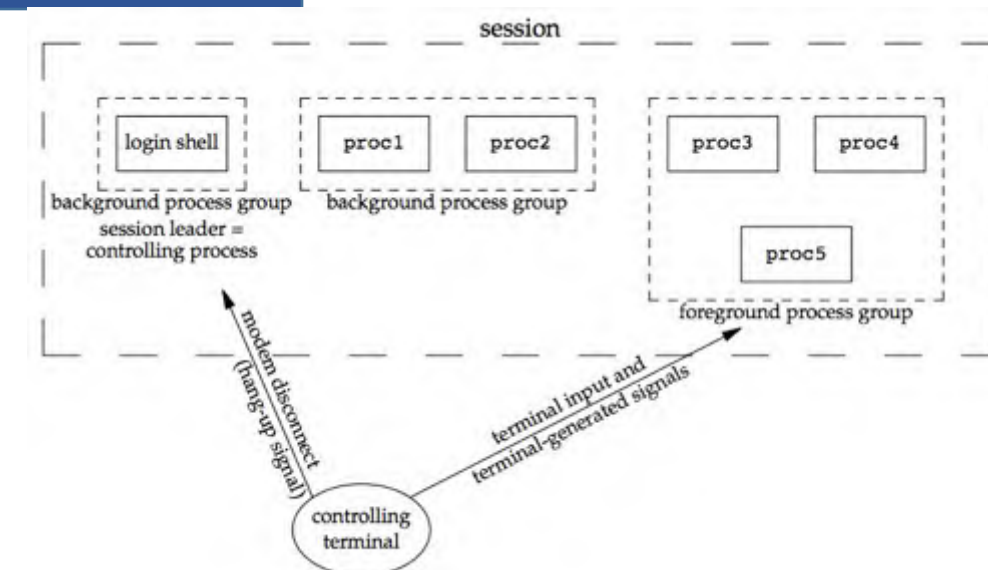
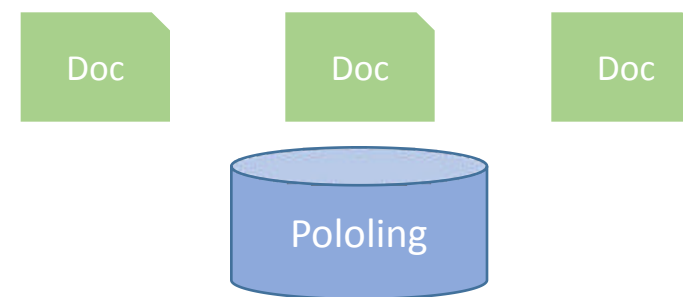


Auto Training.
No Needs Manual Annotation.

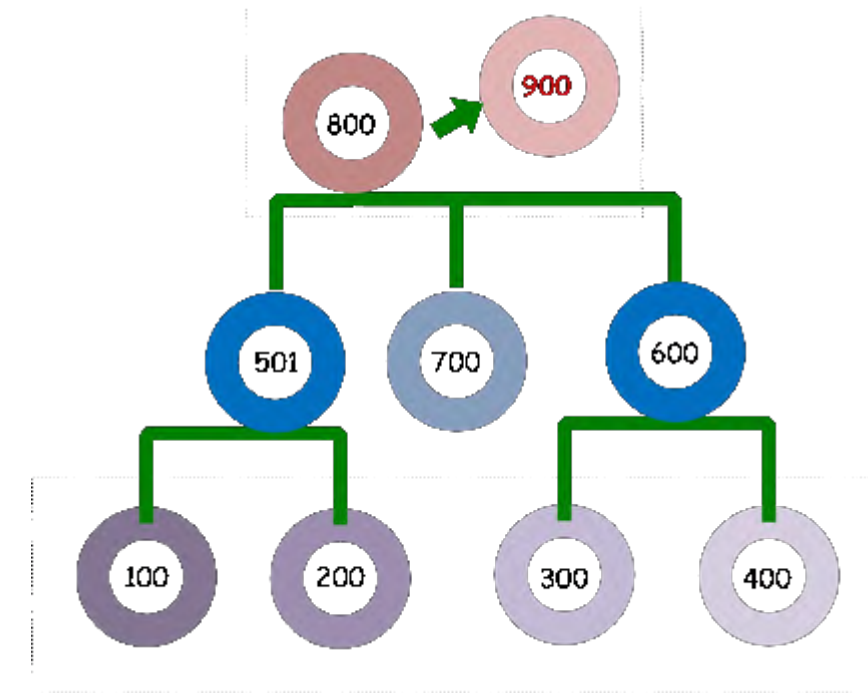
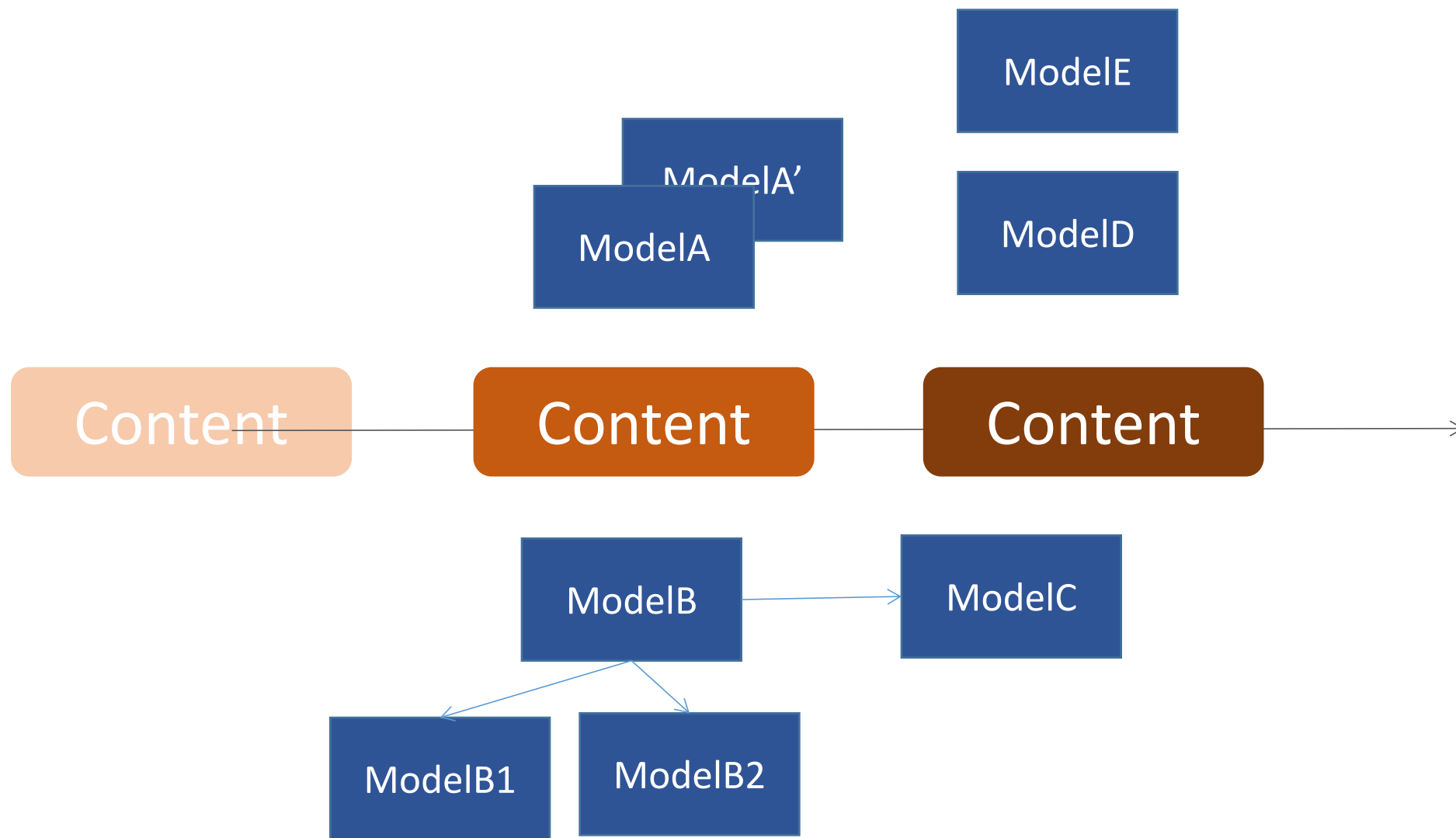




Group Modeling.
Attrs inherit from Group.
Ranking by ContentQuality.
Scheduler Control Content taste.



$$\text{ContentQuality} = \text{Document Score} + \text{Media Score} + \text{PCTR} + \dots$$



Switch Model

Configure

```
<switch name "st_xxx", scan_time = "600" >
    <dependence switch = "st_name">
    <dependence switch = "st_name2">
</switch>
<switch name "st_vvv", scan_time = "200" >
    <dependence switch = "st_xxx">
    <dependence switch = "st_name2">
</switch>
```

SDCC 2017 | 上海

互联网应用架构实战峰会

CSDN

Welcome Join Us

NLP, ML, DL

Text, Photo, Video

Search, Recomm, Ads