SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

万亿数据实时接入与基于SQL的实时应用开发

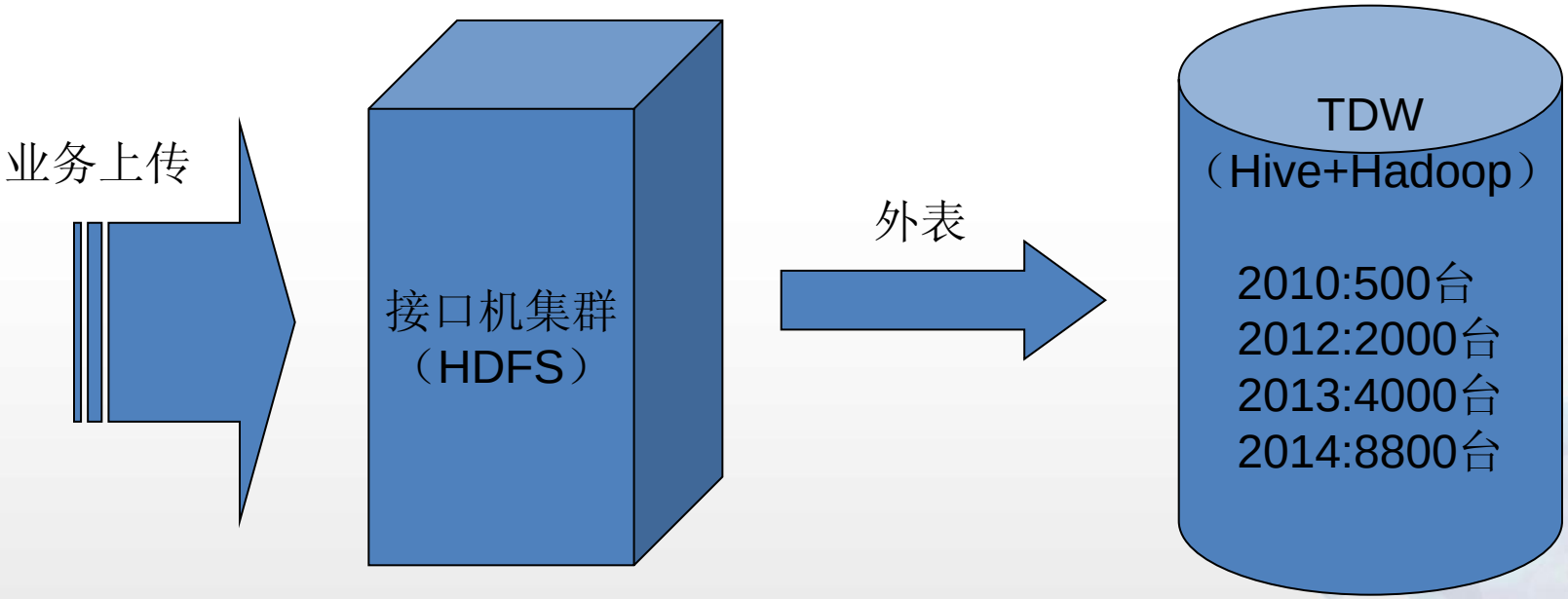
腾讯数据平台部
田万鹏
2014.09.17

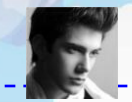
关于我

- 2010清华大学软件学院硕士毕业加入腾讯数据平台部
- 专注于海量分布式数据计算系统
  - TDW（腾讯分布式数据仓库）
  - TDBank（腾讯数据接入平台）
  - TRC（腾讯实时计算平台）
- 个人技术专长，分布式存储，分布式流式计算，基于MR和DAG模型的分布式SQL引擎。

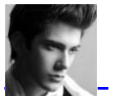
TDW的发展

- 2009：开始研发分布式数据仓库系统TDW
- 2010：正式上线1.0版本，集群规模500台
- 2011：形成**接口机+TDW**的模式，业务接入迅速增长
- 2012：**数据接入遇到瓶颈**



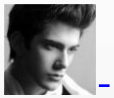


数据分布在全国100多个IDC，
有文件，有DB，还有消息



数据不能通过专网传输，为节约成本，需要走**公网**，还要加密

另外其他几个兄弟部门也需要这份数据，你们要帮忙**转发**，要**实时**的哟



对了，顺便说下，所有接口表数据都是混在一起的，需要提前**分拣**开

土豪，把数据放在接口机上，
剩下的事情就交给我们吧

还真不少，没关系，我们提供工具

土豪也差钱吗，不过这是个问题

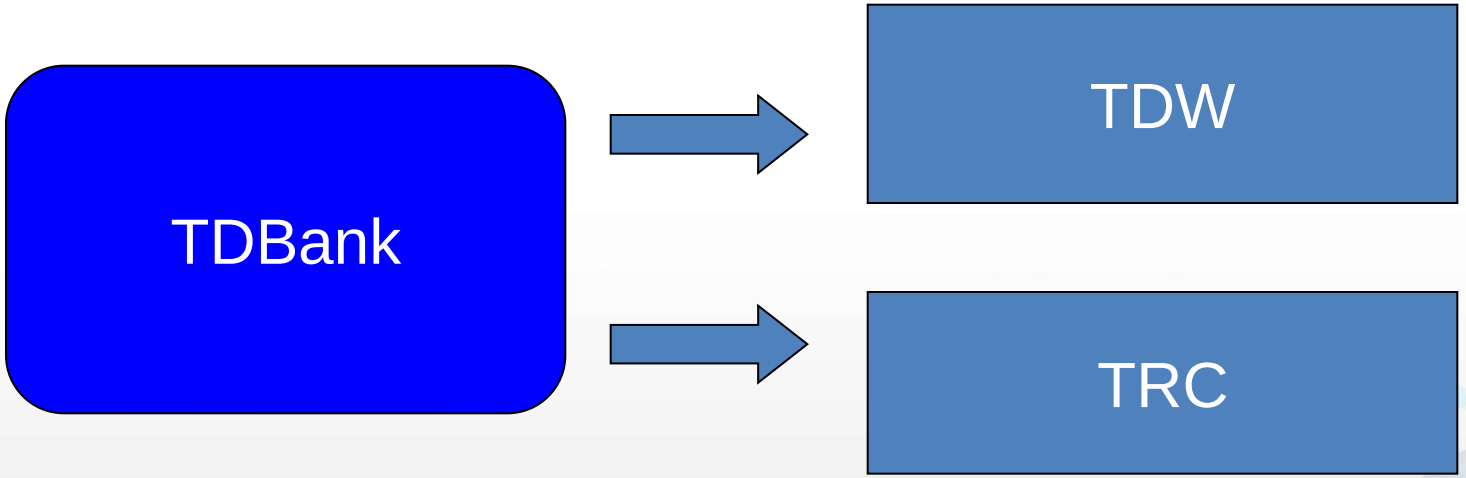
这事没有想过，需要仔细研究一下

。 。 。 。

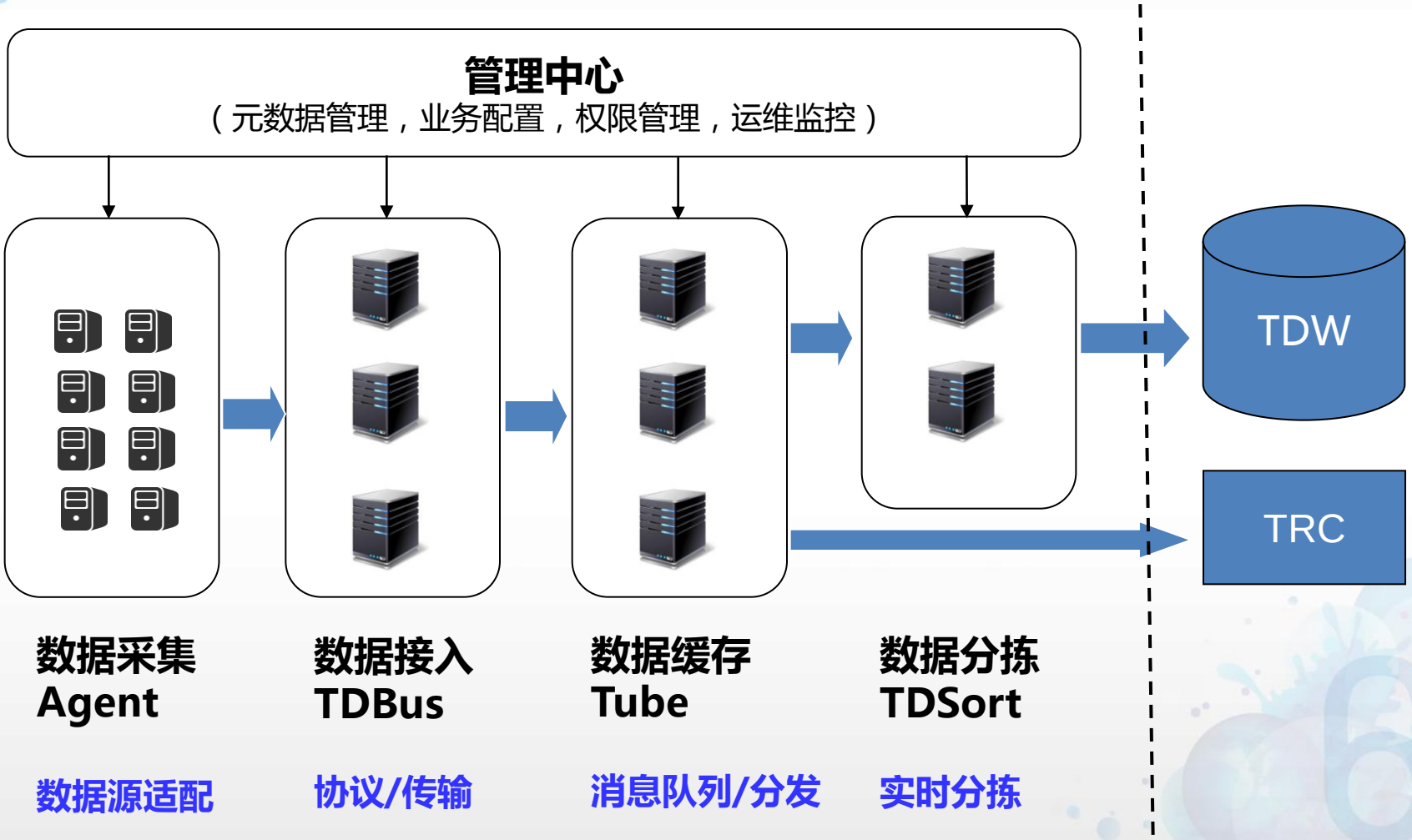
- 异构数据源
- 网络环境复杂
- 多路实时复用
- 实时分拣

统一接入 数据融合产生价值

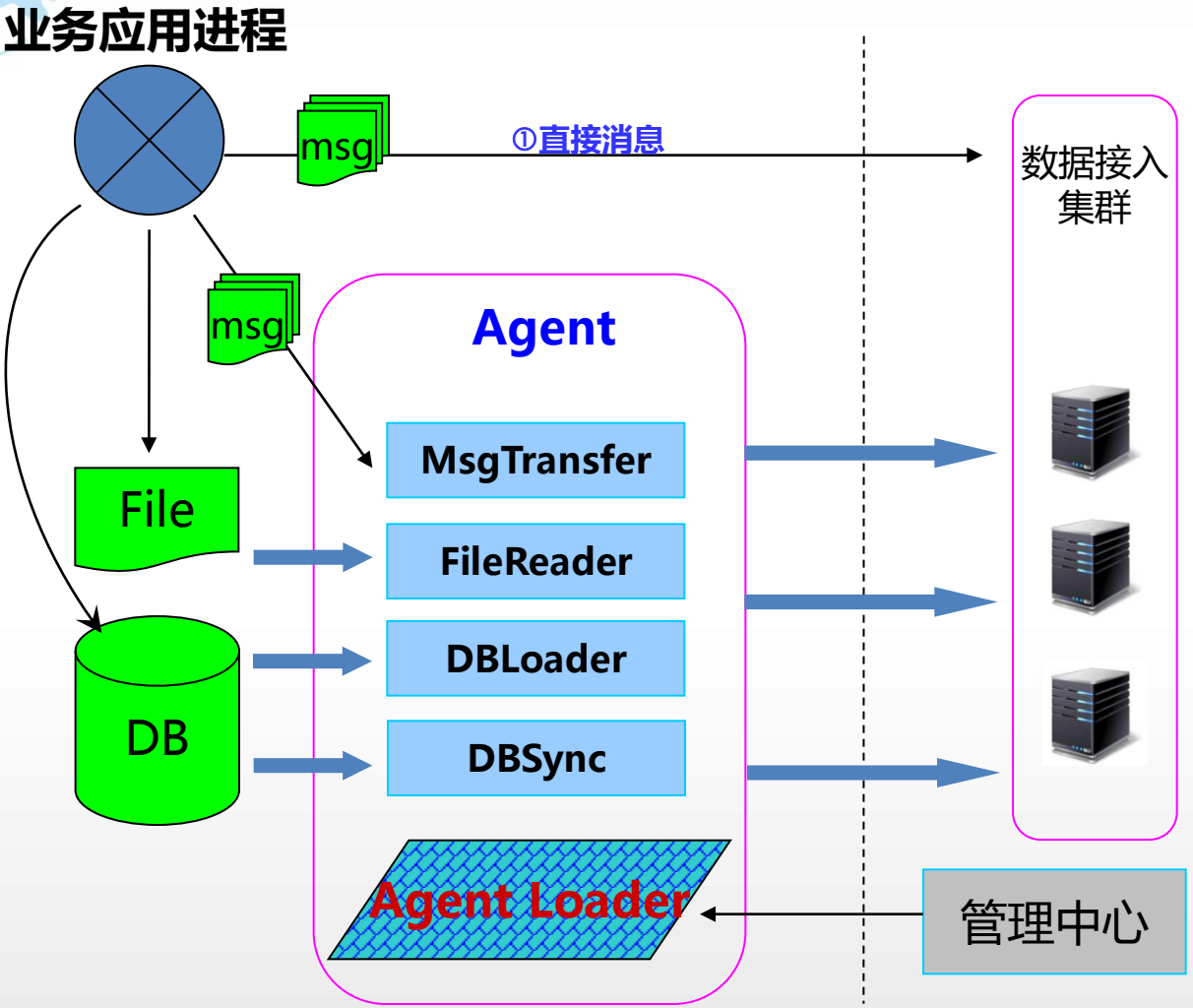
实时分发 实时数据更大价值



TDBank系统架构

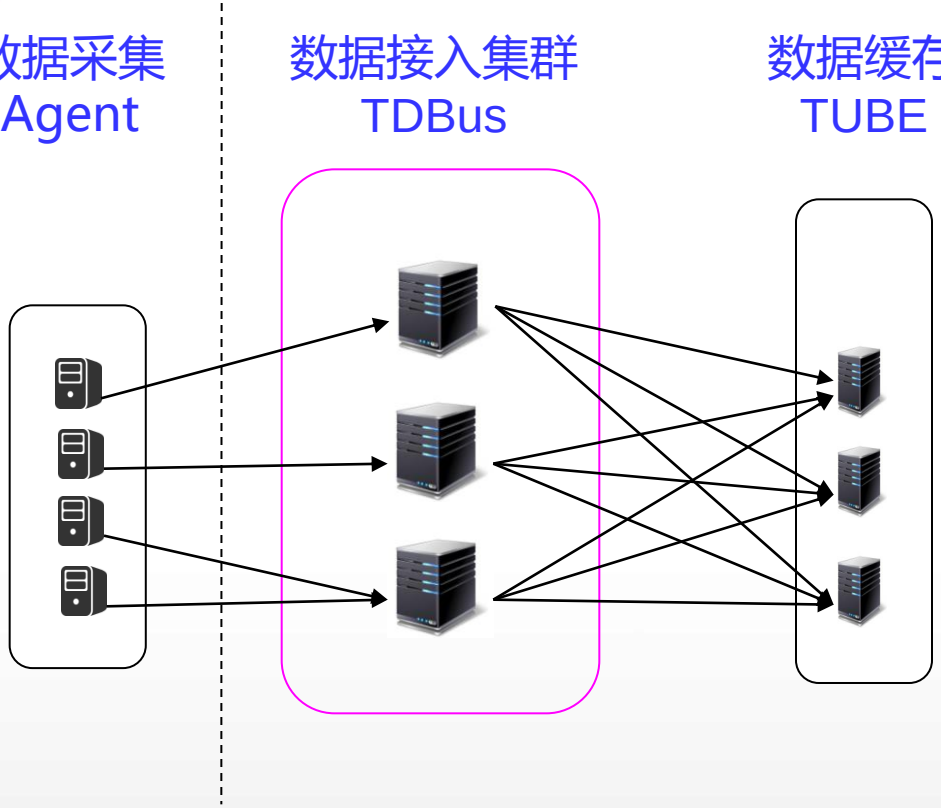

管理中心
（元数据管理，业务配置，权限管理，运维监控）
TDW
TRC
数据采集
Agent
数据接入
TDBus
数据缓存
Tube
数据分拣
TDSort
数据源适配
协议/传输
消息队列/分发
实时分拣

1. 数据采集



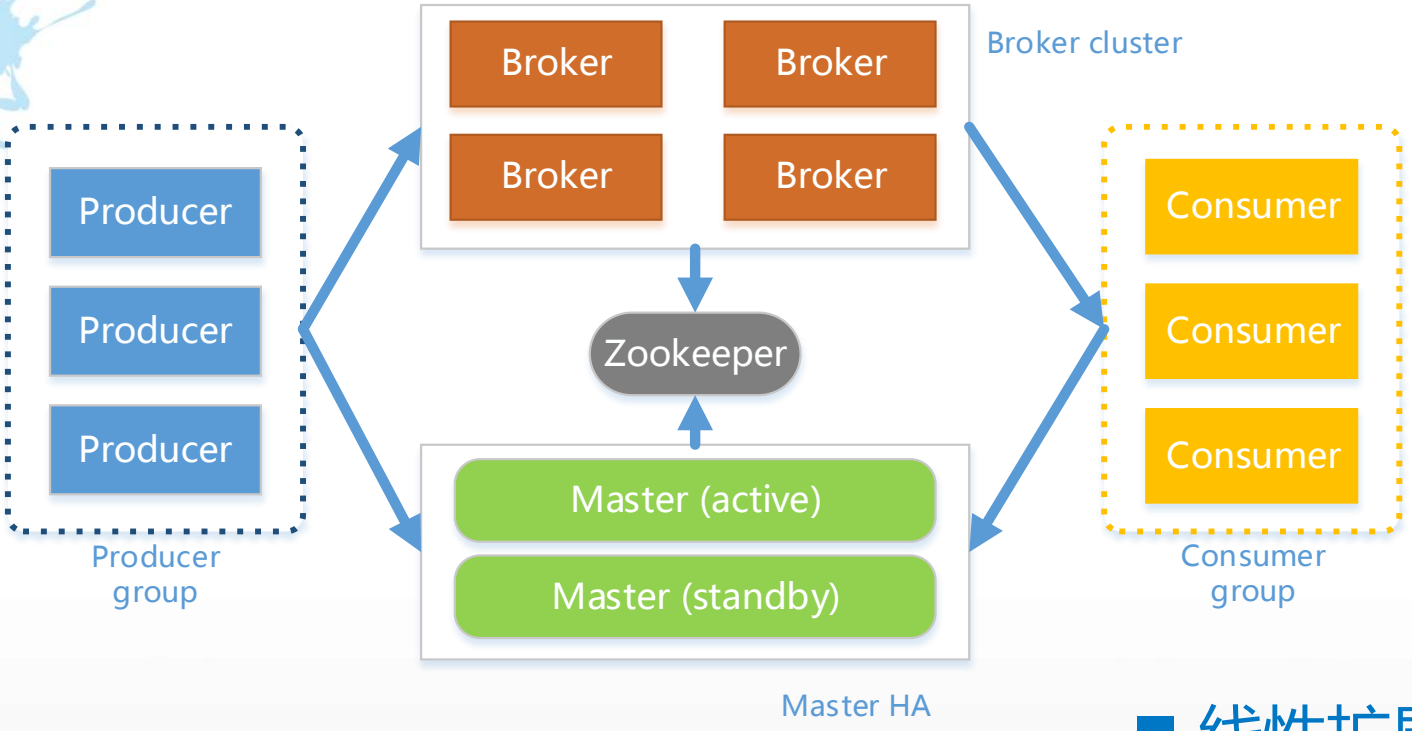
- 异构数据源适配
  - 直接消息
  - 转发消息
  - 本地文件
  - DB全量
  - DB增量
- AgentLoader
  - 统一管理
  - 批量部署

2. 数据接入



- 协议适配
- 高效网络传输
  - 公网/内网切换
  - 局部去重
  - 打包压缩加密
- 数据分散化

3. 数据缓存

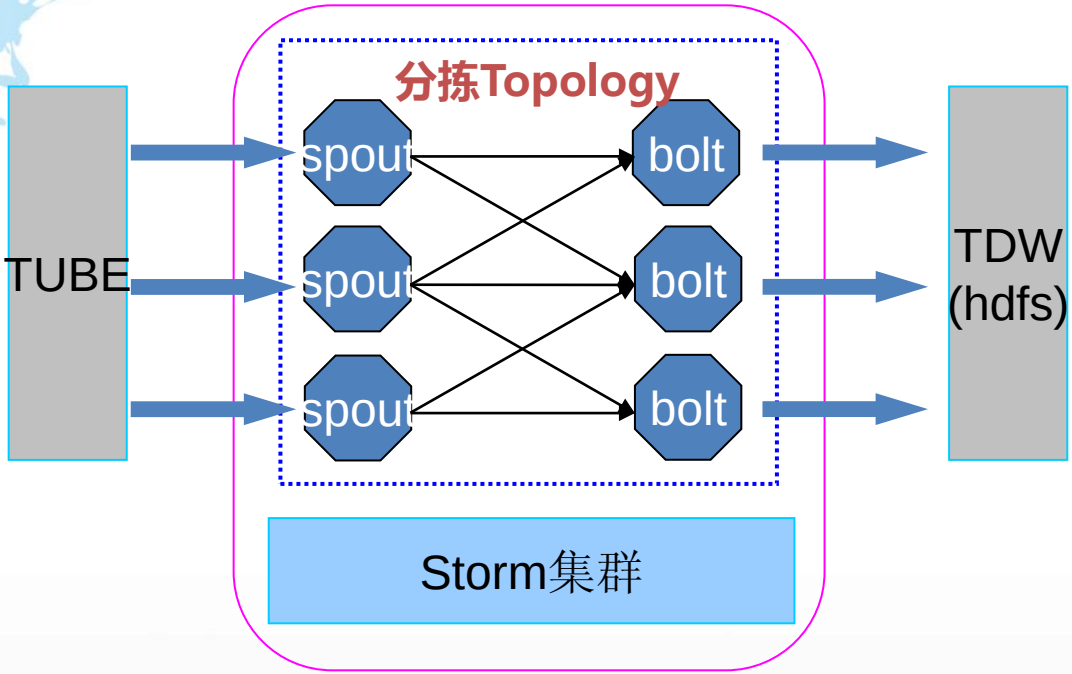


- 多路分发
- 数据回溯

分布式消息队列系统TUBE

- 线性扩展
- 数据持久化
- 高吞吐

4. 数据分拣

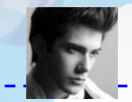


- 实时分拣
- 万级接口并发上传
- 容灾恢复
- 流量控制

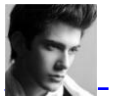
基于Storm流式计算平台

TDBank运营现状

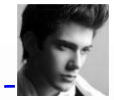
- 10,000亿　接入消息数/天
- 200TB　接入数据量/天
- 10,000个　并发分拣业务接口表/天
- 1-2s　采集平均延时
- 99.999%　可用度



爽是爽，不过最近有了新的麻烦



老板说了：
数据**报表** 要 分钟粒度
业务**监控** 要 秒级搞定
道具**推荐** 要 动态实时
系统异常 要 在线**分析**



要是能像TDW那样，一条SQL搞定，那就好了

土豪，有了TDBank，是不是很爽

神马情况，说说看呗？

你们老板胃口越来越大了，
那你想怎么办呢？

有见地，稍等一下，让我来试试看

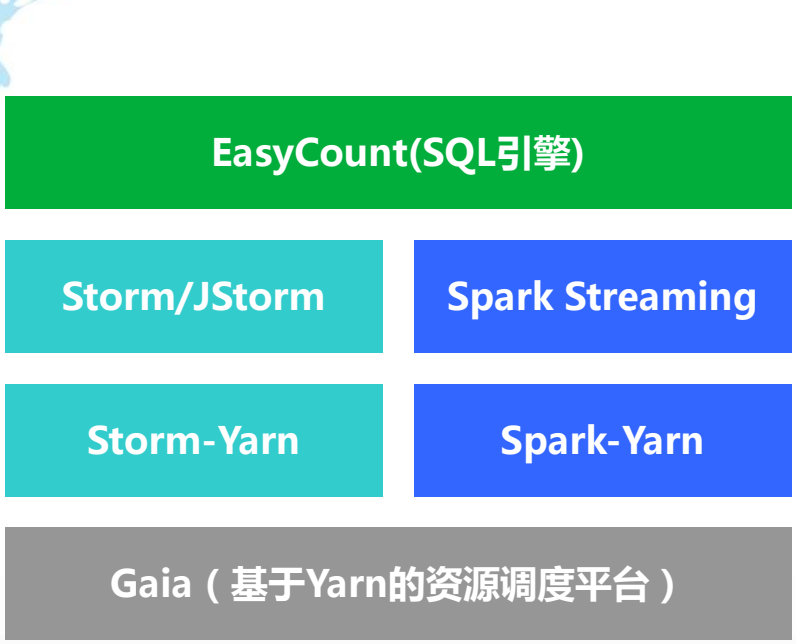
SQL与实时流式计算

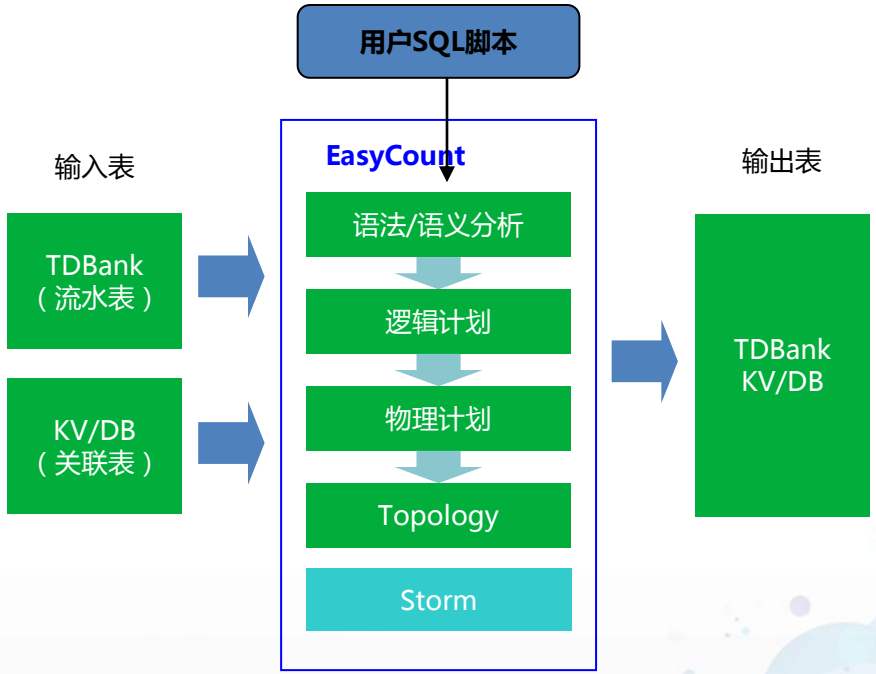
- SQL是数据的语言
- 流式SQL描述计算过程而非计算结果
- 流式计算中的数据存储方式灵活多样，无专用存储
- 一切数据都抽象为表，表的定义有两种：流水表，关联表

流式计算处理平台

- Storm
- SparkStreaming

EasyCount系统架构





- 支持with，where，groupby，union，join等基本sql语法
- 兼容HIVE所有的函数，聚合函数
- 支持复杂数据类型，map，array，struct

EasyCount-SQL几个重要问题

- 表的分类
- JOIN
- 聚合计算
- 去重统计
- 复杂逻辑计算

表的分类

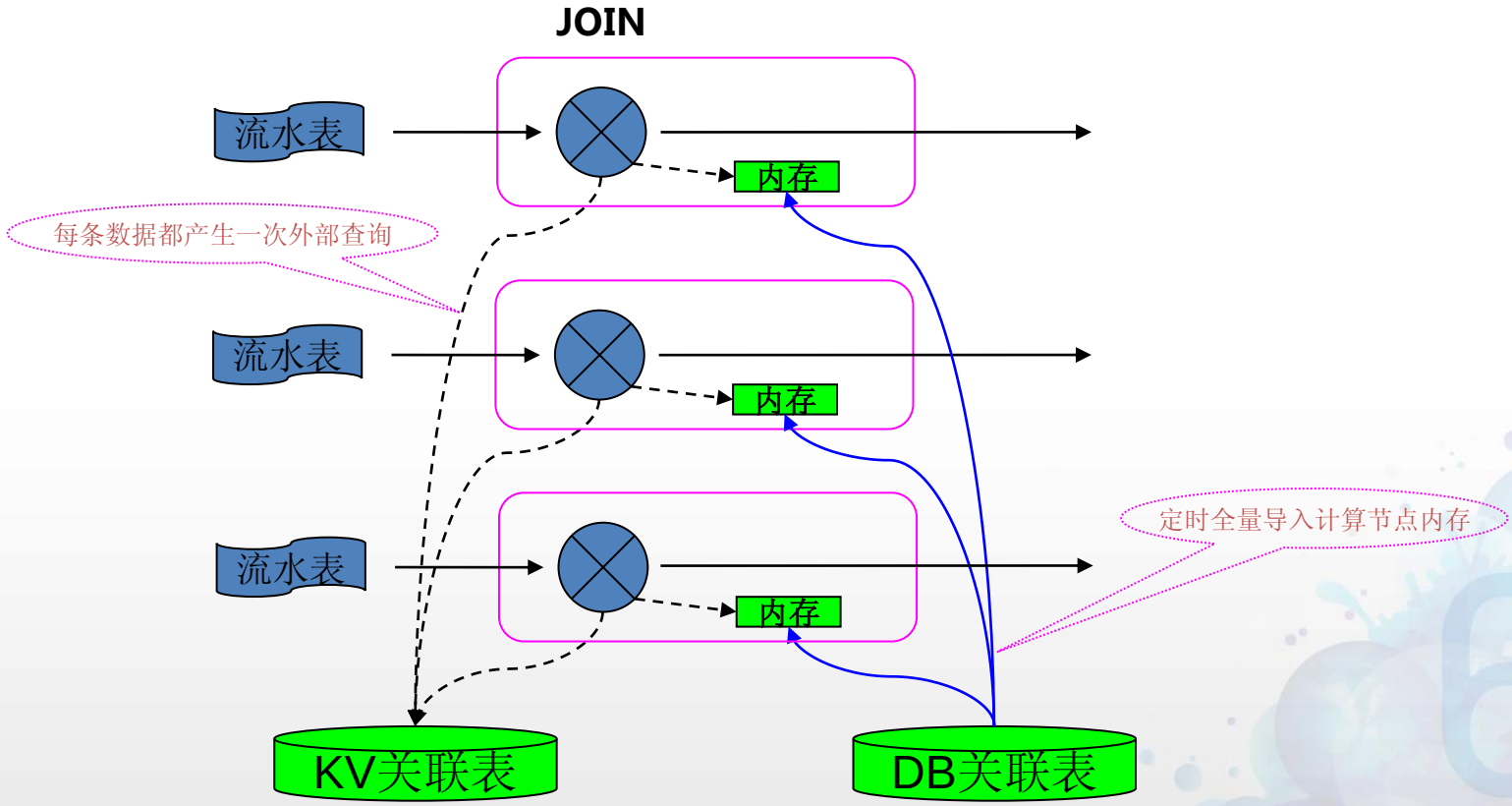
- **流水表**
  - TDBank流水数据表
- **关联表**
  - DB：mysql，pg
  - KV：tde，redis
  - HBase

JOIN

- JOIN的数学定义
  - 内关联，**左外**，右外，全外
  - 等值，非等值
- EasyCount支持：**左外等值关联**
  - 左表：流水表
  - 右表：关联表
- 流水表之间的关联需求怎么办？
  - 将其中一张表转化为关联表
  - UNION + Group By

JOIN的两种方式

- KV关联表：数据较大或即时更新
- DB关联表：数据较小且允许延迟更新



聚合计算

- select count(uin) ... group by appid
- 数据时间
- 聚合窗口
- 累加窗口及滑动窗口
- 语法级别的支持
  - group by appid coordinate by dtime
  - with aggr interval 60 seconds
  - with accu/sw interval 180 seconds

去重统计

- count(distinct uin)
- 精确的去重统计资源消耗较大
- 两个问题
  - 除了count以外的去重统计是否有意义？
  - 精确的去重统计有多么重要？
- 基于HyperLogLog基数统计算法的非精确去重统计函数
  - countd，hllp，hllp_merge，hllp_get
  - 1亿对象，99.5%精确度，45KB内存占用

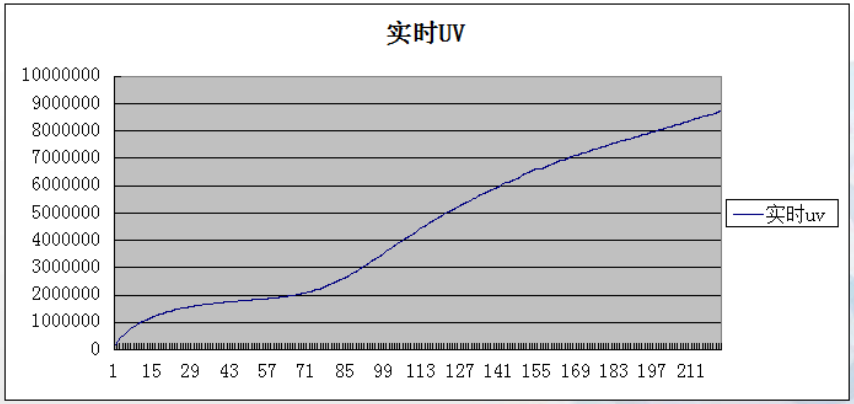
复杂逻辑计算

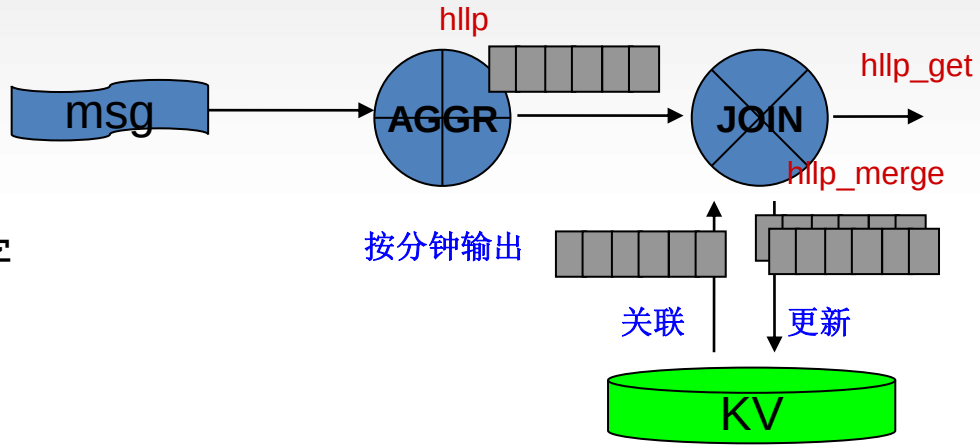
- SQL表达式
  - select expr from ....
- 复杂数据类型的处理
  - 循环，判断等
- 自定义UDF？
  - 运营成本，代码可维护性，系统稳定性
- execute表达式，嵌入在sql内部的过程式处理

```
select execute
(
   DEFINE x as int, ....
   {
      FOR(
         condexpr
         x := expr
      )
      IF(
         condexpr
         x := expr
      )
      ....
   }
   EMIT $x
) from tbl
```

APP累加UV统计案例分析

- **背景及需求**：
  - 对**100个APP在10个地区**每分钟统计累加UV。
  - 累加UV：从**当天0点开始到当前分钟**的独立登录数
  - 假设每个APP在每个地区的天UV约为**1千万**





使用**KV作为中间状态**关联表，以聚合分组字段为key，以聚合基数数组为value

```
with (select appid, areaid, hllp(qq) b from tbl group by appid, areaid ... 60 seconds) t1 \
     (select appid, areaid, jkv.k, hllp_merge(t1.b, j.ball) ball, t1.g g from t1 left join jkv \
                                   on concat(appid, areaid)=jkv.k)) jt \
insert into jkv select k, ball from jt \
insert into dest select appid, areaid, hllp_get(ball) from jt
```

非精确UV统计（99.5%准确率）

KV内存占用：100*10*45KB=45MB

KV查询：100*10*1440=150万次，20次/s

EasyCount运营现状

- 单日输入数据量：2000亿，计算量5000亿
- 业务类型包括，报表，监控，推荐，分析等
- 覆盖几乎所有业务BG

土豪，今天下班很早啊！

是啊，有了EasyCount啥事都迅速搞定，老板好久没有找我的麻烦了

哈哈哈，那你看还有什么可以为您效劳的？

暂时没想到，工作太轻松了，人就有点空虚，对了，你们那边**还招人吗**？

招啊，招啊，你快来吧

QQ：364787069

email：tianwp10@qq.com

微博：steventian-腾讯

腾讯大数据

微信公众号：tencentbigdata

Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB