Mobvista.

Qcon 2016

Mobvista 海外移动变现核心技术

李海波
Mobvista首席算法架构师

www.mobvista.com

Mobvista.

关于 我

过去式

Microsoft

2007-2008
AdCenter
MSN广告user segment

Baidu百度

- 2008-2014
- 商务搜索，高级项目经理
- 从无到有构建百度风险防控系统
- 百度商业关键词搜索推荐系统。凤巢消费占比50%+(5kw+/天)

AutoNavi

- 2014-2015
- 个性化推荐
- 构建高德商户系统
- LBS推荐系统

现在

Mobvista.

SSP
AD Network
DSP
DMP
反作弊
新闻推荐
机器学习大数据平台
海外子公司数据&算法整合广告user segment

Wechat:dustinsea

http://semocean.com/

Mobvista.

大纲

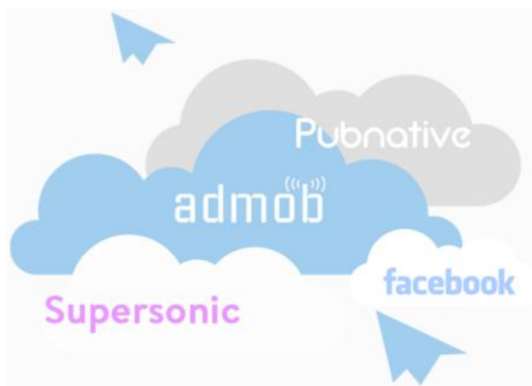
1. 海外移动变现挑战
2. 核心算法框架及数据流
3. 特征&模型定制优化
4. 效果&总结

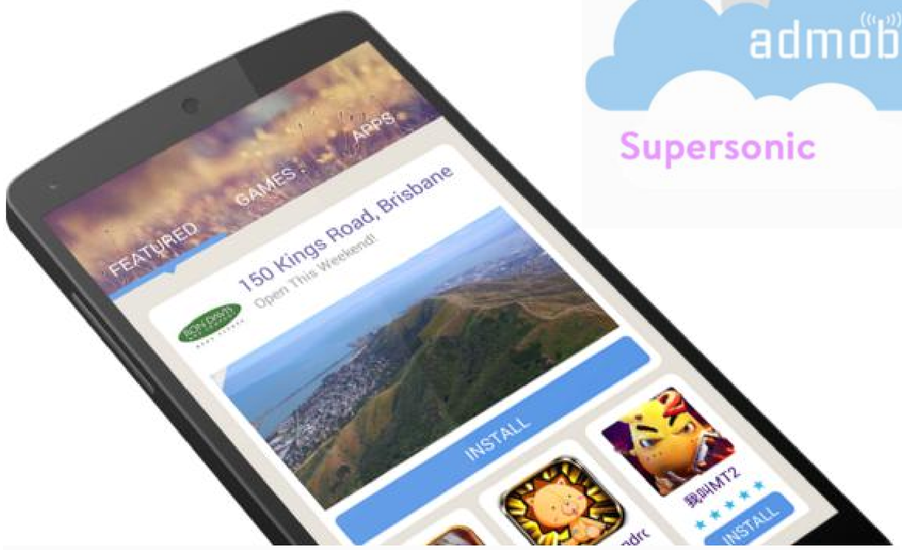
Mobvista 全球业务

AFFILIATES

- 国内外多家广告主独家优质Offer
- 业界领先的CVR和EPC

广告变现M系统
高效变现体系

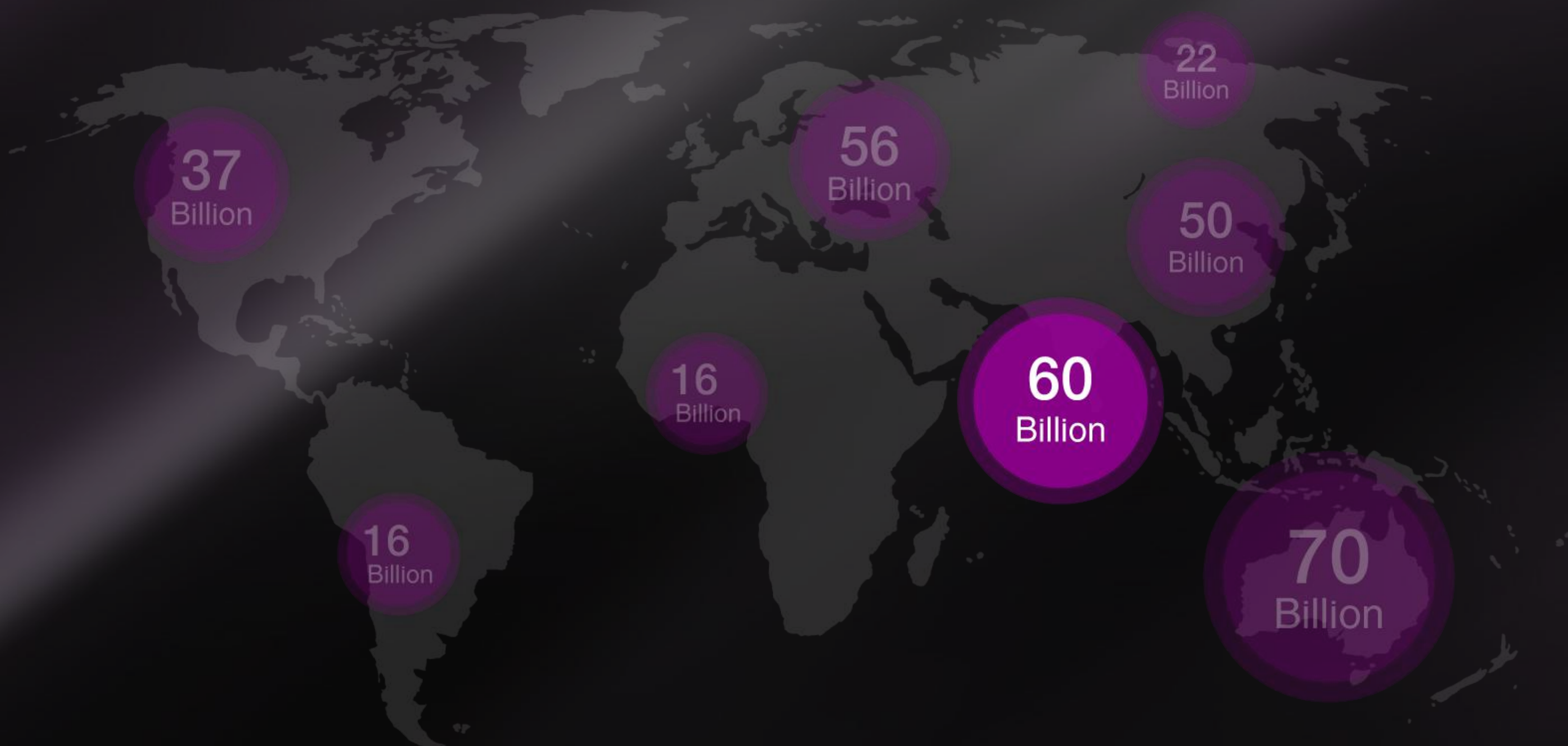




Mobvista.

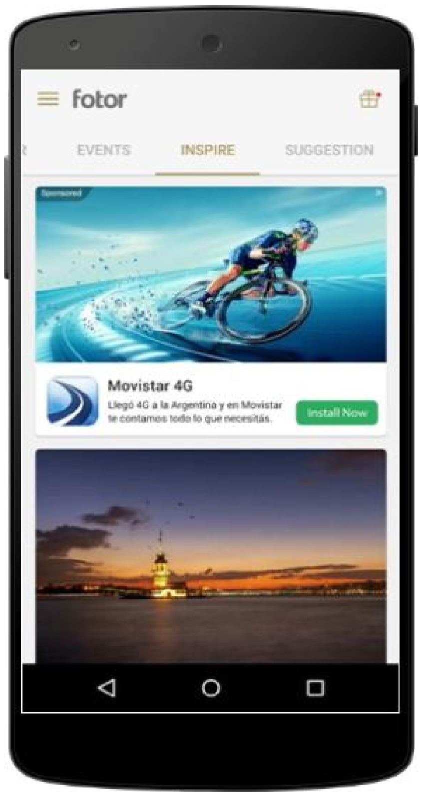
Mobvista 全球变现业务

- 流量覆盖230+ 国家和地区
- 流量：100亿+/天



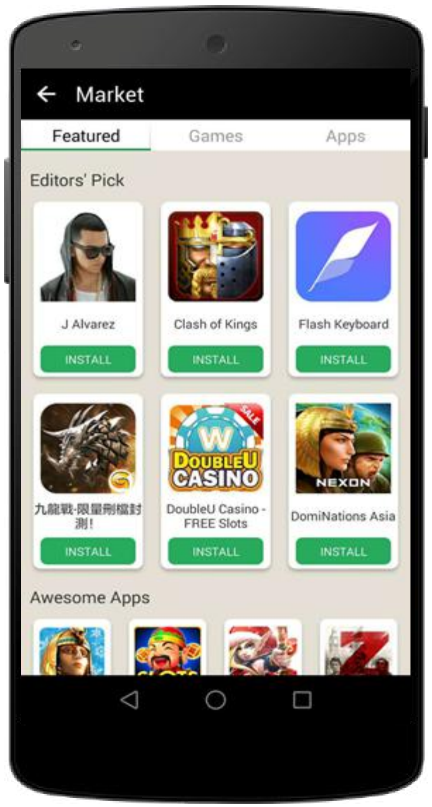
Mobvista.

挑战-广告样式众多



Native

- 广告植入内容设计
- 不打断用户交互，减少干扰
- 推送有价值内容



APPWALL

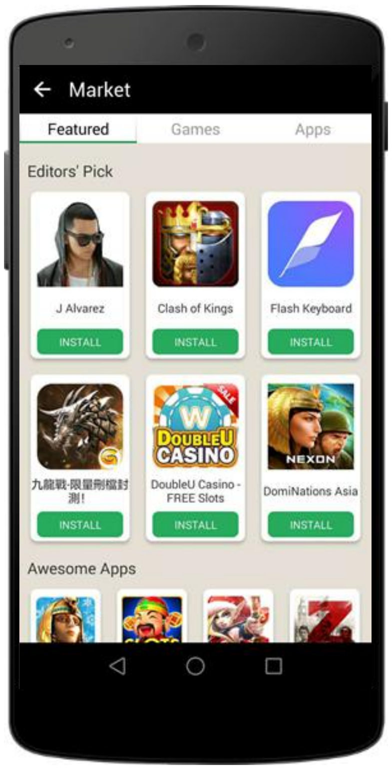
- 入口图表可定制
- 与应用市场相似
- 高CPM

Video

- 适用于游戏开发者
- 激励观看

1. 插屏
2. 全屏
3. 信息流
4. Banner
5. Native
6. 应用墙
7. Video

挑战-转化路径较长



用户看到广告位至广告展现的时间
用户等待的耐心是有限的

用户是否点击广告：CTR预估问题
用户是否被广告内容/素材吸引

点击广告后，能否到GooglePlay安装页面
不同国家，地域网络状况的复杂性与差异

用户是否真的需要该app并进行安装
CVR预估

挑战-国际化带来的多样性

- 国家间网络基础建设水平参差不齐
  e.g. India VS South Korea
- 流量质量，特性差异.
  e.g. 最大200倍差异
- 流量波动较大；开发者流量差异
  巨头 VS 小开发者

Region	Unique IPv4 Addresses	Average Connection Speed (Mbps)	Average Peak Connection Speed (Mbps)	% Above 4 Mbps	% Above 10 Mbps	% Above 15 Mbps
ASIA PACIFIC						
Australia	9,746,046	7.8	41.9	72%	18%	7.4%
China	126,077,926	3.7	23.1	33%	1.6%	0.3%
Hong Kong	3,194,113	15.8	101.1	92%	59%	36%
India	17,901,010	2.5	18.7	14%	2.3%	0.8%
Indonesia	3,589,975	3.0	31.0	17%	0.9%	0.4%
Japan	46,074,201	15.0	78.4	90%	54%	32%
Malaysia	2,048,354	4.9	38.3	52%	4.0%	0.9%
New Zealand	2,075,373	8.7	42.0	87%	22%	8.2%
Philippines	1,354,527	2.8	25.3	10%	0.9%	0.3%
Singapore	1,783,055	12.5	135.4	87%	51%	27%
South Korea	23,871,054	20.5	86.6	96%	68%	45%
Sri Lanka	227,183	5.1	33.5	76%	2.2%	0.6%
Taiwan	10,532,581	10.1	77.9	88%	29%	13%
Thailand	3,431,736	8.2	58.3	93%	18%	5.8%
Vietnam	6,002,849	3.4	25.5	31%	0.6%	0.1%

图：国家间网速差异
From:akamai’ s [state of the internet]

ecpm

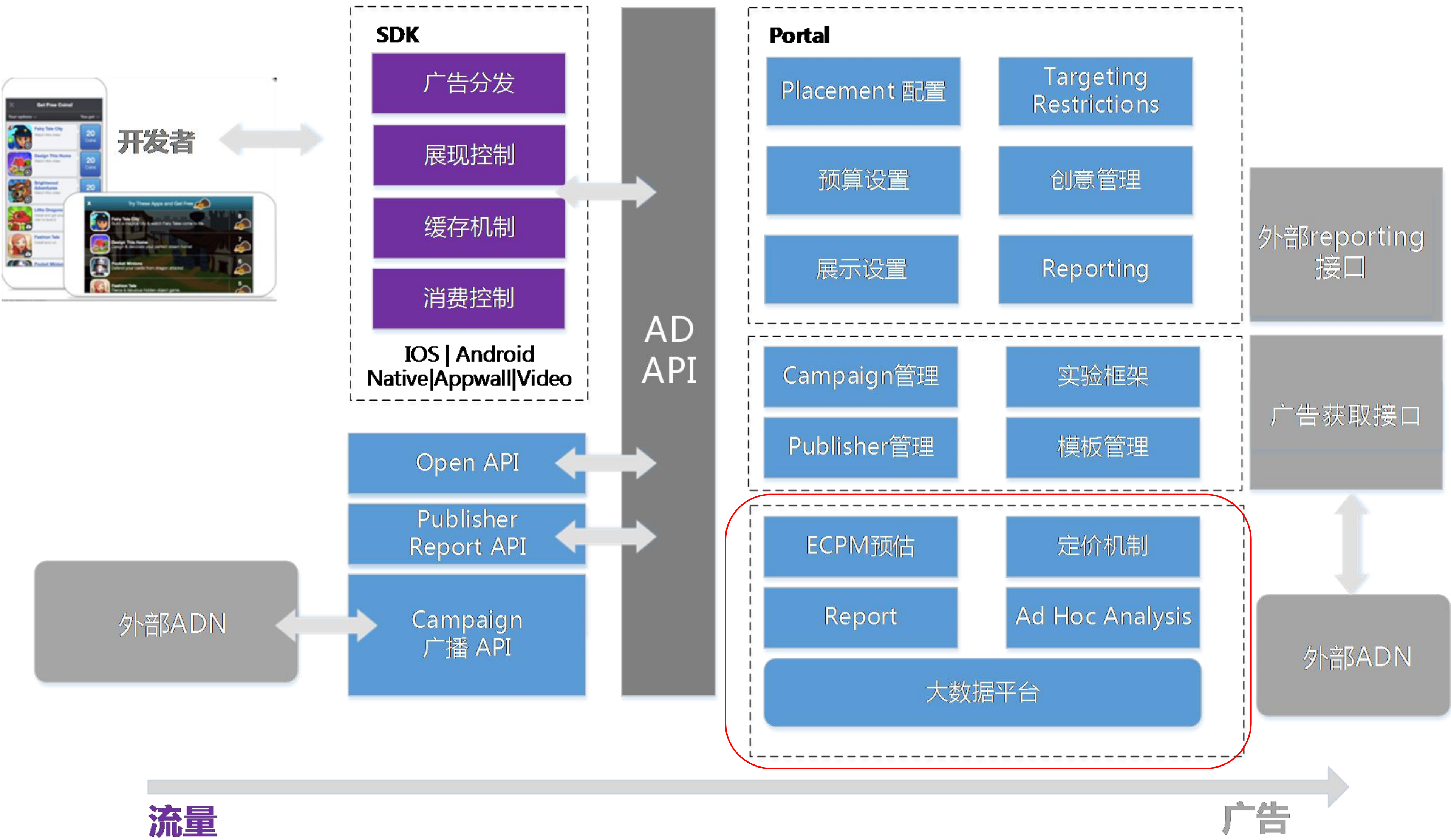
表：国家间ecpm差异

traffic

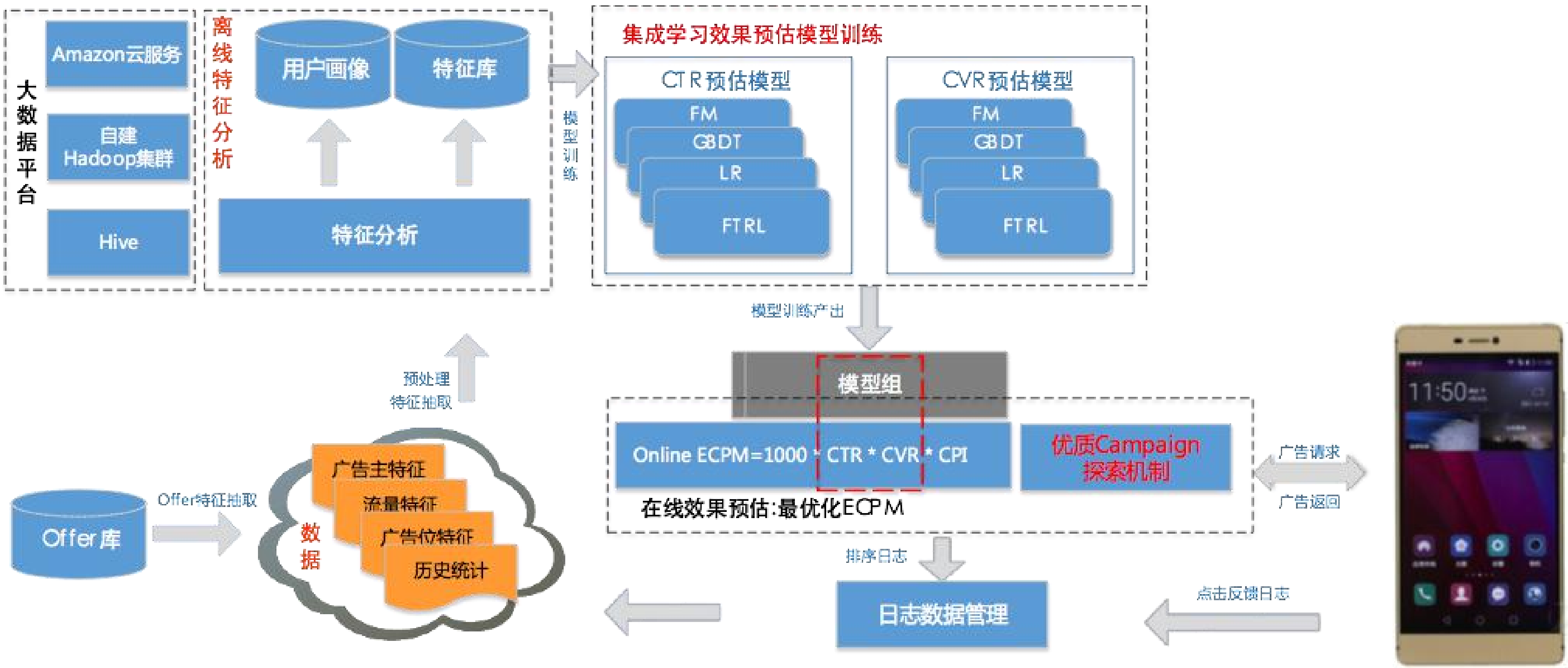
表：开发者间流量差异

Mobvista.

Mobvista 变现架构



变现策略架构



Mobvista.

预测算法框架

Mobvista ecpm目标

ecpm = 1000 * ctr * cvr * price

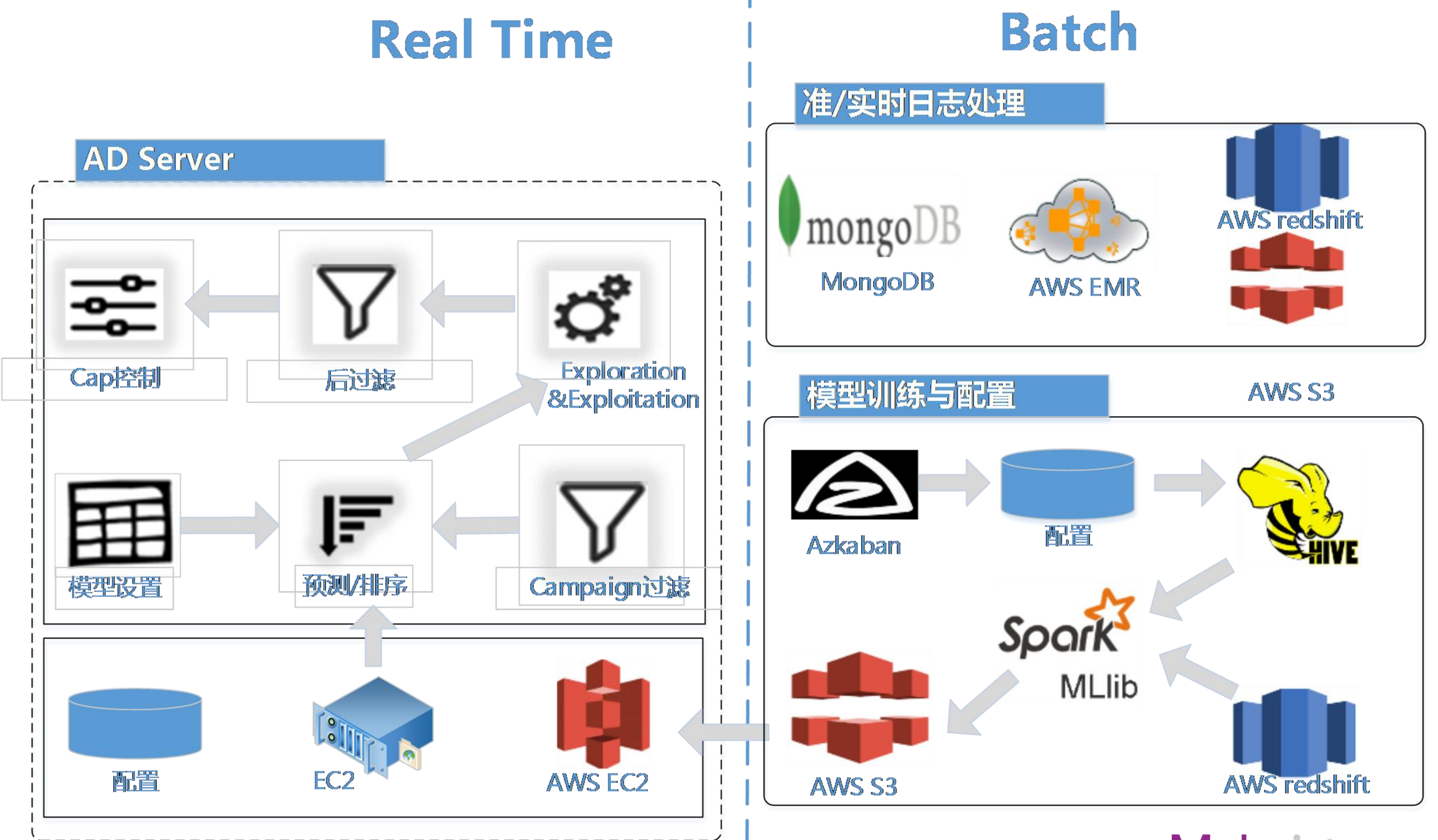
ad	publisher	adviserer	f1	f2	fn	*click	conv	predicted ctr	predicted cvr	price	predicted ecpm
1	1	6	f1_v1	f2_v2		5	6	0.1	0.02	$2	$4.00
2	1	7	f1_v1	f2_v2		3	7	0.2	0.01	$1	$2.00
4	3	8			fn_v1	12	9	0.2	0.01	$1	$2.00
6	4	9				10	4	0.1	0.01	$1	$1.00

Features　　Lables　　f(x)　　Predict　　Win

- 以ecpm为业务目标进行排序
- 分别训练CTR，CVR模型
- A/B test进行效果测试

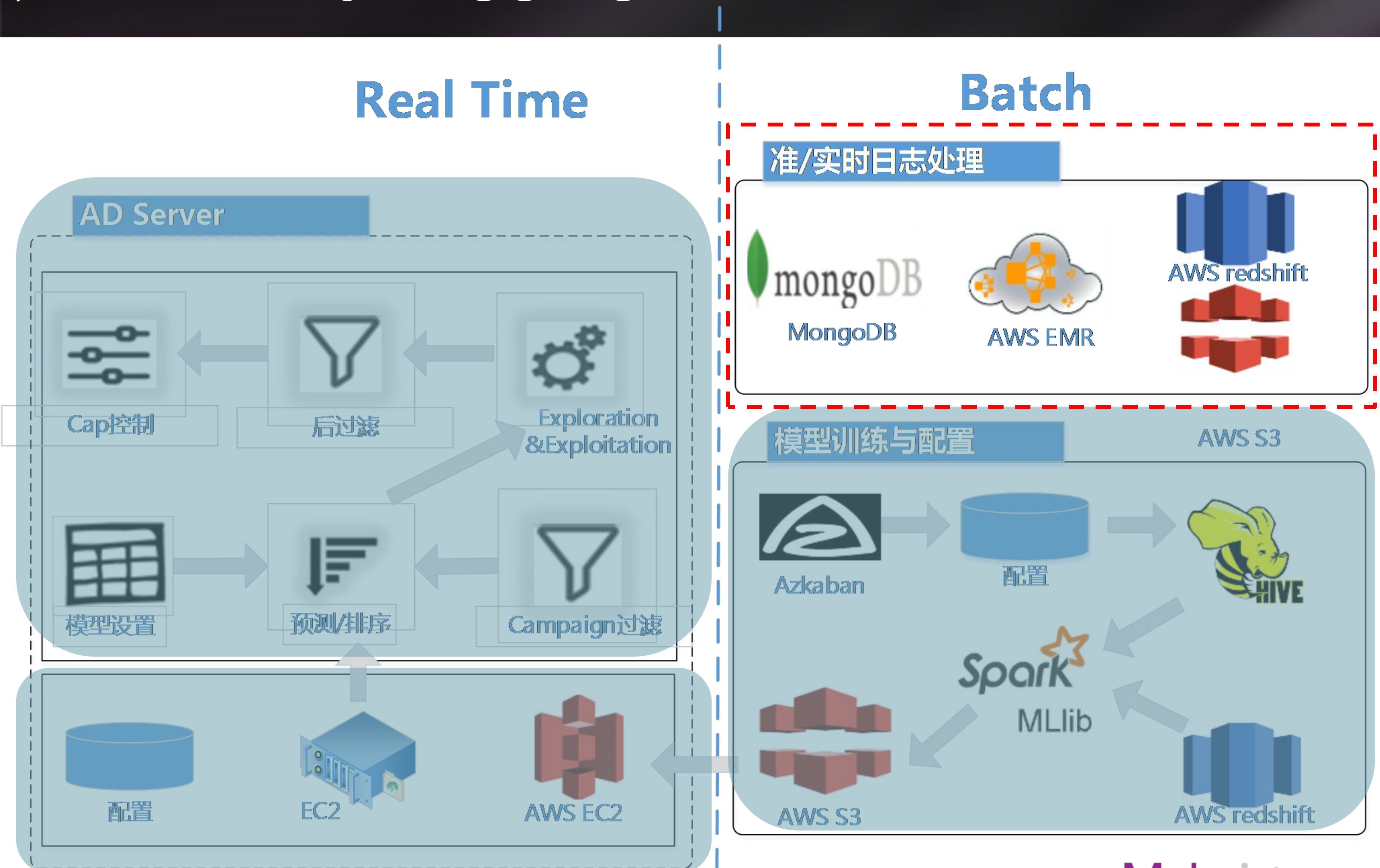
Mobvista.

预测算法框架



Mobvista.

Activity Logging

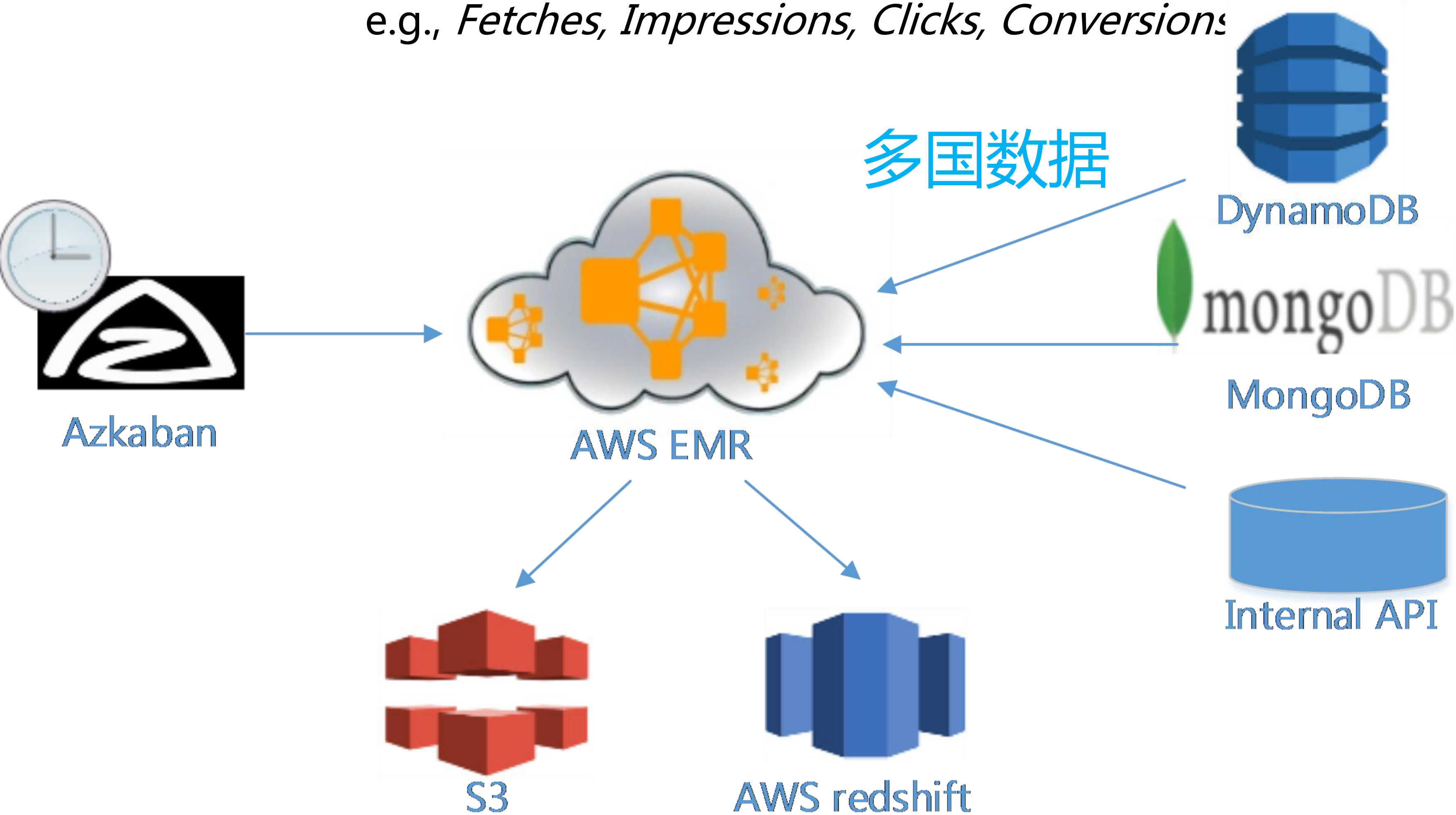


Mobvista.

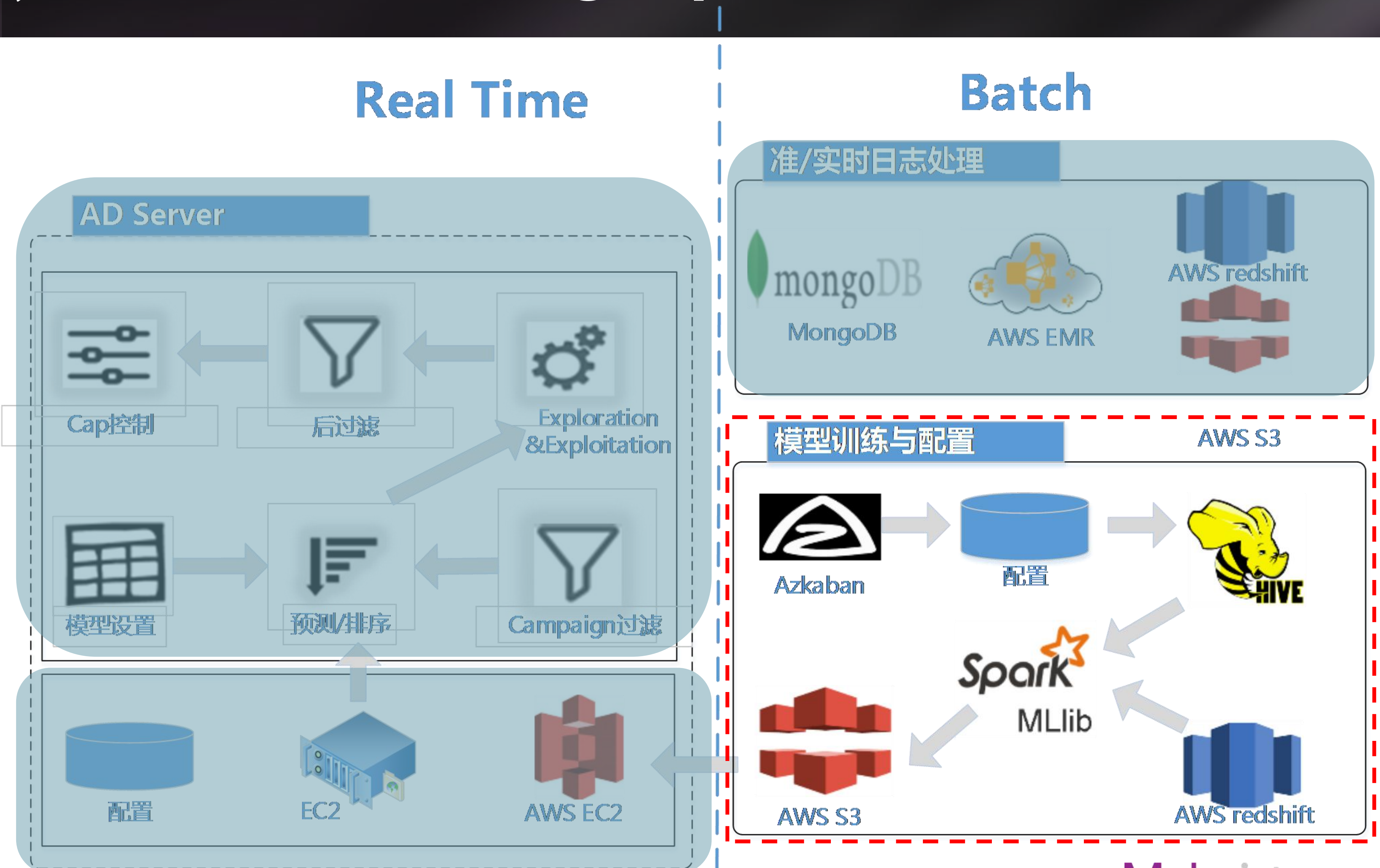
Activity Tracking Data Pipeline

All device activity (events) stored in Mongo&Dynamo as a JSON string

e.g., *Fetches, Impressions, Clicks, Conversions*

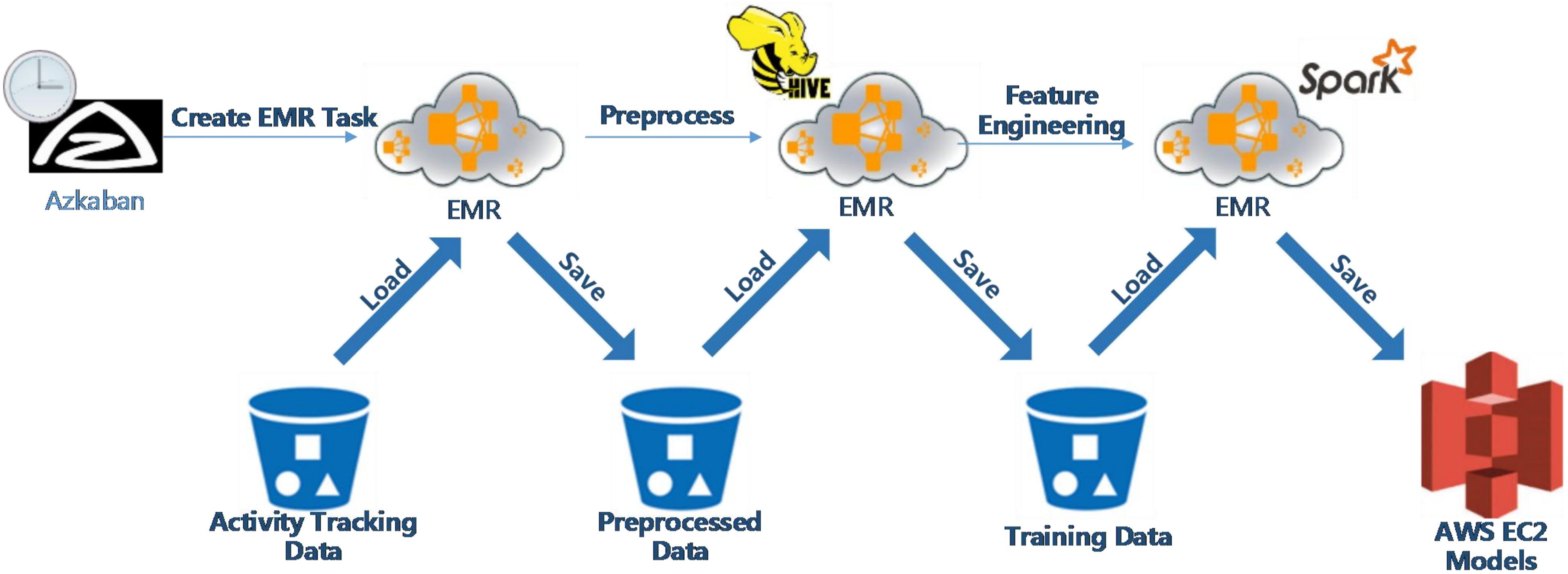


Model Training Pipeline

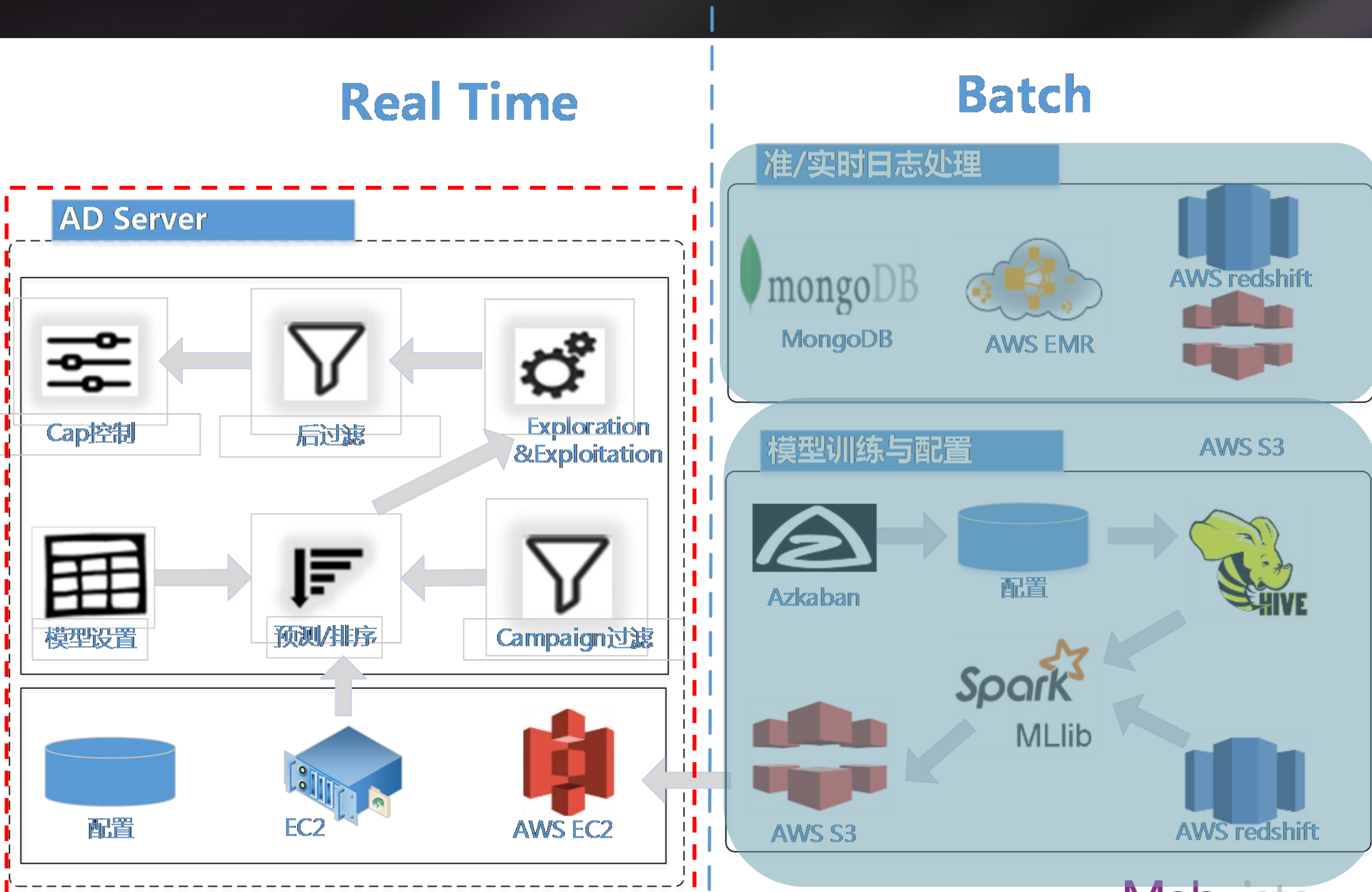


Mobvista.

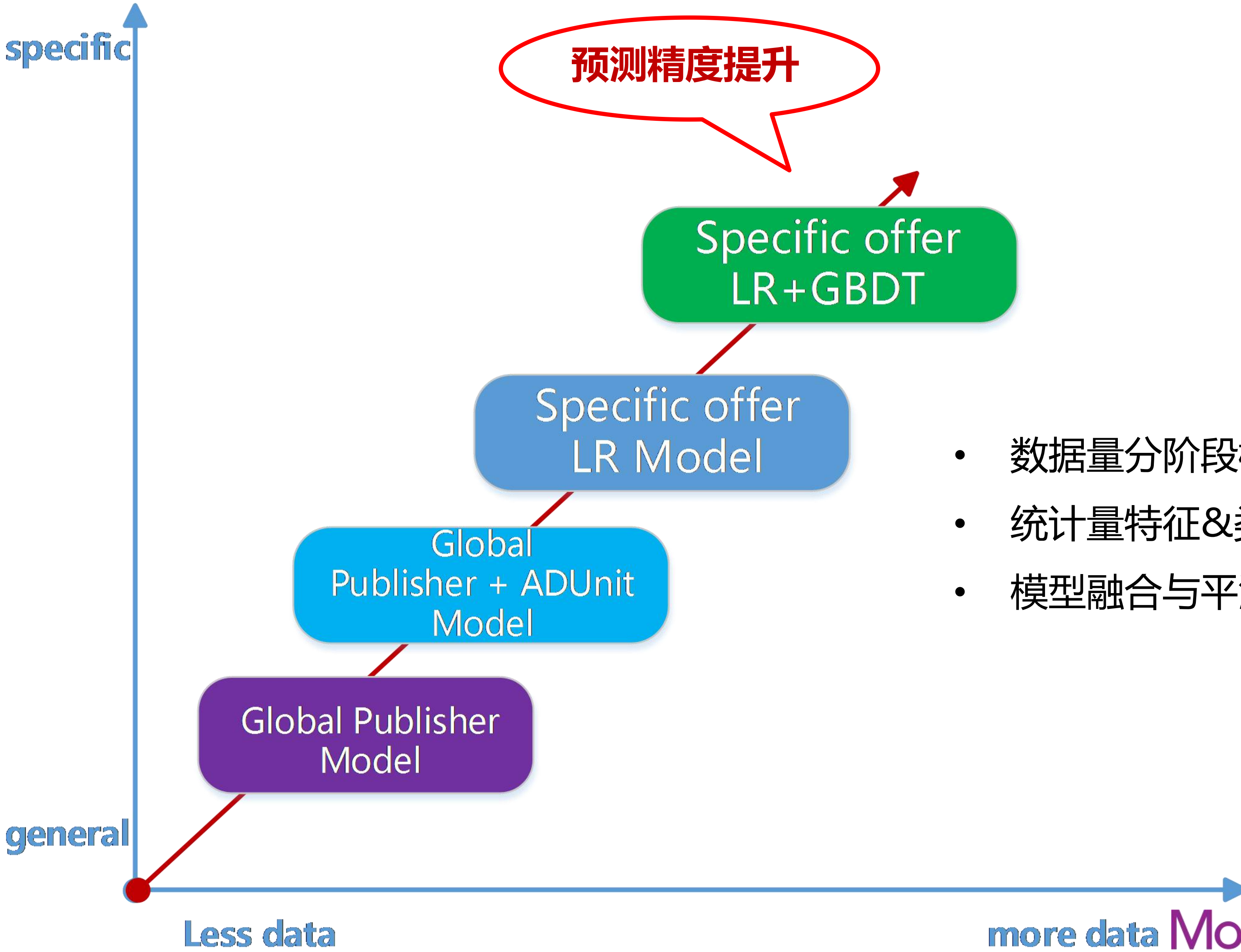
Modeling Pipeline



Ad Sort


Real Time
Batch
AD Server
Cap控制
后过滤
Exploration
&Exploitation
模型设置
预测/排序
Campaign过滤
配置
EC2
AWS EC2
准/实时日志处理
MongoDB
AWS EMR
AWS redshift
模型训练与配置
AWS S3
Azkaban
配置
HIVE
Spark
MLlib
AWS S3
AWS redshift

模型分级



- 数据量分阶段模型
- 统计量特征&类别特征
- 模型融合与平滑

Mobvista.

类别特征

Feature	Description
PublisherID	ID of publisher
APPID	ID of app
SessionCountryID	country id from ip address supplied by publisher on create session
DeviceOSVersionId	Version id for a given device operating system
SessionIsOnWifi	Flag indicating whether the device uses wi-fi or cellular
PublisherSDKVersionId	SDK version id or API integration flag
PlacementCategory	category of placement
TimeCategory	Grouping of impressions by time of day it happened (UTC)
DeviceModel	Model of Device(.e.g. Huawei,Samsung)
CampaignID	ID of Campaign
AdvertiserID	ID of Advertiser
language	language of system

Nominal Scale

ID Feature

- ID类特征
- 行为类特征
- 用户类特征

不同特性如何区别对待？

Mobvista.

行为特征

Interval Scale

Feature	Description
FrequencyImpression	Number of impressions for a given device for a given channel unit in the last 30 days
RecencyImpression	The last time the device was active for a given channel unit in the last 30 days (Minutes)
RecencyClicks	The last time the device generated a click for a given channel unit in the last 30 days (Minutes)
FrequencyClicks	Number of clicks for a given device for a given channel unit in the last 30 days
RecencyConversion	The last time the device converted for a given channel unit in the last 30 days (Minutes)
Total Revenue	Total revenue per the CU for this device in the last 6 months
TimeSinceLastOfferView	Time from the last impression on this device for this CU in seconds
Minute Frequencies	Number of impressions for this device for given CU in the last minutes
Hour Frequencies	Number of impressions for this device for given CU in the last hour
Day Frequencies	Number of impressions for this device for given CU in the last day

- 统计量特征&类别特征
- 模型融合与平滑

数据举例

nominal	ratio scale	high miss	binary	skew	
country	total revenue	age	OnWifi	fea5	label
India	$4.02	male	On	1.5	1
India	$3.19	null	On	1.7	0
Japan	$0.71	null	On	1.5	1
Japan	$3.82	female	Off	2.5	1
Brazil	$0.05	null	On	1.5	0
Philippines	$2.71	null	Off	1.5	0
Philippines	$1.99	null	On	1.5	1
Mexico	$3.10	female	Off	1.5	0
Philippines	$1.69	null	Off	1.5	0
Malaysia	$3.30	null	Off	1.5	0
Korea	$0.80	male	On	1.5	1
Philippines	$1.30	null	On	14.8	0

不同类型特征预处理方式?

基本处理方法

Feature	Description
FrequencyImpression	Number of impressions for a given device for a given channel unit in the last 30 days
RecencyImpression	The last time the device was active for a given channel unit in the last 30 days (Minutes)
RecencyClicks	The last time the device generated a click for a given channel unit in the last 30 days (Minutes)
FrequencyClicks	Number of clicks for a given device for a given channel unit in the last 30 days
RecencyConversion	The last time the device converted for a given channel unit in the last 30 days (Minutes)
Total Revenue	Total revenue per the CU for this device in the last 6 months
TimeSinceLastOfferView	Time from the last impression on this device for this CU in seconds
Minute Frequencies	Number of impressions for this device for given CU in the last minutes
Hour Frequencies	Number of impressions for this device for given CU in the last hour
Day Frequencies	Number of impressions for this device for given CU in the last day

Unit/Weighted Wised

Mean
Max
Min
Skewness
Kurtosis
Missing Rate

Unit/Weighted Binning

Boundaries
Counts
Positive Counts
Negative Counts

Level2 Statistics

Positive Rate
Negative Rate
Information Values
Weight of Evidence Values
Kolmogorov-Smirnov Values

Mobvista.

Binning Methods

Equal Interval

Equal Total(Each bin with 3 elements)

e.g. RecencyClicks,TotalRevenue,RecencyConversion

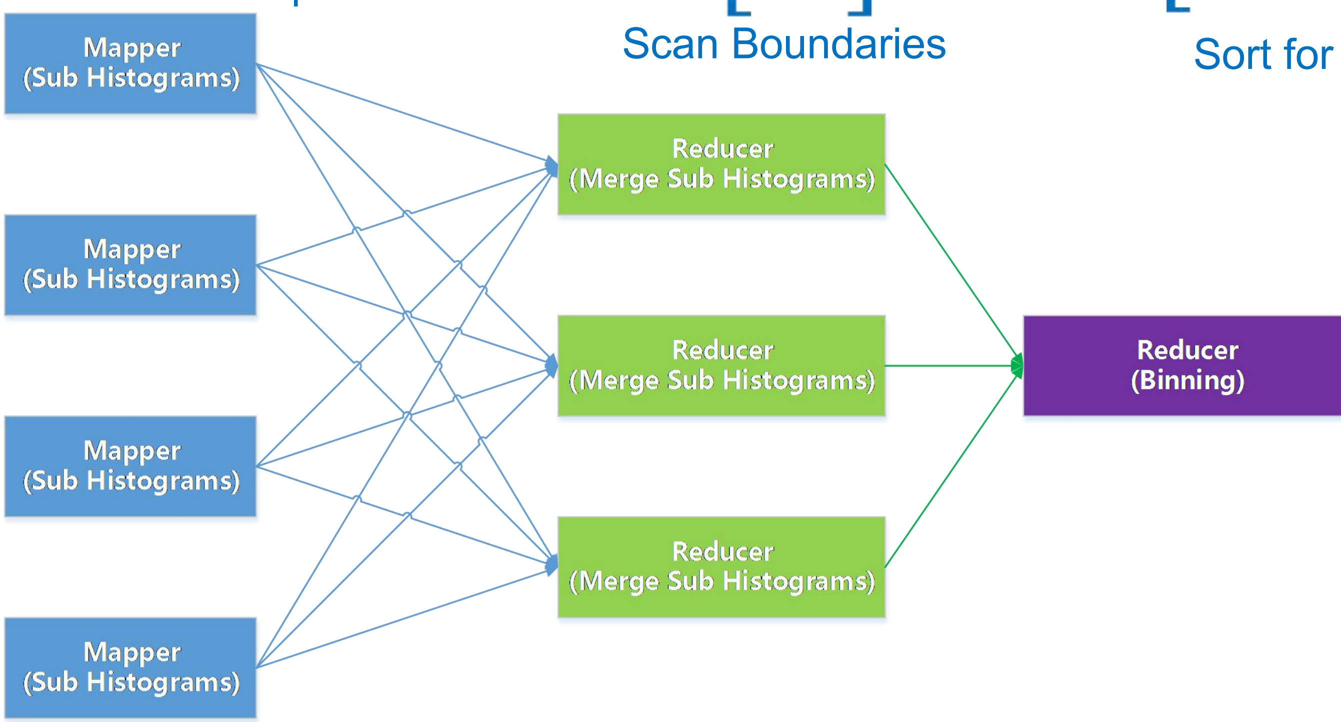
Equal Positive/Negative(Each bin with 3 positive/negative elements)

现实中的效果如何？

Binning Methods

$$A \rightarrow A^T \rightarrow \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} \rightarrow \begin{bmatrix} Sort(A_1) \\ Sort(A_2) \\ Sort(A_3) \end{bmatrix}$$

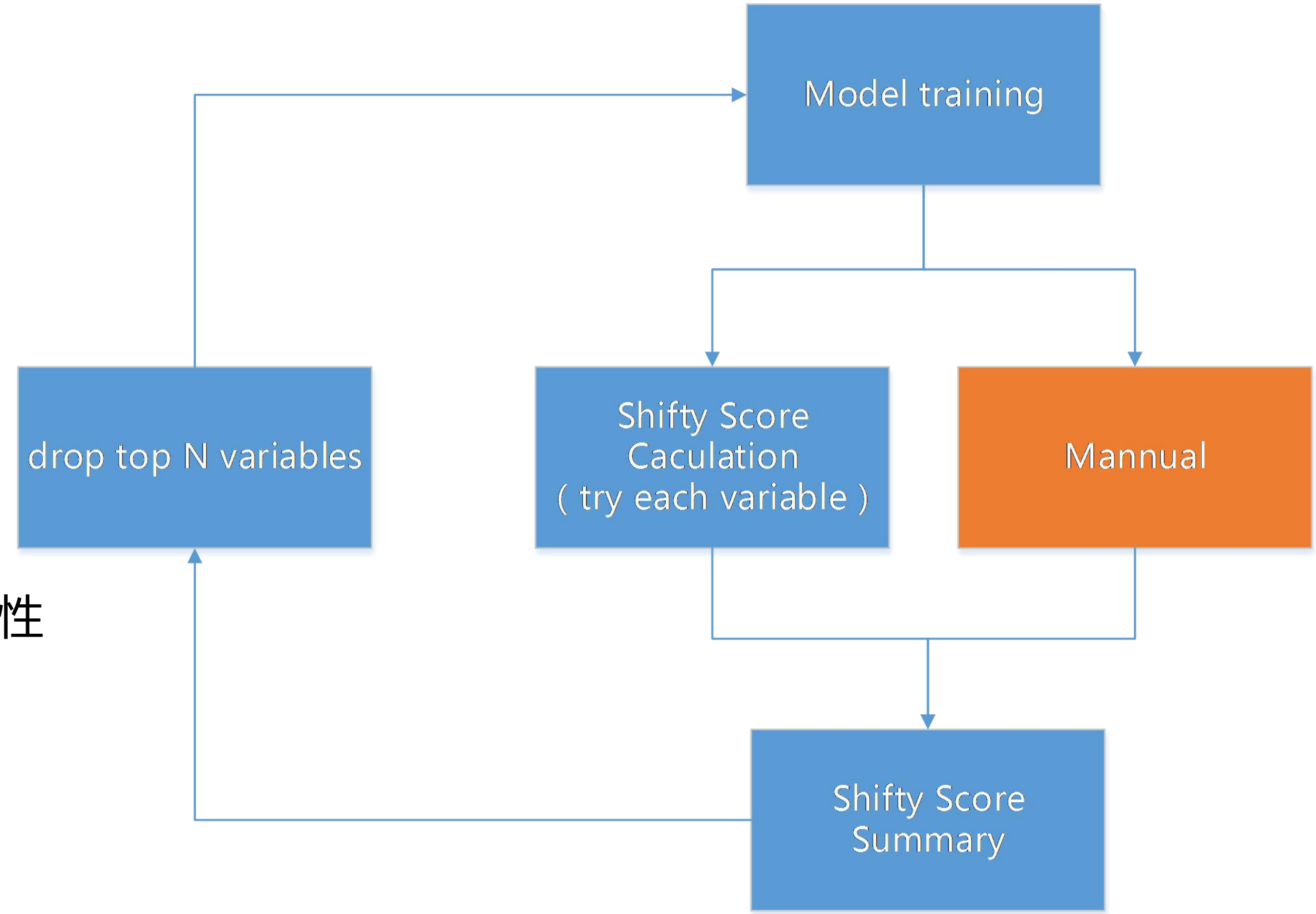
Original Features · Transposition · Scan Boundaries · Sort for Binning



e.g. RecencyClicks,TotalRevenue,RecencyConversion

Variable Selection

- 迭代特征选择
- 重要性度量
- 效率/效果权衡
- 人工经验的重要性



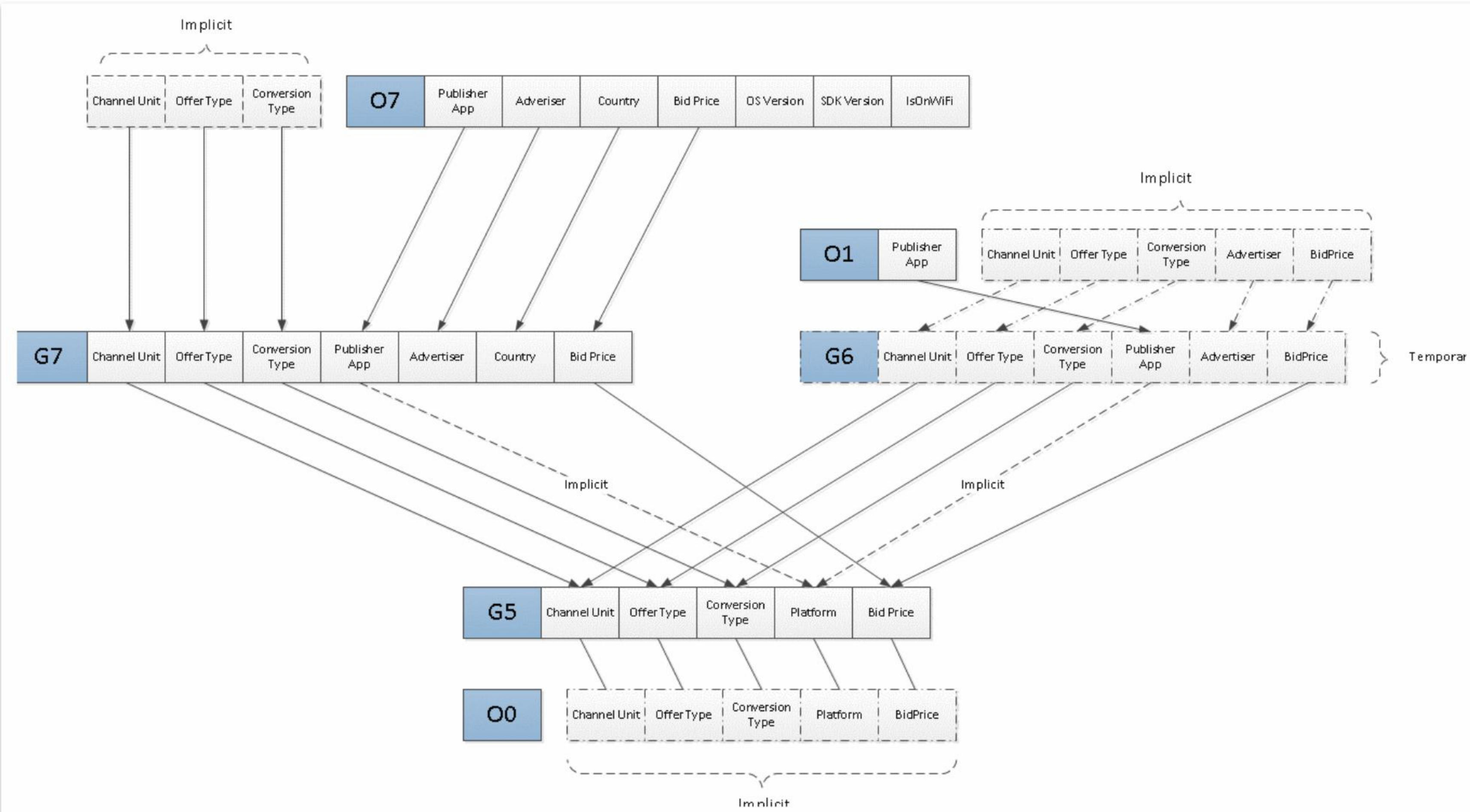
e.g. RecencyClicks,TotalRevenue,RecencyConversion

模型分级

- 数据量分阶段模型
- 统计量特征&类别特征
- 模型融合与平滑

Mobvista.

Global模型



Mobvista.

Global模型

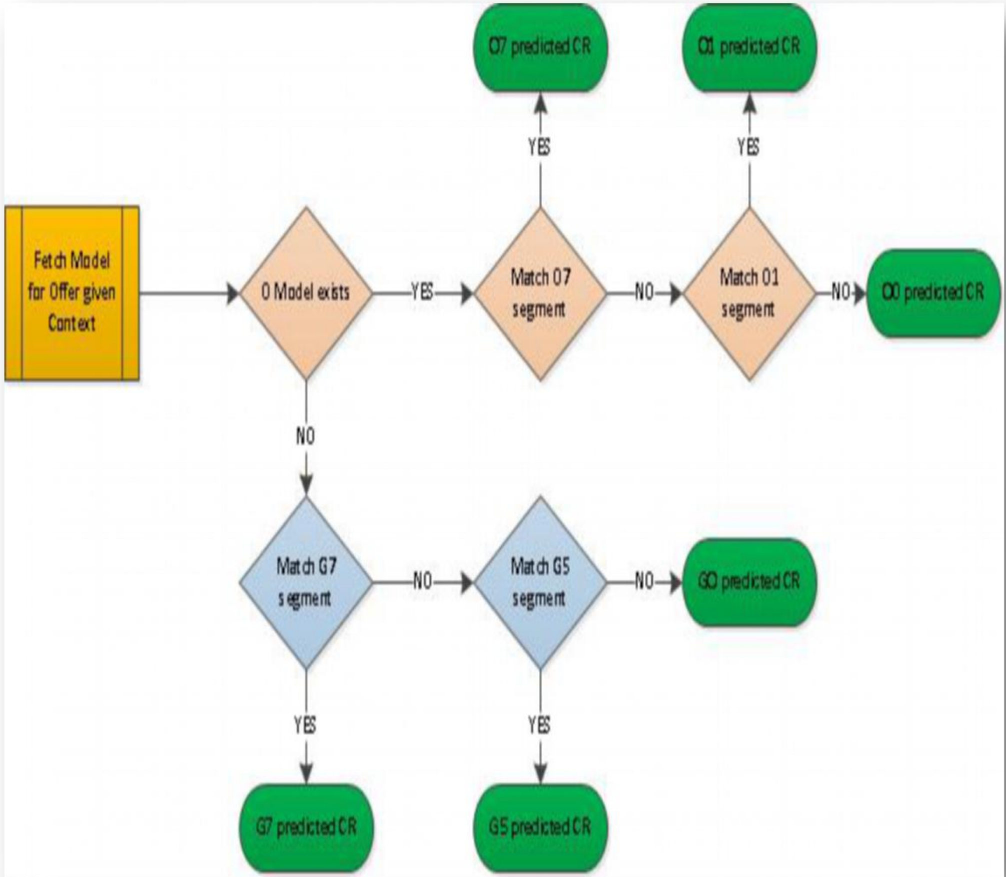
Specific

Regular Segment
- A regular segment consists of a concatenation of specific combination of values of the model's regular segmentation variables, plus a predicted conversion rate.

Fallback Segments
- A fallback segment prediction is looked for when a regular segment prediction isn't found.

Default Prediction
- A single default predicted conversion rate value which is used as the prediction when regular and fallback segment predictions aren't found.

General



树模型

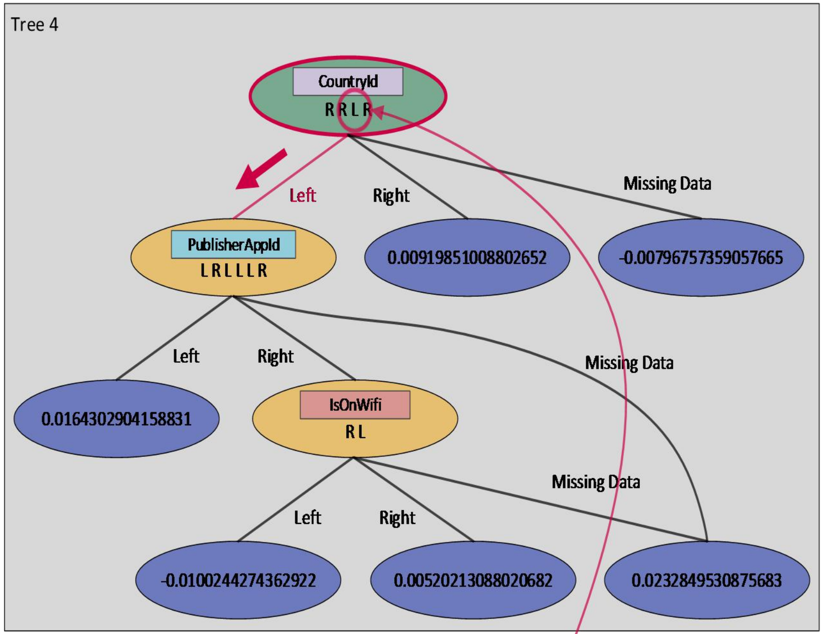
GBDT对数据充分CASE进行预估

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta, \mathbf{a}} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a})\right)$$

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; \mathbf{a}_m).$$

##Calculating an Offer’ s Predictive Score with a DT Model

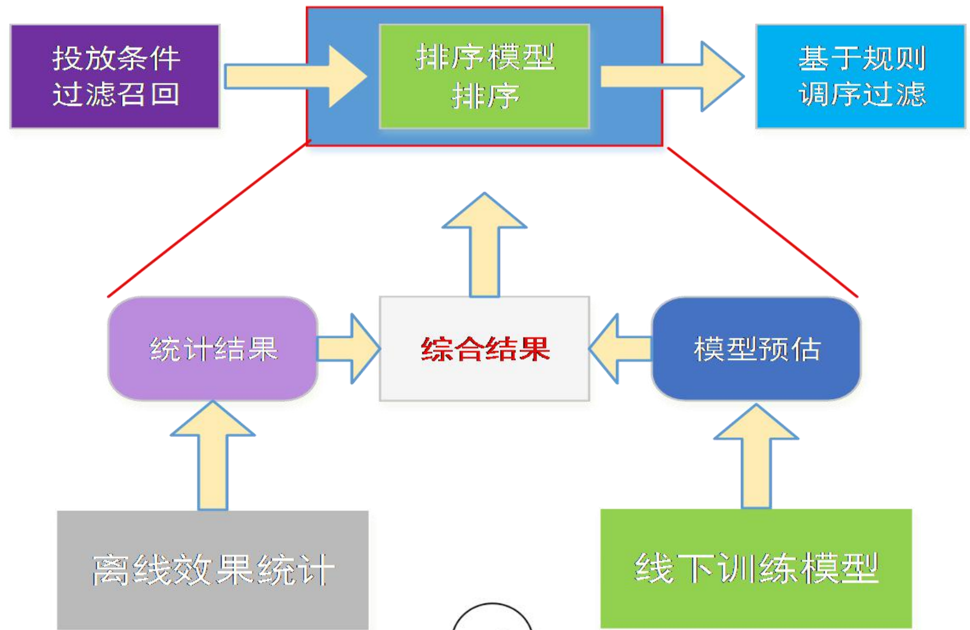
```
1.BiasOffset: -7.02642680869964
2.Decision Tree Results
                 - Tree 1:                    0.00576330069452524
                 - Tree 2:                    -0.0100230621173978
3.Logit of Predicted CR
                 =  BiasOffset + SUM( [result of each tree] )  =  -
6.966206
4.Predicted Conversion Rate
                 =  1  /  ( EXP( -1 * Logit of Predicted CR) + 1 )
                 =  0.000942278186909201
5.Predictive Score for Sorting (Predicted Revenue)
                 = Predicted Conversion Rate * Bid Price
                 = 0.000942278186909201 * $2.00
                 = 0.00188455637381840 (* 1000 = $1.88 eCPM)
```



Variable Value Lists

PublisherAppId	11678 11781 12232 12668 13385 13944
DeviceOSVersionId	1080 1141 1172 1270 1449 1578 60091 60113 60137
CountryId	14 229 37 76
IsOnWifi	false true
PublisherSdkVersionId	110 133 145 152 155 156 161

Smothing CTR Estimation



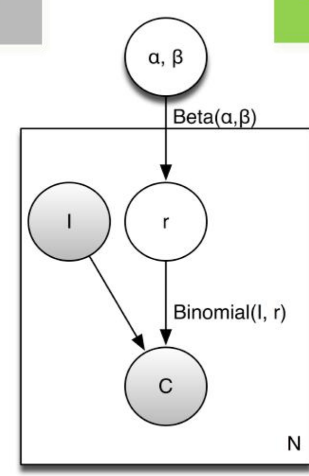
- Bayes Methods

$$CTR_i = \frac{(C_i+\alpha)}{(C_i+\alpha+\beta)}$$

- 指数平滑

$$\text{CTRi} = C_i \qquad if\ i = 1$$

$$CTR_i = \gamma C_i + (1-\gamma)C_{i-1} \quad if\ i >= 2$$



Exploration&Exploitation

Parameters for each offer:

Predicted Conversion Rate ***CR$_p$*** and total number of Impressions

Given true conversion rate . The likelihood is a binomial distribution:

$$\binom{N_{imp}}{N_{imp}\cdot CR_p} CR^{N_{imp}\cdot CR_p}(1-CR)^{N_{imp}-N_{imp}\cdot CR_p}$$

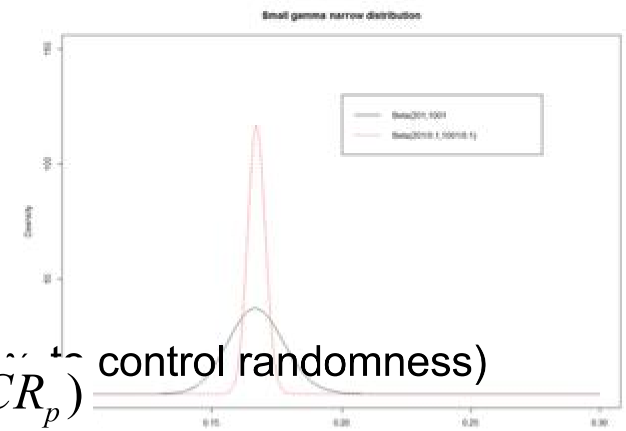
Conjugate Prior of true conversion rate

$$CR \sim Beta(\alpha,\beta)$$

Posterior Distribution of true c

Draw Conversion rate ***CR*** from Beta (add reshaping parameter γ to control randomness)

$$CR \sim Beta(\alpha + N_{imp}\cdot CR_p, \beta + N_{imp} - N_{imp}\cdot CR_p)$$



Here γ is the reshaping parameter

$$CR \sim \left(\frac{\alpha + N_{imp}\cdot CR_p}{\gamma}, \frac{\beta + N_{imp} - N_{imp}\cdot CR_p}{\gamma}\right)$$

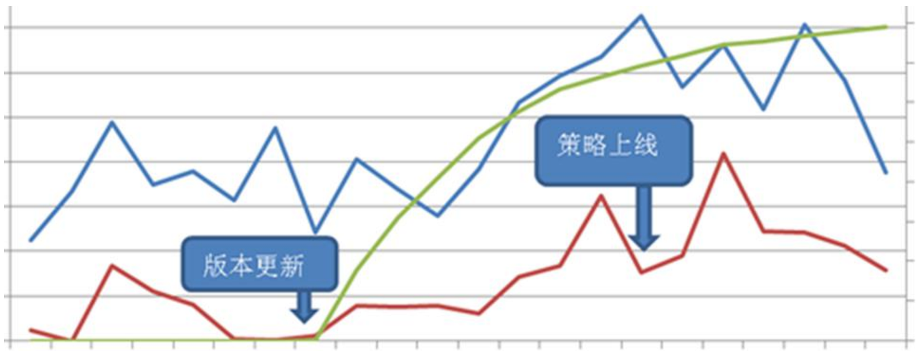
Select the offer w

$$CR_S \cdot \text{bid price}$$

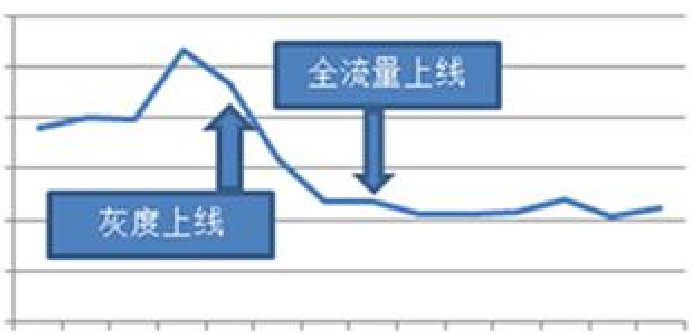
效果&总结

经验：模型很重要，但不是全部

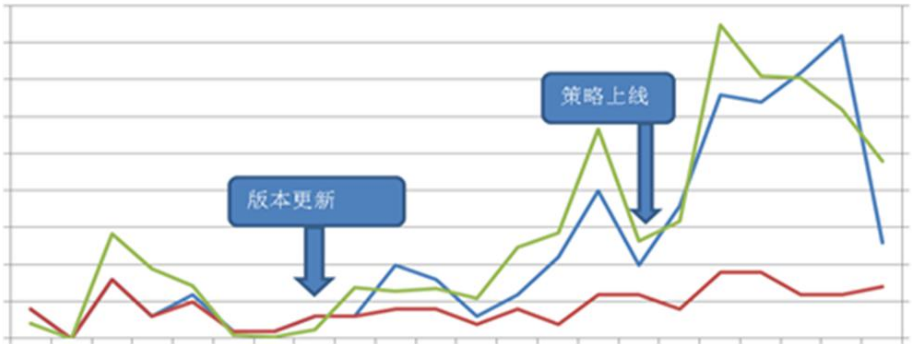
- 展现样式
- 网络连接跳转优化
- 各环节互相配合（e.g. SDK&算法）



Ecpm 模型配合SDK升级1



Ecpm FB/MV



Ecpm 模型配合SDK升级2

效果&总结

小流量 扩大流量 50%流量

经验：学不出来，可考虑拆分

- 分置信度
- 数据丰富程度
- 流量大小
- DEBUG？

模型优化到一定程度，效果最优的策略也许是找bug

Mobvista.

Localization in Depth
World in Hand

Thank You

Wechat:dustinsea

Mobvista.

www.mobvista.com