

SACC 第八届中国系统架构师大会
2016 SYSTEM ARCHITECT CONFERENCE CHINA 2016

架构 创新之路

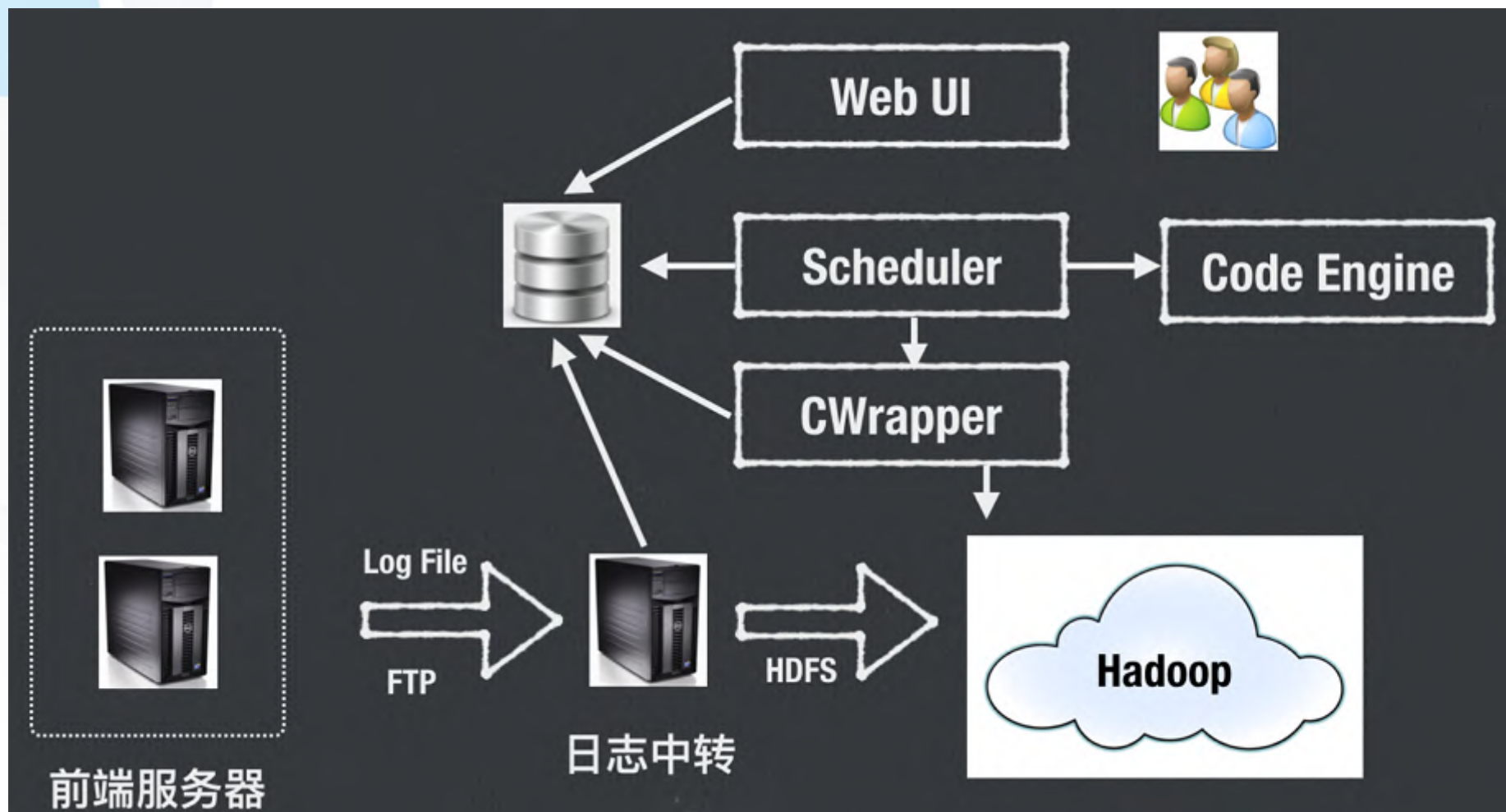
从日志统计到大数据分析

曹犟 @ 神策数据

- 2008：日志统计平台
- 2011：用户数据仓库
- 2013：数据源管理

- 需求响应周期长
- 运维成本高
- 运行速度慢
- 员工个人成长受限

基于 Hadoop 的日志统计平台



[+ 添加数据源](#) [隐藏高级选项](#)

1. 数据源: 统计输出 {0, 1 day} [删除](#)

统计编辑区

[+ 计数](#)

[+ Top统计](#)

[+ 去重计数](#)

[+ 数据过滤](#)

[+ 更多](#) [全部收起](#)

[帮助](#)

1. 去重计数

[复制](#) [删除](#)

*总数变量名 *中文名 [使用扩展属性](#)

*去重字段:

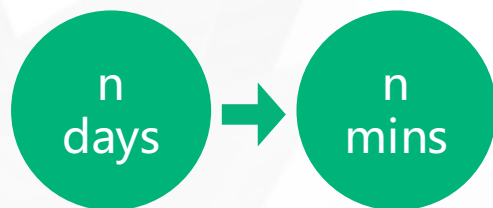
[+ 添加限制条件](#)

[使用公共限制条件组](#)

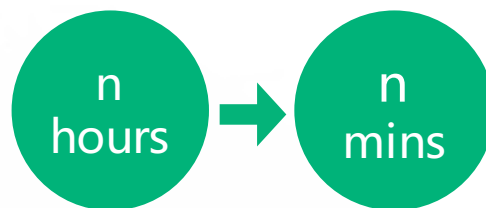
- | | | | | | |
|----|---|---|------------------------------------|--|----------------------|
| 1. | <input type="text" value="_UrlFields"/> | <input pid"]"="" type="text" value="["/> | <input type="text" value="字符串等于"/> | <input type="text" value="102"/> | 且 删除 |
| 2. | <input type="text" value="_UrlFields"/> | <input type="text" type"]"="" value="["/> | <input type="text" value="字符串等于"/> | <input type="text" value="2070"/> | 且 删除 |
| 3. | <input type="text" value="_UrlFields"/> | <input area"]"="" type="text" value="["/> | <input type="text" value="字符串等于"/> | <input type="text" value="na"/> | 且 删除 |
| 4. | <input type="text" value="_UrlFields"/> | <input page"]"="" type="text" value="["/> | <input type="text" value="字符串等于"/> | <input type="text" value="spring-sale"/> | 且 删除 |
| 5. | <input type="text" value="_UrlFields"/> | <input action"]"="" type="text" value="["/> | <input type="text" value="字符串等于"/> | <input type="text" value="pv"/> | 且 删除 |

为该统计添加注释(选填):

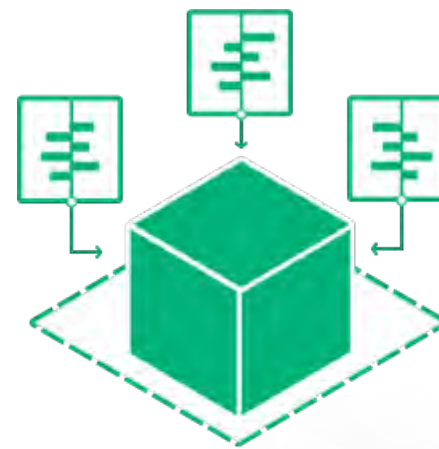
```
DQuery::input()  
->select(array('_Url'))  
->group('_Url')  
->countEach('_Url', '_UrlCnt')  
->outputAsFile('pv_per_url', '分URL访问统计');
```



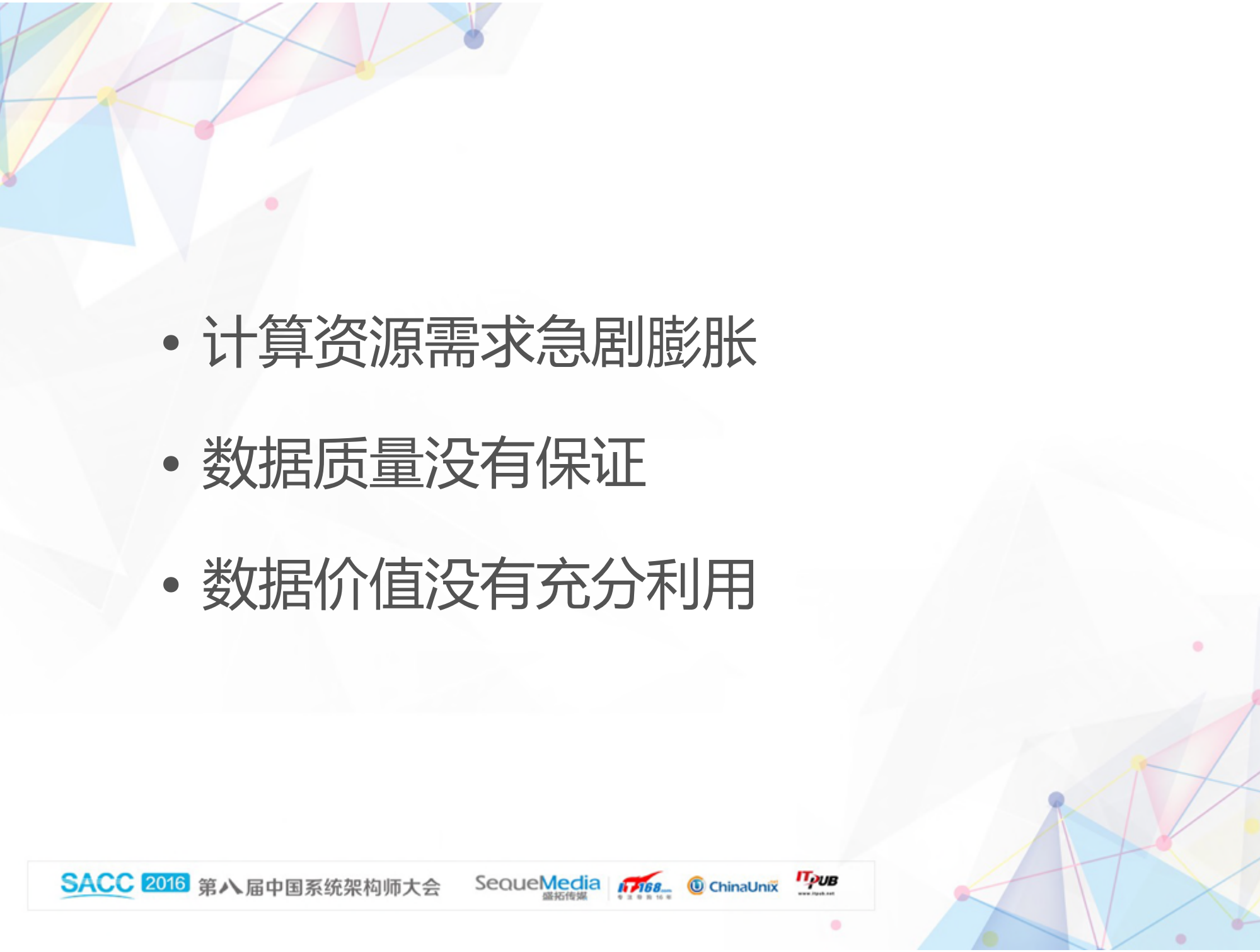
开发周期：
天级 到 分钟级



运行时间：
小时级 到 分钟级



经过1年半的时间，
整个公司的统计任务
都统一到了平台

- 
- 计算资源需求急剧膨胀
 - 数据质量没有保证
 - 数据价值没有充分利用

统一的用户数据仓库

Insight

DataMart

User Data Warehouse

网页搜索

凤巢

网盟

知道

百科

结构化数据

网页搜索

凤巢

网盟

知道

百科

文本日志

全公司所有业务线的用户行为数据统一到一张表，通过用户 ID 可以访问到用户在全百度上的所有行为

用户ID	事件类型	时间	国家	省份	URL
ID01	注册	...	...	...	...
ID02	登陆	...	...	...	...
ID03	搜索	...	...	...	...





数据源混乱



入库周期长



数据时效性低

直接从数据源头治理

非结构日志源

- 格式变更影响下游
- 数据解析效率低
- 数据内容不可理解

结构化日志源 (Google Protocol Buffer) :

- 格式变更向前向后兼容
- 数据解析效率高
- 数据带有Schema
- 节省存储和带宽



日志源的结构化、
元数据审核管理



传输实时化



查询引擎直接可分
析源结构化数据

- 数据源统一管理，覆盖数万台服务器
- 数据源头产生，既可实时 SQL 分析

- 数据源很重要
- Event 模型很有效

- 数据采集手段在增强
- 大数据计算能力的软硬件环境在成熟
- 竞争加剧，整个行业的数据意识在提升

- 人力成本
- 时间成本

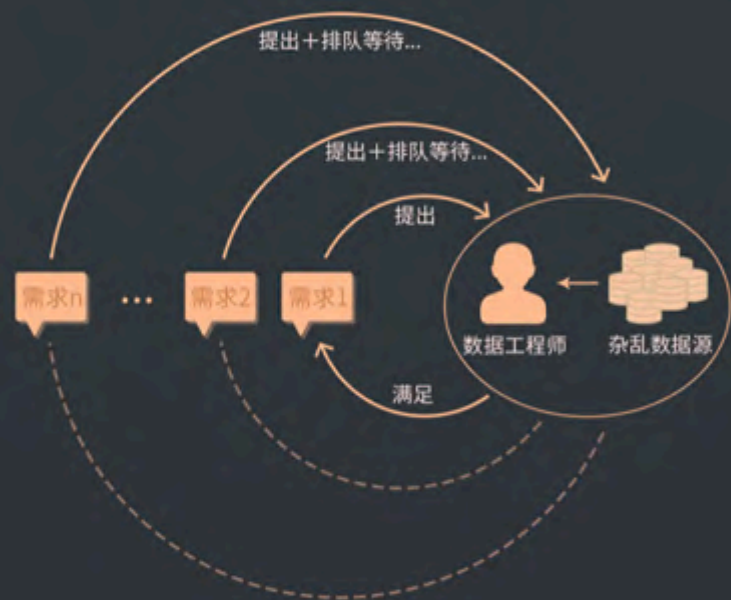
- 数据采集能力不足
- 分析能力比较基础
- 数据资产顾虑

神策分析

可以私有化部署的 用户行为分析产品



现实中的数据驱动



时间都等没了！
时机都错过了！

理想中的数据驱动



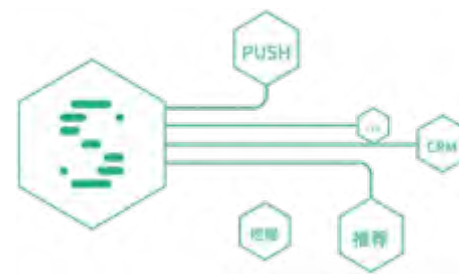
业务人员自助满足需求，
把时机握在手中！



私有化部署



全端数据接入

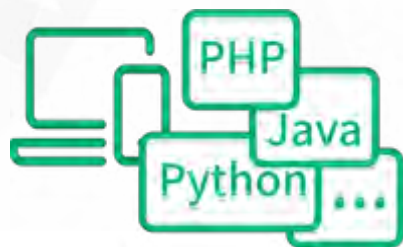


PaaS 平台

全端数据接入



Web	添加埋点	禁用埋点
iOS	编辑埋点	删除埋点
Android	:	:
Server	:	:



代码埋点



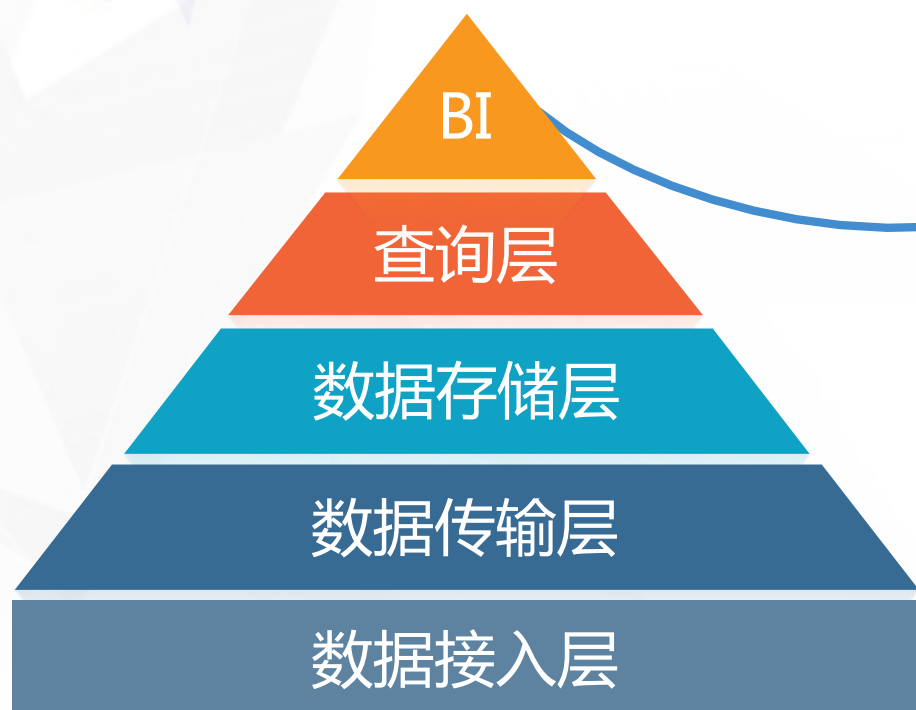
可视化埋点



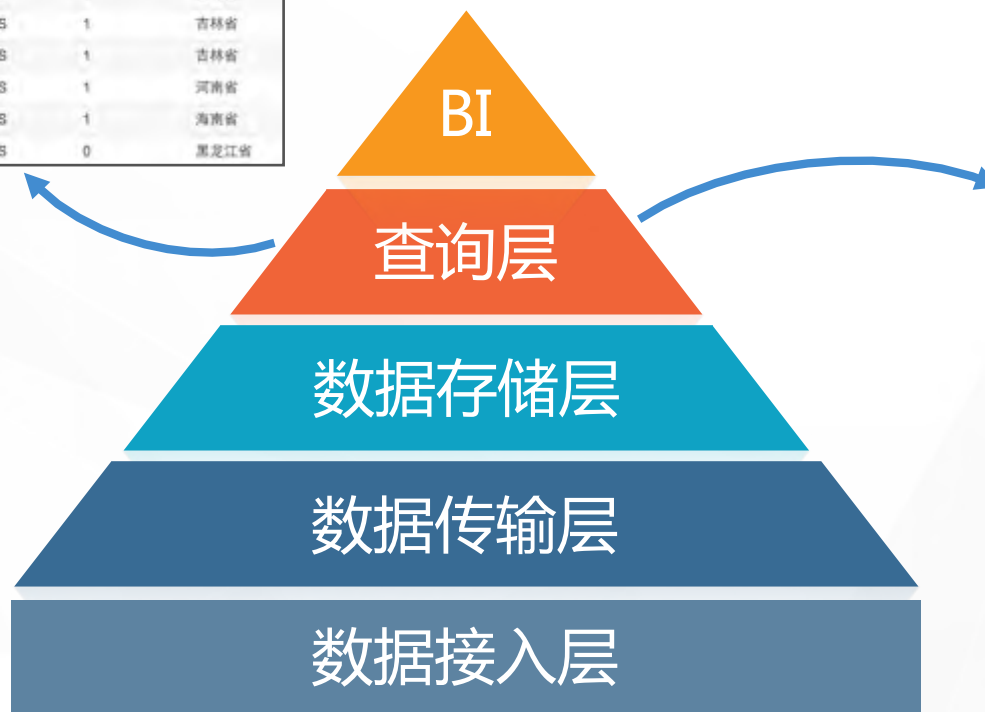
BatchImporter LogAgent FormatImporter

导入辅助工具

PaaS 平台



event	distinct_id	\$os	\$wifi	\$province
ViewProduct	649585174d6e8b36	iOS	1	甘肃省
ViewProduct	5b166a4a059679b0	iOS	1	吉林省
ViewProduct	1812549065	iOS	1	吉林省
ViewProduct	1812549065	iOS	1	吉林省
ViewProduct	1812549065	iOS	1	吉林省
ViewProduct	723143964	iOS	1	河南省
ViewProduct	2011903106	iOS	1	海南省
ViewProduct	712991814	iOS	0	黑龙江省

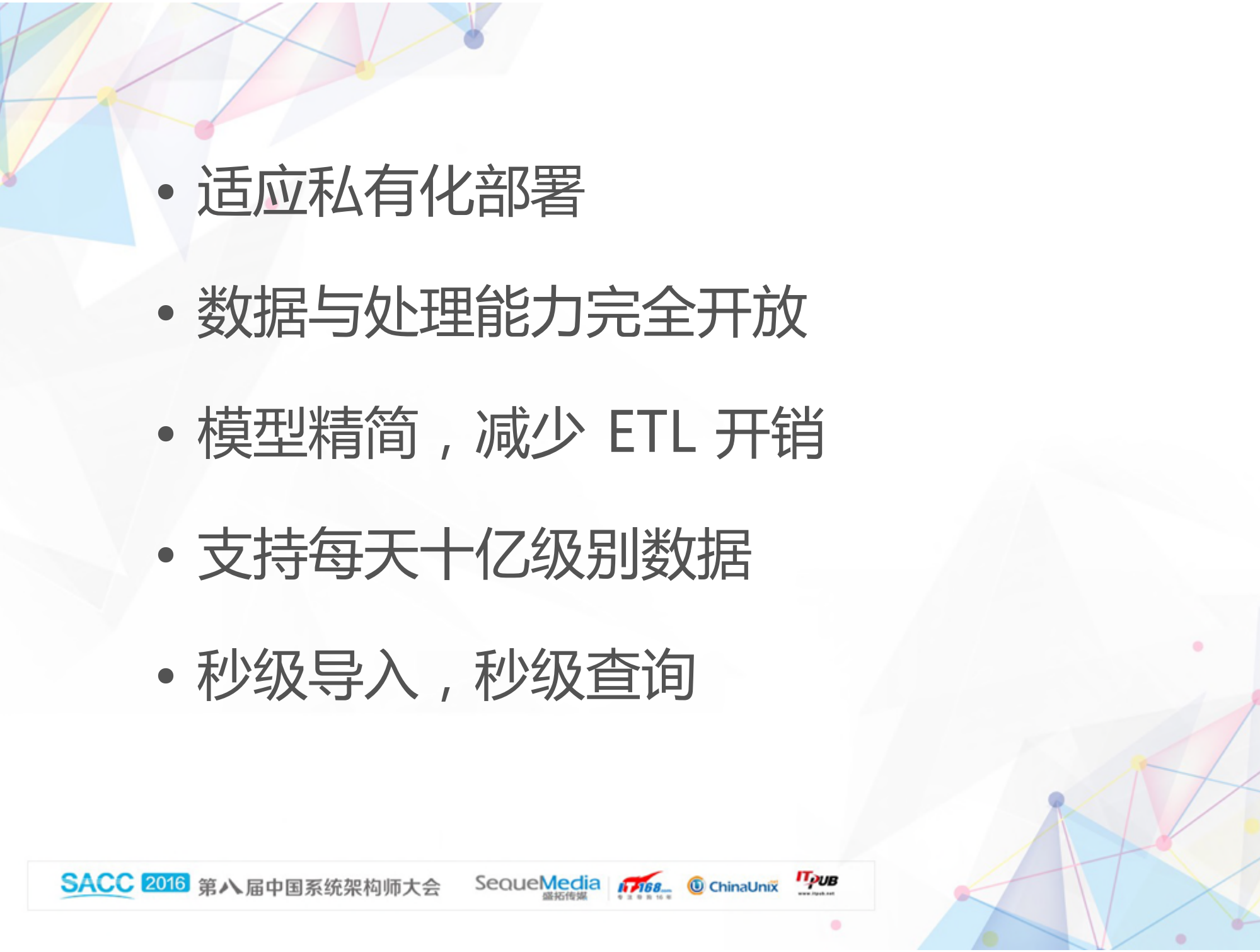


```

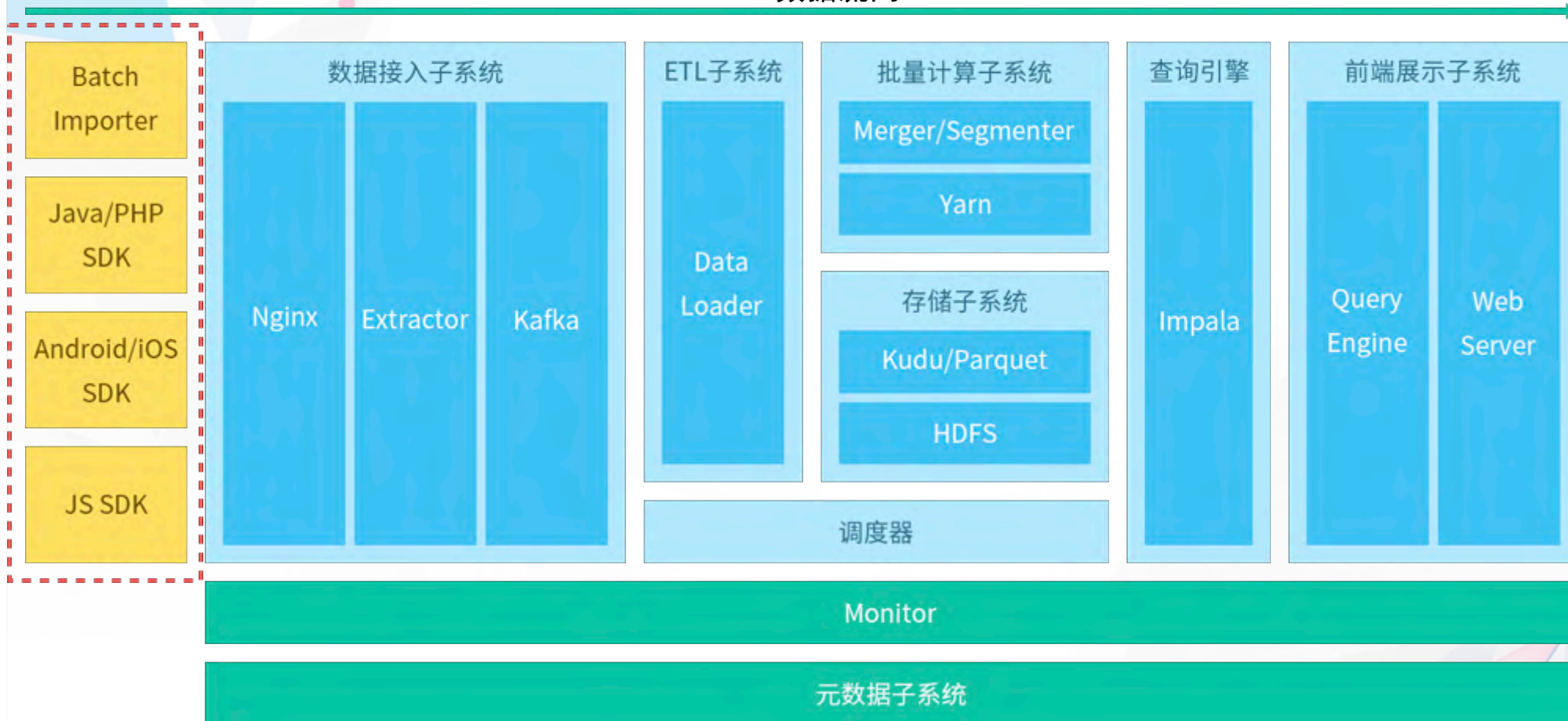
01 {
02   "by_fields": [
03     "event.BuyGold.$wifi",
04     "user.Gender"
05   ],
06   "series": [
07     "2015-06-06 00:00:00",
08     "2015-06-07 00:00:00",
09     "2015-06-08 00:00:00",
10     "2015-06-09 00:00:00",
11     "2015-06-10 00:00:00",
12     "2015-06-11 00:00:00",
13     "2015-06-12 00:00:00",
14     "2015-06-13 00:00:00"
15   ],
16   "rows": [
17     {
18       "values": [
19         86015190,
20         81952822,
21         87961512,
22         81244661,
23         90635719,
24         87182989,
25         86979903,
26         84155806
27       ],

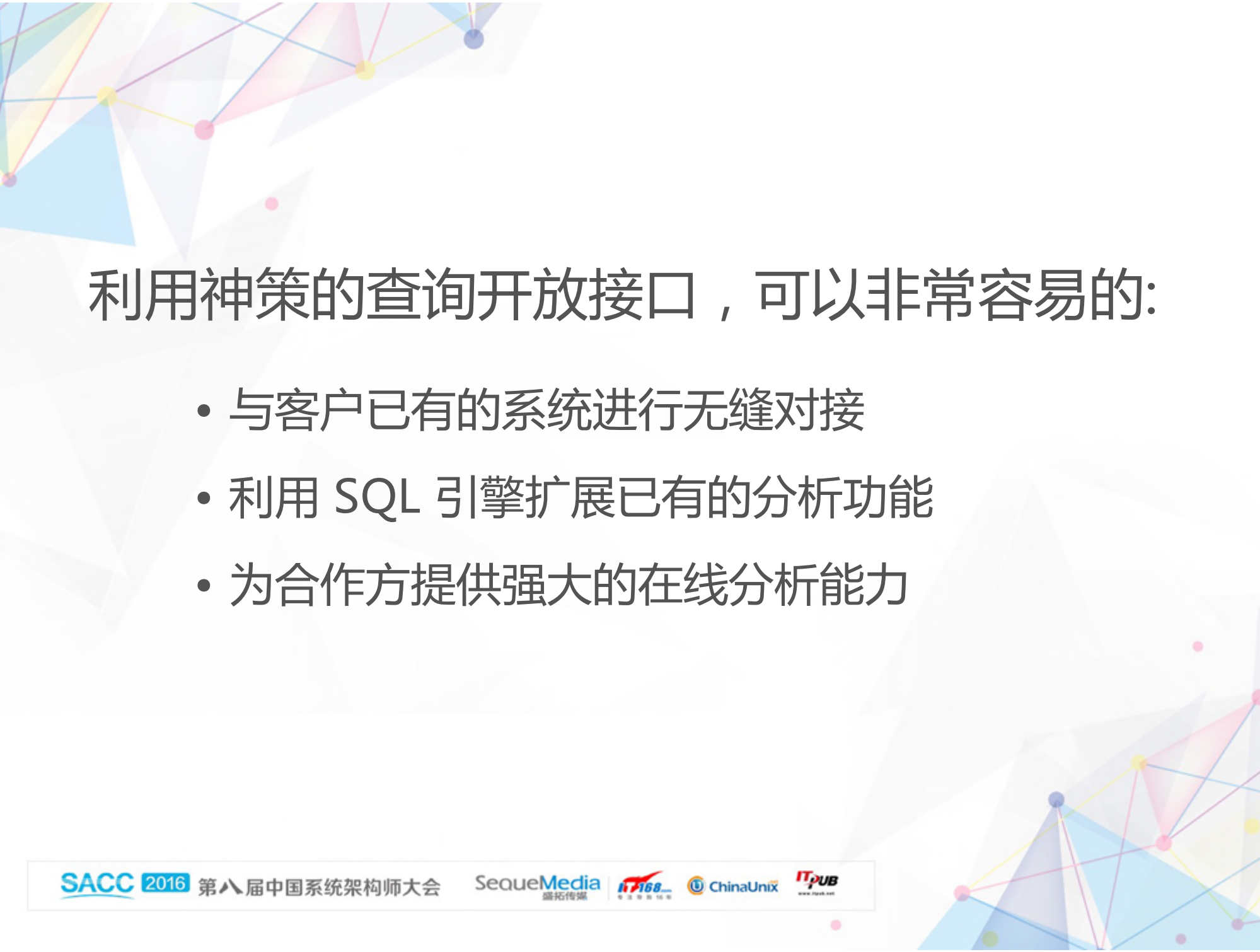
```

架构实现

- 
- 适应私有化部署
 - 数据与处理能力完全开放
 - 模型精简，减少 ETL 开销
 - 支持每天十亿级别数据
 - 秒级导入，秒级查询

数据流向





利用神策的查询开放接口，可以非常容易的：

- 与客户已有的系统进行无缝对接
- 利用 SQL 引擎扩展已有的分析功能
- 为合作方提供强大的在线分析能力

在线分析

搜索优化

个性化推荐

用户画像

精准广告

文本挖掘

反作弊

100%



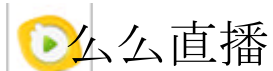
金融



电商



视频



企业服务



在线教育



在线医疗



新媒体



健身



游戏



互联网 +



其他



神策分析，帮你实现数据驱动！



THANKS

SequeMedia
盛拓传媒

IT168.com
专注 导购 10 年

ChinaUnix

ITPUB
www.itpub.net