

大数据实时体系的架构和应用

数据平台部/实时计算中心/业务开发组

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



关于我



2010年加入腾讯数据平台部负责分布式计算平台，集群调度的开发，现负责实时计算体系基础建设和基于实时计算平台的推荐系统建设和业务推广。

微信：[tshirt](#)

邮箱：gabyzhang@tencent.com

数据平台目标-促进公司各业务数据共享

- ✓ 生活化电商, 微店
- ✓ 用户行为(交易、收藏)
- ✓ 产品类目信息
- ✓ 财付通



- ✓ 计费、营销
- ✓ 搜索、地图LBS
- ✓ 邮箱
- ✓ 输入法



- ✓ QQ、手Q
- ✓ qzone、朋友
- ✓ 会员、超Q、QQ秀
- ✓ 开放平台
- ✓ 微信



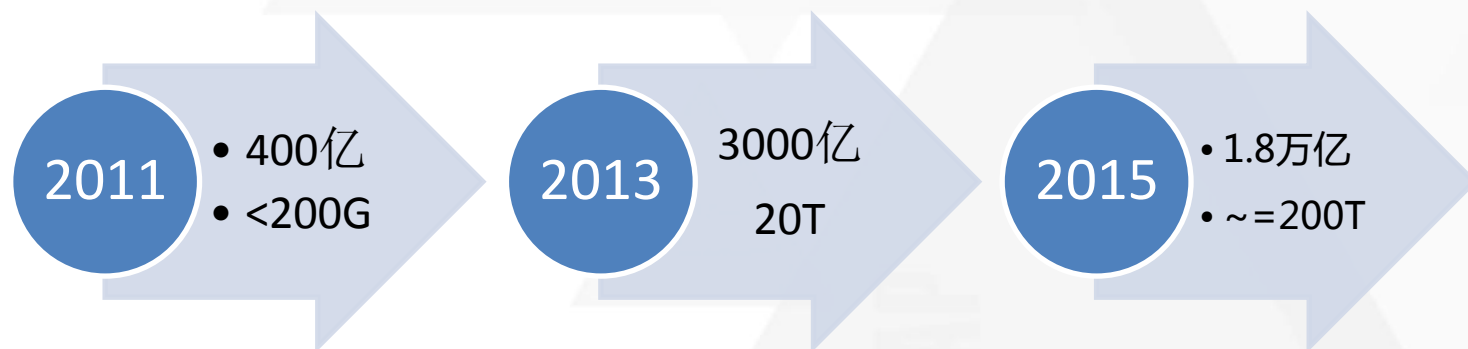
- ✓ CF、DNF等几十款
- ✓ 注册、登录
- ✓ 付费、充值
- ✓ 游戏内个性数据



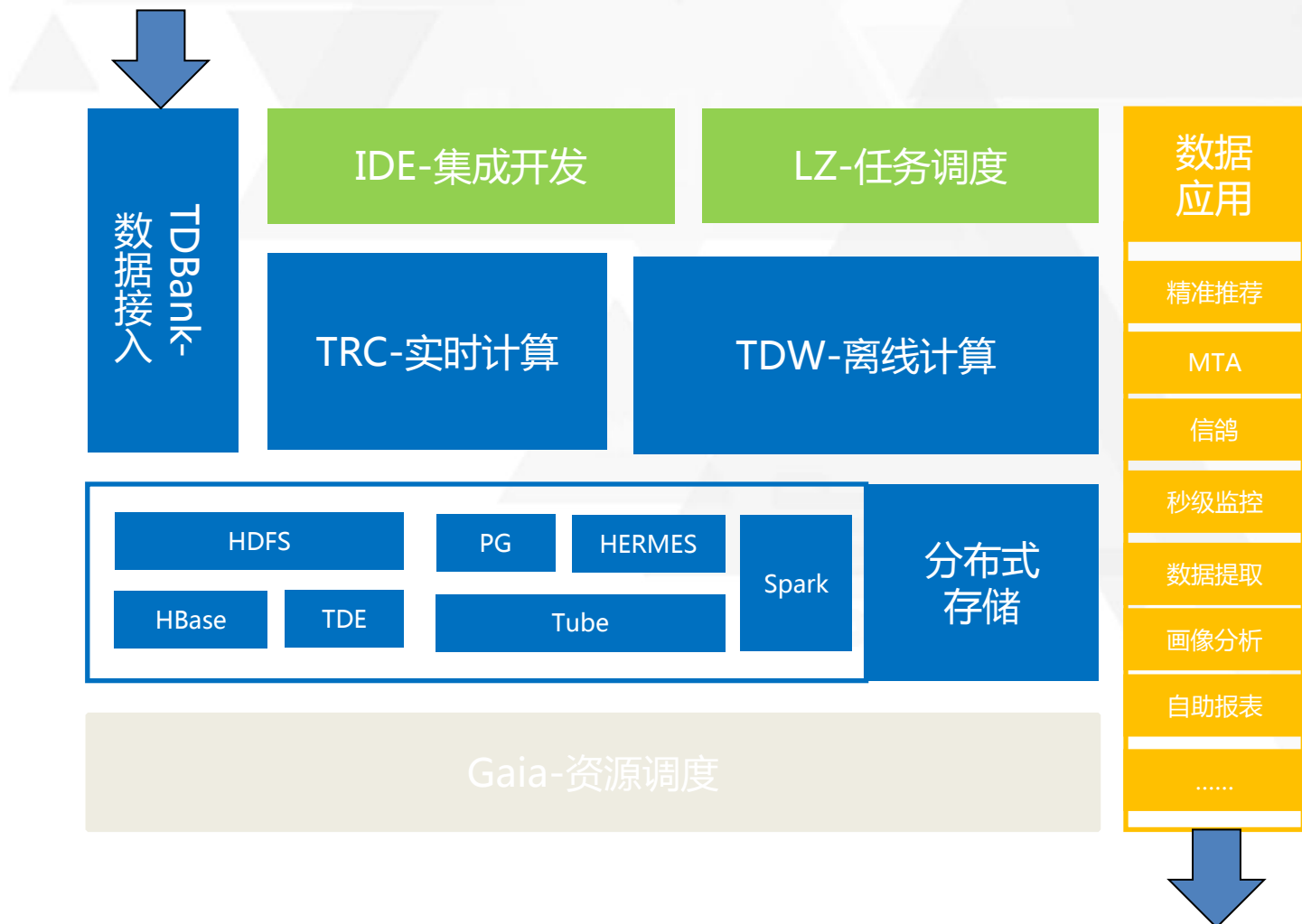
- ✓ 腾讯网网站行为
- ✓ 视频、音乐
- ✓ 新闻
- ✓ 广告



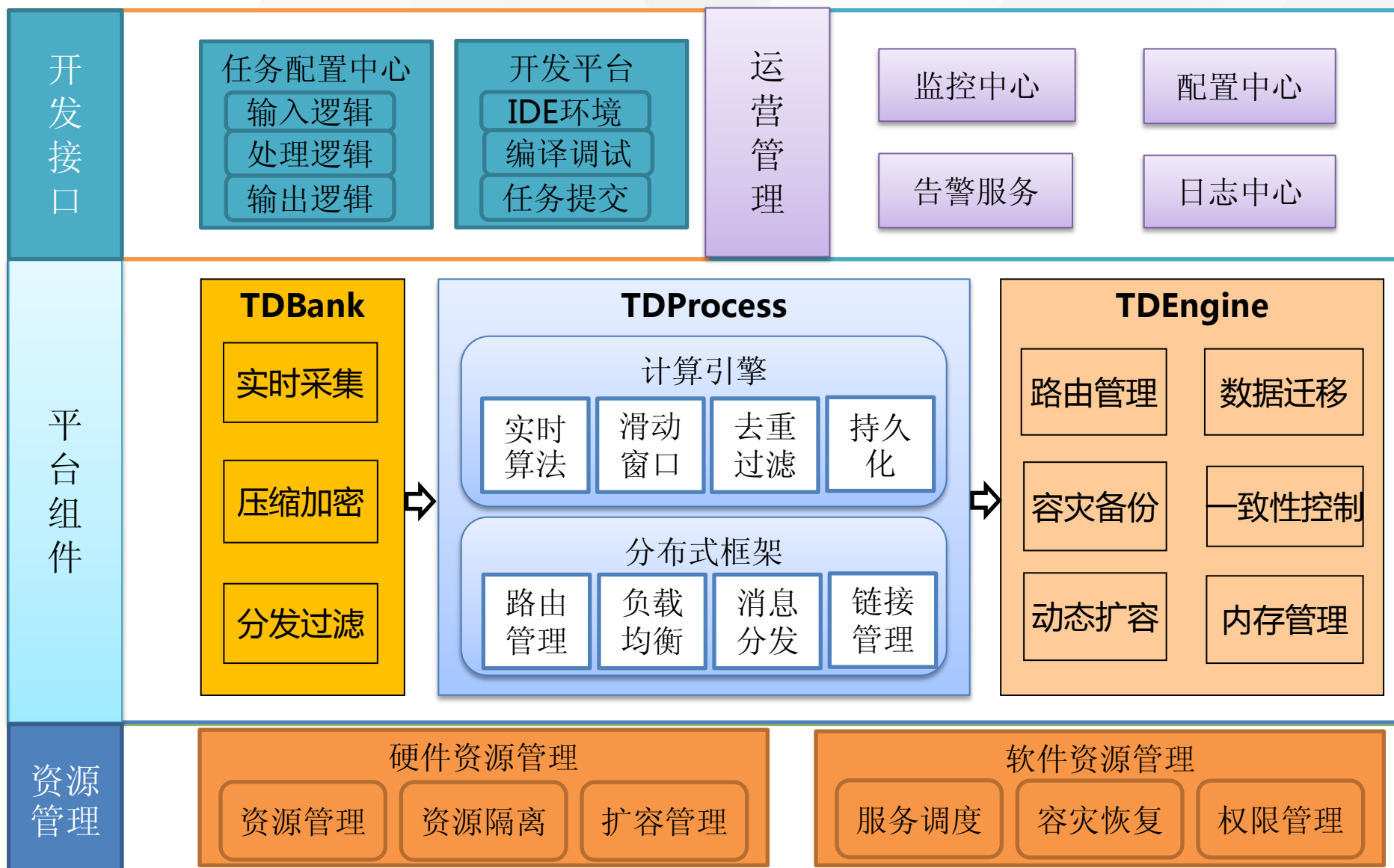
日均接入数平的数据



数据平台部大数据体系基础架构



TRC的整体架构



数据接入主要问题

□ 主要矛盾

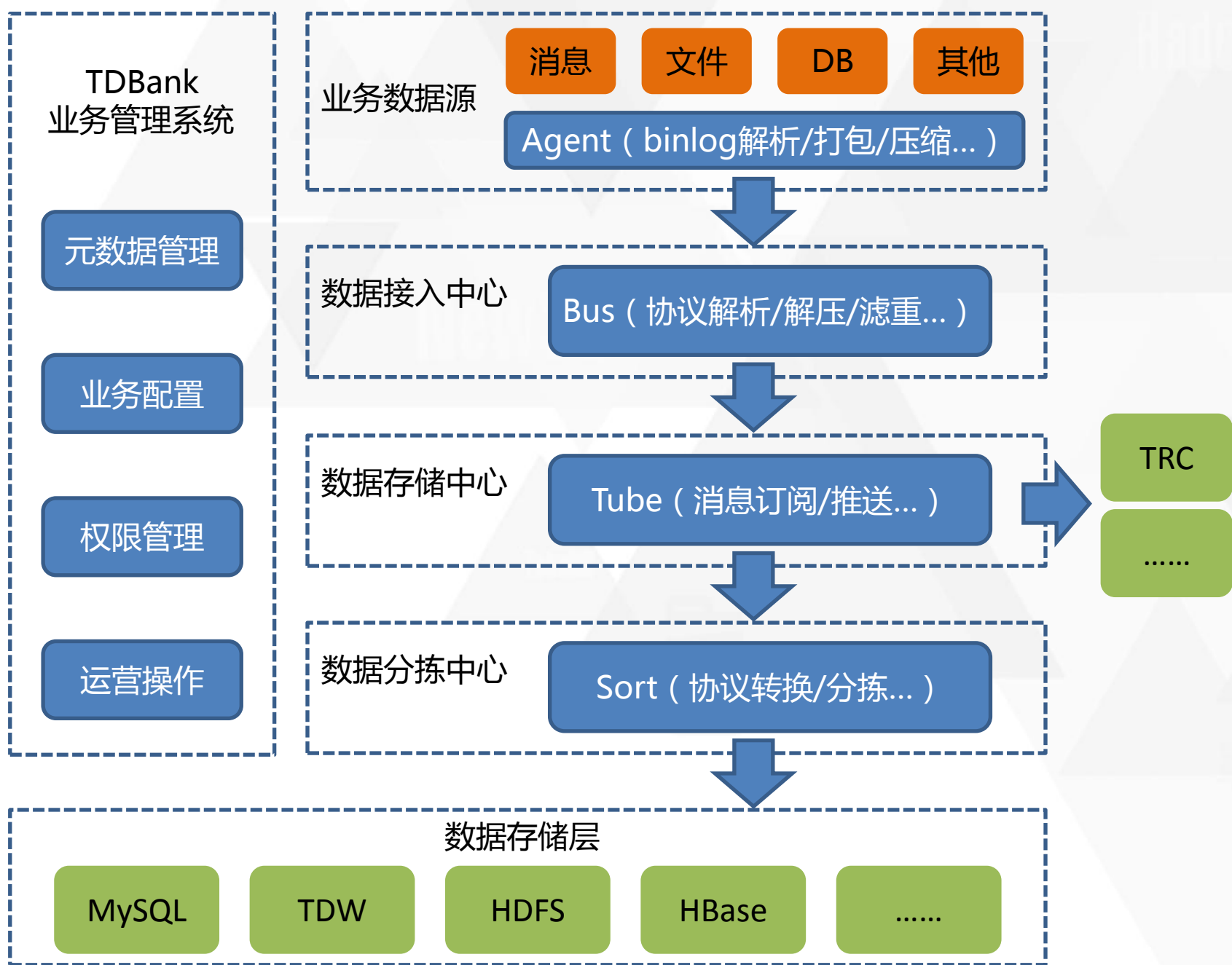
- 数据总量巨大
- 数据源种类繁多
- 数据格式各异
- 数据分布IDC众多

□ 核心需求

- 秒级接入延时
- 成本、效率、安全
- 方便数据管理和使用

□ 特色功能

- 自助接入
- 多种格式适配
- 公网加密传输



TDProcess流式处理引擎

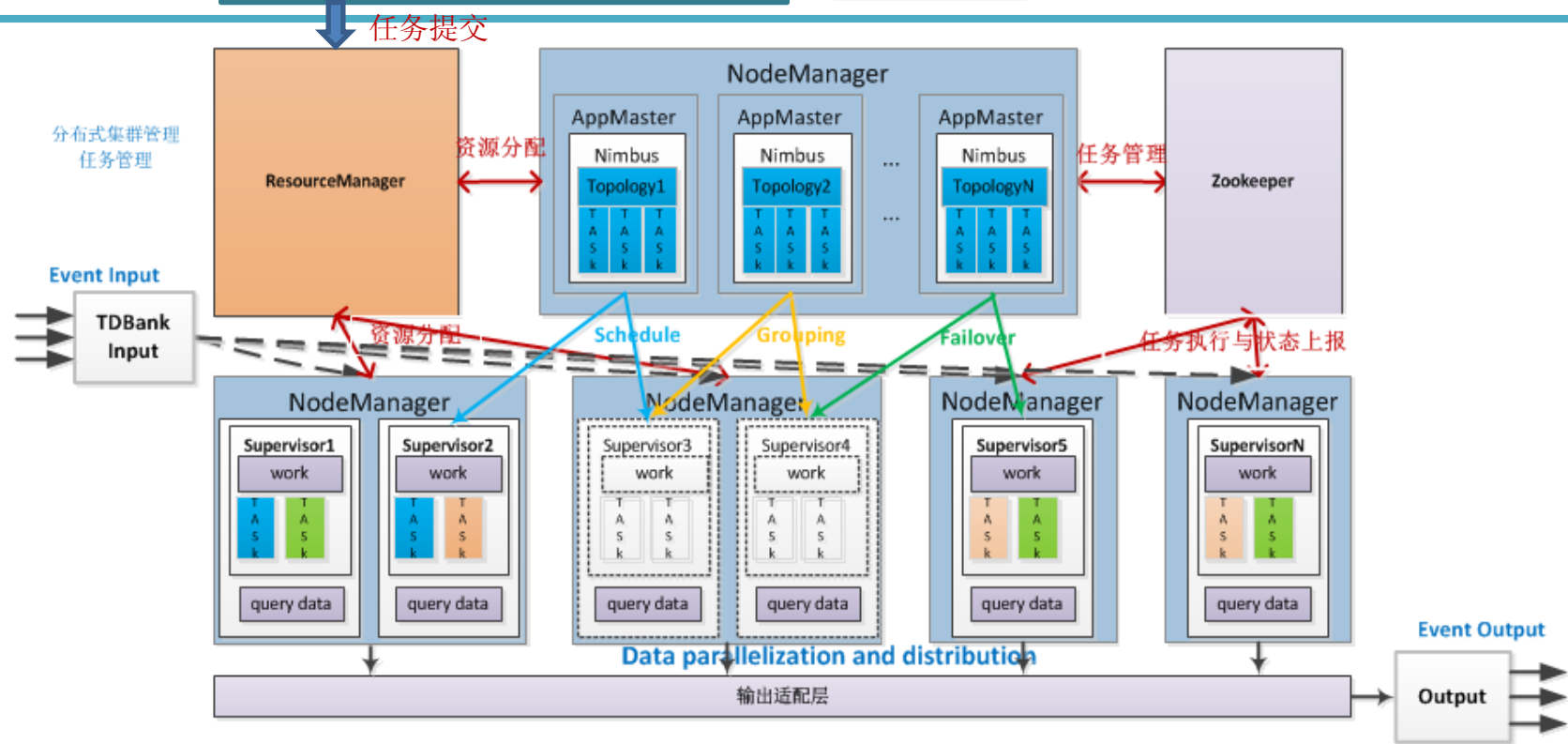
开发工具



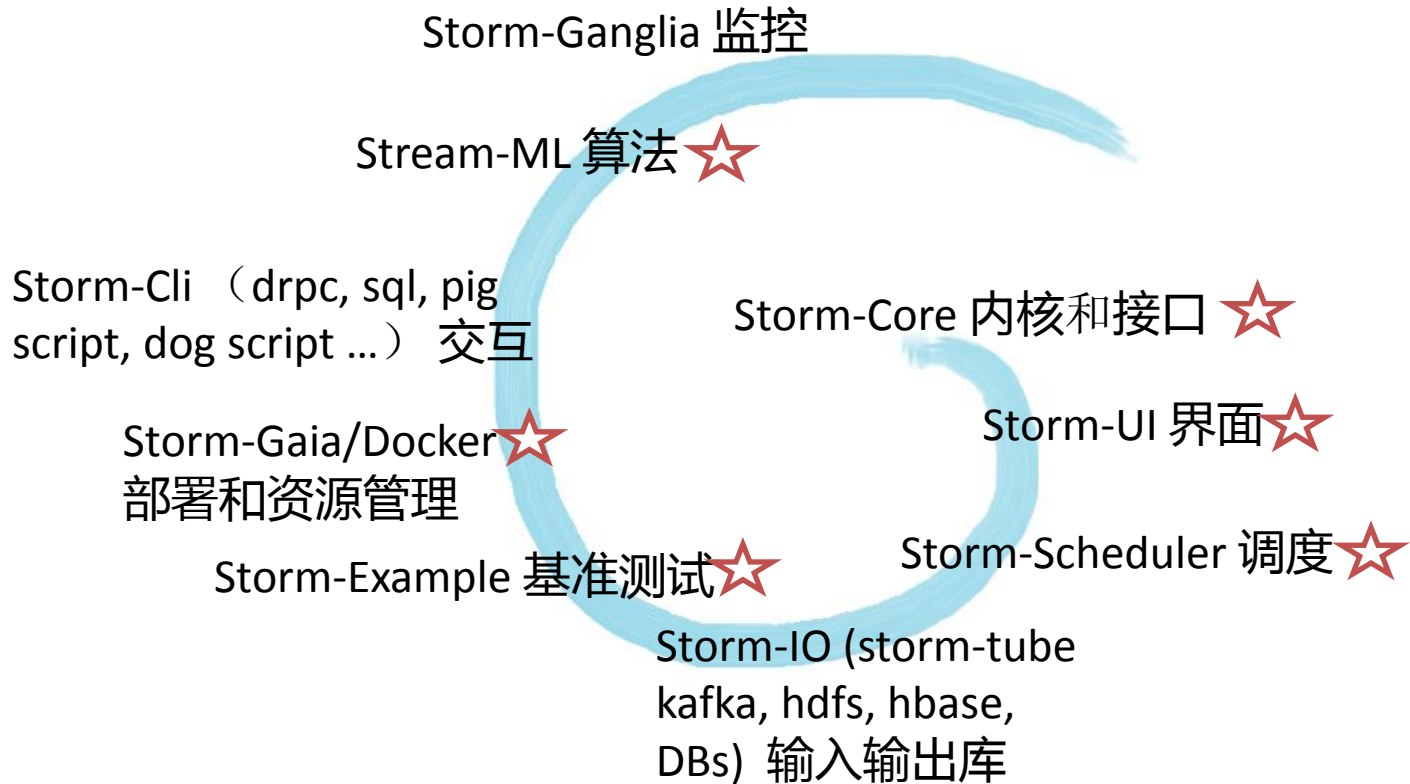
语言扩展



计算引擎



Storm Ecosystem

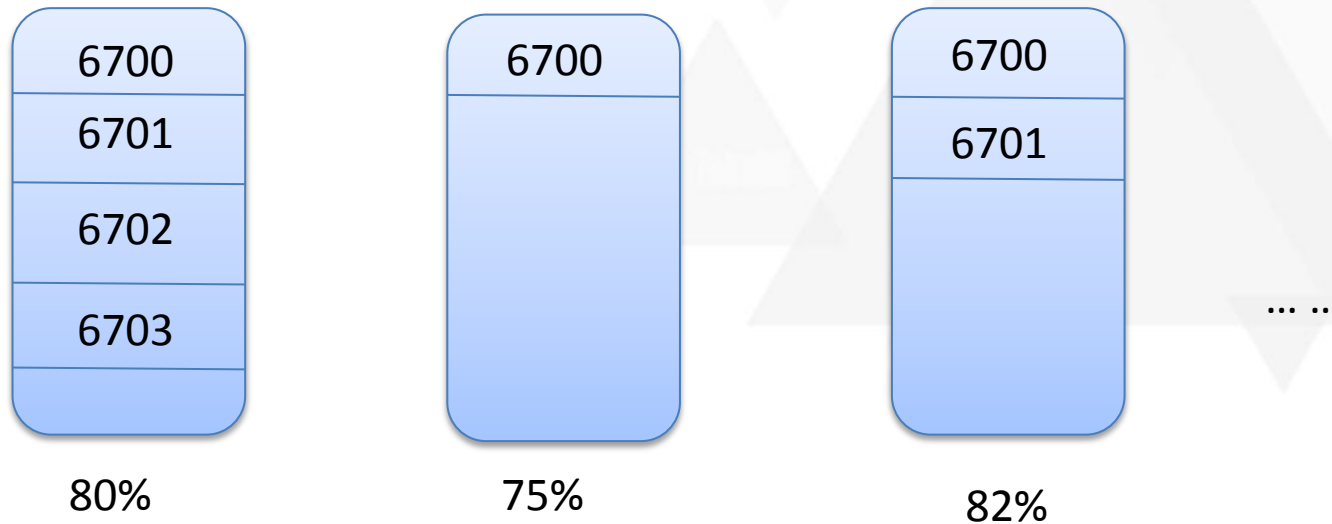


Scheduler Impls

What about resource negotiation?

基于物理机器负荷的调度策略：

按照机器的CPU/MEM资源使用百分比进行调度，理想结果是集群中每天机的CPU/MEM使用百分比是相近的



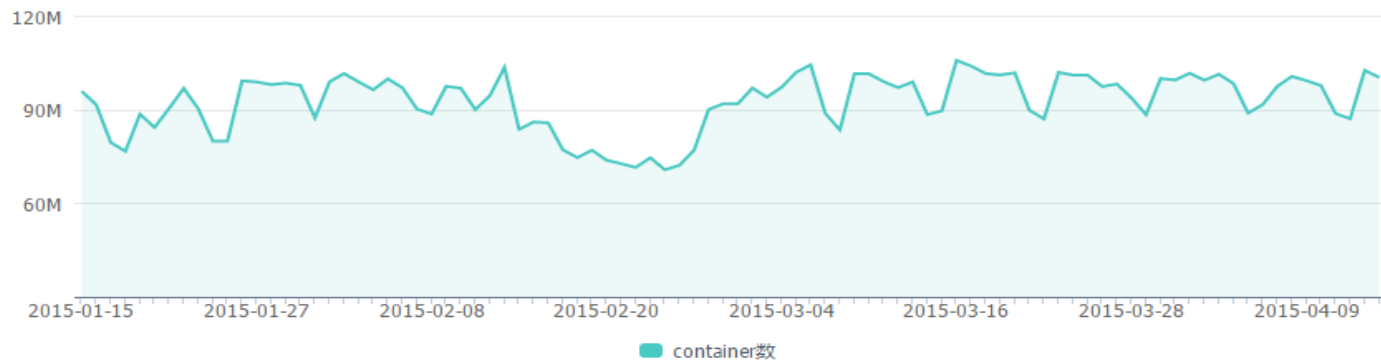
But ...



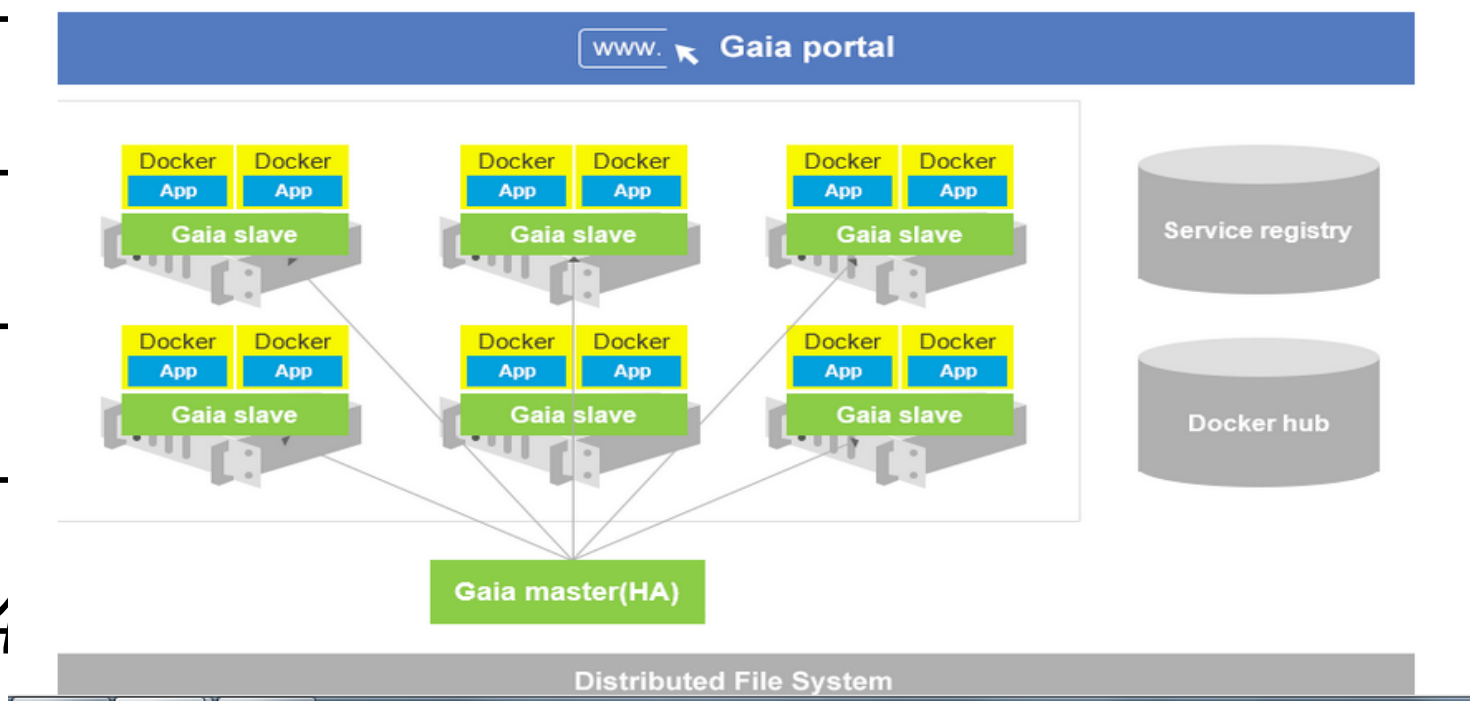
Ga

一服

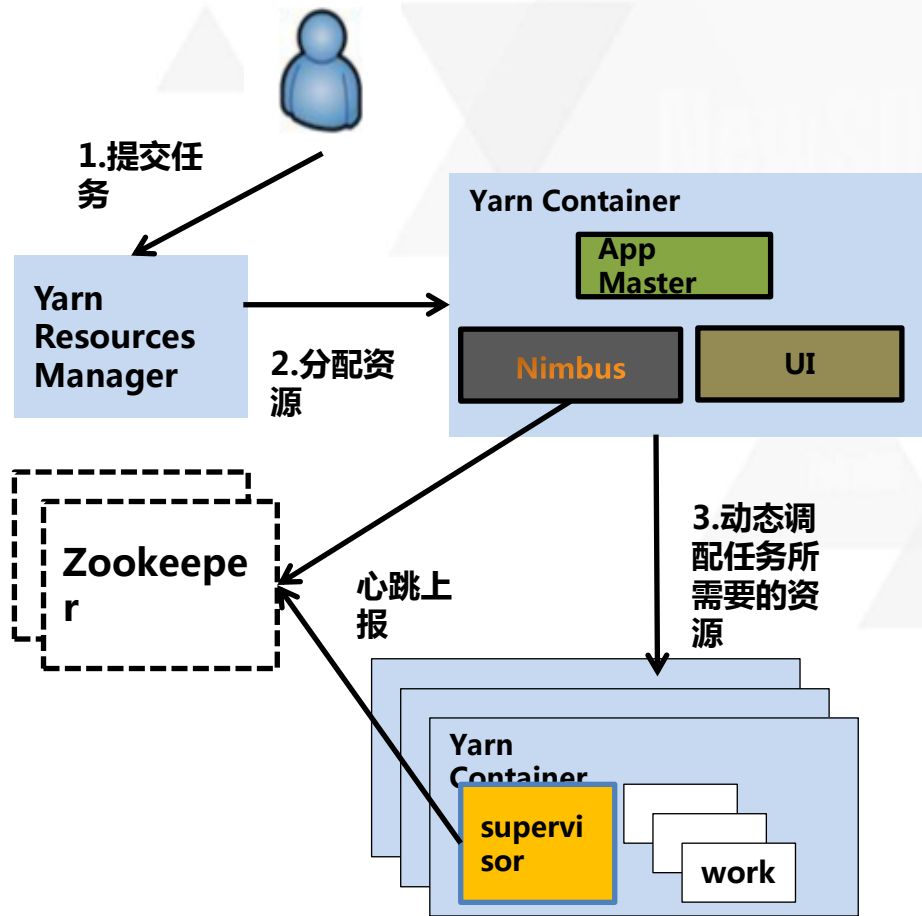
运行实例数量统计图



平台架构



基于Gaia订制storm



基于Gaia的Nimbus HA

One topology One storm

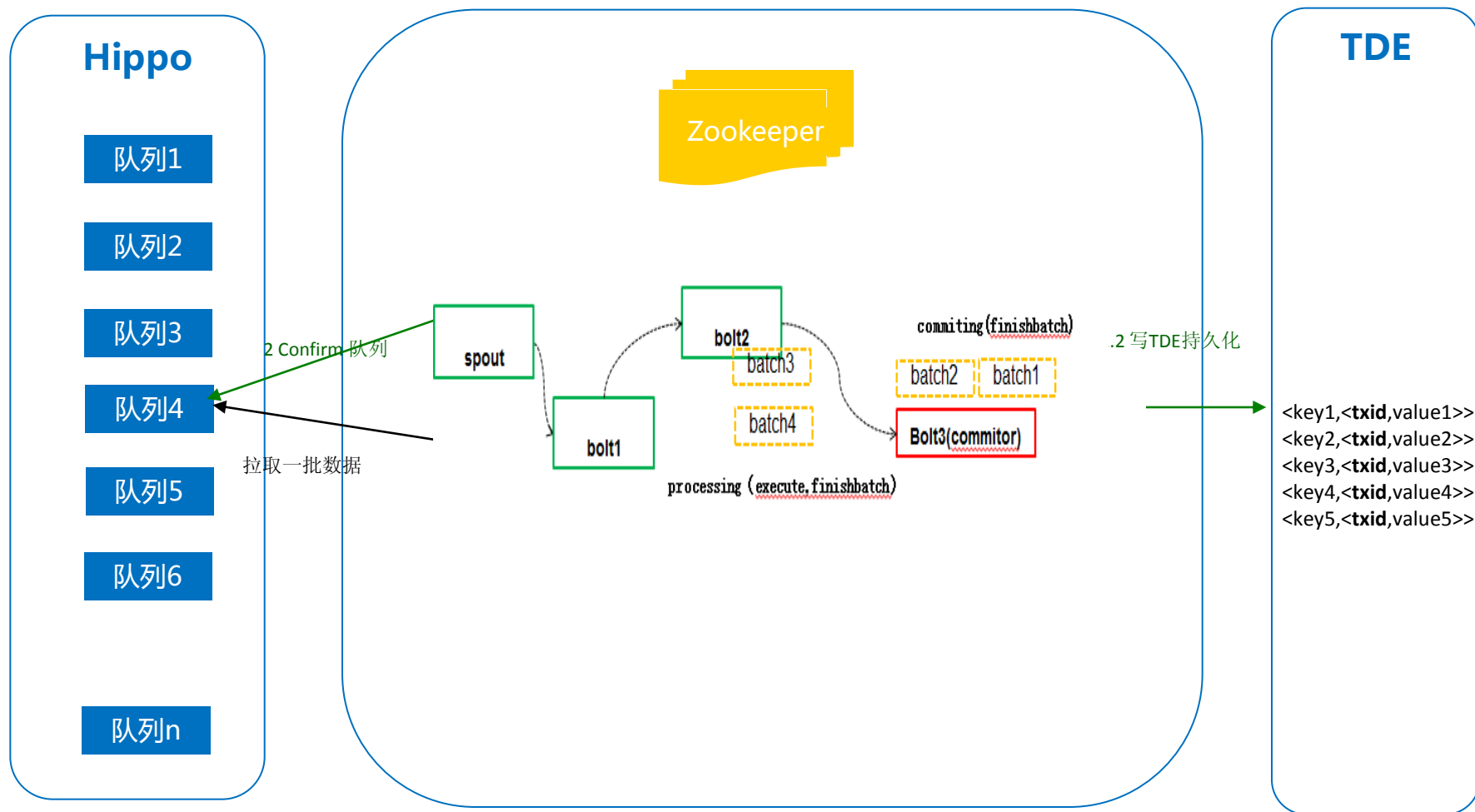
Nimbus Supervisor的数量不受限于物理机器

Gaia负责资源调度，
Nimbus负责任务分配
(task)

扩容缩容逻辑由Nimbus的
rebalance实现



事务 Topologies



- ❑ 一个事务分两个阶段完成，batch和commit，batch并发执行，commit顺序执行。
- ❑ Hippo队列由多个EmitBolt均衡读取，在没有confirm之前，队列的数据可重新读取。
- ❑ TDE存储数据，在Value里面存储了事务ID，如果TDE里面的数据的事务ID大于或等于当前事务ID，则不做写操作。

Ganglia



现网引流测试平台

精准推荐现网引流系统

广点通

指标展示

模块信息

广点通-Computer

推荐引擎Compute_adpos_计算_

[基]请求量

[测]请求量

推荐引擎Compute_adpos_计算_

[基]系统失败率

[测]系统失败率

广点通-task

推荐引擎Task_adpos_计算_现网

[基]请求量

[测]请求量

推荐引擎Task_adpos_计算_现网

[基]系统失败率

[测]系统失败率

广点通TDE集群

TDE请求量

[基]请求量

[测]请求量

DataSet进程数

[基]请求量

平台运行正常 [查看详情](#)

指标类型

业务指标

公共指标

详细指标图展

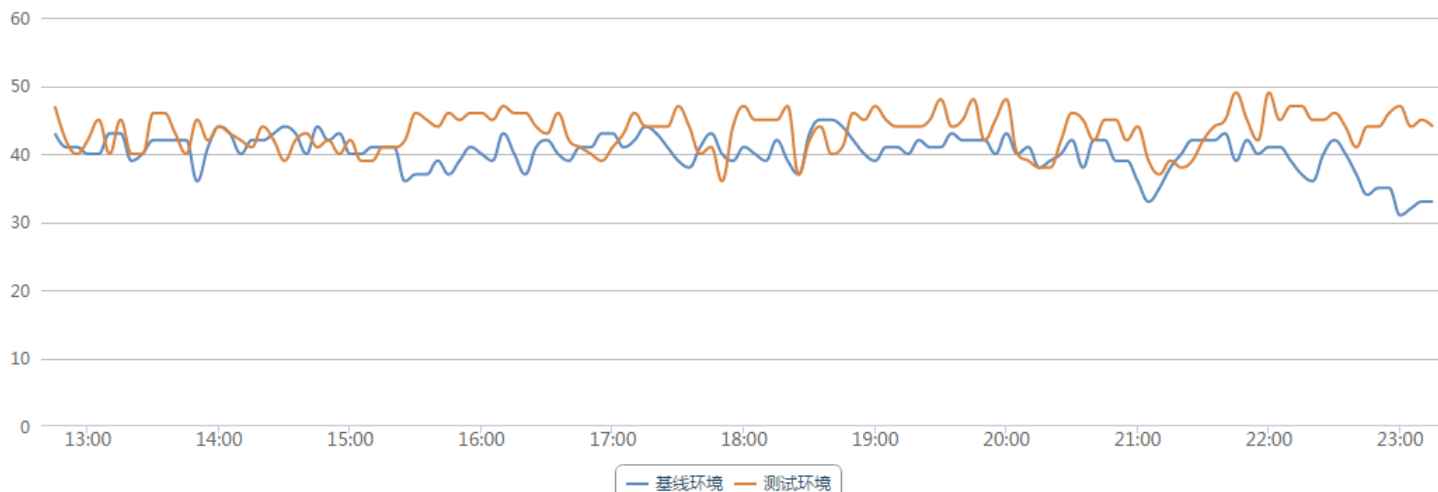
接口: CPU使用率

时间: 2014-12-23 12:45

-至- 2014-12-23 23:20

查询

公共指标曲线图



DTCC

2015年中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2015

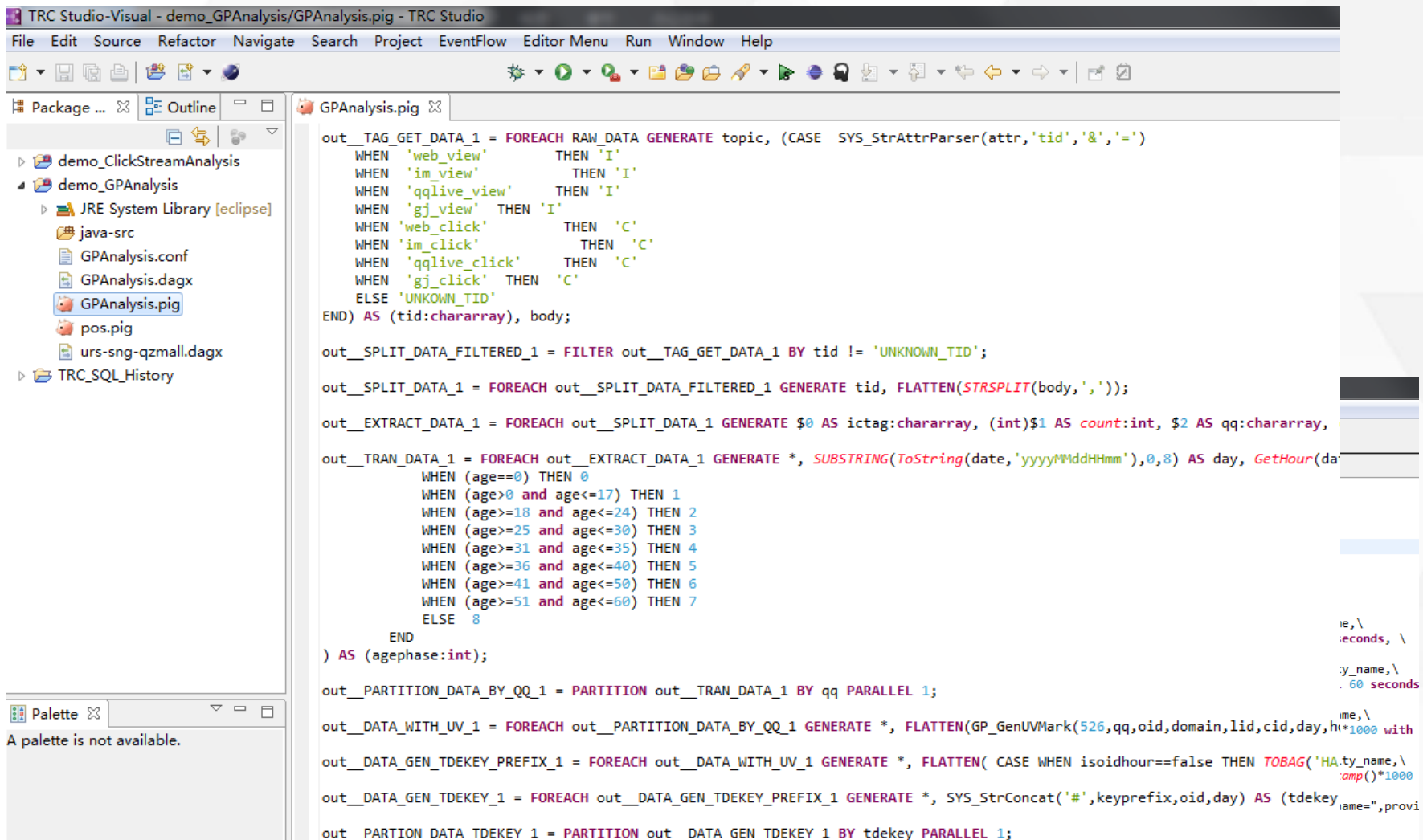


易用性？

- ❑ 编程接口复杂，对开发人员技能要求高，现有模式难以对外开放；
- ❑ 随着承接业务的增多，维护投入越来越大；
- ❑ 业务需求变化（例如算法调优）频繁，响应速度慢；
- ❑ 各业务独立开发，共享度不高，不能充分利用历史智力资产。



DSL on Storm



TRC Studio-Visual - demo_GPAnalysis/GPAnalysis.pig - TRC Studio

File Edit Source Refactor Navigate Search Project EventFlow Editor Menu Run Window Help

Package ... Outline GPAnalysis.pig

demo_ClickStreamAnalysis
demo_GPAnalysis
JRE System Library [eclipse]
java-src
GPAnalysis.conf
GPAnalysis.dagx
GPAnalysis.pig
pos.pig
urs-sng-qzmall.dagx
TRC_SQL_History

```
out_TAG_GET_DATA_1 = FOREACH RAW_DATA GENERATE topic, (CASE SYS_StrAttrParser(attr,'tid','&','=')
    WHEN 'web_view' THEN 'I'
    WHEN 'im_view' THEN 'I'
    WHEN 'qqlive_view' THEN 'I'
    WHEN 'gj_view' THEN 'I'
    WHEN 'web_click' THEN 'C'
    WHEN 'im_click' THEN 'C'
    WHEN 'qqlive_click' THEN 'C'
    WHEN 'gj_click' THEN 'C'
    ELSE 'UNKNOWN_TID'
END) AS (tid:chararray), body;

out_SPLIT_DATA_FILTERED_1 = FILTER out_TAG_GET_DATA_1 BY tid != 'UNKNOWN_TID';

out_SPLIT_DATA_1 = FOREACH out_SPLIT_DATA_FILTERED_1 GENERATE tid, FLATTEN(STR2SPRINT(body,''));

out_EXTRACT_DATA_1 = FOREACH out_SPLIT_DATA_1 GENERATE $0 AS ictag:chararray, (int)$1 AS count:int, $2 AS qq:chararray,

out_TRAN_DATA_1 = FOREACH out_EXTRACT_DATA_1 GENERATE *, SUBSTRING(ToString(date,'yyyyMMddHHmm'),0,8) AS day, GetHour(da
    WHEN (age==0) THEN 0
    WHEN (age>0 and age<=17) THEN 1
    WHEN (age>=18 and age<=24) THEN 2
    WHEN (age>=25 and age<=30) THEN 3
    WHEN (age>=31 and age<=35) THEN 4
    WHEN (age>=36 and age<=40) THEN 5
    WHEN (age>=41 and age<=50) THEN 6
    WHEN (age>=51 and age<=60) THEN 7
    ELSE 8
END
) AS (agephase:int);

out_PARTITION_DATA_BY_QQ_1 = PARTITION out_TRAN_DATA_1 BY qq PARALLEL 1;

out_DATA_WITH_UV_1 = FOREACH out_PARTITION_DATA_BY_QQ_1 GENERATE *, FLATTEN(GP_GenUVMark(526,qq,oid,domian,lid,cid,day,h*1000 with
out_DATA_GEN_TDEKEY_PREFIX_1 = FOREACH out_DATA_WITH_UV_1 GENERATE *, FLATTEN( CASE WHEN isoidhour==false THEN TOBAG('HA.ty_name,\
    amp()*1000
out_DATA_GEN_TDEKEY_1 = FOREACH out_DATA_GEN_TDEKEY_PREFIX_1 GENERATE *, SYS_StrConcat('#',keyprefix,oid,day) AS (tdekey
    ame=",provi
out_PARTION_DATA_TDEKEY_1 = PARTITION out_DATA_GEN_TDEKEY_1 BY tdekey PARALLEL 1;
```

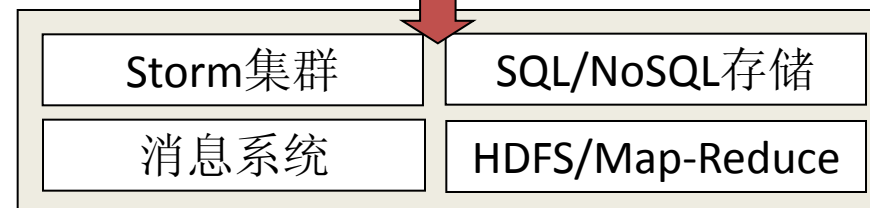
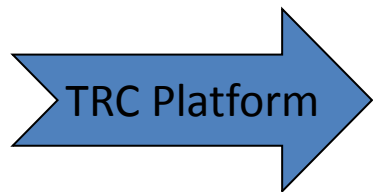
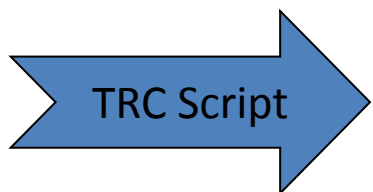
ie,\
seconds, \
y_name,\
60 seconds
me,\
1000 with
HA.ty_name,\
amp()*1000
ame=",provi

Palette
A palette is not available.

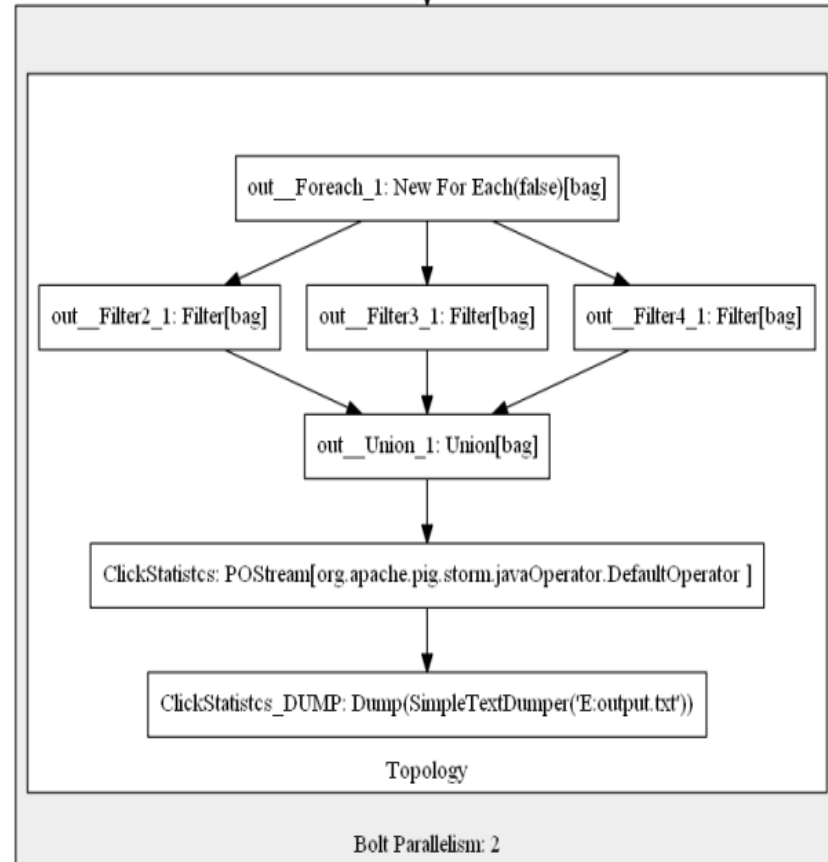
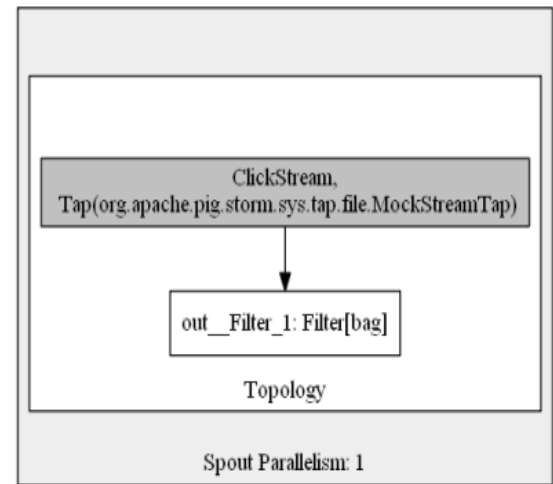
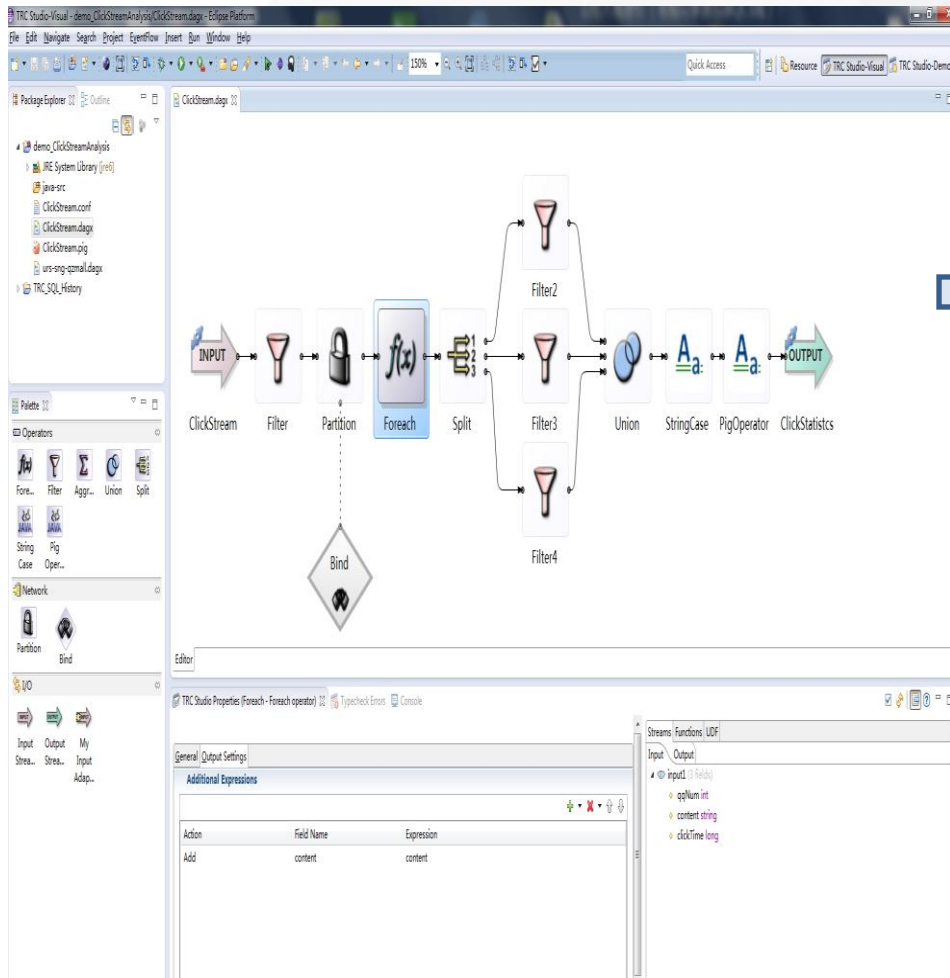


开发语言：SQL or Pig-Latin?

	比较项	SQL Like	Pig-Latin Like
业务需求	外部存储访问	支持	支持
	嵌套数据结构处理能力	弱	强
	多维度组合交叉计算	不支持	支持
	复杂业务支持能力	弱	强
	UDF	不支持	支持
	时间窗	不支持	支持
	join	支持	支持
	其他（Top,Sort等）	支持	支持
非业务需求	学习成本	低	中
	实现复杂度	高	中
	语言扩展能力	低	高



如何降低Storm开发的复杂度




可视化DSL语言

The screenshot displays the StreamVisualWorks IDE interface. On the left is a **Palette** with three categories: **Operators** (containing Fore..., Filter, Aggr..., and Split), **Network** (containing Partition and Bind), and **I/O** (containing an INPUT component and a green arrow component labeled 'itp...' and 'rea...'). The main **Editor** area shows a single component labeled **InputStream** with an **INPUT** arrow icon. A red banner at the top of the editor reads **从TDBank输入点击流** (Clickstream input from TDBank). At the bottom, the **StreamVisualWorks Properties (InputStream - Input Stream)** panel is open, showing the **General** tab with the **Java function:** `clickStreamTDBank`. To the right of the properties panel, a **Streams** tab is active, showing an **Input** section with the message: "There are no available input schemas for this component to display."


可视化DSL语言

Palette


Open




Fore...




Filter2




Aggr...




Split




Net...




Partition




Bind




I/O




Input



Outp...



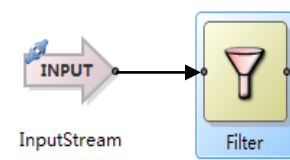
Input



Outp...

*ClickStream.dagx

根据itemId进行过滤



```
graph LR; InputStream[INPUT] --> Filter[Filter]
```

Editor

StreamVisualWorks Properties (Filter - Filter operator)

Typecheck Errors Problems

Specify at least one predicate by adding and completing rows in the table below.

General

Predicate Settings

☐ Create output port for non-matching tuples

Predicates:

Output Port	Predicate
1	itemId==207

Streams

Functions

Expression QuickRef

Input

Output

input1 (2 fields)

itemId int [点击商品ID]

time timestamp [点击时间]

可视化DSL语言

Palette

Operators

$f(x)$

Filter

Aggr...

Split

Foreach

Partition

Bind

I/O

Input Stream...

Output Stream...

*ClickStream.dagx

对数据进行预处理

InputStream

Filter

Foreach

Editor

StreamVisualWorks Properties (Foreach - Foreach operator)

Typecheck Errors

Problems

General

Output Settings

Input Fields

Additional Expressions

Action	Field Name	Expression
Add	minute	get_minute(time)%5

Streams

Functions

Expression QuickRef

Input

Output

input1 (2 fields)

itemId int [点击商品ID]

time timestamp [点击时间]

可视化DSL语言

Palette

Operators

- Fore... $f(x)$
- Filter
- Aggregate Σ
- Split

Network

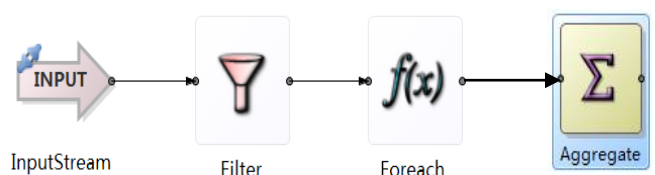
- Partition
- Bind

I/O

- Input
- Output

*ClickStream.dagx

进行5分钟点击聚合计数



```
graph LR; InputStream --> Filter; Filter --> Foreach; Foreach --> Aggregate
```

Editor

StreamVisualWorks Properties (Aggregate - Aggregate operator)

General Group Options Aggregate Functions

Additional Expressions

Action	Field Name	Expression
Add	count	count(itemId)

Streams Functions Expression QuickRef

Input Output

- input1 (3 fields)
 - itemId int
 - time timestamp
 - minute int

可视化DSL语言

Palette

Operators

Fore... Filter Aggr... Split

Network

Partition Bind

I/O

Input Stream Output Stream

*ClickStream.dagx

计算结果输出到TDE

InputStream Filter Foreach Aggregate OutputStream2

Editor

StreamVisualWorks Properties (OutputStream2 - Output Stream)

Typecheck Errors Problems

General Schema Advanced

Output Schema: ☒ Let typechecking define ☐ Declare:

Current Schema:

Field Name	Type
minute	int
itemId	int
time	timestamp
count	int

Streams Functions Expression QuickRef

Input Output

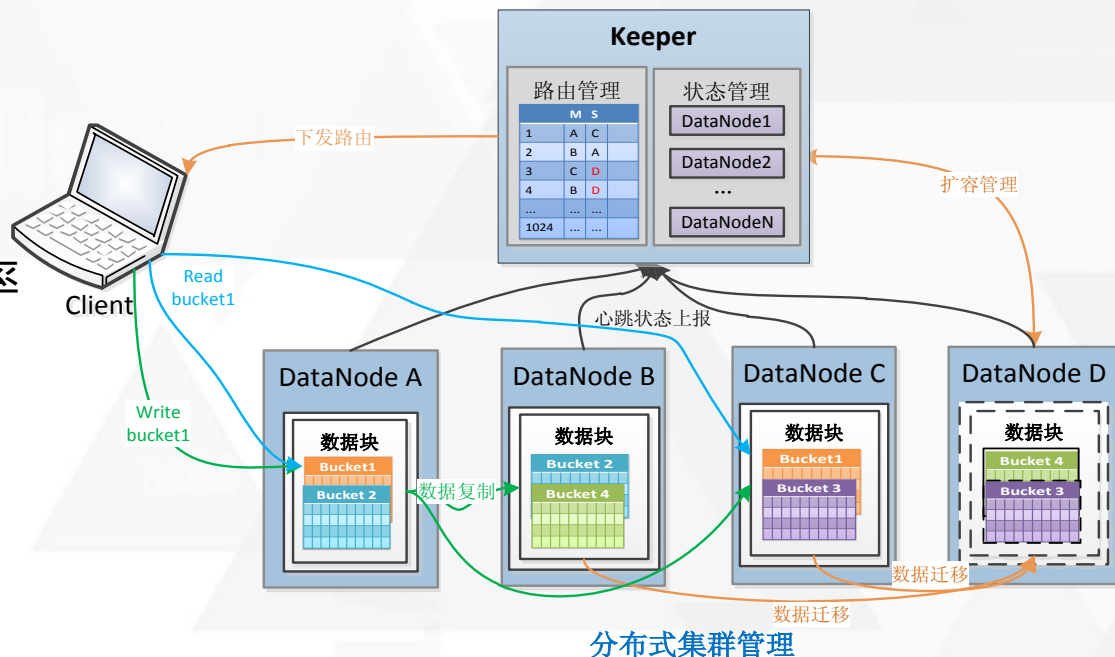
input1 (4 fields)

- minute int
- itemId int
- time timestamp
- count int

TDEngine存储引擎

□ 核心需求

- 高并发，低延迟
- 高可用性，数据安全
- 关注成本，关注资源利用率
- 线性扩展



□ 特色功能

- 支持多副本数据备份，确保数据安全
- 主备机同时提供服务，提升集群资源利用率
- 集群高可用，容灾切换过程中仍然提供读写服务
- 全内存设计，多引擎支持

每天支撑万亿数据访问请求



TRC在腾讯

每天，

万亿实时消息接入，**万亿**次实时计算，**万亿**次存储访问

覆盖，

SNG、IEG、MIG、CDG 等各大BG

涵盖，

广告、视频、游戏、文学、新闻、微信等多个业务

涉及，

个性化**精准推荐**、实时分析统计、秒级监控告警 等多个领域



★ QQ空间

个人中心

我的主页

应用

装扮

搜影视/音乐/小说/漫画

11:21

+

添加新应用

猜你感兴趣

QQ网购



货到付款

抗寒零下40°C

加绒冲锋衣 再不买就冷了

大家在玩



机甲旋风

好友阿敏在玩



梦幻Q仙

1376961人在玩



天神传奇

2867044人在玩

应用中心

动】系列同一个印章满/大

间抽奖次数不限,且100%中

奖品

时间: 活动结束60个工作

用户信息收集: 届时将通

集地址信息。

10:19 赞(184) 评论(91) 转发(3)

184人觉得很赞

我也说一句



樊小兰-刘璐老婆

我怎么迷糊成这样呢。。。。

09:52 赞 评论 转发 浏览(16)

我也说一句



叶凡

12.10现货白银、大圆银操作建

1、银价于3905—3925区域依然

2、下方回落3805—3825做多止

新闻

订阅

图片

视频

推广

快乐齐临门。祝福话语

色。点击链接,领取

上你的祝福语动起来!

乐。

中国联通 10:20

动态 好友动态

rayes发 radar_idemmed: 20°C; 压力

肯定有的,一切按程序来,每人在席位

一小时一换岗,一个席位有协调席和管制

席,还有一个人监控,一定程度上可以分

担压力

我也说一句...

帮 妈妈帮 — 备孕、怀孕、...

怀男孩和怀女孩有什么症状? 特征大

汇总就在此。

怀男孩和怀女孩

有什么症状?

特征大汇总就在此!

妈妈圈

下载

专利

昨天13:35

点赞 评论

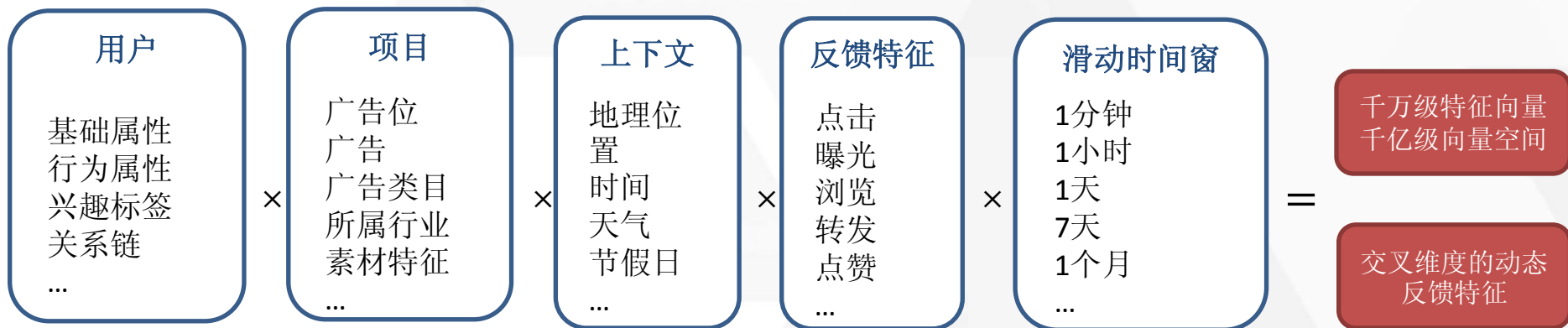
昨天11:24



CTR流式处理

预测用户A最可能点击广告，如何准备好预测相关数据？

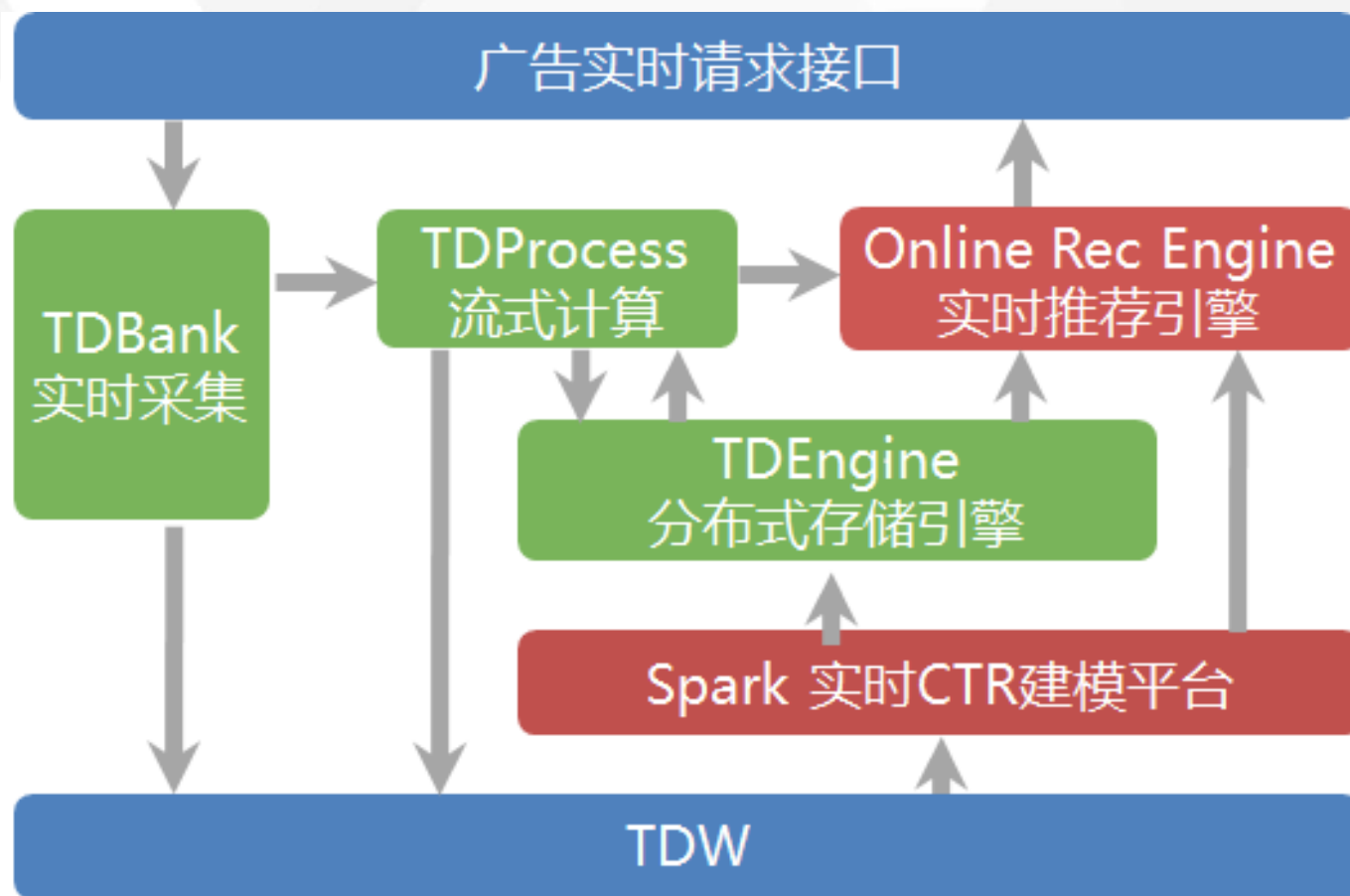
- 对每个广告，实时计算“用户↔广告”多个不同维度组合的相关度指标



- 日均200亿请求对应的每1条曝光日志，平均计算50多种交叉特征
- 仅广点通业务每天实时计算量超过万亿次
- 整个集群的计算量超过十万亿次



基于实时计算的点击预估模型架构



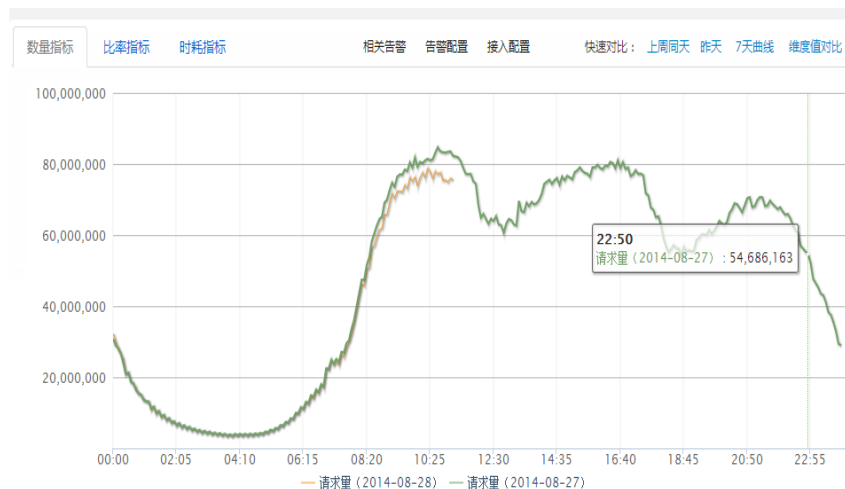
TRC的应用- 概览



对**微信的性能**优化、IDC部署、运营商选择等有着十分重要的作用



告警准确性大幅度提高；对监控对象进行全纬度组合分析，实现了**监控的100%覆盖**。





THANKS