云智未来 9th

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

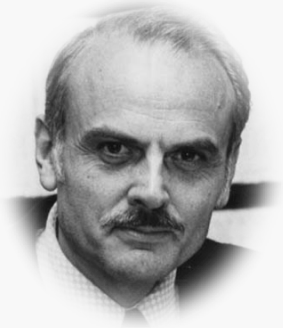
Welcome to the Online and Open Big Data Analytics Era

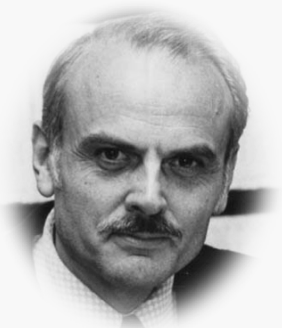
离哲（@flyinweb） 资深技术专家

1 发展趋势

2 典型案例

3 解决方案





Google

1970

Relational
[SQL+OLTP]

1993

12 Rules
[OLAP]

2005

GFS+MR
[Big Data]

More...

Autonomous

HTAP

Federation

Cube

Vector

Text

Time Series

Graph

5M-More Accessible

内部服务 VS 外部产品

~100

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 Ways

数据驱动

数据展现

流程化

业务驱动

数据洞见

敏捷化

漏斗模型/标签模型/...

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+

安全风控
1M+

预测
100K+

推荐
1M+

电商业务

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...

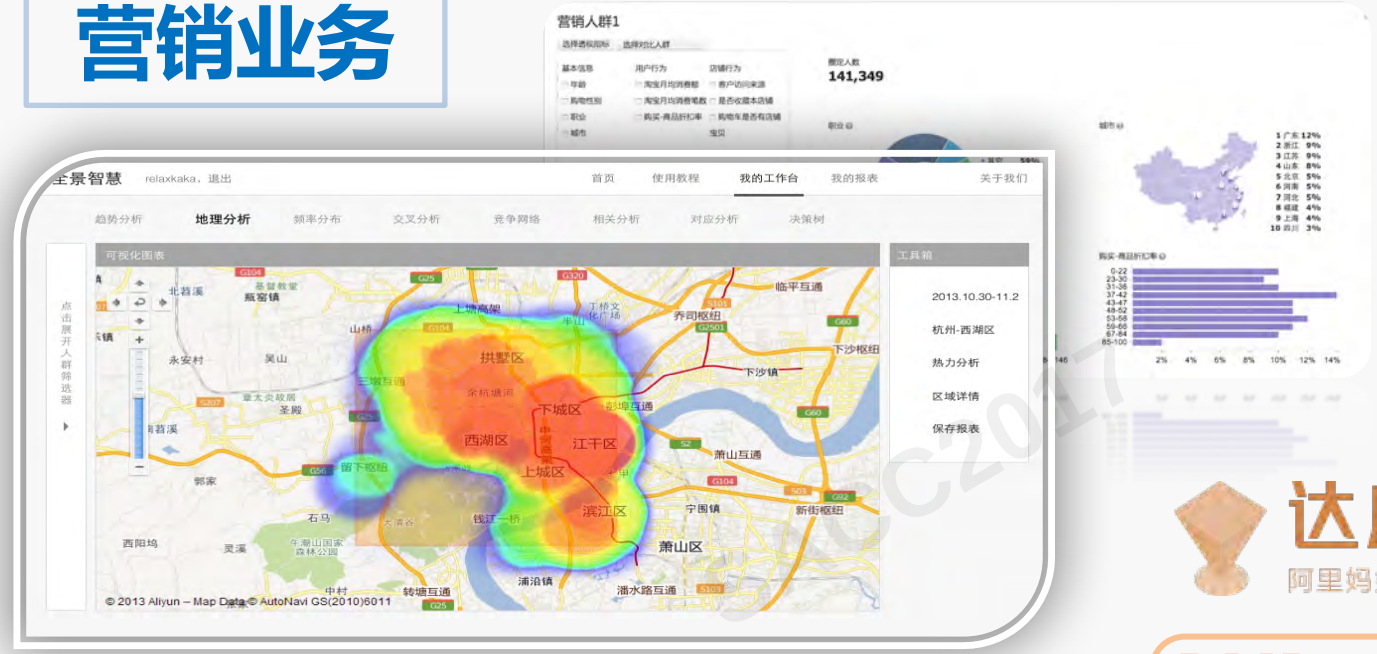
数据实时性：半个小时内的数据波动

电商业务

可视化大屏——洞悉业绩

海量实时数据：每笔订单、每次浏览都要立即呈现
交互式展现：收货地域、分品类...
复杂的计算：top排行、区域热点...

营销业务



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...

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　……

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

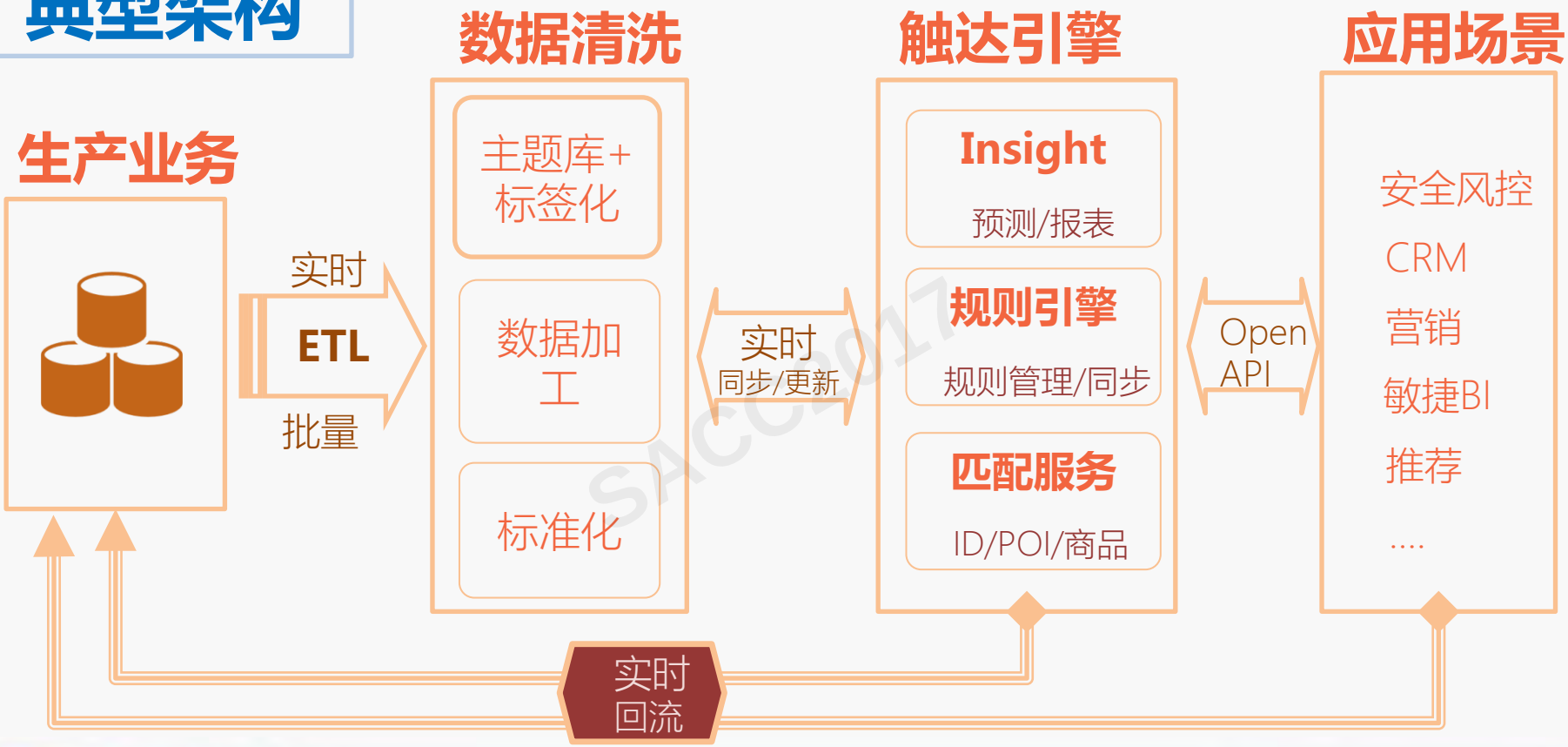
网络舆情监测系统

舆情导控

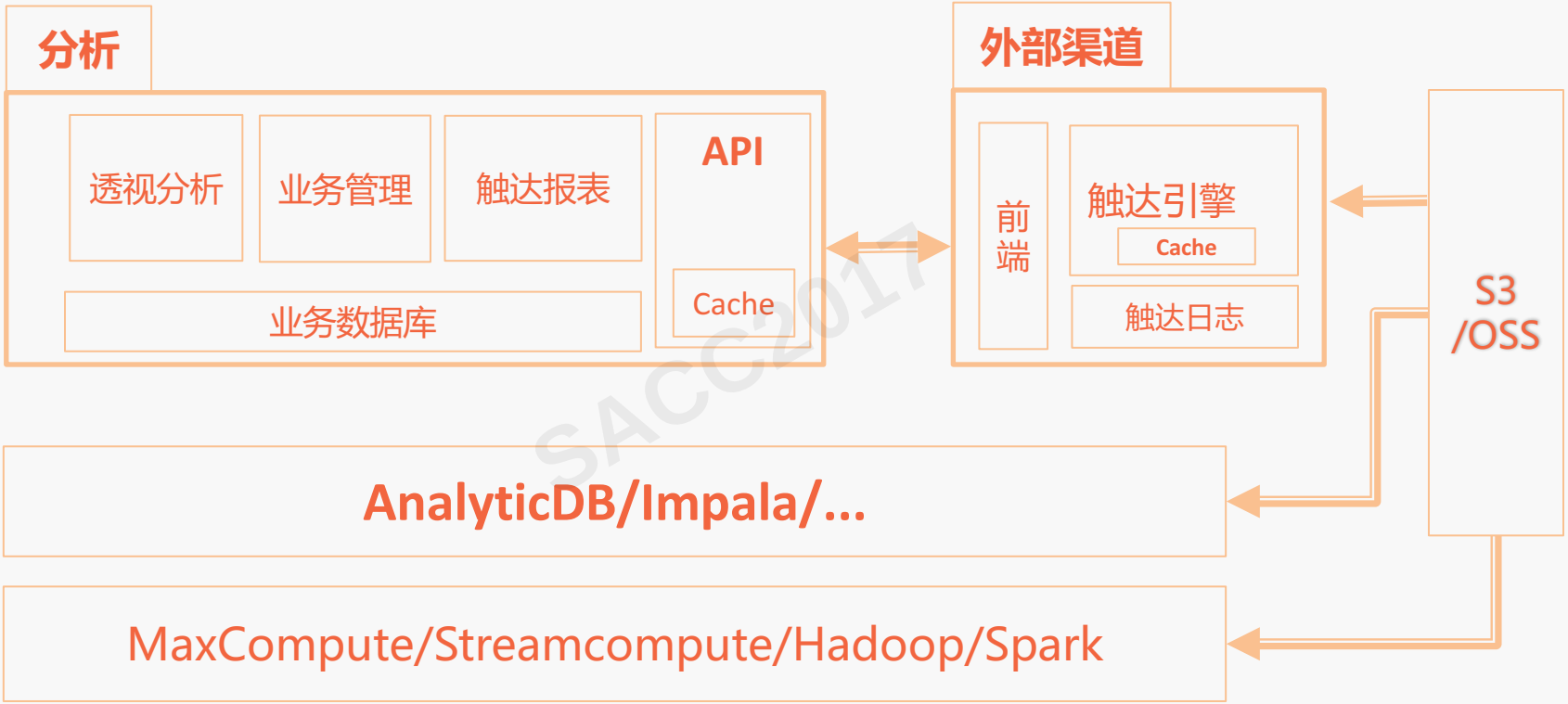
标签系统

对象标签化

典型架构



典型架构

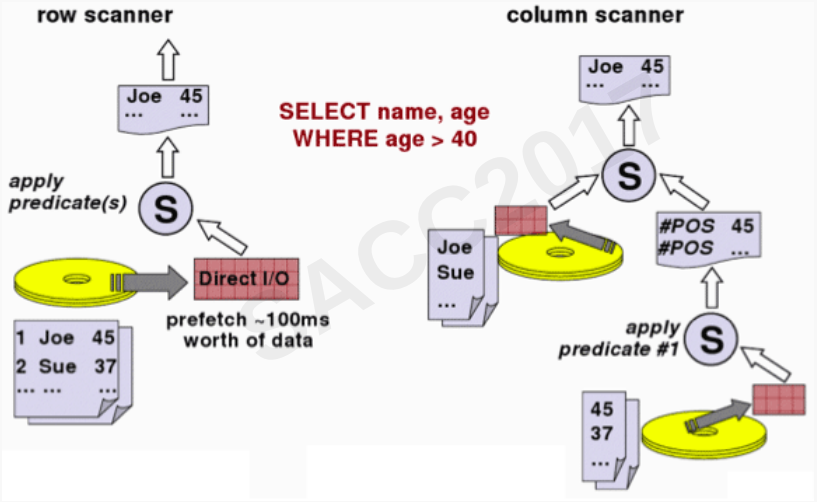


挑战一：高并发访问-存储性能

内部测试

20X
列存

3X
压缩

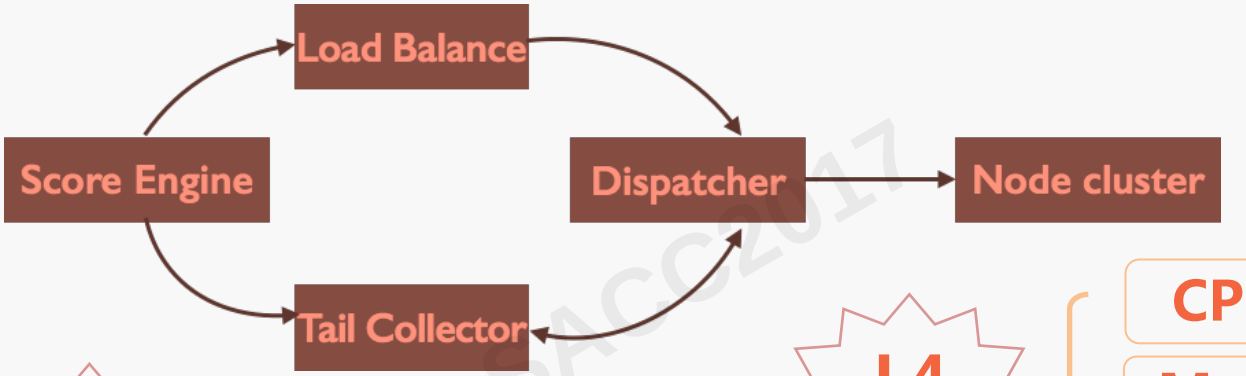


1000X
预排序

100X
全索引

挑战一：高并发访问-低延时

内部测试



13%

时延方差

3秒法则

L4

混合负载

- CPU
- Mem
- Net
- IO

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

挑战二：实时数据&明细查询

内部测试

40x
行列混存

10x
Block索引

挑战三：多集合交并差

内部测试

漏斗、标签模型

100+

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

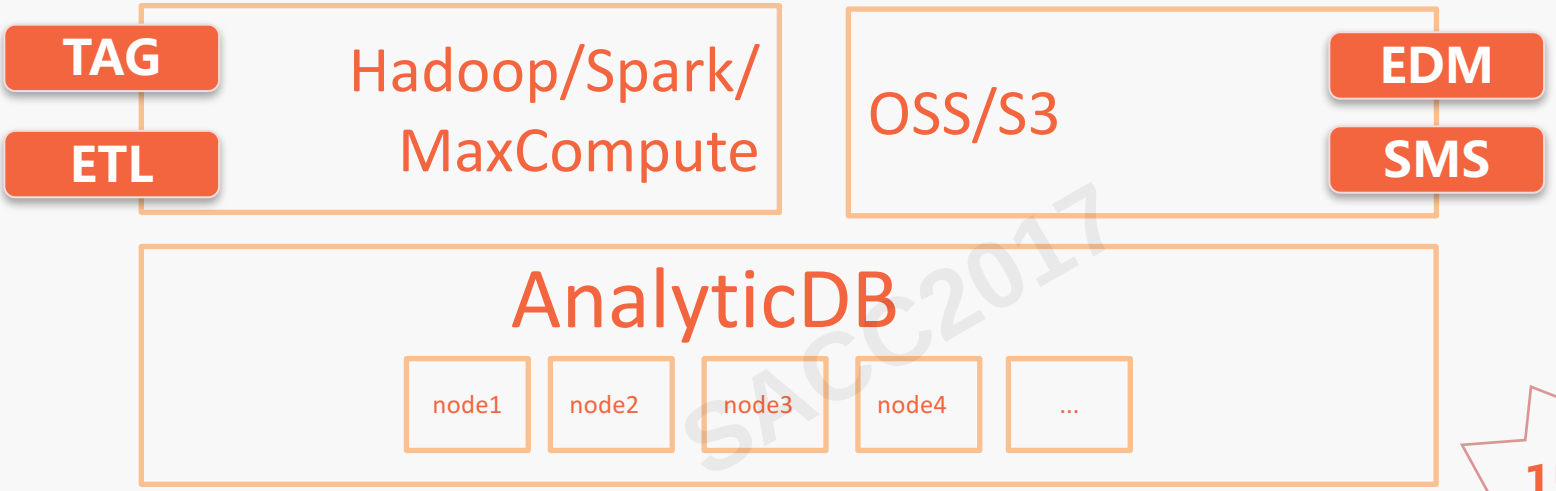
100x

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

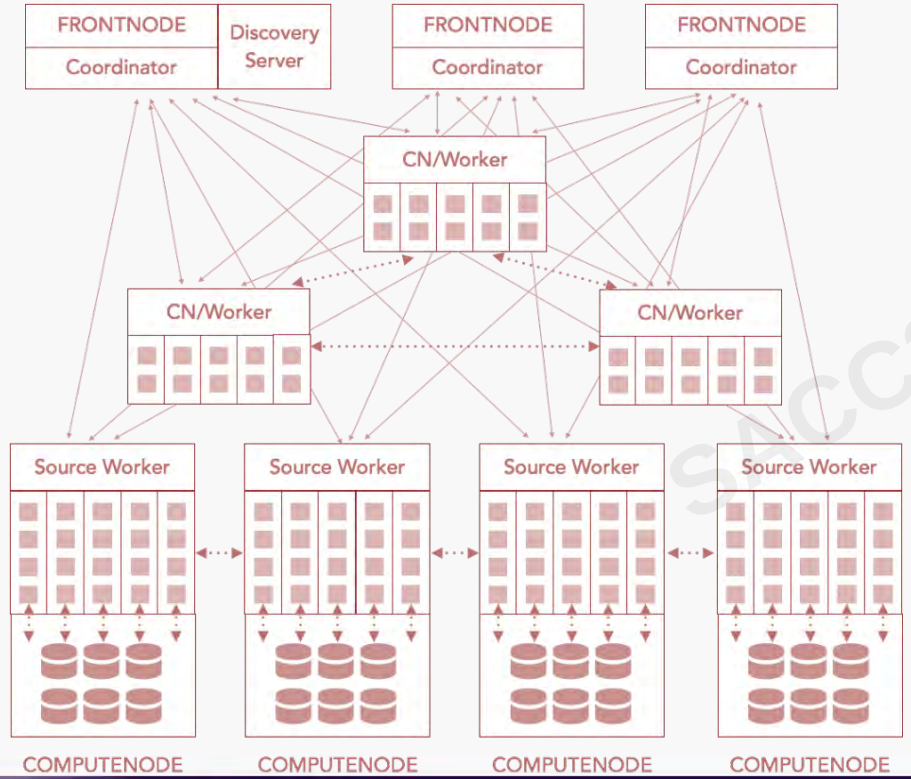
内部测试



150万 exp
500万 imp

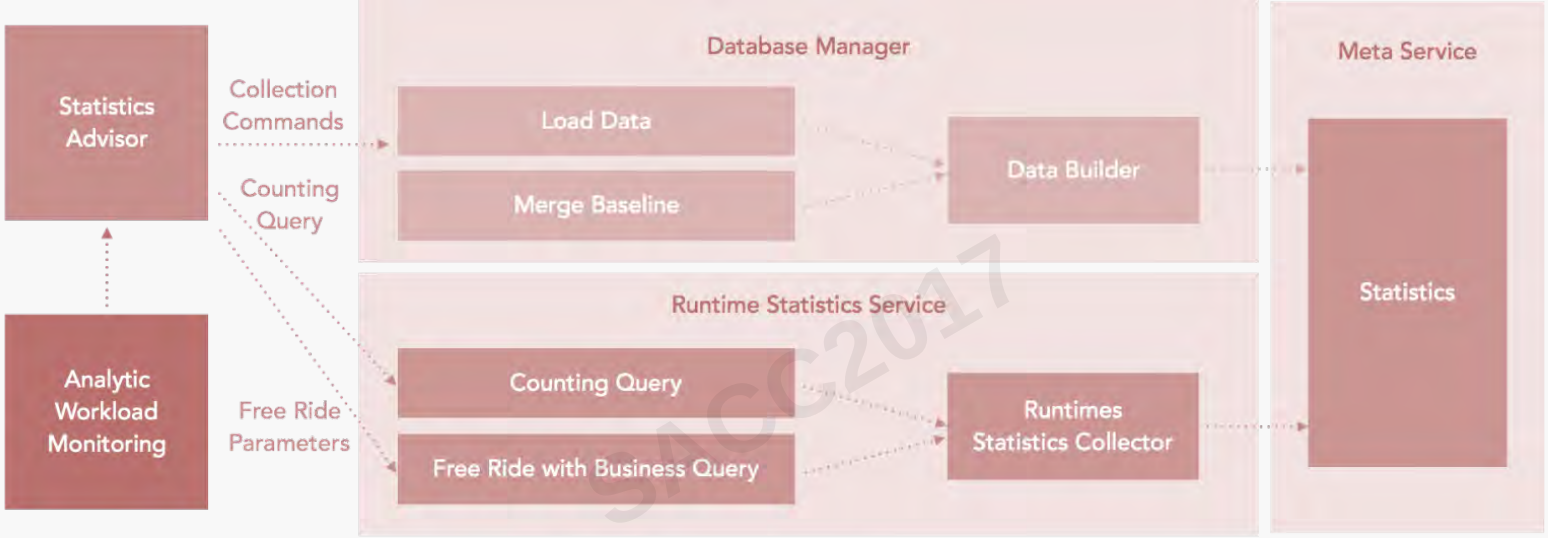
双边并行

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


THANKS