PPTV的大数据应用

金昀（YUNJIN@PPTV.COM）

提纲

- PPTV的数据
- 数据的价值？
- 大数据工程系统
- 大数据应用

PPTV的数据-用户规模

- 超过**10亿**的客户端下载量，**3.4亿**全平台月度活跃用户
- 每个活跃用户平均每天使用**2小时45分钟**
- 所有活跃用户一天使用总时长约为**10980年**
- 重大直播事件**1000万**用户同时在线

月度用户数(UV)	每天覆盖人数(UV)	直播在线峰值
3.4亿	5000万	1000万

PPTV的数据

300万视频
视频属性，评分，标签，
评论，榜单…

结构化，
半结构化
数据

3.4亿活跃用户
用户属性，播放历史，
播放列表，收藏…

非结构化
数据

服务器日志 （10+T/日）
Web server, LB, CDN server，
广告投放，搜索，P2P
server， application trace…

客户端日志 （10+T/日）
PC客户端，Flash/网页客户
端，移动App，P2P引擎…

数据的价值？

?

投入
版权，带宽，推广，运营…

执行

产出
流量，广告，会员，增值…

数据

分析挖掘

优化
算法
策略

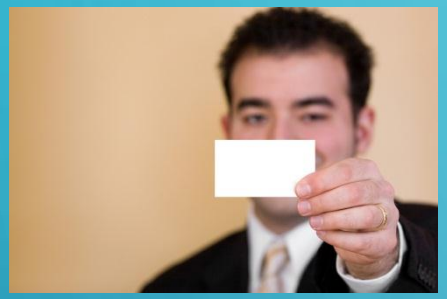
数据的价值 – 商业运营

- 香港动作电影在上个月产生了多少次播放？上周呢？昨天呢？前一个小时呢？
- 不同渠道带来的独立用户有多少？他们的停留时间和留存率如何？
- 每一部视频的投入（版权、带宽）和产出（广告收入，付费点播）比如何？分地区分析？分终端分析？
- 过去6个月月林志玲在视频观众里受关注程度变化趋势如何？山东地区观众对广州恒大队看法是否正面？

数据的价值 – 广告投放

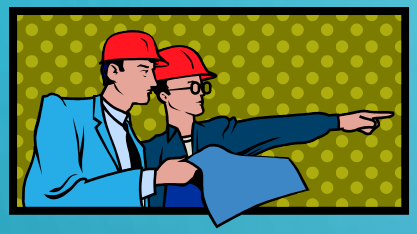
- 容量预测：下个月有多少北京地区的独立用户有可能看成龙电影3次以上？
- 广告投放策略优化及效果跟踪：针对任一次视频观看，怎样投放保证最终投放次数最大化？上海地区前一个小时有多少3+用户被投放了广告X？
- 人群定向：
  - 性别，年龄
  - 用户标签：“白领”，“家庭主妇”，“爱车族”，等等

数据的价值 – 视频推荐



- 精准化推送：一部新的“虐恋剧”上线了，给所有曾经对类似视频感兴趣但是2个月内没有登录的用户发送消息推送
- 个性化推荐：猜你喜欢

数据的价值 – 工程优化



- 上个月Iphone 应用2.2.2版本在武汉地区消耗了多少带宽？直播和点播带宽使用比例如何？VIP用户和普通用户相比带宽保证增加多少？
- 哪些影片应该被推送到三级缓存可以在有限的存储空间内保持最好的服务效果？
- P2P算法调优
- CDN布局调优

安全可用

响应时间

计算复杂性

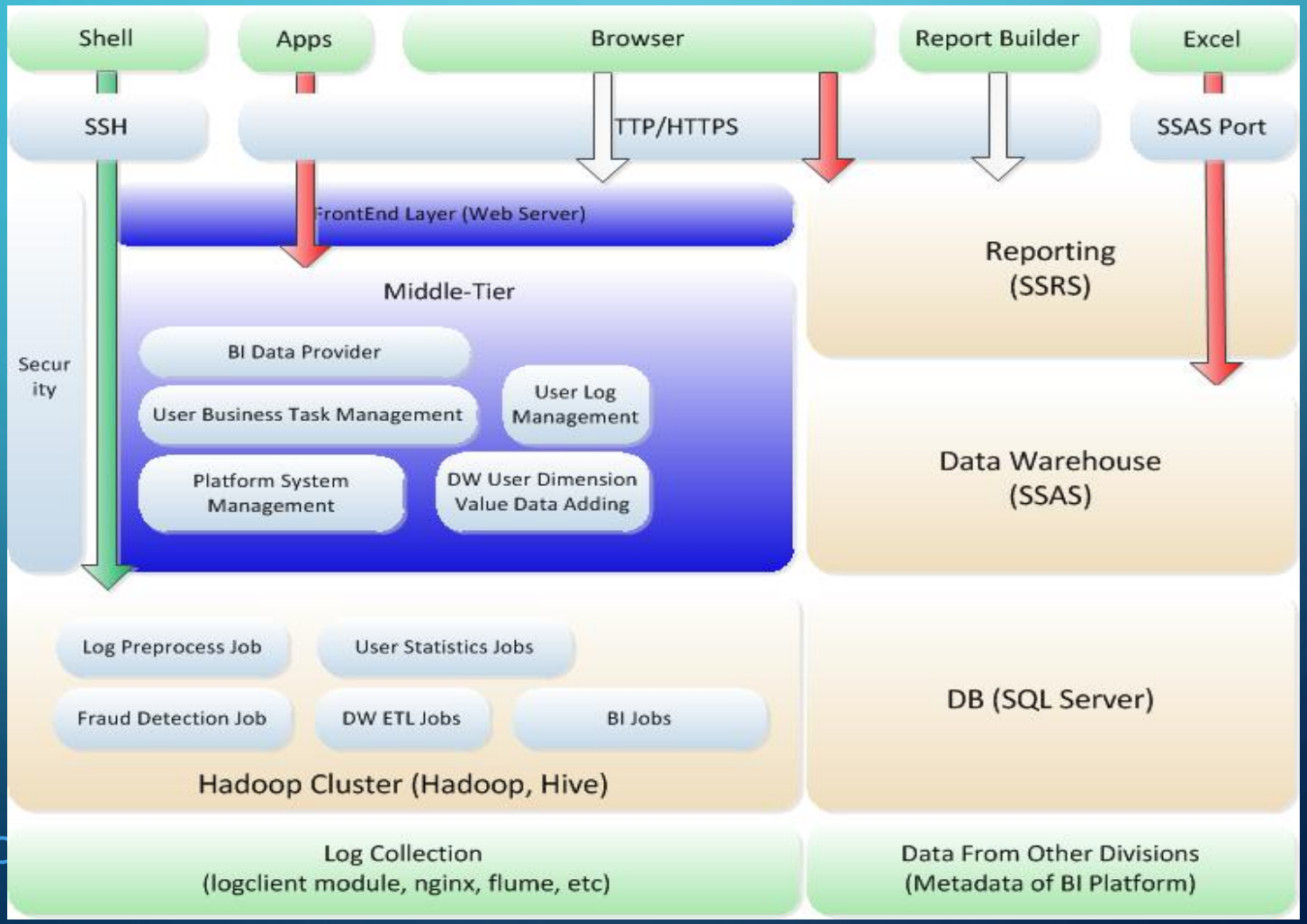
大数据

灵活性

涵盖的业务范围

开发成本

PPBIP – PPTV的大数据平台



Hadoop Ecosystem Map

1

2

Workflow

10

12

Support

More High Level Interfaces

13

Unstructured Data

6

Flume

Scribe

Engine + Logic

5

High Level Interfaces

8

JAQL

4

9

File system

3

7

Structured Data

hiho

Sqoop

RDBMS

OLTP

14

Monitor/manage Hadoop ecosystem

11

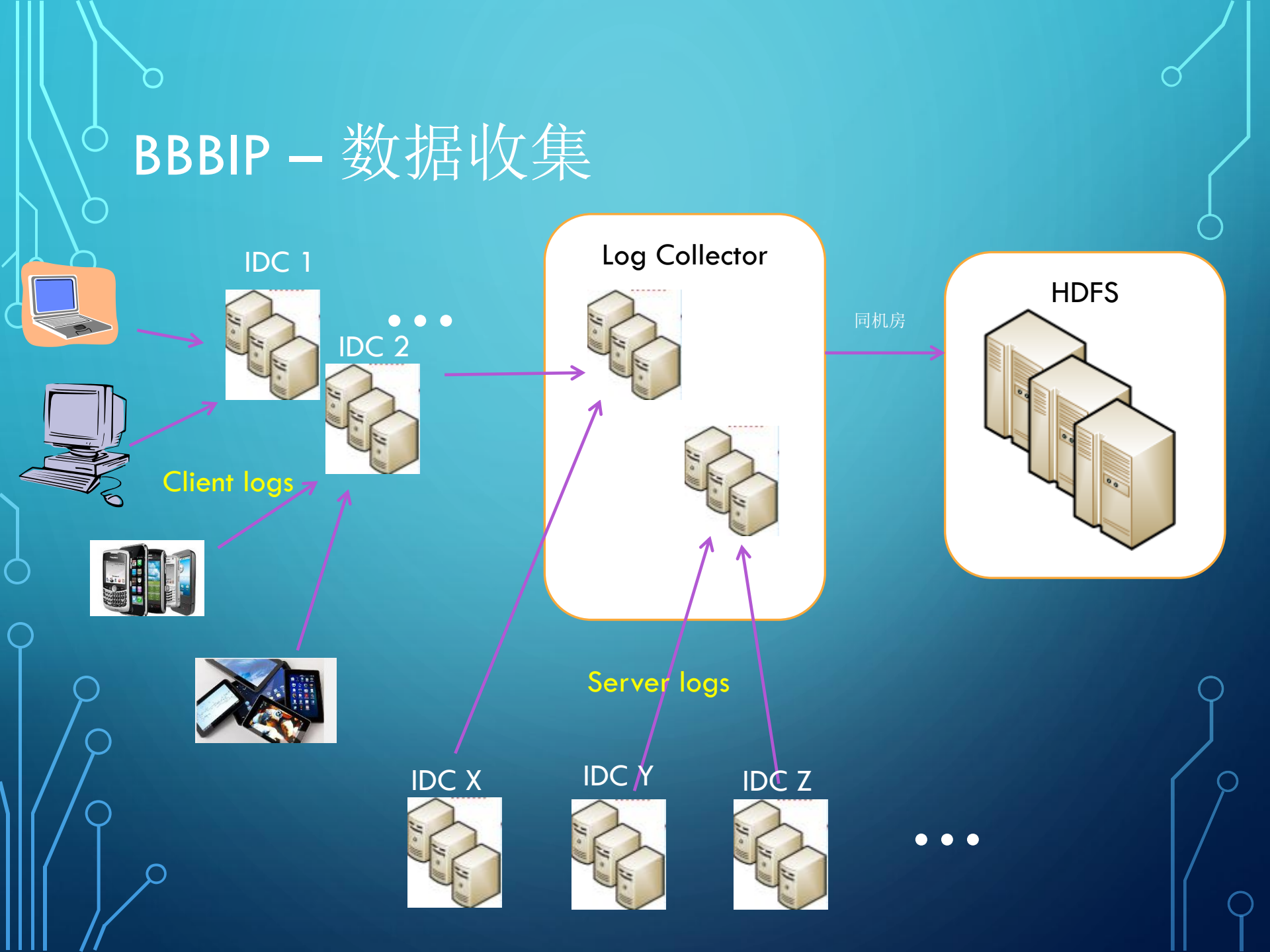
Hue

Karmasphere

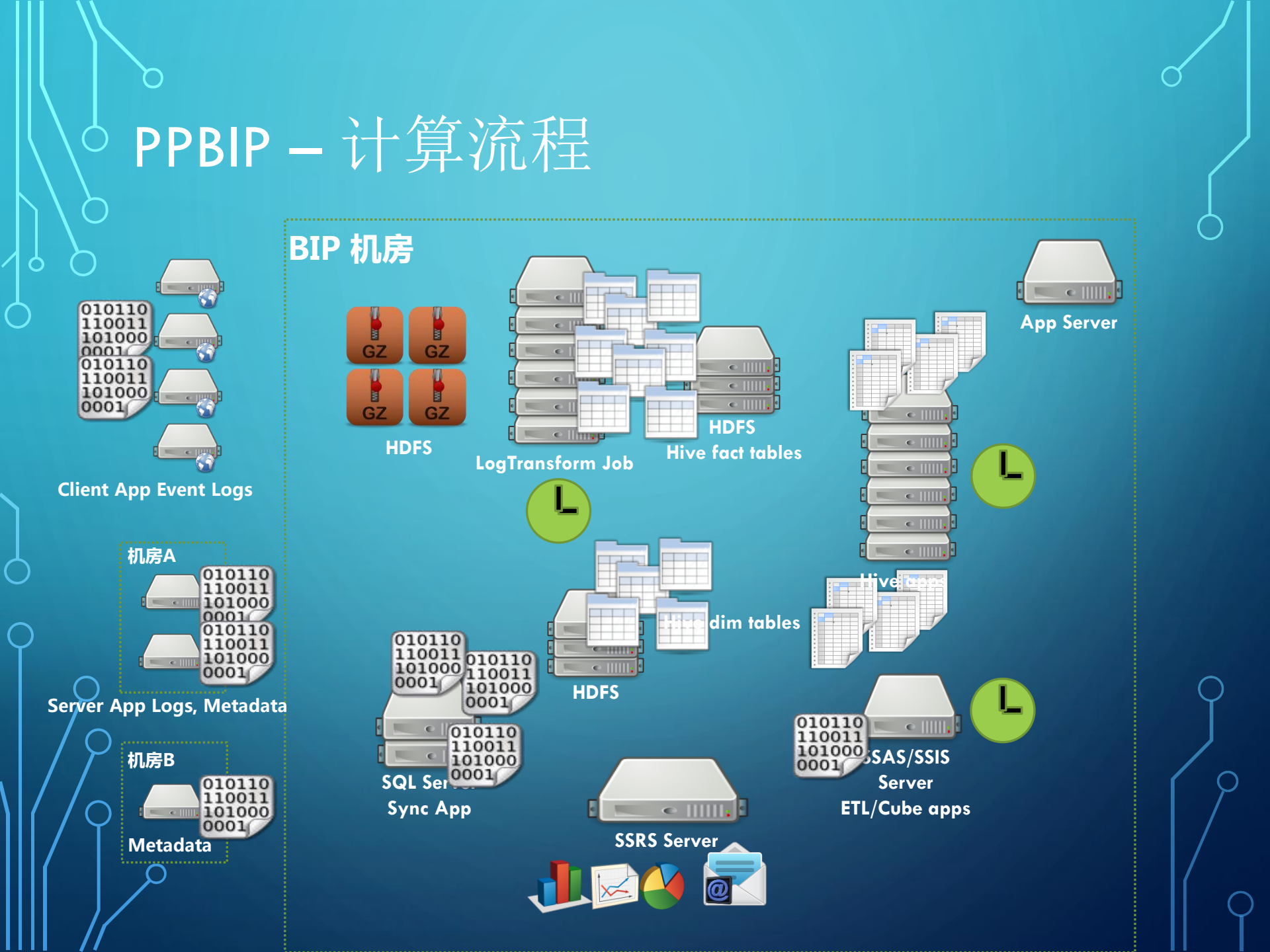
Ganglia

Java Applications

BBBIP – 数据收集


IDC 1
IDC 2
Client logs
Log Collector
同机房
HDFS
Server logs
IDC X
IDC Y
IDC Z

PPBIP – 计算流程

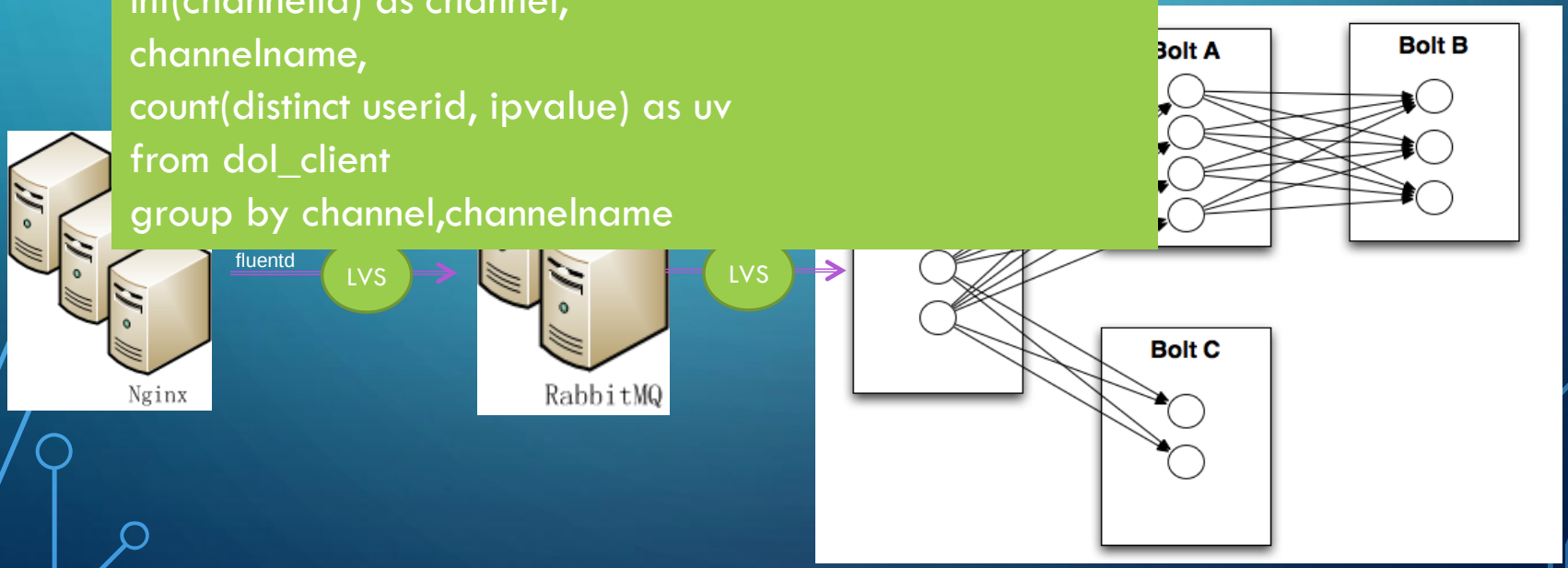

BIP 机房
Client App Event Logs
机房A
Server App Logs, Metadata
机房B
Metadata
GZ
GZ
GZ
GZ
HDFS
LogTransform Job
HDFS
Hive fact tables
App Server
HDFS
SQL Ser
Sync App
SSRS Server
SSAS/SSIS
Server
ETL/Cube apps

PPBIP – 规模

- Hadoop 集群
  - 节点: 60=>120=>200
  - 存储空间: 4PB 容量
- 日负载
  - 新增数据： 30+TB (来自 200+ IDC和数千万客户端)
  - Hadoop Job: 4万
  - 新增Hive 记录: 300亿
  - 读字节数： 2.5PB
- 数据仓库
  - Cubes: 200+
  - Dimensions: 100+
  - Measure: 50+

基于STORM的实时统计系统


```
withperiod 5m
select
dt(5*60) as dt,
int(channelid) as channel,
channelname,
count(distinct userid, ipvalue) as uv
from dol_client
group by channel,channelname
```



应用设计


手工查询
自助报表
全局报表
数据仓库
APIs 调用
系统集成
并行机器集群 + 业务计算程序 (批处理，实时)
每天数据 (用户端, 服务器)
同步元数据

应用场景 – BI报表，CUBE

- Hadoop ETL=>Data warehouse
- Daily job

应用场景 – AD HOC查询，自动化报表

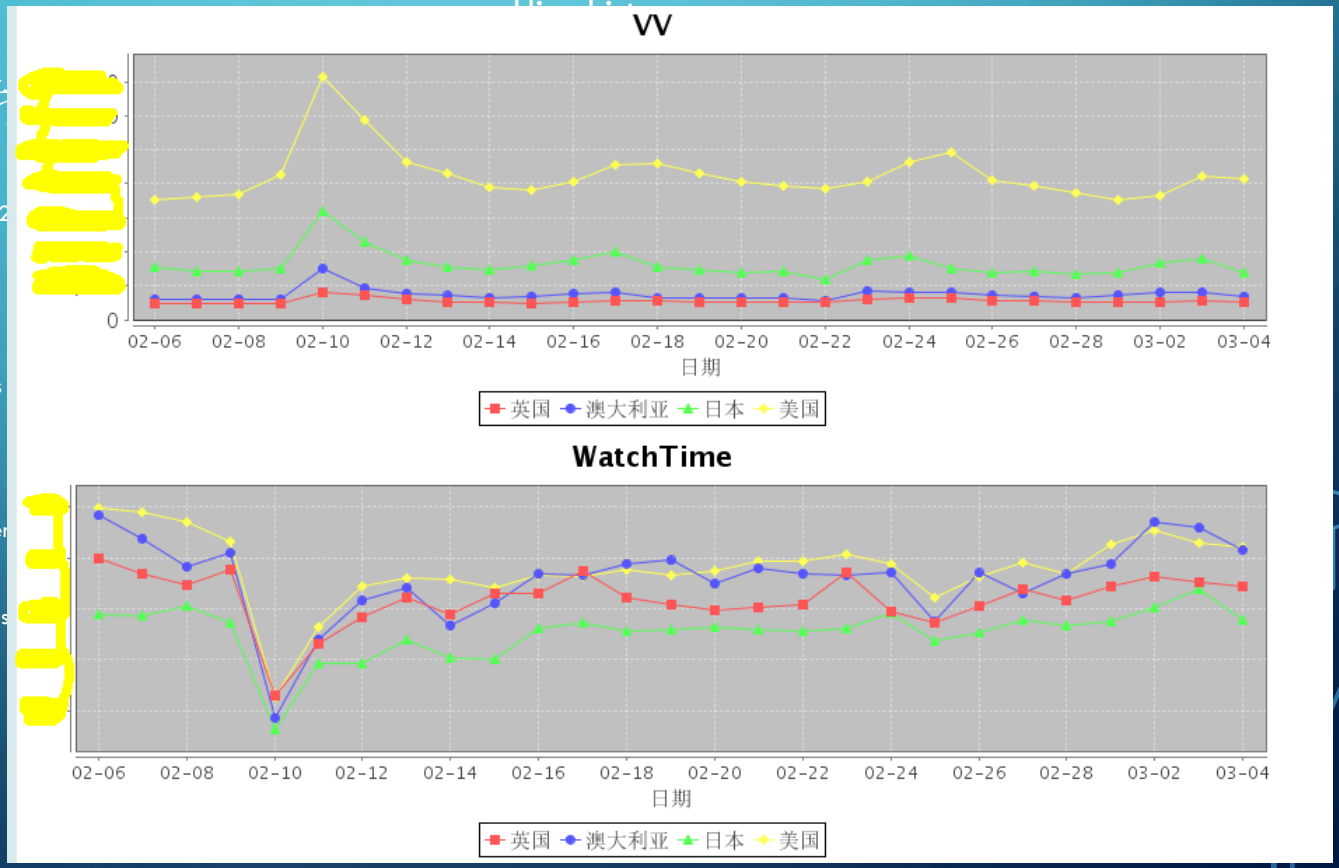
- 使用HQL直接查询Hive
- 每天数千次
- 平均每分钟执

[cloud@SHBNJ-BIPHIVE-HADOOP-20-83 ~]$ hive -f yunjin/tempquery.hql > yunjin/tempquery-result.txt

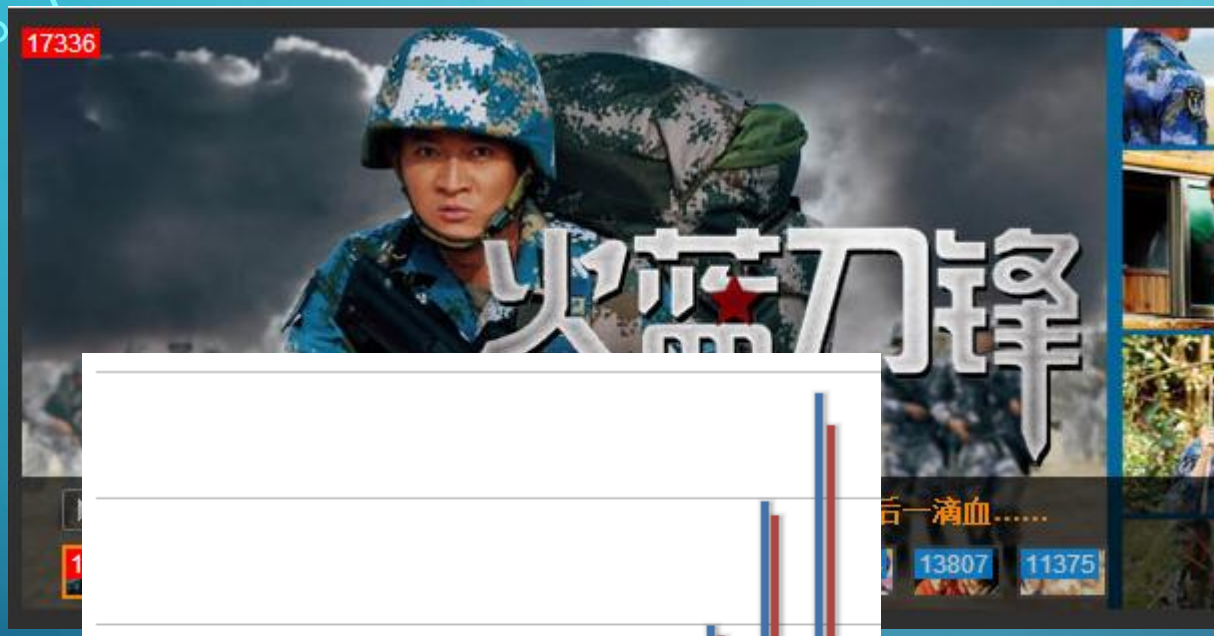
```
[cloud@SHBNJ-BIPHIVE-HADOOP-2
select   dt,
         CP,
       platform,
         sum(vv_play) as vv,
         sum(uv_play) as uv,
                        sum(channel) as
from
(
         select /*+mapjoin(t2)*/ dt,
                  'client' as platform,
                                  t2.conter

                  count(1) as VV_Play,
                  count(distinct userid) as
           from cl_play t1

                  ...
```



应用场景 – 实时流量统计


17336
13807
11375

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

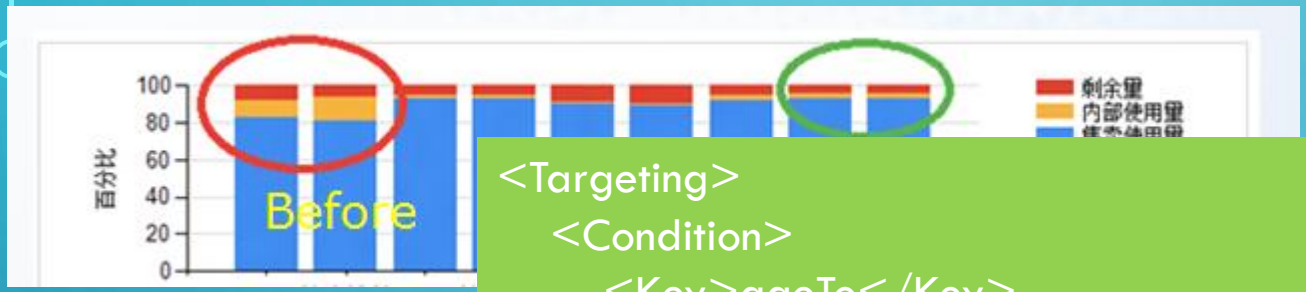

大家都在看
1
直播港澳台
深圳卫视[高清]
7219人在看
2
全星全益
安徽卫视
4874人在看
3
金牌调解
江西卫视
4160人在看
4
电视剧:我是特种兵之利刃出鞘
山东卫视[高清]
2178人在看
5
财富相对论
宁夏卫视
1293人在看
6
电视剧:老有所依
天津卫视[高清]
1106人在看

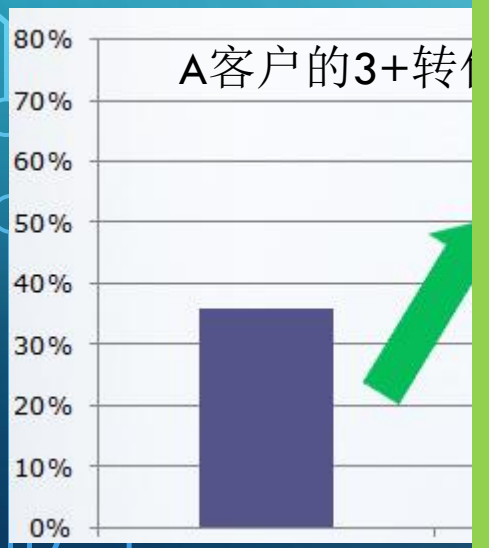
应用场景 – 个性化推荐


为您推荐
1 2 3 4
[VIP]普林斯顿公开课
因为您收藏了 [VIP]巴黎高等商学院：直观的智慧
善良的男人
因为您看过 想你
MayQueen
因为您看过 想你
[VIP]普林斯顿公开课
因为您收藏了 [VIP]耶鲁大学-全球人口增长问题
浪漫满屋2
因为您看过 想你


为您推荐
1 2 3 4
快乐大本营
我看过 不喜欢
因为您看过 康熙来了
赛德克巴莱(下)彩虹桥
我看过 不喜欢
因为您看过 赛德克巴莱(上)太阳旗
23rd GOLDEN MELODY AWARD
最佳作曲人
我看过 不喜欢
因为您看过 2012诺亚方舟(第23届流行音乐金曲奖)
跟踪孔令学
我看过 不喜欢
因为您收藏了 幸福盛宴
BTV档案-刘江楼决胜金
我看过 不喜欢
因为您看过 BTV档案-血战三湘 浴血衡阳战日寇(上)-20120926

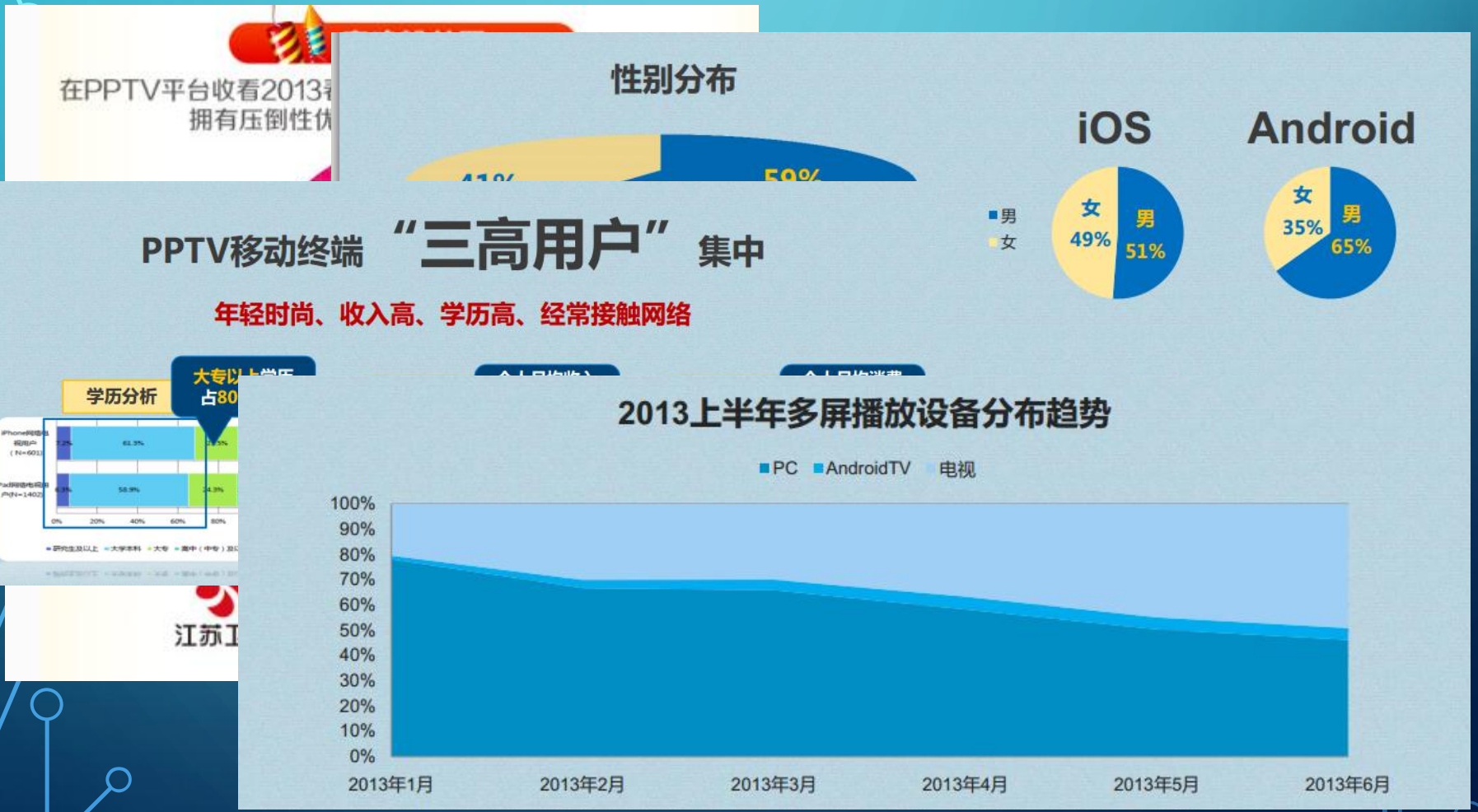
应用场景 – 广告投放策略优化






```
<Targeting>
    <Condition>
        <Key>ageTo</Key>
        <NumericalValueRange>
            <GreaterThan>20</GreaterThan>
        </NumericalValueRange>
    </Condition>
    <Condition>
        <Key>gender</Key>
        <EnumValueRange>
            <EqualsTo>20</EqualsTo>
        </EnumValueRange>
    </Condition>
</Targeting>
```

应用场景 – PP指数


在PPTV平台收看2013
拥有压倒性优
性别分布
59%
iOS
Android
男
女
女
49%
男
51%
女
35%
男
65%
PPTV移动终端 “三高用户” 集中
年轻时尚、收入高、学历高、经常接触网络
学历分析
大专以上
占80
2013上半年多屏播放设备分布趋势
PC
AndroidTV
电视
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2013年1月
2013年2月
2013年3月
2013年4月
2013年5月
2013年6月

谢谢！