



第九届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2017

# 迎接在线化与开放化分析时代

Welcome to the Online and Open Big Data Analytics Era

离哲 ( @flyinweb ) 资深技术专家

SACC  
2017

北京·新云南皇冠假日酒店



1

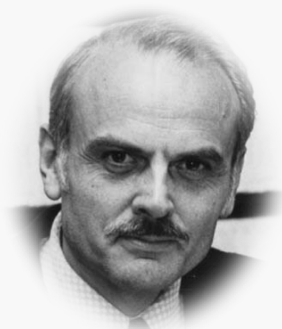
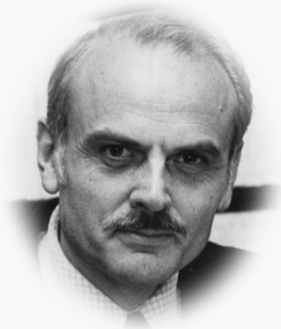
发展趋势

2

典型案例

3

解决方案



1970

1993

2005

Relational  
[ SQL+OLTP ]

12 Rules  
[ OLAP ]

GFS+MR  
[ Big Data ]

More...

Autonomous

HTAP

Federation

Cube

Text

Vector

Graph

Time Series

# 5M-More Accessible

## 内部服务



~100

VS

## 外部产品



100,000+



~1s +

# 5M-More Data



抽样 VS 全量 *PB +*

多数据源(DB/HD/HD/File/..) 毫秒级

多场景 (Table/Graph/GIS/TS/Matrix..)



场景优化



混合云



# 5M-More Realtime



批量装载 VS 实时写入  $10,000,000/s +$

预建模 VS 即时  $10,000 QPS +$

自服务



人人都是分析师

# 5M-More Action



**Insight**  
5000+

**营销管理  
(VCRM...)**  
10000+

**安全风险**  
1M+

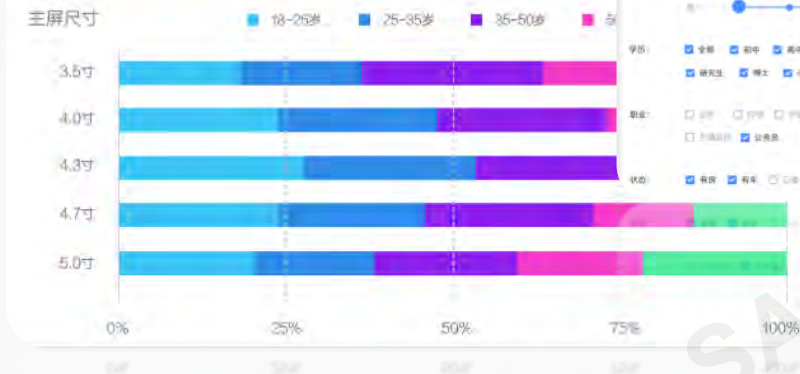
**预测**  
100K+

**推荐**  
1M+



# 电商业务

时间: 2013.9.20-2014.3.20 人群: 男性  
行列信息: “手机主屏尺寸”和“年龄”的关系



## CRM——洞察用户

用户属性多样化：几十甚至数百个用户标签

筛选条件多样化：“买了又买”，“买了又看”，“看了不买” ...

洞察指标多样化：性别分布、浏览次数、城市分布....

# 电商业务



## 经营分析——洞察经营情况

维度多：品类、品牌、产品、型号

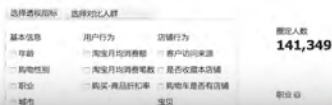
周期长：半年、一年、两年...

数据实时性：半个小时内的数据波动



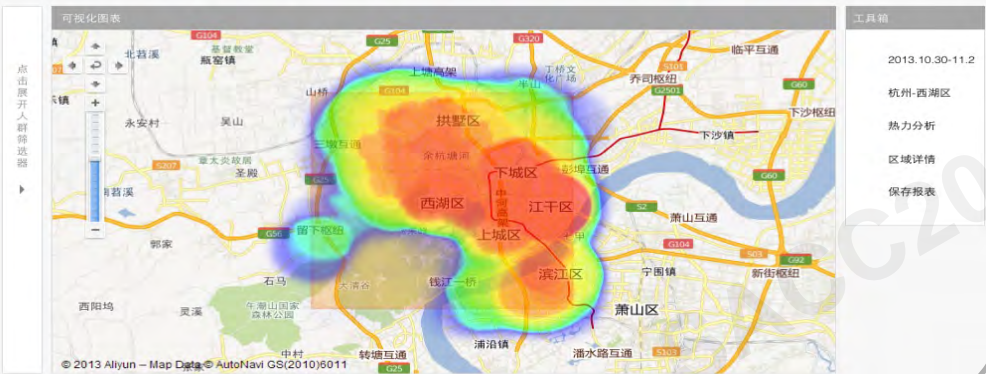
# 营销业务

营销人群1

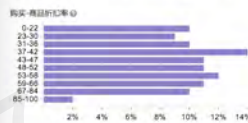


全景智慧 relaxkaka. 退出 首页 使用教程 我的工作台 我的报表 关于我们

趋势分析 地理分析 频率分布 交叉分析 竞争网络 相关分析 对应分析 决策树



城市排名



## ROI 787%



## 达摩盘

阿里妈妈旗下产品

### DMP——精准营销

海量数据：万亿级的互联网行为数据

海量维度：自由组合上千标签，快速圈选人群

复杂功能：人群扩展、自有数据上传...

SACC  
2017

云智未来 9<sup>th</sup>

IT168.com

ChinaUnix

ITPUB

# O2O



**O2O-CRM——数据体现价值**

海量会员：线下门店、餐饮的刷卡客户/预定客户/咨询客户...  
多种来源：交易数据、营销数据、wifi连接...  
实时干预：针对刚刚到访的、刚刚路过的客户进行分析、投放

# 交通

监控大屏

违章识别

套牌车

防控车辆

.....

车辆轨迹

过车统计

精确卡口查询

跟车关联分析

模糊查询

车辆  
维度  
表

过车事实表

车辆  
维度  
表

多维度关联查询

300亿行数据规模

频繁出入

区域碰撞

短时过车

多并发查询



安全

## 智慧搜索系统

一站式多维搜索



## 档案系统

知识图谱



## 时空分析系统

一切皆有迹可循



## 碰撞比对系统

多源极速碰撞



## 网络舆情监测系统

舆情导控



## 标签系统

对象标签化



# 典型架构

## 数据清洗

## 触达引擎

## 应用场景

## 生产业务



实时

ETL

批量

主题库+  
标签化

数据加  
工

标准化

实时  
同步/更新

Insight

预测/报表

规则引擎

规则管理/同步

匹配服务

ID/POI/商品

Open  
API

安全风险控

CRM

营销

敏捷BI

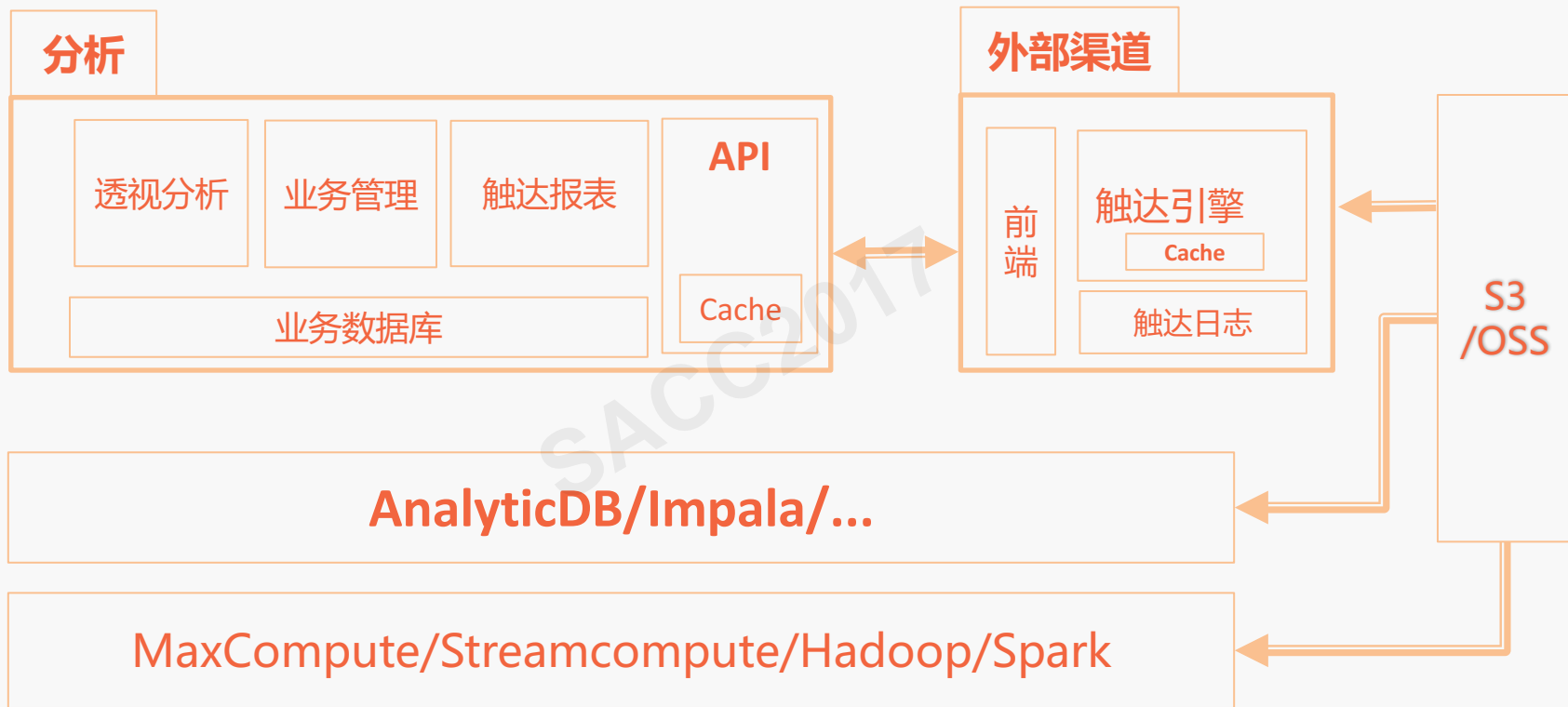
推荐

....

实时  
回流



# 典型架构



# 挑战一：高并发访问-存储性能

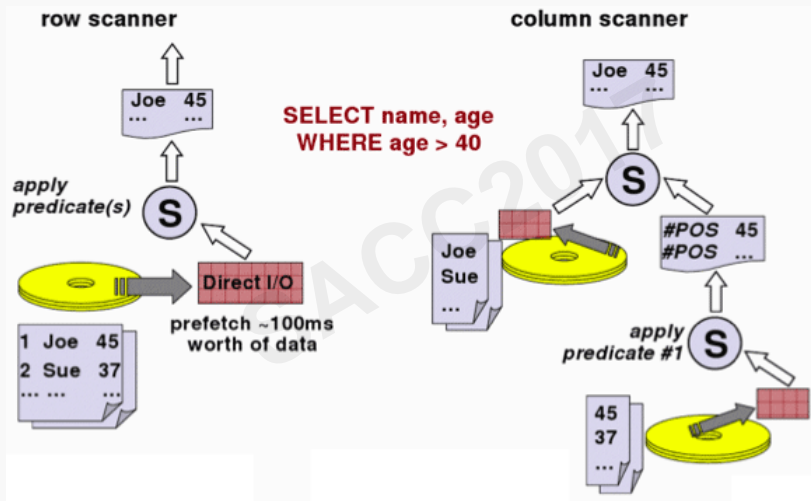
内部测试

20X

列存

3X

压缩



1000X

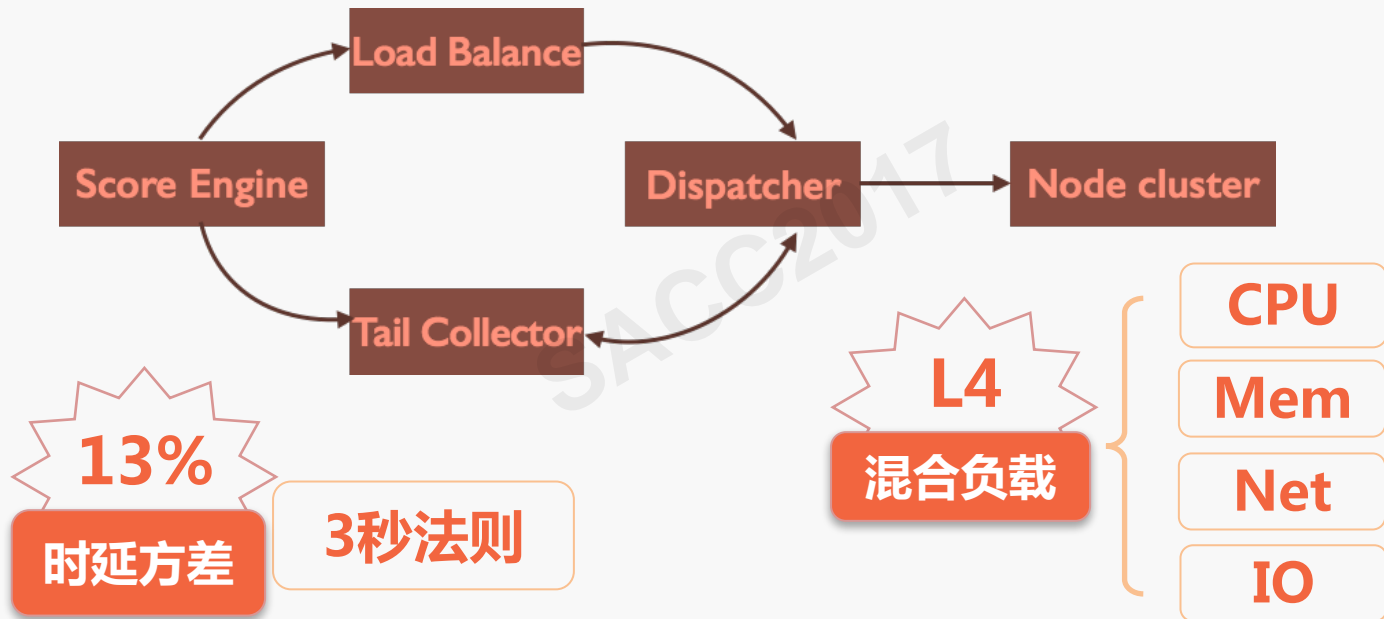
预排序

100X

全索引

# 挑战一：高并发访问-低延时

内部测试



## 挑战二：实时数据&明细查询

*Select TA.\*,TB.\* from..  
Order by ... Desc  
Limit 100*

Drill Down

TOP N

*Insert into TA Values ...*

10,000,000+ /s

SACC2017

## 挑战二：实时数据&明细查询

内部测试



# 挑战三：多集合交并差

内部测试

## 漏斗、标签模型

*Select \* from subquery1*  
*Minus*  
*(Select \* from subquery2*  
*Intersect*  
*Select \* from subquery3*  
*Union*  
*Select \* from subquery4)*

100+

100x

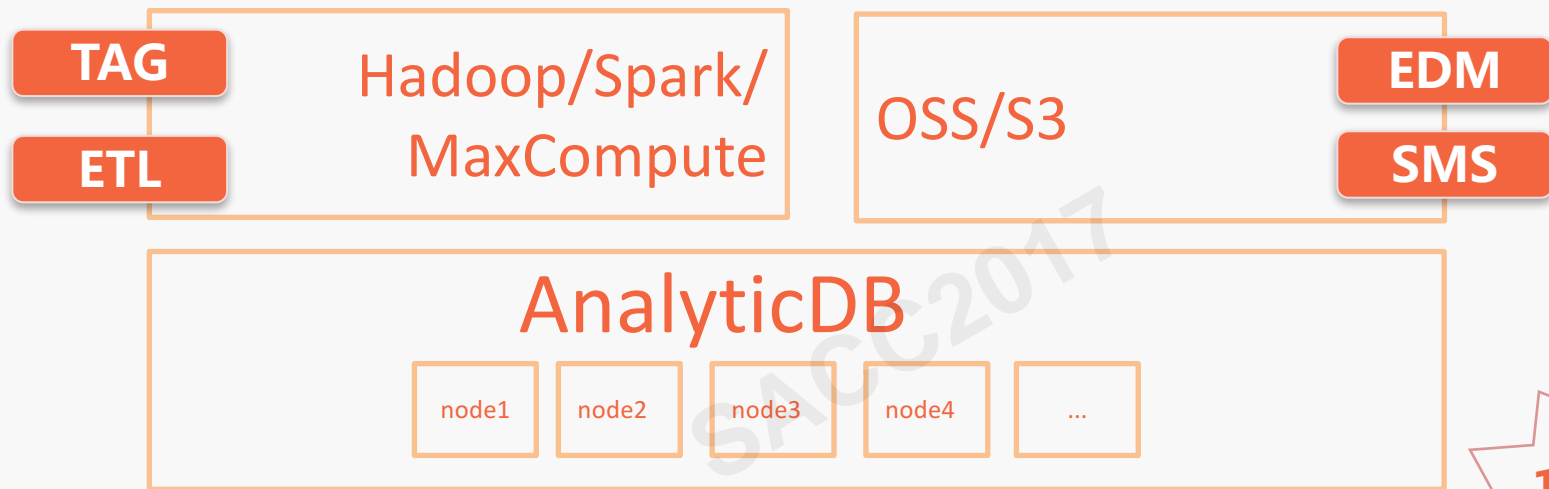
局部并行

表组级多版本

多层缓存

# 挑战四：海量数据实时ETL&同步

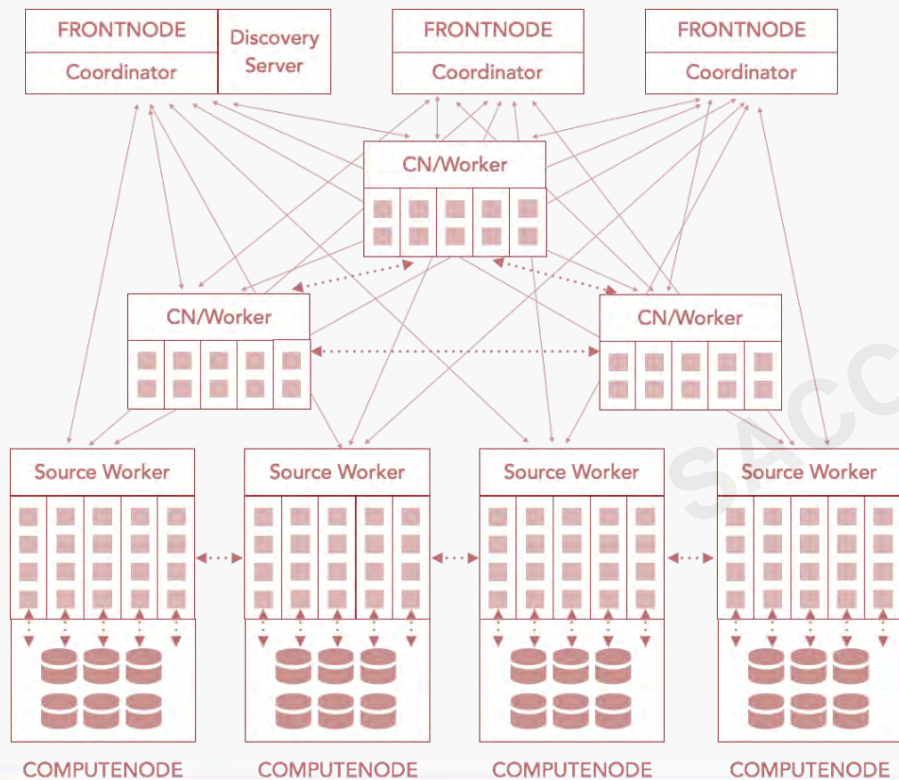
内部测试



150万 exp  
500万 imp

双边并行

# 挑战五：执行引擎



**MPP+DAG双引擎**

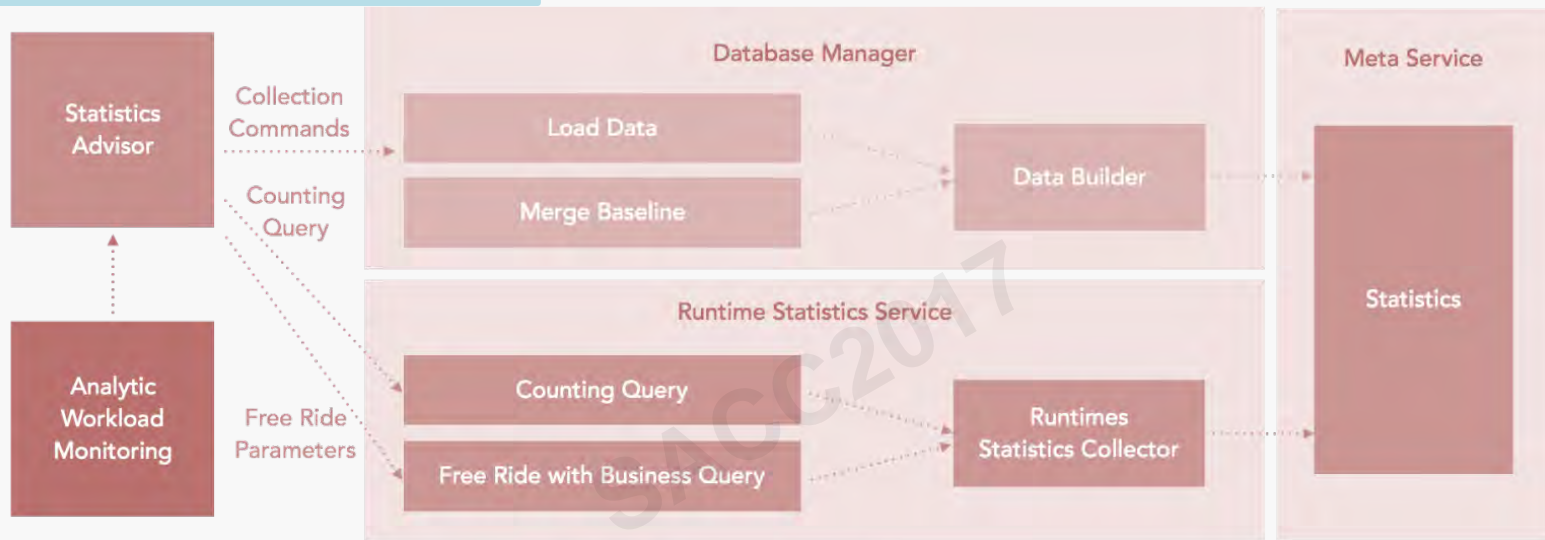
**流式分时执行**

**细粒度资源分级**

**TPC-H/TPC-DS支持**



# 挑战六：优化器



- Volume
- Correlation
- Data Skew (Point)
- Data Skew (Range)

Single Column

- Cardinality - Volume
- Max(HIGH2KEY)/Min(LOW2KEY)
- NDV - Data Skew (Point)
- NNV - Data Skew (Point)
- Frequency - Data Skew (Point)
- Histogram - Data Skew (Range)

Multiple Columns

- Cardinality - Correlation
- Frequency - Data Skew (Point) & Correlation
- Histogram - Data Skew (Range) & Correlation

More



钉钉群



微信群

SACC  
2017

云智未来<sup>9th</sup>



THANKS

The background features a dark, almost black space filled with numerous small, bright blue particles. These particles are arranged in several distinct, curved paths that sweep across the frame from the bottom left towards the top right. A bright, white-to-blue gradient light source is positioned behind the word 'THANKS', creating a lens flare effect and illuminating the nearby particles.